



Entwicklerhandbuch

Amazon Machine Learning



Version Latest

Copyright © 2022 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon Machine Learning: Entwicklerhandbuch

Copyright © 2022 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Die Marken und Handelsmarken von Amazon dürfen nicht in einer Weise in Verbindung mit nicht von Amazon stammenden Produkten oder Services verwendet werden, die geeignet ist, Kunden irrezuführen oder Amazon in irgendeiner Weise herabzusetzen oder zu diskreditieren. Alle anderen Handelsmarken, die nicht Eigentum von Amazon sind, gehören den jeweiligen Besitzern, die möglicherweise zu Amazon gehören oder nicht, mit Amazon verbunden sind oder von Amazon gesponsert werden.

Table of Contents

.....	ix
Was ist Amazon Machine Learning?	1
Amazon Machine Learning Learning-Schlüsselkonzepte	1
Datenquellen	1
ML-Modelle	4
Auswertungen	4
Stapelvoraussagen	6
Echtzeitvoraussagen	6
Zugriff auf Amazon Machine Learning	7
Regionen und Endpunkte	7
Preise für Amazon ML	8
Einschätzen von Stapelvoraussagekosten	8
Einschätzen von Echtzeit-Voraussagekosten	10
Machine Learning-Konzepte	12
Lösen von wirtschaftlichen Problemen mit Amazon Machine Learning	12
Einsatzgebiete von Machine Learning	13
Die Erstellung einer Machine Learning-Anwendung	14
Erarbeitung des Problems	14
Sammeln von Daten mit Bezeichnung	15
Analysieren Ihrer Daten	16
Funktionsverarbeitung	16
Aufteilung von Daten in Schulungs- und Evaluierungsdaten	18
Schulen des Modells	19
Evaluation der Modellrichtigkeit	22
Optimierung der Modellrichtigkeit	27
Mit dem Modell Voraussagenerstellen	28
Modelle auf neue Daten umschulen	29
Der Amazon Machine Learning Learning-Prozess	29
Einrichten von Amazon Machine Learning	32
Anmelden bei AWS	32
Tutorial: Verwenden von Amazon ML zum Voraussagen der Reaktionen auf ein Marketingangebot	33
Voraussetzung	33
Schritte	33

Schritt 1: Vorbereiten der Daten	34
Schritt 2: Erstellen einer Schulungsdatenquelle	36
Schritt 3: Erstellen eines ML-Modells	42
Schritt 4: Überprüfen der Voraussageleistung des ML-Modells und Festlegen eines Punktzellenwerts	43
Schritt 5: Verwenden des ML-Modells zum Generieren von Prognosen	47
Schritt 6: Bereinigen	54
Erstellen und Verwenden von Datenquellen	56
Erwähnungen zum Datenformat für Amazon ML	56
Attribute	57
Anforderungen an das Eingabedateiformat	57
Verwenden mehrerer Dateien als Dateneingabe für Amazon ML	58
Zeilenendezeichen im CSV-Format	59
Das Erstellen eines Datenschemas für Amazon ML	60
Beispielschema	60
Verwenden des Felds targetAttributeName	62
Verwenden des Felds rowID	62
Verwenden des Felds AttributeType	63
Bereitstellen eines Schemas für Amazon ML	65
Aufteilen Ihrer Daten	66
Vorabtrennung Ihrer Daten	67
Sequenzielle Aufteilung Ihrer Daten	67
Zufällige Aufteilung Ihrer Daten	68
Dateneinblicke	70
Beschreibende Statistiken	70
Zugriff auf Dateneinblicke in der Amazon ML-Konsole	71
Verwenden von Amazon S3 mit Amazon ML	80
Hochladen Ihrer Daten in Amazon S3	81
Berechtigungen	82
Erstellen einer Amazon ML-Datenquelle aus Daten in Amazon Redshift	82
Erforderliche Parameter für den Assistenten Datenquelle erstellen	83
Erstellen einer Datenquelle mit Amazon Redshift Data (Konsole)	87
Fehlersuche bei Amazon Redshift	91
Verwenden von Daten aus einer Amazon RDS-Datenbank zum Erstellen einer Amazon ML- Datenquelle	97
RDS-Datenbank-Instance-Kennung	98

MySQL-Datenbankname	98
Benutzeranmeldeinformationen für die Datenbank	99
Sicherheitsinformationen zur AWS Data Pipeline	99
Amazon RDS-Sicherheitsinformationen	100
MySQL-SQL-Abfragen	100
S3-Ausgabespeicherort	100
Schulung von ML-Modellen	101
ML-Modelltypen	101
Binäres Klassifizierungsmodell	101
Mehrklassen-Klassifizierungsmodell	102
Regressionsmodell	102
Schulungsprozess	102
Schulungsparameter	103
Maximale Modellgröße	103
Maximale Anzahl von Datendurchläufen	104
Art der Mischung von Schulungsdaten	105
Regularisationstyp und -umfang	106
Schulungsparameter: Typen und Standardwerte	107
Erstellen eines ML-Modells	108
Voraussetzungen	109
Erstellen eines ML-Modells mit Standardoptionen	109
Erstellen eines ML-Modells mit benutzerdefinierten Optionen	110
Datentransformationen für maschinelles Lernen	113
Bedeutung der Funktionstransformation	113
Funktionstransformation mit Datenrezepten	114
Referenz zum Rezeptformat	114
Gruppen	115
Zuweisungen	116
Outputs	116
Beispiel eines vollständigen Rezepts	119
Empfohlene Rezepte	120
Referenz zur Datentransformation	120
N-Gramm-Transformation	121
Orthogonal Sparse Bigram (OSB)-Transformation	122
Umwandlung in Kleinbuchstaben	123
Transformation zum Entfernen von Satzzeichen	123

Quartile-Binning-Transformation	124
Normierungstransformation	125
Kartesische Produkt-Transformation	125
Neuordnung von Daten	127
DataRearrangement-Parameter	128
Evaluation von ML-Modellen	132
Einblicke in ML-Modelle	133
Einblicke in binäre Modelle	133
Interpretieren der Voraussagen	133
Einblicke in Mehrklassen-Modelle	138
Interpretieren der Voraussagen	138
Regressionsmodell-Einblicke	140
Interpretieren der Voraussagen	140
Verhindern von Overfitting	142
Kreuzvalidierung	143
Anpassen Ihrer Modelle	145
Auswertungswarnungen	146
Generieren und Interpretieren von Voraussagen	148
Erstellen einer Stapelvoraussage	148
Erstellen einer Stapelvoraussage (Konsole)	149
Erstellen einer Stapelvoraussage (API)	149
Überprüfen von Stapelvoraussage-Metriken	150
Überprüfen von Stapelvoraussage-Metriken (Konsole)	151
Überprüfen von Stapelvoraussage-Metriken und -Details (API)	151
Lesen der Ausgangsdateien für Stapelvoraussagen	151
Suchen der Manifestdatei für Stapelvoraussagen	152
Lesen der Manifestdatei	152
Abrufen der Ausgangsdateien für Stapelvoraussagen	153
Den Inhalt von Stapelvoraussagedateien für ein binäres ML-Klassifikationsmodell interpretieren	153
Den Inhalt von Stapelvoraussagedateien für ein Multiclass-ML-Klassifikationsmodell interpretieren	154
Den Inhalt von Stapelvoraussagedateien für ein Regressions-ML-Modell interpretieren	156
Anfordern von Echtzeitvoraussagen	156
Testen von Echtzeitvoraussagen	157
Erstellen eines Echtzeitendpunkts	159

Auffinden des Endpunkts für Echtzeitvoraussagen (Konsole)	161
Auffinden des Endpunkts für Echtzeitvoraussagen (API)	161
Erstellen einer Echtzeitvoraussage-Anforderung	162
Löschen eines Echtzeitendpunkts	164
Verwalten von Amazon ML -Objekten	166
Auflisten von Objekten	166
Auflisten von Objekten (Konsole)	167
Auflisten von Objekten (API)	168
Das Abrufen von Objektbeschreibungen	169
Detaillierte Beschreibungen in der Konsole	169
Detaillierte Beschreibungen von der API	169
Aktualisieren von Objekten	170
Löschen von Objekten	170
Löschen von Objekten (Konsole)	171
Löschen von Objekten (API)	172
Überwachen von Amazon ML mit Amazon CloudWatch-Metriken	173
Protokollieren von Amazon-ML-API-Aufrufen mit AWS CloudTrail	174
Amazon-ML-Informationen in CloudTrail	174
Beispiel: Amazon-ML-Protokolldateieinträge	176
Markieren von -Objekten	180
Grundlagen zu Tags	180
Tag-Einschränkungen	181
Markieren von Amazon ML -Objekten (Konsole)	182
Markieren von Amazon ML-Objekten (API)	184
Amazon Machine Learning-Referenz	185
Gewähren von Amazon ML-Berechtigungen zum Lesen von Daten aus Amazon S3	185
Gewähren von Berechtigungen für Amazon ML zwecks Ausgabe von Voraussagen in Amazon S3	187
Steuern des Zugriffs auf Amazon ML-Ressourcen – mit IAM	189
IAM-Richtliniensyntax	190
Festlegen von IAM-Richtlinienaktionen für Amazon ML	191
Angaben von ARNs für Amazon ML-Ressourcen in IAM-Richtlinien	192
Beispielhafte Richtlinien für Amazon	192
Dienstübergreifende Confused-Deputy-Prävention	196
Dependency Management von asynchrone Operationen	197
Das Überprüfen des Status einer Anfrage	198

Systemeinschränkungen	199
Namen und IDs für alle Objekte	201
Objektlebensdauer	201
Ressourcen	203
Dokumentverlauf	204

Wir aktualisieren den Amazon Machine Learning Learning-Service nicht mehr und akzeptieren keine neuen Benutzer mehr dafür. Diese Dokumentation ist für bestehende Benutzer verfügbar, wir aktualisieren sie jedoch nicht mehr. Weitere Informationen finden Sie unter [Was Amazon Machine Learning](#).

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.

Was ist Amazon Machine Learning?

Wir aktualisieren den Amazon Machine Learning (Amazon ML) Service nicht mehr oder akzeptieren neue Benutzer dafür. Diese Dokumentation steht bestehenden Benutzern zur Verfügung, aktualisieren sie jedoch nicht mehr.

AWS Amazon SageMaker bietet jetzt einen robusten, cloudbasierten Service - Amazon SageMaker, damit Entwickler aller Qualifikationsstufen Machine Learning-Technologie nutzen können. SageMaker ist ein vollständig verwalteter Machine Learning-Service, der Ihnen hilft, leistungsstarke Machine-Learning-Modelle Mit SageMaker können Datenwissenschaftler und Entwickler Machine-Learning-Modelle erstellen und trainieren und diese dann direkt in einer produktionsbereiten gehosteten Umgebung bereitstellen.

Weitere Informationen finden Sie im [.SageMaker-Dokumentation](#) aus.

Themen

- [Amazon Machine Learning Learning-Schlüsselkonzepte](#)
- [Zugriff auf Amazon Machine Learning](#)
- [Regionen und Endpunkte](#)
- [Preise für Amazon ML](#)

Amazon Machine Learning Learning-Schlüsselkonzepte

Dieser Abschnitt bietet eine Übersicht über die folgenden wesentlichen Konzepte und beschreibt ausführlich, wie sie innerhalb von Amazon ML verwendet werden:

- [Datenquellen](#) enthalten Metadaten, die mit Dateneingaben in Amazon ML verknüpft sind
- [ML-Modelle](#) generieren Voraussagen mithilfe der aus den Eingabedaten extrahierten Muster
- [Auswertungen](#) messen die Qualität von ML-Modellen
- [Stapelvoraussagen](#) generieren Voraussagen asynchron für mehrere Eingabedatenbeobachtungen
- [Echtzeitvoraussagen](#) generieren Voraussagen synchron für einzelne Datenbeobachtungen

Datenquellen

Eine Datenquelle ist ein Objekt, das Metadaten über Ihre Eingabedaten enthält. Amazon ML liest die Eingabedaten, berechnet beschreibende Statistiken zu deren Attributen und speichert die Statistiken

zusammen mit einem Schema und weiteren Informationen als Bestandteil des Datenquellenobjekts. Datenquellen werden von Amazon ML anschließend für das Lernen und die Auswertung eines maschinellen Lernmodells sowie für die Generierung von Stapelvoraussagen genutzt.

Important

Eine Datenquelle speichert keine Kopie Ihrer Eingabedaten. Stattdessen wird ein Verweis auf den Speicherort in Amazon S3 gespeichert, an dem sich Ihre Eingabedaten befinden. Wenn Sie die Amazon S3 S3-Datei verschieben oder ändern, kann Amazon ML nicht auf sie zugreifen oder sie verwenden, um ein ML-Modell zu erstellen, Auswertungen zu generieren oder Voraussagen zu machen.

In der folgenden Tabelle sind Bedingungen definiert, die im Zusammenhang mit Datenquellen stehen.

Laufzeit	Definition
Attribut	<p>Eine eindeutige und benannte Eigenschaft innerhalb einer Beobachtung. In tabellarischen Daten (z. B. Kalkulationstabellen oder Dateien im CSV-Format (durch Komma getrennte Werte)) stellen die Spaltenüberschriften die Attribute dar, in den Zeilen sind Werte für diese Attribute enthalten.</p> <p>Synonyme: Variable, Variablenname, Feld, Spalte</p>
Datenquellenname	<p>(Optional) Sie können einen lesbaren Namen für eine Datenquelle definieren. Anhand dieser Namen können Sie Ihre Datenquellen in der Amazon ML-Konsole suchen und verwalten.</p>
Eingabedaten	<p>Sammelbezeichnung für alle Beobachtungen, auf die von einer Datenquelle verwiesen wird.</p>
Ort	<p>Speicherort der Eingabedaten. Derzeit kann Amazon ML Daten verwenden, die innerhalb von Amazon S3 S3-Buckets, Amazon Redshift Redshift-Datenbanken oder MySQL-Datenbanken in Amazon Relational Database Service (RDS) gespeichert sind.</p>
Beobachtung	<p>Eine einzelne Einheit von Eingabedaten. Wenn Sie beispielsweise ein ML-Modell erstellen, um betrügerische Transaktionen zu ermitteln, bestehen</p>

Laufzeit	Definition
	<p>Ihre Eingabedaten aus vielen Beobachtungen, von denen jede eine einzelne Transaktion darstellt.</p> <p>Synonyme: Datensatz, Beispiel, Instanz, Zeile</p>
Zeilen-ID	<p>(Optional) – Ein Flag, das, falls angegeben, ein Attribut in den Eingabedaten identifiziert, das in das Voraussageergebnis eingeschlossen werden soll. Anhand dieses Attributs kann einfacher zugeordnet werden, welche Voraussage welcher Beobachtung entspricht.</p> <p>Synonyme: Zeilen-ID</p>
Schema	<p>Die Informationen, die zur Deutung der Eingabedaten benötigt werden, einschließlich Attributnamen und ihre zugeordneten Datentypen sowie die Namen besonderer Attribute.</p>
Statistiken	<p>Zusammenfassende Statistik für jedes Attribut in den Eingabedaten. Diese Statistiken dienen zwei Zwecken:</p> <p>Die Amazon ML-Konsole zeigt sie in Diagrammen an, damit Sie die Daten auf einen Blick besser verstehen und Unregelmäßigkeiten oder Fehler erkennen können.</p> <p>Amazon ML verwendet sie während des Schulungsprozesses, um die Qualität des resultierenden ML-Modells zu verbessern.</p>
Status	<p>Gibt den aktuellen Status der Datenquelle an, beispielsweise Laufend, Abgeschlossen oder Fehlgeschlagen.</p>
Zielattribut	<p>Im Kontext der Schulung eines ML-Modells identifiziert das Zielattribut den Namen des Attributs in den Eingabedaten, welche die "richtigen" Antworten enthalten. Amazon ML verwendet diese, um Muster in den Eingabedaten zu entdecken und ein ML-Modell zu generieren. Im Kontext des Auswertens und Generierens von Voraussagen, ist das Zielattribut das Attribut, dessen Wert vorhergesagt von einem qualifizierten ML-Modell vorhergesagt wird.</p> <p>Synonyme: Ziel</p>

ML-Modelle

Ein ML-Modell ist ein mathematisches Modell, bei dem Voraussagen durch die Ermittlung von Mustern in Ihren Daten generiert werden. Amazon ML unterstützt drei ML-Modelltypen: binäre Klassifizierung, Mehrklassen-Klassifizierung und Regression.

In der folgenden Tabelle sind Begriffe definierte, die im Zusammenhang mit ML-Modellen stehen.

Laufzeit	Definition
Regression	Das Ziel der Schulung eines Regressions-ML-Modells besteht darin, einen numerischen Wert vorherzusagen.
Mehrklassen	Das Ziel der Schulung eines Mehrklassen-ML-Modells besteht darin, Werte vorherzusagen, die zu einem begrenzten und vordefinierten Satz an zulässigen Werten gehören.
Binary	Das Ziel der Schulung eines Binär-ML-Modells besteht darin, Werte vorherzusagen, die nur einen von zwei Status aufweisen können, z. B. "true" oder "false".
Modellgröße	ML-Modelle erfassen und speichern Muster. Je mehr Muster in einem ML-Modell gespeichert sind, desto größer ist es. Die ML-Modellgröße wird in MB beschrieben.
Anzahl der Durchläufe	Wenn Sie ein ML-Modell schulen, verwenden Sie Daten aus einer Datenquelle. Es ist manchmal von Vorteil, jeden Datensatz im Lernprozess mehrmals zu verwenden. Die Anzahl von Malen, die Sie Amazon ML dieselben Datensätze verwenden lassen, wird als die Anzahl von Durchläufen bezeichnet.
Regularisation	Die Regularisation ist eine Technik des maschinellen Lernens (ML), die verwendet werden kann, um ML-Modelle mit höherer Qualität zu erhalten. Amazon ML bietet eine Standardeinstellung, die in den meisten Fällen funktioniert.

Auswertungen

Eine Auswertung misst die Qualität Ihres ML-Modells und bestimmt, ob es gute Leistungen bringt.

In der folgenden Tabelle sind Begriffe im Zusammenhang mit Auswertungen definiert.

Laufzeit	Definition
Einblicke in Modelle	Amazon ML stellt Ihnen eine Metrik und eine Reihe von Erkenntnissen bereit, die Sie verwenden können, um die prädiktive Leistung Ihres Modells zu bewerten.
AUC	AUC (Area Under the ROC Curve) misst die Fähigkeit eines binären ML-Modells, eine höhere Bewertung für positive Beispiele im Vergleich zu negativen Beispielen vorherzusagen.
F1-Bewertung mit Makro-Durchschnitt	Die F1-Bewertung mit Makro-Durchschnitt wird zum Auswerten der prädiktiven Leistung von Mehrklassen-ML-Modellen verwendet.
RMSE	Der Root Mean Square Error (RMSE) ist eine Metrik zur Bewertung der prädiktiven Leistung von Regressions-ML-Modellen.
Grenzwert	ML-Modelle arbeiten durch Generierung von numerischen Voraussageergebnissen. Durch Anwenden eines Grenzwerts konvertiert das System diese Werte in 0- und 1-Bezeichnungen.
Accuracy	Die Richtigkeit misst den Anteil der richtigen Voraussagen.
Genauigkeit	„Precision“ zeigt den Prozentsatz der tatsächlichen positiven Instances (im Gegensatz zu Fehlalarmen) unter den Instances an, die abgerufen wurden (diejenigen, die als positiv vorausgesagt wurden). Mit anderen Worten: Wie viele ausgewählte Elemente sind positiv?
Wiedererkennung	„Recall“ zeigt den Prozentsatz der tatsächlichen positiven Instances in der Gesamtanzahl der betreffenden Instances an (tatsächliche positive Instances). Mit anderen Worten: Wie viele positive Elemente sind ausgewählt?

Stapelvoraussagen

Stapelvoraussagen werden für eine Reihe von Beobachtungen verwendet, die alle gleichzeitig ausgeführt werden können. Diese Lösung eignet sich optimal für prädiktive Analysen, die keine Echtzeitanforderung aufweisen.

In der folgenden Tabelle sind Begriffe im Zusammenhang mit Stapelvoraussagen definiert.

Laufzeit	Definition
Ausgabespeicherort	Die Ergebnisse einer Stapelvoraussage werden in einem S3-Bucket-Ausgabespeicherort gespeichert.
Manifestdatei	Diese Datei verknüpft die Eingabedatendatei mit den zugehörigen Ergebnissen der Stapelvoraussage. Sie wird am S3-Ausgabespeicherort gespeichert.

Echtzeitvoraussagen

Echtzeitvoraussagen werden für Anwendungen mit geringer Latenzanforderung verwendet, z. B. interaktive Webanwendungen, mobile Anwendungen oder Desktopanwendungen. Jedes ML-Modell kann im Hinblick auf Voraussagen mithilfe der latenzarmen Echtzeitvoraussage-API abgefragt werden.

In der folgenden Tabelle sind Begriffe im Zusammenhang mit Echtzeitvoraussagen definiert.

Laufzeit	Definition
Echtzeitvoraussage-API	Die Echtzeitvoraussage-API akzeptiert eine einzelne Eingabebeobachtung in der Nutzlast der Anforderung und gibt die Voraussage synchron in der Antwort zurück.
Endpunkt für Echtzeitvoraussagen	Um ein ML-Modell mit einer Echtzeitvoraussage-API zu verwenden, müssen Sie einen Endpunkt für Echtzeitvoraussagen erstellen. Nach der Erstellung enthält der Endpunkt die URL, die Sie verwenden können, um Echtzeitvoraussagen anzufordern.

Zugriff auf Amazon Machine Learning

Sie können mit einer der folgenden Möglichkeiten auf Amazon ML zugreifen:

Amazon ML Konsole

Sie können mit der AWS-Managementkonsole auf zugreifen, indem Sie sich bei der AWS-Managementkonsole anmelden und unter <https://console.aws.amazon.com/machinelearning/> aus.

AWS-CLI

Informationen zur Installation und Konfiguration der AWS-CLI finden Sie unter Einrichtung der AWS-Befehlszeilenschnittstelle im [AWS Command Line Interface-Benutzerhandbuch](#) aus.

Amazon ML API

Weitere Informationen zur Amazon ML API finden Sie unter [Amazon ML API-Referenz](#) aus.

AWS-SDKs

Weitere Informationen über AWS SDKs finden Sie unter [Tools for Amazon Web Services](#).

Regionen und Endpunkte

Amazon Machine Learning (Amazon ML) unterstützt Echtzeitprognose-Endpunkte in den folgenden zwei Regionen:

Name der Region	Region	Endpunkt	Protocol (Protokoll)
US East (N. Virginia)	us-east-1	machinelearning.us-east-1.amazonaws.com	HTTPS
Europa (Irland)	eu-west-1	machinelearning.eu-west-1.amazonaws.com	HTTPS

Sie können Datensätze hosten, Modelle schulen und auswerten und Voraussagen in einer beliebigen Region auslösen.

Wir empfehlen, dass Sie alle Ressourcen in derselben Region belassen. Wenn sich Ihre Eingabedaten in einer anderen Region als Ihre Amazon ML-Ressourcen befinden, fallen regionsübergreifende Datenübertragungsgebühren an. Sie können einen Echtzeitvoraussage-Endpunkt aus einer beliebigen Region aufrufen, das Aufrufen eines Endpunkts aus einer Region, die nicht über den Endpunkt verfügt, den Sie aufrufen, kann zu Latenzen bei der Echtzeitvoraussage führen.

Preise für Amazon ML

Bei AWS Services bezahlen Sie nur für das, was Sie tatsächlich nutzen. Es fallen keine Mindestgebühren oder Vorausleistungen an.

Amazon Machine Learning (Amazon ML) berechnet eine Gebühr auf Stundenbasis für die Zeit, die benötigt wird, um die Datenstatistiken zu berechnen und Modelle zu schulen und zu bewerten; zudem bezahlen Sie für die Anzahl Voraussagen, die für Ihre Anwendung erzeugt wurden. Für Echtzeit-Voraussagen können Sie auch für eine reservierte Kapazität bezahlen, deren Größe sich nach der Größe Ihres Modells richtet.

Amazon ML schätzt die Kosten für Voraussagen nur in der [Amazon ML -Konsole](#) aus.

Weitere Informationen zu Amazon ML finden Sie unter [Amazon Machine Learning Learning-Preise](#) aus.

Themen

- [Einschätzen von Stapelvoraussagekosten](#)
- [Einschätzen von Echtzeit-Voraussagekosten](#)

Einschätzen von Stapelvoraussagekosten

Wenn Sie Batch-Voraussagen von einem Amazon ML-Modell mithilfe des Create Batch Prediction-Assistenten abrufen, schätzt Amazon ML die Kosten für diese Voraussagen. Die Methode zum Berechnen der Schätzung variiert je nach Art der verfügbaren Daten.

Einschätzen der Kosten für eine Stapelvoraussage mit verfügbaren Datenstatistiken

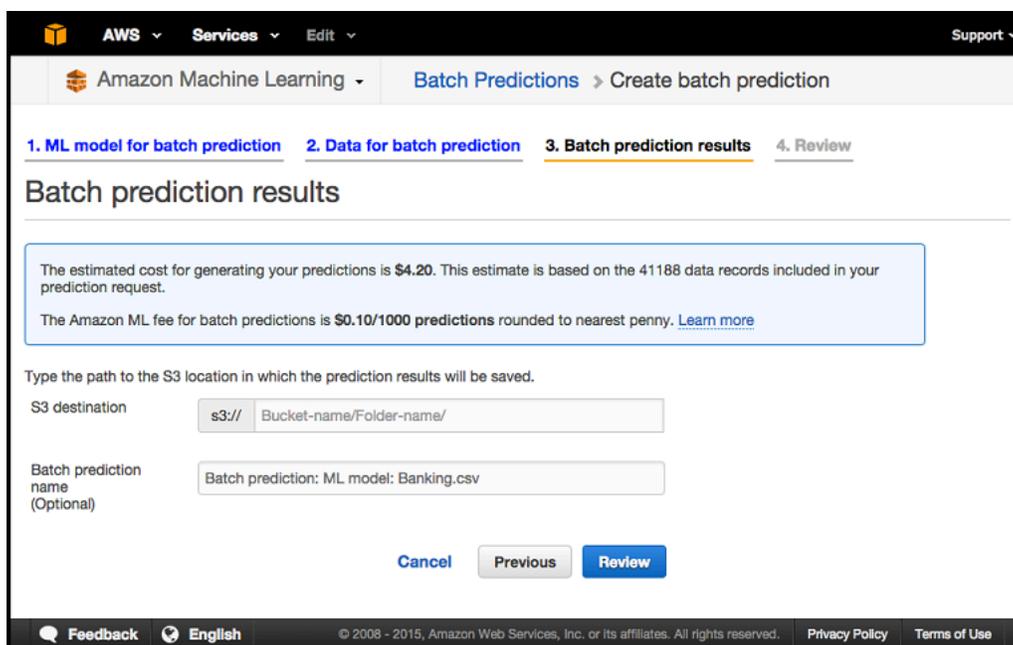
Die genaueste Kostenschätzung erhalten Sie, wenn Amazon ML bereits zusammenfassende Statistiken zur verwendeten Datenquelle berechnet hat, die verwendet wird, um Voraussagen abzufragen. Diese Statistiken werden immer für Datenquellen berechnet, die mit der Amazon

ML-Konsole erstellt wurden. API-Benutzer müssen die Markierung `ComputeStatistics` auf `True` setzen, wenn sie Datenquellen mit den APIs [CreateDataSourceFromS3](#), [CreateDataSourceFromRedshift](#) oder [CreateDataSourceFromRDS](#) erstellen. Die Datenquelle muss den Status `READY` aufweisen, damit die Statistiken zur Verfügung stehen.

Eine der Statistiken, die Amazon ML berechnet, ist die Anzahl der Datensätze. Wenn die Anzahl der Datensätze verfügbar ist, schätzt der Amazon ML Create Batch Prediction-Assistent die Anzahl der Voraussagen, indem er die Anzahl der Datensätze mit der [Gebühr für Chargenprognosen](#) aus.

Ihre tatsächlichen Kosten können von dieser Schätzung aus folgenden Gründen abweichen:

- Einige der Datensätze werden nicht verarbeitet. Datensätze von nicht verarbeiteten Voraussagen werden nicht in Rechnung gestellt.
- Die Schätzung berücksichtigt keine Gutschriften oder andere Anpassungen, die von AWS angewendet werden.



The screenshot shows the 'Batch prediction results' page in the Amazon Machine Learning console. The page is divided into four steps: 1. ML model for batch prediction, 2. Data for batch prediction, 3. Batch prediction results (highlighted), and 4. Review. The main content area displays the estimated cost for generating predictions as \$4.20, based on 41188 data records. Below this, it states the Amazon ML fee for batch predictions is \$0.10/1000 predictions rounded to nearest penny. The user is prompted to enter the S3 destination path and the batch prediction name. The 'Review' button is highlighted.

Einschätzen der Kosten für eine Stapelvoraussage mit verfügbarer Datengröße

Wenn Sie die Batch-Voraussage anfordern und die Datenstatistiken für die angeforderte Datenquelle sind nicht verfügbar, schätzt Amazon ML die Kosten basierend auf Folgendem:

- Die Größe der gesamten Daten, die während der Datenquellenvalidierung berechnet und beibehalten werden.

- Die durchschnittliche Datensatzgröße, die Amazon ML durch das Lesen und Parsen der ersten 100 MB Ihrer Datendatei schätzt.

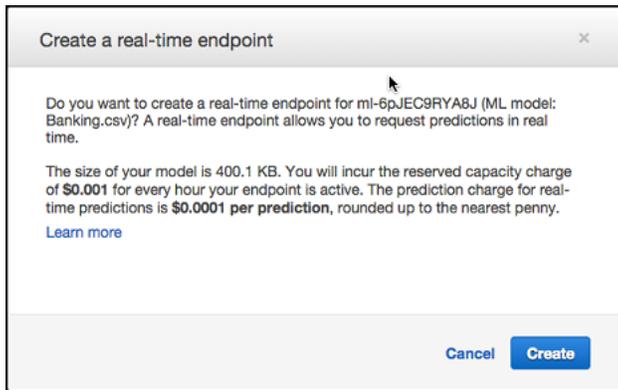
Um die Kosten Ihrer Stapelvoraussage zu schätzen, teilt Amazon ML die Gesamtdatengröße durch die durchschnittliche Datensatzgröße. Diese Methode der Kostenvoraussage ist weniger genau als die Methode, die verwendet wird, wenn die Anzahl der Datensätze verfügbar ist, da die ersten Datensätze der Datendatei die durchschnittliche Datensatzgröße möglicherweise nicht genau wiedergeben.

Einschätzen der Kosten für eine Stapelvoraussage ohne Datenstatistiken und Datengröße

Wenn weder Datenstatistiken noch Datengröße verfügbar sind, kann Amazon ML die Kosten für Ihre Stapelvoraussagen nicht einschätzen. Dies ist häufig der Fall, wenn die Datenquelle, die Sie verwenden, um Stapelvoraussagen anzufordern, noch nicht von Amazon ML validiert wurde. Dies kann der Fall sein, wenn Sie eine Datenquelle erstellt haben, die auf einer Amazon Redshift -Abfrage (Amazon Redshift) oder Amazon Relational Database Service (Amazon RDS) basiert und die Datenübertragung noch nicht abgeschlossen ist, oder wenn sich die Erstellung der Datenquelle in einer Warteschlange hinter anderen Operationen auf Ihrem Konto befindet. In diesem Fall informiert Sie die Amazon ML -Konsole über die Gebühren für eine Stapelvoraussage. Sie können dann entweder ohne Einschätzung mit der Anfrage der Stapelvoraussage fortfahren, oder Sie können den Assistenten beenden und zurückkehren, nachdem die für Voraussagen verwendete Datenquelle den Status INPROGRESS oder READY aufweist.

Einschätzen von Echtzeit-Voraussagekosten

Beim Erstellen eines Echtzeit-Voraussageendpunkts mit der Amazon ML -Konsole werden Ihnen die geschätzten Gebühren für die reservierte Kapazität angezeigt; dies ist eine laufende Gebühr für die Reservierung eines Endpunkts für die Verarbeitung einer Voraussage. Diese Gebühr variiert je nach Größe des Modells, wie auf der [Service-Seite mit den Preisen](#) erläutert. Zudem werden Sie über die Gebühr für eine standardmäßige Amazon ML -Echtzeit-Voraussage informiert.



Machine Learning-Konzepte

Mittels Machine learning (ML) können Sie archivierte Daten verwenden, um bessere Business-Entscheidungen zu treffen. ML-Algorithmen entdecken Muster in Daten und konstruieren anhand dieser Entdeckungen mathematische Modelle. Anschließend können Sie die Modelle verwenden, um Voraussagen für zukünftige Daten zu erstellen. Eine mögliche Anwendung des Machine Learning-Modells ist beispielsweise die Voraussage der Wahrscheinlichkeit, mit der ein Kunde auf der Grundlage seines vergangenen Verhaltens ein bestimmtes Produkt kaufen wird.

Themen

- [Lösen von wirtschaftlichen Problemen mit Amazon Machine Learning](#)
- [Einsatzgebiete von Machine Learning](#)
- [Die Erstellung einer Machine Learning-Anwendung](#)
- [Der Amazon Machine Learning Learning-Prozess](#)

Lösen von wirtschaftlichen Problemen mit Amazon Machine Learning

Mit Amazon Machine Learning können Sie maschinelles Lernen auf Probleme anwenden, für die Sie bereits Beispiele mit tatsächlichen Antworten haben. Wenn Sie beispielsweise Amazon Machine Learning verwenden möchten, um vorherzusagen, ob eine E-Mail Spam ist, müssen Sie E-Mail-Beispiele sammeln, die korrekt als Spam-Nachrichten oder kein Spam gekennzeichnet sind. Sie können dann diese E-Mail-Beispiele für das maschinelle Lernen nutzen, um vorherzusagen, wie wahrscheinlich eine neue E-Mail Spam ist oder nicht. Dieser Ansatz zum Lernen aus Daten, die mit der tatsächlichen Antwort markiert sind, wird als überwacht maschinelles Lernen bezeichnet.

Sie können den überwachten ML-Ansatz für folgende spezifischen Machine Learning-Aufgaben verwenden: binäre Klassifizierung (Vorhersage von einem aus zwei möglichen Ergebnissen), Mehrklassen-Klassifizierung (Vorhersage von einem aus mehr als zwei Ergebnissen) und Regression (Vorhersage eines numerischen Werts).

Beispiele für binäre Klassifizierungsprobleme:

- Wird der Kunde das Produkt kaufen oder nicht?
- Ist diese E-Mail Spam oder nicht?

- Ist Ihr Produkt ein Buch oder ein Nutztier?
- Wurde diese Bewertung von einem Kunden oder einer Maschine geschrieben?

Beispiele für Mehrklassen-Klassifizierungsprobleme:

- Ist das Produkt ein Buch, ein Film oder Kleidung?
- Ist dieser Film ein Liebeskomödie, eine Dokumentation oder ein Thriller?
- Welche Kategorie von Produkten für diesen Kunden am interessantesten?

Beispiele für Regressions-Klassifizierungsprobleme:

- Wie wird die Temperatur in Seattle morgen sein?
- Wie viele Einheiten dieses Produkts werden wir verkaufen?
- Wie viele Tage dauert es, bis der Kunde die Anwendung nicht mehr verwendet?
- Für welchen Preis wird dieses Haus verkauft?

Einsatzgebiete von Machine Learning

Beachten Sie, dass ML nicht für jede Art von Problemen eine Lösung ist. Es gibt bestimmte Fälle, in denen robuste Lösungen ohne ML-Techniken entwickelt werden können. Beispielsweise benötigen Sie kein ML, wenn Sie einen Zielwert bestimmen können, indem Sie einfache Regeln, Berechnungen oder vorbestimmte Schritte anwenden, die ohne datengesteuertes Lernen programmiert werden können.

Verwenden Sie das maschinelle Lernen für folgende Situationen:

- Sie können die Regeln nicht codieren: Viele Aufgaben, die von Menschen durchgeführt werden (z. B. das Erkennen von Spam-Nachrichten) können mit einer einfachen (deterministischen), regelbasierten Lösung nicht ausreichend gelöst werden. Eine große Anzahl von Faktoren beeinflusst das Ergebnis. Wenn zu viele Regeln von zu vielen Faktoren abhängen und sich viele dieser Regeln überlappen oder sehr fein abgestimmt werden müssen, wird es schnell schwierig, diese Regeln präzise zu codieren. Mit ML können Sie dieses Problem effektiv lösen.
- Sie können nicht skalieren: Sie können manuell ein paar Hundert E-Mails erkennen und entscheiden, ob sie Spam sind oder nicht. Doch diese Aufgabe wird bei Millionen von E-Mails zu aufwendig. ML-Lösungen sind für die Verarbeitung von Problemen in großem Umfang effektiv.

Die Erstellung einer Machine Learning-Anwendung

Die Erstellung von ML-Anwendungen ist ein iterativer Prozess mit mehreren Schritten. Um eine ML-Anwendung zu erstellen, führen Sie die folgenden Schritte durch:

1. Stellen Sie auf der Grundlage Ihrer Beobachtungen das/die wichtigste(n) ML-Problem(e) heraus und welche Antwort das Modell voraussagen soll.
2. Erheben und bereinigen sie Daten und bereiten Sie diese auf, sodass sie von Schulungsalgorithmen für das ML-Modell verwendet werden können. Visualisieren und analysieren Sie die Daten, und führen Sie Kontrollprüfungen durch, um die Qualität der Daten sicherzustellen, und um die Daten zu verstehen.
3. Oftmals werden die Rohdaten (Eingabevariablen) und die Antwort (Ziel) nicht so dargestellt, dass sie zur Schulung eines Voraussagemodells verwendet werden können. Daher sollten Sie versuchen, mehr Eingaben oder Funktionen aus den Rohvariablen mit voraussagendem Charakter zu erstellen.
4. Geben Sie die daraus hervorgehenden Funktionen in den Lernalgorithmus ein, um Modelle zu erstellen und die Qualität der Modelle anhand der Daten auszuwerten, die nicht für die Erstellung des Modells verwendet wurden.
5. Verwenden Sie das Modell zum Generieren von Voraussagen der Zielantwort für neue Daten-Instances.

Erarbeitung des Problems

Der erste Schritt im Machine Learning besteht darin, zu entscheiden, was Sie voraussagen möchten; dies ist das Label oder die Zielantwort. Stellen Sie sich vor, Sie möchten Produkte herstellen, aber die Entscheidung für die Herstellung eines Produkts hängt von der Anzahl der Verkaufschancen ab. In diesem Szenario möchten Sie voraussagen, wie oft jedes Produkt erworben werden wird (Anzahl Verkäufe voraussagen). Es gibt mehrere Möglichkeiten, dieses Problem mittels Machine Learning zu definieren. Die Art und Weise, wie Sie das Problem definieren, hängt von Ihrem Anwendungsfall bzw. Ihren Geschäftsbedürfnissen ab.

Möchten Sie die Anzahl Käufe voraussagen, die Ihre Kunden für jedes Produkt tätigen werden (in diesem Fall ist das Ziel numerisch und Sie lösen ein Regressionsproblem)? Oder möchten Sie voraussagen, welche Produkte mehr als 10 Mal gekauft werden (in diesem Fall ist das Ziel binär und Sie lösen ein binäres Klassifikationsproblem)?

Es ist wichtig, das Problem nicht zu verkomplizieren und die einfachste Lösung zu erarbeiten, die auf Ihre Bedürfnisse zugeschnitten ist. Allerdings ist es ebenso wichtig, keine Informationen zu verlieren, insbesondere Informationen in den historischen Antworten. Durch das Konvertieren einer vergangenen Verkaufszahl in die binäre Variable "über 10" statt "weniger" würden wertvolle Informationen verloren gehen. Investieren Sie Zeit in Ihre Entscheidung für die Ziele, die für Ihre Voraussage am meisten Sinn machen, um Modelle zu erstellen, die Ihre Frage nicht beantworten.

Sammeln von Daten mit Bezeichnung

ML-Probleme starten mit den Daten – vorzugsweise viele Daten (Beispiele oder Beobachtungen), deren Zielantwort Ihnen bereits bekannt ist. Daten, deren Zielantwort Ihnen bereits bekannt ist, werden bezeichnete Daten genannt. Im überwachten ML lernt der Algorithmus selbst, wie er aus bezeichneten Beispielen, die wir bereitstellen, lernen muss.

Jede(s) Beispiel/Beobachtung in Ihren Daten muss zwei Elemente enthalten:

- Das Ziel – Die Antwort, die Sie voraussagen möchten. Sie stellen dem ML-Algorithmus zum Lernen Daten bereit, die mit dem Ziel (richtige Antwort) bezeichnet sind. Anschließend verwenden Sie das geschulte ML-Modell für Daten, deren Zielantwort Sie nicht kennen, um diese Antwort vorauszusagen.
- Variablen/Funktionen – Hierbei handelt es sich um Attribute des Beispiels, die verwendet werden können, um Muster zu erkennen und die Zielantwort vorauszusagen.

Beispielsweise ist beim E-Mail-Klassifizierungsproblem das Ziel eine Bezeichnung, die angibt, ob eine E-Mail Spam ist oder nicht. Beispiele für Variablen sind der Absender der E-Mail, der Text im Textkörper der E-Mail, der Text in der Betreff-Zeile, der Zeitpunkt, zu dem die E-Mail gesendet wurde, und vorangegangene Korrespondenz zwischen Sender und Empfänger.

Häufig stehen die Daten nicht als bezeichnete Daten zur Verfügung. Das Sammeln und Vorbereiten von Variablen und Ziel ist oft der wichtigste Schritt für die Lösung eines ML-Problems. Das Beispieldaten sollten die Daten repräsentieren, die Ihnen vorliegen, wenn Sie das Modell für eine Voraussage verwenden. Beispiel: Wenn Sie voraussagen möchten, ob eine E-Mail Spam ist oder nicht, müssen Sie sowohl positive (Spam-E-Mails) als auch negative (keine Spam-E-Mails) sammeln, damit der Machine Learning-Algorithmus Muster erkennen kann, die diese beiden Arten von E-Mails voneinander unterscheiden.

Sobald Sie über die bezeichneten Daten verfügen, müssen Sie diese möglicherweise in einem Format konvertieren, das Ihr Algorithmus oder Ihre Software akzeptiert. Für Amazon ML müssen Sie

die Daten beispielsweise in das CSV-Format (CSV) bringen, wobei jedes Beispiel eine Zeile in der CSV-Datei darstellt und jede Spalte eine Eingabevariable und eine Spalte die Zielantwort enthält.

Analysieren Ihrer Daten

Bevor Sie den ML-Algorithmus mit Ihren bezeichneten Daten füttern, sollten Sie Ihre Daten überprüfen, um Probleme festzustellen und einen Einblick in die Daten zu erhalten, die Sie verwenden. Die Voraussagekraft Ihres Modells ist nur so gut wie die Daten, mit denen Sie es füttern.

Beim Analysieren Ihrer Daten sollten Sie die Folgendes beachten:

- **Variablen- und Zieldatenzusammenfassung** – Es ist nützlich, die Werte zu verstehen, die Ihre Variablen annehmen, und welche Werte in Ihren Daten dominant sind. Sie können diese Zusammenfassungen von einem Experte für das zu lösende Problem erstellen lassen. Fragen Sie sich oder den Experten: Erfüllen die Daten Ihre Erwartungen? Haben Sie allem Anschein nach ein Problem mit dem Sammeln von Daten? Kommt eine Klasse in Ihrem Ziel häufiger vor als die anderen Klassen? Gibt es mehrere fehlende Werte oder ungültige Daten als erwartet?
- **Variable-Ziel-Korrelationen** – Es ist nützlich, die Korrelation zwischen Variablen und Zielklassen zu kennen, da eine hohe Korrelation darauf hindeutet, dass zwischen der Variablen und der Ziel-Klasse eine Beziehung besteht. Im Allgemeinen sollten Sie Variablen mit hoher Korrelation verwenden, da diese eine stärkere Voraussagekraft (Signal) haben, und Variablen mit niedriger Korrelation auslassen, da sie wahrscheinlich nicht relevant sind.

In Amazon ML können Sie Ihre Daten analysieren, indem Sie den Datenbericht prüfen und eine Datenquelle erstellen.

Funktionsverarbeitung

Nachdem Sie sich mithilfe von Datenzusammenfassungen und Visualisierungen mit Ihren vertraut gemacht haben, möchten Sie Ihre Variablen möglicherweise weiter transformieren, damit sie aussagekräftiger sind. Dieser Vorgang wird Funktionsverarbeitung genannt. Beispiel: Sie haben eine Variable, die Datum und Uhrzeit eines Ereignisses erfasst. Dieses Datum und diese Uhrzeit treten nie wieder auf und sind daher nicht für eine Voraussage Ihres Ziels geeignet. Wenn Sie diese Variable jedoch in Funktionen transformieren, welche die Stunden eines Tags, den Wochentag und den Monat angeben, können diese Variablen nützlich sein, um zu erfahren, ob das Ereignis zu einer bestimmten Stunde, an einem bestimmten Wochentag oder in einem bestimmten Monat auftritt. Eine solche Funktionsverarbeitung für generalisierbare Datenpunkte können das Voraussagemodell deutlich verbessern.

Weitere Beispiele für eine gängige Funktionsverarbeitung:

- Ersetzen fehlender oder ungültiger Daten durch aussagekräftige Werte (wenn Sie z. B. wissen, dass ein fehlender Wert für eine Produktart-Variable bedeutet, dass es sich um ein Buch handelt, können Sie alle fehlenden Werte in der Produktart durch den Wert für Buch ersetzen). Eine gängige Strategie für das Ersetzen fehlender Werte ist das Austauschen der fehlenden Werte mit einem Mittel- oder Durchschnittswert. Es ist wichtig, dass Sie Ihre Daten verstehen, bevor Sie sich für eine Strategie für das Austauschen fehlender Werte entscheiden.
- Bilden kartesischer Produkte aus einer Variable mit einer anderen. Wenn Sie beispielsweise über zwei Variablen verfügen, nämlich Bevölkerungsdichte (Stadt, Vorort, Land) und Staat (Washington, Oregon, Kalifornien), können sich in den Funktionen, die aus einem kartesischen Produkt aus diesen beiden Variablen zu neuen Funktionen geformt werden (urban_Washington, suburban_Washington, rural_Washington, urban_Oregon, suburban_Oregon, rural_Oregon, urban_California, suburban_California, rural_California), nützliche Informationen verbergen.
- Nicht-lineare Transformationen wie das Binning von numerischen Variablen zu Kategorien. In vielen Fällen ist die Beziehung zwischen einer numerischen Funktion und dem Ziel nicht linear (der numerische Funktionswert wird nicht gleichmäßig mit dem Ziel erhöht oder verringert). In solchen Fällen kann es nützlich sein, die numerische Funktion in kategorische Funktionen zu packen, um verschiedene Bereiche der numerischen Funktion darzustellen. Jede kategorische Funktion (Bin) kann dann mit einer eigenen linearen Beziehung zum Ziel im Modell dargestellt werden. Nehmen wir an, Sie wissen, dass die kontinuierliche numerische Funktion "age" nicht linear mit der Wahrscheinlichkeit verläuft, ein Buch zu kaufen. Sie können die Dauer also in kategorische Funktionen packen, die in der Lage sind, die Beziehung zum Ziel genauer zu erfassen. Die optimale Anzahl von Paketen für eine numerische Variable hängt von den Eigenschaften der Variablen und ihrer Beziehung mit dem Ziel ab und wird am besten durch Experimente bestimmt. Amazon ML schlägt die optimale Paketanzahl für eine numerische Funktion basierend auf den Datenstatistiken im vorgeschlagenen Rezept vor. Weitere Informationen zum empfohlenen Rezept finden Sie im Developer-Handbuch.
- Domain-spezifische Funktionen (z. B. können Sie mit Länge, Breite und Höhe als separate Variablen eine neue Volume-Funktion als Produkt dieser drei Variablen erstellen).
- Variable-spezifische Funktionen. Einige Variable-Typen, z. B. SMS-Funktionen oder Funktionen, welche die Struktur einer Webseite oder die Struktur eines Satzes erfassen, haben generische Verarbeitungsmöglichkeiten, welche die Extraktion von Struktur und Kontext unterstützen. Beispielsweise kann das Bilden von n-grams aus dem Text "the fox jumped over the fence" mit unigrams dargestellt werden: the, fox, jumped, over, fence, oder bigrams: the fox, fox jumped, jumped over, over the, the fence.

Das Einbeziehen relevanter Funktionen verbessert die Voraussagekraft. Natürlich ist es nicht immer möglich, die Funktionen mit "signal"- oder Voraussagekraft im Voraus zu kennen. Deshalb ist es gut, dass alle Funktionen, die möglicherweise einen Bezug zur Zielbezeichnung haben, einzubeziehen und den Modellschulungsalgorithmus die stärksten Korrelationen wählen zu lassen. In Amazon ML kann die Funktionsverarbeitung beim Erstellen eines Modells im Rezept festgelegt werden. Eine Liste der verfügbaren Funktionsprozessoren finden Sie im Developer-Handbuch.

Aufteilung von Daten in Schulungs- und Evaluierungsdaten

Grundlegendes Ziel von ML ist es, über die Daten-Instances, die für die Schulung von Modellen verwendet werden, hinaus zu generalisieren. Wir möchten das Modell so evaluieren, dass es die Qualität seiner Mustergeneralisierung für Daten, für die das Modell nicht geschult wurde, einschätzt. Da zukünftige Instances unbekannte Zielwerte enthalten können wir die Richtigkeit unserer Voraussagen für zukünftige Instances jetzt nicht prüfen können, müssen wir die Daten, deren Antwort wir bereits kennen, als Proxy für zukünftige Daten verwenden. Das Testen des Modells mit denselben Daten, die für die Schulung verwendet wurden, ist nicht sinnvoll, weil sich Modelle an spezifische Schulungsdaten „erinnern“ anstatt sie zu verallgemeinern.

Eine gängige Strategie ist es, alle verfügbaren bezeichneten Daten in Schulungs- und Evaluierungssätze aufzuteilen; in der Regel erfolgt dies mit einem Verhältnis von 70 bis 80 Prozent für Schulungen und 20-30 Prozent für die Evaluation. Das ML-System verwendet die Schulungsdaten, um Modelle auf die Mustererkennung zu schulen, und verwendet die Evaluierungsdaten, um die Voraussagequalität der geschulten Modell zu bewerten. Das ML-System bewertet die Voraussageleistung durch Vergleichen der Voraussagen auf der Grundlage eines Evaluierungsdatensatzes mit den tatsächlichen Werten (bekannt als Referenzwert) und mithilfe einer Vielzahl von Metriken. In der Regel verwenden Sie die Modelle, die den Evaluierungsdatensatz am besten für ihre Voraussagen verwendet haben, für zukünftige Instances, deren Zielantwort Sie nicht kennen.

Amazon ML teilt die für die Schulung eines Modells über die Amazon ML-Konsole gesendeten Daten in 70 Prozent für Schulungen und 30 Prozent für die Bewertung auf. Standardmäßig verwendet Amazon ML die ersten 70 Prozent der Eingabedaten in der Reihenfolge, in der sie in der Datenquelle stehen, als Schulungsdatenquelle und die verbleibenden 30 Prozent der Daten als Evaluierungsdatenquelle. Mit Amazon ML können Sie zudem zufällige 70 Prozent der Quelldaten anstelle der ersten 70 Prozent für Schulungen und den Rest dieser zufälligen Teilmenge zur Bewertung verwenden. Sie können Amazon ML-APIs verwenden, um benutzerdefinierte Split-APIs festzulegen und Schulungs- und Bewertungsdaten zur Verfügung zu stellen, die außerhalb von

Amazon ML aufgeteilt wurden. Amazon ML bietet zudem drei Optionen für das Aufteilen der Daten. Weitere Informationen zu Aufteilungsoptionen finden Sie unter [Aufteilen Ihrer Daten](#).

Schulen des Modells

Sie sind nun bereit, den ML-Algorithmus (also den Lernalgorithmus) mit den Schulungsdaten zu füttern. Der Algorithmus lernt von den Schulungsdatenmustern, welche die Variablen dem Ziel zuweisen, und gibt ein Modell aus, das diese Beziehungen erfasst. Das ML-Modell kann verwendet werden, um Voraussagen für neue Daten zu erhalten, bei denen Sie die Zielantwort nicht kennen.

Lineare Modelle

Es steht eine große Anzahl von ML-Modellen zur Verfügung. Amazon ML lernt eine Art von ML-Modell: lineare Modelle. Der Begriff lineares Modell deutet darauf hin, dass das Modell als eine lineare Kombination von Funktionen spezifiziert ist. Basierend auf den Schulungsdaten berechnet der Lernprozess eine Gewichtung für jede Funktion, woraus sich ein Modell ergibt, das den Zielwert voraussagen oder einschätzen kann. Beispiel: Wenn Ihr Ziel die Höhe der Versicherung ist, die ein Kunde kaufen wird, und Ihre Variablen sind Alter und Einkommen, wäre ein einfaches lineares Modell wie folgt:

```
Estimated target = 0.2 + 5·age + 0.0003·income
```

Lernalgorithmus

Der Lernalgorithmus soll die Gewichtungen für ein Modell lernen. Die Gewichtungen beschreiben die Wahrscheinlichkeit, dass die Muster, die das Modell lernt, die tatsächlichen Beziehungen in den Daten widerspiegeln. Ein Lernalgorithmus besteht aus einer Verlustfunktion und einer Optimierungsmethode. Der Verlust ist die Strafe dafür, wenn die vom ML-Modell bereitgestellte Einschätzung des Ziels nicht genau dem Ziel entspricht. Eine Verlustfunktion quantifiziert diese Strafe als einzelnen Wert. Eine Optimierungsmethode dient dazu, den Verlust zu minimieren. In Amazon Machine Learning verwenden wir drei Verlustfunktionen, eine für jeden der drei Typen von Voraussageproblemen. Die in Amazon ML verwendete Optimierungsmethode ist der online-basierte Stochastic Gradient Descent (SGD). Der SGD macht sequenzielle Durchgänge durch die Schulungsdaten und aktualisiert in jedem Durchgang die Funktionsgewichtungen der einzelnen Beispiele, um eine optimale Gewichtung zu erzielen und den Verlust zu minimieren.

Amazon ML verwendet die folgenden Algorithmen:

- Für die binäre Klassifizierung verwendet Amazon ML die logistische Regression (logistische Verlustfunktion+SGD).

- Für die Mehrklassen-Klassifizierung verwendet Amazon ML die multinominale logistische Regression (multinominale logistische Verlustfunktion+SGD).
- Für die Regression verwendet Amazon ML die lineare Regression (quadratische Verlustfunktion +SGD).

Schulungsparameter

Der Amazon ML-Lernalgorithmus akzeptiert Parameter, genannt Hyper-Parameter oder Schulungsparameter, mit denen Sie die Qualität des resultierenden Modells kontrollieren können. Je nach Hyperparameter wählt Amazon ML automatisch Einstellungen oder bietet statische Standardwerte für die Hyperparameter. Obwohl die Einstellungen der Standard-Hyperparameter in der Regel nützliche Modelle produzieren, können Sie die Voraussageleistung Ihrer Modelle verbessern, indem Sie die Hyperparameterwerte ändern. In den folgenden Abschnitten wird häufig Parameter im Zusammenhang mit Lernalgorithmen für lineare Modelle, z. B. die von Amazon ML.

Lernrate

Bei der Lernrate handelt es sich um einen konstanten Wert im Algorithmus des Stochastic Gradient Descent (SGD). Die Lernrate wirkt sich die Geschwindigkeit aus, mit welcher der Algorithmus die optimalen Gewichtungen erreicht bzw. sich diesen annähert. Der SGD-Algorithmus aktualisiert die Gewichtungen des linearen Modells für jedes greifbare Datenbeispiel. Die Größe dieser Aktualisierungen wird von der Lernrate bestimmt. Eine zu große Lernrate kann verhindern, dass sich die Gewichtungen der optimalen Lösung annähern. Eine zu kleine Lernrate führt dazu, dass der Algorithmus viele Durchläufe benötigt, um eine optimale Gewichtung zu erzielen.

In Amazon ML wird die Lernrate automatisch basierend auf Ihre Daten ausgewählt.

Modellgröße

Wenn Sie über viele Eingabefunktionen verfügen, kann die Anzahl der möglichen Muster in den Daten zu einem großen Modell führen. Große Modelle haben praktische Implikationen, sie erfordern z. B. mehr RAM für das Modell während der Schulung und beim Generieren von Voraussagen. In Amazon ML können Sie die Modellgröße mithilfe der L1-Regularisation oder durch spezifisches Einschränken der Modellgröße durch Festlegen der Maximalgröße reduzieren. Beachten Sie, dass wenn Sie die Modellgröße zu sehr verringern, die Voraussagekraft Ihres Modells eingeschränkt sein kann.

Weitere Informationen zur Standard-Modellgröße finden Sie unter [Schulungsparameter: Typen und Standardwerte](#). Weitere Informationen zur Regularisation finden Sie unter [Regularisation](#).

Anzahl der Durchläufe

Der SGD-Algorithmus macht sequenzielle Durchgänge durch die Schulungsdaten. Der `Number of passes`-Parameter steuert die Anzahl von Durchgängen, die der Algorithmus durch die Schulungsdaten vornimmt. Mehr Durchgänge führen dazu, dass das Modell besser auf die Daten abgestimmt ist (sofern die Lernrate nicht zu hoch ist). Mit der deutlichen Zunahme der Anzahl der Durchgänge jedoch schrumpft dieser Vorteil wieder. Bei kleineren Datensätzen können Sie die Anzahl von Durchläufen deutlich erhöhen, sodass der Lernalgorithmus effektiv auf die Daten abgestimmt werden kann. Bei besonders großen Datensätzen ist ein einzelner Durchgang möglicherweise ausreichend.

Weitere Informationen zur standardmäßigen Anzahl an Durchläufen finden Sie unter [Schulungsparameter: Typen und Standardwerte](#).

Daten-Shuffling

In Amazon ML müssen Sie Ihre Daten mischen, da der SGD-Algorithmus durch die Reihenfolge der Zeilen in den Schulungsdaten beeinflusst wird. Das Mischen oder Shuffling Ihrer Schulungsdaten führt zu besseren ML-Modellen, da der SGD-Algorithmus Lösungen vermeidet, die zwar für den ersten Datentyp aber nicht für alle Daten optimal sind. Beim Mischen wird die Reihenfolge der Daten so geändert, dass der SGD-Algorithmus nacheinander nicht nur einen Datentyp bei zahlreichen Beobachtungen erkennt. Wenn für mehrere aufeinanderfolgende Durchgänge nur eine Art von Daten erkannt werden, kann der Algorithmus die Modellgewichtungen möglicherweise nicht für einen neuen Datentyp korrigieren, da die Aktualisierung zu groß sein kann. Wenn die Daten zudem nicht in zufälliger Reihenfolge präsentiert werden, ist es für den Algorithmus schwierig, schnell die optimale Lösung für alle Datentypen zu finden; in einigen Fällen findet der Algorithmus möglicherweise überhaupt keine optimale Lösung. Das Mischen der Schulungsdaten hilft dem Algorithmus, die optimale Lösung schneller zu finden.

Angenommen, Sie möchten ein ML-Modell so schulen, dass es eine Produktart voraussagt, und Ihre Schulungsdaten enthalten die Produktarten Film, Spielzeug und Videospiel. Wenn Sie die Daten vor dem Hochladen in Amazon S3 nach Produktartspalte sortieren, erkennt der Algorithmus die Daten alphabetisch nach Produkttyp. Der Algorithmus erkennt alle Daten für Filme zuerst, und das ML-Modell beginnt, Muster für Filme zu erlernen. Wenn das Modell dann Daten zu Spielsachen erkennt, würde jedes Update, das der Algorithmus vornimmt, das Modell an den Produkttyp "Spielzeug" anpassen, auch wenn diese Updates die Muster herabsetzen, die Filmen entsprechen. Durch diesen

plötzlichen Wechsel vom Typ "Film" zu "Spielzeug" kann ein Modell erzeugen, das nicht lernt, wie Produkttypen korrekt vorhergesagt werden.

Weitere Informationen zur Mischart finden Sie unter [Schulungsparameter: Typen und Standardwerte](#).

Regularisation

Die Regularisation hilft dabei, zu verhindern, dass lineare Modelle Schulungsdatenbeispiele übermäßig anpassen (d. h. sich Muster merken statt sie zu verallgemeinern), indem Werte mit extremer Gewichtung mit einer Strafe belegt werden. Die L1-Regularisation mindert Anzahl von Funktionen, die im Modell verwendet werden, indem sie die Gewichtungen von Funktionen mit kleinen Gewichtungen auf Null setzt. Infolgedessen führt die L1-Regularisation zu platzsparenden Modellen und reduziert die Störungsmenge im Modell. Die L2-Regularisation führt zu kleineren Gesamtgewichtungswerten und stabilisiert die Gewichtungen, wenn zwischen den Eingabefunktionen eine hohe Korrelation besteht. Mithilfe der Parameter `Regularization type` und `Regularization amount` steuern Sie die Höhe der angewendeten L1- und L2-Regularisation. Durch einen extrem hohen Regularisationswert kann die Gewichtung aller Funktionen Null sein, sodass ein Modell keine Muster mehr lernen kann.

Weitere Informationen zu Regularisationswerten finden Sie unter [Schulungsparameter: Typen und Standardwerte](#).

Evaluation der Modellrichtigkeit

Ziel des ML-Modells ist es, Muster zu lernen, die gut für unbekannte Daten verallgemeinert werden können, statt sich die Daten aus einer Schulung zu merken. Sobald Sie über ein Modell verfügen, überprüfen Sie, ob Ihr Modell unbekannte Beispiele, die Sie nicht für die Schulung des Modells verwendet haben, gut verarbeitet. Verwenden Sie also das Modell, um die Antwort für einen Evaluationsdatensatz (zurückgehaltene Daten) vorauszusagen, und vergleichen Sie dann das vorausgesagte Ziel mit der tatsächlichen Antwort (Referenzwert).

Das ML verwendet eine Reihe von Metriken, um die Voraussagerichtigkeit eines Modells zu messen. Die Wahl für eine Richtigkeitsmetrik ist abhängig von der Art der ML-Aufgabe. Es ist wichtig, diese Metriken zu prüfen, um zu entscheiden, ob Ihr Modell gut funktioniert.

Binäre Klassifizierung

Die tatsächliche Ausgabe von vielen binären Klassifizierungsalgorithmen ist eine Voraussagepunktzahl. Die Punktzahl gibt die Sicherheit des Systems an, dass die angegebene Beobachtung der positiven Klasse angehört. Sie können die Punktzahl interpretieren, indem

Sie einen Klassifizierungsschwellenwert oder Grenzwert festlegen und die Punktzahl damit vergleichen, um zu entscheiden, ob die Beobachtung als positiv oder negativ klassifiziert wird. Alle Beobachtungen mit Punkteständen höher als der Schwellenwert werden als positive Klasse vorausgesagt, und Punktestände unter dem Schwellenwert werden als negative Klasse vorausgesagt.

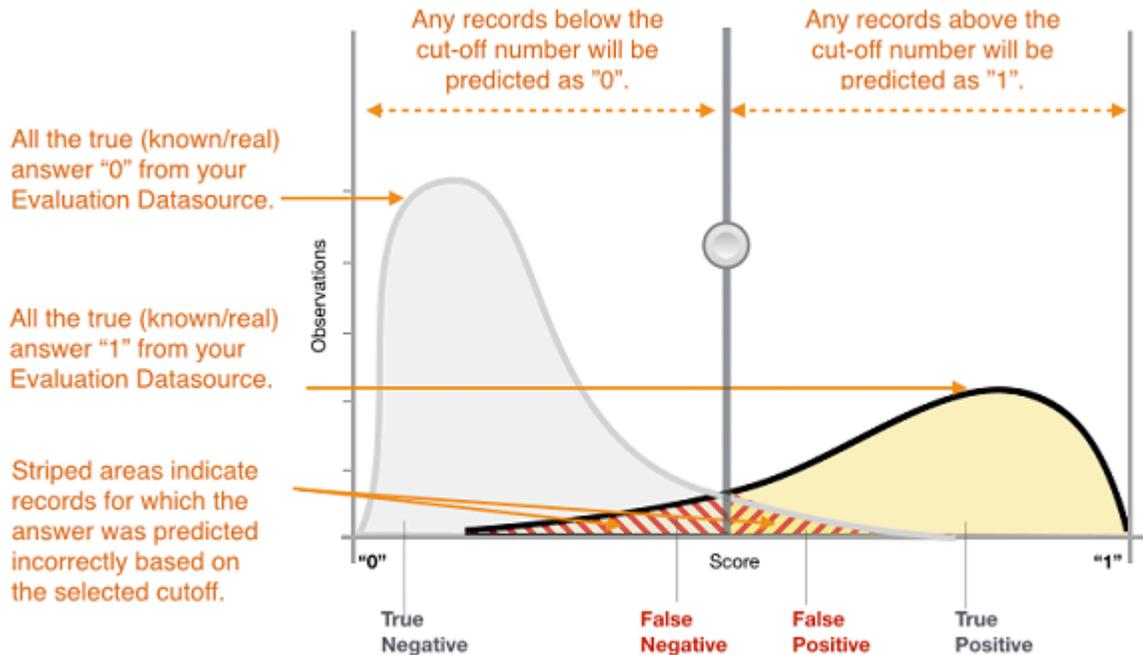


Abbildung 1: Verteilung der Punkte für ein binäres Klassifikationsmodell

Die Voraussagen werden nun basierend auf der tatsächlichen bekannten Antwort und der vorausgesagten Antwort in vier Gruppen unterteilt: richtige positive Voraussagen (echte Positive), richtige negative Voraussagen (echte Negative), falsche positive Voraussagen (falsche Positive) und falsche negative Voraussagen (falsche Negative).

Richtigkeitsmetriken für die binäre Klassifizierung quantifizieren zwei Arten von richtigen Voraussagen und zwei Arten von Fehlern. Typische Metriken sind Richtigkeit (ACC), Präzision, Wiedererkennung, Falschpositivrate, F1-measure. Jede Metrik misst einen anderen Aspekt des Voraussagemodells. Die Richtigkeit (ACC) misst den Anteil der richtigen Voraussagen. Genauigkeit misst den Anteil der tatsächlichen Positiva unter den Beispielen, die als positiv vorausgesagt wurden. Wiedererkennung misst, wie viele tatsächliche Positive als positiv vorausgesagt wurden. F1-measure ist das harmonische Mittel von Genauigkeit und Wiedererkennung.

AUC ist eine andere Art der Metrik. Sie misst die Fähigkeit des Modells, eine höhere Bewertung für positive Beispiele im Vergleich zu negativen Beispielen vorherzusagen. Da die AUC unabhängig vom

ausgewählten Schwellenwert ist, bekommen Sie ein Gefühl für die Voraussageleistung Ihres Modells aus der AUC-Metrik, ohne einen Schwellenwert auszuwählen.

Abhängig von Ihrem Unternehmensproblem benötigen Sie vielleicht eher ein Modell, das für eine bestimmte Teilmenge dieser Metriken gut funktioniert. Zwei Unternehmensanwendungen können beispielsweise sehr unterschiedliche Anforderungen an ihre ML-Modelle haben:

- Eine Anwendung muss vielleicht sehr sicher sein, dass die positiven Voraussagen tatsächlich positiv sind (hohe Präzision) und kann es verkraften, dass einige positive Beispiele falsch als negativ klassifiziert werden (moderate Wiedererkennung).
- Eine andere Anwendung soll so viele positive Beispiele wie möglich korrekt voraussagen (hohe Wiedererkennung) und nimmt es in Kauf, dass einige negative Beispiele falsch als positiv klassifiziert werden (moderate Genauigkeit).

In Amazon ML erhalten Beobachtungen eine vorausgesagte Punktzahl im Bereich $[0,1]$. Der Schwellenwert für die Entscheidung für die Klassifizierung von Beispielen als 0 oder 1 ist standardmäßig auf 0,5 festgelegt. Mit Amazon ML können Sie die Auswirkungen der Auswahl von unterschiedlichen Schwellenwerten prüfen und einen geeigneten Schwellenwert wählen, das Ihren geschäftlichen Anforderungen entspricht.

Mehrklassen-Klassifizierung

Im Gegensatz zum Prozess für binäre Klassifizierungsprobleme müssen Sie keinen Schwellenwert wählen, um Voraussagen zu treffen. Die vorausgesagte Antwort ist die Klasse (z. B. Bezeichnung) mit der höchsten vorausgesagten Punktzahl. In einigen Fällen möchten Sie die vorausgesagte Antwort vielleicht nur dann verwenden, wenn sie mit einer hohen Punktzahl vorausgesagt wurde. In diesem Fall können Sie einen Schwellenwert für die vorausgesagten Punktzahlen wählen, anhand dessen Sie die vorausgesagte Antwort akzeptieren oder nicht.

Typische in Mehrklassen verwendete Metriken sind dieselben wie die Metriken, die für den binären Klassifizierungsfall verwendet werden. Die Metrik wird für jede einzelne Klasse berechnet, indem sie nach der Gruppierung aller anderen Klassen in die zweite Klasse als binäres Klassifizierungsproblem behandelt wird. Dann wird die binäre Klasse gemittelt, um entweder eine Makro-Mittelmetrik (jede Klasse wird gleich behandelt) oder gewichtete Mittelmetrik (gewichtet nach Klassenfrequenz) zu erhalten. In Amazon ML wird das Makro-Mittel F1-Measure zur Evaluierung des Voraussageerfolgs eines Mehrklassen-Klassifizierers verwendet.

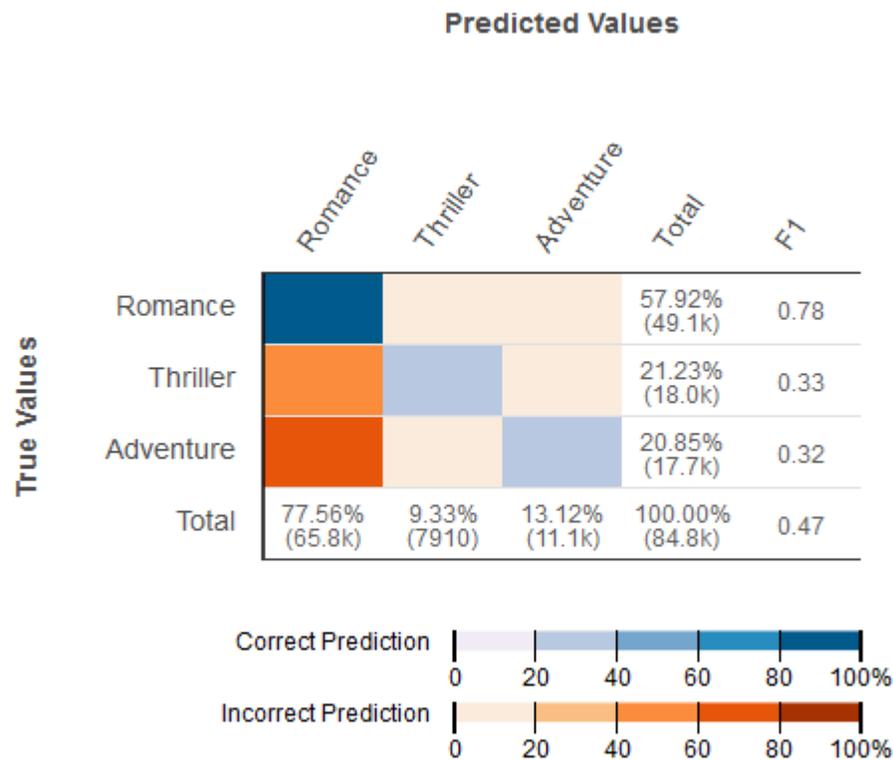


Abbildung 2: Verwirrungsmatrix für ein Mehrklassen-Klassifizierungsmodell

Es ist hilfreich, die Konfusionsmatrix auf Mehrklassen-Probleme zu prüfen. Die Konfusionsmatrix ist eine Tabelle, die jeden Klasse in den Evaluationsdaten und die Anzahl oder den Prozentsatz der richtigen und falschen Voraussagen darstellt.

Regression

Die typischen Richtigkeitsmetriken für Regressionsaufgaben sind root mean square error (RMSE) und mean absolute percentage error (MAPE). Diese Metriken messen die Abweichung zwischen dem vorausgesagten numerischen Ziel und der tatsächlichen numerischen Antwort (Referenzdaten). In Amazon ML wird die RMSE-Metrik zur Evaluierung der Voraussagerichtigkeit eines Regressionsmodells verwendet.

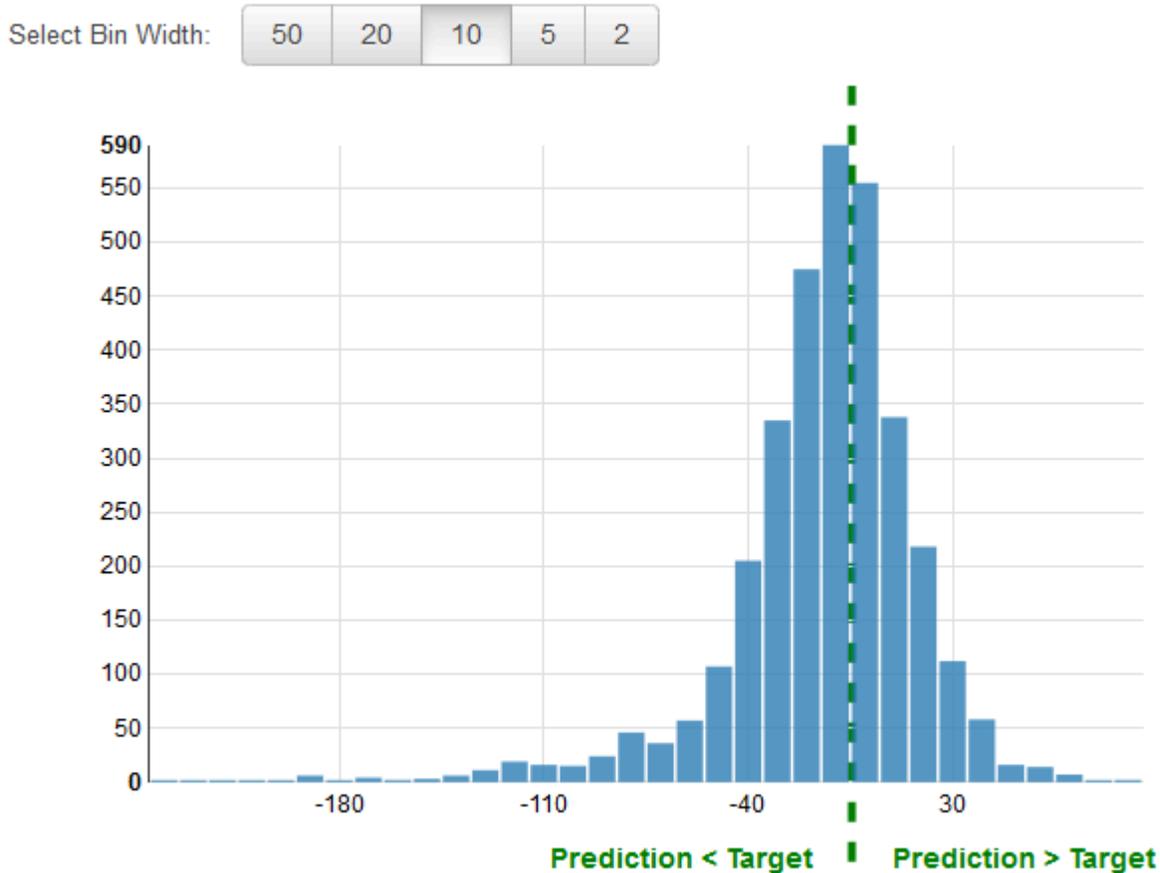


Abbildung 3: Verteilung der Residuen für ein Regressionsmodell

Es ist eine gängige Vorgehensweise, den Rest bei Regressionsproblemen zu überprüfen. Ein Rest für eine Beobachtung in der Evaluierungsdaten ist der Unterschied zwischen dem wahren Ziel und dem vorausgesagten Ziel. Reste stellen den Teil des Ziels dar, den das Modell nicht voraussagen konnte. Ein positiver Rest deutet darauf hin, dass das Modell das Ziel unterschätzt (das tatsächliche Ziel ist größer als das vorausgesagte Ziel). Ein negativer Rest deutet auf eine Überbewertung hin (das tatsächliche Ziel ist kleiner als das vorausgesagte Ziel). Das Histogramm der Reste für die Evaluierungsdaten deutet bei glockenförmiger Anordnung und Zentrierung auf Null darauf hin, dass das Modell willkürliche Fehler macht und keinen spezifischen Zielwertbereich systematisch über- oder unterschätzt. Wenn die Reste keine Glockenform mit Zentrierung auf Null bilden, gibt es eine gewisse Struktur bei den Voraussagefehlern des Modells. Das Hinzufügen von weiteren Variablen kann es dem Modell ermöglichen, Muster zu erfassen, die vom aktuellen Modell nicht erfasst werden.

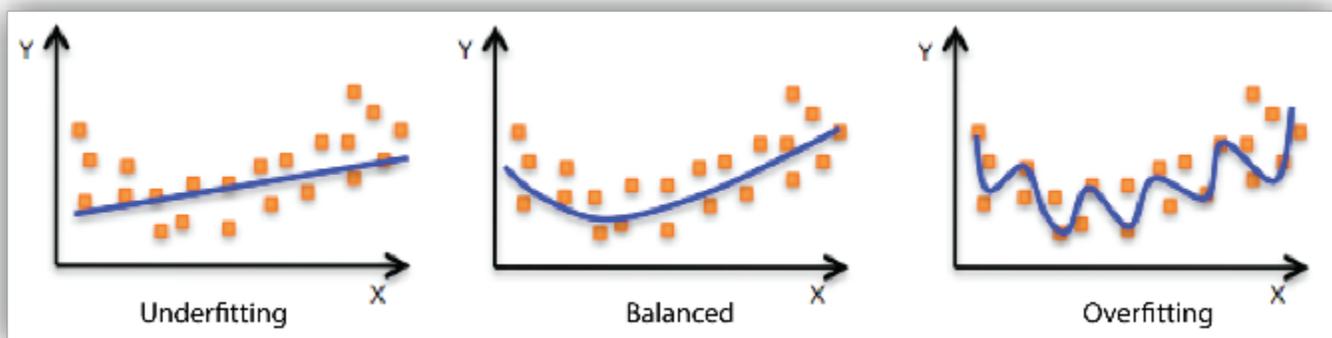
Optimierung der Modellrichtigkeit

Um ein ML-Modell zu erhalten, das Ihren Anforderungen genügt, müssen Sie für gewöhnlich diesen ML-Prozess durchlaufen und ein paar Varianten ausprobieren. Vielleicht erhalten Sie nach dem ersten Durchlauf kein voraussagestarkes Modell, oder Sie möchten Ihr Modell optimieren, um noch bessere Voraussagen treffen zu können. Um die Leistung zu verbessern, führen Sie die folgenden Schritte aus:

1. Sammeln von Daten: Erhöhen der Anzahl der Schulungsbeispiele
2. Funktionsverarbeitung: Fügen Sie zusätzlichen Variablen und eine bessere Funktionsverarbeitung hinzu
3. Modellparameter-Tuning: Berücksichtigen Sie alternative Werte für die Schulungsparameter, die von Ihrem Lernalgorithmus verwendet werden.

Modellanpassung: Unteranpassung vs. Overfitting

Wenn Sie die Modellanpassung verstehen, verstehen Sie auch die Ursache für eine schlechte Modellrichtigkeit. Auf der Grundlage dieses Verständnisses können Sie korrigierende Maßnahmen ergreifen. Wir können bestimmen, ob ein Voraussagemodell die Schulungsdaten zu viel oder zu wenig anpasst, indem wir uns den Voraussagefehler für die Schulungsdaten und Evaluationsdaten ansehen.



Ihr Modell führt eine Unteranpassung der Schulungsdaten durch, wenn das Modell die Schulungsdaten nicht gut verarbeitet. Der Grund hierfür ist, dass das Modell die Beziehung zwischen den Eingabebeispielen (häufig als "X" bezeichnet) und den Zielwerten (häufig als "Y" bezeichnet) nicht erfassen kann. Ihr Modell führt eine Überanpassung Ihrer Schulungsdaten durch, wenn die Leistung des ML-Modells an den Lerndaten selbst gut ist, es jedoch an der Datenauswertung

scheitert. Der Grund hierfür ist, dass das Modell sich die bekannten Daten merkt und diese nicht auf unbekannte Beispiele anwenden kann.

Eine schlechte Leistung an den Lerndaten kann daran liegen, dass das Modell zu einfach ist (die Eingabefunktionen sind nicht ausdrucksstark genug), um das Ziel gut zu beschreiben. Die Leistung kann verbessert werden, indem die Flexibilität des Modells erhöht wird. Um die Flexibilität des Modells zu erhöhen, führen Sie die folgenden Schritte aus:

- Fügen Sie neue Domain-spezifische Funktionen und mehr kartesischen Produkte hinzu, und ändern Sie die Art der verwendeten Funktionsverarbeitung (z. B. Erhöhen der n-grams-Größe)
- Verringern Sie den Umfang der verwendeten Regularisation

Wenn Ihr Modell die Schulungsdaten übermäßig stark anpasst, ist es durchaus sinnvoll, entsprechende Maßnahmen zu ergreifen, um die Flexibilität des Modells zu reduzieren. Um die Flexibilität des Modells zu reduzieren, führen Sie die folgenden Schritte aus:

- Funktionsauswahl: Verwenden Sie weniger Funktionskombinationen, reduzieren Sie die n-grams-Größe, und reduzieren Sie die Anzahl der Bins für numerische Attribute.
- Erhöhen Sie den Umfang der verwendeten Regularisation.

Die Richtigkeit von Schulungs- und Prüfdaten kann gering sein, da der Lernalgorithmus nicht ausreichend Daten zum lernen zur Verfügung hatte. Sie können die Leistung verbessern, indem Sie die folgenden Schritte ausführen:

- Erhöhen Sie die Anzahl der Schulungsdatenbeispiele.
- Erhöhen Sie die Anzahl von Durchläufen durch die vorhandenen Schulungsdaten.

Mit dem Modell Voraussagenerstellen

Nachdem Sie nun über ein gut funktionierendes ML-Modell verfügen, verwenden Sie es, um Voraussagen zu treffen. In Amazon Machine Learning gibt es zwei Möglichkeiten für die Verwendung eines Modells für Voraussagen:

Stapelvoraussagen

Stapelvoraussagen sind nützlich, wenn Sie für eine Reihe von Beobachtungen gleichzeitig Voraussagen erstellen und anschließend auf einen bestimmten Prozentsatz oder an einer

bestimmten Anzahl Beobachtungen Maßnahmen ergreifen möchten. In der Regel benötigen Sie für eine solche Anwendung keine niedrige Latenz. Wenn Sie beispielsweise entscheiden möchten, welche Kunden Sie als Teil einer Werbekampagne für ein Produkt ansprechen möchten, und Sie erhalten Voraussagepunktzahlen für alle Kunden, sortieren Sie die Voraussagen Ihres Modells so, dass die Kunden identifiziert werden, die am ehesten kaufen werden, und nehmen Sie dann die top 5 %, die am wahrscheinlichsten kaufen werden.

Online-Voraussagen

Online-Voraussageszenarien sind für Fälle bestimmt, in denen Sie für jedes Beispiel einzeln und unabhängig von anderen Beispielen in einer Umgebung mit geringer Latenz eine Voraussage erzeugen möchten. Sie können Voraussagen beispielsweise verwenden, um direkte Entscheidungen darüber zu treffen, ob eine bestimmte Transaktion eine betrügerische Transaktion ist.

Modelle auf neue Daten umschulen

Damit ein Modell eine richtige Voraussage treffen kann, müssen die Daten, anhand derer die Voraussage getroffen wird, eine ähnliche Verteilung haben wie die Daten, mit denen das Modell geschult wurde. Da sich die Datenverteilung im Laufe der Zeit verschiebt, ist die Bereitstellung eines Modells keine einmalige Sache, sondern ein kontinuierlicher Prozess. Es empfiehlt sich, die eingehenden Daten fortlaufend zu überwachen und Ihr Modell neu zu schulen, wenn Sie feststellen, dass die Datenverteilung deutlich von der Datenverteilung der ursprünglichen Schulung abgewichen ist. Wenn die Überwachung der Daten zur Erkennung von Änderungen in der Datenverteilung im Rückstand ist, empfiehlt sich eine regelmäßige Schulung des Modells, beispielsweise täglich, wöchentlich oder monatlich. Um Modelle in Amazon ML neu zu schulen, müssen Sie auf der Grundlage Ihrer neuen Schulungsdaten ein neues Modell erstellen.

Der Amazon Machine Learning Learning-Prozess

In der folgenden Tabelle ist beschrieben, wie die Amazon ML-Konsole für die Ausführung des in diesem Dokument dargelegten ML-Prozesses verwendet wird.

ML-Prozess	Amazon-ML-Aufgaben
Analysieren Ihrer Daten	Um Ihre Daten in Amazon ML zu analysieren, erstellen Sie eine Datenquelle und überprüfen Sie die Seite zu den Dateneinblicken.

ML-Prozess	Amazon-ML-Aufgaben
Aufteilung von Daten in Schulungs- und Evaluierungsdatenquellen	<p>Amazon ML kann die Datenquelle aufteilen, sodass 70% der Daten für die Modellschulung und 30% für die Evaluierung der Voraussage-Performance Ihres Modell verwendet werden.</p> <p>Wenn Sie den Assistenten zur ML-Modellerstellung mit den Standardinstellungen wählen, teilt Amazon ML die Daten für Sie auf.</p> <p>Wenn Sie den Assistenten zur ML-Modellerstellung mit den benutzerdefinierten Einstellungen verwenden und sich für die Evaluierung des ML-Modells entscheiden, wird Ihnen eine Option angezeigt, mit der Sie es Amazon ML erlauben können, die Daten für Sie aufzuteilen und eine Evaluierung auf Höhe von 30% der Daten durchzuführen.</p>
Ihre Schulungsdaten mischen	<p>Wenn Sie den Assistenten zur ML-Modellerstellung mit den Standardinstellungen wählen, mischt Amazon ML Ihre Daten für Sie. Sie können Ihre Daten auch vor dem Import in Amazon ML mischen.</p>
Prozessfunktionen	<p>Der Prozess des Zusammenstellens von Schulungsdaten in einem optimalen Format für das Lernen und die Generalisierung wird als Funktionstransformation bezeichnet. Wenn Sie den Assistenten zur ML-Modellerstellung mit den Standardeinstellungen verwenden, schlägt Amazon ML Funktionsverarbeitungseinstellungen für Ihre Daten vor.</p> <p>Um Funktionsverarbeitungseinstellungen festzulegen, wählen Sie die Option Benutzerdefiniert des Assistenten zur ML-Modellerstellung aus und geben Sie ein Funktionsverarbeitungsrezept an.</p>
Schulen des Modells	<p>Wenn Sie den Assistenten zur ML-Modellerstellung für die Erstellung eines Modells in Amazon ML verwenden, schult Amazon ML Ihr Modell.</p>
Auswählen von Modellparametern	<p>In Amazon ML können Sie vier Parameter einstellen, die sich auf die Voraussage-Performance Ihres Modells auswirken: Modellgröße, Anzahl der Durchläufe, Art der Verbindung und Regularisation. Sie können diese Parameter einstellen, wenn Sie den Assistenten zur ML-Modellerstellung für die Erstellung eines ML-Modells verwenden und die Option Custom wählen.</p>

ML-Prozess	Amazon-ML-Aufgaben
Evaluieren der Modell-Performance	Verwenden Sie den Assistenten zum Erstellen von Evaluierungen, um die Voraussage-Performance Ihres Modells zu bewerten.
Funktionsauswahl	Der Amazon ML-Lernalgorithmus kann Funktionen auslassen, die nicht wirklich zum Lernprozess beitragen. Um anzugeben, dass Sie diese Funktionen auslassen möchten, wählen Sie bei der Erstellung des ML-Modells den <code>L1 regularization</code> -Parameter aus.
Festlegen eines Schwellenwerts für die Voraussagegenauigkeit	Überprüfen Sie Voraussage-Performance Ihre Modells im Evaluierungsbericht für verschiedene Schwellenwerte und legen Sie den Schwellenwert auf Basis Ihrer Business-Anwendung fest. Der Schwellenwert bestimmt, wie das Modell eine Voraussageübereinstimmung definiert. Passen Sie die Zahl an, um Fehleinstufungen zu steuern.
Verwenden des Modells	<p>Verwenden Sie Ihr Modell, um Voraussagen für einen Stapel von Beobachtungen zu erhalten, indem Sie den Assistenten zur Erstellung von Stapelvoraussagen verwenden.</p> <p>Alternativ können Sie mithilfe der <code>Predict</code>-API auch Voraussagen für individuelle Beobachtungen auf Abruf erhalten, indem Sie es Ihrem ML-Modell ermöglichen, Echtzeitvoraussagen zu verarbeiten.</p>

Einrichten von Amazon Machine Learning

Sie benötigen ein AWS-Konto, bevor Sie Amazon Machine Learning verwenden können. Falls Sie noch nicht über ein AWS-Konto verfügen, finden Sie weitere Informationen bei der Anmeldung zu AWS.

Anmelden bei AWS

Bei der Registrierung für Amazon Web Services (AWS) wird Ihr AWS-Konto automatisch für alle Dienste in AWS einschließlich Amazon ML registriert. Berechnet werden Ihnen aber nur die Services, die Sie nutzen. Wenn Sie bereits ein AWS-Konto haben, überspringen Sie diesen Schritt. Wenn Sie kein AWS-Konto haben, führen Sie die folgenden Schritte zum Erstellen eines Kontos aus.

Registrieren Sie sich für ein AWS-Konto wie folgt:

1. Gehen Sie zu <http://aws.amazon.com> und wählen Sie Anmeldung.
2. Folgen Sie den Anweisungen auf dem Bildschirm.

Der Anmeldeprozess beinhaltet auch einen Telefonanruf und die Eingabe einer PIN über die Telefontastatur.

Tutorial: Verwenden von Amazon ML zum Voraussagen der Reaktionen auf ein Marketingangebot

Mit Amazon Machine Learning (Amazon ML) können Sie Voraussagemodelle erstellen und schulen und Ihre Anwendungen in einer skalierbaren Cloudlösung hosten. In diesem Tutorial zeigen wir Ihnen, wie Sie die Amazon ML-Konsole zum Erstellen einer Datenquelle verwenden, eine Machine Learning (ML) -Modell entwickeln und das Modell zum Generieren von Voraussagen zweck Verwendung in Ihren Anwendungen nutzen können.

Unsere Beispielübung zeigt, wie potenzielle Kunden für eine gezielte Marketingkampagne identifiziert werden. Sie können aber dieselben Prinzipien anwenden, um eine Vielfalt von ML-Modellen zu erstellen und zu verwenden. Für die Beispielübung verwenden Sie öffentlich verfügbare Banking- und Marketingdatensätze aus dem [University of California at Irvine \(UCI\) Machine Learning Repository](#). Diese Datensätze enthalten allgemeine Informationen über Kunden sowie deren Reaktion auf vorherige Marketingkontakte. Sie werden diese Daten verwenden, um zu ermitteln, welche Kunden mit der größten Wahrscheinlichkeit Ihr neues Produkt, eine Banktermineinlage, auch bekannt als Einlagenzertifikat, abonnieren werden.

Warning

Dieses Tutorial ist nicht im kostenlosen AWS-Kontingent enthalten. Weitere Informationen zu Amazon ML-Preisen für finden Sie unter [Amazon Machine Learning Learning-Preiseaus](#).

Voraussetzung

Für das Tutorial benötigen Sie ein AWS-Konto. Wenn Sie noch kein AWS-Konto haben, informieren Sie sich unter [Einrichten von Amazon Machine Learning](#).

Schritte

- [Schritt 1: Vorbereiten der Daten](#)
- [Schritt 2: Erstellen einer Schulungsdatenquelle](#)
- [Schritt 3: Erstellen eines ML-Modells](#)
- [Schritt 4: Überprüfen der Voraussageleistung des ML-Modells und Festlegen eines Punktzellenwerts](#)

- [Schritt 5: Verwenden des ML-Modells zum Generieren von Prognosen](#)
- [Schritt 6: Bereinigen](#)

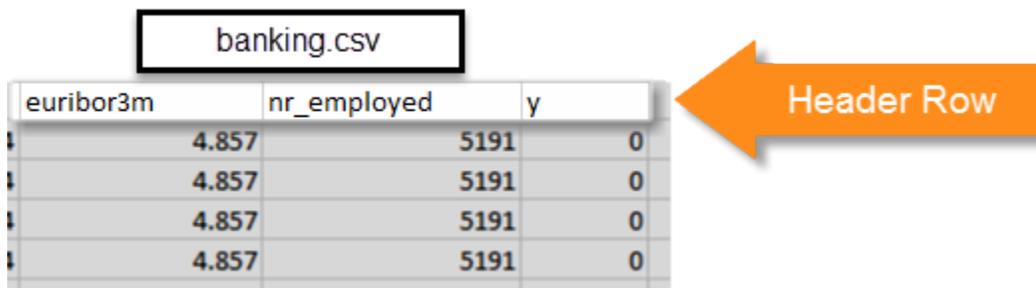
Schritt 1: Vorbereiten der Daten

Beim Machine Learning erhalten Sie in der Regel die Daten und stellen sicher, dass sie richtig formatiert sind, bevor Sie mit dem Schulungsprozess beginnen. Für die Zwecke dieses Tutorials haben wir einen Beispieldatensatz von der [UCI Machine Learning Learning-Repository](#), formatierte ihn gemäß den Amazon ML-Richtlinien und stellte ihn dann für Sie zum Download bereit. Laden Sie den Datensatz von unserem Amazon Simple Storage Service (Amazon S3) -Speicherort herunter und laden Sie ihn in Ihren eigenen S3-Bucket hoch, indem Sie die Verfahren in diesem Thema befolgen.

Die Amazon ML--Formatierungsanforderungen finden Sie unter [Erwähnungen zum Datenformat für Amazon ML](#) aus.

Vorgehensweise zum Herunterladen der Datensätze

1. Laden Sie die Datei mit den Verlaufsdaten für Kunden herunter, die Produkte ähnlich Ihrer Banktermineinlage gekauft haben, indem Sie auf [banking.zip](#) klicken. Entpacken Sie den Ordner und speichern Sie die Datei banking.csv-Datei auf Ihrem Computer.
2. Sie laden die Datei, mit der Sie vorhersagen können, ob potenzielle Kunden auf Ihr Angebot reagieren, durch Klicken auf [banking-batch.zip](#) herunter. Entpacken Sie den Ordner und speichern Sie die Datei banking-batch.csv auf Ihrem Computer.
3. Öffnen `banking.csv`. Sie sehen Zeilen und Spalten mit Daten. Die Kopfzeile enthält die Attributnamen für jede Spalte. Ein Attribut ist eine eindeutige, benannte Eigenschaft, die ein bestimmtes Merkmal der einzelnen Kunden beschreibt. So gibt `nr_employed` beispielsweise den Anstellungsstatus des Kunden an. Jede Zeile stellt die Sammlung von Beobachtungen zu einem einzelnen Kunden dar.



euribor3m	nr_employed	y
4.857	5191	0
4.857	5191	0
4.857	5191	0
4.857	5191	0

Sie möchten, dass Ihr ML-Modell die Frage "Wird dieser Kunde mein neues Produkt abonnieren?" beantwortet. Im `banking.csv`-Datensatz ist die Antwort auf diese Frage das Attribut `y`, welches einen Wert von 1 (für "Ja") oder 0 (für "Nein") enthält. Das Attribut, von dem Amazon ML lernen soll, wird als Zielattribut aus.

 Note

Das Attribut `y` ist ein binäres Attribut. Es kann nur einen von zwei Werten enthalten, in diesem Fall 0 oder 1. Im ursprünglichen UCI-Datensatz ist das `y`-Attribut entweder "Yes" (Ja) oder "No" (Nein). Wir haben den ursprünglichen Datensatz für Sie bearbeitet. Alle Werte des Attributs `y`, die für "Ja" stehen, sind jetzt 1, und alle Werte, die für "Nein" stehen, sind jetzt 0. Wenn Sie Ihre eigenen Daten verwenden, können Sie andere Werte für ein binäres Attribut verwenden. Weitere Informationen zu gültigen Werten finden Sie unter [Verwenden des Felds AttributeType](#).

Die folgenden Beispiele zeigen die Daten bevor und nachdem wir die Werte in Attribut `y` in die binären Attribute 0 und 1 geändert haben.

Before transformation

banking.csv



euribor3m	nr_employed	y
4.857	5191	no
4.857	5191	no
4.857	5191	yes
4.857	5191	yes
4.857	5191	no

After transformation

banking.csv



euribor3m	nr_employed	y
4.857	5191	0
4.857	5191	0
4.857	5191	1
4.857	5191	1
4.857	5191	0

Die `banking-batch.csv`-Datei enthält das `y`-Attribut nicht. Nachdem Sie ein ML-Modell erstellt haben, werden Sie das Modell zur Voraussage von `y` für jeden Datensatz in dieser Datei verwenden.

Laden Sie als Nächstes die `banking.csv` und `banking-batch.csv`-Dateien an Amazon S3.

Hochladen der Dateien an einen Amazon S3 -Speicherort

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon-S3-Konsole unter <https://console.aws.amazon.com/s3/>.
2. Erstellen Sie in der Liste All Buckets (Alle Buckets) einen Bucket oder wählen Sie den Speicherort aus, an dem Sie die Dateien hochladen möchten.
3. Wählen Sie in der Navigationsleiste die Option Hochladen aus.
4. Wählen Sie Add Files (Dateien hinzufügen) aus.
5. Navigieren Sie im Dialogfeld zu Ihrem Desktop, wählen Sie `banking.csv` und `banking-batch.csv` aus und klicken Sie dann auf Öffnen.

Jetzt können Sie Ihre [Schulungsdatenquelle erstellen](#).

Schritt 2: Erstellen einer Schulungsdatenquelle

Nachdem Sie das hochgeladen haben `banking.csv`-Datensatz an Ihren Amazon Simple Storage Service (Amazon S3) -Speicherort (Amazon S3), erstellen Sie damit eine Schulungsdatenquelle. Eine Datenquelle ist ein Amazon Machine Learning (Amazon ML) -Objekt, das den Speicherort Ihrer Input-Daten sowie wichtige Metadaten zu Ihren Input-Daten enthält. Amazon ML verwendet die Datenquelle für Operationen wie die ML-Modellschulung und -evaluierung.

Geben Sie Folgendes an, um eine Datenquelle zu erstellen:

- Den Amazon S3 S3-Speicherort Ihrer Daten und die Berechtigung für den Zugriff auf die Daten
- Das Schema, das die Namen der Attribute in den Daten und den Typ der einzelnen Attribute (numerisch, Text, kategorisch oder Binary) enthält
- Den Namen des Attributs, das die Antwort enthält, deren Voraussage Amazon ML lernen soll, also das Zielattribut

 Note

Die Datenquelle speichert Ihre Daten nicht, sondern verweist nur darauf. Vermeiden Sie es, die in Amazon S3 gespeicherten Dateien zu verschieben oder zu ändern. Wenn Sie sie verschieben oder ändern, kann Amazon ML nicht auf sie zugreifen, um ein ML-Modell zu erstellen, Evaluierungen zu generieren oder Voraussagen zu machen.

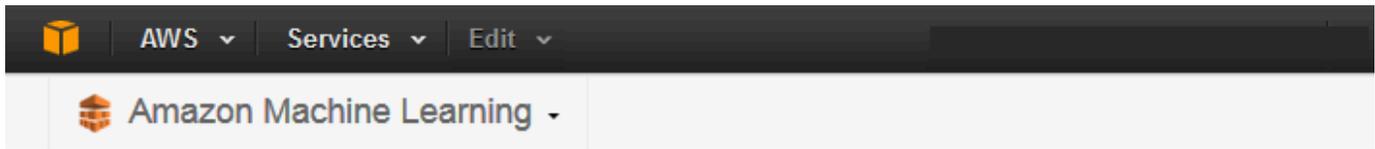
Vorgehensweise zum Erstellen der Schulungsdatenquelle

1. Öffnen Sie die Amazon Machine Learning Konsole unter <https://console.aws.amazon.com/machinelearning/> aus.
2. Wählen Sie Get started.

 Note

In diesem Tutorial wird davon ausgegangen, dass Sie zum ersten Mal Amazon ML verwenden. Wenn Sie bereits Amazon ML verwendet haben, können Sie die Erstellen eines neuen... Dropdownliste auf dem Amazon ML -Dashboard, um eine neue Datenquelle zu erstellen.

3. Auf der Ersten Schritte mit Amazon Machine Learning-Seite, wählen Starten aus.



Get started with Amazon Machine Learning



Standard setup

Start creating your first ML model. If you don't have your data ready, you can use our sample dataset.
[Amazon Machine Learning Tutorial](#)

Launch



Dashboard

Skip straight to the Amazon Machine Learning dashboard.

View Dashboard

4. Stellen Sie auf der Seite Eingabedaten sicher, dass bei Where is your data located? (Wo befinden sich Ihre Daten?) die Option S3 ausgewählt ist.

Where is your data located? S3 Redshift

5. Für S3-Speicherort, geben Sie den vollständigen Standort der `banking.csv` Datei aus Schritt 1: Vorbereiten der Daten. Beispiel: *Ihr Bucket*/**banking.csv** aus. Amazon ML stellt Ihrem Bucket-Namen `s3://` voran.
6. Geben Sie bei Datenquellenname den Wert **Banking Data 1** ein.

S3 location *

s3:// aml-sample-data/banking.csv

Enter the path to a single file or folder in Amazon S3. You need to grant Amazon ML permission to read this data. [Learn more](#).

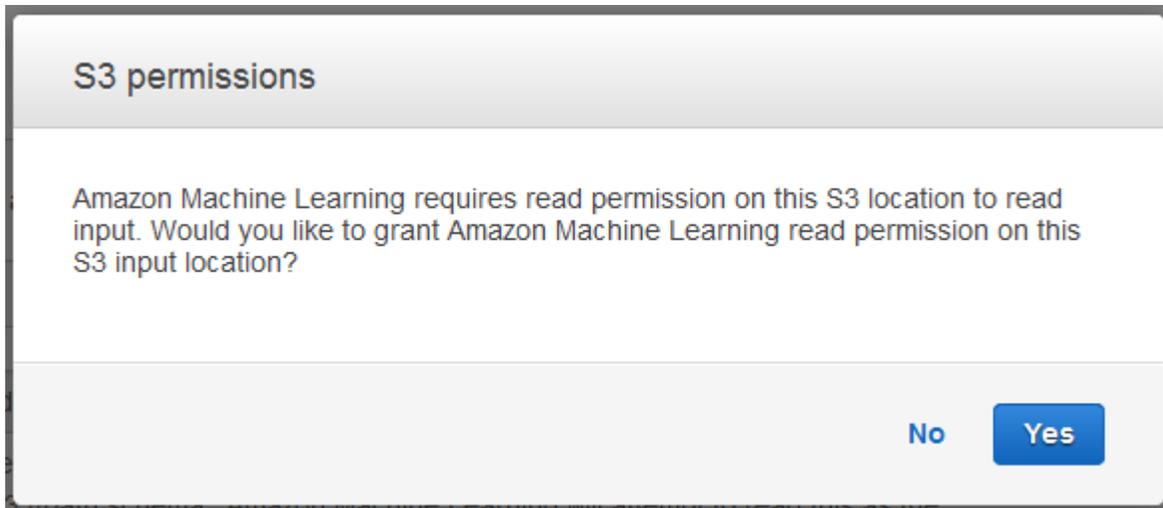
If you already have a schema for this data, provide it in a file at `s3://<path-of-input-data>.schema`. If you don't have a schema, Amazon ML will help you create one on the next page. 

Datasource name

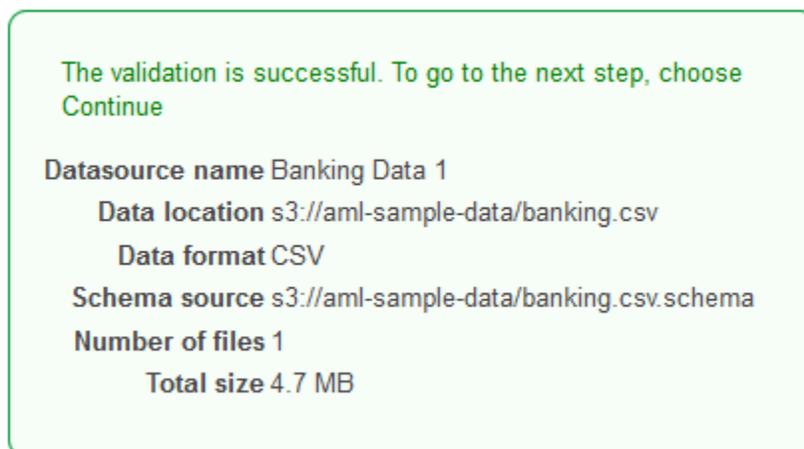
Banking Data 1

7. Wählen Sie Überprüfen.

8. Klicken Sie im Dialogfeld S3 permissions (S3-Berechtigungen) auf Ja.



9. Wenn Amazon ML auf die Datendatei am S3-Speicherort zugreifen und diese lesen kann, wird Ihnen eine Seite wie die folgende angezeigt. Überprüfen Sie die Eigenschaften und wählen Sie dann Weiter aus.



Als Nächstes erstellen Sie ein Schema. EINSchema stellt die Informationen dar, die Amazon ML zur Deutung der Input-Daten für ein ML-Modell benötigt, einschließlich Attributnamen und ihre zugeordneten Datentypen sowie die Namen besonderer Attribute. Es gibt zwei Möglichkeiten, Amazon ML ein Schema bereitzustellen:

- Stellen Sie eine separate Schemadatei bereit, wenn Sie Ihre Amazon S3 S3-Daten hochladen.
- Erlauben Sie es Amazon ML, die Attributtypen abzuleiten und ein Schema für Sie zu erstellen.

In diesem Tutorial werden wir Amazon ML ein Schema ableiten lassen.

Weitere Informationen zum Erstellen einer separaten Schemadatei finden Sie unter [Das Erstellen eines Datenschemas für Amazon ML](#).

So erlauben Sie es Amazon ML, das Ableiten eines Schemas

1. Auf der Schema-Seite zeigt Amazon ML Ihnen das abgeleitete Schema. Überprüfen Sie die Datentypen, die Amazon ML für die Attribute abgeleitet hat. Es ist wichtig, dass Attribute dem richtigen Datentyp zugeordnet sind, damit Amazon ML die Daten richtig aufnehmen kann und eine ordnungsgemäße Funktionsverarbeitung für die Attribute ermöglicht wird.
 - Attribute, für die es nur zwei mögliche Status gibt wie "Ja" oder "Nein", sollten als Binary markiert werden.
 - Attribute, die Zahlen oder Zeichenfolgen zur Kennzeichnung einer Kategorie sind, sollten als Categorical markiert werden.
 - Attribute, die numerischen Mengen sind und bei denen die Reihenfolge wichtig ist, sollten als Numeric markiert werden.
 - Attribute, die Zeichenfolgen sind und als durch Leerzeichen getrennte Wörter gehandhabt werden sollen, sollten als Text markiert werden.

<input type="checkbox"/>	Name	Data Type	Sample Field Value 1
<input type="checkbox"/>	age	Numeric	56
<input type="checkbox"/>	campaign	Numeric	1
<input type="checkbox"/>	cons_conf_idx	Numeric	-36.4
<input type="checkbox"/>	cons_price_idx	Numeric	93.994
<input type="checkbox"/>	contact	Categorical	telephone
<input type="checkbox"/>	day_of_week	Categorical	mon
<input type="checkbox"/>	default	Categorical	no
<input type="checkbox"/>	duration	Numeric	261
<input type="checkbox"/>	education	Categorical	basic.4y
<input type="checkbox"/>	emp_var_rate	Numeric	1.1

2. In diesem Tutorial hat Amazon ML die Datentypen für alle Attribute richtig identifiziert, also klicken wir auf **Continue** aus.

Wählen Sie als Nächstes ein Zielattribut aus.

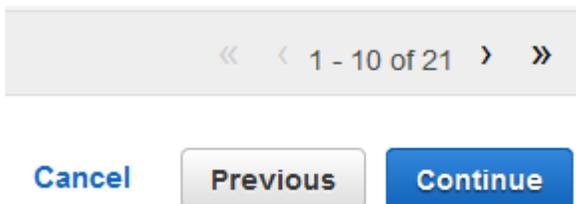
Denken Sie daran, dass das Zielattribut das Attribut ist, dessen Voraussage das ML-Modell lernen soll. Attribut *y* gibt an, ob eine Person in der Vergangenheit eine Kampagne abonniert hat: 1 (Ja) oder 0 (Nein).

Note

Wählen Sie ein Zielattribut nur aus, wenn Sie die Datenquelle für die Schulung und Evaluierung von ML-Modellen verwenden werden.

Vorgehensweise zum Auswählen von *y* als Zielattribut

1. Klicken Sie unten rechts in der Tabelle auf den einzelnen Pfeil, um zur letzten Seite der Tabelle zu gelangen, auf der das Attribut *y* angezeigt wird.



2. Wählen Sie in der Spalte Ziel den Wert *y* aus.



Amazon ML bestätigt, dass *y* als Ziel ausgewählt.

3. Klicken Sie auf **Continue**.

4. Vergewissern Sie sich, dass auf der Seite Zeilen-ID bei Does your data contain an identifier? (Enthalten Ihre Daten eine ID?) die Standardeinstellung Nein ausgewählt ist.
5. Klicken Sie auf Review und dann auf Continue.

Nun, da Sie eine Schulungsdatenquelle haben, können Sie [Ihr Modell erstellen](#).

Schritt 3: Erstellen eines ML-Modells

Nachdem Sie die Schulungsdatenquelle erstellt haben, können Sie diese verwenden, um ein ML-Modell zu erstellen, das Modell zu schulen und die Ergebnisse auszuwerten. Das ML-Modell ist eine Sammlung von Mustern, die Amazon ML während der Schulung in Ihren Daten findet. Sie verwenden das Modell, um Voraussagen zu erstellen.

Erstellen eines ML-Modells

1. Da der Assistent Erste Schritte sowohl eine Schulungsdatenquelle als auch ein Modell erstellt, verwendet Amazon Machine Learning (Amazon ML) automatisch die Schulungsdatenquelle, die Sie gerade erstellt haben, und führt Sie direkt auf die Einstellungen ML-Modell angezeigt. Stellen Sie auf der Seite ML-Modelleinstellungen sicher, dass für ML-Modellname der Standardname **ML model: Banking Data 1** angezeigt wird.

Verwenden Sie einen benutzerfreundlichen Namen, wie z. B. den Standardnamen, damit Sie das ML-Modell leicht identifizieren und verwalten können.

2. Stellen Sie sicher, dass in den Schulungs- und Auswertungseinstellungen der Wert Standard ausgewählt ist.

Select training and evaluation settings

Recipes and training parameters control the ML model training process. You can select these settings for your ML model or use the defaults provided by Amazon ML. In either case, you can choose to have Amazon ML reserve a portion of the input data for evaluation. [Learn more.](#)

Default (Recommended)

Choose this option if you want to use Amazon ML's recommended recipe, training parameters, and evaluation settings. 

Name this evaluation (Optional)

Evaluation: ML model: Banking Data 1

3. Bestätigen Sie für Diese Auswertung benennen die Standardeinstellung **Evaluation: ML model: Banking Data 1**.
4. Wählen Sie Prüfen, überprüfen Sie die Einstellungen und klicken Sie dann auf Beenden.

Nachdem Sie sich entschieden haben, fügt Amazon ML Ihr Modell zu der Verarbeitungswarteschlange hinzu. Wenn Amazon ML Ihr Modell erstellt, werden die Standardwerte angewandt und die folgenden Aktionen ausgeführt:

- Die Schulungsdatenquelle wird in zwei Abschnitte aufgeteilt, von denen einer 70 % der Daten und der andere die verbleibenden 30 % enthält.
- Schult das ML-Modell im Abschnitt, der 70 % der Eingabedaten enthält
- Wertet das Modell mit den verbleibenden 30 % der Eingabedaten aus

Während sich das Modell in der Warteschlange befindet, meldet Amazon ML den Status als **Ausstehend** aus. Während Amazon ML Ihr Modell erstellt, meldet es den Status als **In Bearbeitung** aus. Wenn alle Aktionen abgeschlossen wurden, meldet es den Status als **Abgeschlossen**. Warten Sie, bis die Auswertung abgeschlossen wurde, bevor Sie fortfahren.

Sie können jetzt [die Leistung Ihres Modells überprüfen und eine Grenzwertpunktzahl festlegen](#).

Weitere Informationen zu Schulungs- und Auswertungsmodellen finden Sie unter [Schulung von ML-Modellen](#) und [evaluate an ML model](#).

Schritt 4: Überprüfen der Voraussageleistung des ML-Modells und Festlegen eines Punktzellenwerts

Nachdem Sie Ihr ML-Modell erstellt haben und Amazon Machine Learning (Amazon ML) es bewertet hat, können Sie noch feststellen, ob es einsatzbereit ist. Während der Bewertung hat Amazon ML eine hochwertige Qualitätsmetrik berechnet, die sogenannte Area Under a Curve (AUC), welche die Leistungsqualität Ihres ML-Modells ausdrückt. Amazon ML interpretiert die AUC-Metrik außerdem, um Ihnen mitzuteilen, ob die Qualität des ML-Modells für die meisten Machine Learning-Anwendungen geeignet ist. (Weitere Informationen zur AUC finden Sie unter [Messung der ML-Modellgenauigkeit](#).) Betrachten wir nun die AUC-Metrik näher, und passen wir den Grenzwert an, um die Voraussageleistung Ihres Modells zu optimieren.

So überprüfen Sie die AUC-Metrik Ihres ML-Modells

1. Auf der Zusammenfassung ML-Modell-Seite, im ML-Modell Bericht Wählen Sie (Navigationsbereich) Auswertungen, wählen Bewertung: ML-Modell: Banking Modell 1 Klicken Sie auf und danach auf Übersicht aus.
2. Prüfen Sie auf der Seite Auswertungszusammenfassung die Auswertungszusammenfassung einschließlich der AUC-Leistungsmetrik des Modells.

ML model performance metric

On your most recent evaluation, **ev-3fF6uP2W5VL**, the ML model's quality score is considered **extremely good** for most machine learning applications. ⓘ



A horizontal bar chart showing the AUC score. The bar is a gradient from red (0) to green (1). A green triangle points to the current score of 0.94. A grey triangle points to the baseline score of 0.50.

AUC: 0.94
Baseline AUC: 0.50
Difference: 0.44

Next step: If you want to use this ML model to generate predictions, explore trade-offs to optimize the performance of your ML model first. ⓘ



A small graph showing a distribution curve with a vertical line representing the score threshold. The area under the curve to the right of the threshold is shaded red, indicating the proportion of positive predictions.

Score threshold: 0.5

[Adjust score threshold](#)

Das ML-Modell erzeugt für jeden Datensatz in einer Voraussagedatenquelle eine numerische Voraussage und wendet dann einen Grenzwert an, um diese Punktzahlen in binäre Kennzeichnungen von 0 (für Nein) und 1 (für Ja) zu konvertieren. Durch das Ändern des Punktzahlgrenzwerts können Sie anpassen, wie das ML-Modell diese Kennzeichnungen zuweist. Legen Sie nun den Punktzahlgrenzwert fest.

So legen Sie einen Punktzahlgrenzwert für Ihr ML-Modell fest

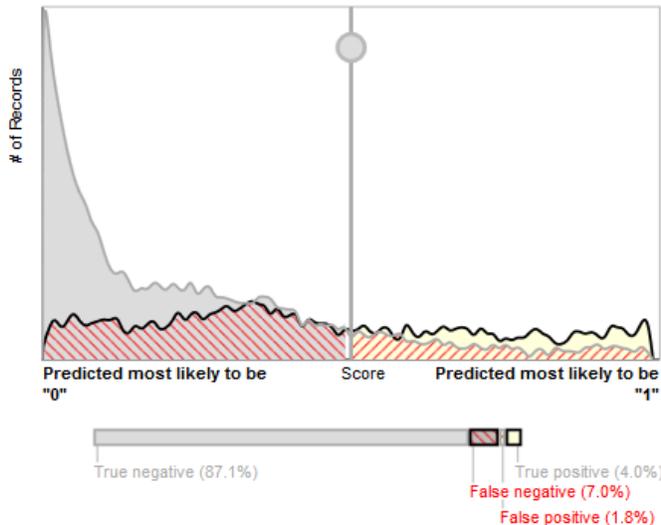
1. Wählen Sie auf der Seite Auswertungszusammenfassung die Option Punktzahlgrenzwert anpassen.

ML model performance

This chart shows the distributions of your predicted answers for the actual "1" and "0" records in your evaluation data. Any overlap of the actual "1"  & "0"  is where your ML model guesses wrong. [Learn more](#).

Adjust the slider to indicate how much error you can tolerate from your ML model based on your needs. Moving the score threshold to the right decreases the number of false positives and increases the number of false negatives.

[Explain this chart](#)



Trade-off based on score threshold

[Reset score threshold \(0.5\)](#)

- **91% are correct**
500 true positive
10,766 true negative
- **9% are errors**
226 false positive
863 false negative

- 6% of the records are predicted as "1"
- 94% of the records are predicted as "0"

[Save score threshold at 0.50](#)

▼ Advanced metrics

Accuracy 0.9119	0	<input type="range"/>	1
False positive rate 0.0206	0	<input type="range"/>	1
Precision 0.6887	0	<input type="range"/>	1
Recall 0.3668	0	<input type="range"/>	1

Sie können die Leistungsmetriken Ihres ML-Modells einstellen, indem Sie den Punktzahlgrenzwert anpassen. Durch die Anpassung dieses Wertes ändert sich das Vertrauen, dass das Modell in eine Voraussage haben muss, bevor es die Voraussage als positiv erachtet. Außerdem ändert sich, wie viele Falschmeldungen Sie in Ihren Voraussagen tolerieren.

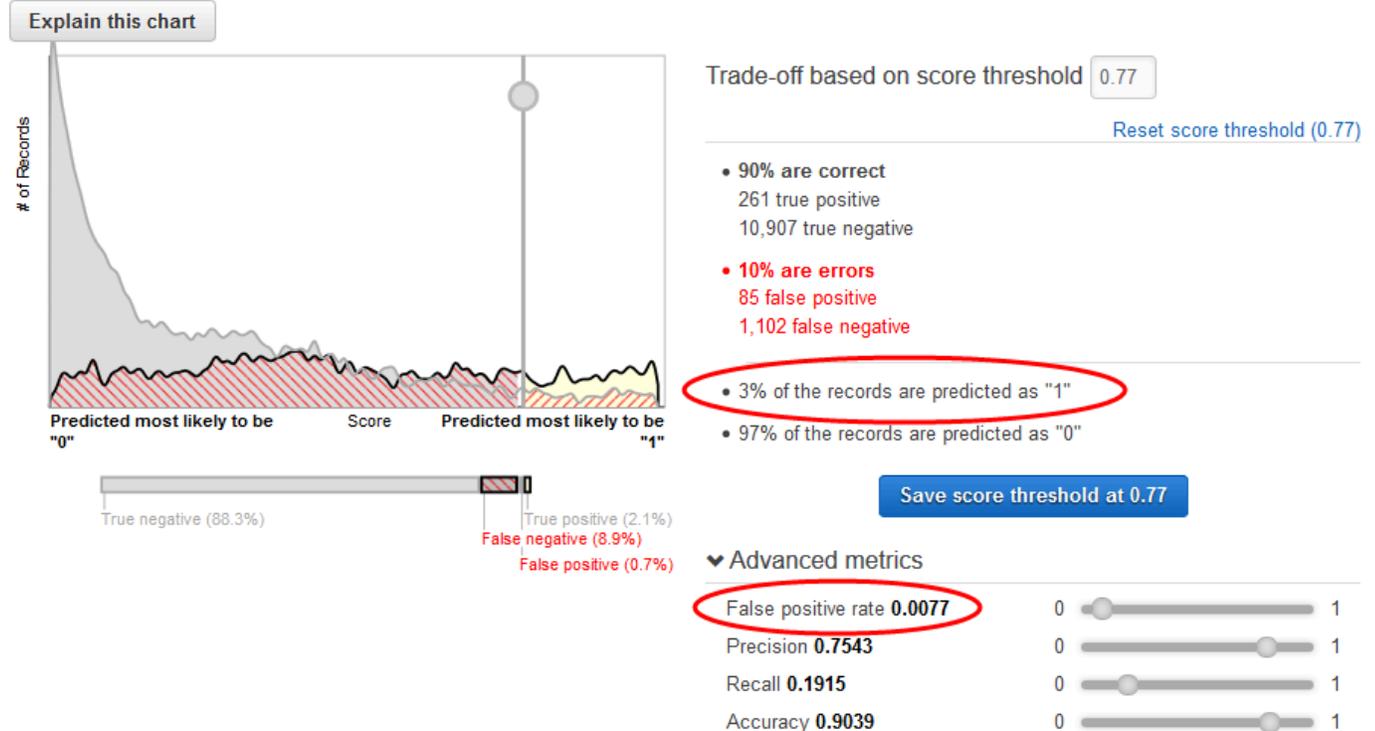
Sie können den Grenzwert für eine positive Voraussage kontrollieren, indem Sie den Punktzahlgrenzwert erhöhen, bis nur die Voraussagen als positiv erachtet werden, die am wahrscheinlichsten echte positive Voraussagen sind. Sie können den Punktzahlgrenzwert auch so weit reduzieren, bis keine negativen Falschmeldungen mehr auftreten. Wählen Sie Ihren Grenzwert entsprechend Ihren betrieblichen Anforderungen. Für dieses Tutorial kostet jede positive Falschmeldung das Unternehmen Geld; wir möchten also ein möglichst hohes Verhältnis von positiven zu negativen Falschmeldungen.

2. Angenommen, Sie möchten die obersten 3 % der Kunden berücksichtigen, die das Produkt abonnieren werden. Verschieben Sie den vertikalen Selektor so, dass der Wert des Punktzahlgrenzwerts 3 % der als "1" vorhergesagten Datensätze entspricht.

ML model performance

This chart shows the distributions of your predicted answers for the actual "1" and "0" records in your evaluation data. Any overlap of the actual "1"  & "0"  is where your ML model guesses wrong. [Learn more](#).

Adjust the slider to indicate how much error you can tolerate from your ML model based on your needs. Moving the score threshold to the right decreases the number of false positives and increases the number of false negatives.



Beachten Sie die Auswirkungen dieses Punktzahlgrenzwerts auf die Leistung des ML-Modells: Die Rate der positiven Falschmeldungen beträgt 0,007. Angenommen, diese Rate ist akzeptabel.

3. Wählen Sie Punktzahlgrenzwert bei 0,77 speichern.

Jedes Mal, wenn Sie dieses ML-Modell nutzen, um Voraussagen zu machen, werden Datensätze mit Punktzahlen über 0,77 als "1" und der Rest der Datensätze als "0" gekennzeichnet.

Weitere Informationen zum Punktzahlgrenzwert finden Sie unter [Binäre Klassifizierung](#).

Jetzt können Sie [Mit dem Modell Voraussagen erstellen](#).

Schritt 5: Verwenden des ML-Modells zum Generieren von Prognosen

Amazon Machine Learning (Amazon ML) kann zwei Arten von Prognosen generieren — Batch und Echtzeit.

EINEchtzeitprognose ist eine Prognose für eine einzelne Beobachtung, die Amazon ML bei Bedarf generiert. Echtzeitvoraussagen sind ideal für mobile Apps, Websites und andere Anwendungen, die Ergebnisse interaktiv verwenden sollen.

EINStapelprognose ist eine Reihe von Prognosen für eine Gruppe von Beobachtungen. Amazon ML verarbeitet die Datensätze in einer Batch-Prognose zusammen, sodass die Verarbeitung einige Zeit in Anspruch nehmen kann. Verwenden Sie Stapelvoraussagen für Anwendungen, die Voraussagen für Gruppen von Beobachtungen benötigen oder Voraussagen, die Ergebnisse nicht interaktiv verwenden.

Für dieses Tutorial generieren Sie eine Echtzeit-Voraussage, die vorhersagt, ob ein potenzieller Kunde das neue Produkt abonnieren wird. Zudem können Sie Voraussagen für einen großen Batch potenzieller Kunden generieren. Für die Stapelvoraussage verwenden Sie die Datei `banking-batch.csv`, die Sie in [Schritt 1: Vorbereiten der Daten](#) hochgeladen haben.

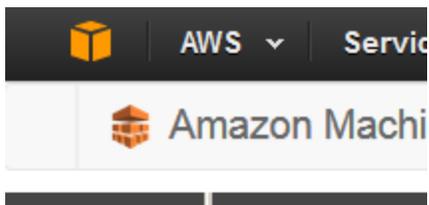
Lassen Sie uns mit einer Echtzeitvoraussage beginnen.

Note

Für Anwendungen, die Echtzeitvoraussagen erfordern, müssen Sie einen Echtzeit-Endpunkt für das ML-Modell erstellen. Es fallen Kosten für Sie an, während ein Echtzeit-Endpunkt verfügbar ist. Bevor Sie Echtzeitvoraussagen tatsächlich nutzen und dabei Kosten anfallen, können Sie die Echtzeitvoraussage-Funktion testweise in Ihrem Web-Browser verwenden, ohne Echtzeit-Endpunkt. Das reicht für dieses Tutorial aus.

So testen Sie eine Echtzeitvoraussage

1. Klicken Sie im Navigationsbereich ML model report auf Try real-time predictions.



ML model report

Summary

Settings

Monitoring

Tools

Try real-time predictions

2. Wählen Sie Paste a record aus.

Try real-time predictions

Try generating real-time predictions for free using the web browser on this page. To request a real-time prediction, complete the following form or provide a single data record in CSV format. To provide a data record, choose the **Paste a record** button.

Paste a record

<input type="text" value="Attribute name"/>	Items per page: <input type="text" value="10"/>	« < 1 - 10 of 21 > »
Name	Type	Value

3. Fügen Sie im Dialogfeld Paste a record die folgende Beobachtung ein:

32, services, divorced, basic.9y, no, unknown, yes, cellular, dec, mon, 110, 1, 11, 0, nonexistent, -1.8, 9

4. In der Fügen Sie einen Datensatz ein wählen Sie Sendenum zu bestätigen, dass Sie eine Prognose für diese Beobachtung generieren möchten. Amazon ML füllt die Werte in Echtzeitprognose-Form auf.

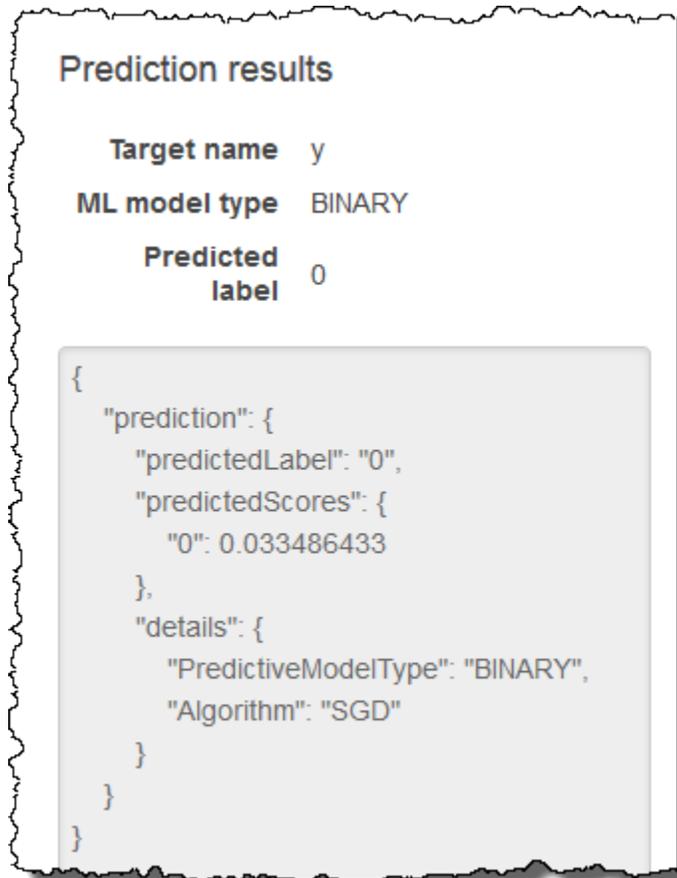
<input type="text" value="Attribute name"/>	Items per page: <input type="text" value="10"/>	« < 1 - 10 of 21 > »	
Name	Type	Value	
1	age	Numeric	<input type="text" value="32.0"/>

Note

Sie können auch die Wert-Felder auffüllen, indem Sie einzelne Werte eingeben. Unabhängig von der Methode, die Sie auswählen, sollten Sie eine Beobachtung bereitstellen, die nicht zur Modellschulung verwendet wurde.

5. Klicken Sie unten auf der Seite auf Create prediction.

Die Voraussage wird im Bereich Prediction results auf der rechten Seite angezeigt. Diese Voraussage hat eine Predicted (Voraussage)-Kennung von 0, was bedeutet, dass diese potenziellen Kunden wahrscheinlich nicht auf die Kampagne antworten. Eine Predicted (Voraussage)-Kennung von 1 würde bedeuten, dass der Kunde wahrscheinlich auf die Kampagne antwortet.

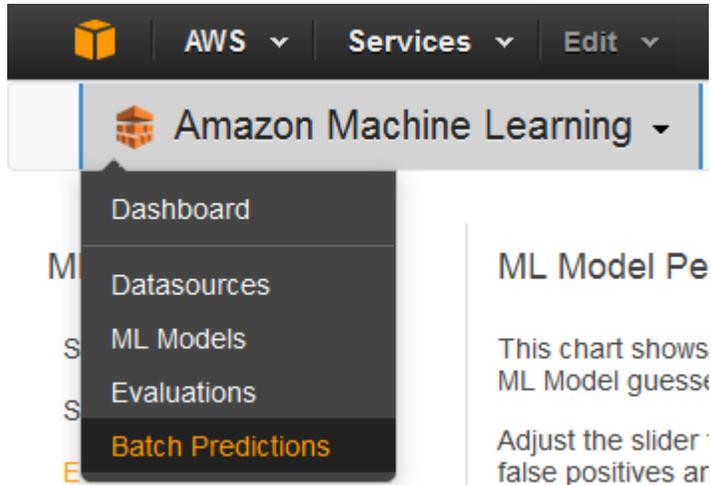


Erstellen Sie nun eine Stapelvoraussage. Sie geben Amazon ML den Namen des ML-Modells an, das Sie verwenden; das Verzeichnis in Amazon Simple Storage Service (Amazon S3) der Eingabedaten,

für die Sie Prognosen generieren möchten (Amazon ML wird aus diesen Daten eine Batch-Prognose-Datenquelle erstellen); und das Amazon-S3-Verzeichnis für die Speicherung der Ergebnisse.

So erstellen Sie eine Stapelvoraussage

1. Wählen Sie Amazon Machine Learning und dann Stapelvoraussagen aus.



2. Wählen Sie Create new batch prediction.
3. Auf der ML-Modell für Chargenprognosen-Seite wählen ML-Modell: Banking Daten 1 aus.

Amazon ML zeigt den ML-Modellnamen, die ID, den Zeitpunkt der Erstellung und die zugehörige Datenquellen-ID an.

4. Klicken Sie auf Continue.
5. Zum Generieren von Prognosen müssen Sie Amazon ML die Daten bereitstellen, für die Sie Prognosen benötigen. Diese werden als Eingabedaten bezeichnet. Geben Sie die Eingabedaten zunächst in eine Datenquelle ein, sodass Amazon darauf zugreifen kann.

Wählen Sie unter Locate the input data (Eingabedaten lokalisieren) die Option My data is in S3, and I need to create a datasource (Meine Daten befinden sich in S3 und ich muss eine Datenquelle erstellen).

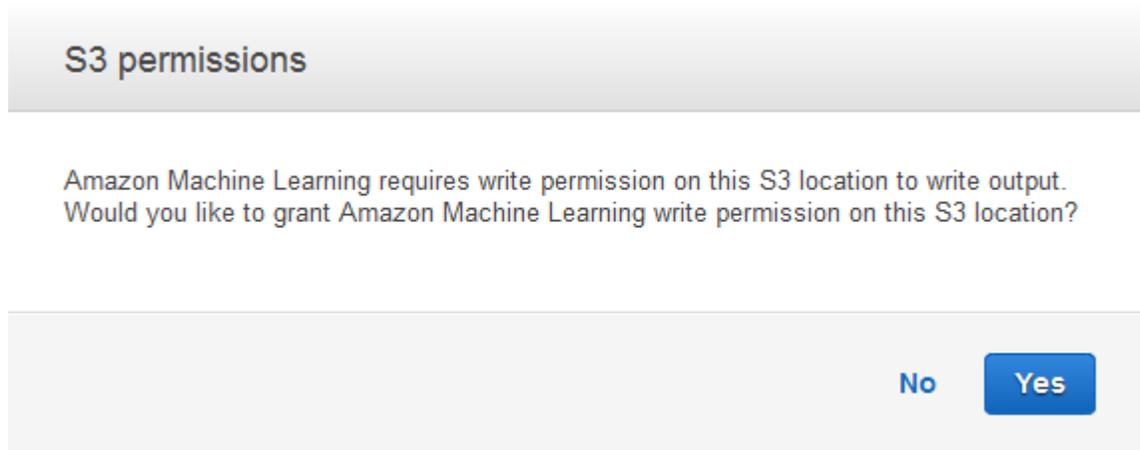
Locate the input data I already created a datasource pointing to my S3 data
 My data is in S3, and I need to create a datasource

6. Geben Sie bei Datenquellenname den Wert **Banking Data 2** ein.
7. Geben Sie als S3 Location (S3-Speicherort) den vollständigen Speicherort der Datei `banking-batch.csv` ein: *Ihr-Bucket/banking-batch.csv*.

8. Legen Sie für Does the first line in your CSV contain the column names? (Enthält die erste Zeile Ihrer CSV-Datei die Spaltennamen?) den Wert Ja fest.
9. Wählen Sie Überprüfen.

Amazon ML validiert den Speicherort Ihrer Daten.

10. Klicken Sie auf Continue.
11. Für S3-Ziel Geben Sie in „Schritt 1“ den Namen des Amazon-S3-Speicherorts an, in den Sie die Dateien hochgeladen haben: Vorbereiten der Daten. Amazon ML lädt die Prognoseergebnisse dorthin.
12. Für Stapelvoraussage-Name akzeptieren Sie die Standardeinstellung, **Batch prediction: ML model: Banking Data 1** aus. Amazon ML wählt den Standardnamen basierend auf dem Modell aus, das zum Erstellen von Prognosen verwendet wird. In diesem Tutorial wird das Modell sowie die Voraussagen nach dem Namen Schulungsdatenquelle benannt: Banking Data 1.
13. Wählen Sie Review.
14. Klicken Sie im Dialogfeld S3 permissions (S3-Berechtigungen) auf Ja.

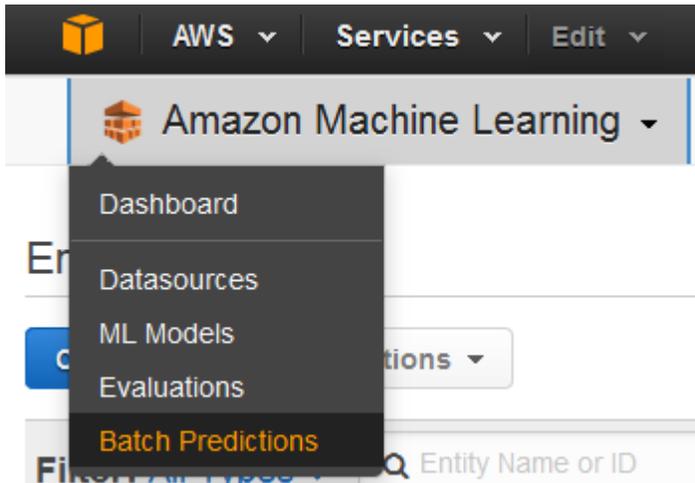


15. Klicken Sie auf der Seite Prüfen auf Beenden.

Die Batch-Prognose-Anforderung wird an Amazon ML gesendet und in eine Warteschlange gestellt. Die Zeit, die Amazon ML für die Verarbeitung einer Batch-Prognose benötigt, hängt von der Größe der Datenquelle und der Komplexität Ihres ML-Modells ab. Während Amazon ML die Anforderung verarbeitet, lautet der Status von In Bearbeitung aus. Nachdem die Stapelvoraussage abgeschlossen ist, ändert sich der Status der Anforderung in Abgeschlossen. Jetzt können Sie die Ergebnisse anzeigen.

So zeigen Sie die Voraussagen an

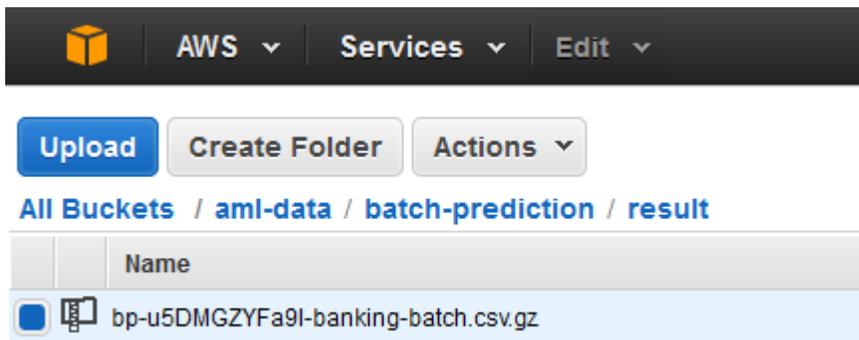
1. Wählen Sie Amazon Machine Learning und dann Stapelvoraussagen aus.



2. Klicken Sie in der Liste der Prognosen auf Stapelprognose: ML-Modell: Banking Daten 1 aus. Die Seite Batch prediction info (Informationen zur Stapelvoraussage) wird angezeigt.

Name	Subscription propensity Predictions 
ID	bp-u5DMGZYFa9l
Creation Time	Mar 5, 2015 3:28:33 PM
Status	Completed
Log	Download Log
Datasource ID	ds-33Rqgz9w3ee
ML Model ID	ml-u7ljoShX2kX
Input S3 URL	s3://aml-data/banking-batch.csv
Output S3 URL	s3://aml-data/

3. Öffnen Sie zum Anzeigen der Ergebnisse der Batch-Prognose die Amazon-S3-Konsole unter <https://console.aws.amazon.com/s3/> und navigieren Sie zum Amazon-S3-Speicherort, auf den im S3-URL ausgegebenfeld. Von dort navigieren Sie zum Ergebnisordner, der etwa folgenden Namen hat: s3://aml-data/batch-prediction/result.



Die Voraussage Schlüssel wird in einer komprimierten gzip-Datei mit der Erweiterung .gz gespeichert.

- Laden Sie die Voraussagedatei auf Ihren Desktop herunter, wo Sie sie entpacken und öffnen können.

bestAnswer	score
0	0.06046
0	0.00507
0	0.01410
0	0.00170
0	0.00184
0	0.07133
0	0.30811

Die Datei verfügt über zwei Spalten: bestAnswer und score, sowie eine Zeile für jede Beobachtung in der Datenquelle. Die Ergebnisse in der Spalte bestAnswer basieren auf dem Punktzahlgrenzwert von 0,77, den Sie in [Schritt 4: Überprüfen der Voraussageleistung des ML-Modells und Festlegen eines Punktzellenwerts](#) festgelegt haben. Eine Punktzahl größer als 0,77 führt zu einer bestAnswer von 1. Dabei handelt es sich um eine positive Antwort oder Voraussage. Eine Punktzahl von weniger als 0,77 ergibt als bestAnswer 0. Dabei handelt es sich um eine negative Antwort oder Voraussage.

Die folgenden Beispiele zeigen positive und negative Voraussagen basierend auf den Schwellenwert von 0,77.

Positive Voraussage:

bestAnswer	score
1	0.8228876

In diesem Beispiel beträgt der Wert für `bestAnswer` 1 und die Punktzahl ist 0,8228876. Der Wert für `bestAnswer` ist 1, da die Punktzahl größer als der Schwellenwert von 0,77 ist. Eine `bestAnswer` von 1 gibt an, dass der Kunde Ihr Produkt wahrscheinlich kaufen wird, und ist somit eine positive Voraussage.

Negative Voraussage:

<code>bestAnswer</code>	<code>score</code>
0	0.7695356

In diesem Beispiel ist der Wert von `bestAnswer` 0, da die Punktzahl 0,7695356 ist und somit unter dem Schwellenwert von 0,77 liegt. Die `bestAnswer` 0 gibt an, dass der Kunden Ihr Produkt wahrscheinlich nicht kaufen wird, und ist somit eine negative Voraussage.

Jede Zeile des Stapelergebnisses entspricht einer Zeile in Ihrer Stapeleingabe (einer Beobachtung in Ihrer Datenquelle).

Nach der Analyse der Voraussagen können Sie die gewünschte Marketing-Kampagne durchführen, indem Sie z. B. allen mit einer vorausgesagten Punktzahl von 1 einen Flyer schicken.

Nachdem Sie Ihr Modell erstellt, geprüft und verwendet haben, [bereinigen Sie die erstellten Daten und AWS-Ressourcen](#), um zu vermeiden, dass hierfür unnötige Gebühren anfallen, und damit Ihr Arbeitsbereich übersichtlich bleibt.

Schritt 6: Bereinigen

Um zu vermeiden, dass Ihnen weitere Amazon Simple Storage Service (Amazon S3) -Gebühren berechnet werden, können Sie die in Amazon S3 gespeicherten Daten löschen. Ihnen werden keine Gebühren für andere ungenutzte Amazon ML-Ressourcen berechnet, aber wir empfehlen Ihnen, sie zu löschen, damit Ihr Workspace ordentlich bleibt.

So löschen Sie die in Amazon S3 gespeicherten Eingabedaten

1. Öffnen Sie die Amazon S3-Konsole unter <https://console.aws.amazon.com/s3/>.
2. Navigieren Sie zum Amazon S3 S3-Verzeichnis, an dem Sie `denbanking.csv` und `denbanking-batch.csv` Dateien.
3. Wählen Sie die `banking.csv`-, `banking-batch.csv`- und `.writePermissionCheck.tmp`-Dateien aus.
4. Wählen Sie Aktionen und anschließend Löschen.

5. Wenn Sie zur Bestätigung aufgefordert werden, klicken Sie auf OK.

Es fallen zwar keine Kosten für Sie an, wenn Sie den Datensatz der von Amazon ML durchgeführten Stapelvoraussage oder die Datenquellen, das Modell und die Evaluierung behalten, die Sie im Tutorial erstellt haben; wir empfehlen Ihnen aber dennoch, sie zu löschen, um Ordnung in Ihre Workspace herzustellen.

Vorgehensweise zum Löschen von Stapelvoraussagen

1. Navigieren Sie zum Amazon S3 S3-Verzeichnis, an dem Sie die Ausgabe der Stapelvoraussagen gespeichert haben.
2. Wählen Sie den Ordner `batch-prediction` aus.
3. Wählen Sie Aktionen und anschließend Löschen.
4. Wenn Sie zur Bestätigung aufgefordert werden, klicken Sie auf OK.

So löschen Sie die Amazon ML-Ressourcen

1. Wählen Sie die folgenden Ressourcen im Amazon ML-Dashboard aus.
 - Die Banking Data 1-Datenquelle
 - Die Banking Data 1_[percentBegin=0, percentEnd=70, strategy=sequential]-Datenquelle
 - Die Banking Data 1_[percentBegin=70, percentEnd=100, strategy=sequential]-Datenquelle
 - Die Banking Data 2-Datenquelle
 - Das ML model: Banking Data 1-ML-Modell
 - Die Evaluation: ML model: Banking Data 1-Evaluierung
2. Wählen Sie Aktionen und anschließend Löschen.
3. Klicken Sie im Dialogfeld auf Delete, um alle ausgewählten Ressourcen zu löschen.

Sie haben das Tutorial jetzt erfolgreich abgeschlossen. Informieren Sie sich für die weitere Verwendung der Konsole zum Erstellen von Datenquellen, Modellen und Voraussagen im [Amazon Machine Learning Learning-Entwicklerleit](#)aus. Weitere Informationen zur Verwendung der API finden Sie in der [Amazon Machine Learning-API-Referenz](#).

Erstellen und Verwenden von Datenquellen

Sie können Amazon ML-Datenquellen verwenden, um ein ML-Modell zu schulen, auszuwerten und Stapelvoraussagen mithilfe eines ML-Modells zu generieren. Datenquellenobjekte enthalten Metadaten über Ihre Eingabedaten. Wenn Sie eine Datenquelle erstellen, liest Amazon ML die Eingabedaten, berechnet beschreibende Statistiken zu deren Attributen und speichert die Statistiken zusammen mit einem Schema und weiteren Informationen als Bestandteil des Datenquellenobjekts. Nachdem Sie eine Datenquelle erstellt haben, können Sie das [Dateneinblicke in Amazon ML](#) um statistische Eigenschaften Ihrer Eingabedaten zu erkunden, können Sie die Datenquelle verwenden, um [ML-Modell trainieren](#) aus.

Note

In diesem Abschnitt wird davon ausgegangen, dass Sie mit vertraut sind [Amazon Machine Learning Learning-Konzept](#) aus.

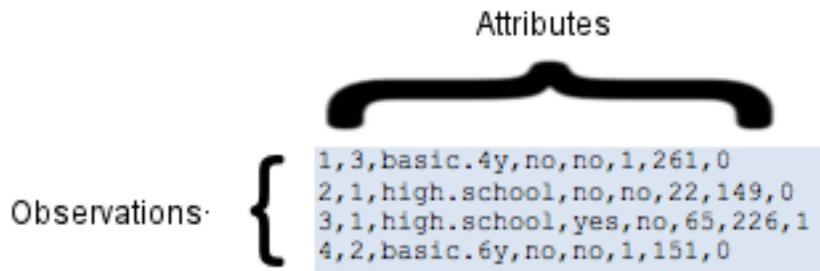
Themen

- [Erwähnungen zum Datenformat für Amazon ML](#)
- [Das Erstellen eines Datenschemas für Amazon ML](#)
- [Aufteilen Ihrer Daten](#)
- [Dateneinblicke](#)
- [Verwenden von Amazon S3 mit Amazon ML](#)
- [Erstellen einer Amazon ML-Datenquelle aus Daten in Amazon Redshift](#)
- [Verwenden von Daten aus einer Amazon RDS-Datenbank zum Erstellen einer Amazon ML-Datenquelle](#)

Erwähnungen zum Datenformat für Amazon ML

Eingabedaten sind die Daten, die Sie zur Erstellung einer Datenquelle verwenden. Speichern Sie Ihre Eingabedaten als durch Kommas getrennte Werte (CSV). Jede Zeile in der CSV-Datei ist ein einzelner Datensatz bzw. eine Beobachtung. Jede Spalte in der CSV-Datei enthält ein Attribut der Beobachtung. Die folgende Abbildung zeigt den Inhalt einer CSV-Datei, die vier Beobachtungen enthält, die jeweils in einer eigenen Zeile stehen. Jede Beobachtung enthält acht Attribute, die durch ein Komma getrennt sind. Die Attribute stehen für die folgenden

Informationen zu jedem einzelnen Objekt, das durch eine Beobachtung dargestellt ist: `customerId,jobId,education,housing,loan,campaign,duration,willRespondToCampaign`.



Attribute

Für Amazon ML erfordert Namen für jedes Attribut. Sie können Attributnamen wie folgt angeben:

- Einbeziehen der Attributnamen in der ersten Zeile der CSV-Datei (auch als Kopfzeile bezeichnet), die Sie für Eingabedaten verwenden
- Einbeziehen der Attributnamen in einer separaten Schemadatei im selben S3-Bucket wie die Eingabedaten

Weitere Informationen zur Verwendung von Schemadateien finden Sie unter [Erstellen eines Datenschemas](#).

Im folgenden Beispiel für eine CSV-Datei sind die Namen der Attribute in der Kopfzeile enthalten.

```

customerId,jobId,education,housing,loan,campaign,duration,willRespondToCampaign

1, 3, basic.4y, no, no, 1, 261, 0

2, 1, high.school, no, no, 22, 149, 0

3, 1, high.school, yes, no, 65, 226, 1

4, 2, basic.6y, no, no, 1, 151, 0
  
```

Anforderungen an das Eingabedateiformat

Die CSV-Datei, die Ihre Eingabedaten enthält, muss die folgenden Anforderungen erfüllen:

- Muss reiner Text mit einem Zeichensatz wie z. B. ASCII, Unicode oder EBCDIC sein.
- Besteht aus Beobachtungen, eine Beobachtung pro Zeile.
- Für jede Beobachtung müssen die Attributwerte durch Komma getrennt werden.
- Wenn ein Attributwert ein Komma enthält (das Trennzeichen), muss der gesamte Attributwert in Anführungszeichen gesetzt werden.
- Jede Beobachtung muss mit einem Zeilenendezeichen beendet werden. Dabei handelt es sich um ein Sonderzeichen oder eine Zeichenfolge, die das Ende einer Zeile kennzeichnet.
- Attributwerte dürfen keine Zeilenendezeichen enthalten, auch wenn der Attributwert in Anführungszeichen eingeschlossen wird.
- Jede Beobachtung muss die gleiche Anzahl von Attributen und Folge von Attributen aufweisen.
- Jede Beobachtung darf nicht größer als 100 KB sein. Amazon ML lehnt jede Beobachtung, die größer als 100 KB ist, während der Verarbeitung ab. Wenn Amazon ML mehr als 10.000 Beobachtungen ablehnt, wird die gesamte CSV-Datei abgelehnt.

Verwenden mehrerer Dateien als Dateneingabe für Amazon ML

Sie können Ihre Eingabedateien für Amazon ML als einzelne Datei oder als Sammlung von Dateien bereitstellen. Sammlungen müssen folgende Bedingungen erfüllen:

- Alle Dateien müssen dasselbe Datenschema haben.
- Alle Dateien müssen im selben Amazon Simple Storage Service (Amazon S3) -Präfix gespeichert sein, und der Pfad, den Sie für die Sammlung angeben, muss mit einem Schrägstrich („/“) enden.

Wenn Ihre Datendateien zum Beispiel `input1.csv`, `input2.csv` und `input3.csv` heißen und Ihr S3-Bucket-Name `"s3://examplebucket"` lautet, könnten Ihre Dateipfade wie folgt aussehen:

```
s3://examplebucket/path/to/data/input1.csv
```

```
s3://examplebucket/path/to/data/input2.csv
```

```
s3://examplebucket/path/to/data/input3.csv
```

Sie würden den folgenden S3-Speicherort als Eingabe für Amazon ML verwenden:

```
's3://examplebucket/path/to/data/'
```

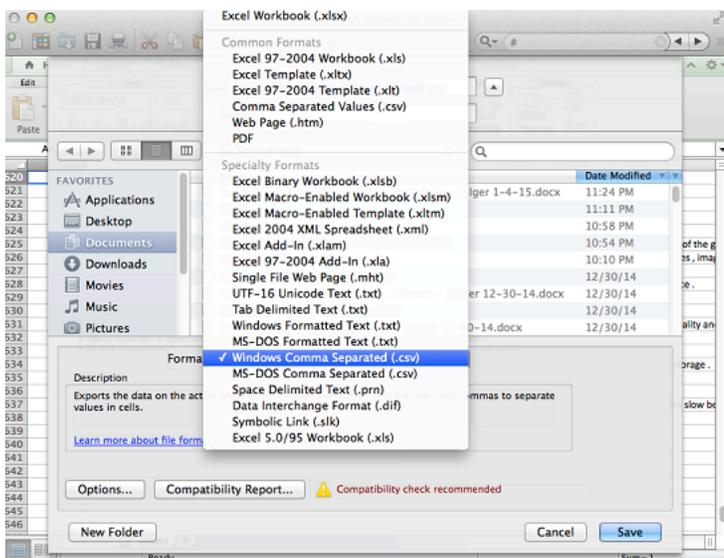
Zeilenendezeichen im CSV-Format

Wenn Sie Ihre CSV-Datei erstellen, wird jede Beobachtung mit einem speziellen Zeilenendezeichen beendet. Dieses Zeichen ist nicht sichtbar, wird aber automatisch am Ende jeder Beobachtung eingefügt, wenn Sie die Eingabetaste oder die Enter-Taste drücken. Das Sonderzeichen, welches das Ende der Zeile kennzeichnet, variiert je nach Betriebssystem. Unix-Systeme, wie z. B. Linux oder OS X, verwenden ein Zeilenvorschubzeichen, das durch "\n" (ASCII-Code 10 in Dezimalcode oder 0x0a in Hexadezimalcode) dargestellt wird. Microsoft Windows verwendet zwei Zeichen namens Wagenrücklauf und Zeilenvorschub, die mit "\r\n" (ASCII-Codes 13 und 10 in Dezimalcode oder 0x0d und 0x0a in Hexadezimalcode) dargestellt werden.

Wenn Sie OS X und Microsoft Excel zum Erstellen der CSV-Datei verwenden möchten, führen Sie die im Folgenden beschriebene Vorgangsweise aus. Stellen Sie sicher, dass Sie das richtige Format wählen.

So speichern Sie eine CSV-Datei, wenn Sie OS X und Excel verwenden

1. Beim Speichern der CSV-Datei wählen Sie Format und dann Windows Comma Separated (.csv).
2. Wählen Sie Save (Speichern) aus.



! Important

Speichern Sie die .csv-Datei nicht mithilfe des Durch Kommas getrennte Werte (.csv) oder MS-DOS Komma getrennt (.csv) Formate, weil Amazon ML sie nicht lesen kann.

Das Erstellen eines Datenschemas für Amazon ML

Ein Schema umfasst alle Attribute der Eingabedaten und die entsprechenden Datentypen. Hiermit kann Amazon ML die Daten in der Datenquelle verstehen. Amazon ML verwendet die Informationen im Schema zum Lesen und Interpretieren der Eingabedaten, zur Berechnung der Statistiken, zur Anwendung der richtigen Attributtransformationen und zur Einstellung der Lernalgorithmen. Wenn Sie kein Schema bereitstellen, leitet Amazon ML eines aus den Daten ab.

Beispielschema

Damit Amazon ML die Eingabedaten richtig lesen und genaue Voraussagen treffen kann, muss jedes Attribut dem richtigen Datentyp zugewiesen werden. Im Folgenden finden Sie ein Beispiel für die Zuweisung von Datentypen zu Attributen und die Berücksichtigung von Attributen und Datentypen in einem Schema. Wir nennen unser Beispiel "Kundenkampagne", da wir voraussagen möchten, welche Kunden auf unsere E-Mail-Kampagne reagieren werden. Unsere Eingabedatei ist eine CSV-Datei mit neun Spalten:

```
1,3,web developer,basic.4y,no,no,1,261,0
2,1,car repair,high.school,no,no,22,149,0
3,1,car mechanic,high.school,yes,no,65,226,1
4,2,software developer,basic.6y,no,no,1,151,0
```

Dies ist das Schema für diese Daten:

```
{
  "version": "1.0",
  "rowId": "customerId",
  "targetAttributeName": "willRespondToCampaign",
  "dataFormat": "CSV",
  "dataFileContainsHeader": false,
  "attributes": [
    {
      "attributeName": "customerId",
      "attributeType": "CATEGORICAL"
    },
    {
      "attributeName": "jobId",
```

```
        "attributeType": "CATEGORICAL"
    },
    {
        "attributeName": "jobDescription",
        "attributeType": "TEXT"
    },
    {
        "attributeName": "education",
        "attributeType": "CATEGORICAL"
    },
    {
        "attributeName": "housing",
        "attributeType": "CATEGORICAL"
    },
    {
        "attributeName": "loan",
        "attributeType": "CATEGORICAL"
    },
    {
        "attributeName": "campaign",
        "attributeType": "NUMERIC"
    },
    {
        "attributeName": "duration",
        "attributeType": "NUMERIC"
    },
    {
        "attributeName": "willRespondToCampaign",
        "attributeType": "BINARY"
    }
]
}
```

In der Schemadatei für dieses Beispiel hat `rowId` den Wert `customerId`:

```
"rowId": "customerId",
```

Das Attribut `willRespondToCampaign` ist als Zielattribut festgelegt:

```
"targetAttributeName": "willRespondToCampaign ",
```

Das Attribut `customerId` und der Datentyp `CATEGORICAL` sind der ersten Spalte zugewiesen, das Attribut `jobId` und der Datentyp `CATEGORICAL` sind der zweiten Spalte zugewiesen, das Attribut `jobDescription` und der Datentyp `TEXT` sind der dritten Spalte zugewiesen, das Attribut `education` und der Datentyp `CATEGORICAL` sind der vierten Spalte zugewiesen usw. Die neunte Spalte ist dem Attribut `willRespondToCampaign` und dem Datentyp `BINARY` zugewiesen, und dieses Attribut ist auch als Zielattribut festgelegt.

Verwenden des Felds `targetAttributeName`

Der Wert `targetAttributeName` ist der Name des Attributs, das Sie voraussagen möchten. Sie müssen `targetAttributeName` bei der Erstellung oder der Bewertung eines Modells zuweisen.

Wenn Sie ein ML-Modell trainieren oder bewerten, wird die `targetAttributeName` identifiziert den Namen des Attributs in den Eingabedaten, welche die „richtigen“ Antworten für das Zielattribut enthalten. Amazon ML verwendet das Ziel mit den richtigen Antworten, um Muster zu erkennen und ein ML-Modell zu erstellen.

Wenn Sie Ihr Modell bewerten, verwendet Amazon ML das Ziel, um die Richtigkeit Ihrer Voraussagen zu prüfen. Nachdem Sie das ML-Modell erstellt und bewertet haben, können Sie Daten ohne zugewiesenen `targetAttributeName` verwenden, um Voraussagen mit Ihrem ML-Modell zu erzeugen.

Sie definieren das Zielattribut in der Amazon ML-Konsole, sobald Sie eine Datenquelle erstellen, oder in einer Schema-Datei. Wenn Sie eine eigene Schemadatei erstellen, verwenden Sie die folgende Syntax, um das Zielattribut festzulegen:

```
"targetAttributeName": "exampleAttributeTarget",
```

In diesem Beispiel ist `exampleAttributeTarget` der Name des Attributs in Ihrer Eingabedatei, welches das Zielattribut ist.

Verwenden des Felds `rowID`

Die `row ID` ist ein optionales Flag, das einem Attribut in den Eingabedaten zugewiesen ist. Wenn festgelegt, ist das als `row ID` markierte Attribut im Voraussageergebnis enthalten. Anhand dieses Attributs kann einfacher zugeordnet werden, welche Voraussage welcher Beobachtung entspricht. Ein Beispiel für eine gute `row ID` ist eine Kunden-ID oder ein ähnliches eindeutiges Attribut.

Note

Die Zeilen-ID dient nur zu Ihrer Referenz. Amazon ML verwendet sie nicht für die Schulung eines ML-Modells. Wenn ein Attribut als Zeilen-ID festgelegt wird, kann es nicht mehr zur Schulung eines ML-Modells verwendet werden.

Sie definieren die `row ID` Wenn Sie eine Datenquelle erstellen, oder in einer Schema-Datei. Wenn Sie eine eigene Schemadatei erstellen, verwenden Sie die folgende Syntax, um die `row ID` festzulegen:

```
"rowId": "exampleRow",
```

Im voranstehenden Beispiel ist `exampleRow` der Name des Attributs in Ihrer Eingabedatei, welches als Zeilen-ID festgelegt ist.

Beim Generieren von Stapelvoraussagen erhalten Sie möglicherweise folgendes Ergebnis:

```
tag,bestAnswer,score  
55,0,0.46317  
102,1,0.89625
```

In diesem Beispiel stellt `RowID` das Attribut `customerID` dar. Beispiel: `customerID` 55 wird auf unsere E-Mail-Kampagne mit geringer Wahrscheinlichkeit (0,46317) reagieren, während `customerID` 102 auf unsere E-Mail-Kampagne mit hoher Wahrscheinlichkeit (0,89625) reagieren wird.

Verwenden des Felds `AttributeType`

In Amazon ML gibt es vier Datentypen für Attribute:

Binary

Wählen Sie `BINARY` für ein Attribut, das nur zwei möglichen Status haben kann, z. B. `yes` oder `no`.

So kann beispielsweise das Attribut `isNew`, das nachverfolgt, ob eine Person ein neuer Kunde ist, den Wert `true` aufweisen, um anzuzeigen, dass der Kunde ein Neukunde ist, und den Wert `false`, um anzuzeigen, dass er oder sie kein neuer Kunde ist.

Gültige negative Werte sind `0`, `n`, `no`, `f` und `false`.

Gültige positive Werte sind `1`, `y`, `yes`, `t` und `true`.

Amazon ML ignoriert die Groß-/Kleinschreibung von Binäreingaben und entfernt die umgebenden Leerzeichen. " FaLSe " ist beispielsweise ein gültiger Binärwert. Sie können die binären Werte, die Sie in derselben Datenquelle verwenden, mischen, Sie können beispielsweise `true`, `no`, und `1` aus. Nur Amazon ML-Ausgänge `0` und `1` für binäre Attribute.

Kategorisch

Wählen Sie `CATEGORICAL` für ein Attribut, das eine begrenzte Anzahl eindeutiger Zeichenfolgenwerte aufnimmt. Beispielsweise eine Benutzer-ID, der Monat und eine ZIP-Code sind kategorische Werte. kategorische Attribute werden als einzelne Zeichenfolge behandelt und nicht weiter als Token verwendet.

Numerischer Wert

Wählen Sie `NUMERIC` für ein Attribut, das eine Menge als Wert verwendet.

Beispiel: Temperatur, Gewicht und Klickrate sind numerische Werte.

Nicht alle Attribute mit Zahlen sind numerisch. Kategorische Attribute, z. B. Tage des Monats und IDs, werden häufig als Zahlen dargestellt. Um als numerische Zahl zu gelten, muss eine Zahl mit einer anderen Zahl vergleichbar sein. Die Kunden-ID 664727 enthält beispielsweise keine Informationen zur Kunden-ID 124552, aber eine Gewichtung von 10 informiert Sie darüber, dass das Attribut schwerer gewichtet ist als ein Attribut mit einer Gewichtung von 5. Tage des Monats sind nicht numerischen, da der Erste eines Monats vor oder nach dem Zweiten eines anderen Monat auftreten kann.

Note

Wenn Sie Ihr Schema mit Amazon ML erstellen, wird die `Numeric` Datentyp für alle Attribute, die diese Zahlen verwenden. Wenn Amazon ML Ihr Schema erstellt, prüfen Sie auf falsche Zuweisungen, und setzen Sie diese Attribute auf `CATEGORICAL` aus.

Text

Wählen Sie `TEXT` für ein Attribut, das eine Zeichenfolge von Wörtern ist. Wenn Sie Textattribute lesen, konvertiert Amazon ML sie in Token, die mit Leerzeichen voneinander getrennt werden.

Beispielsweise wird `email subject` zu `email` und `subject` und `email-subject` `here` wird zu `email-subject` und `here`.

Wenn der Datentyp für eine Variable im Schulungsschema nicht dem Datentyp für die Variable im Auswertungsschema entspricht, ändert Amazon ML den Auswertungsdatentyp so, dass er dem Schulungsdatentyp entspricht. Wenn das Trainingsdatenschema beispielsweise einen Datentyp von `TEXT` zur Variablen `zurage`, aber das Auswertungsschema weist einen Datentyp von `NUMERIC` zu `age`, dann behandelt Amazon ML das Alter in den Auswertungsdaten als `TEXT` Variablen anstelle von `NUMERIC` aus.

Weitere Informationen zu Statistiken zu den einzelnen Datentypen finden Sie unter [Beschreibende Statistiken](#).

Bereitstellen eines Schemas für Amazon ML

Jede Datenquelle braucht ein Schema. Sie haben die Wahl zwischen zwei Möglichkeiten zum Bereitstellen eines Schemas für Amazon ML:

- Erlauben Sie Amazon ML, die Datentypen der einzelnen Attribute in der Eingabedaten-Datei zu verwenden, und erstellen Sie automatisch ein Schema für Sie.
- Stellen Sie eine Schemadatei bereit, sobald Sie Ihre Amazon Simple Storage Service (Amazon S3) -Daten hochladen.

Amazon ML die Erstellung Ihres Schemas erlauben

Wenn Sie die Amazon ML-Konsole zum Erstellen einer Datenquelle verwenden, verwendet Amazon ML basierend auf den Werten Ihrer Variablen einfache Regeln, um Ihr Schema zu erstellen. Es wird ausdrücklich empfohlen, dass Sie das von Amazon ML erstellte Schema überprüfen und die Datentypen gegebenenfalls korrigieren.

Bereitstellen eines Schemas

Nachdem Sie Ihre Schemadatei erstellt haben, müssen Sie sie Amazon ML zur Verfügung stellen. Sie haben hierfür zwei Möglichkeiten:

1. Stellen Sie das Schema über die Amazon ML-Konsole bereit.

Verwenden Sie die Konsole zum Erstellen Ihrer Datenquelle, und berücksichtigen Sie dabei die Schemadatei, indem Sie die Erweiterung `.schema` an den Namen Ihrer Eingabedatendatei

anhängen. Ist beispielsweise die Amazon Simple Storage Service -URI (Amazon S3) zu Ihren Eingabedaten `s3://my-bucket-name/data/input.csv`, ist die URI zu Ihrem Schema entsprechend `s3://my-bucket-name/data/input.csv.schema`. Amazon ML findet die von Ihnen bereitgestellte Schemadatei automatisch und versucht nicht, das Schema aus Ihren Daten zu übernehmen.

Um eine Verzeichnisdatei als Dateneingabe für Amazon ML zu verwenden, hängen Sie die Erweiterung `.schema` an Ihren Verzeichnispfad an. Wenn sich Ihre Datendateien beispielsweise unter `s3://examplebucket/path/to/data/` befinden, lautet die URI zu Ihrem Schema entsprechend `s3://examplebucket/path/to/data/.schema`.

2. Stellen Sie das Schema über die Amazon ML -API bereit.

Wenn Sie die Amazon ML-API aufrufen möchten, um Ihre Datenquelle zu erstellen, können Sie die Schemadatei in Amazon S3 hochladen und anschließend die URI zu der Datei in der `DataSchemaLocationS3Attribut` der `CreateDataSourceFromS3API`. Weitere Informationen finden Sie unter [CreateDataSourceFromS3](#).

Sie können das Schema direkt in den Nutzdaten von `CreateDataSource*API` anstatt es zuerst auf Amazon S3 zu speichern. Hierzu fügen Sie die vollständige Schemazeichenfolge in das Attribut `DataSchema` von `CreateDataSourceFromS3-`, `CreateDataSourceFromRDS-` oder `CreateDataSourceFromRedshift-APIs` ein. Weitere Informationen finden Sie unter [Amazon Machine Learning API Reference](#).

Aufteilen Ihrer Daten

Das wesentliche Ziel eines ML-Modells ist es, genaue Voraussagen über zukünftige Daten-Instances über die Schulungsmodelle hinaus zu erreichen. Bevor Sie ein ML-Modell verwenden, um Voraussagen zu erstellen, müssen Sie die Leistung der Voraussagen des Modells bewerten. Zur Einschätzung der Qualität eines ML-Modells für Voraussagen mit Daten, die es noch nicht gesehen hat, können wir einen Teil der Daten, für die wir bereits die Antwort kennen, als Vertreter für zukünftige Daten reservieren oder aufteilen und auswerten, wie gut das ML-Modell die richtigen Antworten für diese Daten antizipiert. Teilen Sie die Datenquelle in einen Teil als Schulungsdatenquelle und einen Teil als Auswertungsdatenquelle auf.

Amazon ML bietet drei Optionen für das Aufteilen der Daten:

- Teilen Sie die Daten vor- Sie können die Daten vor dem Hochladen in Amazon Simple Storage Service (Amazon S3) in zwei Eingabeverzeichnisse aufteilen und zwei separate Datenquellen damit erstellen.

- Amazon ML sequenzielle Aufteilung- Sie können Amazon ML bei der Erstellung der Schulungs- und Auswertungsdatenquellen sequenziell aufteilen.
- Amazon ML zufällige Aufteilung- Sie können Amazon ML bei der Erstellung der Schulungs- und Auswertungsdatenquellen mit einer Zufallsmethode mit Startwert (Seed) aufteilen.

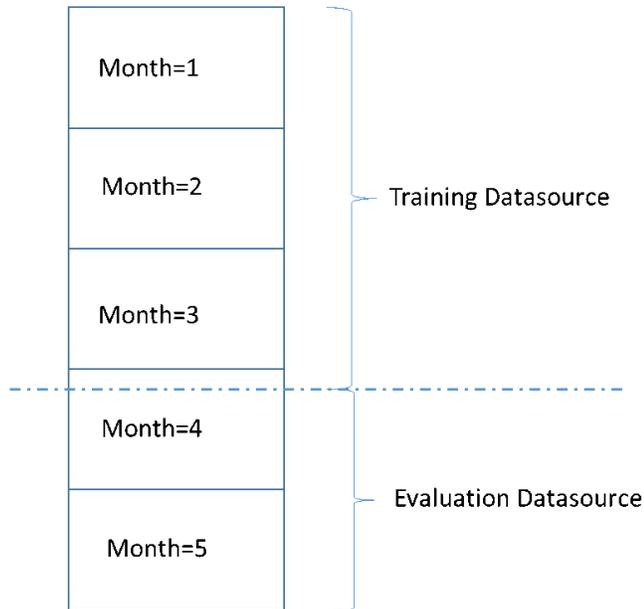
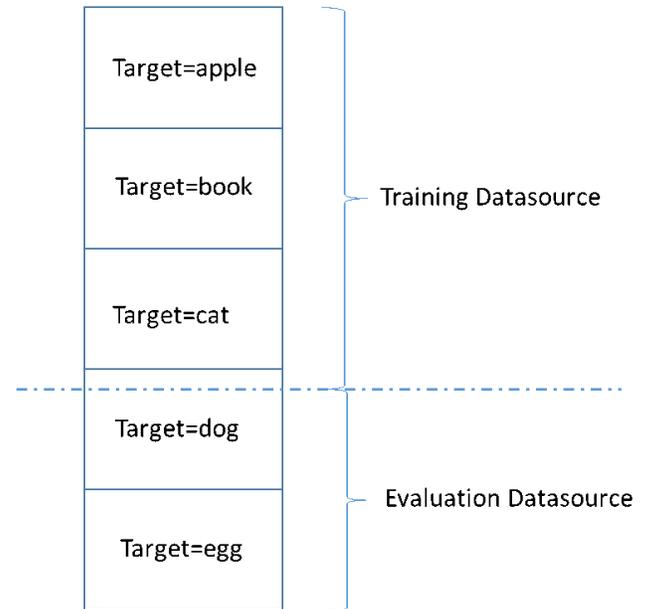
Vorabtrennung Ihrer Daten

Wenn Sie explizite Kontrolle über die Daten in den Schulungs- und Auswertungsdatenquellen wünschen, teilen Sie die Daten in separate Datenverzeichnisse auf und erstellen Sie separate Datenquellen für die Eingabe- und Auswertungsverzeichnisse.

Sequenzielle Aufteilung Ihrer Daten

Eine einfache Möglichkeit, Ihre Eingabedaten für Schulung und Auswertung aufzuteilen, ist die Auswahl nicht überlappender Teilmengen Ihrer Daten, wobei die Reihenfolge der Datensätze beibehalten wird. Dieser Ansatz ist nützlich, wenn Sie Ihre ML-Modelle mit Daten von einem bestimmten Datum oder innerhalb eines bestimmten Zeitraums auswerten möchten. Angenommen, Sie haben die Kundenbindungsdaten der letzten fünf Monate, und Sie möchten diese historischen Daten nutzen, um die Kundenbindung für den nächsten Monat vorauszusagen. Mit der Nutzung der Daten aus dem Anfangsbereich für die Schulung und der Daten aus dem Endbereich für die Auswertung erhalten Sie wahrscheinlich eine genauere Einschätzung der Modellqualität als bei Verwendung von Datensätzen aus dem gesamten Datumsbereich.

In der folgenden Abbildung finden Sie Beispiele dafür, wann Sie eine sequenzielle Aufteilungsstrategie statt einer Zufallsmethode verwenden sollten.

Case 1: Sequential split is the **correct** strategyCase 2: Sequential split is the **wrong** strategy

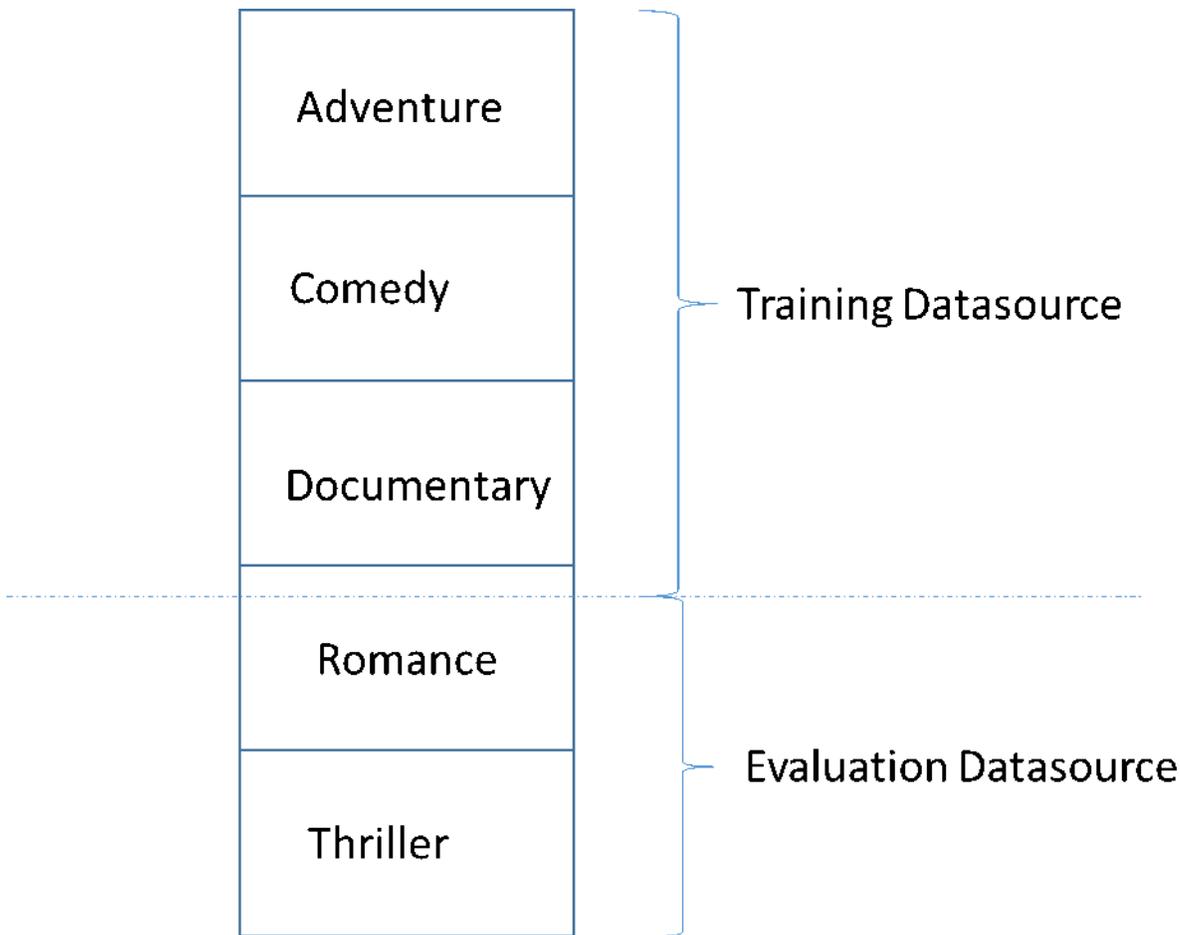
Wenn Sie eine Datenquelle erstellen, können Sie Ihre Datenquelle sequenziell aufteilen, und Amazon ML verwendet die ersten 70 Prozent Ihrer Daten für die Schulung und die verbleibenden 30 Prozent der Daten zur Auswertung. Dies ist die Standardmethode bei der Verwendung der Amazon ML-Konsole zur Aufteilung Ihrer Daten.

Zufällige Aufteilung Ihrer Daten

Zufälliges Aufteilen der Eingabedaten in Schulungs- und Auswertungsdatenquellen stellt sicher, dass die Verteilung von Daten in der Schulungs- und Auswertungsdatenquelle ähnlich ist. Wählen Sie diese Option, wenn Sie die Reihenfolge der Eingabedaten nicht beibehalten müssen.

Amazon ML verwendet eine Pseudo-Zufallszahl-Methode mit Startwert für die Aufteilung Ihrer Daten. Der Seed-Startwert basiert teilweise auf einer Eingabezeichenfolge und teilweise auf dem Inhalt der Daten. Die Amazon ML-Konsole verwendet standardmäßig das S3-Verzeichnis der Eingabedaten als Zeichenfolge. API-Benutzer können eine benutzerdefinierte Zeichenfolge festlegen. Das bedeutet, dass Amazon ML bei gleichem S3-Bucket und gleichen Daten die Daten jedes Mal auf die gleiche Weise aufteilt. Um zu ändern, wie Amazon ML die Daten aufteilt, können Sie `createDataSourceFromS3`, `createDataSourceFromRedshift`, oder `createDataSourceFromRDSAPI` und geben Sie einen Wert für die Seed-Zeichenfolge an. Wenn Sie diese APIs verwenden, um separate Datenquellen für Schulung und Auswertung zu erstellen, ist es wichtig, denselben Seed-Zeichenfolgewert für beide Datenquellen und die

Ergänzungsmarkierung für eine Datenquelle anzugeben, um sicherzustellen, dass es keine Überschneidung zwischen den Schulungs- und Auswertungsdaten gibt.



Ein häufiges Problem bei der Entwicklung eines hochwertigen ML-Modells ist die Auswertung der ML-Modells mit Daten, die nicht den Daten für die Schulung entsprechen. Angenommen, Sie verwenden ML, um ein Filmgenre vorauszusagen, und Ihre Schulungsdaten enthalten Filme aus den Segmenten Abenteuer, Komödie und Dokumentation. Ihre Auswertungsdaten enthalten jedoch nur Daten aus dem Genre Liebesfilm und Thriller. In diesem Fall konnte das ML-Modell keine Informationen über die Filmgenre Liebesfilm und Thriller lernen, und die Auswertung konnte nicht feststellen, wie gut das Modell die Muster für Abenteuer, Komödie und Dokumentation gelernt hat. Demzufolge sind die Genre-Informationen nutzlos, und die Qualität der ML-Modellvoraussagen für alle Genres ist beeinträchtigt. Das Modell und die Auswertung sind zu unterschiedlich (extrem unterschiedliche beschreibende Statistiken), als dass sie nützlich wären. Dies kann der Fall sein, wenn die Eingabedaten nach einer der Spalten im Datensatz sortiert werden und dann sequenziell aufgeteilt werden.

Wenn Ihre Schulungs- und Auswertungsdatenquellen unterschiedliche Datenverteilungen haben, sehen Sie eine Auswertungswarnung in der Modellauswertung. Weitere Informationen zu Auswertungswarnungen finden Sie unter [Auswertungswarnungen](#).

Sie müssen die zufällige Aufteilung in Amazon ML nicht verwenden, wenn Ihre Eingabedaten bereits randomisiert sind, z. B. durch zufällige Umstellung Ihrer Eingabedaten in Amazon S3 oder mithilfe der Amazon Redshift SQL-Abfragefunktion oder mithilfe der Amazon Redshift SQL-Abfragefunktion `random()` Funktion oder eine MySQL SQL-Abfragefunktion `rand()` Funktion beim Erstellen der Datenquellen. In diesen Fällen können Sie die sequenzielle Aufteilungsoption zum Erstellen von Schulungs- und Auswertungsdatenquellen mit ähnlichen Verteilungen verwenden.

Dateneinblicke

Amazon ML berechnet beschreibende Statistiken für Ihre Eingabedaten, mit denen Sie Ihre Daten besser verstehen.

Beschreibende Statistiken

Amazon ML berechnet die folgenden beschreibenden Statistiken für verschiedene Attributtypen:

Numerischer Wert:

- Verteilungshistogramme
- Anzahl ungültiger Werte
- Minimale, mittlere, durchschnittliche und maximale Werte

Binär und kategorisch:

- Anzahl (eindeutige Werte pro Kategorie)
- Wertverteilungshistogramm
- Häufigste Werte
- Anzahl eindeutiger Werte
- Prozentsatz des tatsächlichen Werts (nur binär)
- Bedeutendste Wörter
- Häufigste Wörter

Text:

- Name des Attributs
- Korrelation zum Ziel (wenn ein Ziel festgelegt ist)
- Wörter gesamt
- Eindeutige Wörter
- Umfang der Anzahl der Wörter in einer Zeile
- Umfang der Wortlängen
- Bedeutendste Wörter

Zugriff auf Dateneinblicke in der Amazon ML-Konsole

In der Amazon ML-Konsole können Sie den Namen oder die ID einer Datenquelle auswählen, um die zugehörige Dateneinblicke angezeigt. Auf dieser Seite finden Sie Metriken und Visualisierungen, mit deren Hilfe Sie mehr über die Eingabedaten in Verbindung mit der Datenquelle erfahren, einschließlich der folgenden Informationen:

- Datenzusammenfassung
- Zielverteilungen
- Fehlende Werte
- Ungültige Werte
- Zusammenfassende Statistik der Variablen nach Datentyp
- Verteilung der Variablen nach Datentyp

In den folgenden Abschnitten werden die Metriken und Visualisierungen im Detail beschrieben.

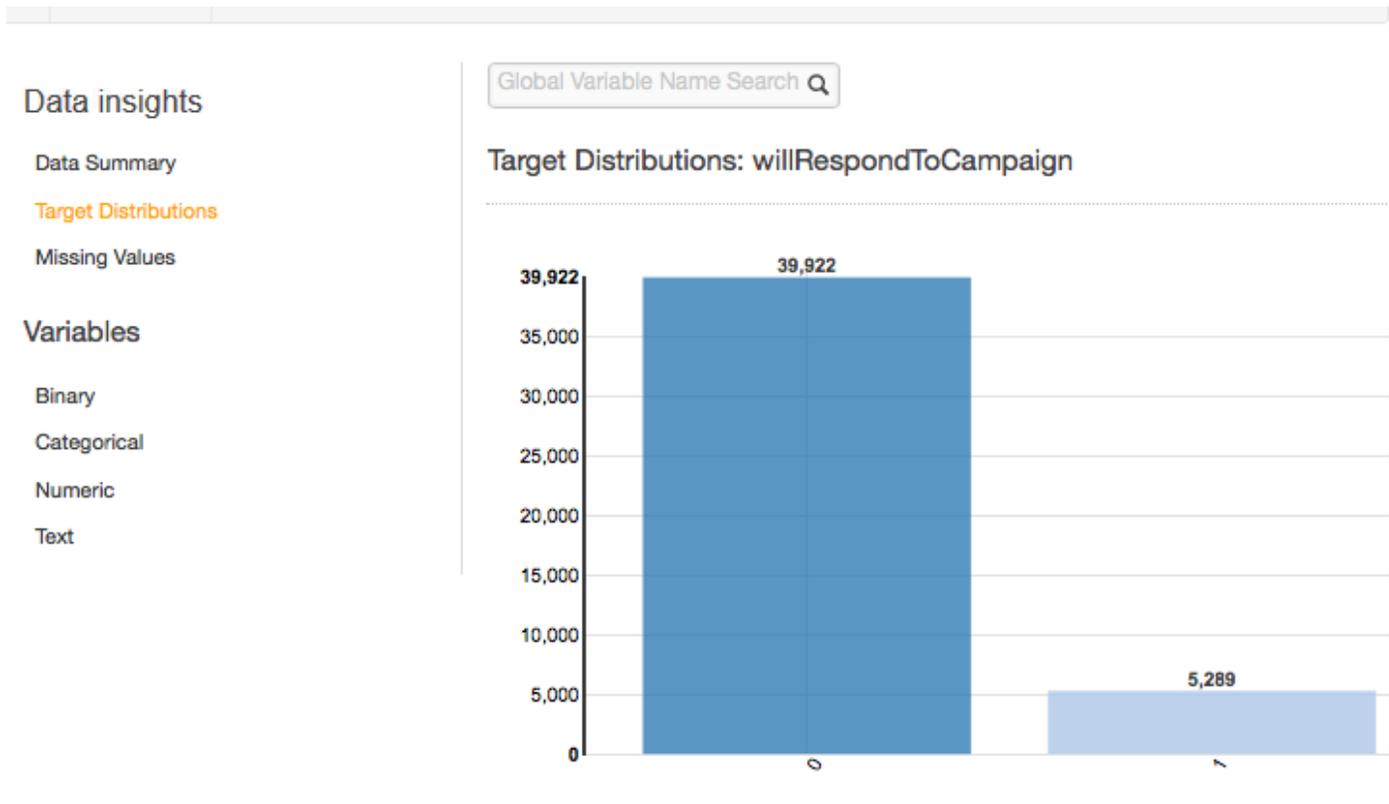
Datenzusammenfassung

Der zusammenfassende Datenbericht einer Datenquelle enthält zusammenfassende Informationen, einschließlich der Datenquellen-ID, Name, wo sie abgeschlossen wurde, aktueller Status, Zielattribut, Eingabedateninformationen (S3-Bucket-Speicherort, Datenformat, Anzahl der verarbeiteten Datensätze und Anzahl der fehlerhaften Datensätze während der Verarbeitung) sowie die Anzahl der Variablen nach Datentyp.

Zielverteilungen

Der Zielverteilungsbericht zeigt die Verteilung von Zielattributen der Datenquelle. Im folgenden Beispiel gibt es 39.922 Beobachtungen, in denen das Zielattribut "willRespondToCampaign" gleich 0 ist. Dies ist die Anzahl der Kunden, die nicht auf die E-Mail-Kampagne reagiert haben. Es gibt 5.289

Beobachtungen, in denen "willRespondToCampaign" 1 entspricht. Dies ist die Anzahl der Kunden, die auf die E-Mail-Kampagne reagiert haben.



Fehlende Werte

Der Bericht über fehlende Werte listet die Attribute in den Eingabedaten auf, für die Werte fehlen. Nur Attribute mit numerischen Datentypen können fehlende Werte haben. Da fehlende Werte sich auf die Qualität der Schulung eines ML-Modells auswirken können, empfehlen wir, dass fehlende Werte angegeben werden, falls möglich.

Wenn während der ML-Modellschulung das Zielattribut fehlt, lehnt Amazon ML den entsprechenden Datensatz ab. Wenn das Zielattribut im Datensatz vorhanden ist, aber der Wert für ein anderes numerisches Attribut fehlt, übergeht Amazon ML den fehlenden Wert. In diesem Fall erstellt Amazon ML ein Ersatzattribut und setzt es auf 1, um anzugeben, dass dieses Attribut fehlt. Auf diese Weise lernt Amazon ML Muster aus dem Auftreten von fehlenden Werten.

Ungültige Werte

Ungültige Werte können nur bei numerischen und binären Datentypen auftreten. Sie können ungültige Werte suchen, indem Sie die zusammenfassenden Statistiken von Variablen in den

Datentypberichten anzeigen. In den folgenden Beispielen gibt es nur einen ungültigen Wert im numerischen Attribut "Duration" und zwei ungültige Werte im binären Datentyp (einer im Attribut "Housing" und einer im Attribut "Loan").

Numeric Variables

Variables ^	Correlations to Target ⇅	Missing Values ⇅	Invalid Values ⇅	Range ⇅	Mean ⇅	Median ⇅	Preview
duration	0.05165	2 (0%)	1 (0%)	0 - 4918	258.1618	180	

Binary Variables

Variables ^	Correlations to Target ⇅	Percent True ⇅	Invalid Values ⇅	Preview
campaign	NA	100%	27667 (61%)	
housing	0.01842	56%	1 (0%)	
loan	0.00656	16%	1 (0%)	
willRespondToCampaign	NA	12%	0 (0%)	

Korrelation zwischen Variable und Ziel

Nach dem Erstellen einer Datenquelle kann Amazon ML die Datenquelle auswerten und die Korrelation oder die Auswirkungen zwischen Variablen und Ziel ermitteln. Beispielsweise ist es möglich, dass der Preis eines Produkts signifikante Auswirkungen darauf hat, ob es ein Bestseller wird, während die Abmessungen des Produkts wahrscheinlich wenig Vorhersagekraft haben.

Eine bewährte Methode ist es, so viele Variablen wie möglich in die Schulungsdaten einzuschließen. Das Datenrauschen durch viele Variablen mit wenig Vorhersagekraft kann jedoch die Qualität und die Richtigkeit Ihres ML-Modells negativ beeinflussen.

Sie können die prädiktive Leistung Ihres Modells verbessern, indem Sie Variablen mit wenig Auswirkungen während der Modellschulung entfernen. Sie können in einem definieren, welche Variablen für den maschinellen Lernprozess zur Verfügung gestellt werden. Rezept, was ein Transformationsmechanismus von Amazon ML ist. Weitere Informationen zu Rezepten finden Sie unter [Data Transformation for Machine Learning](#).

Zusammenfassende Statistik der Attribute nach Datentyp

Im Bericht zu Dateneinblicken können Sie zusammenfassende Attributstatistiken nach den folgenden Datentypen anzeigen:

- Binary
- Kategorisch
- Numerischer Wert
- Text

Zusammenfassende Statistiken für binären Datentypen zeigen alle binären Attribute. Die Spalte **Correlations to target** zeigt die zwischen der Zielspalte und der Attributspalte gemeinsam genutzten Informationen. Die Spalte **Percent true** zeigt den Prozentsatz der Beobachtungen mit dem Wert "1" an. In der Spalte **Invalid values** wird die Anzahl der ungültigen Werte und der Prozentsatz der ungültigen Werte für jedes Attribut angezeigt. In der Spalte **Preview** finden Sie einen Link zu einer grafischen Verteilung für jedes Attribut.

Binary Variables

Variables	Correlations to Target	Percent True	Invalid Values	Preview
campaign	NA	100%	27667 (61%)	
housing	0.01842	56%	1 (0%)	
loan	0.00656	16%	1 (0%)	
willRespondToCampaign	NA	12%	0 (0%)	

Zusammenfassende Statistiken für den kategorischen Datentyp zeigen aller kategorischen Attribute mit der Anzahl eindeutiger Werte, häufigster Wert und seltenster Wert. In der Spalte **Preview** finden Sie einen Link zu einer grafischen Verteilung für jedes Attribut.

Categorical Variables

Variables	Correlations to Target	Unique Values	Most Frequent	Least Frequent	Preview
campaign	0.00433	49	1	39	
customerid	NA	45211	45211	1	
education	0.00355	5	secondary		
housing	0.01846	4	1		
jobid	0.00671	13	blue-collar		
willRespondToCampaign	NA	3	0		

Zusammenfassende Statistiken für den numerischen Datentyp zeigen alle numerischen Attributen mit der Anzahl fehlender Werte, ungültiger Werte, Wertebereich, durchschnittlicher und mittlerer Wert. In der Spalte Preview finden Sie einen Link zu einer grafischen Verteilung für jedes Attribut.

Numeric Variables

Variables	Correlations to Target	Missing Values	Invalid Values	Range	Mean	Median	Preview
duration	0.05165	2 (0%)	1 (0%)	0 - 4918	258.1618	180	

Zusammenfassende Statistiken für den Datentyp Text zeigen alle Textattribute, die Gesamtanzahl der Wörter in diesem Attribut, die Anzahl der eindeutigen Wörter in diesem Attribut, den Umfang von Wörtern in einem Attribut, den Umfang der Wortlängen und die bedeutendsten Wörter. In der Spalte Preview finden Sie einen Link zu einer grafischen Verteilung für jedes Attribut.

Text attributes

Attributes	Correlations to target *	Total words	Unique words	Words in attribute (range)	Word length (range)	Most prominent words
Phrase	0.07118	751741	12811	0 - 48	1 - 18	enters, trust ...

« < 1 - 1 of 1 Attributes > »

* Correlations to Target is an approximate statistic for text attributes.

Das folgende Beispiel zeigt Statistiken für den Datentyp Text für eine Textvariable mit dem Namen "Review", mit vier Datensätzen.

1. The fox jumped over the fence.
2. This movie is intriguing.
- 3.
4. Fascinating movie.

Die Spalten für dieses Beispiel würden die folgenden Informationen anzeigen.

- Die Spalte **Attributes** zeigt den Namen der Variablen an. In diesem Beispiel heißt die Spalte "Review".
- Die Spalte **Correlations to target** existiert nur, wenn ein Ziel angegeben ist. Korrelationen messen die Menge an Informationen, die dieses Attribut über das Ziel bietet. Je höher die Korrelation, desto mehr erfahren Sie aus diesem Attribut über das Ziel. Korrelation wird in Form von gegenseitigen Informationen zwischen einer vereinfachten Darstellung des Textattributs und des Ziels gemessen.
- Die Spalte **Total words** zeigt die Anzahl der Wörter, die durch Aufgliederung jeden Datensatzes in Token generiert werden, wobei Wörter durch Leerzeichen getrennt werden. In diesem Beispiel enthält die Spalte "12".
- Die Spalte **Unique words** zeigt die Anzahl der eindeutigen Wörter für ein Attribut an. In diesem Beispiel enthält die Spalte "10".
- Die Spalte **Words in attribute (range)** zeigt die Anzahl der Wörter in einer einzigen Zeile im Attribut an. In diesem Beispiel enthält die Spalte "0-6".
- Die Spalte **Word length (range)** zeigt den Umfang an, wie viele Zeichen in den Wörtern enthalten sind. In diesem Beispiel enthält die Spalte "2-11".
- Die Spalte **Most prominent words** zeigt eine Rangliste der Wörter, die im Attribut vorhanden sind. Wenn es ein Zielattribut gibt, werden Wörter anhand ihrer Korrelation zum Ziel sortiert. Das bedeutet, dass die Wörter mit der höchsten Korrelation zuerst aufgelistet werden. Wenn kein Ziel in den Daten vorhanden ist, werden die Wörter anhand ihres mittlerer Informationsgehalts in Rangfolge gebracht.

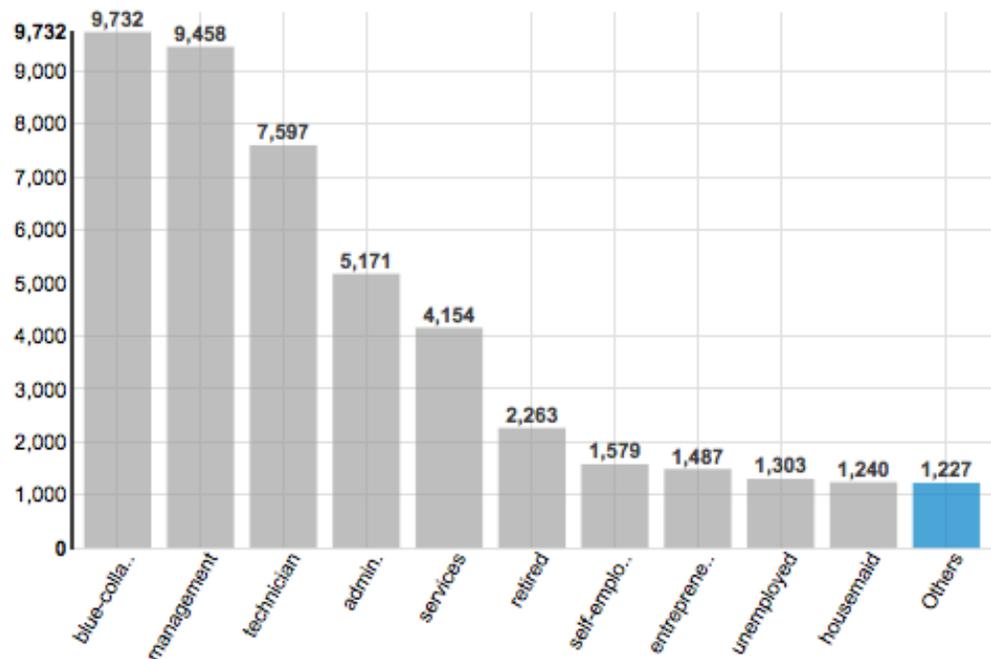
Erläuterungen zur Verteilung von kategorischen und binären Attributen

Durch Klicken auf den Link **Preview** für das kategorische oder binäre Attribut können Sie die Verteilung des Attributs sowie die Beispieldaten aus der Eingabedatei für jeden kategorischen Wert des Attributs anzeigen.

Die folgende Abbildung zeigt die Verteilung für das kategorische Attribut `jobId`. Die Verteilung zeigt die 10 häufigsten kategorischen Werte, alle anderen Werte sind als "other" gruppiert. Sie stuft jeden der 10 häufigsten kategorischen Werte nach der Anzahl der Beobachtungen in der Eingabedatei ein, die diesen Wert enthält, sowie einem Link zur Anzeige von Beispielbeobachtungen aus der Eingabedatendatei.

Categorical Variables: `jobId`

Top 10 `jobId`



All Categories

Ranking	Category	Count	
1	blue-collar	9732	Sample data
2	management	9458	Sample data
3	technician	7597	Sample data

Erläuterungen zur Verteilung von numerischen Attributen

Um die Verteilung eines numerischen Attributs anzuzeigen, klicken Sie auf den Link `Preview` für das Attribut. Wenn Sie die Verteilung eines numerischen Attributs erstellen, können Sie Bin-Größen von 500, 200, 100, 50 oder 20 festlegen. Die größer die Bin-Größe, um so kleiner die Anzahl der Balkendiagramme, die angezeigt werden. Darüber hinaus wird die Auflösung der Verteilung für

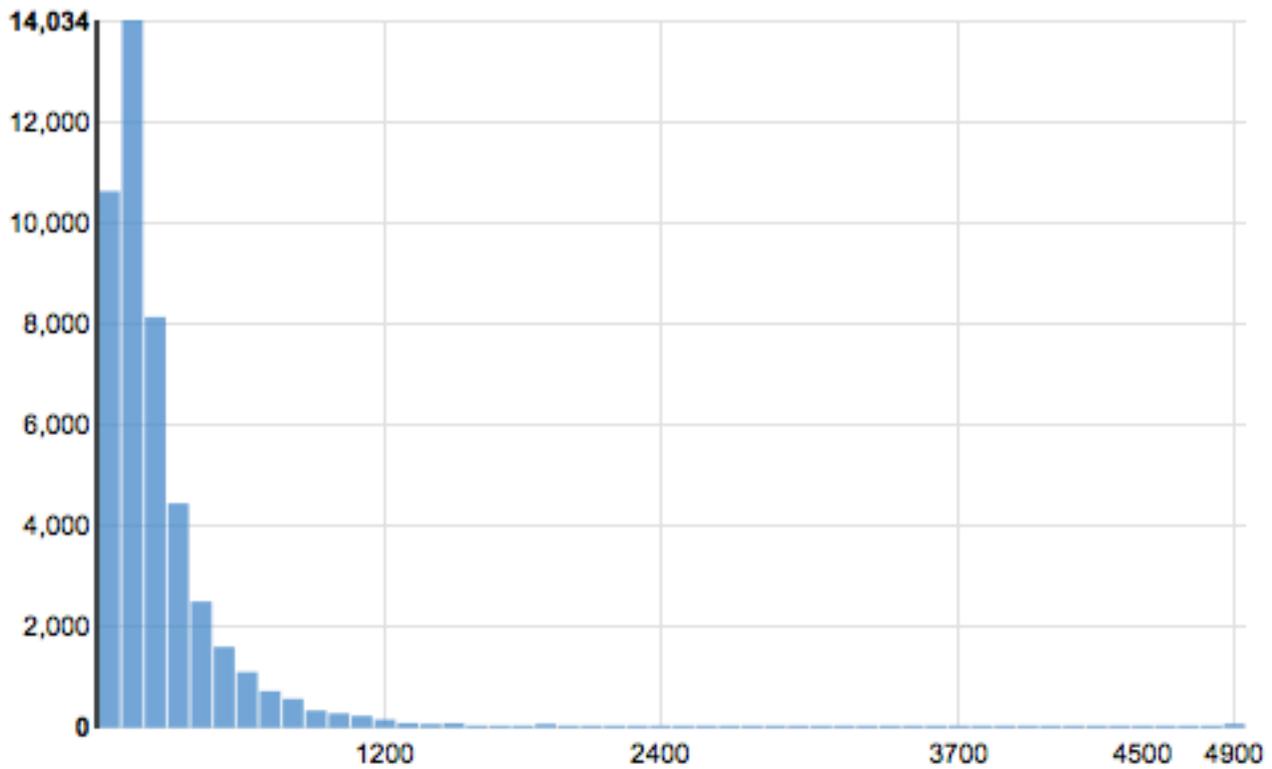
große Bin-Größen recht grob. Im Gegensatz dazu erhöht das Festlegen der Bucket-Größe auf 20 die Auflösung der angezeigten Verteilung.

Die minimalen, mittleren und maximalen Werte werden ebenfalls angezeigt (siehe Abbildung).

Numeric Variables: duration

Select Bin Width:

500 200 100 50 20



Min: 0 Mean: 258.1618 Max: 4918

Erläuterungen zur Verteilung von Textattributen

Um die Verteilung eines Textattributs anzuzeigen, klicken Sie auf den Link Preview für das Attribut. Bei der Anzeige der Verteilung eines Textattributs sehen Sie die folgenden Informationen.

Text attributes: Phrase

Ranking	Token	Word prominence	Count	
1	enters	0.01105	7	0.0%
2	trust	0.00884	28	0.0%
3	bad	0.00735	833	0.2%
4	film	0.00669	4747	1.3%
5	movie	0.00611	4242	1.2%
6	unwieldy	0.00605	11	0.0%
7	good	0.00574	1620	0.5%
8	ashamed	0.00551	7	0.0%
9	funny	0.00550	1078	0.3%
10	wankery	0.00498	9	0.0%

« < 1 - 10 of 11091 > »

Ranking

Texttoken werden anhand der Menge der übermittelten Informationen angeordnet, von am meisten informativen bis zu am wenigsten informativ.

Token

Das Token zeigt das Wort aus dem Eingabetext, auf das sich die Statistikzeile bezieht.

Wortbedeutung

Wenn es ein Zielattribut gibt, werden Wörter anhand ihrer Korrelation zum Ziel sortiert. Das bedeutet, dass die Wörter mit der höchsten Korrelation zuerst aufgelistet werden. Wenn kein Ziel

in den Daten vorhanden ist, werden die Wörter anhand ihrer Entropie in Rangfolge gebracht, d. h. die Menge an Informationen, die sie kommunizieren können.

Anzahl

Die Anzahl zeigt die Anzahl der Eingabedatensätze, in denen das Token vorhanden ist.

Prozentsatz

Der Prozentsatz zeigt den prozentualen Anteil der Eingabedatenzeilen, in denen das Token vorhanden ist.

Verwenden von Amazon S3 mit Amazon ML

Amazon Simple Storage Service (Amazon S3) ist ein Speicher für das Internet. Mit Amazon S3 können Sie jederzeit beliebige Mengen von Daten von überall aus im Internet speichern und aufrufen. Amazon ML verwendet Amazon S3 als primäres Daten-Repository für die folgenden Aufgaben:

- Für den Zugriff auf Ihre Eingabedateien zum Erstellen von Datenquellenobjekten für die Schulung und die Auswertung Ihrer ML-Modelle.
- Für den Zugriff auf Ihre Eingabedateien zum Generieren von Stapelvoraussagen.
- Wenn Sie Stapelvoraussagen mithilfe Ihrer ML-Modelle generieren zum Ausgeben der Voraussagedatei an einen S3-Bucket, den Sie angeben.
- So kopieren Sie Daten, die Sie in Amazon Redshift oder Amazon Relational Database Service (Amazon RDS) gespeichert haben, in eine CSV-Datei und zum Hochladen in Amazon S3.

Um Amazon ML die Ausführung dieser Aufgaben zu ermöglichen, müssen Sie Amazon ML Berechtigungen für den Zugriff auf Ihre Amazon S3 S3-Daten gewähren.

Note

Sie können keine Stapelvoraussagedateien in einen S3-Bucket ausgeben, der nur serverseitige verschlüsselte Dateien akzeptiert. Stellen Sie sicher, dass Ihre Bucket-Richtlinie das Hochladen unverschlüsselter Dateien zulässt, in dem Sie bestätigen, dass die Richtlinie keinen Deny-Effekt für die `s3:PutObject`-Aktion umfasst, wenn kein `s3:x-amz-server-side-encryption`-Header in der Anforderung vorhanden ist. Weitere Informationen zu serverseitigen S3-Verschlüsselungs-Bucket-Richtlinien finden Sie unter [Schützen von Daten](#)

[mithilfe serverseitiger Verschlüsselung im Benutzerhandbuch für Amazon Simple Storage Service](#) aus.

Hochladen Ihrer Daten in Amazon S3

Sie müssen Ihre Eingabedaten in den Amazon Simple Storage Service (Amazon S3) hochladen, da Amazon ML Daten von Amazon S3-Standorten liest. Sie können Ihre Daten direkt in Amazon S3 hochladen (z. B. von Ihrem Computer), oder Amazon ML kann Daten, die Sie in Amazon Redshift oder Amazon Relational Database Service (RDS) gespeichert haben, in eine CSV-Datei kopieren und in Amazon S3 hochladen.

Weitere Informationen über das Kopieren Ihrer Daten von Amazon Redshift oder Amazon RDS finden Sie unter [Using Amazon Redshift with Amazon ML](#) bzw. [Using Amazon RDS with Amazon ML](#).

Im Rest dieses Abschnitts wird beschrieben, wie Ihre Eingabedaten direkt von Ihrem Computer in Amazon S3 hochgeladen werden. Bevor Sie die Verfahren in diesem Abschnitt beginnen, müssen sich Ihre Daten in einer CSV-Datei befinden. Weitere Informationen zum korrekten Formatieren der CSV-Datei derart, dass Amazon ML diese verwenden kann, finden Sie unter [Erläuterungen zum Datenformat für Amazon ML](#) aus.

So laden Sie Ihre Daten von Ihrem Computer in Amazon S3 hoch

1. Melden Sie sich bei der AWS-Managementkonsole an und öffnen Sie die Amazon S3-Konsole unter <https://console.aws.amazon.com/s3>.
2. Erstellen Sie einen Bucket, oder wählen Sie einen vorhandenen Bucket aus.
 - a. Wählen Sie die Option Create Bucket aus, um einen Bucket zu erstellen. Benennen Sie Ihren Bucket, wählen Sie eine Region aus (Sie können eine beliebige verfügbare Region auswählen), und wählen Sie dann Create aus. Weitere Informationen dazu erhalten Sie unter [Create a Bucket](#) im Amazon-Handbuch Erste Schritte.
 - b. Um einen vorhandenen Bucket zu verwenden, suchen Sie nach dem Bucket, indem Sie den Bucket in der Liste All Buckets (Alle Buckets) auswählen. Wenn der Bucket-Name angezeigt wird, wählen Sie ihn aus, und klicken Sie dann auf Upload.
3. Klicken Sie im Dialogfeld Upload auf Add Files.
4. Navigieren Sie zu dem Ordner, der die CSV-Eingabedatei enthält, und wählen Sie dann Öffnen aus.

Berechtigungen

Um Berechtigungen für Amazon ML für den Zugriff auf einen Ihrer S3-Buckets zu gewähren, müssen Sie die Bucket-Richtlinie bearbeiten.

Weitere Informationen zum Erteilen von Amazon ML-Berechtigungen zum Lesen von Daten aus Ihrem Bucket in Amazon S3 finden Sie unter [Gewähren von Amazon ML-Berechtigungen zum Lesen von Daten aus Amazon S3](#) aus.

Weitere Informationen zum Erteilen von Berechtigungen für Amazon ML zur Ausgabe der Stapelvoraussageergebnisse in Ihren Bucket in Amazon S3 finden Sie unter [Gewähren von Berechtigungen für Amazon ML zwecks Ausgabe von Voraussagen in Amazon S3](#) aus.

Weitere Informationen zum Verwalten von Zugriffsberechtigungen auf Amazon S3 S3-Ressourcen finden Sie unter [Amazon S3 Entwicklerhandbuch](#) aus.

Erstellen einer Amazon ML-Datenquelle aus Daten in Amazon Redshift

Wenn Sie in Amazon Redshift Daten gespeichert haben, können Sie die Datenquelle erstellen-Assistent in der Amazon Machine Learning (Amazon ML) -Konsole zum Erstellen eines Datenquellenobjekts. Wenn Sie eine Datenquelle aus Amazon Redshift Redshift-Daten erstellen, geben Sie den Cluster an, der Ihre Daten und die SQL-Abfrage zum Abrufen Ihrer Daten enthält. Amazon ML führt die Abfrage durch Aufrufen der Amazon Redshift ausUnLoad-Befehl auf dem Cluster. Amazon ML speichert die Ergebnisse am Amazon Simple Storage Service (Amazon S3) - Standort Ihrer Wahl und verwendet dann die in Amazon S3 gespeicherten Daten, um die Datenquelle zu erstellen. Die Datenquelle, der Amazon Redshift -Cluster und der S3-Bucket müssen sich alle in derselben Region befinden.

Note

Amazon ML unterstützt das Erstellen von Datenquellen von Amazon Redshift -Clustern in privaten VPCs nicht. Der Cluster muss über eine öffentliche IP-Adresse verfügen.

Themen

- [Erforderliche Parameter für den Assistenten Datenquelle erstellen](#)
- [Erstellen einer Datenquelle mit Amazon Redshift Data \(Konsole\)](#)

- [Fehlersuche bei Amazon Redshift](#)

Erforderliche Parameter für den Assistenten Datenquelle erstellen

Damit Amazon ML eine Verbindung zu Ihrer Amazon Redshift Redshift-Datenbank aufbauen und Daten in Ihrem Auftrag lesen kann, müssen Sie Folgendes angeben:

- Der Amazon RedshiftClusterIdentifier
- Der Name der Amazon Redshift Redshift-Datenbank
- Die Anmeldeinformationen für die Amazon Redshift Redshift-Datenbank (Benutzername und Passwort) für die Datenbank
- Der Amazon ML Amazon RedshiftAWS Identity and Access Management(IAM) -Rolle
- Die Amazon Redshift SQL-Abfrage
- (Optional) Den Speicherort des Amazon-ML-Schemas
- Den Amazon S3 S3-Staging-Standort (wobei Amazon ML die Daten vor dem Erstellen der Datenquelle ablegt)

Darüber hinaus müssen Sie sicherstellen, dass die IAM-Benutzer oder -Rollen, die Amazon Redshift -Datenquellen erstellen (über die Konsole oder mithilfe derCreateDataSourceFromRedshiftaction) habe dasiam:PassRoleDie Erlaubnis.

Amazon RedshiftClusterIdentifier

Verwenden Sie bei diesem Parameter, bei dem die Groß- und Kleinschreibung beachtet werden muss, um Ihren Cluster zu finden und eine Die Cluster-ID (Name) können Sie über Ihre Amazon Redshift Redshift-Konsole abrufen. Weitere Informationen zu Clustern finden Sie unter[Amazon-Redshift-Cluster](#)aus.

Amazon-Redshift-Datenbankname

Verwenden Sie diesen Parameter, um Amazon ML mitzuteilen, welche Datenbank im Amazon Redshift -Cluster die Daten enthält, die Sie als Datenquelle verwenden möchten.

Anmeldeinformationen für Amazon Redshift Redshift-Datenbank

Verwenden Sie diesen Parameter, um den Benutzernamen und das Passwort des Amazon Redshift Redshift-Datenbankbenutzers anzugeben, in dessen Kontext die Sicherheitsabfrage ausgeführt wird.

Note

Amazon ML benötigt einen Amazon Redshift Redshift-Benutzernamen und ein -Passwort, um eine Verbindung zu Ihrer Amazon Redshift Redshift-Datenbank aufzubauen. Nach dem Entladen der Daten auf Amazon S3 verwendet Amazon ML Ihr Passwort nie wieder, und das Passwort wird auch nicht gespeichert.

Amazon ML Amazon Redshift Redshift-Rolle

Verwenden Sie diesen Parameter, um den Namen der IAM-Rolle anzugeben, den Amazon ML zum Konfigurieren der Sicherheitsgruppen für den Amazon Redshift Redshift-Cluster und der Bucket-Richtlinie für den Amazon S3 S3-Staging-Speicherort verwenden soll.

Wenn Sie nicht über eine IAM-Rolle verfügen, die auf Amazon Redshift zugreifen kann, kann Amazon ML eine Rolle für Sie erstellen. Wenn Amazon ML eine Rolle erstellt, wird eine Kundenverwaltete Richtlinie erstellt und an eine IAM-Rolle angefügt. Die Richtlinie, die Amazon ML erstellt, erteilt Amazon ML die Berechtigung für den Zugriff auf die von Ihnen angegebenen Cluster.

Wenn Sie bereits eine IAM-Rolle für den Zugriff auf Amazon Redshift besitzen, können Sie den ARN der Rolle eingeben oder die Rolle aus der Dropdown-Liste wählen. IAM-Rollen mit Amazon Redshift Redshift-Zugriff finden Sie oben in der Dropdown-Liste.

Die IAM-Rolle muss den folgenden Inhalt haben:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "machinelearning.amazonaws.com"
      },
      "Action": "sts:AssumeRole",
      "Condition": {
        "StringEquals": { "aws:SourceAccount": "123456789012" },
        "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:datasource/*" }
      }
    }
  ]
}
```

```
}]
}
```

Weitere Informationen zu kundenverwalteten Richtlinien finden Sie unter [Vom Kunden verwaltete Richtlinien](#) im IAM User Guide aus.

Amazon-Redshift-SQL-Abfrage

Verwenden Sie diesen Parameter, um die SQL SELECT-Abfrage anzugeben, die Amazon ML auf Ihrer Amazon Redshift Redshift-Datenbank ausführt, um Ihre Daten auszuwählen. Amazon ML verwendet Amazon Redshift [ENTLADEN](#)-Aktion, um die Ergebnisse Ihrer Abfrage sicher in einen Amazon S3 S3-Speicherort zu kopieren.

Note

Amazon ML eignet sich hervorragend dann, wenn Eingabedatensätze in zufälliger Reihenfolge (gemischt) bereitgestellt werden. Sie können problemlos die Ergebnisse Ihrer Amazon Redshift -SQL-Abfrage mischen, indem Sie Amazon Redshift verwenden.random ()Funktion. Beispiel: Angenommen, dies ist die ursprüngliche Abfrage:

```
"SELECT col1, col2, ... FROM training_table"
```

Sie können durch Aktualisierung der Abfrage zufällig mischen:

```
"SELECT col1, col2, ... FROM training_table ORDER BY random()"
```

Schemaspeicherort (Optional)

Verwenden Sie diesen Parameter, um den Amazon S3-Pfad zu Ihrem Schema für die Amazon Redshift Redshift-Daten, die Amazon ML exportieren wird, anzugeben.

Wenn Sie kein Schema für Ihre Datenquelle angeben, erstellt die Amazon ML-Konsole automatisch ein Amazon ML-Schema, das auf dem Datenschema der Amazon Redshift -SQL-Abfrage basiert. Amazon ML-Schemata verfügen über weniger Datentypen als Amazon Redshift Redshift-Schemata, sodass es sich nicht um eine Eins-zu-Eins-Konvertierung handelt. Die Amazon ML-Konsole konvertiert Amazon Redshift -Datentypen nach dem folgenden Konvertierungsschema in Amazon ML-Datentypen.

Amazon Redshift-Datentypen	Amazon Redshift Redshift-Aliase	Amazon ML-Datentyp
SMALLINT	INT2	NUMERIC
INTEGER	INT, INT4	NUMERIC
BIGINT	INT8	NUMERIC
DECIMAL	NUMERIC	NUMERIC
REAL	FLOAT4	NUMERIC
DOUBLE PRECISION	FLOAT8, FLOAT	NUMERIC
BOOLEAN	BOOL	BINARY
CHAR	CHARACTER, NCHAR, BPCHAR	CATEGORICAL
VARCHAR	CHARACTER VARYING, NVARCHAR, TEXT	TEXT
DATUM		TEXT
TIMESTAMP	TIMESTAMP WITHOUT TIME ZONE	TEXT

Wird in Amazon ML konvertiert Binary-Datentypen, müssen die Werte der Amazon Redshift Booleschen in Ihren Daten unterstützte Amazon ML Binärwerte sein. Wenn Ihr boolescher Datentyp nicht unterstützte Werte besitzt, konvertiert Amazon ML diese in einen möglichst spezifischen Datentyp. Wenn beispielsweise ein Amazon Redshift Boolean die Werte hat 0, 1, und 2 konvertiert Amazon ML den booleschen Wert in eine Numeric-Datentyp. Weitere Informationen zu unterstützten binären Werten finden Sie unter [Verwenden des Felds AttributeType](#).

Wenn Amazon ML keinen Datentyp herausfinden kann, lautet der Standardwert Text aus.

Nachdem Amazon ML das Schema konvertiert hat, können Sie die zugewiesenen Amazon ML-Datentypen im Assistenten Datenquelle erstellen überprüfen und korrigieren und das Schema überarbeiten, bevor Amazon ML die Datenquelle erstellt.

Speicherort für Amazon S3 Staging

Verwenden Sie diesen Parameter, um den Namen des Amazon S3 S3-Staging-Speicherorts anzugeben, an dem Amazon ML die Ergebnisse der Amazon Redshift -SQL-Abfrage speichert. Nach dem Erstellen der Datenquelle verwendet Amazon ML die Daten im Staging-Speicherort anstatt zu Amazon Redshift zurückzukehren.

Note

Da Amazon ML die von der Amazon ML Amazon Redshift definierte IAM-Rolle voraussetzt, besitzt Amazon ML die Berechtigungen, auf alle am Amazon S3 S3-Staging-Speicherort angegebenen Objekte zuzugreifen. Aus diesem Grund empfehlen wir, dass Sie nur Dateien im Amazon S3 S3-Staging-Speicherort speichern, die keine sensiblen Informationen enthalten. Zum Beispiel, wenn Ihr Root-Buckets3://mybucket/empfehlen wir, dass Sie einen Speicherort erstellen, in dem Sie nur die Dateien speichern, auf die Amazon ML Zugriff erhalten soll, z. B. s3://mybucket/AmazonMLInput/aus.

Erstellen einer Datenquelle mit Amazon Redshift Data (Konsole)

Die Amazon ML -Konsole bietet zwei Möglichkeiten zum Erstellen einer Datenquelle Amazon Redshift -Daten. Sie können eine Datenquelle erstellen, indem Sie den Assistenten Datenquelle erstellen abschließen, oder, falls Sie bereits über eine mit Amazon Redshift Redshift-Daten erstellte Datenquelle verfügen, die ursprüngliche Datenquelle kopieren und die Einstellungen bearbeiten. Das Kopieren einer Datenquelle ermöglicht Ihnen die einfache Erstellung mehrerer ähnlicher Datenquellen.

Weitere Informationen zum Erstellen einer Datenquelle mit der API finden Sie unter [CreateDataSourceFromRedshift](#).

Weitere Information zu den Parametern der folgenden Verfahren finden Sie unter [Erforderliche Parameter für den Assistenten Datenquelle erstellen](#).

Themen

- [Erstellen einer Datenquelle \(Konsole\)](#)
- [Kopieren einer Datenquelle \(Konsole\)](#)

Erstellen einer Datenquelle (Konsole)

Um Daten von Amazon Redshift in eine Amazon ML-Datenquelle zu entladen, verwenden Sie den Assistenten Datenquelle erstellen.

So erstellen Sie eine Datenquelle aus Daten in Amazon Redshift

1. Öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/> aus.
2. Im Amazon ML-Dashboard unter Entitäten, wählen Neues erstellen... Klicken Sie auf und danach auf Eine Datenquelle vorhanden sein aus.
3. Auf der Eingabedatenangezeigten, wählen Sie Amazon Redshift aus.
4. Geben Sie im Assistenten „Datenquelle erstellen“ unter Cluster-ID den Namen des Clusters ein.
5. Für Datenbankname geben Sie den Namen der Amazon Redshift Redshift-Datenbank ein.
6. Geben Sie im Feld Datenbank-Benutzername den Benutzernamen für die Datenbank ein.
7. Geben Sie im Feld Datenbankpasswort das Passwort für die Datenbank ein.
8. Wählen Sie unter IAM-Rolle Ihre IAM-Rolle. Wenn Sie noch kein AWS-Konto haben, wählen Sie Create a new role (Neue Rolle erstellen) aus. Amazon ML erstellt eine IAM Amazon Redshift Redshift-Rolle für Sie.
9. Um Ihre Amazon Redshift Redshift-Einstellungen zu testen, wählen Sie Zugriff testen (neben IAM-Rolle) enthalten. Wenn Amazon ML mit den angegebenen Einstellungen keine Verbindung zu Amazon Redshift aufbauen kann, können Sie nicht mit der Erstellung der Datenquelle fortfahren. Hilfe zur Problembeseitigung finden Sie unter [Fehlersuche](#).
10. Geben Sie unter SQL-Abfrage die SQL-Abfrage ein.
11. Für Speicherort Schema Wählen Sie, ob Amazon ML ein Schema für Sie erstellen soll. Wenn Sie selbst kein Schema erstellt haben, geben Sie den Amazon S3-Pfad zu Ihrer Schemadatei ein.
12. Für Speicherort für Amazon S3 S3-Kommissionierung geben Sie den Amazon S3-Pfad zu dem Bucket ein, in den Amazon ML die von Amazon Redshift entladene Daten ablegen soll.
13. (Optional) Geben Sie unter Datenquellennamen einen Namen für die Datenquelle ein.
14. Wählen Sie Überprüfen. Amazon ML überprüft die Verbindung zu Ihrer Amazon Redshift Redshift-Datenbank.

15. Überprüfen Sie auf der Seite Schema die Datentypen für alle Attribute und korrigieren Sie sie falls erforderlich.
16. Klicken Sie auf Continue.
17. Wenn Sie diese Datenquelle verwenden möchten, um ein ML-Modell zu erstellen oder auszuwerten, wählen Sie für Beabsichtigen Sie, dieses Datenset für die Erstellung oder Auswertung eines ML-Modells zu verwenden? die Antwort Ja. Wenn Sie Ja gewählt haben, wählen Sie die Zeile für die Ziele. Weitere Informationen über Ziele finden Sie unter [Verwenden des Felds targetAttributeName](#).

Wenn Sie diese Datenquelle mit einem Modell verwenden möchten, das Sie bereits erstellt haben, um Voraussagen zu erstellen, wählen Sie Nein.

18. Klicken Sie auf Continue.
19. Wählen Sie für Enthalten Ihre Daten eine ID? die Antwort Nein, wenn Ihre Daten keine Zeilen-ID enthalten.

Wenn Ihre Daten eine Zeilen-ID enthalten, wählen Sie Ja. Weitere Information zu Zeilen-IDs finden Sie unter [Verwenden des Felds rowID](#).

20. Wählen Sie Review.
21. Prüfen Sie auf der Seite Prüfen Ihre Einstellungen und wählen Sie Fertig.

Nachdem Sie eine Datenquelle erstellt haben, können Sie diese folgendermaßen verwenden: [create an ML model](#). Wenn Sie bereits ein Modell erstellt haben, können Sie mit der Datenquelle [evaluate an ML model](#) oder [generate predictions](#).

Kopieren einer Datenquelle (Konsole)

Wenn Sie eine Datenquelle erstellen möchten, die einer vorhandenen Datenquelle ähnelt, können Sie die Amazon ML-Konsole verwenden, um die ursprüngliche Datenquelle zu kopieren und die Einstellungen zu bearbeiten. Sie können beispielsweise mit einer vorhandenen Datenquelle beginnen und das Datenschema dann an Ihre Daten anpassen; ändern Sie die für das Entladen der Daten von Amazon Redshift verwendete SQL-Abfrage oder geben Sie eine andere AWS Identity and Access Management (IAM) -Benutzer, der auf den Amazon Redshift Redshift-Cluster zugreifen kann.

So kopieren und ändern Sie eine Amazon Redshift Redshift-Datenquelle

1. Öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/aus>.

2. Im Amazon ML-Dashboard unter Entitäten, wählen Neues erstellen... Klicken Sie auf und danach auf Eine Datenquelle vorhanden sein aus.
3. Auf der Eingabedaten-Seite, für Wo sind Ihre Daten?, wählen Amazon Redshift aus. Wenn Sie bereits eine Datenquelle aus Amazon Redshift Redshift-Daten erstellt haben, haben Sie die Möglichkeit, Einstellungen von anderen Datenquellen zu kopieren.

Where is your data?



S3



Amazon Redshift

Do you want to copy the settings from another Amazon Redshift datasource to create a new datasource? To copy settings, choose [Find a datasource](#).

Wenn Sie noch keine Datenquelle aus Amazon Redshift Redshift-Daten erstellt haben, wird diese Option nicht angezeigt.

4. Wählen Sie Eine Datenquelle suchen.
5. Wählen Sie die Datenquelle, die Sie kopieren möchten, und wählen Sie Einstellungen kopieren aus. Amazon ML übernimmt automatisch die meisten Datenquelleneinstellungen der ursprünglichen Datenquelle. Das Passwort für die Datenbank, der Speicherort des Schemas und der Name der Datenquelle werden nicht von der ursprünglichen Datenquelle kopiert.
6. Ändern Sie die automatisch vorgenommenen Einstellungen nach Bedarf. Wenn Sie beispielsweise die Daten ändern möchten, die Amazon ML von Amazon Redshift entlädt, ändern Sie die SQL-Abfrage.
7. Geben Sie im Feld Datenbankpasswort das Passwort für die Datenbank ein. Amazon ML speichert keine Passwörter und verwendet diese auch nicht erneut, sodass Sie immer ein Passwort angeben müssen.
8. (Optional) Für Speicherort Schemawählt Amazon ML vorausich möchte, dass Amazon ML ein empfohlenes Schema generiertFür dich. Wenn Sie bereits ein Schema erstellt haben, wählen Sieich möchte das Schema verwenden, dass ich in Amazon S3 erstellt und gespeichert habeund geben Sie den Pfad zu Ihrer Schemadatei in Amazon S3 ein.
9. (Optional) Geben Sie unter Datenquellennamen einen Namen für die Datenquelle ein. Andernfalls generiert Amazon ML einen neuen Datenquellennamen für Sie.
10. Wählen Sie Überprüfen. Amazon ML überprüft die Verbindung zu Ihrer Amazon Redshift Redshift-Datenbank.

11. (Optional) Wenn Amazon ML das Schema für Sie abgeleitet hat, finden Sie im Schemaklicker Sie auf, überprüfen Sie die Datentypen für alle Attribute und korrigieren Sie sie falls erforderlich.
12. Klicken Sie auf Continue.
13. Wenn Sie diese Datenquelle verwenden möchten, um ein ML-Modell zu erstellen oder auszuwerten, wählen Sie für Beabsichtigen Sie, dieses Datenset für die Erstellung oder Auswertung eines ML-Modells zu verwenden? die Antwort Ja. Wenn Sie Ja gewählt haben, wählen Sie die Zeile für die Ziele. Weitere Informationen über Ziele finden Sie unter [Verwenden des Felds targetAttributeName](#).

Wenn Sie diese Datenquelle mit einem Modell verwenden möchten, das Sie bereits erstellt haben, um Voraussagen zu erstellen, wählen Sie Nein.

14. Klicken Sie auf Continue.
15. Wählen Sie für Enthalten Ihre Daten eine ID? die Antwort Nein, wenn Ihre Daten keine Zeilen-ID enthalten.

Wenn Ihre Daten eine Zeilen-ID enthalten, wählen Sie Ja und wählen Sie die Zeile aus, die Sie als ID verwenden möchten. Weitere Information zu Zeilen-IDs finden Sie unter [Verwenden des Felds rowID](#).

16. Wählen Sie Review.
17. Überprüfen Sie die Einstellungen und klicken Sie anschließend auf Fertig.

Nachdem Sie eine Datenquelle erstellt haben, können Sie diese folgendermaßen verwenden: [create an ML model](#). Wenn Sie bereits ein Modell erstellt haben, können Sie mit der Datenquelle [evaluate an ML model](#) oder [generate predictions](#).

Fehlersuche bei Amazon Redshift

Während Sie Ihre Amazon Redshift Redshift-Datenquelle, ML-Modelle und Evaluierung erstellen, meldet Amazon Machine Learning (Amazon ML) den Status Ihrer Amazon ML-Objekte in der Amazon ML-Konsole. Wenn Amazon ML Fehlermeldungen zurückgibt, verwenden Sie die folgenden Informationen und Ressourcen, um die Probleme zu beheben.

Antworten auf allgemeine Fragen zu Amazon ML finden Sie in der [Amazon Machine Learning FAQs](#). Sie können auch in der [Amazon-Forum für Machine Learning](#).

Themen

- [Fehlersuche](#)
- [Den AWS-Support kontaktieren](#)

Fehlersuche

Das Format der Rolle ist ungültig. Geben Sie eine gültige IAM-Rolle an. Zum Beispiel `arn:aws:iam::YourAccountID:Rolle/YourRedshiftRole`.

Ursache

Das Format des Amazon Resource Name (ARN) der IAM-Rolle ist nicht korrekt.

Lösung

Korrigieren Sie im Assistenten zum Erstellen von Datenquellen den ARN für Ihre Rolle. Weitere Informationen zur Formatierung von Rollen-ARNs finden Sie [IAM-ARNs](#) in der IAM User Guide. Die Region ist optional für ARNs der IAM-Rolle.

Die Rolle ist ungültig. Amazon ML kann die <role ARN>IAM-Rolle nicht übernehmen. Geben Sie eine gültige IAM-Rolle an und machen Sie sie für Amazon ML zugänglich.

Ursache

Ihre Rolle ist nicht so eingerichtet, dass Amazon ML sie übernehmen kann.

Lösung

In der [IAM-Konsole](#) bearbeiten Sie Ihre Rolle so, dass sie über eine Vertrauensrichtlinie verfügt, die es Amazon ML ermöglicht, die ihr zugeordnete Rolle zu übernehmen.

Der Benutzer <Benutzer-ARN> ist nicht berechtigt, die IAM-Rolle <Rollen-ARN> weiterzugeben.

Ursache

Ihr IAM-Benutzer verfügt nicht über eine Berechtigungsrichtlinie, die es ihm ermöglicht, eine Rolle an Amazon ML zu übergeben.

Lösung

Fügen Sie Ihrem IAM-Benutzer eine Berechtigungsrichtlinie hinzu, mit der Sie Rollen an Amazon ML übergeben können. Sie können Ihrem IAM-Benutzer in der [IAM-Konsole](#) eine Berechtigungsrichtlinie zuordnen.

Das Übergeben einer IAM-Rolle über Konten ist nicht zulässig. Die IAM-Rolle muss zu diesem Konto gehören.

Ursache

Sie können keine Rolle übergeben, die zu einem anderen IAM-Konto gehört.

Lösung

Melden Sie sich bei dem AWS-Konto, das Sie verwendet haben, um die Rolle zu erstellen. Ihre IAM-Rollen finden Sie in Ihrer [IAM-Konsole](#).

Die angegebene Rolle ist nicht berechtigt, die Operation durchzuführen. Stellen Sie eine Rolle bereit, die über eine Richtlinie verfügt, die Amazon ML die erforderlichen Berechtigungen gewährt.

Ursache

Ihre IAM-Rolle ist nicht berechtigt, die gewünschte Operation durchzuführen.

Lösung

Ändern Sie die Ihrer Rolle zugewiesene Berechtigungsrichtlinie in der [IAM-Konsole](#) dahingehend ab, dass die erforderlichen Berechtigungen zur Verfügung stehen.

Amazon ML kann in diesem Amazon Redshift Redshift-Cluster keine Sicherheitsgruppe mit der angegebenen IAM-Rolle konfigurieren.

Ursache

Ihre IAM-Rolle verfügt nicht über die erforderlichen Berechtigungen, um einen Amazon Redshift Redshift-Sicherheitscluster zu konfigurieren.

Lösung

Ändern Sie die Ihrer Rolle zugewiesene Berechtigungsrichtlinie in der [IAM-Konsole](#) dahingehend ab, dass die erforderlichen Berechtigungen zur Verfügung stehen.

Beim Versuch von Amazon ML, eine Sicherheitsgruppe in Ihrem Cluster zu konfigurieren, ist ein Fehler aufgetreten. Bitte versuchen Sie es später erneut.

Ursache

Als Amazon ML versuchte, eine Verbindung zu Ihrem Amazon-Redshift-Cluster herzustellen, ist ein Problem aufgetreten.

Lösung

Überprüfen Sie, ob die IAM-Rolle, die Sie im Assistenten Datenquelle erstellen angegeben haben, über alle erforderlichen Berechtigungen verfügt.

Das Format der Cluster-ID ist ungültig. Cluster-IDs müssen mit einem Buchstaben beginnen und dürfen nur alphanumerische Zeichen und Bindestriche enthalten. Sie dürfen nicht zwei aufeinanderfolgende Bindestriche enthalten oder mit einem Bindestrich enden.

Ursache

Ihr Amazon Redshift Redshift-Cluster-ID-Format ist falsch.

Lösung

Korrigieren Sie im Assistenten Datenquelle erstellen Ihre Cluster-ID, sodass nur alphanumerische Zeichen und Bindestriche und keine zwei aufeinander folgenden Bindestriche enthalten sind und die Cluster-ID nicht mit einem Bindestrich endet.

Es gibt keinen <Amazon Redshift cluster name>Cluster, oder der Cluster befindet sich nicht in derselben Region wie Ihr Amazon ML-Service. Geben Sie einen Cluster in derselben Region wie diese Amazon-ML an.

Ursache

Amazon ML kann Ihren Amazon Redshift Redshift-Cluster nicht finden, da er sich nicht in der Region befindet, in der Sie eine Amazon ML-Datenquelle erstellen.

Lösung

Überprüfen Sie, ob Ihr Cluster in der Amazon-Redshift-Konsole vorhanden ist. [Cluster](#), dass Sie eine Datenquelle in derselben Region erstellen, in der sich Ihr Amazon Redshift Redshift-Cluster befindet, und dass die im Assistenten „Datenquelle erstellen“ angegebene Cluster-ID korrekt ist.

Amazon ML kann die Daten in Ihrem Amazon-Redshift-Cluster nicht lesen. Geben Sie die richtige Amazon Redshift Redshift-Cluster-ID an.

Ursache

Amazon ML kann die von Ihnen angegebenen Daten im Amazon Redshift Redshift-Cluster nicht lesen.

Lösung

Geben Sie im Assistenten zum Erstellen einer Datenquelle die richtige Amazon Redshift-Cluster-ID an und stellen Sie sicher, dass Sie eine Datenquelle in derselben Region erstellen, in der sich Ihr Amazon Redshift Redshift-Cluster befindet, und ob Ihr Cluster in Amazon Redshift aufgeführt ist. [Cluster](#)Seite.

Der <Amazon Redshift cluster name>Cluster ist nicht öffentlich zugänglich.

Ursache

Amazon ML kann nicht auf Ihren Cluster zugreifen, da der Cluster nicht öffentlich zugänglich ist und keine öffentliche IP-Adresse besitzt.

Lösung

Machen Sie den Cluster öffentlich zugänglich und geben Sie eine öffentliche IP-Adresse ein. Hinweise darüber, wie Cluster öffentlich zugänglich gemacht werden, finden Sie unter [Modifizieren eines Clusters](#) in der Amazon-Redshift-Verwaltungsanleitung.

Der <Redshift>Clusterstatus ist für Amazon ML nicht verfügbar. Verwenden Sie die Amazon-Redshift-Konsole, um dieses Clusterstatusproblem anzuzeigen und zu beheben. Der Cluster-Status muss "available" lauten.

Ursache

Amazon ML kann den Clusterstatus nicht sehen.

Lösung

Stellen Sie sicher, dass Ihr Cluster verfügbar ist. Weitere Informationen zum Überprüfen des Status Ihres Clusters finden Sie unter [Einen Überblick über den Clusterstatus erhalten](#) in der Amazon-Redshift-Verwaltungsanleitung. Informationen zum Neustarten des Clusters, damit er verfügbar ist, finden Sie unter [Neustart eines Clusters](#) in der Amazon-Redshift-Verwaltungsanleitung.

In diesem Cluster ist keine Datenbank <Datenbankname> vorhanden. Überprüfen Sie, ob der Datenbanknamen korrekt ist oder geben Sie einen anderen Cluster bzw. eine andere Datenbank an.

Ursache

Amazon ML kann die angegebene Datenbank im angegebenen Cluster nicht finden.

Lösung

Überprüfen Sie, ob der im Assistenten Datenbank erstellen angegebene Datenbankname korrekt ist, oder geben Sie den richtigen Cluster- und Datenbanknamen an.

Amazon ML konnte nicht auf Ihre Datenbank zugreifen. Geben Sie ein gültiges Passwort für den Datenbankbenutzer <Benutzername> ein.

Ursache

Das Passwort, das Sie im Assistenten zum Erstellen einer Datenquelle angegeben haben, damit Amazon ML auf Ihre Amazon Redshift Redshift-Datenbank zugreifen kann, ist falsch.

Lösung

Geben Sie das richtige Passwort für Ihren Amazon Redshift Redshift-Datenbankbenutzer ein.

Beim Versuch von Amazon ML, die Abfrage zu validieren, ist ein Fehler aufgetreten.

Ursache

Es besteht ein Problem mit Ihrer SQL-Abfrage.

Lösung

Überprüfen Sie, ob Ihre SQL-Abfrage gültig ist.

Beim Ausführen der SQL-Abfrage ist ein Fehler aufgetreten. Überprüfen Sie den Datenbanknamen und die angegebene Abfrage. Fehlerursache: {serverMessage}.

Ursache

Amazon Redshift konnte Ihre Abfrage nicht ausführen.

Lösung

Überprüfen Sie, ob Sie den richtigen Datenbanknamen im Assistenten Datenquelle erstellen angegeben haben und dass Ihre SQL-Abfrage gültig ist.

Beim Ausführen der SQL-Abfrage ist ein Fehler aufgetreten. Fehlerursache: {serverMessage}.

Ursache

Amazon Redshift konnte die angegebene Tabelle nicht finden.

Lösung

Stellen Sie sicher, dass die Tabelle, die Sie im Assistenten zum Erstellen einer Datenquelle angegeben haben, in Ihrer Amazon Redshift Redshift-Cluster-Datenbank vorhanden ist und dass Sie die richtige Cluster-ID, den Datenbanknamen und die richtige SQL-Abfrage eingegeben haben.

Den AWS-Support kontaktieren

Wenn Sie AWS Premium Support haben, können Sie einen technischen Support-Fall unter [AWS Support Center](#) eröffnen.

Verwenden von Daten aus einer Amazon RDS-Datenbank zum Erstellen einer Amazon ML-Datenquelle

Amazon ML ermöglicht Ihnen die Erstellung eines Datenquellenobjekts mithilfe von Daten, die in einer MySQL-Datenbank im Amazon Relational Database Service (Amazon RDS) gespeichert sind. Wenn Sie diese Aktion ausführen, erstellt Amazon ML ein AWS Data Pipeline Pipeline-Objekt, das die von Ihnen angegebene SQL-Abfrage ausführt, und platziert die Ausgabe in einem S3-Bucket Ihrer Wahl. Amazon ML verwendet diese Daten zum Erstellen der Datenquelle.

Note

Amazon ML unterstützt nur MySQL-Datenbanken in VPCs.

Damit Amazon ML Ihre Eingabedaten lesen kann, müssen Sie die Daten zuvor in Amazon Simple Storage Service (Amazon S3) exportieren. Sie können Amazon ML mithilfe der API so einrichten, dass es den Export für Sie durchführt. (RDS ist auf die API beschränkt und nicht von der Konsole aus verfügbar.)

Damit Amazon ML eine Verbindung mit Ihrer MySQL-Datenbank in Amazon RDS herstellt und Daten in Ihrem Auftrag liest, müssen Sie Folgendes angeben:

- RDS DB-Instance-Kennung
- Name der MySQL-Datenbank
- Die AWS Identity and Access Management (IAM) -Rolle, die zum Erstellen, Aktivieren und Ausführen der Daten-Pipeline verwendet wird
- Benutzeranmeldeinformationen für die Datenbank:
 - Benutzername
 - Passwort

- Sicherheitsinformationen zur AWS Data Pipeline:
 - Die IAM-Ressourcenrolle
 - Die IAM-Servicerolle
- Die Amazon RDS-Sicherheitsinformationen:
 - Die Subnetz-ID
 - Die Sicherheitsgruppen-IDs
- Die SQL-Abfrage, welche die Daten angibt, die Sie verwenden möchten, um die Datenquelle zu erstellen
- Der S3-Ausgabespeicherort (Bucket) zum Speichern der Ergebnisse der Abfrage
- (Optional) Der Speicherort der Datenschemadatei

Darüber hinaus müssen Sie sicherstellen, dass die IAM-Benutzer und -Rollen zum Erstellen der Amazon RDS-Datenquellen mithilfe des [CreateDataSourceFromRDS](#) Betrieb haben die `iam:PassRole` Erlaubnis. Weitere Informationen finden Sie unter [Steuern des Zugriffs auf Amazon ML-Ressourcen – mit IAM](#).

Themen

- [RDS-Datenbank-Instance-Kennung](#)
- [MySQL-Datenbankname](#)
- [Benutzeranmeldeinformationen für die Datenbank](#)
- [Sicherheitsinformationen zur AWS Data Pipeline](#)
- [Amazon RDS-Sicherheitsinformationen](#)
- [MySQL-SQL-Abfragen](#)
- [S3-Ausgabespeicherort](#)

RDS-Datenbank-Instance-Kennung

Die RDS-DB-Instance-Kennung ist ein eindeutiger Name, den Sie festlegen und der die Datenbank-Instance identifiziert, welche Amazon ML zur Interaktion mit Amazon RDS verwenden soll. Sie finden die RDS-DB-Instance-Kennung in der Amazon RDS-Konsole.

MySQL-Datenbankname

Der MySQL-Datenbankname gibt den Namen der MySQL-Datenbank in der RDS-DB-Instance an.

Benutzeranmeldeinformationen für die Datenbank

Zum Herstellen einer Verbindung mit der RDS-DB-Instance müssen Sie den Benutzernamen und das Kennwort des Datenbankbenutzers angeben, der über die erforderlichen Berechtigungen zum Ausführen der SQL-Abfrage verfügt, die Sie bereitgestellt haben.

Sicherheitsinformationen zur AWS Data Pipeline

Zum Aktivieren des sicheren Zugriffs auf die AWS Data Pipeline müssen Sie die Namen der IAM-Ressourcenrolle und der IAM-Servicerolle angeben.

Eine EC2-Instance nimmt die Ressourcenrolle zum Kopieren von Daten von Amazon RDS in Amazon S3 an. Die einfachste Möglichkeit zum Erstellen dieser Ressourcenrolle ist mithilfe der `DataPipelineDefaultResourceRole`-Vorlage und Auflisten von `machinelearning.aws.com` als vertrauenswürdiger Service. Weitere Informationen zur Vorlage finden Sie unter [Einrichten von IAM-Rollen](#) im AWS Data Pipeline-Entwicklerhandbuch.

Wenn Sie eine eigene Rolle erstellen, muss sie folgende Inhalte haben:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "machinelearning.amazonaws.com"
      },
      "Action": "sts:AssumeRole",
      "Condition": {
        "StringEquals": { "aws:SourceAccount": "123456789012" },
        "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-east-1:123456789012:datasource/*" }
      }
    }
  ]
}
```

AWS Data Pipeline nimmt die Servicerolle zur Überwachung des Fortschritts beim Kopieren der Daten von Amazon RDS in Amazon S3 an. Die einfachste Möglichkeit zum Erstellen dieser Ressourcenrolle ist mithilfe der `DataPipelineDefaultRole`-Vorlage und Auflisten von `machinelearning.aws.com` als vertrauenswürdiger Service. Weitere Informationen zur Vorlage finden Sie unter [Einrichten von IAM-Rollen](#) im AWS Data Pipeline-Entwicklerhandbuch.

Amazon RDS-Sicherheitsinformationen

Zum Aktivieren des sicheren Amazon RDS-Zugriffs müssen Sie die VPC Subnet ID und RDS Security Group IDs aus. Darüber hinaus müssen Sie geeignete Eingangsregeln für das VPC-Subnetz einrichten, auf das vom Parameter Subnet ID verwiesen wird, und die ID der Sicherheitsgruppe angeben, die über diese Berechtigung verfügt.

MySQL-SQL-Abfragen

Der Parameter MySQL SQL Query gibt die SQL SELECT-Abfrage an, die Sie für die MySQL-Datenbank ausführen möchten. Die Ergebnisse der Abfrage werden an den S3-Ausgabespeicherort (Bucket) kopiert, den Sie angeben.

Note

Machine Learning-Technologie eignet sich besonders dann, wenn Eingabedatensätze in zufälliger Reihenfolge (gemischt) bereitgestellt werden. Sie können problemlos die Ergebnisse Ihrer MySQL-SQL-Abfrage mischen, indem Sie die Funktion `rand()` verwenden.

Beispiel: Angenommen, dies ist die ursprüngliche Abfrage:

```
"SELECT col1, col2, ... FROM training_table"
```

Sie können durch Aktualisierung der Abfrage zufällig mischen:

```
"SELECT col1, col2, ... FROM training_table ORDER BY rand()"
```

S3-Ausgabespeicherort

Der Parameter S3 Output Location gibt den Namen des Amazon S3-Speicherorts zum Staging an, in den die Ergebnisse der MySQL-SQL-Abfrage ausgegeben werden.

Note

Sie müssen sicherstellen, dass Amazon ML Berechtigungen zum Lesen von Daten von diesem Speicherort hat, nachdem die Daten von Amazon RDS exportiert wurden. Informationen zur Einrichtung dieser Berechtigungen finden Sie unter "Gewähren von Amazon ML-Berechtigungen zum Lesen von Daten aus Amazon S3".

Schulung von ML-Modellen

Für die Schulung eines ML-Modells muss ein ML-Algorithmus (der Lernalgorithmus) mit Schulungsdaten bereitgestellt werden. Der Begriff ML-Modell bezeichnet das Modell-Artefakt, das durch den Schulungsprozess erstellt wird.

Die Schulungsdaten müssen die richtige Antwort enthalten, die als Zeil oder Zielattribut bezeichnet wird. Der Lernalgorithmus findet Muster in den Schulungsdaten, die die Attribute der Input-Daten dem Ziel (die Antwort, die Sie voraussagen möchten) zuordnen, und gibt ein ML-Modell aus, in dem diese Muster erfasst sind.

Sie können das ML-Modell verwenden, um Voraussagen für neue Daten zu erhalten, bei denen Sie das Ziel nicht kennen. Nehmen wir beispielsweise an, dass Sie ein ML-Modell schulen möchten, damit es voraussagt, ob eine E-Mail Spam ist oder nicht. Sie stellen Amazon ML Schulungsdaten bereit, die E-Mails enthalten, bei denen Sie das Ziel kennen (eine Kennzeichnung, die angibt, ob eine E-Mail Spam ist oder nicht). Amazon ML schult ein ML-Modell anhand dieser Daten. Das Ergebnis ist dann ein Modell, das versucht, vorauszusagen, ob neue E-Mails Spam sind oder nicht.

Allgemeine Informationen über ML-Modelle und ML-Algorithmen finden Sie unter [Machine Learning-Konzepte](#).

Themen

- [ML-Modelltypen](#)
- [Schulungsprozess](#)
- [Schulungsparameter](#)
- [Erstellen eines ML-Modells](#)

ML-Modelltypen

Amazon ML unterstützt drei ML-Modelltypen: binäre Klassifizierung, Mehrklassen-Klassifizierung und Regression. Wählen Sie den Modelltyp danach aus, welches Ziel Sie voraussagen möchten.

Binäres Klassifizierungsmodell

ML-Modelle für binäre Klassifizierungsprobleme prognostizieren ein binäres Ergebnis (eine von zwei möglichen Klassen). Um binäre Klassifizierungsmodelle zu schulen, verwendet Amazon ML den

branchenüblichen Lern-Algorithmus der logistischen Regression, um binäre Klassifizierungsmodelle zu schulen.

Beispiele für binäre Klassifizierungsprobleme:

- "Ist diese E-Mail Spam oder nicht?"
- "Wird der Kunde das Produkt kaufen?"
- "Ist Ihr Produkt ein Buch oder ein Nutztier?"
- "Wurde diese Bewertung von einem Kunden oder einer Maschine geschrieben?"

Mehrklassen-Klassifizierungsmodell

Mit ML-Modellen für Mehrklassen-Klassifizierungsprobleme können Sie Prognosen für mehrere Klassen generieren (Vorhersage von einem aus mehr als zwei Ergebnissen). Nutzen Sie den branchenüblichen Lern-Algorithmus der multinominalen logistischen Regression, um Mehrklassen-Modelle zu schulen.

Beispiele für Mehrklassen-Probleme:

- "Ist das Produkt ein Buch, ein Film oder Kleidung?"
- "Ist dieser Film ein Liebeskomödie, eine Dokumentation oder ein Thriller?"
- "Welche Kategorie von Produkten für diesen Kunden am interessantesten?"

Regressionsmodell

ML-Modelle für Regressionsprobleme sagen einen numerischen Wert voraus. Nutzen Sie den branchenüblichen Lern-Algorithmus der multinominalen linearen Regression, um Regressionsmodelle zu schulen.

Beispiele für Regressionsprobleme:

- "Wie wird die Temperatur in Seattle morgen sein?"
- "Wie viele Einheiten dieses Produkts werden wir verkaufen?"
- "Für welchen Preis wird dieses Haus verkauft?"

Schulungsprozess

Wenn Sie ein ML-Modell schulen, müssen Sie Folgendes angeben:

- Eingabe-Schulungsdatenquelle
- Name des Datenattributs, welche das vorherzusagende Ziel enthält
- Erforderliche Datentransformationsanweisungen
- Schulungsparameter zur Steuerung des Lern-Algorithmus

Während des Schulungsprozesses wählt Amazon ML automatisch den richtigen Lern-Algorithmus für Sie aus, basierend auf dem Typ des Ziels, das Sie in der Schulungsdatenquelle angegeben haben.

Schulungsparameter

In der Regel akzeptieren Algorithmen für Machine Learning Parameter, die verwendet werden können, um bestimmte Eigenschaften des Schulungsmodells und des resultierenden ML-Modells zu steuern. In Amazon Machine Learning heißen diese Schulungsparameter aus. Sie können diese Parameter mithilfe der Amazon ML-Konsole, einer API oder über die Befehlszeilenschnittstelle (Command Line Interface, CLI) festlegen. Wenn Sie keine Parameter festlegen, verwendet Amazon ML Standardwerte, die bekanntermaßen für eine breite Palette von Machine Learning-Aufgaben funktionieren.

Sie können Werte für die folgenden Schulungsparameter angeben:

- Maximale Modellgröße
- Maximale Anzahl von Durchläufen von Schulungsdaten
- Art der Mischung
- Regularisationstyp
- Regularisationsumfang

In der Amazon ML-Konsole werden die Schulungsparameter standardmäßig festgelegt. Die Standardeinstellungen sind für die meisten ML-Probleme geeignet, Sie können jedoch andere Werte auswählen, um die Leistung zu optimieren. Bestimmte andere Schulungsparameter, z. B. das Lerntempo, werden basierend auf Ihren Daten für Sie konfiguriert.

In den folgenden Abschnitten finden Sie weitere Informationen zu Schulungsparametern.

Maximale Modellgröße

Die maximale Modellgröße ist die Gesamtgröße in Einheiten von Byte, die Amazon ML während der Schulung eines ML-Modells erstellt.

Standardmäßig erstellt Amazon ML ein 100 MB-Modell. Sie können Amazon ML anweisen, ein kleineres oder größeres Modell zu erstellen, indem Sie eine andere Größe angeben. Den Bereich verfügbarer Größen finden Sie unter [ML-Modelltypen](#)

Wenn Amazon ML nicht genügend Muster zum Ausfüllen der Modellgröße finden kann, wird ein kleineres Modell erstellt. Wenn Sie beispielsweise eine maximale Modellgröße von 100 MB angeben, Amazon ML jedoch nur Muster mit einer Gesamtgröße von 50 MB findet, beträgt das resultierende Modell 50 MB. Wenn Amazon ML mehr Muster findet, als in die angegebene Größe passen, wird eine maximale Grenze erzwungen, indem die Muster entfernt werden, die die Qualität des geschulten Modells am wenigsten beeinträchtigen.

Durch Auswählen der Modellgröße können Sie den Kompromiss zwischen der prognostizierten Qualität und den Kosten der Verwendung steuern. Durch kleinere Modelle kann verursacht werden, dass Amazon ML viele Muster entfernt, die in die maximale Größenbegrenzung passen, wodurch die Qualität von Prognosen beeinträchtigt wird. Bei größeren Modellen ist hingegen das Abfragen von Echtzeitvoraussagen kostspieliger.

Note

Wenn Sie ein ML-Modell verwenden, um Echtzeitvoraussagen zu erstellen, fallen Gebühren für die Kapazitätsreservierung an, die auf der Größe des Modells basieren. Weitere Informationen finden Sie unter [Preise für Amazon ML](#).

Größere Eingabedatensätze führen nicht unbedingt zu größeren Modellen, da Modelle Muster speichern, und nicht Eingabedaten. Wenn es nur wenige simple Muster gibt, ist das resultierende Modell klein. Eingabedaten mit einer großen Anzahl von reinen Attributen (Eingabespalten) oder abgeleiteten Funktionen (Ausgaben der Amazon ML-Datentransformationen) weisen mit großer Wahrscheinlichkeit mehr Muster auf, die gefunden und während des Schulungsvorgangs gespeichert werden. Das Auswählen der richtigen Modellgröße für Ihre Daten und Problem erfolgt am besten mit ein paar Experimenten. Das Protokoll des Amazon ML-Modells (das Sie von der Konsole oder über die API herunterladen können) enthält Meldungen dazu, in welchem Ausmaß Kürzungen (falls vorhanden) während des Schulungsvorgangs durchgeführt werden, sodass Sie die potenzielle Prognosequalität besser einschätzen können.

Maximale Anzahl von Datendurchläufen

Um die besten Ergebnisse zu erzielen, benötigt Amazon ML möglicherweise mehrere Durchläufe für Ihre Daten, um Muster zu entdecken. Standardmäßig führt Amazon ML Durchläufe aus, Sie können

den Standardwert jedoch ändern, indem Sie eine Zahl von bis zu 100 festlegen. Amazon ML verfolgt die Qualität der Muster (Modellkonvergenz) und stoppt automatisch die Schulung automatisch, wenn keine weiteren Datenpunkte oder Muster mehr erkannt werden. Wenn Sie beispielsweise die Anzahl von Durchläufen auf 20 festlegen, Amazon ML jedoch feststellt, dass nach Beendigung von 15 Durchläufen keine neuen Muster gefunden werden können, wird die Schulung bei 15 Durchläufen gestoppt.

Im Allgemeinen sind für Datensätze mit nur wenigen Beobachtungen in der Regel mehr Datendurchläufe erforderlich, um eine höhere Modellqualität zu erzielen. Größere Datensätze enthalten häufig viele ähnliche Datenpunkte, sodass keine größere Anzahl von Durchläufen erforderlich ist. Es gibt zwei Auswirkungen der Auswahl mehrerer Datendurchläufe: die Modellschulung dauert länger und sie kostet mehr.

Art der Mischung von Schulungsdaten

In Amazon ML müssen Sie Ihre Schulungsdaten mischen. Beim Mischen wird die Reihenfolge der Daten so geändert, dass der SGD-Algorithmus nacheinander nicht nur einen Datentyp bei zahlreichen Beobachtungen erkennt. Wenn Sie beispielsweise ein ML-Modell so schulen, dass es einen Produkttyp vorhersagen kann, und Ihre Schulungsdaten enthalten die Produkttypen "Film", "Spielzeug" und "Videospiel", so sortiert der Algorithmus die Daten alphabetisch nach Produkttyp, wenn Sie die Daten vor dem Hochladen nach der Spalte für den Produkttyp sortiert haben. Der Algorithmus erkennt alle Daten für Filme zuerst, und das ML-Modell beginnt, Muster für Filme zu erlernen. Wenn das Modell dann Daten zu Spielsachen erkennt, würde jedes Update, das der Algorithmus vornimmt, das Modell an den Produkttyp "Spielzeug" anpassen, auch wenn diese Updates die Muster herabsetzen, die Filmen entsprechen. Durch diesen plötzlichen Wechsel vom Typ "Film" zu "Spielzeug" kann ein Modell erzeugen, das nicht lernt, wie Produkttypen korrekt vorhergesagt werden.

Sie müssen Ihre Trainingsdaten auch dann mischen, wenn Sie die Option für eine zufällige Aufteilung beim Aufteilen der Eingabedatenquelle in Schulungs- und Evaluierungsabschnitte auswählen. Bei der Strategie der zufälligen Aufteilung wird eine zufällige Teilmenge der Daten für jede Datenquelle ausgewählt, die Reihenfolge der Zeilen in der Datenquelle wird jedoch nicht geändert. Weitere Informationen zum Aufteilen der Daten finden Sie unter [Aufteilen Ihrer Daten](#).

Wenn Sie ein ML-Modell mithilfe der Konsole erstellen, werden die Daten in Amazon ML standardmäßig mit einer Pseudo-Zufallsmischtechnik gemischt. Unabhängig von der Anzahl angeforderter Durchläufe mischt Amazon ML die Daten nur einmal, bevor das ML-Modell geschult wird. Wenn Sie Ihre Daten gemischt haben, bevor diese Amazon ML bereitgestellt

wurden, und Sie möchten nicht, dass Amazon ML Ihre Daten erneut mischt, können Sie die Art der Mischung zu neuem Ausgangspunkt. Wenn Sie beispielsweise die Datensätze in Ihrer .csv-Datei nach dem Zufallsprinzip gemischt haben, bevor Sie sie auf Amazon S3 hochgeladen haben, verwenden Sie die `rand()`-Funktion in Ihrer MySQL SQL-Abfrage beim Erstellen Ihrer Datenquelle aus Amazon RDS oder verwendet die `random()`-Funktion in Ihrer Amazon Redshift SQL-Abfrage beim Erstellen Ihrer Datenquelle aus Amazon Redshift, Einstellung der Mischung zu neuem Ausgangspunkt wirkt sich nicht auf die Vorhersagegenauigkeit Ihres ML-Modells aus. Durch einmaliges Mischen der Daten werden die Laufzeit und die Kosten für das Erstellen eines ML-Modells reduziert.

Important

Wenn Sie ein ML-Modell mithilfe der Amazon ML-API erstellen, werden die Daten in Amazon ML in Amazon ML nicht standardmäßig gemischt. Wenn Sie die API anstelle der Konsole zum Erstellen des ML-Modells verwenden, wird dringend empfohlen, dass Sie die Daten mischen, indem Sie den Parameter `sgd.shuffleType` auf `auto` festlegen.

Regularisationstyp und -umfang

Die prädiktive Leistung komplexer ML-Modelle (die Modelle mit vielen Eingabeattributen) wird beeinträchtigt, wenn die Daten zu viele Muster enthalten. Wenn die Anzahl von Mustern steigt, so steigt auch die Wahrscheinlichkeit, dass das Modell unbeabsichtigte Datenartefakte anstelle von echten Datenmustern erlernt. In einem solchen Fall weist das Modell eine hervorragende Leistung bei Schulungsdaten auf, kann aber neue Daten nicht verallgemeinern. Dieses Phänomen wird als Überanpassung der Schulungsdaten bezeichnet.

Mithilfe der Regularisation wird verhindert, dass eine Überanpassung von Schulungsdatenbeispielen durch lineare Modelle stattfindet, indem extreme Gewichtungswerte bestraft werden. Durch die L1-Regularisation wird die Anzahl von Funktionen reduziert, die in dem Modell verwendet werden, indem das Gewicht von Funktionen, die ansonsten ein sehr geringes Gewicht hätten, auf Null gedrückt wird. Die L1-Regularisation erzeugt platzsparende Modelle und reduziert das Rauschen im Modell. Die L2-Regularisation führt zu kleineren Gesamtgewichtungswerten, wodurch die Gewichtungen stabilisiert werden, wenn eine hohe Korrelation zwischen den Funktionen besteht. Sie können den Umfang der L1- oder der L2-Regularisation steuern, indem Sie den Parameter `Regularization amount` verwenden. Das Festlegen eines besonders hohen `Regularization amount`-Werts kann dazu führen, dass alle Funktionen kein Gewicht aufweisen.

Das Auswählen und Einstellen des optimalen Regularisierungswerts ist ein aktives Thema bei der Erforschung von Machine Learning. Wahrscheinlich ist es von Vorteil, einen mittleren Umfang für die L2-Regularisation auszuwählen, was dem Standard in der Amazon ML-Konsole entspricht. Fortgeschrittene Benutzer haben die Wahl zwischen drei Arten von Regularisation (Keine, L1 oder L2) und deren Umfang. Weitere Informationen zur Regularisation erhalten Sie unter [Regularisation \(Mathematik\)](#).

Schulungsparameter: Typen und Standardwerte

In der folgenden Tabelle sind die Amazon ML-Schulungsparameter zusammen mit den Standardwerten sowie dem jeweils zulässigen Bereich aufgeführt.

Schulungsparameter	Typ	Default Value (Standardwert)	Beschreibung
maxMLMode ISizeInBytes	Ganzzahl	100.000.000 Byte (100 MiB)	Zulässiger Bereich: 100.000 (100 KiB) bis 2.147.483.648 (2 GiB) In Abhängigkeit von den Eingabedaten kann sich die Modellgröße auf die Leistung auswirken.
sgd.maxPasses	Ganzzahl	10	Zulässiger Bereich: 1-100
sgd.shuffleType	Zeichenfolge	auto	Zulässige Werte: auto oder none
sgd.l1RegularizationAmount	Double	0 (L1 wird standardmäßig nicht verwendet)	Zulässiger Bereich: 0 bis MAX_DOUBLE L1-Werte zwischen 1E-4 und 1E-8 erzeugen bekanntermaßen gute Ergebnisse. Höhere Werte erzeugen möglicherweise Modell, die nicht sehr hilfreich sind. Sie können sowohl L1 als auch L2 festlegen. Sie müssen sich für eine Option entscheiden.

Schulungsparameter	Typ	Default Value (Standardwert)	Beschreibung
sgd.l2RegularizationAmount	Double	1E-6 (L2 wird standardmäßig für diesen Regularisierungssumfang verwendet)	<p>Zulässiger Bereich: 0 bis MAX_DOUBLE</p> <p>L2-Werte zwischen 1E-2 and 1E-6 erzeugen bekanntermaßen gute Ergebnisse. Höhere Werte erzeugen möglicherweise Modell, die nicht sehr hilfreich sind.</p> <p>Sie können sowohl L1 als auch L2 festlegen. Sie müssen sich für eine Option entscheiden.</p>

Erstellen eines ML-Modells

Nachdem Sie eine Datenquelle erstellt haben, können Sie ein ML-Modell erstellen. Wenn Sie die Amazon Machine Learning Learning-Konsole zum Erstellen eines Modells verwenden, können Sie die Standardeinstellungen verwenden oder Ihr Modell durch Anwenden benutzerdefinierter Optionen anpassen.

Zu den benutzerdefinierten Optionen gehören:

- **Einstellungen für die Bewertung:** Sie können festlegen, dass Amazon ML einen Teil der Eingabedaten zum Auswerten der prädiktiven Qualität des ML-Modells reserviert. Weitere Informationen zu Auswertungen finden Sie unter [Evaluation von ML-Modellen](#).
- **Ein Rezept:** Ein Rezept weist Amazon ML an, welche Attribute und Attributtransformationen für die Modellschulung verfügbar sind. Weitere Informationen über Amazon ML-Recipes finden Sie unter [Funktionstransformation mit Datenrezepten](#) aus.
- **Schulungsparameter:** Parameter steuern bestimmte Eigenschaften des Schulungsprozesses und des resultierenden ML-Modells. Weitere Informationen zu Schulungsparametern finden Sie unter [Schulungsparameter](#).

Um Werte für diese Einstellungen auszuwählen oder anzugeben, wählen Sie die Option Benutzerdefiniert aus, wenn Sie den Assistenten zum Erstellen des ML-Modells verwenden. Wenn Sie möchten, dass Amazon ML die Standardeinstellungen anwendet, wählen Sie Standard aus.

Wenn Sie ein ML-Modell erstellen, wählt Amazon ML die Art des verwendeten Lernalgorithmus basierend auf dem Attributtyp Ihres Zielattributs aus. (Das Zielattribut ist das Attribut, das die "richtigen" Antworten enthält.) Wenn Ihr Zielattribut „Binär“ ist, erstellt Amazon ML ein binäres Klassifizierungsmodell, das den logistischen Regressionsalgorithmus verwendet. Wenn Ihr Zielattribut „Kategorisch“ ist, erstellt Amazon ML ein Mehrklassen-Modell, das einen Algorithmus der multinominalen logistischen Regression verwendet. Wenn Ihr Zielattribut „Numerisch“ ist, erstellt Amazon ML ein Regressionsmodell, das einen linearen Regressionsalgorithmus verwendet.

Themen

- [Voraussetzungen](#)
- [Erstellen eines ML-Modells mit Standardoptionen](#)
- [Erstellen eines ML-Modells mit benutzerdefinierten Optionen](#)

Voraussetzungen

Bevor Sie die Amazon ML-Konsole zum Erstellen eines ML-Modells verwenden, müssen Sie zwei Datenquellen erstellen, eine für die Schulung des Modells und eine für die Auswertung des Modells. Wenn Sie nicht zwei Datenquellen erstellt haben, lesen Sie [Schritt 2: Erstellen einer Schulungsdatenquelle](#) im Tutorial.

Erstellen eines ML-Modells mit Standardoptionen

Wählen Sie das Symbol Standard Wenn Sie möchten, dass Amazon ML Folgendes tut:

- Aufteilen der Eingabedaten so, dass die ersten 70 Prozent für die Schulung und die verbleibenden 30 Prozent für die Auswertung verwendet werden
- Vorschlagen eines Rezeptes basierend auf Statistiken, die in der Schulungsdatenbank erfasst wurden, was 70 Prozent der Eingabedatenbank entspricht.
- Auswählen der Standardschulungsparameter

So wählen Sie Standardoptionen aus

1. Wählen Sie in der Amazon ML-Konsole Amazon Machine Learning Klicken Sie auf und danach auf ML-Modelle aus.
2. Wählen Sie auf der Zusammenfassungsseite der ML-Modelle die Option zum Erstellen eines neuen ML-Modells aus.
3. Stellen Sie auf der Seite mit den Eingabedaten sicher, dass die Option ausgewählt ist, dass Sie bereits eine Datenquelle erstellt haben, die auf Ihre S3-Daten zeigt.
4. Wählen Sie in der Tabelle Ihre Datenquelle aus, und wählen Sie dann Continue aus.
5. Geben Sie auf der Seite ML-Modelleinstellungen für ML-Modellname einen Namen für Ihr ML-Modell ein.
6. Stellen Sie sicher, dass in den Schulungs- und Auswertungseinstellungen die Option Standard ausgewählt ist.
7. Für Nennen Sie diese Auswertung Geben Sie einen Namen für die Auswertung ein und wählen Sie dann Prüfen aus. Amazon ML umgeht den Rest des Assistenten und führt Sie zum Prüfen angezeigt.
8. Überprüfen Sie Ihre Daten, löschen Sie alle aus der Datenquelle kopierten Tags, die nicht auf Ihr Modell und die Auswertungen angewendet werden sollen, und wählen Sie dann Finish aus.

Erstellen eines ML-Modells mit benutzerdefinierten Optionen

Durch Anpassen Ihres ML-Modells können Sie:

- Ihr eigenes Rezept bereitstellen. Weitere Informationen dazu, wie Sie Ihre eigenes Rezept bereitstellen, finden Sie unter [Referenz zum Rezeptformat](#).
- Wählen Sie Schulungsparameter aus. Weitere Informationen zu Schulungsparametern finden Sie unter [Schulungsparameter](#).
- Wählen Sie ein anderes Aufteilungsverhältnis für Schulungen/Auswertungen als das standardmäßige Verhältnis 70/30 aus, oder geben Sie eine andere Datenquelle an, die Sie bereits für die Auswertung vorbereitet haben. Weitere Informationen über Aufteilungsstrategien finden Sie unter [Aufteilen Ihrer Daten](#).

Sie können auch die Standardwerte für diese Einstellungen auswählen.

Wenn Sie bereits ein Modell mithilfe der Standardoptionen erstellt haben und die prädiktive Leistung Ihres Modells verbessern möchten, verwenden Sie die Option, um ein neues Modell mit einigen benutzerdefinierten Einstellungen zu erstellen. Vielleicht möchten Sie weitere Funktionstransformationen zu dem Rezept hinzufügen oder die Anzahl von Durchläufen im Schulungsparameter erhöhen.

So erstellen Sie ein Modell mit benutzerdefinierten Optionen

1. Wählen Sie in der Amazon ML-Konsole **Amazon Machine Learning** Klicken Sie auf **+** und danach auf **ML-Modelle** aus.
2. Wählen Sie auf der Zusammenfassungsseite der ML-Modelle die Option zum Erstellen eines neuen ML-Modells aus.
3. Wenn Sie bereits eine Datenquelle erstellt haben, wählen Sie auf der Seite **Input Data** die Option **I already created a datasource pointing to my S3 data** aus. Wählen Sie in der Tabelle Ihre Datenquelle aus, und wählen Sie dann **Continue** aus.

Wenn Sie eine Datenquelle erstellen müssen, wählen Sie **My data is in S3, and I need to create a datasource** aus, und wählen Sie dann **Continue (Weiter)**. Sie werden zum Assistenten zum Erstellen einer Datenquelle weitergeleitet. Geben Sie an, ob sich Ihre Daten in S3 oder Redshift befinden, und klicken Sie anschließend auf **Verify**. Führen Sie die Schritte zum Erstellen einer Datenquelle aus.

Nachdem Sie eine Datenquelle erstellt haben, werden Sie automatisch zum nächsten Schritt im Assistenten zum Erstellen eines ML-Modells weitergeleitet.

4. Geben Sie auf der Seite **ML-Modelleinstellungen** für **ML-Modellname** einen Namen für Ihr ML-Modell ein.
5. Wählen Sie unter **Schulungs- und Auswertungseinstellungen** die Option **Benutzerdefiniert** aus, und klicken Sie dann auf **Continue**.
6. Auf der Seite **Rezept** können Sie [customize a recipe](#). Wenn Sie kein Rezept anpassen möchten, schlägt Amazon ML eines für Sie vor. Klicken Sie auf **Continue**.
7. Geben Sie auf der Seite **Erweiterte Einstellungen** die maximale Größe des ML-Modells, die maximale Anzahl von Datendurchläufen, die Mischungsart für Schulungsdaten, den Regularisationstyp sowie den Regularisationsumfang an. Wenn Sie dies nicht angeben, verwendet Amazon ML die Standardschulungsparameter.

Weitere Informationen zu diesen Parametern und ihren Standardwerten finden Sie unter [Schulungsparameter](#).

Klicken Sie auf Continue.

8. Geben Sie auf der Auswertung an, ob das ML-Modell sofort ausgewertet werden soll. Wenn das ML-Modell nicht jetzt ausgewertet werden soll, wählen Sie Review aus.

Wenn das ML-Modell jetzt ausgewertet werden soll:

- a. Geben Sie unter Diese Auswertung benennen einen Namen für die Auswertung ein.
 - b. FürAuswertungsdaten auswählenWählen Sie aus, ob Amazon ML einen Teil der Eingabedaten für die Auswertung reservieren soll, und, wenn dies der Fall ist, wie Sie die Datenquelle aufteilen möchten. Oder wählen Sie eine andere Datenquelle für die Auswertung aus.
 - c. Wählen Sie Review.
9. Bearbeiten Sie auf der Seite Review Ihre Auswahl, löschen Sie alle aus der Datenquelle kopierten Tags, die nicht auf Ihr Modell und die Auswertungen angewendet werden sollen, und wählen Sie dann Finish.

Lesen Sie [Schritt 4: Überprüfen der Voraussageleistung des ML-Modells und Festlegen eines Punktzellenwerts](#), nachdem Sie das Modell erstellt haben.

Datentransformationen für maschinelles Lernen

Machine Learning-Modelle sind nur so gut wie die verwendeten Daten für ihre Schulung. Ein Schlüsselmerkmal guter Schulungsdaten ist, dass sie auf eine Weise bereitgestellt werden, die für das Lernen und die Generalisierung optimiert ist. Der Vorgang, bei dem die Daten in diesem optimalen Format zusammengestellt werden, wird in der Branche als Funktionstransformation bezeichnet.

Themen

- [Bedeutung der Funktionstransformation](#)
- [Funktionstransformation mit Datenrezepten](#)
- [Referenz zum Rezeptformat](#)
- [Empfohlene Rezepte](#)
- [Referenz zur Datentransformation](#)
- [Neuordnung von Daten](#)

Bedeutung der Funktionstransformation

Betrachten wir ein Machine Learning-Modell, das entscheiden soll, ob eine Kreditkartentransaktion betrügerisch ist oder nicht. Basierend auf Ihrem Wissen über die Anwendung und Ihre Datenanalyse können Sie entscheiden, welche Datenfelder (oder Funktionen) in den Eingabedaten enthalten sein sollten. Beispielsweise sollten Transaktionsbetrag, Name des Händlers, Adresse und Adresse des Eigentümers der Kreditkarte im Lernprozess enthalten sein. Eine zufällig erstellte Transaktionsnummer hingegen enthält keine Informationen (sofern wir wissen, dass sie wirklich zufällig ist) und ist nicht nützlich.

Sobald Sie sich entschieden haben, welche Felder verwendet werden sollen, transformieren Sie diese Funktionen so, dass sie den Lernprozess unterstützen. Transformationen ergänzen die Eingabedaten mit Hintergrunderfahrung, sodass das Machine Learning-Modell aus dieser Erfahrung lernen kann. Beispielsweise steht die folgende Händleradresse in einer Zeichenfolge:

```
"123 Main Street, Seattle, WA 98101"
```

Die Adresse verfügt über begrenzte Aussagekraft – sie dient nur dem Erlernen von Mustern für diese spezielle Adresse. Durch das Herunterbrechen der Adresse in mehrere Teile jedoch können weitere Funktionen wie "Adresse" (123 Main Street), "Stadt" (Seattle), "Staat" (WA) und "ZIP" (98101) erstellt

werden. Der Lernalgorithmus kann nun mehrere separate Transaktionen zusammenführen und größere Muster erkennen – möglicherweise liegen zu bestimmten Händler-ZIPs mehr betrügerische Erfahrungen vor als zu anderen.

Weitere Informationen zur Funktionstransformation finden Sie unter [Machine Learning-Konzepte](#).

Funktionstransformation mit Datenrezepten

Es gibt zwei Möglichkeiten zum Umwandeln von Funktionen vor dem Erstellen von ML-Modellen mit Amazon ML: Sie können Ihre Eingabedaten direkt transformieren, bevor Sie sie an Amazon ML weitergeben, oder Sie können die integrierte Datentransformation von Amazon ML nutzen. Sie können Amazon ML-Rezepte verwenden, die vorformatierte Anweisungen für gängige Transformationen enthalten. Mit Rezepten können Sie Folgendes ausführen:

- Wählen Sie aus einer Liste von integrierten gängigen Machine Learning-Transformationen und wenden Sie diese auf einzelne Variablen oder Gruppen von Variablen an.
- Wählen Sie, welche der Eingabevariablen und Transformationen für den maschinellen Lernprozess zur Verfügung gestellt werden.

Die Verwendung von Amazon ML-Rezepten bietet mehrere Vorteile. Amazon ML führt die Datentransformationen für Sie aus. Es ist also nicht erforderlich, sie selbst zu implementieren. Außerdem sind sie schnell, da Amazon ML die Transformationen beim Lesen der Eingabedaten anwendet und die Ergebnisse an den Lernprozess weiterleitet, ohne dass die Ergebnisse als Zwischenschritt auf die Festplatte gespeichert werden.

Referenz zum Rezeptformat

Amazon ML-Rezepte enthalten Anweisungen zum Umwandeln Ihrer Daten im Rahmen des maschinellen Lernprozesses. Rezepte werden mit einer JSON-ähnlichen Syntax definiert, weisen aber zusätzliche Einschränkungen über die normalen JSON-Einschränkungen hinaus auf. Rezepte weisen die folgenden Abschnitte auf, die in der hier dargestellten Reihenfolge angezeigt werden müssen:

- Gruppen ermöglichen die Gruppierung mehrerer Variablen zur einfacheren Anwendung von Transformationen. Sie können beispielsweise eine Gruppe aller Variablen erstellen, die im Zusammenhang mit Freitextteilen einer Webseite (Titel, Textkörper) stehen, und dann eine Transformation für alle Teile gleichzeitig ausführen.

- Zuweisungen ermöglichen die Erstellung von benannten Zwischenvariablen, die bei der Verarbeitung wiederverwendet werden können.
- Ausgaben definieren, welche Variablen im Lernprozess verwendet werden und welche Transformationen ggf. für diese Variablen gelten.

Gruppen

Sie können Gruppen von Variablen definieren, um alle Variablen innerhalb der Gruppen gemeinsam zu transformieren oder um diese Variablen für maschinelles Lernen ohne Transformation zu verwenden. Amazon ML erstellt standardmäßig die folgenden Gruppen für Sie:

ALL_TEXT, ALL_NUMERIC, ALL_CATEGORICAL, ALL_BINARY – Typspezifische Gruppen basierend auf Variablen, die im Datenquellschema definiert sind.

Note

Mit ALL_INPUTS kann keine Gruppe erstellt werden.

Diese Variablen können im Ausgabenabschnitts Ihres Rezepts verwendet werden, ohne definiert zu werden. Sie können auch benutzerdefinierte Gruppen erstellen, indem Sie Variablen zu bzw. von vorhandenen Gruppen addieren bzw. subtrahieren, oder direkt aus einer Sammlung von Variablen. Im folgenden Beispiel zeigen werden alle drei Ansätze sowie die Syntax für die Gruppenzuweisung veranschaulicht:

```
"groups": {  
  
  "Custom_Group": "group(var1, var2)",  
  "All_Categorical_plus_one_other": "group(ALL_CATEGORICAL, var2)"  
  
}
```

Gruppennamen müssen mit einem Buchstaben beginnen und können zwischen 1 und 64 Zeichen lang sein. Wenn der Gruppenname nicht mit einem Buchstaben beginnt oder Sonderzeichen (""\t\r\n () \) enthält, muss der Name in Anführungszeichen eingeschlossen werden, damit er in das Rezept aufgenommen werden kann.

Zuweisungen

Zur besseren Lesbarkeit und der Einfachheit halber können Sie eine oder mehrere Transformationen einer Zwischenvariablen zuweisen. Wenn Sie beispielsweise eine Textvariable mit dem Namen "email_subject" haben und Sie die Kleinbuchstaben-Transformation darauf anwenden, können Sie der resultierenden Variable den Namen "email_subject_lowercase" geben, sodass diese an anderer Stelle im Rezept leicht auffindbar ist. Zuweisungen können auch verkettet werden, sodass Sie mehrere Transformationen in einer angegebenen Reihenfolge anwenden können. Das folgende Beispiel zeigt einzelne und verkettete Zuordnungen in der Rezeptsyntax:

```
"assignments": {  
  
  "email_subject_lowercase": "lowercase(email_subject)",  
  
  "email_subject_lowercase_ngram": "ngram(lowercase(email_subject), 2)"  
  
}
```

Namen von Zwischenvariablen müssen mit einem Buchstaben beginnen und können zwischen 1 und 64 Zeichen lang sein. Wenn der Name nicht mit einem Buchstaben beginnt oder Sonderzeichen ("\"t\r\n () \) enthält, muss der Name in Anführungszeichen eingeschlossen werden, damit er in das Rezept aufgenommen werden kann.

Outputs

Der Ausgabeabschnitt steuert, welche Eingabevariablen für den Lernprozess verwendet werden und welche Transformationen darauf angewendet werden. Ein leerer oder nicht vorhandener Ausgabeabschnitt stellt einen Fehler dar, da keine Daten an den Lernprozess übergeben werden.

Der einfachste Ausgabeabschnitt umfasst die vordefinierte Gruppe ALL_INPUTS, die Amazon ML anweist, alle Variablen für den Lernprozess zu verwenden, die in der Datenquelle definiert sind:

```
"outputs": [  
  
  "ALL_INPUTS"  
  
]
```

Der Ausgabeabschnitt kann auch auf die anderen vordefinierten Gruppen verweisen, indem Amazon ML angewiesen wird, alle Variablen in diesen Gruppen zu verwenden:

```
"outputs": [  
  "ALL_NUMERIC",  
  "ALL_CATEGORICAL"  
]
```

Der Ausgabeabschnitt kann auch auf benutzerdefinierte Gruppen verweisen. Im folgenden Beispiel wird nur eine der benutzerdefinierten Gruppen für maschinelles Lernen verwendet, die im Abschnitt mit den Gruppierungszuweisungen im vorhergehenden Beispiel definiert sind. Alle anderen Variablen werden gelöscht:

```
"outputs": [  
  "All_Categorical_plus_one_other"  
]
```

Der Ausgabeabschnitt kann auch auf Variablenzuweisungen verweisen, die im Zuweisungsabschnitt definiert sind:

```
"outputs": [  
  "email_subject_lowercase"  
]
```

Und Eingabevariablen oder Transformationen können direkt im Ausgabenabschnitt definiert werden:

```
"outputs": [  
  "var1",  
  "lowercase(var2)"  
]
```

```
]
```

Die Ausgabe muss explizit alle Variablen und transformierten Variablen angeben, von denen erwartet wird, dass sie für den Lernprozess verfügbar sind. Angenommen, Sie schließen ein kartesisches Produkt aus `var1` und `var2` in die Ausgabe ein. Wenn Sie auch die unformatierten Variablen `var1` und `var2` einschließen möchten, müssen Sie die unformatierten Variablen im Ausgabeabschnitt hinzufügen:

```
"outputs": [  
  "cartesian(var1,var2)",  
  "var1",  
  "var2"  
]
```

Ausgaben können zur besseren Lesbarkeit Kommentare umfassen, indem der Kommentartext zusammen mit der Variablen hinzugefügt wird:

```
"outputs": [  
  "quantile_bin(age, 10) //quantile bin age",  
  "age // explicitly include the original numeric variable along with the  
  binned version"  
]
```

Sie können alle diese Ansätze innerhalb des Ausgabeabschnitts miteinander mischen.

 Note

In der Amazon ML-Konsole sind beim Hinzufügen eines Rezepts keine Kommentare zulässig.

Beispiel eines vollständigen Rezepts

Das folgende Beispiel bezieht sich auf mehrere integrierten Datenprozessoren, die den vorherigen Beispielen eingeführt wurden:

```
{
  "groups": {
    "LONGTEXT": "group_remove(ALL_TEXT, title, subject)",
    "SPECIALTEXT": "group(title, subject)",
    "BINCAT": "group(ALL_CATEGORICAL, ALL_BINARY)"
  },
  "assignments": {
    "binned_age" : "quantile_bin(age,30)",
    "country_gender_interaction" : "cartesian(country, gender)"
  },
  "outputs": [
    "lowercase(no_punct(LONGTEXT))",
    "ngram(lowercase(no_punct(SPECIALTEXT)),3)",
    "quantile_bin(hours-per-week, 10)",
    "hours-per-week // explicitly include the original numeric variable
    along with the binned version",
    "cartesian(binned_age, quantile_bin(hours-per-week,10)) // this one is
    critical",
    "country_gender_interaction",
    "BINCAT"
```

```
]
}
```

Empfohlene Rezepte

Wenn Sie eine neue Datenquelle in Amazon ML erstellen und Statistiken für die Datenquelle berechnet werden, erstellt Amazon ML außerdem eine Rezeptempfehlung, um ein neues ML-Modell aus der Datenquelle zu erstellen. Die empfohlene Datenquelle basiert auf den Daten und in den vorhandenen Daten Zielattributen und bietet einen nützlichen Ausgangspunkt für das Erstellen und Optimieren der ML-Modelle.

Zur Verwendung des empfohlenen Rezepts in der Amazon ML-Konsole wählen Sie Datasource oder Datasource and ML model aus der Dropdown-Liste Create new. Für ML-Modelleinstellungen haben Sie verschiedene Standard- oder benutzerdefinierte Schulungs- und Auswertungseinstellungen im Schritt ML Model Settings des Assistenten für Create ML Model. Wenn Sie die Option "Default" auswählen, verwendet Amazon ML automatisch das empfohlene Rezept. Wenn Sie die Option "Custom" auswählen, wird der Rezepteditor im nächsten Schritt angezeigt, und Sie haben die Möglichkeit, das empfohlenen Rezept zu prüfen oder zu ändern.

Note

Amazon ML ermöglicht Ihnen die Erstellung einer Datenquelle und die sofortige Verwendung zum Erstellen eines ML-Modells, bevor die Statistikberechnung abgeschlossen ist. In diesem Fall können Sie das empfohlene Rezept nicht in der Option "Custom" sehen, aber Sie können über diesen Schritt hinaus fortfahren und Amazon ML das Standardrezept für die Modellschulung verwenden lassen.

Um das empfohlene Rezept mit der Amazon ML-API zu verwenden, können Sie eine leere Zeichenfolge in den Parametern "Recipe" und "RecipeUri API" übergeben. Es ist nicht möglich, das empfohlene Rezept mit der Amazon ML-API abzurufen.

Referenz zur Datentransformation

Themen

- [N-Gramm-Transformation](#)

- [Orthogonal Sparse Bigram \(OSB\)-Transformation](#)
- [Umwandlung in Kleinbuchstaben](#)
- [Transformation zum Entfernen von Satzzeichen](#)
- [Quartile-Binning-Transformation](#)
- [Normierungstransformation](#)
- [Kartesische Produkt-Transformation](#)

N-Gramm-Transformation

Bei einer N-Gramm-Transformation wird eine Textvariable als Eingabe erfasst und Zeichenfolgen erzeugt, die einem Fenster mit "n" Wörtern (benutzerdefiniert) entsprechen. Dabei werden Ausgaben generiert. Im folgenden Beispiel wird die Textzeichenfolge "Ich las dieses Buch wirklich gern".

Durch Angabe der N-Gramm-Transformation mit Fenstergröße = 1 erhalten Sie alle einzelnen Wörter in dieser Zeichenfolge:

```
{"I", "really", "enjoyed", "reading", "this", "book"}
```

Wird für eine N-Gramm-Transformation die Fenstergröße = 2 angegeben, erhalten Sie alle Kombinationen aus zwei Wörtern sowie alle Einzelwortkombinationen:

```
{"I really", "really enjoyed", "enjoyed reading", "reading this", "this book", "I", "really", "enjoyed", "reading", "this", "book"}
```

Durch Angabe der N-Gramm-Transformation mit Fenstergröße = 3 werden der Liste Kombinationen aus drei Wörtern hinzugefügt. Dies führt zu folgendem Ergebnis:

```
{"I really enjoyed", "really enjoyed reading", "enjoyed reading this", "reading this book", "I really", "really enjoyed", "enjoyed reading", "reading this", "this book", "I", "really", "enjoyed", "reading", "this", "book"}
```

Sie können N-Gramme mit einer Größe von 2-10 Wörtern anfordern. N-Gramme mit der Größe 1 werden implizit für alle Eingaben generiert, deren Typ als Text im Datenschema markiert ist, sodass

Sie dies nicht extra angeben müssen. Außerdem sollten Sie bedenken, dass N-Gramme generiert werden, indem die Eingabedaten durch Leerzeichen unterbrochen werden. Das bedeutet, dass beispielsweise Interpunktionszeichen werden als Teil des Worttokens betrachtet werden: Generieren eines N-Gramms mit einem Fenster von 2 für die Zeichenfolge "Rot, Grün, Blau" ergibt {"Rot,", "Grün,", "Blau,", "Rot, Grün", "Grün, Blau"}. Sie können den Satzzeichenentfernungs-Prozessor (siehe weiter unten in diesem Dokument) verwenden, um die Satzzeichen zu entfernen, wenn Sie dies nicht wünschen.

So berechnen Sie N-Gramme mit der Fenstergröße 3 für Variable var1:

```
"ngram(var1, 3)"
```

Orthogonal Sparse Bigram (OSB)-Transformation

Das OSB-Transformation soll die Textzeichenfolgenanalyse unterstützen und ist eine Alternative zur Bi-Gramm-Transformation (N-Gramm mit Fenstergröße 2). Bei der Generierung von OSBs wird ein Fenster der Größe N über den Text geschoben, und es werden alle Wortpaare ausgegeben, die das erste Wort im Fenster enthalten.

Zum Erstellen von OSB werden die enthaltenen Wörter mit einem Unterstrich "_" verbunden und alle übersprungenen Token durch Angabe eines weiteren Unterstrichs im OSB angegeben. Daher codiert die OSB-Transformation nicht nur die Token in einem Fenster, sondern auch die Anzahl der übersprungenen Token in diesem Fenster.

Nehmen Sie beispielsweise die Zeichenfolge "Franz jagt im komplett verwahrlosten Taxi quer durch Bayern" und OSBs der Größe 4. Die sechs Vier-Wort-Fenster und die letzten beiden kleineren Fenster vom Ende der Zeichenfolge werden im folgenden Beispiel angezeigt, zusammen mit den jeweils generierten OSBs:

Fenster, {generierte OSBs}

```
"The quick brown fox", {The_quick, The__brown, The___fox}
```

```
"quick brown fox jumps", {quick_brown, quick__fox, quick___jumps}
```

```
"brown fox jumps over", {brown_fox, brown__jumps, brown___over}
```

```
"fox jumps over the", {fox_jumps, fox__over, fox___the}

"jumps over the lazy", {jumps_over, jumps__the, jumps___lazy}

"over the lazy dog", {over_the, over__lazy, over___dog}

"the lazy dog", {the_lazy, the__dog}

"lazy dog", {lazy_dog}
```

OSBs sind eine Alternative für N-Gramme, die ggf. in einigen Fällen besser funktionieren. Wenn Ihre Daten große Textfelder (10 oder mehr Wörter) haben, experimentieren Sie damit, um zu sehen, welche Option besser geeignet ist. Bitte beachten Sie: Was ein „großes Textfeld“ ist, kann je nach Situation unterschiedlich sein. Bei großen Textfeldern hat sich jedoch empirisch erwiesen, dass OSBs den Text aufgrund des speziellen Überspringen-Symbols (dem Unterstrich) eindeutig darstellen.

Sie können eine Fenstergröße von 2 bis 10 für OSB-Transformationen von Eingabetextvariablen angeben.

So berechnen Sie OSBs mit der Fenstergröße 5 für Variable var1:

```
"osb(var1, 5)"
```

Umwandlung in Kleinbuchstaben

Der Kleinbuchstaben-Transformationsprozessor wandelt Texteingaben in Kleinbuchstaben um. Die Eingabe "Franz jagt im komplett verwaahlsten Taxi quer durch Bayern" wird vom Prozessor beispielsweise in die Ausgabe "franz jagt im komplett verwaahlsten taxi quer durch bayern" umgewandelt.

So wenden Sie Kleinbuchstaben-Transformation auf die Variable var1 an:

```
"lowercase(var1)"
```

Transformation zum Entfernen von Satzzeichen

Amazon ML unterteilt als Text markierte Eingaben im Datenschema implizit bei Leerzeichen. Satzzeichen in der Zeichenfolge ergeben entweder angrenzende Wort-Token oder vollständig separate Token, abhängig von den umgebenden Leerzeichen. Wenn Sie dies verhindern möchten, können Sie die Transformation zum Entfernen der Satzzeichen verwenden, um Satzzeichen aus

generierten Funktionen zu entfernen. Beispielsweise werden für die Zeichenfolge "Willkommen bei AML – bitte verwenden Sie Sicherheitsgurte!" implizit die folgenden Token generiert:

```
{"Welcome", "to", "Amazon", "ML", "-", "please", "fasten", "your", "seat-belts!"}
```

Die Anwendung des Satzzeichenentfernungs-Prozessors auf diese Zeichenfolge ergibt folgenden Satz:

```
{"Welcome", "to", "Amazon", "ML", "please", "fasten", "your", "seat-belts"}
```

Beachten Sie, dass nur vorangestellte und nachfolgende Satzzeichen entfernt werden. Satzzeichen, die mitten in einem Token stehen, z. B. ein Bindestrich in "Sicherheits-Gurt", werden nicht entfernt.

So wenden Sie Satzzeichenentfernung auf die Variable `var1` an:

```
"no_punct(var1)"
```

Quartile-Binning-Transformation

Der Quartile-Binning-Prozessor verwendet zwei Eingaben, nämlich eine numerische Variable und ein als Bin-Nummer bezeichneter Parameter. Die Ausgabe besteht aus einer kategorischen Variable. Der Zweck ist das Erkennen von Nicht-Linearität in der Variablenverteilung durch Gruppierung der beobachteten Werte.

In vielen Fällen ist die Beziehung zwischen einer numerischen Variablen und dem Ziel nicht linear (der numerische Variablenwert wird nicht gleichmäßig mit dem Ziel erhöht oder verringert). In solchen Fällen kann es nützlich sein, die numerische Funktion in eine kategorische Funktion zu packen, um verschiedene Bereiche der numerischen Funktion darzustellen. Jeder kategorische Funktionswert (Bin) kann dann mit einer eigenen linearen Beziehung zum Ziel im Modell dargestellt werden. Nehmen wir an, Sie wissen, dass die kontinuierliche numerische Funktion Kontodauer nicht linear mit der Wahrscheinlichkeit verläuft, ein Buch zu kaufen. Sie können die Dauer also in kategorische Funktionen packen, die in der Lage sind, die Beziehung zum Ziel genauer zu erfassen.

Der Quartile-Binning-Prozessor kann verwendet werden, um Amazon ML anzuweisen, `n` Pakete gleicher Größe basierend auf der Verteilung aller Eingabewerte der Dauer-Variable zu erstellen und dann jede Zahl durch ein Text-Token mit dem darin enthaltenen Paket zu ersetzen. Die optimale Anzahl von Paketen für eine numerische Variable hängt von den Eigenschaften der Variablen und ihrer Beziehung mit dem Ziel ab und wird am besten durch Experimente bestimmt. Amazon ML

schlägt die optimale Paketanzahl für eine numerische Funktion basierend auf den Datenstatistiken im [vorgeschlagenen Rezept](#) vor.

Sie können zwischen 5 und 1000 Quantil-Pakete für jede numerische Eingabevariable berechnet lassen.

Im folgenden Beispiel wird gezeigt, wie 50 Pakete anstelle der numerischen Variablen var1 berechnet und verwendet werden:

```
"quantile_bin(var1, 50)"
```

Normierungstransformation

Die Normierungstransformation normalisiert numerische Variablen auf einen Mittelwert von 0 und die Varianz 1. Die Normierung von numerischen Variablen kann den Lernprozess unterstützen, wenn es sehr große Bereichsunterschiede zwischen numerischen Variablen gibt, da Variablen mit der höchsten Größenordnung das ML-Modell dominieren könnten, unabhängig davon, ob die Funktion in Bezug auf das Ziel informativ ist oder nicht.

Um diese Transformation auf die numerische Variable var1 anzuwenden, fügen Sie Folgendes zum Rezept hinzu:

```
normalize(var1)
```

Diese Transformation kann auch eine benutzerdefinierte Gruppe von numerischen Variablen oder die voreingestellte Gruppe für alle numerischen Variablen (ALL_NUMERIC) als Eingabe verwenden:

```
normalize(ALL_NUMERIC)
```

Hinweis

Es ist nicht erforderlich, den Normierungsprozessor für numerische Variablen zu verwenden.

Kartesische Produkt-Transformation

Die kartesische Transformation generiert Permutationen von zwei oder mehr Text- oder kategorischen Eingabevariablen. Diese Transformation wird verwendet, wenn eine Interaktion zwischen Variablen vermutet wird. Nehmen Sie zum Beispiel den Bank-Marketing-Datensatz im Tutorial: Verwenden von Amazon ML zum Voraussagen der Reaktionen auf ein Marketingangebot. Mit diesem Datensatz möchten wir voraussagen, ob eine Person positiv auf ein Bankangebot

reagiert, basierend auf wirtschaftlichen und demografischen Informationen. Wir könnten vermuten, dass die Person einen wichtigen Job hat (vielleicht gibt es einen Zusammenhang zwischen einer Anstellung in bestimmten Bereichen und verfügbarem Geld), und auch der höchste erreichte Bildungsabschluss ist ebenfalls wichtig. Wir könnten auch eine tiefergehende Intuition haben, dass es eine klare Botschaft in der Interaktion dieser beiden Variablen gibt, z. B., dass die Werbeaktion besonders für Kunden geeignet ist, die Unternehmer mit Hochschulabschluss sind.

Die kartesische Produkt-Transformation nimmt kategorische Variablen oder Text als Eingabe und erzeugt neue Funktionen, um die Interaktion zwischen diesen Eingabevariablen zu erfassen. Insbesondere für jedes Schulungsbeispiel erstellt es eine Kombination aus Funktionen und fügt sie dann als eigenständige Funktion hinzu. Angenommen, unsere vereinfachten Eingabezeilen sehen so aus:

Ziel, Bildung, Job

0, Universität.Abschluss, Techniker

0, Fach.Hochschule, Service

1, Universität.Abschluss, Admin

Wenn wir angeben, dass die kartesischen Transformation auf die kategorischen Variablen "Bildung" und "Job" angewendet wird, sieht die resultierende Funktion Bildung_Job_Interaktion wie folgt aus:

Ziel, Bildung_Job_Interaktion

0, Universität.Abschluss_Techniker

0, Fach.Hochschule_Service

1, Universität.Abschluss_Admin

Die kartesische Transformation ist sogar noch leistungsfähiger, wenn es um die Bearbeitung von Token-Sequenzen geht. Dies ist der Fall, wenn es sich bei einem der Argumente um eine Textvariable handelt, die implizit oder explizit in Token aufgeteilt ist. Nehmen Sie beispielsweise die Klassifizierungsaufgabe, ob ein Buch als Lehrbuch eingesetzt wird oder nicht. Intuitiv denken wir, dass etwas im Buchtitel uns verrät, ob es sich um ein Lehrbuch handelt (bestimmte Wörter können häufiger in Lehrbuchtiteln auftauchen), und wir vermuten auch den Bucheinband (Fachbücher sind mit höherer Wahrscheinlichkeit gebunden), aber es ist in Wirklichkeit die Kombination von einigen Wörtern im Titel und der Einband, die die beste Voraussage ergeben. Für ein praktisches Beispiel

finden Sie in der folgenden Tabelle die Ergebnisse der Anwendung des kartesischen Prozessors auf die Eingabevariablen für Einband und Titel:

Lehr	Titel	Einband	Kartesisches Produkt aus no_punct(Titel) und Einband
1	Wirtschaft: Prinzipien, Probleme, Richtlinien	Hardcover	{"Wirtschaft_Hardcover", "Prinzipien_Hardcover", "Probleme_Hardcover", "Richtlinien_Hardcover"}
0	Das unsichtbare Herz: Eine Wirtschaftromanze	Taschenbuch	{"Das_Taschenbuch", "unsichtbare_Taschenbuch", "Herz_Taschenbuch", "Eine_Taschenbuch", "Wirtschaftromanze_Taschenbuch"}
0	Spaß mit Problemen	Taschenbuch	{"Spaß_Taschenbuch", "mit_Taschenbuch", "Problemen_Taschenbuch"}

Das folgende Beispiel zeigt, wie Sie die kartesische Transformation auf var1 und var2 anwenden:

```
cartesian(var1, var2)
```

Neuordnung von Daten

Mit der Funktionalität Neuordnung von Daten können Sie eine Datenquelle erstellen, die lediglich auf einem Teil der Eingabedaten basiert, auf die sie verweist. Wenn Sie beispielsweise ein ML-Modell mit dem Erstellen eines ML-Modells Assistenten auf der Amazon ML-Konsole und wählen Sie die Standard-Bewertungsoption. Amazon ML behält automatisch 30% Ihrer Daten für die ML-Modellauswertung und verwendet die übrigen 70% für die Schulung. Diese Funktionalität wird durch die -Funktion Neuordnung von Daten aktiviert.

Wenn Sie die Amazon ML-API verwenden, um Datenquellen zu erstellen, können Sie angeben, auf welchem Teil der Eingabedaten die neue Datenquelle basieren soll. Dazu übergeben Sie im Parameter `DataRearrangement` Anweisungen an die APIs `CreateDataSourceFromS3`, `CreateDataSourceFromRedshift` oder `CreateDataSourceFromRDS`. Der Inhalt der `DataRearrangement`-Zeichenfolge ist eine JSON-Zeichenfolge, welche die Anfangs- und End-Standorte Ihrer Daten in Prozentsätzen, eine Ergänzungsmarkierung und eine Aufteilungsstrategie enthält. Die folgende `DataRearrangement`-

Zeichenfolge gibt beispielsweise an, dass die ersten 70 % der Daten verwendet werden, um die Datenquelle zu erstellen:

```
{
  "splitting": {
    "percentBegin": 0,
    "percentEnd": 70,
    "complement": false,
    "strategy": "sequential"
  }
}
```

DataRearrangement-Parameter

Wenn Sie die Vorgehensweise bei der Erstellung von Datenquellen bei Amazon ML ändern möchten, verwenden Sie die folgenden Parameter.

PercentBegin (Optional)

Verwenden Sie `percentBegin`, um anzugeben, wo die Daten für die Datenquelle beginnen. Wenn Sie nicht einschließen `percentBegin` und `percentEnd` enthält Amazon ML bei der Erstellung der Datenquelle alle Daten.

Gültige Werte sind 0 bis einschließlich 100.

PercentEnd (Optional)

Verwenden Sie `percentEnd`, um anzugeben, wo die Daten für die Datenquelle enden. Wenn Sie nicht einschließen `percentBegin` und `percentEnd` enthält Amazon ML bei der Erstellung der Datenquelle alle Daten.

Gültige Werte sind 0 bis einschließlich 100.

Complement (Optional)

Der Parameter `complement` weist Amazon ML an, Daten zu verwenden, die nicht im Bereich von `percentBegin` zu `percentEnd` eine Datenquelle zu erstellen. Der Parameter `complement` ist nützlich, wenn Sie ergänzende Datenquellen zu Schulungs- und Auswertungszwecken erstellen müssen. Um eine ergänzende Datenquelle zu erstellen, verwenden Sie die gleichen Werte für `percentBegin` und `percentEnd` mit dem Parameter `complement`.

Die beiden folgenden Datenquellen teilen beispielsweise keine Daten und können verwendet werden, um ein Modell zu schulen und auszuwerten. Die erste Datenquelle besteht aus 25 % und die zweite aus 75 % der Daten.

Auswertungsdatenquelle:

```
{
  "splitting":{
    "percentBegin":0,
    "percentEnd":25
  }
}
```

Schulungsdatenquelle:

```
{
  "splitting":{
    "percentBegin":0,
    "percentEnd":25,
    "complement":"true"
  }
}
```

Gültige Werte sind `true` und `false`.

Strategy (Optional)

Wenn Sie die Vorgehensweise bei Amazon ML die Daten für eine Datenquelle ändern möchten, verwenden Sie den `strategy`-Parameter.

Der Standardwert für den `strategy`-Parameter ist `sequential`, was bedeutet, dass Amazon ML alle Datensätze zwischen dem `percentBegin` und `percentEnd` Parameter für die Datenquelle in der Reihenfolge, in der sie in den Eingabedaten erscheinen.

Die folgenden beiden `DataRearrangement`-Zeilen sind Beispiele für sequentiell geordnete Schulungs- und Auswertungsdatenquellen:

Auswertungsdatenquelle: `{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential"}}`

Schulungsdatenquelle: `{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential", "complement":"true"}}`

Wenn Sie eine Datenquelle aus einer Zufallsauswahl von Daten erstellen möchten, setzen Sie den Parameter `strategy` auf `random` und geben Sie eine Zeichenfolge an, die als Ausgangswert für die zufällige Datenaufteilung verwendet wird (z. B. den S3-Pfad zu Ihren Daten als zufällige Seed-Zeichenfolge). Wenn Sie die zufällige Verteilungsstrategie wählen, weist Amazon ML jeder Datenzeile eine Pseudo-Zufallszahl zu und wählt dann die Zeilen mit einer zugeordneten Zahl zwischen `percentBegin` und `percentEnd` aus. Pseudo-Zufallszahlen werden mit dem Byte-Offset als Seed zugewiesen, sodass die Datenergebnisse anders aufgeteilt werden. Alle vorhandenen Reihenfolgen bleiben erhalten. Die zufällige Aufteilungsstrategie stellt sicher, dass die Variablen der Schulungs- und Auswertungsdaten gleichmäßig verteilt werden. Dies ist nützlich, wenn die Eingabedaten möglicherweise eine implizite Sortierreihenfolge besitzen, was ansonsten dazu führen würde, dass Schulungs- und Auswertungsdatenquellen nicht-ähnliche Datensätze enthalten würden.

Die folgenden beiden `DataRearrangement`-Zeilen sind Beispiele für nicht-sequentiell geordnete Schulungs- und Auswertungsdatenquellen:

Auswertungsdatenquelle:

```
{
  "splitting":{
    "percentBegin":70,
    "percentEnd":100,
    "strategy":"random",
    "strategyParams": {
      "randomSeed":"RANDOMSEED"
    }
  }
}
```

Schulungsdatenquelle:

```
{
  "splitting":{
    "percentBegin":70,
    "percentEnd":100,
    "strategy":"random",
    "strategyParams": {
      "randomSeed":"RANDOMSEED"
    }
  }
  "complement":"true"
}
```

```
}
```

Gültige Werte sind `sequential` und `random`.

(Optional) Strategy:RandomSeed

Amazon ML verwendet die `RandomSeed` um die Daten zu trennen. Der Standard-Seed für die API ist eine leere Zeichenfolge. Um einen Seed für die zufällige Aufteilungsstrategie anzugeben, übergeben Sie eine Zeichenfolge. Weitere Information zufällige Seeds finden Sie unter [Zufällige Aufteilung Ihrer Daten](#) im Amazon Machine Learning Learning-Entwicklerleitfaden.

Einen Beispielcode für die Verwendung der Cross-Validierung mit Amazon ML finden Sie unter [Github Machine Learning Learning-Beispiele](#).

Evaluation von ML-Modellen

Sie sollten ein Modell evaluieren, um festzustellen, ob es das Ziel in neuen und zukünftigen Daten gut voraussagen kann. Da künftigen Instances unbekannte Zielwerte haben, müssen Sie die Richtigkeitsmetrik des ML-Modells für Daten prüfen, deren Zielantwort Sie bereits kennen, und diese Bewertung als Proxy für die Voraussagerichtigkeit für zukünftige Daten verwenden.

Um ein Modell ordnungsgemäß zu bewerten, sollten Sie eine Stichprobe der Daten zurückhalten, die mit dem Ziel (Referenzdaten) aus der Schulungsdatenquelle gekennzeichnet wurden. Das Testen der Voraussagerichtigkeit eines ML-Modell mit denselben Daten, die für die Schulung verwendet wurden, ist nicht sinnvoll, weil sich Modelle an spezifische Schulungsdaten „erinnern“ anstatt sie zu verallgemeinern. Sobald Sie die Schulung des ML-Modells abgeschlossen haben, senden Sie die dem Modell die zurückgehaltenen Beobachtungen, deren Zielwerte Sie kennen. Anschließend vergleichen Sie die vom ML-Modell zurückgegebenen Voraussagen mit dem bekannten Zielwert. Schließlich berechnen Sie eine Zusammenfassungsmetrik, die Ihnen sagt, wie gut die prognostizierten und tatsächlichen Werte übereinstimmen.

In Amazon ML bewerten Sie ein ML-Modell, indem Sie Erstellen einer Bewertung aus. Um eine Bewertung für ein ML-Modell zu erstellen, benötigen Sie ein zu bewertendes ML-Modell und gekennzeichnete Daten, die nicht für Schulungszwecke verwendet wurden. Erstellen Sie zuerst eine Datenquelle für die Bewertung, indem Sie mit den zurück gehaltenen Daten eine Amazon ML-Datenquelle erstellen. Die bei der Evaluation verwendeten Daten müssen dasselbe Schema aufweisen wie die in der Schulung verwendeten Daten, und sie müssen tatsächlichen Werte für die Zielvariable enthalten.

Wenn sich Ihre Daten in einer einzigen Datei oder in einem Verzeichnis befinden, können Sie sie mithilfe der Amazon ML-Konsole trennen. Der Standard-Pfad im Create ML-Modellassistenten trennt die eingegebenen Datenquelle und verwendet die ersten 70 % als Schulungsdatenquelle sowie die verbleibenden 30 % als Evaluationsdatenquelle. Sie können das Teilungsverhältnis auch anpassen, indem Sie die Option Custom im Create ML-Modellassistenten verwenden, mit der Sie ein zufälliges Beispiel von 70 % für Schulungen und die verbleibenden 30 % für die Evaluation verwenden können. Um weitere benutzerdefinierte Teilungsverhältnisse anzugeben, verwenden Sie die Zeichenfolge zur Neuerstellung von Daten in der [Create Datasource](#)-API. Sobald Sie über eine Evaluationsdatenquelle und ein ML-Modell verfügen, können Sie eine Bewertung erstellen und die Ergebnisse davon überprüfen.

Themen

- [Einblicke in ML-Modelle](#)
- [Einblicke in binäre Modelle](#)
- [Einblicke in Mehrklassen-Modelle](#)
- [Regressionsmodell-Einblicke](#)
- [Verhindern von Overfitting](#)
- [Kreuzvalidierung](#)
- [Auswertungswarnungen](#)

Einblicke in ML-Modelle

Zur Auswertung eines ML-Modells bietet Amazon ML eine Metrik nach Branchenstandard sowie eine Reihe von Einblicken, um die Prognosegenauigkeit Ihres Modells zu prüfen. In Amazon ML umfasst das Ergebnis einer Auswertung Folgendes:

- Eine Prognosegenauigkeits-Metrik, einen Bericht zum Gesamterfolg des Modells
- Visualisierungen, mit denen die Genauigkeit Ihres Modells über die Prognosegenauigkeits-Metrik hinaus dargestellt wird
- Die Möglichkeit, die Auswirkungen der Einstellung eines Schwellenwerts zu überprüfen (nur für binäre Klassifizierung)
- Warnungen zu Kriterien, um die Gültigkeit der Auswertung zu überprüfen

Die Wahl der Metrik und Visualisierung ist abhängig von der Art des ML-Modells, das Sie testen. Es ist wichtig, diese Visualisierungen zu überprüfen, um zu entscheiden, wann Ihr Modell gut genug ist, um Ihre geschäftlichen Anforderungen zu erfüllen.

Einblicke in binäre Modelle

Interpretieren der Voraussagen

Die tatsächliche Ausgabe von vielen binären Klassifizierungsalgorithmen ist eine Voraussagepunktzahl. Die Punktzahl gibt die Sicherheit des Systems an, dass die angegebene Beobachtung der positiven Klasse angehört (der tatsächliche Zielwert ist 1). Binäre Klassifizierungsmodelle in Amazon ML geben ein Ergebnis im Bereich von 0 bis 1 aus. Als Verbraucher dieser Bewertung können Sie die Punktzahl interpretieren, indem Sie einen

Klassifizierungsschwellenwert oder Grenzwert festlegen und die Punktzahl damit vergleichen, um zu entscheiden, ob die Beobachtung als 1 oder 0 klassifiziert wird. Alle Beobachtungen mit Ergebnissen über dem Grenzwert werden als "Ziel = 1" vorausgesagt; Beobachtungen mit Ergebnissen unter dem Grenzwert werden als "Ziel = 0" vorausgesagt.

In Amazon ML ist der standardmäßige Ergebnisgrenzwert 0,5. Sie können diesen Grenzwert Ihren geschäftlichen Anforderungen entsprechend ändern. Sie können die Visualisierungen in der Konsole verwenden, um zu verstehen, wie sich die Auswahl des Grenzwerts auf Ihre Anwendung auswirkt.

Messung der ML-Modellgenauigkeit

Amazon ML bietet eine Genauigkeitsmetrik nach Branchenstandard für binäre Klassifizierungsmodelle, genannt Fläche unter der Kurve (AUC) oder Receiver Operating Characteristic (ROC). AUC misst die Fähigkeit des Modells, eine höhere Bewertung für positive Beispiele im Vergleich zu negativen Beispielen vorherzusagen. Da es unabhängig vom Grenzwert ist, bekommen Sie ein Gefühl für die Voraussagegenauigkeit Ihres Modells aus der AUC-Metrik, ohne einen Schwellenwert auszuwählen.

Die AUC-Metrik gibt einen Dezimalwert zwischen 0 und 1 zurück. AUC-Werte nahe 1 weisen auf ein sehr genaues ML-Modell hin. Werte um 0,5 weisen auf ein ML-Modell hin, das nicht besser als zufälliges Raten ist. Werte um 0 sind ungewöhnlich und weisen in der Regel auf ein Problem mit den Daten hin. Im Wesentlichen gibt ein AUC in der Nähe von 0 an, dass das ML-Modell die richtigen Muster gelernt hat, aber diese verwendet, um Voraussagen zu geben, die im Gegensatz zur Realität stehen (0 wird als 1 vorhergesagt und umgekehrt). Weitere Informationen zur AUC finden Sie auf der Seite [Receiver Operating Characteristic](#) in Wikipedia.

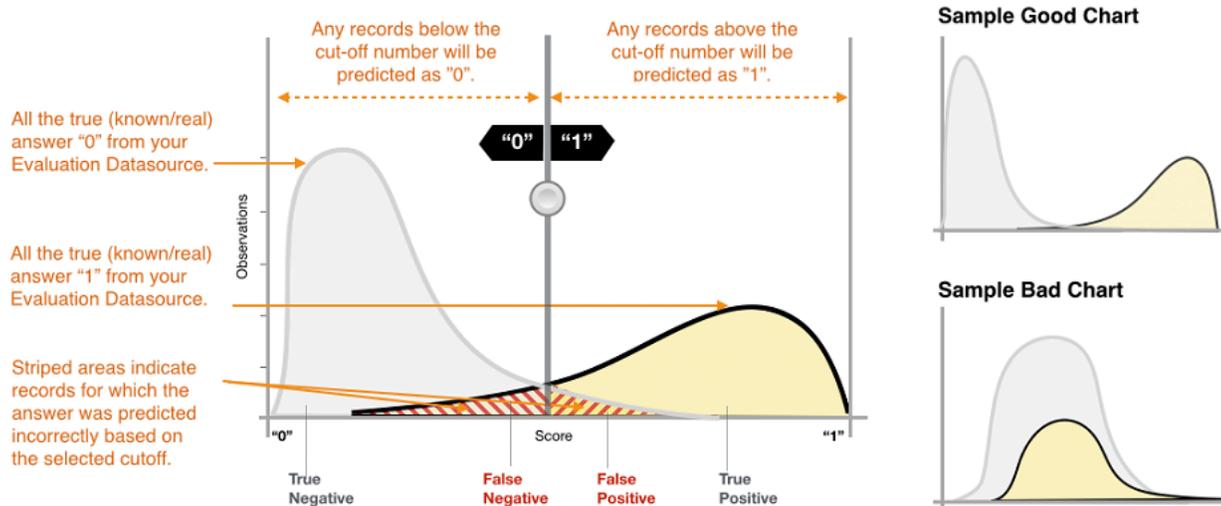
Die grundlegende AUC-Metrik für ein binäres Modell ist 0,5. Dies ist der Wert für ein hypothetisches ML-Modell, das zufällig eine Antwort von 1 oder 0 vorausagt. Ihr binäres ML-Modell sollte besser als das funktionieren, um einen Mehrwert zu bieten.

Verwenden der Performance-Visualisierung

Sehen Sie zur Prüfung der Genauigkeit des ML-Modells die Diagramme in der Bewertung-Seite auf der Amazon ML-Konsole. Diese Seite zeigt Ihnen zwei Histogramme: a) ein Histogramm der Punktzahlen für die tatsächlichen positiven Ergebnisse (das Ziel ist 1) und b) ein Histogramm der Punktzahlen für die tatsächlichen negativen Ergebnisse (das Ziel ist 0) in den Auswertungsdaten.

Ein ML-Modell mit guter Voraussagegenauigkeit sagt höhere Ergebnisse für die tatsächliche 1 und niedrigere Ergebnisse für die tatsächliche 0 voraus. Ein perfektes Modell zeigt in den zwei Histogramme an den zwei verschiedenen Enden der X-Achse, dass alle tatsächlichen

positiven Ergebnisse hohe Punktzahlen und die tatsächlichen negativen Ergebnisse niedrige Punktzahlen haben. ML-Modelle machen jedoch auch Fehler, und ein typisches Diagramm zeigt Überschneidungen an bestimmten Punktzahlen. Ein extrem schlechtes Modell unterscheidet nicht zwischen den positiven und negativen Klassen, und beide Klassen haben vorwiegend überlappende Histogramme.



Anhand von Visualisierungen können Sie die Anzahl der Voraussagen bestimmen, die in zwei Arten von richtigen Voraussagen und zwei Arten von falschen Voraussagen unterteilt werden.

Richtige Voraussagen

- Wirkliches Positiv (TP): Amazon ML hat den Wert 1 vorausgesagt, und der tatsächliche Wert ist 1.
- Wirkliches Negativ (TN): Amazon ML hat den Wert 0 vorausgesagt, und der tatsächliche Wert ist 0.

Falsche Voraussagen

- Falsches Positiv (FP): Amazon ML hat den Wert 1 vorausgesagt, aber der tatsächliche Wert ist 0.
- Falsches Negativ (FN): Amazon ML hat den Wert 0 vorausgesagt, aber der tatsächliche Wert ist 1.

i Note

Die Anzahl von TP, TN, FP und FN richtet sich nach dem ausgewählten Schwellenwert, und die Optimierung für eine dieser Zahlen würde einen Kompromiss bei den anderen bedeuten. Eine hohe Anzahl von TPs führt in der Regel zu einer hohen Anzahl von FPs und einer geringen Anzahl von TNs.

Anpassen des Ergebnisgrenzwerts

ML-Modelle arbeiten mittels Generierung von numerischen Voraussagepunktzahlen und Anwenden eines Grenzwertes, um diese Punktzahlen in binäre 0/1-Etiketten zu konvertieren. Durch das Ändern des Grenzwerts können Sie das Verhalten des Modells anpassen, wenn es einen Fehler macht. Auf der Bewertung in der Amazon ML-Konsole können Sie die Auswirkungen von verschiedenen Ergebnisgrenzwerten überprüfen und den Ergebnisgrenzwert speichern, den Sie für Ihr Modell nutzen möchten.

Beachten Sie bei der Anpassung des Schwellenwerts für den Ergebnisgrenzwert den Kompromiss zwischen den beiden Arten von Fehlern. Wenn Sie den Grenzwert nach links verschieben, werden mehr positive Ergebnisse erfasst, aber der Kompromiss besteht in der Erhöhung der Anzahl von falschen Positivangaben. Verschieben nach rechts ergibt weniger falsche Positivangaben, aber der Kompromiss ist, dass einige tatsächliche positive Ergebnisse nicht erfasst werden. Für die Voraussageanwendung können Sie entscheiden, welche Art von Fehler eher akzeptabel ist, indem Sie einen geeigneten Ergebnisgrenzwert festlegen.

Überprüfen von erweiterten Metriken

Amazon ML bietet die folgenden zusätzlichen Metriken, um die Voraussagegenauigkeit des ML-Modells zu messen: Richtigkeit, Genauigkeit, Wiedererkennbarkeit und Falschpositivrate.

Accuracy

Die Richtigkeit (ACC) misst den Anteil der richtigen Voraussagen. Der Bereich liegt zwischen 0 und 1. Ein größerer Wert gibt eine bessere Richtigkeit an:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Genauigkeit

Genauigkeit misst den Anteil der tatsächlichen Positiva unter den Beispielen, die als positiv vorausgesagt wurden. Der Bereich liegt zwischen 0 und 1. Ein größerer Wert gibt eine bessere Richtigkeit an:

$$Precision = \frac{TP}{TP + FP}$$

Wiedererkennung

Wiedererkennung misst den Anteil der tatsächlichen Positiva, die als positiv vorausgesagt wurden. Der Bereich liegt zwischen 0 und 1. Ein größerer Wert gibt eine bessere Richtigkeit an:

$$Recall = \frac{TP}{TP + FN}$$

Falschpositivrate

Die Falschpositivrate (FPR) misst den Anteil falscher Alarme oder den Anteil tatsächlicher negativer Ergebnisse, die als positiv vorhergesagt wurden. Der Bereich liegt zwischen 0 und 1. Ein kleinerer Wert gibt eine bessere Richtigkeit der Voraussage an:

$$FPR = \frac{FP}{FP + TN}$$

Abhängig von Ihrem Unternehmensproblem benötigen Sie vielleicht eher ein Modell, das für eine bestimmte Teilmenge dieser Metriken gut funktioniert. Zwei Unternehmensanwendungen können beispielsweise sehr unterschiedliche Anforderungen an ihr ML-Modell haben:

- Eine Anwendung muss vielleicht sehr sicher sein, dass die positiven Voraussagen tatsächlich positiv sind (hohe Präzision) und kann es verkräften, dass einige positive Beispiele falsch als negativ klassifiziert werden (moderate Wiedererkennung).
- Eine andere Anwendung soll so viele positive Beispiele wie möglich korrekt voraussagen (hohe Wiedererkennung) und nimmt es in Kauf, dass einige negative Beispiele falsch als positiv klassifiziert werden (moderate Genauigkeit).

In Amazon ML können Sie einen Ergebnismesswert festlegen, der einem bestimmten Wert von einer der oben genannten erweiterten Metriken entspricht. Außerdem werden die Kompromisse durch Optimierung für eine der Metriken angezeigt. Wenn Sie beispielsweise einen Messwert auswählen, der eine hohe Genauigkeit erzielt, erhalten Sie dafür in der Regel eine geringere Wiedererkennung.

Note

Sie müssen den Ergebnismesswert speichern, damit er für die Klassifizierung künftiger Voraussagen Ihrer ML-Modelle wirksam wird.

Einblicke in Mehrklassen-Modelle

Interpretieren der Voraussagen

Die tatsächliche Ausgabe eines Mehrklassen-Klassifizierungsalgorithmus ist ein Satz von Voraussagepunktzahlen. Die Punktzahl gibt die Gewissheit des Modells an, dass die angegebene Beobachtung zu jeder der Klassen gehört. Im Gegensatz zu binären Klassifizierungsproblemen müssen Sie für Voraussagen keinen Ergebnismittelwert auswählen. Die vorhergesagte Antwort ist die Klasse (z. B. Bezeichnung) mit der höchsten vorhergesagten Punktzahl.

Messung der ML-Modellgenauigkeit

Typische Metriken, die in Mehrklassen-Modellen verwendet werden, sind dieselben Metriken, die auch bei der binären Klassifizierung verwendet werden, nachdem über alle Klassen hinweg der Durchschnitt hierfür berechnet wurde. In Amazon ML wird die F1-Bewertung mit Makro-Durchschnitt zum Auswerten der prädiktiven Richtigkeit einer Mehrklassenmetrik verwendet.

F1-Bewertung mit Makro-Durchschnitt

Die F1-Bewertung ist eine binäre Klassifizierungsmetrik, die sowohl die binäre Metrikpräzision als auch Rückruf berücksichtigt. Sie ist das harmonische Mittel zwischen Präzision und Rückruf. Der Bereich liegt zwischen 0 und 1. Ein größerer Wert gibt eine bessere Richtigkeit an:

$$F1 \text{ score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Die F1-Bewertung mit Makro-Durchschnitt ist der nicht gewichtete Durchschnitt der F1-Bewertung über alle Klassen in dem Mehrklassen-Modell. Sie berücksichtigt nicht die Frequenz des Auftretens der Klassen im Auswertungsdatensatz. Ein größerer Wert gibt eine bessere prädiktive Richtigkeit an: Das folgende Beispiel zeigt K Klassen in der Auswertungsdatenquelle:

$$\text{Macro average F1 score} = \frac{1}{K} \sum_{k=1}^K \text{F1 score for class } k$$

Grundlegende F1-Bewertung mit Makro-Durchschnitt

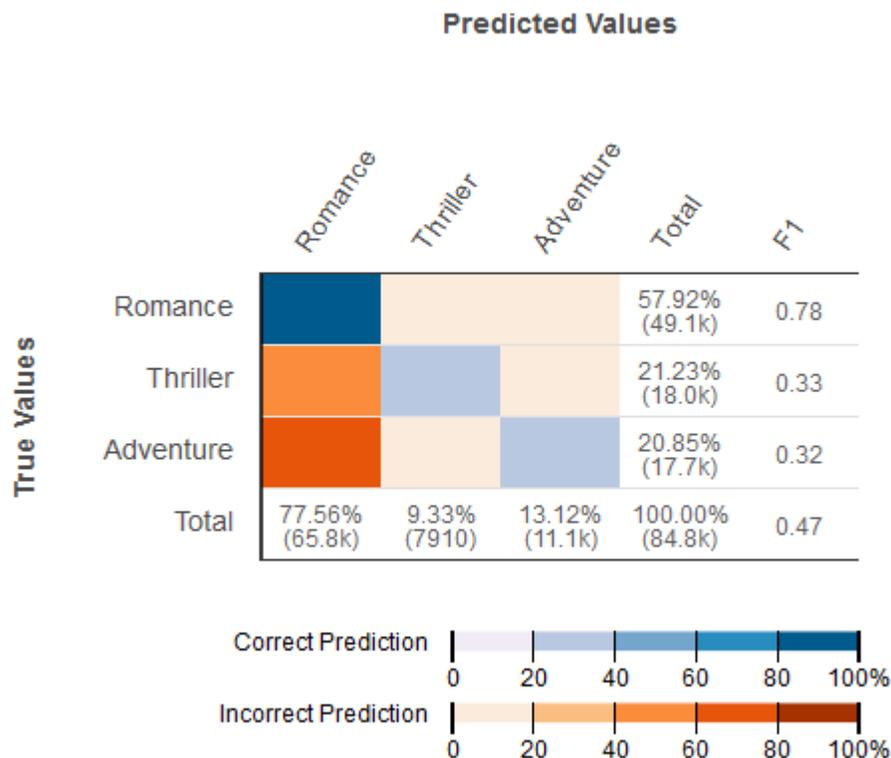
Amazon ML bietet eine grundlegende Metrik für Mehrklassen-Modelle. Es handelt sich um die F1-Bewertung mit Makro-Durchschnitt für ein hypothetisches Mehrklassen-Modell, das immer die häufigste Klasse als Antwort vorhersagen würde. Wenn Sie beispielsweise das Genre eines Films

vorhersagen würden, und das gängigste Genre in Ihren Schulungsdaten ist "Romanze", so würde das grundlegende Modell das Genre immer als "Romanze" vorhersagen. Sie würden das ML-Modell mit dieser Grundlage vergleichen, um auszuwerten, ob das ML-Modell besser als ein ML-Modell ist, das diese konstante Antwort vorhersagt.

Verwenden der Performance-Visualisierung

Amazon ML bietet eine Konfusionsmatrix als Möglichkeit zur Visualisierung der Richtigkeit von prädiktiven Mehrklassen-Klassifizierungsmodellen. Die Konfusionsmatrix veranschaulicht in einer Tabelle die Anzahl oder den Prozentwert richtiger und falscher Voraussagen für jede Klasse, indem die vorhergesagte Klasse einer Beobachtung mit der tatsächlichen Klasse verglichen wird.

Wenn Sie beispielsweise versuchen, einen Film in ein Genre zu klassifizieren, sagt das prädiktive Modell möglicherweise hervor, dass das Genre (Klasse) "Romanze" ist. Der tatsächliche Genre ist jedoch möglicherweise "Thriller". Wenn Sie die Richtigkeit eines Mehrklassen-Klassifizierungsmodells auswerten, identifiziert Amazon ML diese Fehlklassifikationen und zeigt die Ergebnisse in der Konfusionsmatrix an, wie in der folgenden Abbildung dargestellt.



Die folgenden Informationen werden in einer Konfusionsmatrix angezeigt:

- Anzahl der korrekten und falschen Vorhersagen für jede Klasse: Jede Zeile in der Konfusionsmatrix entspricht den Metriken für eine der tatsächlichen Klassen. In der ersten Zeile wird beispielsweise angezeigt, dass für Filme, die sich eigentlich im Genre "Romanze" befinden, die Voraussagen des Mehrklassen-ML-Modells in über 80 % der Fälle korrekt sind. Das Genre "Thriller" wird in weniger als 20 % der Fälle falsch vorhergesagt, genau wie das Genre "Abenteuer".
- Klassenweise F1-Bewertung mit Klassenweise: In der letzten Spalte wird die F1-Bewertung für die einzelnen Klassen angezeigt.
- Echte Klassenfrequenzen in den Auswertungsdaten: In der vorletzten Spalte wird angezeigt, dass im Auswertungsdatensatz 57,92% der Beobachtungen in den Auswertungsdaten unter „Romanze“ fallen, 21,23% unter „Thriller“ und 20,85% unter „Abenteuer“.
- Vorhergesagte Klassenfrequenzen für die Auswertungsdaten: In der letzten Zeile wird die Häufigkeit jeder Klasse in den Vorhersagen angezeigt. 77,56% der Beobachtungen werden „Romanze“ vorhergesagt, 9,33% als „Thriller“ und 13,12% als „Abenteuer“.

Die Amazon ML-Konsole bietet eine visuelle Anzeige, die bis zu 10 Klassen in der Konfusionsmatrix aufnimmt, aufgeführt von der häufigsten bis zur am wenigsten häufigen Klasse in den Auswertungsdaten. Wenn Ihre Auswertungsdaten mehr als 10 Klassen aufweisen, werden die 9 am häufigsten auftretenden Klassen in der Konfusionsmatrix angezeigt, und alle anderen Klassen werden in einer Klasse mit der Bezeichnung "Weitere" reduziert. Amazon ML bietet auch die Möglichkeit, die vollständige Konfusionsmatrix über einen Link auf der Mehrklassen-Visualisierungsseite herunterzuladen.

Regressionsmodell-Einblicke

Interpretieren der Voraussagen

Die Ausgabe eines ML-Regressionsmodells ist ein numerischer Wert für die Modellvoraussage des Ziels. Wenn Sie beispielsweise Wohnungspreise voraussagen, kann die Voraussage des Modells ein Wert wie 254 013 sein.

Note

Der Ergebnisbereich der Voraussagen muss nicht mit dem Bereich des Ziels in den Schulungsdaten übereinstimmen. Nehmen wir beispielsweise an, dass Sie Wohnungspreise voraussagen und das Ziel in den Schulungsdaten Werte zwischen 0 und 450 000 hatte. Das vorausgesagte Ziel muss nicht in diesem Bereich liegen. Es kann jeden beliebigen positiven

Wert (über 450 000) oder negativen Wert (kleiner als 0) haben. Es ist wichtig, einen Plan für die Handhabung von Voraussagewerten außerhalb des akzeptablen Bereichs für Ihre Anwendung zu haben.

Messung der ML-Modellgenauigkeit

Für Regressionsaufgaben verwendet Amazon ML die RMSE-Metrik (Root Mean Square Error, mittlerer quadratischer Vorhersagefehler) nach Branchenstandard. Es handelt sich um ein Maß für die Abweichung zwischen dem vorausgesagten numerischen Ziel und der tatsächlichen numerischen Antwort (Referenzdaten). Je kleiner der RMSE-Wert ist, umso höher ist die Voraussagegenauigkeit des Modells. Ein Modell mit absolut richtigen Voraussagen hat einen RMSE von 0. Das folgende Beispiel zeigt Evaluierungsdaten mit n Datensätzen:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\text{actual target} - \text{predicted target})^2}$$

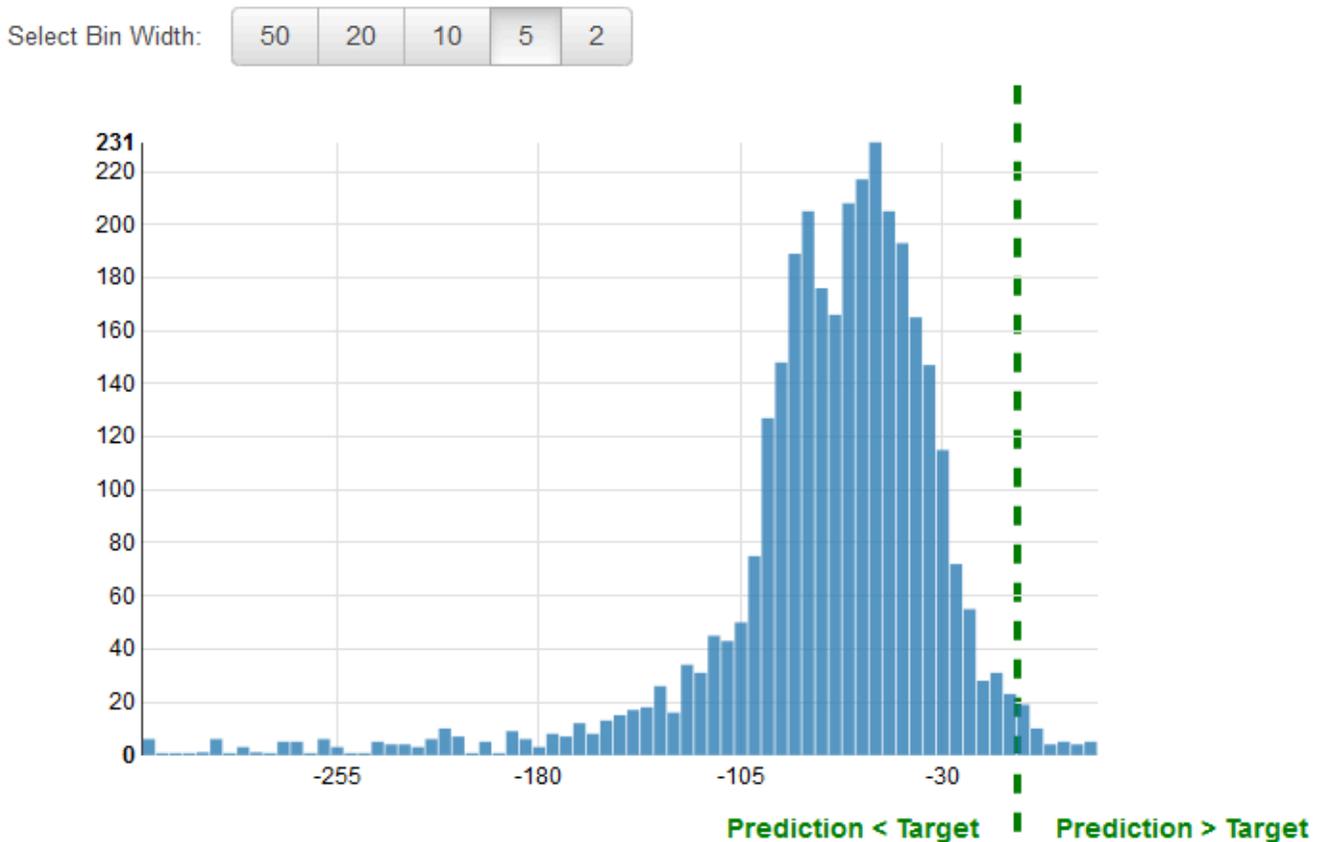
Basis-RMSE

Amazon ML bietet eine Basis-Metrik für Regressionsmodelle. Dabei handelt es sich um den RMSE für ein hypothetisches Regressionsmodell, das immer den Durchschnitt des Ziel als Antwort voraussagt. Wenn Sie beispielsweise das Alter eines Hauskäufer voraussagen und das Durchschnittsalter für die Beobachtungen in Ihren Schulungsdaten 35 ist, sagt das Basismodell immer 35 als Antwort voraus. Sie können Ihr ML-Modell dann mit diesem Basismodell vergleichen, um zu ermitteln, ob Ihr ML-Modell besser ist als ein ML-Modell, das diese konstante Antwort vorhersagt.

Verwenden der Performance-Visualisierung

Es ist eine gängige Vorgehensweise, den Rest bei Regressionsproblemen zu überprüfen. Ein Rest für eine Beobachtung in der Evaluierungsdaten ist der Unterschied zwischen dem wahren Ziel und dem vorausgesagten Ziel. Reste stellen den Teil des Ziels dar, den das Modell nicht voraussagen konnte. Ein positiver Rest deutet darauf hin, dass das Modell das Ziel unterschätzt (das tatsächliche Ziel ist größer als das vorausgesagte Ziel). Ein negativer Rest deutet auf eine Überbewertung hin (das tatsächliche Ziel ist kleiner als das vorausgesagte Ziel). Das Histogramm der Reste für die Evaluierungsdaten deutet bei glockenförmiger Anordnung und Zentrierung auf Null darauf hin, dass das Modell willkürliche Fehler macht und keinen spezifischen Zielwertbereich systematisch über- oder unterschätzt. Wenn die Reste keine Glockenform mit Zentrierung auf Null bilden, gibt es eine gewisse

Struktur bei den Voraussagefehlern des Modells. Das Hinzufügen von weiteren Variablen kann es dem Modell ermöglichen, Muster zu erfassen, die vom aktuellen Modell nicht erfasst werden. Die folgende Abbildung zeigt Reste, die nicht um Null zentriert sind.



Verhindern von Overfitting

Beim Erstellen und Schulen eines ML-Modells soll das Modell ausgewählt werden, das die besten Voraussagen liefert, das bedeutet, es wird das Modell mit den besten Einstellungen (ML-Modelleinstellungen oder Hyperparameter) gewählt. In Amazon Machine Learning können Sie vier Hyperparameter einstellen: Anzahl Durchläufe, Regularisation, Modellgröße und Shuffle-Typ. Wenn Sie jedoch Modellparametereinstellungen wählen, welche die „beste“ Voraussageleistung für Evaluationsdaten produzieren, können Sie Ihr Modell leicht „überanpassen“. Overfitting tritt auf, wenn ein Modell über gelernte Muster verfügt, die bei der Schulung und in Evaluationsdatenquellen auftreten, diese aber nicht auf allgemein auf Daten angewendet werden können. Dies geschieht häufig, wenn die Schulungsdaten alle Daten bei der Evaluation verwendeten Daten umfassen. Ein überangepasstes Modell zeigt bei einer Evaluation eine hervorragende Leistung, versagt jedoch bei der Anwendung auf ungesehene Daten.

Um zu verhindern, dass ein überangepasstes Modell als beste Modell ausgewählt wird, können Sie zusätzliche Daten reservieren, um die Leistung des ML-Modells zu validieren. Sie können Ihre Daten beispielsweise in Blöcke von 60 Prozent für Schulungen, 20 Prozent für die Bewertung und weitere 20 Prozent für die Validierung aufteilen. Nachdem Sie die Modellparameter ausgewählt haben, die für die Bewertungsdaten gut funktionieren, führen Sie eine zweite Bewertung mit den Validierungsdaten durch, um zu sehen, wie gut die ML-Modell die Validierungsdaten verarbeitet. Wenn das Modell Ihren Erwartungen hinsichtlich der Validierungsdaten entspricht, passt es die Daten nicht übermäßig an.

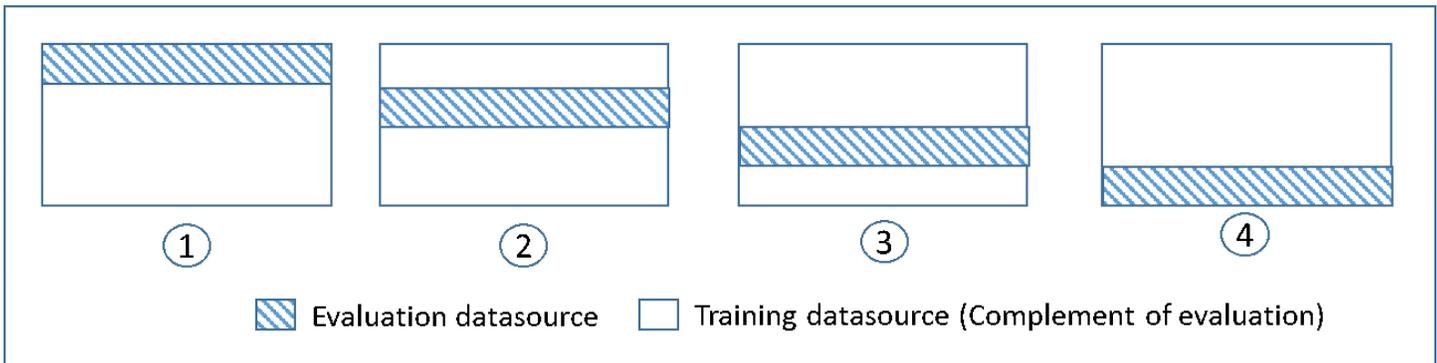
Unter Verwendung einer dritten Gruppe von Daten für die Validierung können Sie geeignete ML-Modellparameter wählen, um eine Überanpassung zu verhindern. Durch das Zurückhalten von Daten für Evaluation und Validierung stehen jedoch weniger Daten für die Schulung zur Verfügung. Dies ist besonders bei geringen Datenmengen ein Problem, da stets so viele Daten wie möglich für die Schulung verwendet werden sollten. Um dieses Problem zu lösen, können Sie eine Cross-Validierung durchführen. Weitere Informationen zu Cross-Validierungen finden Sie unter [Kreuzvalidierung](#).

Kreuzvalidierung

Die Kreuzvalidierung ist eine Methode zur Bewertung von ML-Modellen, indem Sie mehrere ML-Modelle mit Teilsätzen der verfügbaren Eingabedaten schulen und diese in der ergänzenden Teilmenge der Daten auswerten. Verwenden Sie die Kreuzvalidierung, um eine Überanpassung zu erkennen, d. h. wenn ein Muster nicht verallgemeinert wird.

In Amazon ML können Sie die Methode der k-fachen Kreuzvalidierung verwenden, um die Kreuzvalidierung durchzuführen. Bei der k-fachen Kreuzvalidierung teilen Sie die Eingabedaten in k-Datenteilmengen auf. Ein ML-Modell wird, bis auf eine Teilmenge (k-1), auf alle Teilmengen geschult. Das Modell wird dann anhand der Teilmenge ausgewertet, die nicht für die Schulung verwendet wurde. Dieser Vorgang wird k-Mal wiederholt, wobei jedes Mal eine andere Teilmenge für die Auswertung verwendet wird (die für Schulungen ausgeschlossen ist).

Das folgende Diagramm zeigt ein Beispiel der Schulungsteilmengen und der ergänzenden Auswertungsteilmengen, die für jedes der vier Modelle generiert werden, die während einer 4-fachen Kreuzvalidierung erstellt und geschult werden. Modell 1 verwendet die ersten 25 Prozent der Daten für die Auswertung und die verbleibenden 75 Prozent für die Schulung. Modell 2 verwendet die zweite Teilmenge von 25 Prozent (25 Prozent bis 50 Prozent) für die Auswertung und die verbleibenden drei Teilmengen der Daten für Schulungen und so weiter.



Jedes Modell wird mithilfe der ergänzenden Datenquellen geschult und ausgewertet – die Daten in der Auswertungsdatenquelle umfassen und sind auf alle Daten beschränkt, die sich nicht in der Schulungsdatenquelle befinden. Sie erstellen die Datenquellen für jede dieser Untergruppen mit dem Parameter `DataRearrangement` in den `createDataSourceFromS3`-, `createDataSourceFromRedShift`- und `createDataSourceFromRDS`-APIs. Geben Sie im Parameter `DataRearrangement` an, welche Teilmenge von Daten in eine Datenquelle eingeschlossen werden soll, indem Sie angeben, wo jedes Segment angefangen und beendet werden soll. Um die ergänzenden Datenquellen zu erstellen, die für eine 4000-fache Kreuzvalidierung erforderlich sind, geben Sie den Parameter `DataRearrangement` wie im folgenden Beispiel an:

Modell 1:

Auswertungsdatenquelle:

```
{"splitting":{"percentBegin":0, "percentEnd":25}}
```

Schulungsdatenquelle:

```
{"splitting":{"percentBegin":0, "percentEnd":25, "complement":"true"}}
```

Modell 2:

Auswertungsdatenquelle:

```
{"splitting":{"percentBegin":25, "percentEnd":50}}
```

Schulungsdatenquelle:

```
{"splitting":{"percentBegin":25, "percentEnd":50, "complement":"true"}}
```

Modell 3:

Auswertungsdatenquelle:

```
{"splitting":{"percentBegin":50, "percentEnd":75}}
```

Schulungsdatenquelle:

```
{"splitting":{"percentBegin":50, "percentEnd":75, "complement":"true"}}
```

Modell 4:

Auswertungsdatenquelle:

```
{"splitting":{"percentBegin":75, "percentEnd":100}}
```

Schulungsdatenquelle:

```
{"splitting":{"percentBegin":75, "percentEnd":100, "complement":"true"}}
```

Durch eine 4-fache Kreuzvalidierung werden vier Modelle, vier Datenquellen zum Schulen der Modelle, vier Datenquellen zum Auswerten der Modelle und vier Auswertungen (eine für jedes Modell) generiert. Amazon ML generiert für jede Auswertung eine Leistungsmetrik für das Modell. Bei einer 4-fachen Kreuzvalidierung für ein binäres Klassifizierungsproblem erstellt jede der Auswertungen einen Bericht für eine sogenannte Area Under a Curve (AUC). Die Gesamtleistung erhalten Sie, indem Sie den Durchschnitt der vier AUC-Metriken berechnen. Weitere Informationen zur AUC-Metrik finden Sie unter [Messung der ML-Modellgenauigkeit](#).

Beispielcode, der veranschaulicht, wie eine Kreuzvalidierung erstellt wird und der Durchschnitt der Modelle durchgeführt wird, finden Sie im [Amazon ML-Beispiel-Code](#) aus.

Anpassen Ihrer Modelle

Nachdem Sie die Modelle kreuzvalidiert haben, können Sie die Einstellungen für das nächste Modell anpassen, wenn das Modell nicht Ihren Standards entspricht. Weitere Informationen zur Überanpassung finden Sie unter [Modellanpassung: Unteranpassung vs. Overfitting](#). Weitere Informationen zur Regularisation finden Sie unter [Regularisation](#). Weitere Informationen zum Ändern

der Regularisationseinstellungen finden Sie unter [Erstellen eines ML-Modells mit benutzerdefinierten Optionen](#).

Auswertungswarnungen

Amazon ML bietet Einblicke, damit Sie prüfen können, ob Sie das Modell korrekt ausgewertet haben. Wenn eines der Validierungskriterien nicht von der Auswertung erfüllt wurde, gibt die Amazon ML-Konsole eine Warnung aus, aus der Sie wie nachfolgend erklärt erkennen können, in welcher Weise gegen die Kriterien verstoßen wurde.

- Auswertung des ML-Modells erfolgt mit zurückgehaltenen Daten

Amazon ML benachrichtigt Sie, wenn Sie für die Schulung und Auswertung dieselbe Datenquelle verwenden. Wenn Sie Amazon ML verwenden, um Ihre Daten aufzuteilen, erfüllen Sie dieses Validierungskriterium. Wenn Sie Amazon ML nicht verwenden, um Ihre Daten aufzuteilen, stellen Sie sicher, dass Sie Ihr ML-Modell mit einer anderen Datenquelle als der Schulungsdatenquelle bewerten.

- Genügend Daten wurden für die Auswertung des Voraussagemodells verwendet

Amazon ML warnt Sie, wenn die Anzahl der Beobachtungen/Datensätze in den Auswertungsdaten weniger als 10% der Anzahl an Beobachtungen in Ihrer Schulungsdatenquelle beträgt. Um Ihr Modell korrekt zu bewerten, ist es wichtig, ein ausreichend großes Datenbeispiel bereitzustellen. Mit diesem Kriterium wird überprüft, ob Sie genügend Daten verwenden. Die für die Auswertung Ihres ML-Modells notwendige Datenmenge ist subjektiv. 10% ist hier als vorläufiger Richtwert eingestellt.

- Schema abgeglichen

Amazon ML warnt Sie, wenn das Schema für die Schulung und die Auswertung der Datenquelle nicht gleich sind. Wenn Sie über bestimmte Attribute verfügen, die nicht in der Datenquelle vorhanden sind, oder zusätzliche Attribute besitzen, zeigt Amazon ML diese Warnung an.

- Alle Datensätze aus Auswertungsdateien wurden für eine vorausschauende Modelleistungsbewertung verwendet

Es ist wichtig zu wissen, ob alle für die Auswertung bereitgestellten Datensätze tatsächlich für die Auswertung des Modells verwendet wurden. Amazon ML warnt Sie, wenn einige Datensätze in der Auswertungsdatenquelle ungültig waren und nicht für die genaue metrische Berechnung verwendet wurden. Wenn beispielsweise die Zielvariable für einige der Beobachtungen in der Auswertungsdatenquelle fehlt, kann Amazon ML nicht prüfen, ob die Voraussagen des ML-Modells

für diese Beobachtungen richtig sind. In diesem Fall beträgt werden die Datensätze mit fehlenden Zielwerten als ungültig betrachtet.

- Verteilung der Zielvariable

Amazon ML zeigt die Verteilung die Verteilung des Zielattributs der Schulungs- und Auswertungsdatenquelle an, damit Sie prüfen können, ob das Ziel in beiden Datenquellen ähnlich verteilt wurde. Wenn das Modell mit einer Schulungsdatenverteilung geschult wurde, die von der Verteilung des Ziels in den Auswertungsdaten abweicht, kann die Qualität der Auswertung leiden, da sie mithilfe von Daten mit sehr unterschiedlichen Statistiken berechnet wurde. Die Daten sollten in den Schulungs- und Auswertungsdaten ähnlich verteilt sein, und diese Datasets sollten so gut wie möglich den Daten des Modells beim Treffen von Voraussagen entsprechen.

Wenn diese Warnung ausgelöst wird, versuchen Sie, mit der zufälligen Verteilungsstrategie die Daten in Schulungs- und Auswertungsdatenquellen aufzuteilen. In seltenen Fällen kann eine Warnung über Verteilungsunterschiede ausgegeben werden, obwohl Sie Ihre Daten zufällig verteilt haben. Amazon ML verwendet ungefähre Datenstatistiken, um die Datenverteilungen auszuwerten, wodurch diese Warnung gelegentlich irrtümlich ausgelöst werden kann.

Generieren und Interpretieren von Voraussagen

Amazon ML bietet zwei Methoden zum Generieren von Voraussagen: asynchrone (Batch-basierte) und synchrone (einzeln).

Verwenden Sie asynchrone Voraussagen oder Stapelvoraussagen, wenn Sie eine Reihe von Beobachtungen haben und Voraussagen für alle Beobachtungen gleichzeitig erhalten möchten. Der Prozess verwendet eine Datenquelle als Eingabe und gibt Voraussagen in eine CSV-Datei aus, die in einem S3-Bucket Ihrer Wahl gespeichert wird. Sie müssen warten, bis der Stapelvoraussageprozess abgeschlossen ist, bevor Sie auf die Voraussageergebnisse zugreifen können. Die maximale Größe einer Datenquelle, die Amazon ML in einer Batch-Datei verarbeiten kann, ist 1 TB (ca. 100 Millionen Datensätze). Wenn Ihre Datenquelle größer als 1 TB ist, wird der Vorgang fehlschlagen und Amazon ML gibt einen Fehlercode zurück. Um dies zu verhindern, teilen Sie Ihre Daten in mehrere Batches auf. Wenn Ihre Datensätze in der Regel länger sind, erreichen Sie den Grenzwert von 1 TB, bevor 100 Millionen Datensätze verarbeitet wurden. In diesem Fall empfehlen wir Ihnen, sich an den [AWS-Support](#) zu wenden, um die Auftragsgröße für Ihre Stapelvoraussage zu erhöhen.

Verwenden Sie synchrone oder Echtzeit-Voraussagen, wenn Sie Voraussagen mit niedriger Latenz erhalten möchten. Die Echtzeit-Voraussage-API akzeptiert eine einzelne Eingabebeobachtung, serialisiert als JSON-Zeichenfolge, gibt die Voraussage und die zugehörigen Metadaten synchron als Teil der API-Antwort zurück. Sie können die API mehrmals gleichzeitig aufrufen, um synchrone Voraussagen parallel zu erhalten. Weitere Informationen zur Durchsatzkapazität der Echtzeit-Voraussage-API finden Sie unter den Echtzeit-Voraussage-Limits in der [Amazon ML-API-Referenz](#).

Themen

- [Erstellen einer Stapelvoraussage](#)
- [Überprüfen von Stapelvoraussage-Metriken](#)
- [Lesen der Ausgangsdateien für Stapelvoraussagen](#)
- [Anfordern von Echtzeitvoraussagen](#)

Erstellen einer Stapelvoraussage

Um eine Batch-Voraussage zu erstellen, erstellen Sie ein `BatchPredictionObject`, das entweder die Amazon Machine Learning (Amazon ML) -Konsole oder die API verwendet. Ein `BatchPrediction`-Objekt beschreibt eine Gruppe von Voraussagen, die Amazon ML

mit Ihrem ML-Modell und einer Reihe eingegebener Beobachtungen erstellt. Wenn Sie eine `createBatchPrediction`-Objekt startet Amazon ML einen asynchronen Workflow, der die Voraussagen berechnet.

Sie müssen für die Datenquelle dasselbe Schema verwenden, das Sie verwendet haben, um Stapelvoraussagen zu erhalten, sowie dieselbe Datenquelle, die Sie verwendet haben, um das ML-Modell, mit dem Sie Voraussagen abfragen, zu schulen. Die einzige Ausnahme ist, dass die Datenquelle für eine Batch-Voraussage nicht das Zielattribut enthalten muss, da Amazon ML das Ziel voraussagt. Wenn Sie das Zielattribut angeben, ignoriert Amazon ML den Wert.

Erstellen einer Stapelvoraussage (Konsole)

Um eine Batch-Voraussage mit der Amazon ML-Konsole zu erstellen, verwenden Sie den Assistenten `CreateBatchPrediction`.

Erstellen einer Stapelvoraussage (Konsole)

1. Melden Sie sich bei der AWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/>.
2. Im Amazon ML-Dashboard unter `Objekte`, wählen `Neue erstellen...` Klicken Sie auf `Neue erstellen...` und danach auf `Stapelprognose`.
3. Wählen Sie die Amazon ML-Modell, das Sie verwenden möchten, um die Batch-Voraussage zu erstellen.
4. Um zu bestätigen, dass Sie dieses Modell verwenden möchten, klicken Sie auf `Fortfahren`.
5. Wählen Sie die Datenquelle, für die Sie Voraussagen erstellen möchten. Die Datenquelle muss dasselbe Schema wie Ihr Modell besitzen, muss allerdings nicht das Zielattribut enthalten.
6. Klicken Sie auf `Continue`.
7. Geben Sie unter `S3-Ziel` den Namen Ihres S3-Buckets ein.
8. Wählen Sie `Review`.
9. Überprüfen Sie die Einstellungen und klicken Sie auf `Stapelvoraussage erstellen`.

Erstellen einer Stapelvoraussage (API)

Erstellen eines `BatchPrediction` Wenn Sie die Amazon ML-API verwenden, müssen Sie die folgenden Parameter angeben:

ID der Datenquelle

Die ID der Datenquelle, die auf die Beobachtungen hinweist, für die Voraussagen wünschen. Wenn Sie beispielsweise Voraussagen für Daten wünschen, die in einer Datei mit dem Namen `s3://examplebucket/input.csv` enthalten sind, erstellen Sie ein Datenquellenobjekt, das auf die Datendatei hinweist, und übergeben Sie dann die ID dieser Datenquelle mit diesem Parameter.

Batch-Voraussage-ID

Die ID zur Zuweisung der Stapelvoraussage.

ML-Modell-ID

Die ID des ML-Modells, das Amazon ML für die Voraussage abfragen soll.

Ausgabe-Uri

Die URI des S3-Buckets, in dem die Ausgabe der Voraussage speichern. Amazon ML muss über die Berechtigung verfügen, Daten in diesen Bucket zu schreiben.

Der Parameter `OutputUri` muss sich auf einen S3-Pfad beziehen, der mit einem Schrägstrich ("/") endet, z. B.:

```
s3://examplebucket/examplepath/
```

Informationen zu S3-Berechtigungen finden Sie unter [Gewähren von Berechtigungen für Amazon ML zwecks Ausgabe von Voraussagen in Amazon S3](#).

(Optional) Batch-Voraussagenname

(Optional) Ein lesbarer Name für Ihre Stapelvoraussage.

Überprüfen von Stapelvoraussage-Metriken

Nachdem Amazon Machine Learning (Amazon ML) eine Batch-Voraussage erstellt hat, werden zwei Metriken angezeigt: `Records seen` und `Records failed to process`. `Records seen` gibt an, wie viele Datensätze sich Amazon ML für die Batch-Voraussage angesehen hat. `Records failed to process` gibt an, wie viele Datensätze Amazon ML nicht verarbeiten konnte.

Um Amazon ML die Verarbeitung von fehlgeschlagenen Datensätze zu ermöglichen, überprüfen Sie die Formatierung der Datensätze in den Daten, die für die Erstellung der Datenquelle verwendet

wurden, und stellen Sie sicher, dass alle erforderlichen Attribute vorhanden und alle Daten korrekt sind. Nachdem Sie Ihre Daten berichtigt haben, können Sie entweder Ihre Stapelvoraussage erneut erstellen oder eine neue Datenquelle mit den fehlgeschlagenen Datensätzen erstellen, um dann eine neue Stapelvoraussage mit der neuen Datenquelle zu erstellen.

Überprüfen von Stapelvoraussage-Metriken (Konsole)

Um die Metriken in der Amazon ML-Konsole anzuzeigen, öffnen Sie die Zusammenfassung der Stapelprognoseseite und schauen Sie im Verarbeitete Info-Abschnitts erstellt.

Überprüfen von Stapelvoraussage-Metriken und -Details (API)

Sie können die Amazon ML-APIs verwenden, um Details über BatchPrediction-Objekte, einschließlich der Datensatzmetriken. Amazon ML bietet die folgenden Batch-Voraussage-API-Aufrufe:

- CreateBatchPrediction
- UpdateBatchPrediction
- DeleteBatchPrediction
- GetBatchPrediction
- DescribeBatchPredictions

Weitere Informationen finden Sie im [.Amazon ML API-Referenz](#)aus.

Lesen der Ausgangsdateien für Stapelvoraussagen

Führen Sie die folgenden Schritte durch, um die Ausgangsdateien von Stapelvoraussagen abzurufen:

1. Suchen Sie die Manifestdatei für Stapelvoraussagen.
2. Lesen Sie die Manifestdatei, um die Speicherorte der Ausgabedateien zu ermitteln.
3. Rufen Sie die Ausgabedateien mit den Voraussagen ab.
4. Interpretieren Sie den Inhalt der Ausgabedateien. Der Inhalt variiert je nach Art des für die Erzeugung von Voraussagen verwendeten ML-Modells.

In den folgenden Abschnitte beschreiben diese Schritte ausführlicher.

Suchen der Manifestdatei für Stapelvoraussagen

Die Manifestdateien der Stapelvoraussagen enthalten Informationen, die Ihre Eingabedateien den Voraussage-Ausgabedateien zuordnen.

Um eine Manifestdatei zu suchen, starten Sie mit dem Ausgabespeicherort, den Sie beim Erstellen des Stapelvoraussageobjekts festgelegt haben. Sie können ein abgeschlossenes Stapelvoraussageobjekt abfragen, um den S3-Speicherort dieser Datei abzurufen [Amazon ML](#) oder das <https://console.aws.amazon.com/machinelearning/aus>.

Die Manifestdatei befindet sich am Ausgabespeicherort unter einem Pfad, der aus der statischen Zeichenfolge `/batch-prediction/` angehängt an den Speicherort und den Namen der Manifestdatei besteht, der wiederum die ID der Stapelvoraussage mit der angehängten Erweiterung `.manifest` ist.

Wenn Sie beispielsweise eine Stapelvoraussageobjekt mit der ID `bp-example` erstellen und den S3-Speicherort `s3://examplebucket/output/` als Ausgabespeicherort angeben, finden Sie Ihre Manifestdatei hier:

```
s3://examplebucket/output/batch-prediction/bp-example.manifest
```

Lesen der Manifestdatei

Der Inhalt der Manifestdatei ist als JSON-Zuweisung codiert, deren Schlüssel eine Zeichenfolge mit dem Namen einer S3-Eingabedatei ist, und der Wert ist eine Zeichenfolge der zugehörigen Stapelvoraussageergebnisdatei. Für jedes Eingabe-/Ausgabedateipaar besteht eine Zuweisungszeile. Wird nun das Beispiel oben weitergeführt und die Eingabe für die Erstellung des `BatchPrediction`-Objekts besteht aus einer einzelnen Datei mit dem Namen `data.csv`, die unter `s3://examplebucket/input/` gespeichert ist, sehen Sie möglicherweise eine Zuweisungszeichenfolge wie diese:

```
{"s3://examplebucket/input/data.csv": "s3://examplebucket/output/batch-prediction/result/bp-example-data.csv.gz"}
```

Wenn die Eingabe für die Erstellung des `BatchPrediction`-Objekts aus drei Dateien mit den Namen `data1.csv`, `data2.csv` und `data3.csv` besteht und alle am S3-Speicherort `s3://examplebucket/input/` gespeichert sind, sehen Sie möglicherweise eine Zuweisungszeichenfolge wie diese:

```
{"s3://examplebucket/input/data1.csv":"s3://examplebucket/output/batch-prediction/
result/bp-example-data1.csv.gz",

"s3://examplebucket/input/data2.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data2.csv.gz",

"s3://examplebucket/input/data3.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data3.csv.gz"}
```

Abrufen der Ausgangsdateien für Stapelvoraussagen

Sie können jede aus der Manifestzuweisung abgerufene Stapelvoraussagedatei herunterladen und lokal verarbeiten. Das Dateiformat ist CSV mit komprimiertem gzip-Algorithmus. In dieser Datei gibt es eine Zeile pro Eingangsbeobachtung in der entsprechenden Eingabedatei.

Um die Voraussagen mit der Eingabedatei der Stapelvoraussage zu verbinden, können Sie eine einfache datensatzweise Zusammenführung der beiden Dateien durchführen. Die Ausgabedatei der Stapelvoraussage enthält immer dieselbe Anzahl Datensätze wie die Voraussage-Eingabedatei, und zwar in derselben Reihenfolge. Wenn eine Eingabebeobachtung bei der Verarbeitung versagt und keine Voraussage erstellt werden kann, enthält die Ausgabedatei der Stapelvoraussage an der entsprechenden Stelle eine Leerzeile.

Den Inhalt von Stapelvoraussagedateien für ein binäres ML-Klassifikationsmodell interpretieren

Die Spalten der Stapelvoraussagedatei für ein binäres Klassifikationsmodell heißen `bestAnswer` und `score`.

Die Spalte `bestAnswer` enthält das Voraussagekennzeichen ("1" oder "0"), das aus der Evaluation der Voraussagepunktzahl im Vergleich zur Grenzwertpunktzahl hervorgeht. Weitere Informationen zu Grenzwertpunktzahlen finden Sie unter [Anpassen des Ergebnisgrenzwerts](#). Sie können eine Grenzwertpunktzahl für das ML-Modell festlegen; verwenden Sie dazu entweder die Amazon ML API oder die Modellbewertungsfunktionalität auf der Amazon ML-Konsole. Wenn Sie keine Grenzwertpunktzahl festlegen, verwendet Amazon ML den Standardwert 0,5.

Die Punktzahl enthält die rohe Voraussagepunktzahl, die das ML-Modell dieser Voraussage zugewiesen hat. Amazon ML nutzt logistische Regressionsmodelle, sodass diese Punktzahl die Wahrscheinlichkeit der Beobachtung zu modellieren versucht, die einem wahren Wert („1“) entspricht.

Beachten Sie, dass `score` in Exponentialschreibweise gemeldet wird, sodass in der ersten Zeile des folgenden Beispiels der Wert `8.7642E-3` gleich `0,0087642` entspricht.

Wenn die Grenzwertpunktzahl für das ML-Modell beispielsweise `0,75` beträgt, kann der Inhalt der Stapelvoraussageausgangsdatei eines binären Klassifikationsmodells folgendermaßen aussehen:

```
bestAnswer,score
0,8.7642E-3
1,7.899012E-1
0,6.323061E-3
0,2.143189E-2
1,8.944209E-1
```

Die zweite und fünfte Beobachtung in der Eingabedatei haben Voraussagepunktzahlen über `0,75` erhalten, sodass die Spalte `bestAnswer` für diese Beobachtungen den Wert `"1"` enthält, während andere Beobachtungen den Wert `"0"` haben.

Den Inhalt von Stapelvoraussagedateien für ein Multiclass-ML-Klassifikationsmodell interpretieren

Die Stapelvoraussagedatei für ein Multiclass-Modell enthält eine Spalte für jede Klasse in den Schulungsdaten. Die Spaltennamen werden in der Kopfzeile der Stapelvoraussagedatei angezeigt.

Wenn Sie Voraussagen aus einem Multiclass-Modell abrufen, berechnet Amazon ML für jede Beobachtung in der Eingabedatei mehrere Voraussagepunktzahlen, eine für jede der im Eingabedatensatz definierten Klassen. Dies entspricht der Frage "Wie hoch ist die Wahrscheinlichkeit (gemessen zwischen 0 und 1), dass diese Beobachtung in diese Klasse und in keine der anderen Klassen fällt?" Jede Punktzahl kann als "Wahrscheinlichkeit, dass die Beobachtung zu dieser Klasse gehört" gelesen werden. Da Voraussagepunktzahlen die zugrunde liegenden Wahrscheinlichkeiten der Beobachtung modellieren, die einer Klasse angehören, beträgt die Summe aller Voraussagepunktzahlen in einer Zeile 1. Sie müssen eine Klasse als vorausgesagte Klasse für das Modell wählen. In den meisten Fällen würden Sie die Klasse mit der höchsten Wahrscheinlichkeit als beste Antwort wählen.

Nehmen wir beispielsweise an, Sie versuchen, die Bewertung eines Kunden für ein bestimmtes Produkt auf Grundlage einer Skala von 1 bis 5 vorzusagen. Wenn die Klassen `1_star`, `2_stars`, `3_stars`, `4_stars` und `5_stars` heißen, sieht die Multiclass-Voraussage-Ausgangsdatei möglicherweise folgendermaßen aus:

```
1_star, 2_stars, 3_stars, 4_stars, 5_stars  
  
8.7642E-3, 2.7195E-1, 4.77781E-1, 1.75411E-1, 6.6094E-2  
  
5.59931E-1, 3.10E-4, 2.48E-4, 1.99871E-1, 2.39640E-1  
  
7.19022E-1, 7.366E-3, 1.95411E-1, 8.78E-4, 7.7323E-2  
  
1.89813E-1, 2.18956E-1, 2.48910E-1, 2.26103E-1, 1.16218E-1  
  
3.129E-3, 8.944209E-1, 3.902E-3, 7.2191E-2, 2.6357E-2
```

In diesem Beispiel hat die erste Beobachtung die höchste Voraussagepunktzahl für die Klasse `3_stars` (Voraussagepunktzahl = $4.77781E-1$), sodass Sie die angezeigten Ergebnisse so interpretieren würden, dass Klasse `3_stars` die beste Antwort auf diese Beobachtung ist. Beachten Sie, dass Voraussagepunktzahlen in Exponentialschreibweise angegeben werden, eine Voraussagepunktzahl von $4.77781E-1$ entspricht also 0,477781.

Unter bestimmten Umständen kann es nützlich sein, wenn Sie nicht die Klasse mit der höchsten Wahrscheinlichkeit wählen. Wenn Sie beispielsweise einen Mindestgrenzwert festlegen möchten, unter dem Sie eine Klasse nicht mehr als beste Antwort erachten, auch wenn sie die höchste Voraussagepunktzahl hat. Nehmen wir an, Sie klassifizieren Filme in Genres, und Sie möchten, dass die Voraussagepunktzahl mindestens $5E-1$ beträgt, bevor Sie das Genre als beste Antwort erachten. Sie erhalten eine Voraussagepunktzahl von $3E-1$ für Komödien, $2.5E-1$ für Dramen, $2.5E-1$ für Dokumentationen und $2E-1$ für Action-Filme. In diesem Fall sagt ML-Modell voraus, das Komödie Ihre wahrscheinlichste Wahl ist, doch Sie entscheiden, dies nicht als beste Antwort zu wählen. Da keine der Voraussagepunktzahlen Ihre Basis-Voraussagepunktzahl von $5E-1$ überschritten hat, entscheiden Sie, dass die Voraussage nicht ausreicht, um das Genre souverän vorherzusagen, und Sie entscheiden sich für etwas anderes. Ihre Anwendung kann das Genre-Feld für diesen Film dann als "unbekannt" festlegen.

Den Inhalt von Stapelvoraussagedateien für ein Regressions-ML-Modell interpretieren

Die Stapelvoraussagedatei für ein Regressionsmodell enthält eine einzelne Spalte mit dem Namen `score`. Diese Spalte enthält die unformatierte numerische Voraussage für jede Beobachtung in den Eingabedaten. Die Werte werden in Exponentialschreibweise angegeben, sodass ein `score`-Wert von `-1.526385E1` in der ersten Reihe im folgenden Beispiel einem Wert von `-15.26835` entspricht.

Dieses Beispiel zeigt eine Ausgabedatei für eine Stapelvoraussage für ein Regressionsmodell:

```
score  
  
-1.526385E1  
  
-6.188034E0  
  
-1.271108E1  
  
-2.200578E1  
  
8.359159E0
```

Anfordern von Echtzeitvoraussagen

Eine Echtzeitvoraussage ist ein synchroner Aufruf an Amazon Machine Learning (Amazon ML). Die Voraussage wird erstellt, wenn Amazon ML die Anforderung erhält und die Antwort sofort zurückgegeben wird. Echtzeitvoraussagen werden häufig verwendet, um Voraussagefunktionen in interaktiven Web-, Mobil- oder Desktopanwendungen zu ermöglichen. Sie können ein mit Amazon ML erstelltes ML-Modell mit der geringen Latenzzeit nach Voraussagen in Echtzeit abfragen, um Voraussagen in Echtzeit zu erhalten. Die `Predict`-API akzeptiert eine einzelne Input-Beobachtung in der Nutzlast der Anforderung und gibt die Voraussage synchron in der Antwort zurück. Dadurch unterscheidet sie sich von der Stapelvoraussage-API, die mit der ID eines Amazon ML-Datenquellenobjekts aufgerufen wird, das auf den Ort der Input-Beobachtungen verweist, und die asynchron einen URI zu einer Datei zurückgibt, die Voraussagen für alle diese Beobachtungen enthält. Amazon ML reagiert innerhalb von 100 Millisekunden auf die meisten Anforderungen für Echtzeitvoraussagen.

Sie können Echtzeitvoraussagen testen, ohne dass Gebühren in der Amazon ML-Konsole entstehen. Wenn Sie dann Echtzeitvoraussagen verwenden möchten, müssen Sie zuerst einen Endpunkt für die

Generierung von Echtzeitvoraussagen erstellen. Sie können dies in der Amazon ML-Konsole tun oder mithilfe der `CreateRealtimeEndpoint` API. Nachdem Sie einen Endpunkt haben, verwenden Sie die API für Echtzeitvoraussagen, um Echtzeitvoraussagen zu generieren.

Note

Nach dem Erstellen eines Echtzeitendpunkts für Ihr Modell fallen Gebühren für die Kapazitätsreservierung an, die auf der Größe des Modells basieren. Weitere Informationen finden Sie unter [- Preise](#). Wenn Sie den Echtzeitendpunkt in der Konsole erstellen, zeigt die Konsole eine Aufschlüsselung der geschätzten Kosten an, die sich für den Endpunkt fortlaufend anfallen werden. Damit Ihnen keine Kosten mehr berechnet werden, wenn Sie keine Echtzeitvoraussagen von diesem Modell mehr benötigen, entfernen Sie den Echtzeitendpunkt mithilfe der Konsole oder der `DeleteRealtimeEndpoint`-Operation.

Beispiele für Beispiele für `Predict` Anforderungen und Antworten, siehe [Voraussagen](#) im Amazon Machine Learning Learning-API-Referenz aus. Ein Beispiel für das exakte Antwortformat Ihres Modells finden Sie unter [Testen von Echtzeitvoraussagen](#).

Themen

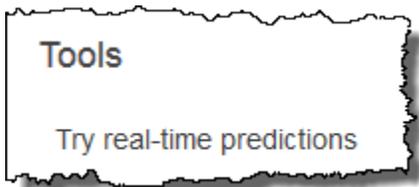
- [Testen von Echtzeitvoraussagen](#)
- [Erstellen eines Echtzeitendpunkts](#)
- [Auffinden des Endpunkts für Echtzeitvoraussagen \(Konsole\)](#)
- [Auffinden des Endpunkts für Echtzeitvoraussagen \(API\)](#)
- [Erstellen einer Echtzeitvoraussage-Anforderung](#)
- [Löschen eines Echtzeitendpunkts](#)

Testen von Echtzeitvoraussagen

Um Ihnen die Entscheidung zu erleichtern, ob Sie Echtzeitvoraussagen nutzen möchten oder nicht, ermöglicht Amazon ML Ihnen, die Generierung von Voraussagen für einzelne Datensätze zu testen, ohne dass die zusätzlichen Gebühren in Verbindung mit der Einrichtung eines Endpunkts für Echtzeitvoraussagen anfallen. Um Echtzeitvoraussagen zu testen, müssen Sie über ein ML-Modell verfügen. Um Echtzeitvoraussagen im größeren Maßstab zu erstellen, verwenden Sie die Option [Voraussagen](#) API im Amazon Machine Learning Learning-API-Referenz aus.

Vorgehensweise zum Testen von Echtzeitvoraussagen

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon Machine Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/>.
2. Wählen Sie im Dropdownmenü Amazon Machine Learning in der Navigationsleiste die Option ML-Modelle aus.
3. Wählen Sie das Modell aus, das Sie zum Testen von Echtzeitvoraussagen verwenden möchten, beispielsweise das `Subscription propensity model` aus dem Tutorial.
4. Wählen Sie auf der ML-Modell-Berichtsseite unter Voraussagen die Option Übersicht und dann Try real-time predictions (Echtzeitvoraussagen ausprobieren) aus.



Amazon ML zeigt eine Liste der Variablen, aus denen die Datensätze bestehen, die Amazon ML für die Schulung Ihres Modells verwendet hat.

5. Sie können fortfahren, indem Sie Daten in die einzelnen Felder im Formular eingeben oder einen einzelnen Datensatz im CSV-Format in das Textfeld einfügen.

Geben Sie bei Verwendung des Formulars in jedes Wert-Feld die Daten ein, die Sie zum Testen Ihrer Echtzeitvoraussagen verwenden möchten. Wenn der Datensatz, den Sie eingeben, keine Werte für ein oder mehrere Datenattribut(e) enthält, lassen Sie die Eingabefelder leer.

Wählen Sie die Option Paste a record, um einen Datensatz bereitzustellen. Fügen Sie eine einzelne Datenzeile im CSV-Format in das Textfeld ein und klicken Sie auf Absenden aus.

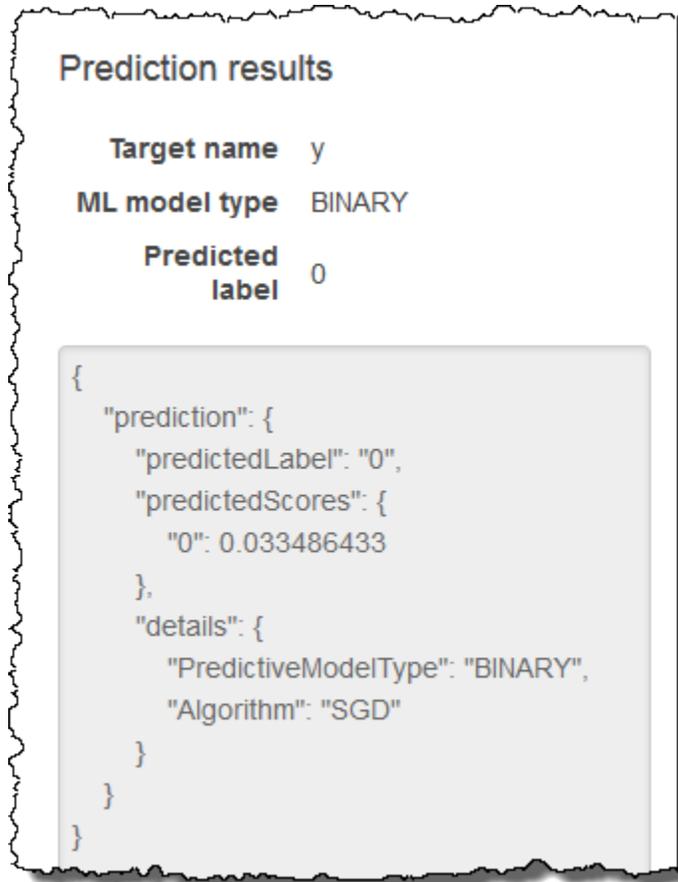
Amazon ML füllt die automatisch aus Value Felder für dich.

Note

Die Daten im Datensatz müssen die gleiche Anzahl von Spalten wie die Schulungsdaten haben und in der gleichen Reihenfolge angeordnet sein. Die einzige Ausnahme ist, dass Sie den Zielwert auslassen sollten. Wenn Sie einen Zielwert angeben, wird dieser von Amazon ML ignoriert.

6. Klicken Sie unten auf der Seite auf Create prediction. Amazon ML gibt die Voraussage sofort zurück.

Im Bereich Prediction results (Voraussageergebnisse) sehen Sie das Voraussageobjekt, das vom Predict-API-Aufruf zurückgegeben wird, sowie den ML-Modelltyp, den Namen der Zielvariable und die vorausgesagte Klasse oder den vorausgesagten Wert. Weitere Informationen zur Interpretation dieser Ergebnisse finden Sie unter [Den Inhalt von Stapelvoraussagedateien für ein binäres ML-Klassifikationsmodell interpretieren](#).



Erstellen eines Echtzeitendpunkts

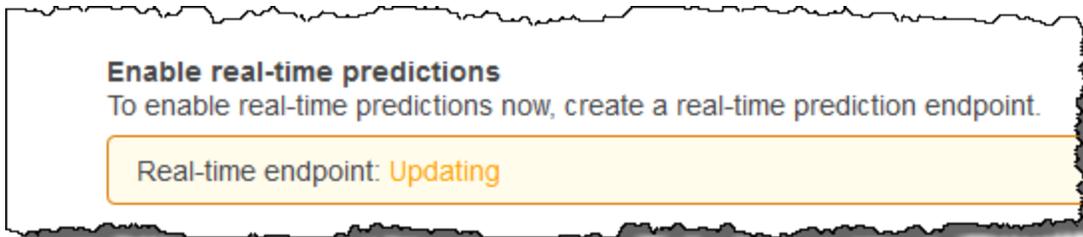
Um Echtzeitvoraussagen generieren zu können, müssen Sie einen Echtzeitendpunkt erstellen. Für die Erstellung eines Echtzeitendpunkt müssen Sie bereits über ein ML-Modell verfügen, für das Sie Echtzeitvoraussagen generieren möchten. Sie können einen Echtzeitendpunkt mithilfe der Amazon ML-Konsole oder durch Aufruf der `CreateRealtimeEndpointAPI`. Weitere Informationen zur Verwendung der finden Sie unter `CreateRealtimeEndpointAPI`, siehe https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_CreateRealtimeEndpoint.html in der Amazon Machine Learning Learning-API-Referenz.

Vorgehensweise zum Erstellen eines Echtzeitendpunkts

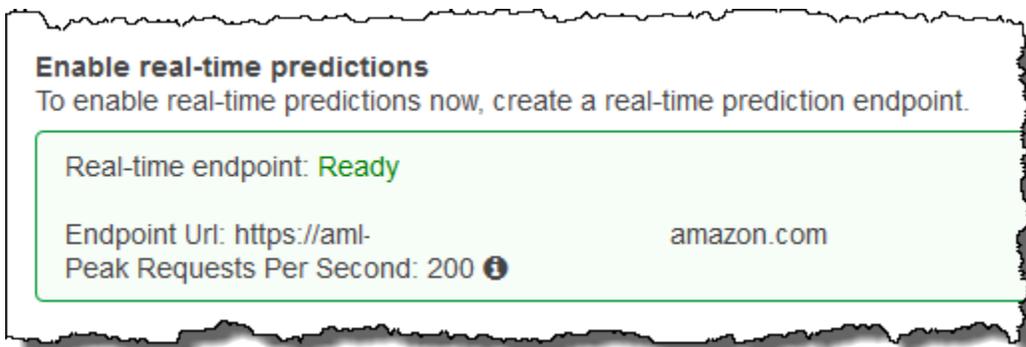
1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon Machine Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/> aus.
2. Wählen Sie im Dropdownmenü Amazon Machine Learning in der Navigationsleiste die Option ML-Modelle aus.
3. Wählen Sie das Modell aus, für das Sie die Echtzeitvoraussagen generieren möchten.
4. Wählen Sie auf der Seite ML model summary unter Predictions die Option Create real-time endpoint aus.

Es wird ein Dialogfeld mit Erläuterungen zur Preisberechnung für Echtzeitvoraussagen angezeigt.

5. Wählen Sie Create (Erstellen) aus. Die Echtzeitendpunkt-Anforderung wird an Amazon ML gesendet und in eine Warteschlange gestellt. Der Status des Echtzeitendpunkts wird aktualisiert.



6. Wenn der Echtzeitendpunkt bereit ist, ändert sich der Status in Bereit, und Amazon ML zeigt die Endpunkt-URL an. Verwenden Sie die Endpunkt-URL, um Echtzeitvoraussagen mit der Predict-API zu erstellen. Weitere Informationen zur Verwendung der Predict-API finden Sie unter Predict-API, siehe https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html in der Amazon Machine Learning API-Referenz.



Auffinden des Endpunkts für Echtzeitvoraussagen (Konsole)

Wenn Sie die Amazon ML-Konsole zum Suchen der Endpunkt-URL für ein ML-Modell verwenden möchten, navigieren Sie zu den Zusammenfassung des ML-Modellsangezeigten.

Vorgehensweise zum Finden einer Echtzeitendpunkt-URL

1. Melden Sie sich beim an AWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/> aus.
2. Wählen Sie im Dropdownmenü Amazon Machine Learning in der Navigationsleiste die Option ML-Modelle aus.
3. Wählen Sie das Modell aus, für das Sie die Echtzeitvoraussagen generieren möchten.
4. Scrollen Sie auf der Seite ML model summary nach unten, bis Sie den Abschnitt Predictions sehen.
5. Die Endpunkt-URL für das Modell wird unter Real-time prediction angezeigt. Verwenden Sie die URL als Endpoint Url-URL für Ihre Echtzeitvoraussage-Aufrufe. Informationen zur Verwendung des Endpunkts zur Generierung von Voraussagen finden Sie unter https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html in der Amazon Machine Learning Learning-API-Referenz.

Auffinden des Endpunkts für Echtzeitvoraussagen (API)

Wenn Sie mithilfe der `CreateRealtimeEndpoint`-Operation einen Echtzeitendpunkt erstellen, werden Ihnen in der Antwort die URL und der Status des Endpunkts mitgeteilt. Wenn Sie einen Echtzeitendpunkt mithilfe der Konsole erstellt haben oder die URL und den Status eines zuvor erstellten Endpunkts abrufen möchten, rufen Sie die `GetMLModel`-Operation mit der ID des Modells auf, das Sie für Echtzeitvoraussagen abfragen möchten. Die Endpunktinformationen sind im `EndpointInfo`-Abschnitt der Antwort enthalten. Bei einem Modell, dem ein Echtzeitendpunkt zugeordnet ist, können die `EndpointInfo` wie folgt aussehen:

```
"EndpointInfo":{
  "CreatedAt": 1427864874.227,
  "EndpointStatus": "READY",
  "EndpointUrl": "https://endpointUrl",
  "PeakRequestsPerSecond": 200
}
```

Ein Modell ohne Echtzeitendpunkt würde zu folgender Antwort führen:

```
EndpointInfo":{
  "EndpointStatus": "NONE",
  "PeakRequestsPerSecond": 0
}
```

Erstellen einer Echtzeitvoraussage-Anforderung

Eine Beispielnutzlast für eine Predict-Anforderung kann wie folgt aussehen:

```
{
  "MLModelId": "model-id",
  "Record":{
    "key1": "value1",
    "key2": "value2"
  },
  "PredictEndpoint": "https://endpointUrl"
}
```

Das `PredictEndpoint`-Feld muss mit der übereinstimmenden `EndpointUrl`-Feld der `EndpointInfo`-Struktur. Amazon ML verwendet dieses Feld, um die Anforderung an die entsprechenden Server in der Echtzeitprognose-Flotte zu leiten.

Das `MLModelId` ist die ID eines zuvor geschulten Modells mit einem Echtzeitendpunkt.

Ein `Record` ist eine Zuordnung von Variablennamen zu Variablenwerten. Jedes Paar stellt eine Beobachtung dar. Die `Record`-Zuordnung enthält die Inputs für Ihr Amazon ML-Modell. Sie ist vergleichbar mit einer einzelnen Datenzeile in Ihrem Schulungsdatensatz, ohne die Zielvariable. Unabhängig von der Art der Werte in den Schulungsdaten enthält `Record` eine Zuordnung von Zeichenfolgen.

Note

Sie können Variablen auslassen, für die Sie keinen Wert haben. Das kann allerdings die Genauigkeit Ihrer Voraussage reduzieren. Je mehr Variablen Sie integrieren können, umso präziser wird Ihr Modell.

Das Format der Antwort von Predict-Anforderungen hängt von der Art des Modells ab, das für die Voraussage abgefragt wird. Das `details`-Feld enthält in jedem Fall Informationen über die Voraussage-Anforderung, darunter insbesondere das `PredictiveModelType`-Feld mit dem Modelltyp.

Das folgende Beispiel zeigt eine Antwort für eine Binärmodell:

```
{
  "Prediction":{
    "details":{
      "PredictiveModelType": "BINARY"
    },
    "predictedLabel": "0",
    "predictedScores":{
      "0": 0.47380468249320984
    }
  }
}
```

Beachten Sie die `predictedLabel`-Feld, das die Voraussagekennzeichnung enthält, in diesem Fall 0. Amazon ML berechnet die Voraussagekennzeichnung, indem die Voraussagepunktzahl mit dem Klassifizierungsgrenzwert verglichen wird:

- Den Klassifizierungsgrenzwert, der aktuell mit einem Modell verbunden ist, finden Sie im `FieldScoreThreshold`-Feld in der Antwort der `GetMLModel`-Betrieb oder durch Anzeigen der Modellinformationen in der Amazon ML-Konsole. Wenn Sie keinen Schwellenwert festlegen, verwendet Amazon ML den Standardwert 0,5.
- Die genaue Voraussagepunktzahl für ein binäres Klassifizierungsmodell erhalten Sie, indem Sie die `predictedScores`-Zuordnung überprüfen. Innerhalb dieser Zuordnung ist die Voraussagekennzeichnung mit der genauen Voraussagepunktzahl verknüpft.

Weitere Informationen zu Binärvoraussagen finden Sie unter [Interpretieren der Voraussagen](#).

Das folgende Beispiel zeigt eine Antwort für ein Regressionsmodell. Beachten Sie, dass der numerische Voraussagewert sich im Feld `predictedValue` befindet:

```
{
  "Prediction":{
    "details":{
```

```
    "PredictiveModelType": "REGRESSION"
  },
  "predictedValue": 15.508452415466309
}
}
```

Das folgende Beispiel zeigt eine Antwort für ein Multiclass-Modell:

```
{
  "Prediction":{
    "details":{
      "PredictiveModelType": "MULTICLASS"
    },
    "predictedLabel": "red",
    "predictedScores":{
      "red": 0.12923571467399597,
      "green": 0.08416014909744263,
      "orange": 0.22713537514209747,
      "blue": 0.1438363939523697,
      "pink": 0.184102863073349,
      "violet": 0.12816807627677917,
      "brown": 0.10336143523454666
    }
  }
}
```

Wie bei binären Klassifizierungsmodellen befindet sich die Voraussagekennzeichnung/-klasse im Feld `predictedLabel`. Die starke Verbindung zwischen der Voraussage und den einzelnen Klassen können Sie besser nachvollziehen, wenn Sie sich die `predictedScores`-Zuordnung ansehen. Je höher die Punktzahl ist, umso stärker ist die Voraussage mit der Klasse verbunden. Der höchste Wert wird schlussendlich als `predictedLabel` ausgewählt.

Weitere Informationen zu Multiclass-Voraussagen finden Sie unter [Einblicke in Mehrklassen-Modelle](#).

Löschen eines Echtzeitendpunkts

Wenn Sie Ihre Echtzeitvoraussagen abgeschlossen haben, löschen Sie den Echtzeitendpunkt, damit keine weiteren Gebühren anfallen. Sobald Sie den Endpunkt gelöscht haben, fallen keine Gebühren mehr an.

Vorgehensweise zum Löschen eines Echtzeitendpunkts

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon Machine Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/>.
2. Wählen Sie im Dropdownmenü Amazon Machine Learning in der Navigationsleiste die Option ML-Modelle aus.
3. Wählen Sie das Modell aus, für das keine Echtzeitvoraussagen mehr erforderlich sind.
4. Wählen Sie auf der ML-Modell-Berichtsseite unter Predictions die Option Summary aus.
5. Wählen Sie Delete real-time endpoint.
6. Klicken Sie im Dialogfeld Delete real-time endpoint auf Delete.

Verwalten von Amazon ML -Objekten

Amazon ML bietet vier Objekte, die Sie über die Amazon ML -Konsole oder die Amazon ML -API verwalten können:

- Datenquellen
- ML-Modelle
- Auswertungen
- Stapelvoraussagen

Jedes Objekt dient einem anderen Zweck im Lebenszyklus zum Erstellen einer maschinellen Lernanwendung, und jedes Objekt hat bestimmte Attribute und Funktionen, die nur für das jeweilige Objekt gelten. Trotz dieser Unterschiede verwalten Sie die Objekte auf ähnliche Weise. Beispielsweise verwenden Sie fast identische Prozesse zum Auflisten von Objekten, Abrufen von Beschreibungen und zum Aktualisieren oder Löschen.

In den folgenden Abschnitten werden die Verwaltungsvorgänge beschreiben, die für alle vier Objekte gelten, und Unterschiede hervorgehoben.

Themen

- [Auflisten von Objekten](#)
- [Das Abrufen von Objektbeschreibungen](#)
- [Aktualisieren von Objekten](#)
- [Löschen von Objekten](#)

Auflisten von Objekten

Listen Sie Ihre Amazon Machine Learning (Amazon ML) -Datenquellen, ML-Modelle, Evaluierungen und Stapelvoraussagen auf, wenn Sie detaillierte Informationen darüber erhalten möchten. Für jedes Objekt wird Ihnen der Name, der Typ, die ID, der Statuscode und der Zeitpunkt der Erstellung angezeigt. Sie können auch spezifische Details für einen bestimmten Objekttyp sehen. Sie können zum Beispiel die Dateneinblicke für eine Datenquelle sehen.

Auflisten von Objekten (Konsole)

Um eine Liste der letzten 1 000 Objekte zu sehen, die Sie erstellt haben, öffnen Sie in der Amazon ML-Konsole die Objekte-Dashboard. So zeigen Sie den an Objekte Melden Sie sich bei der Amazon ML-Konsole an.

Objects ?

Create new... Actions Refresh

Filter: All types Items per page: 10 << 1 - 5 of 5 Objects >>

Name	Type	ID	Status	Creation time	Completion time
<input type="checkbox"/> ▶ Evaluation: ML m...	Evaluation	ev-	Completed	Aug 1, 2016 12:44:48 PM	3 mins.
<input type="checkbox"/> ▶ ML model: Examl...	ML model	ml-	Completed	Aug 1, 2016 12:44:47 PM	2 mins.
<input type="checkbox"/> ▶ Example Datasour...	Datasource	ds-	Completed	Aug 1, 2016 12:44:46 PM	3 mins.
<input type="checkbox"/> ▶ Example Datasour...	Datasource	ds-	Completed	Aug 1, 2016 12:44:46 PM	4 mins.
<input type="checkbox"/> ▶ Example Datasour...	Datasource	ds-	Completed	Aug 1, 2016 12:44:23 PM	3 mins.

Um weitere Details zu einem Objekt sehen zu können, einschließlich spezifischer Details für diesen Objekttyp, wählen Sie den Namen oder die ID des Objekts aus. Wenn Sie beispielsweise Data insights für eine Datenquelle anzeigen möchten, wählen Sie den Namen der Datenquelle aus.

Die Spalten auf dem Dashboard Objekte zeigen die folgenden Informationen zu den einzelnen Objekten.

Name

Der Name des Objekts.

Typ

Der Typ des Objekts. Gültige Werte sind unter anderem Datenquelle, ML-Modell, Auswertung und Stapelvoraussage.

Note

Um zu sehen, ob ein Modell zur Unterstützung von Echtzeitvoraussagen eingerichtet ist, wechseln Sie zur Seite ML model summary (ML-Modellzusammenfassung), indem Sie den Namen oder die Modell-ID auswählen.

ID (ID)

Die ID des Objekts.

Status

Der Status des Objekts. Gültige Werte sind Schwebend, In Bearbeitung, Abgeschlossen und Fehlgeschlagen. Wenn der Status Fehlgeschlagen ist, prüfen Sie Ihre Daten und versuchen Sie es erneut.

Zeitpunkt der Erstellung

Das Datum und die Uhrzeit des Abschlusses der Erstellung dieses Objekt durch Amazon ML.

Fertigstellungszeit

Die Zeit, die Amazon ML für die Erstellung dieses Objekts für die Erstellung dieses Objekts durch Amazon ML brauchte. Sie können die Fertigstellungszeit eines Modells nutzen, um die Schulungszeit für ein neues Modell einzuschätzen.

ID der Datenquelle

Die ID der Datenquelle für Objekte, die unter Verwendung einer Datenquelle erstellt wurde, wie z. B. Modelle und Evaluierungen. Wenn Sie die Datenquelle löschen, können Sie ML-Modelle, die mit dieser Datenquelle erstellt wurden, nicht mehr für die Erstellung von Voraussagen verwenden.

Sie können nach jeder beliebigen Spalte sortieren, indem Sie das doppelte Dreieckssymbol neben der Spaltenüberschrift anklicken.

Auflisten von Objekten (API)

In der [Amazon ML -API](#) mithilfe der folgenden Operationen können Sie Objekte mithilfe der folgenden Operationen nach Typ auflisten:

- `DescribeDataSources`
- `DescribeMLModels`
- `DescribeEvaluations`
- `DescribeBatchPredictions`

Jede Operation beinhaltet Parameter für Filterung, Sortierung und Paginierung durch eine lange Liste von Objekten. Es gibt keine Beschränkungen in Bezug auf die Anzahl an Objekten, auf die Sie über die API zugreifen können. Verwenden Sie den `Limit`-Parameter, um die Größe der Liste zu begrenzen. Der Maximalwert für diesen Parameter ist 100.

Die API-Antwort auf einen `Describe*`-Befehl umfasst ein Paginierungs-Token (`nextPageToken`), sofern zutreffend, sowie kurze Beschreibungen der einzelnen Objekte. Die Objektbeschreibungen enthalten die gleichen Informationen für alle Objekttypen, die in der Konsole angezeigt werden, einschließlich spezifischer Details für einen bestimmten Objekttyp.

Note

Auch wenn die Antwort weniger Objekte als der angegebene Grenzwert umfasst, kann sie ein `nextPageToken` enthalten, das angibt, dass weitere Ergebnisse verfügbar sind. Auch eine Antwort, die 0 Elemente enthält, kann ein `nextPageToken` enthalten.

Weitere Informationen finden Sie im [.Amazon ML API-Referenz](#)aus.

Das Abrufen von Objektbeschreibungen

Sie können detaillierte Beschreibungen eines Objekts über die Konsole oder über die API anzeigen.

Detaillierte Beschreibungen in der Konsole

Um Beschreibungen auf der Konsole anzuzeigen, navigieren Sie zu einem bestimmten Objekttyp (Datenquelle, ML-Modell, Bewertung oder Stapelvoraussage). Als Nächstes suchen Sie entweder mithilfe der Liste oder durch Suchen des Namens oder der ID die Zeile in der Tabelle, die dem Objekt entspricht.

Detaillierte Beschreibungen von der API

Jeder Objekttyp verfügt über eine Funktion, welche die vollständigen Details eines Amazon ML-Objekts abrufen:

- `GetDataSource`
- `GetMLModel`
- `GetEvaluation`
- `GetBatchPrediction`

Jeder Vorgang dauert genau zwei Parameter: die Objekt-ID und ein boolesches Flag mit dem Namen `Verbose`. Abrufe mit `Verbose` auf `"true"` enthalten zusätzliche Details über das Objekt, was zu höherer

Latenzen und umfassenderen Antworten führt. Weitere Informationen zu den Feldern, die durch Setzen des Verbose Flag enthalten sind, finden Sie unter [Amazon ML-API-Referenz](#).

Aktualisieren von Objekten

Jeder Objekttyp verfügt über eine Funktion, welche die Details eines Amazon ML-Objekts aktualisiert (weitere Informationen finden Sie unter [Amazon ML-API-Referenz](#)):

- UpdateDataSource
- UpdateMLModel
- UpdateEvaluation
- UpdateBatchPrediction

Jeder Vorgang erfordert die Objekt-ID, um anzugeben, welches Objekt aktualisiert wird. Sie können die Namen aller Objekte aktualisieren. Sie können jedoch keine anderen Eigenschaften von Objekten für Datenquellen, Auswertungen und Stapelvoraussagen aktualisieren. Für ML-Modelle können Sie das Feld "ScoreThreshold" aktualisieren, solange das ML-Modell nicht über einen zugeordneten Echtzeitvoraussage-Endpunkt verfügt.

Löschen von Objekten

Wenn Sie Ihre Datenquellen, ML-Modelle, Evaluierungen und Stapelvoraussagen nicht mehr benötigen, können Sie sie löschen. Es fallen zwar keine zusätzlichen Kosten für das Behalten nicht mehr benötigter Amazon ML-Objekte an, die keine Stapelvoraussagen sind, doch das Löschen von Objekten sorgt für einen ordentlicher und einfacher zu verwaltenden Workspace. Sie können einzelne oder mehrere Objekte mithilfe der Amazon Machine Learning (Amazon ML) -Konsole oder der API löschen.

Warning

Beim Löschen von Amazon ML-Objekten, sind die Auswirkungen unmittelbar, dauerhaft und unumkehrbar.

Objects ?

Create new... Actions Refresh

Filter: All types Items per page: 10 << < 1 - 5 of 5 Objects > >>

Name	Type	ID	Status	Creation time	Completion time
<input type="checkbox"/> Evaluation: ML m...	Evaluation	ev-	Completed	Aug 1, 2016 12:44:48 PM	3 mins.
<input type="checkbox"/> ML model: Examl...	ML model	ml-	Completed	Aug 1, 2016 12:44:47 PM	2 mins.
<input type="checkbox"/> Example Datasour...	Datasource	ds-	Completed	Aug 1, 2016 12:44:46 PM	3 mins.
<input type="checkbox"/> Example Datasour...	Datasource	ds-	Completed	Aug 1, 2016 12:44:46 PM	4 mins.
<input type="checkbox"/> Example Datasour...	Datasource	ds-	Completed	Aug 1, 2016 12:44:23 PM	3 mins.

Löschen von Objekten (Konsole)

Zum Löschen von Objekten, einschließlich Modellen, können Sie die Amazon ML-Konsole verwenden. Die Vorgehensweise zum Löschen eines Modells hängt davon ab, ob Sie das Modell für die Generierung von Echtzeitvoraussagen verwenden oder nicht. Wenn Sie ein Modell löschen möchten, das zum Generieren von Echtzeitvoraussagen verwendet wird, müssen Sie zuerst den Echtzeitendpunkt löschen.

Vorgehensweise zum Löschen von Amazon ML-Objekten (Konsole)

1. Melden Sie sich beim anAWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/> aus.
2. Wählen Sie das zu löschende Amazon ML-Objekt aus. Verwenden Sie die UMSCHALTTASTE, um mehr als ein Objekt auszuwählen. Verwenden Sie die Schaltflächen  oder  um die Auswahl aller ausgewählten Objekte aufzuheben.
3. Klicken Sie bei Actions auf Delete.
4. Klicken Sie im Dialogfeld auf Löschen, um das Modell zu löschen.

Vorgehensweise zum Löschen eines Amazon ML-Modells mit einem Echtzeitendpunkt (Konsole)

1. Melden Sie sich beim anAWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/> aus.
2. Wählen Sie das Modell aus, das Sie löschen möchten.
3. Wählen Sie unter Aktionen die Option Delete real-time endpoint (Echtzeitendpunkt löschen).

4. Wählen Sie Löschen, um den Endpunkt zu löschen.
5. Wählen Sie das Modell erneut aus.
6. Klicken Sie bei Actions auf Delete.
7. Wählen Sie Löschen, um das Modell zu löschen.

Löschen von Objekten (API)

Sie können Amazon ML-Objekte mit den folgenden API-Aufrufen löschen:

- `DeleteDataSource`: Nutzt den Parameter `DataSourceId`.
- `DeleteMLModel`: Nutzt den Parameter `MLModelId`.
- `DeleteEvaluation`: Nutzt den Parameter `EvaluationId`.
- `DeleteBatchPrediction`: Nutzt den Parameter `BatchPredictionId`.

Weitere Informationen finden Sie unter [Amazon Machine Learning API Reference](#).

Überwachen von Amazon ML mit Amazon CloudWatch-Metriken

Amazon ML sendet automatisch Metriken an Amazon CloudWatch, damit Sie Nutzungsstatistiken für Ihre ML-Modelle sammeln und analysieren können. Wenn Sie beispielsweise Stapel- und Echtzeitvoraussagen nachverfolgen möchten, können Sie die PredictCount-Metrik entsprechend der RequestMode-Dimension überwachen. Die Metriken werden automatisch gesammelt und alle fünf Minuten an Amazon CloudWatch gesendet. Sie können diese Metriken mit der Amazon CloudWatch-Konsole, der AWS-CLI oder AWS-SDKs überwachen.

Für Amazon ML-Metriken, die über CloudWatch gemeldet werden, fallen keine Gebühren an. Wenn Sie Warnmeldung für die Metriken einrichten, wird dies zu standardmäßigen [CloudWatch-Gebühren](#) in Rechnung gestellt.

Weitere Informationen finden Sie in der Amazon ML-Liste der Metriken in der [Referenz zu Amazon CloudWatch-Namespaces, -Dimensionen und -Metriken](#) im Amazon CloudWatch-Developer-Handbuch.

Protokollieren von Amazon-ML-API-Aufrufen mit AWS CloudTrail

Amazon Machine Learning (Amazon ML) ist in integriert AWS CloudTrail, einen -Service, der die Aktionen eines Benutzers, einer Rolle oder eines AWS Service in Amazon ML. CloudTrail erfasst alle API-Aufrufe für Amazon ML als Ereignisse. Zu den erfassten Aufrufen gehören Aufrufe von der Amazon ML-Konsole und Codeaufrufe an die Amazon-ML-API-Operationen. Wenn Sie einen Trail erstellen, können Sie die kontinuierliche Bereitstellung von CloudTrail-Ereignissen an einen Amazon S3 S3-Bucket, einschließlich Ereignissen für Amazon ML, aktivieren. Wenn Sie keinen Trail konfigurieren, können Sie die neuesten Ereignisse in der CloudTrail-Konsole trotzdem in Ereignisverlauf anzeigen. Anhand der von CloudTrail gesammelten Informationen können Sie die Anfrage, die an Amazon ML gestellt wurde, die IP-Adresse, von der die Anfrage gestellt wurde, wer die Anfrage gestellt hat, wann sie gestellt wurde und weitere Details ermitteln.

Weitere Informationen über CloudTrail, einschließlich Konfiguration und Aktivierung, finden Sie im [AWS CloudTrail-Benutzerhandbuch](#).

Amazon-ML-Informationen in CloudTrail

CloudTrail wird beim Erstellen Ihres AWS-Kontos für Sie aktiviert. Wenn in Amazon ML auftretenden unterstützten Aktivitäten werden als CloudTrail-Ereignis zusammen mit anderen aufgezeichnet AWS-Service-Ereignisse in Ereignisverlauf des aus. Sie können die neusten Ereignisse in Ihr AWS-Konto herunterladen und dort suchen und anzeigen. Weitere Informationen finden Sie unter [Anzeigen von Ereignissen mit dem CloudTrail-Ereignisverlauf](#).

Für eine kontinuierliche Aufzeichnung von Ereignissen in AWS-Konto, einschließlich Ereignissen für Amazon ML, erstellen Sie einen Trail. Ein Trail ermöglicht es CloudTrail, Protokolldateien in einem Amazon-S3-Bucket bereitzustellen. Wenn Sie einen Pfad in der Konsole anlegen, gilt dieser für alle AWS-Regionen. Der Trail protokolliert Ereignisse aus allen Regionen in der AWS-Partition und stellt die Protokolldateien in dem von Ihnen angegebenen Amazon S3 Bucket bereit. Darüber hinaus können Sie andere AWS-Services konfigurieren, um die in den CloudTrail-Protokollen erfassten Ereignisdaten weiter zu analysieren und entsprechend zu agieren. Weitere Informationen finden Sie unter:

- [Übersicht zum Erstellen eines Trails](#)
- [Von CloudTrail unterstützte Dienste und Integrationen](#)

- [Konfigurieren von Amazon-SNS-Benachrichtigungen für CloudTrail](#)
- [Empfangen von CloudTrail-Protokolldateien aus mehreren Regionen](#) und [Empfangen von CloudTrail-Protokolldateien aus mehreren Konten](#)

Amazon ML unterstützt die Protokollierung der folgenden Aktionen als Ereignisse in CloudTrail-Protokolldateien:

- [AddTags](#)
- [CreateBatchPrediction](#)
- [CreateDataSourceFromRDS](#)
- [CreateDataSourceFromRedshift](#)
- [CreateDataSourceFromS3](#)
- [CreateEvaluation](#)
- [CreateMLModel](#)
- [CreateRealtimeEndpoint](#)
- [DeleteBatchPrediction](#)
- [DeleteDataSource](#)
- [DeleteEvaluation](#)
- [DeleteMLModel](#)
- [DeleteRealtimeEndpoint](#)
- [DeleteTags](#)
- [DescribeTags](#)
- [UpdateBatchPrediction](#)
- [UpdateDataSource](#)
- [UpdateEvaluation](#)
- [UpdateMLModel](#)

Die folgenden Amazon-ML-Vorgänge verwenden Anforderungsparameter mit Anmeldeinformationen. Bevor diese Anforderungen an CloudTrail gesendet werden, werden die Anmeldeinformationen durch drei Sternchen ersetzt („***“):

- [CreateDataSourceFromRDS](#)

- [CreateDataSourceFromRedshift](#)

Wenn die folgenden Amazon-ML-Vorgänge mit der Amazon ML-Konsole durchgeführt werden, wird das `AttributComputeStatistics` nicht in `enthaltenRequestParameters` Komponente des CloudTrail-Protokolls:

- [CreateDataSourceFromRedshift](#)
- [CreateDataSourceFromS3](#)

Jeder Ereignis- oder Protokolleintrag enthält Informationen zu dem Benutzer, der die Anforderung generiert hat. Anhand der Identitätsinformationen zur Benutzeridentität können Sie Folgendes bestimmen:

- Ob die Anfrage mit Stammbenutzer- oder AWS Identity and Access Management (IAM)-Anmeldeinformationen ausgeführt wurde.
- Ob die Anforderung mit temporären Sicherheitsanmeldeinformationen für eine Rolle oder einen Verbundbenutzer ausgeführt wurde.
- Ob die Anforderung aus einem anderen AWS-Service gesendet wurde

Weitere Informationen finden Sie unter dem [CloudTrail userIdentity-Element](#).

Beispiel: Amazon-ML-Protokolldateieinträge

Ein Trail ist eine Konfiguration, durch die Ereignisse als Protokolldateien an den von Ihnen angegebenen Amazon-S3-Bucket übermittelt werden. CloudTrail-Protokolldateien können einen oder mehrere Einträge enthalten. Ein Ereignis stellt eine einzelne Anfrage aus einer beliebigen Quelle dar und enthält unter anderem Informationen über die angeforderte Aktion, das Datum und die Uhrzeit der Aktion sowie über die Anfrageparameter. CloudTrail-Protokolleinträge sind kein geordnetes Stack-Trace der öffentlichen API-Aufrufe und erscheinen daher in keiner bestimmten Reihenfolge.

Das folgende Beispiel zeigt einen CloudTrail-Protokolleintrag, der die Aktion demonstriert.

```
{
  "Records": [
    {
      "eventVersion": "1.03",
```

```

    "userIdentity": {
      "type": "IAMUser",
      "principalId": "EX_PRINCIPAL_ID",
      "arn": "arn:aws:iam::012345678910:user/Alice",
      "accountId": "012345678910",
      "accessKeyId": "EXAMPLE_KEY_ID",
      "userName": "Alice"
    },
    "eventTime": "2015-11-12T15:04:02Z",
    "eventSource": "machinelearning.amazonaws.com",
    "eventName": "CreateDataSourceFromS3",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "127.0.0.1",
    "userAgent": "console.amazonaws.com",
    "requestParameters": {
      "data": {
        "dataLocationS3": "s3://aml-sample-data/banking-batch.csv",
        "dataSchema": "{\"version\":\"1.0\",\"rowId\":null,\"rowWeight
\":null,
          \"targetAttributeName\":null,\"dataFormat\":\"CSV\",
          \"dataFileContainsHeader\":false,\"attributes\":[
            {\"attributeName\":\"age\",\"attributeType\":\"NUMERIC\"},
            {\"attributeName\":\"job\",\"attributeType\":\"CATEGORICAL
\"},
            {\"attributeName\":\"marital\",\"attributeType\":
\"CATEGORICAL\"},
            {\"attributeName\":\"education\",\"attributeType\":
\"CATEGORICAL\"},
            {\"attributeName\":\"default\",\"attributeType\":
\"CATEGORICAL\"},
            {\"attributeName\":\"housing\",\"attributeType\":
\"CATEGORICAL\"},
            {\"attributeName\":\"loan\",\"attributeType\":\"CATEGORICAL
\"},
            {\"attributeName\":\"contact\",\"attributeType\":
\"CATEGORICAL\"},
            {\"attributeName\":\"month\",\"attributeType\":\"CATEGORICAL
\"},
            {\"attributeName\":\"day_of_week\",\"attributeType\":
\"CATEGORICAL\"},
            {\"attributeName\":\"duration\",\"attributeType\":\"NUMERIC
\"},
            {\"attributeName\":\"campaign\",\"attributeType\":\"NUMERIC
\"},

```

```

        {"attributeName": "pdays", "attributeType": "NUMERIC"},
        {"attributeName": "previous", "attributeType": "NUMERIC"},
        {"attributeName": "poutcome", "attributeType": "CATEGORICAL"},
        {"attributeName": "emp_var_rate", "attributeType": "NUMERIC"},
        {"attributeName": "cons_price_idx", "attributeType": "NUMERIC"},
        {"attributeName": "cons_conf_idx", "attributeType": "NUMERIC"},
        {"attributeName": "euribor3m", "attributeType": "NUMERIC"},
        {"attributeName": "nr_employed", "attributeType": "NUMERIC"}
    ], "excludedAttributeNames": []}
},
"dataSourceId": "exampleDataSourceId",
"dataSourceName": "Banking sample for batch prediction"
},
"responseElements": {
    "dataSourceId": "exampleDataSourceId"
},
"requestID": "9b14bc94-894e-11e5-a84d-2d2deb28fdec",
"eventID": "f1d47f93-c708-495b-bff1-cb935a6064b2",
"eventType": "AwsApiCall",
"recipientAccountId": "012345678910"
},
{
    "eventVersion": "1.03",
    "userIdentity": {
        "type": "IAMUser",
        "principalId": "EX_PRINCIPAL_ID",
        "arn": "arn:aws:iam::012345678910:user/Alice",
        "accountId": "012345678910",
        "accessKeyId": "EXAMPLE_KEY_ID",
        "userName": "Alice"
    },
    "eventTime": "2015-11-11T15:24:05Z",
    "eventSource": "machinelearning.amazonaws.com",
    "eventName": "CreateBatchPrediction",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "127.0.0.1",
    "userAgent": "console.amazonaws.com",

```

```
    "requestParameters": {
      "batchPredictionName": "Batch prediction: ML model: Banking sample",
      "batchPredictionId": "exampleBatchPredictionId",
      "batchPredictionDataSourceId": "exampleDataSourceId",
      "outputUri": "s3://EXAMPLE_BUCKET/BatchPredictionOutput/",
      "mLModelId": "exampleModelId"
    },
    "responseElements": {
      "batchPredictionId": "exampleBatchPredictionId"
    },
    "requestID": "3e18f252-8888-11e5-b6ca-c9da3c0f3955",
    "eventID": "db27a771-7a2e-4e9d-bfa0-59deee9d936d",
    "eventType": "AwsApiCall",
    "recipientAccountId": "012345678910"
  }
]
}
```

Markieren Sie Ihre Amazon ML-Objekte

Organisieren und verwalten Sie Ihre Amazon Machine Learning (Amazon ML) -Objekte, indem Sie ihnen Metadaten mit Tags zuweisen. Ein Tag ist ein Schlüssel-Wert-Paar, das Sie für ein Objekt definieren.

Sie können Tags aber nicht nur verwenden, um Ihre Amazon ML-Objekte zu organisieren und zu verwalten, sondern auch, um Ihre AWS-Kosten zu kategorisieren und zu verfolgen. Wenn Sie Tags auf AWS-Objekte, einschließlich ML-Modellen, anwenden, enthält der AWS-Kostenzuordnungsbericht mit Tags aggregierte Nutzungs- und Kostendaten. Sie können Tags anwenden, die geschäftliche Kategorien (wie Kostenstellen, Anwendungsname oder Eigentümer) darstellen, um die Kosten für mehrere Services zu organisieren. Weitere Informationen finden Sie unter [Verwenden von Kostenzuordnungs-Tags für benutzerdefinierte Fakturierungsberichte](#) im AWS Billing-Benutzerhandbuch.

Inhalt

- [Grundlagen zu Tags](#)
- [Tag-Einschränkungen](#)
- [Markieren von Amazon ML -Objekten \(Konsole\)](#)
- [Markieren von Amazon ML-Objekten \(API\)](#)

Grundlagen zu Tags

Verwenden Sie Tags, um Ihre Objekte zu kategorisieren und somit einfacher zu verwalten. Sie können Objekte beispielsweise nach Zweck, Inhaber oder Umgebung kategorisieren. Sie könnten dann zum Beispiel eine Reihe von Tags definieren, mit der Sie Modelle nach Inhaber und zugehöriger Anwendung nachverfolgen können. Im Folgenden finden Sie einige Beispiele:

- Projekt: Project name
- Eigentümer: Name
- Zweck: Marketing-Prognosen
- Anwendung: Anwendungsname
- Umgebung: Produktion

Sie können die Amazon ML-Konsole oder -API verwenden, um die folgenden Aufgaben auszuführen:

- Hinzufügen von Tags zu einem Objekt
- Anzeigen der Tags für Ihre Objekte
- Bearbeiten der Tags für Ihre Objekte
- Löschen von Tags von einem Objekt

Standardmäßig werden auf ein Amazon ML-Objekt angewandte Tags auf Objekte kopiert, die mit diesem Objekt erstellt wurden. Wenn eine Amazon Simple Storage Service (Amazon S3) - Datenquelle beispielsweise die „Marketingkosten: Zielgerichtete Marketingkampagne „Tag, ein mit dieser Datenquelle erstelltes Modell, hätte auch die „Marketingkosten: Gezielte Marketingkampagne „-Tag, ebenso wie die Bewertung für das Modell. Auf diese Weise können Sie Tags verwenden, um verwandte Objekte nachzuverfolgen, wie z. B. alle Objekte für eine Marketingkampagne. Wenn ein Konflikt zwischen Tag-Quellen, wie zum Beispiel ein Modell mit dem Tag „Marketingkosten: Gezielte Marketingkampagne“ und eine Datenquelle mit dem Tag „Marketingkosten: Target Marketingkunden“ wendet Amazon ML das Tag aus dem Modell an.

Tag-Einschränkungen

Für Tags gelten die folgenden Einschränkungen.

Grundlegende Einschränkungen:

- Die maximale Anzahl an Tags pro Objekt beträgt 50.
- Bei Tag-Schlüsseln und -Werten muss die Groß-/Kleinschreibung beachtet werden.
- Sie können Tags für ein gelöschttes Objekt nicht ändern oder bearbeiten.

Einschränkungen für Tag-Schlüssel:

- Jeder Tag-Schlüssel muss einmalig sein. Wenn Sie einen Tag mit einem Schlüssel hinzufügen, der bereits verwendet wird, wird das vorhandene Schlüssel-Wert-Paar durch den neuen Tag für dieses Objekt überschrieben.
- Sie können einen Tag-Schlüssel nicht mit `aws:` starten, da dieses Präfix für die Verwendung durch AWS reserviert ist. AWS erstellt zwar in Ihrem Namen Tags, die mit diesem Präfix beginnen, Sie können diese jedoch nicht bearbeiten oder löschen.

- Tag-Schlüssel müssen zwischen 1 und 128 Unicode-Zeichen lang sein.
- Tag-Schlüssel müssen aus folgenden Zeichen bestehen: Unicode-Buchstaben, Ziffern, Leerzeichen und die folgenden Sonderzeichen: `_ . / = + - @`aus.

Einschränkungen für den Tag-Wert:

- Tag-Werte müssen zwischen 0 und 255 Unicode-Zeichen lang sein.
- Tag-Werte können leer sein. Andernfalls müssen sie aus folgenden Zeichen bestehen: Unicode-Buchstaben, Ziffern, Leerzeichen und eines der folgenden Sonderzeichen: `_ . / = + - @`aus.

Markieren von Amazon ML -Objekten (Konsole)

Sie können Tags mit der Amazon ML-Konsole anzeigen, hinzufügen, bearbeiten und löschen.

Anzeigen der Tags eines Objekts (Konsole)

1. Melden Sie sich beim anAWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/>aus.
2. Erweitern Sie in der Navigationsleiste die Regionsauswahl und wählen Sie eine aus.
3. Wählen Sie auf der Seite Objekte ein Objekt.
4. Scrollen Sie zum Abschnitt Tags des gewählten Objekts. Die Tags für dieses Objekt sind im unteren Bereich des Abschnitts aufgeführt.

Hinzufügen eines Tags zu einem Objekt (Konsole)

1. Melden Sie sich beim anAWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/>aus.
2. Erweitern Sie in der Navigationsleiste die Regionsauswahl und wählen Sie eine aus.
3. Wählen Sie auf der Seite Objekte ein Objekt.
4. Scrollen Sie zum Abschnitt Tags des gewählten Objekts. Die Tags für dieses Objekt sind im unteren Bereich des Abschnitts aufgeführt.
5. Wählen Sie Hinzufügen und Bearbeiten von Tags.
6. Geben Sie unter Tag hinzufügen in das Feld Schlüssel den Tag-Schlüssel ein. Optional können Sie auch einen Tag-Wert in das Feld Wert eingeben und dann Änderungen anwenden wählen.

Wenn die Schaltfläche **Änderungen anwenden** nicht aktiviert ist, entspricht entweder der angegebene Tag-Schlüssel oder Tag-Wert nicht den Tag-Einschränkungen. Weitere Informationen finden Sie unter [Tag-Einschränkungen](#).

7. Um Ihren neuen Tag in der Liste im Abschnitt Tags anzuzeigen, aktualisieren Sie die Seite.

So bearbeiten Sie ein Tag (Konsole)

1. Melden Sie sich beim anAWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/aus>.
2. Erweitern Sie in der Navigationsleiste die Regionsauswahl und wählen Sie eine Region aus.
3. Wählen Sie auf der Seite Objekte ein Objekt.
4. Scrollen Sie zum Abschnitt Tags des gewählten Objekts. Die Tags für dieses Objekt sind im unteren Bereich des Abschnitts aufgeführt.
5. Wählen Sie **Hinzufügen und Bearbeiten von Tags**.
6. Bearbeiten Sie unter **Angewandte Tags** den Tag-Wert im Feld **Wert** und klicken Sie dann auf **Änderungen anwenden**.

Wenn die Schaltfläche **Änderungen anwenden** nicht aktiviert ist, entspricht der angegebene Tag-Wert nicht den Tag-Einschränkungen. Weitere Informationen finden Sie unter [Tag-Einschränkungen](#).

7. Um Ihr neues Tag in der Liste im Abschnitt Tags anzuzeigen, aktualisieren Sie die Seite.

Löschen eines Tags von einem Objekt (Konsole)

1. Melden Sie sich beim anAWS Management Console und öffnen Sie die Amazon Machine Learning Learning-Konsole unter <https://console.aws.amazon.com/machinelearning/aus>.
2. Erweitern Sie in der Navigationsleiste die Regionsauswahl und wählen Sie eine aus.
3. Wählen Sie auf der Seite Objekte ein Objekt.
4. Scrollen Sie zum Abschnitt Tags des gewählten Objekts. Die Tags für dieses Objekt sind im unteren Bereich des Abschnitts aufgeführt.
5. Wählen Sie **Hinzufügen und Bearbeiten von Tags**.
6. Klicken Sie unter **Angewandte Tags** auf das Tag, das Sie löschen möchten, und klicken Sie dann auf **Änderungen anwenden**.

Markieren von Amazon ML-Objekten (API)

Sie können Tags mit der Amazon ML-API hinzufügen, aufführen und löschen. Beispiele finden Sie in der folgenden Dokumentation:

[AddTags](#)

Fügt Tags für das angegebene Objekt hinzu oder bearbeitet diese.

[DescribeTags](#)

Listet die Tags für das angegebene Objekt auf.

[DeleteTags](#)

Löscht Tags aus dem angegebenen Objekt.

Amazon Machine Learning-Referenz

Themen

- [Gewähren von Amazon ML-Berechtigungen zum Lesen von Daten aus Amazon S3](#)
- [Gewähren von Berechtigungen für Amazon ML zwecks Ausgabe von Voraussagen in Amazon S3](#)
- [Steuern des Zugriffs auf Amazon ML-Ressourcen – mit IAM](#)
- [Dienstübergreifende Confused-Deputy-Prävention](#)
- [Dependency Management von asynchrone Operationen](#)
- [Das Überprüfen des Status einer Anfrage](#)
- [Systemeinschränkungen](#)
- [Namen und IDs für alle Objekte](#)
- [Objektlebensdauer](#)

Gewähren von Amazon ML-Berechtigungen zum Lesen von Daten aus Amazon S3

Um aus Ihren Eingabedaten in Amazon S3 ein Datenquellenobjekt zu erstellen, müssen Sie Amazon ML die folgenden Berechtigungen für den S3-Speicherort erteilen, an dem Ihre Eingabedaten gespeichert sind:

- GetObject-Berechtigung für den S3-Bucket und das Präfix.
- ListBucket-Berechtigung für den S3-Bucket. Im Gegensatz zu anderen Aktionen ListBucket müssen Bucket-weite Berechtigungen erhalten (und nicht für das Präfix). Sie können die Berechtigungen jedoch auf ein bestimmtes Präfix einschränken, indem Sie eine Condition-Klausel verwenden.

Wenn Sie die Amazon ML-Konsole zum Erstellen der Datenquelle verwenden, können diese Berechtigungen für Sie dem Bucket hinzugefügt werden. Beim Ausführen der Schritte im Assistenten werden Sie aufgefordert, zu bestätigen, ob Sie sie hinzufügen möchten. Die folgende Beispielrichtlinie zeigt, wie Amazon ML die Berechtigung zum Lesen von Daten aus dem Beispielspeicherort s3 erteilt wird: `//Beispiel-Bucket/Beispiel-Präfix`, beim Scoping der ListBucket Erlaubnis nur für die `Beispiel-Präfix` Eingabepfad.

```
{
```

```

"Version": "2008-10-17",
"Statement": [
  {
    "Effect": "Allow",
    "Principal": { "Service": "machinelearning.amazonaws.com" },
    "Action": "s3:GetObject",
    "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*"
    "Condition": {
      "StringEquals": { "aws:SourceAccount": "123456789012" }
      "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
    }
  },
  {
    "Effect": "Allow",
    "Principal": {"Service": "machinelearning.amazonaws.com"},
    "Action": "s3:ListBucket",
    "Resource": "arn:aws:s3:::examplebucket",
    "Condition": {
      "StringLike": { "s3:prefix": "exampleprefix/*" }
      "StringEquals": { "aws:SourceAccount": "123456789012" }
      "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
    }
  }
]
}

```

Um diese Richtlinie für Ihre Daten anzuwenden, müssen Sie die Richtlinienanweisung in Zusammenhang mit dem S3-Bucket, in dem Sie Ihre Daten gespeichert haben, bearbeiten.

Vorgehensweise zum Bearbeiten der Berechtigungsrichtlinie für einen S3-Bucket (unter Verwendung der alten Konsole)

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon-S3-Konsole unter <https://console.aws.amazon.com/s3/>.
2. Wählen Sie den Namen des Buckets aus, on dem sich Ihre daten befinden.
3. Wählen Sie Properties (Eigenschaften).
4. Wählen Sie Edit bucket policy.
5. Geben Sie die oben gezeigte Richtlinie ein, die Sie an Ihre Anforderungen anpassen sollten, und wählen Sie dann Save aus.

6. Wählen Sie Save (Speichern) aus.

Vorgehensweise zum Bearbeiten der Berechtigungsrichtlinie für einen S3-Bucket (unter Verwendung der neuen Konsole)

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon-S3-Konsole unter <https://console.aws.amazon.com/s3/>.
2. Wählen Sie den Bucket-Namen und dann Berechtigungen aus.
3. Wählen Sie Bucket Policy aus.
4. Geben Sie die oben gezeigte Richtlinie ein, die Sie an Ihre Anforderungen anpassen sollten.
5. Wählen Sie Save (Speichern) aus.

Gewähren von Berechtigungen für Amazon ML zwecks Ausgabe von Voraussagen in Amazon S3

Um die Ergebnisse der Stapelvoraussage-Operation in Amazon S3 auszugeben, müssen Sie Amazon ML die folgenden Berechtigungen für den Ausgabeort gewähren, die als Input bei der Operation zum Erstellen von Stapelvoraussagen bereitgestellt werden:

- GetObject-Berechtigung für Ihren S3-Bucket und Ihr Präfix.
- PutObject-Berechtigung für Ihren S3-Bucket und Ihr Präfix.
- PutObjectAcl auf Ihrem S3-Bucket und Präfix.
 - Amazon ML benötigt diese Berechtigung, um sicherzustellen, dass es die Konserven gewähren kann [ACL](#) bucket-owner-full-control Berechtigung für Ihr AWS-Konto, nachdem Objekte erstellt wurden.
- ListBucket-Berechtigung für den S3-Bucket. Im Gegensatz zu anderen Aktionen ListBucket müssen Bucket-weite Berechtigungen erhalten (und nicht für das Präfix). Sie können die Berechtigungen jedoch auf ein bestimmtes Präfix einschränken, indem Sie eine Bedingungsklausel verwenden.

Wenn Sie die Amazon ML-Konsole zum Erstellen der Stapelvoraussageanforderung verwenden, können diese Berechtigungen für Sie dem Bucket hinzugefügt werden. Sie werden aufgefordert, zu bestätigen, ob Sie diese hinzufügen möchten, wenn Sie die Schritte im Assistenten ausführen.

Die folgende Beispielrichtlinie zeigt, wie Amazon ML die Berechtigung erteilt wird, Daten an den Beispielspeicherort `s3://examplebucket/exampleprefix` zu schreiben, während

dieListBucketBerechtigung nur für den Eingabepfad `exampleprefix` und Erteilung der Berechtigung für Amazon ML, Put-Objekt-ACLs für das Ausgabepräfix festzulegen:

```
{
  "Version": "2008-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": { "Service": "machinelearning.amazonaws.com" },
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*"
      "Condition": {
        "StringEquals": { "aws:SourceAccount": "123456789012" }
        "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
      }
    },
    {
      "Effect": "Allow",
      "Principal": { "Service": "machinelearning.amazonaws.com" },
      "Action": "s3:PutObjectAcl",
      "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*",
      "Condition": {
        "StringEquals": { "s3:x-amz-acl": "bucket-owner-full-control" }
        "StringEquals": { "aws:SourceAccount": "123456789012" }
        "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
      }
    },
    {
      "Effect": "Allow",
      "Principal": { "Service": "machinelearning.amazonaws.com" },
      "Action": "s3:ListBucket",
      "Resource": "arn:aws:s3:::examplebucket",
      "Condition": {
        "StringLike": { "s3:prefix": "exampleprefix/*" }
        "StringEquals": { "aws:SourceAccount": "123456789012" }
        "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
      }
    }
  ]
}
```

```
} ]  
}
```

Um diese Richtlinie für Ihre Daten anzuwenden, müssen Sie die Richtlinienanweisung in Zusammenhang mit dem S3-Bucket, in dem Sie Ihre Daten gespeichert haben, bearbeiten.

Vorgehensweise zum Bearbeiten der Berechtigungsrichtlinie für einen S3-Bucket (unter Verwendung der alten Konsole)

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon-S3-Konsole unter <https://console.aws.amazon.com/s3/>.
2. Wählen Sie den Namen des Buckets aus, on dem sich Ihre daten befinden.
3. Wählen Sie Properties (Eigenschaften).
4. Wählen Sie Edit bucket policy.
5. Geben Sie die oben gezeigte Richtlinie ein, die Sie an Ihre Anforderungen anpassen sollten, und wählen Sie dann Save aus.
6. Wählen Sie Save (Speichern) aus.

Vorgehensweise zum Bearbeiten der Berechtigungsrichtlinie für einen S3-Bucket (unter Verwendung der neuen Konsole)

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon-S3-Konsole unter <https://console.aws.amazon.com/s3/>.
2. Wählen Sie den Bucket-Namen und dann Berechtigungen aus.
3. Wählen Sie Bucket Policy aus.
4. Geben Sie die oben gezeigte Richtlinie ein, die Sie an Ihre Anforderungen anpassen sollten.
5. Wählen Sie Save (Speichern) aus.

Steuern des Zugriffs auf Amazon ML-Ressourcen – mit IAM

Mit AWS Identity and Access Management (IAM) können Sie den Zugriff auf AWS-Services und -Ressourcen für Ihre Benutzer sicher steuern. AWS IAM spezifische Amazon ML-API-Aktionen.

Mit IAM haben Sie folgende Möglichkeiten:

- Erstellen von Benutzern und Gruppen für Ihr AWS-Konto
- Zuweisen eindeutiger Sicherheitsanmeldeinformationen zu jedem Benutzer in Ihrem AWS-Konto
- Steuern der Berechtigungen der einzelnen Benutzer zum Durchführen von Aufgaben mit AWS-Ressourcen
- Einfache Freigabe Ihrer AWS-Ressourcen unter den Benutzern Ihres AWS-Kontos
- Erstellen von Rollen für Ihr AWS-Konto und Verwalten ihrer Berechtigungen, um festzulegen, welche Benutzer oder Services sie übernehmen können
- Sie können Rollen in IAM erstellen und Berechtigungen verwalten, um zu steuern, welche Operationen von einer Entität oder einem AWS-Service mit der Rolle ausgeführt werden können. Sie können auch bestimmen, welcher Entität die Rolle zugeordnet werden darf.

Wenn Ihre Organisation bereits über IAM-Identitäten verfügt, können Sie diese für die Gewährung von Berechtigungen zur Ausführung von Aufgaben mit AWS-Ressourcen verwenden.

Weitere Informationen zu IAM finden Sie im [IAM-Benutzerhandbuch](#).

IAM-Richtliniensyntax

Eine IAM-Richtlinie ist ein JSON-Dokument, das eine oder mehrere Anweisungen enthält. Jeder Anweisung hat die folgende Struktur:

```
{
  "Statement": [{
    "Effect": "effect",
    "Action": "action",
    "Resource": "arn",
    "Condition": {
      "condition operator": {
        "key": "value"
      }
    }
  }]
}
```

Eine Richtlinienanweisung umfasst die folgenden Elemente:

- Wirkung: Steuert die Berechtigung zur Verwendung von Ressourcen und API-Aktionen, die Sie später in der Anweisung angeben. Gültige Werte sind Allow und Deny. IAM-Benutzer verfügen standardmäßig nicht über die Berechtigung zur Verwendung von Ressourcen und API-Aktionen.

Daher werden alle Anfragen abgelehnt. Der Standardwert wird durch eine explizite Erlaubnis (Allow) überschrieben. Eine explizite Verweigerung (Deny) überschreibt alle Erlaubnisse (Allows).

- **Aktion:** Die spezifische API-Aktion oder -Aktionen, für die Sie Berechtigungen erteilen oder verweigern.
- **Ressource:** Die von einer Aktion betroffene Ressource. Um eine Ressource in der Anweisung anzugeben, verwenden Sie deren Amazon-Ressourcennamen (ARN).
- **Conditions (optional):** Steuert, wann Ihre Richtlinie in Kraft tritt.

Um das Erstellen und Verwalten von IAM-Richtlinien zu vereinfachen, können Sie den AWS Policy Generator und den IAM Policy Simulator verwenden.

Festlegen von IAM-Richtlinienaktionen für Amazon ML

In einer IAM-Richtlinienanweisung können Sie eine API-Aktion für jeden Service, der IAM unterstützt, angeben. Beim Erstellen einer Richtlinienanweisung für `AmazonMachineLearning`: auf den Namen der API-Aktion, wie in den folgenden Beispielen dargestellt:

- `machinelearning:CreateDataSourceFromS3`
- `machinelearning:DescribeDataSources`
- `machinelearning>DeleteDataSource`
- `machinelearning:GetDataSource`

Um mehrere Aktionen in einer einzigen Anweisung anzugeben, trennen Sie sie mit Kommata:

```
"Action": ["machinelearning:action1", "machinelearning:action2"]
```

Sie können auch mehrere Aktionen mittels Platzhaltern angeben. Beispielsweise können Sie alle Aktionen festlegen, deren Name mit dem Wort "Get" beginnt:

```
"Action": "machinelearning:Get*"
```

Um alle Amazon

```
"Action": "machinelearning:*"
```

Eine vollständige Liste der Amazon [Amazon Machine Learning](#).

Angeben von ARNs für Amazon ML-Ressourcen in IAM-Richtlinien

IAM-Richtlinienerklärungen gelten für eine oder mehrere Ressourcen. Ressourcen für Ihre Richtlinien werden mit ihren ARNs angegeben.

Um die ARNs für Amazon

"Ressource": `arn:aws:machinelearning:region:account:resource-type/identifizier`

Die folgenden Beispiele zeigen, wie Sie allgemeine ARNs angeben.

Datenquellen-ID: `my-s3-datasource-id`

```
"Resource":  
arn:aws:machinelearning:<region>:<your-account-id>:datasource/my-s3-datasource-id
```

ML-Modell-ID: `my-ml-model-id`

```
"Resource":  
arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/my-ml-model-id
```

Stapelvoraussage-ID: `my-batchprediction-id`

```
"Resource":  
arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/my-batchprediction-  
id
```

Evaluierungs-ID: `my-evaluation-id`

```
"Resource": arn:aws:machinelearning:<region>:<your-account-id>:evaluation/my-  
evaluation-id
```

Beispielhafte Richtlinien für Amazon

Beispiel 1: Benutzern ermöglichen, Metadaten von Ressourcen für maschinelles Lernen zu lesen

Die folgende Richtlinie ermöglicht es einem Benutzer oder einer Gruppe, die Metadaten von Datenquellen, ML-Modellen, Batch-Vorhersagen und Auswertungen zu lesen,

indem [DescribeDataSources](#), [DescribeMLModels](#), [DescribeBatchPredictions](#), [DescribeEvaluations](#), [GetDataSources](#) und [GetEvaluation](#) Aktionen für die angegebene (n) Ressource (n). Die Describe *-Operationsberechtigungen können nicht auf eine bestimmte Ressource beschränkt werden.

```
{
  "Version": "2012-10-17",
  "Statement": [{
    "Effect": "Allow",
    "Action": [
      "machinelearning:Get*"
    ],
    "Resource": [
      "arn:aws:machinelearning:<region>:<your-account-id>:datasource/S3-DS-ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:datasource/REDSHIFT-DS-
ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL-ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/BP-
ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:evaluation/EV-ID1"
    ]
  },
  {
    "Effect": "Allow",
    "Action": [
      "machinelearning:Describe*"
    ],
    "Resource": [
      "*"
    ]
  }
]}
```

Beispiel 2: Benutzern das Erstellen von Ressourcen für maschinelles Lernen ermöglichen

Die folgende Richtlinie erlaubt es einem Benutzer oder einer Gruppe, Machine Learning-Datenquellen, ML-Modelle, Stapelvoraussagen und Evaluierungen durch Ausführung der `CreateDataSourceFromS3`-, `CreateDataSourceFromRedshift`-, `CreateDataSourceFromRDS`-, `CreateMLModel`-, `CreateBatchPrediction`- und `CreateEvaluation`-Aktionen zu erstellen. Sie können die Berechtigungen für diese Aktionen nicht auf eine bestimmte Ressource einschränken.

```
{
```

```

"Version": "2012-10-17",
"Statement": [{
  "Effect": "Allow",
  "Action": [
    "machinelearning:CreateDataSourceFrom*",
    "machinelearning:CreateMLModel",
    "machinelearning:CreateBatchPrediction",
    "machinelearning:CreateEvaluation"
  ],
  "Resource": [
    "*"
  ]
}]
}

```

Beispiel 3: Benutzer können Echtzeit-Endpunkte erstellen (und löschen) und Vorhersagen in Echtzeit für ein ML-Modell durchführen

Die folgende Richtlinie erlaubt es Benutzern oder Gruppen, Echtzeitendpunkte zu erstellen und zu löschen sowie Echtzeitvoraussagen für ein bestimmtes ML-Modell zu generieren, indem sie `CreateRealtimeEndpoint`-, `DeleteRealtimeEndpoint`- und `Predict`-Aktionen für dieses Modell ausführen.

```

{
  "Version": "2012-10-17",
  "Statement": [{
    "Effect": "Allow",
    "Action": [
      "machinelearning:CreateRealtimeEndpoint",
      "machinelearning>DeleteRealtimeEndpoint",
      "machinelearning:Predict"
    ],
    "Resource": [
      "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL "
    ]
  }]
}

```

Beispiel 4: Benutzern erlauben, bestimmte Ressourcen zu aktualisieren und zu löschen

Die folgende Richtlinie erlaubt es einem Benutzer oder einer Gruppe, bestimmte Ressourcen in Ihrem AWS-Konto zu aktualisieren oder zu löschen, indem die Berechtigungen zum Ausführen von

UpdateDataSource-, UpdateMLModel-, UpdateBatchPrediction-, UpdateEvaluation-, DeleteDataSource-, DeleteMLModel-, DeleteBatchPrediction- und DeleteEvaluation-Aktionen für diese Ressourcen in Ihrem Konto gewährt werden.

```
{
  "Version": "2012-10-17",
  "Statement": [{
    "Effect": "Allow",
    "Action": [
      "machinelearning:Update*",
      "machinelearning>DeleteDataSource",
      "machinelearning>DeleteMLModel",
      "machinelearning>DeleteBatchPrediction",
      "machinelearning>DeleteEvaluation"
    ],
    "Resource": [
      "arn:aws:machinelearning:<region>:<your-account-id>:datasource/S3-DS-ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:datasource/REDSHIFT-DS-
ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL-ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/BP-
ID1",
      "arn:aws:machinelearning:<region>:<your-account-id>:evaluation/EV-ID1"
    ]
  }]
}
```

Beispiel 5: Beliebige Amazon MLAction zulassen

Mit der folgenden Richtlinie können Benutzer oder Gruppen jede beliebige Amazon Da diese Richtlinie vollen Zugriff auf alle Ihre Ressourcen für maschinelles Lernen gewährt, sollten Sie sie auf Administratoren beschränken.

```
{
  "Version": "2012-10-17",
  "Statement": [{
    "Effect": "Allow",
    "Action": [
      "machinelearning:*"
    ],
    "Resource": [
      "*"
    ]
  }]
}
```

```
]
  }]
}
```

Dienstübergreifende Confused-Deputy-Prävention

Das Confused-Deputy-Problem ist ein Sicherheitsproblem, bei dem eine juristische Stelle, die nicht über die Berechtigung zum Ausführen einer Aktion verfügt, eine privilegiere juristische Stelle zwingen kann, die Aktion auszuführen. In AWS kann der dienstübergreifende Identitätswechsel zu Confused-Deputy-Problem führen. Ein dienstübergreifender Identitätswechsel kann auftreten, wenn ein Dienst (der Anruf-Dienst) einen anderen Dienst anruft (den aufgerufenen Dienst). Der aufrufende Service kann manipuliert werden, um seine Berechtigungen zu verwenden, um Aktionen auf die Ressourcen eines anderen Kunden auszuführen, für die er sonst keine Zugriffsberechtigung haben sollte. Um dies zu verhindern, bietet AWS Tools, mit denen Sie Ihre Daten für alle Services mit Serviceprinzipalen schützen können, die Zugriff auf Ressourcen in Ihrem Konto erhalten haben.

Wir empfehlen die Verwendung [aws:SourceArn](#) und [aws:SourceAccount](#) globale Wenn der `aws:SourceArn`-Wert die Konto-ID nicht enthält, z. B. einen Amazon-S3-Bucket-ARN, müssen Sie beide globale Bedingungskontextschlüssel verwenden, um Berechtigungen einzuschränken. Wenn Sie beide globale Bedingungskontextschlüssel verwenden und der `aws:SourceArn`-Wert die Konto-ID enthält, müssen der `aws:SourceAccount`-Wert und das Konto im `aws:SourceArn`-Wert dieselbe Konto-ID verwenden, wenn sie in der gleichen Richtlinienanweisung verwendet wird. Verwenden Sie `aws:SourceArn`, wenn Sie nur eine Ressource mit dem betriebsübergreifenden Zugriff verknüpfen möchten. Verwenden Sie `aws:SourceAccount`, wenn Sie zulassen möchten, dass Ressourcen in diesem Konto mit der betriebsübergreifenden Verwendung verknüpft werden.

Der effektivste Weg, um sich vor dem Confused-Deputy-Problem zu schützen, ist die Verwendung des globalen Bedingungskontextschlüssels `aws:SourceArn` mit dem vollständigen ARN der Ressource. Wenn Sie den vollständigen ARN der Ressource nicht kennen oder wenn Sie mehrere Ressourcen angeben, verwenden Sie den globalen Kontextbedingungsschlüssel mit Platzhaltern (`aws:SourceArn`) * für die unbekannt Teile des ARN. Zum Beispiel `arn:aws:service:*:123456789012:*`.

Das folgende Beispiel zeigt, wie Sie `aws:SourceArn` und `aws:SourceAccount` globale Bedingungskontextschlüssel in Amazon ML, um das verwirrte Deputy Problem beim Lesen von Daten aus einem Amazon S3 S3-Bucket zu verhindern.

```
{
```

```

"Version": "2008-10-17",
"Statement": [
{
  "Effect": "Allow",
  "Principal": { "Service": "machinelearning.amazonaws.com" },
  "Action": "s3:GetObject",
  "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*"
  "Condition": {
    "StringEquals": { "aws:SourceAccount": "123456789012" }
    "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
  }
},
{
  "Effect": "Allow",
  "Principal": {"Service": "machinelearning.amazonaws.com"},
  "Action": "s3:ListBucket",
  "Resource": "arn:aws:s3:::examplebucket",
  "Condition": {
    "StringLike": { "s3:prefix": "exampleprefix/*" }
    "StringEquals": { "aws:SourceAccount": "123456789012" }
    "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*" }
  }
}]
}

```

Dependency Management von asynchrone Operationen

Batch-Operationen in Amazon ML hängen von anderen Operationen ab, um erfolgreich abgeschlossen werden zu können. Um diese Abhängigkeiten zu verwalten, identifiziert Amazon ML Anfragen, die Abhängigkeiten aufweisen, und prüft, ob die Operationen abgeschlossen wurden. Wenn die Operationen nicht abgeschlossen wurden, wird die entsprechende Anfrage von Amazon ML so lange zurückgestellt, bis die Operationen, von denen sie abhängen, abgeschlossen wurden.

Es gibt einige Abhängigkeiten zwischen Batch-Operationen. Bevor Sie beispielsweise ein ML-Modell erstellen können, müssen Sie eine Datenquelle erstellt haben, mit der Sie das ML-Modell schulen können. Amazon ML kann ein ML-Modell nicht ohne verfügbare Datenquelle schulen.

Amazon ML unterstützt jedoch das Abhängigkeiten-Management für asynchrone Operationen. Beispielsweise müssen Sie nicht warten, bis die Statistiken berechnet wurden, bevor Sie eine

Anforderung senden können, um ein ML-Modell für die Datenquelle zu schulen. Stattdessen können Sie, sobald die Datenquelle erstellt wurde, eine Anforderung senden, um ein ML-Modell mit der Datenquelle zu schulen. Amazon ML startet den Schulungsvorgang erst, nachdem die Datenquellenstatistiken berechnet wurden. Die `createMLModel`-Anforderung wird in eine Warteschlange verschoben, bis die Statistiken berechnet wurden; sobald dies der Fall ist, versucht Amazon ML umgehend, die `createMLModel`-Operation durchzuführen. Ebenso können Sie Anfragen für Stapelvoraussagen und Überprüfungen für ML-Modelle senden, welche die Schulung noch nicht beendet haben.

Die folgende Tabelle zeigt die Anforderungen zur Weiterverarbeitung von verschiedenen AmazonML Aktionen

Um ...	Muss ...
Ein ML-Modell (<code>createMLModel</code>) zu erstellen	Eine Datenquelle mit berechneten Datenstatistiken vorhanden sein
Erstellen Sie eine Batch-Vorhersage (<code>createBatchPrediction</code>)	Eine Datenquelle vorhanden sein ML-Modell
Erstellen Sie eine Batch-Auswertung (<code>createBatchEvaluation</code>)	Eine Datenquelle vorhanden sein ML-Modell

Das Überprüfen des Status einer Anfrage

Wenn Sie eine Anfrage einreichen, können Sie ihren Status mit der Amazon Machine Learning (Amazon ML) API überprüfen. Zum Beispiel, wenn Sie eine `createMLModel` können Sie seinen Status überprüfen, indem Sie `describeMLModel` aufrufen. Amazon ML antwortet mit einem der folgenden Status.

Status	Definition
PENDING	Amazon ML validiert die Anfrage. ODER

Status	Definition
	<p>Amazon ML wartet vor der Ausführung der Anfrage auf freiwerdende Rechenressourcen. Dies kann auftreten, wenn Ihr Konto die maximale Anzahl gleichzeitig ausgeführter Batch-Operation-Anfragen überschreitet. Wenn dies der Fall ist, wechselt der Status auf <code>InProgress</code>, wenn andere laufende Anforderungen abgeschlossen wurden oder abgebrochen wurden.</p> <p>ODER</p> <p>Amazon ML wartet eine Batch-Operation, die Ihre Anfrage benötigt, um abgeschlossen zu werden.</p>
<code>INPROGRESS</code>	Ihre Anfrage wird noch ausgeführt.
<code>COMPLETED</code>	Die Anfrage wurde abgeschlossen, und das Objekt ist bereit zur Verwendung (ML-Modelle und Datenquellen) oder Überprüfung (Stapelvoraussagen und -Auswertungen).
<code>FEHLGESCHLAGEN</code>	Es besteht ein Problem mit den Daten, die Sie zur Verfügung gestellt haben, oder Sie haben den Vorgang abgebrochen. Wenn Sie beispielsweise versuchen, Datenstatistiken auf einer Datenquelle zu berechnen, die nicht abgeschlossen wurde, erhalten Sie möglicherweise die Statusmeldung <code>Invalid</code> oder <code>Failed</code> . Die Fehlermeldung erklärt, warum der Vorgang nicht erfolgreich abgeschlossen wurde.
<code>DELETED</code>	Das Objekt wurde bereits gelöscht.

Amazon ML bietet auch Informationen zu einem Objekt, z. B. wann Amazon ML die Erstellung dieses Objekts abgeschlossen hat. Weitere Informationen finden Sie unter [Auflisten von Objekten](#).

Systemeinschränkungen

Um einen robuste, zuverlässigen Service bereitzustellen, gelten Amazon ML bestimmte Einschränkungen hinsichtlich der Anfragen, die an das System gestellt werden. Ein Großteil der ML-Probleme fällt in diese Einschränkungen. Wenn Sie jedoch der Ansicht sind, dass Ihre Nutzung

von Amazon ML durch diese Einschränkungen eingeschränkt wird, können Sie sich an den [AWS-Kundenservice](#) wenden und die Anfrage stellen, dass eine Einschränkung gelockert wird. Vielleicht gibt es beispielsweise eine Begrenzung von 5 für die Anzahl von Aufträgen, die Sie gleichzeitig ausführen können. Wenn Sie feststellen, dass sich häufig Aufträge in der Warteschlange befinden, die aufgrund dieser Einschränkung auf Ressourcen warten, so ist es wahrscheinlich sinnvoller, diese Einschränkung für Ihr Konto zu lockern.

In der folgenden Tabelle sind standardmäßige Einschränkungen pro Konto in Amazon ML aufgeführt. Nicht all diese Einschränkungen können vom AWS-Kundenservice gelockert werden.

Einschränkungstyp	Systemeinschränkungen
Größe der einzelnen Beobachtungen	100 KB
Größe der Schulungsdaten*	100 GB
Eingabegröße für Stapelvoraussage	1 TB
Eingabegröße für Stapelvoraussage (Anzahl Datensätze)	100 Mio.
Anzahl der Variablen in einer Datendatei (Schema)	1.000
Rezeptkomplexität (Anzahl verarbeiteter Ausgabevariablen)	10.000
TPS für jeden Echtzeitvoraussage-Endpunkt	200
TPS insgesamt für alle Echtzeitvoraussage-Endpunkte	10.000
RAM insgesamt für alle Echtzeitprognose-Endpunkte	10 GB
Anzahl gleichzeitiger Aufträge	25
Längste Laufzeit für jeden Auftrag	7 Tage
Anzahl Klassen für Mehrklassen-ML-Modelle	100
Größe ML-Modell	Mindestens 1 MB, maximal 2 GB
Anzahl von Tags pro Objekt	50

- Die Größe Ihrer Datendateien ist begrenzt, um sicherzustellen, dass Aufträge rechtzeitig abgeschlossen werden. Aufträge, die seit mehr als sieben Tagen ausgeführt werden, werden automatisch mit dem Status FEHLGESCHLAGEN beendet.

Namen und IDs für alle Objekte

Alle Objekte in Amazon ML müssen über eine ID verfügen. Die Amazon ML-Konsole generiert ID-Werte für Sie. Wenn Sie die API verwenden, müssen Sie jedoch Ihre eigenen generieren. Jede ID muss eindeutig innerhalb der Amazon ML-Objekte desselben Typs in Ihrem AWS-Konto sein. Sie können also nicht zwei Evaluierungen mit derselben ID haben. Es ist möglich, eine Evaluierung und eine Datenquelle mit der gleichen ID zu haben, dies wird aber nicht empfohlen.

Wir empfehlen, dass Sie zufällig generierte IDs für Ihre Objekte verwenden, mit einer kurzen Zeichenfolge zur Identifizierung des Typs als Präfix. Wenn die Amazon ML-Konsole beispielsweise eine Datenquelle generiert, weist sie der Datenquelle eine zufällige, eindeutige ID wie „DS-zSCW“ zuluWiOxF“. Diese ID ist ausreichend zufällig, um Kollisionen für einen einzelnen Benutzer zu vermeiden, und gleichzeitig kompakt und lesbar. Das Präfix "ds" ist aus praktischen Gründen und für die Klarheit vorhanden, aber nicht erforderlich. Wenn Sie sich nicht sicher sind, was Sie für Ihre ID-Strings verwenden sollen, empfehlen wir die Verwendung von hexadezimalen UUID-Werten (wie 28b1e915-57e5-4e6c-a7bd-6fb4e729cb23), die in modernen Programmierungsumgebungen problemlos verfügbar sind.

ID-Zeichenfolgen können ASCII-Buchstaben, Ziffern, Bindestriche und Unterstriche enthalten und bis zu 64 Zeichen lang sein. Es ist möglich und vielleicht auch praktisch, Metadaten in eine ID-Zeichenfolge zu codieren. Es wird aber nicht empfohlen, da nach Erstellung eines Objekts seine ID nicht geändert werden kann.

Objektnamen bieten eine einfache Möglichkeit für Sie, den einzelnen Objekten benutzerfreundliche Metadaten zuzuordnen. Namen können nach der Erstellung eines Objekts geändert werden. Dadurch ist es möglich, einige Aspekte Ihres ML-Workflows im Namen eines Objekts wiederzugeben. Sie können einem ML-Modell beispielsweise zu Beginn den Namen "Experiment 3" geben und es dann in "Finales Produktionsmodell" umbenennen. Namen können aus einer beliebigen Zeichenfolge mit bis zu 1 024 Zeichen bestehen.

Objektlebensdauer

Sie können von Ihnen mit Amazon ML erstellte Datenquellen, ML-Modelle, Auswertungen oder Batch-Voraussageobjekte mindestens zwei Jahre lang verwenden. Amazon ML entfernt möglicherweise

automatisch Objekte, auf die länger als zwei Jahre nicht zugegriffen wurde bzw. die in diesem Zeitraum nicht verwendet wurden.

Ressourcen

Die folgenden verwandten Ressourcen bieten Ihnen nützliche Informationen für die Arbeit mit diesem Service.

- [Amazon ML Produktinformationen](#)— Erfasst alle zugehörigen Produktinformationen über Amazon ML an einem zentralen Ort.
- [Häufig gestellte Fragen zu Amazon ML](#)— Deckt alle wichtigen Fragen ab, die Entwickler zu diesem Produkt gestellt haben.
- [Amazon ML Beispielcode](#)— Beispielanwendungen, für die Amazon ML verwendet werden. Sie können den Beispielcode als Ausgangspunkt zur Erstellung Ihrer eigenen ML-Anwendungen verwenden.
- [Amazon ML API-Referenz](#)— Beschreibt sämtliche API-Vorgänge für Amazon ML im Detail. Enthält darüber hinaus Beispiele für Anforderungen und Antworten für die unterstützten Web-Service-Protokolle.
- [AWS-Entwicklerressourcenzentrum](#)— Bietet einen zentralen Ausgangspunkt für Dokumentationen, Codebeispiele, Versionshinweise und sonstigen Informationen für die Entwicklung innovativer Anwendungen mit AWS.
- [AWS-Schulungen und Kurse](#)— Links zu rollenbasierten und speziellen Kursen sowie Übungsgruppen zum Selbststudium zur Verbesserung Ihrer AWS-Kompetenzen und für praktische Erfahrung.
- [AWS-Entwicklertools](#) – Links zu Entwicklertools und -Ressourcen mit Dokumentationen, Codebeispielen, Versionshinweisen und sonstigen Informationen für die Entwicklung innovativer Anwendungen mit AWS.
- [AWS Support Center](#)— der Hub für die Erstellung und Verwaltung Ihrer AWS-Supportfälle. Bietet außerdem Links zu hilfreichen Ressourcen wie z. B. Foren, technischen Fragen und Antworten, Übersicht zum Servicestatus und AWS Trusted Advisor.
- [AWS-Support](#) – Die Hauptwebsite mit Informationen zum AWS Support ist ein persönlicher und reaktionsschneller Support-Kanal, der Sie beim Konfigurieren und Ausführen von Anwendungen in der Cloud unterstützt.
- [Kontakt](#)— Ein zentraler Kontaktpunkt für Anfragen zur AWS-Abrechnung, zu Ihrem Konto, Ereignissen, Missbrauch und anderen Problemen.
- [Nutzungsbedingungen für die AWS-Website](#) – Detaillierte Informationen zu unseren Copyright- und Markenbestimmungen, Ihrem Konto, den Lizenzen und anderen Themen.

Dokumentverlauf

Die folgende Tabelle enthält wichtige Änderungen an der Dokumentation in dieser Version von Amazon Machine Learning (Amazon ML).

- API-Version: 2015-04-09
- Letzte Aktualisierung der Dokumentation: 2016-08-02

Änderung	Beschreibung	Änderungsdatum
Metriken hinzugefügt	Diese Version von Amazon ML fügt neue Metriken für Amazon ML-Objekte hinzu. Weitere Informationen finden Sie unter Auflisten von Objekten .	2. August 2016
Löschen mehrerer Objekte	Diese Version von Amazon ML ermöglicht das Löschen mehrerer Amazon ML-Objekte. Weitere Informationen finden Sie unter Löschen von Objekten .	20. Juli 2016
Tagging hinzugefügt	Diese Version von Amazon ML ermöglicht das Anwenden von Tags auf Amazon ML-Objekte. Weitere Informationen finden Sie unter Markieren Sie Ihre Amazon ML-Objekte .	23. Juni 2016
Kopieren von Amazon Redshift Redshift-Datenquellen	Diese Version von Amazon ML fügt die Möglichkeit des Kopierens von Amazon Redshift Redshift-Datenquelleneinstellungen zu einer neuen Amazon Redshift Redshift-Datenquelle hinzu. Weitere Informationen zum Kopieren von Amazon Redshift-Datenquelleneinstellungen finden Sie unter Kopieren einer Datenquelle (Konsole) aus.	11. April 2016
Shuffling hinzugefügt	Diese Version von Amazon ML ermöglicht das Mischen Ihrer Eingabedaten.	5. April 2016

Änderung	Beschreibung	Änderungsdatum
	<p>Weitere Informationen zur Verwendung des Parameters <code>Art der Mischung</code> finden Sie unter Art der Mischung von Schulungsdaten.</p>	
<p>Verbesserte Datenquellen-erstellung mit Amazon Redshift</p>	<p>Diese Version von Amazon ML ermöglicht das Testen Ihrer Amazon Redshift Redshift-Einstellungen, wenn Sie eine Amazon ML-Datenquelle in der Konsole erstellen, um zu überprüfen, dass die Verbindung funktioniert. Weitere Informationen finden Sie unter Erstellen einer Datenquelle mit Amazon Redshift Data (Konsole).</p>	<p>21. März 2016</p>
<p>Verbesserte Amazon Redshift Redshift-Datenschemakonvertierung</p>	<p>Diese Version von Amazon ML verbessert die Konvertierung von Amazon Redshift (Amazon Redshift) -Datenschemas in Amazon ML-Datenschemas.</p> <p>Weitere Informationen zur Verwendung von Amazon Redshift mit Amazon ML finden Sie unter Erstellen einer Amazon ML-Datenquelle aus Daten in Amazon Redshift aus.</p>	<p>9. Februar 2016</p>
<p>CloudTrail-Protokollierung hinzugefügt</p>	<p>Diese Version von Amazon ML ermöglicht das Protokollieren von Anfragen mit AWS CloudTrail (CloudTrail).</p> <p>Weitere Informationen zur Verwendung der CloudTrail-Protokollierung finden Sie unter Protokollieren von Amazon-ML-API-Aufrufen mit AWS CloudTrail aus.</p>	<p>10. Dezember 2015</p>
<p>Zusätzliche DataRearrangement-Optionen hinzugefügt</p>	<p>Diese Version von Amazon ML ermöglicht die zufällige Aufteilung Ihrer Eingabedaten und die Erstellung ergänzender Datenquellen.</p> <p>Weitere Informationen zur Verwendung der <code>DataRearrangement</code> Parameter, siehe Neuordnung von Daten aus.</p> <p>Weitere Informationen zur Verwendung der neuen Optionen für die Kreuzvalidierung finden Sie unter Kreuzvalidierung.</p>	<p>3. Dezember 2015</p>

Änderung	Beschreibung	Änderungsdatum
Testen von Echtzeitvoraussagen	<p>Diese Version von Amazon ML ermöglicht das Testen von Echtzeitvoraussagen in der Servicekonsole.</p> <p>Weitere Informationen zum Testen von Echtzeitvoraussagen finden Sie unter Anfordern von Echtzeitvoraussagen im Amazon Machine Learning Learning-Entwicklerhaus.</p>	19. November 2015
Neue Region	<p>Diese Version von Amazon ML ermöglicht Unterstützung für die Region EU (Irland).</p> <p>Weitere Informationen zu Amazon ML in der Region EU (Irland) finden Sie unter Regionen und Endpunkte im Amazon Machine Learning Learning-Entwicklerhaus.</p>	20. August 2015
Erstversion	Dies ist die erste Version des Amazon ML-Entwicklerhandbuchs.	9. April 2015