aws

开发人员指南

Amazon Machine Learning

版本 Latest

Copyright © 2023 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon Machine Learning: 开发人员指南

Copyright © 2023 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon 的商标和商业外观不得用于任何非 Amazon 的商品或服务，也不得以任何可能引起客户混淆、贬低或诋毁 Amazon 的方式使用。所有非 Amazon 拥有的其它商标均为各自所有者的财产，这些所有者可能附属于 Amazon、与 Amazon 有关联或由 Amazon 赞助，也可能不是如此。

Table of Contents

我们不再更新 Amazon Machine Learning 服务，也不再接受新用户使用该服务。本文档可供现有用户使用，但我们不会再对其进行更新。有关更多信息，请参阅什么是 Amazon Machine Learning。

什么是 Amazon Machine Learning？

我们不再更新 Amazon Machine Learning (Amazon ML) 服务，也不再接受新用户使用该服务。本文档可供现有用户使用，但我们不会再对其进行更新。

AWS 现在提供基于云的稳健服务，即 Amazon SageMaker，能够让各种技能水平的开发人员都能使用机器学习技术。SageMaker 是一项完全托管式机器学习服务，能够帮助您创建强大的机器学习模型。借助 SageMaker，数据科学家和开发人员可以构建和训练机器学习模型，然后直接将模型部署到生产就绪托管环境中。

有关更多信息，请参阅 SageMaker 文档。

主题

Amazon Machine Learning 关键概念

本部分总结了以下关键概念并详细介绍了如何在 Amazon ML 中使用这些概念：

- 数据源包含与 Amazon ML 输入数据相关的元数据
- ML 模型使用从输入数据中提取的模式生成预测
- 评估衡量 ML 模型的质量
- 批量预测可异步 为多个输入数据观察生成预测
- 实时预测可同步 为单个数据观察生成预测

数据源

数据源是包含有关输入数据的元数据的对象。Amazon ML 读取您的输入数据、计算其属性的描述性统计数据，并将统计数据与架构和其他信息一起存储为数据源对象的一部分。接下来，Amazon ML 使用数据源训练和评估 ML 模型并生成批量预测。

> **Important**
>
> 数据源不存储输入数据的副本。而是存储输入数据所在的 Amazon S3 位置的引用。如果您移动或更改 Amazon S3 文件，Amazon ML 无法访问或使用该文件来创建 ML 模型、生成评估或生成预测。

下表定义了与数据源相关的术语。

期限	定义
属性	观察中唯一的指定属性。在采用表格格式的数据（例如，电子表格或逗号分隔的值 (CSV) 文件）中，列标题代表属性，而行包含每个属性的值。 同义词：变量、变量名称、字段、列
数据源名称	（可选）允许您为数据源定义一个便于阅读的名称。这些名称便于您在 Amazon ML 控制台中查找和管理您的数据源。
输入数据	数据源引用的所有观察的总称。
位置	输入数据的位置。目前，Amazon ML 可以使用存储在 Amazon S3 存储桶、Amazon Redshift 数据库或 Amazon Relational Database Service (RDS) 中的 MySQL 数据库中的数据。
观察	单个输入数据单位。例如，如果您创建的是检测欺诈交易的 ML 模型，您的输入数据将包含许多观察，每个观察表示单个交易。 同义词：记录、示例、实例、行
行 ID	（可选）此标记（如果指定）用于标识输入数据的将包含在预测输出中的属性。借助此属性，您可以更轻松地将预测与对应的观察进行关联。 同义词：行标识符
架构	解释输入数据时所需的信息，包括属性名及其分配的数据类型和特殊属性名。
统计数据	输入数据中每个属性的摘要统计信息。这些统计数据有两种用途：

期限	定义
	Amazon ML 控制台将以图形方式显示这些数据，以帮助您了解您的数据概况和识别违规行为或错误。 Amazon ML 在训练过程中使用它们来改进生成的 ML 模型的质量。
状态	指示数据源的当前状态，例如正在进行、已完成或失败。
目标属性	在训练 ML 模型的上下文中，目标属性会标识包含“正确”答案的输入数据中属性的名称。Amazon ML 使用此属性在输入数据中发现模式并生成 ML 模型。在评估和生成预测的上下文中，目标属性是值将由经过训练的 ML 模型进行预测的属性。 同义词：目标

ML 模型

ML 模型是通过在数据中查找模式来生成预测的数学模型。Amazon ML 支持三种类型的 ML 模型：二进制分类、多类别分类和回归。

下表定义了与 ML 模型相关的术语。

期限	定义
回归	训练回归 ML 模型的目标是预测数字值。
多类别	训练多类别 ML 模型的目标是预测属于有限的、预定义的允许值集的值。
二进制	训练二进制 ML 模型的目标是预测只能有两种状态之一的值，例如 true 或 false。
模型大小	ML 模型可以捕获和存储模式。ML 模型存储的模式越多，就会变得越大。ML 模型的大小以 MB 为单位。
扫描次数	训练 ML 模型时，您可以使用数据源中的数据。在学习过程中多次使用每个数据记录有时比一次使用更加有用。您让 Amazon ML 使用相同数据记录的次数称为扫描次数。

期限	定义
正则化	正则化是一种机器学习方法，可用来获得更高质量的模型。Amazon ML 提供适用于大多数情况的默认设置。

评估

评估可衡量您的 ML 模型的质量，并确定它是否表现良好。

下表定义了与评估相关的术语。

期限	定义
模型洞察	Amazon ML 会为您提供一个指标和许多洞察，您可以用这些来评估模型的预测性能。
AUC	ROC 曲线下面积 (AUC) 测量二进制 ML 模型为正面示例预测比负面示例更高分数的能力。
宏平均 F1 分数	宏平均 F1 分数用于评估多类别 ML 模型的预测性能。
RMSE	均方根误差 (RMSE) 是用于评估回归 ML 模型的预测性能的指标。
截断	ML 模型通过生成数字预测分数来工作。通过应用截断值，系统可将这些分数转换为 0 和 1 标签。
准确度	准确度可测量正确预测的百分比。
精度	精度显示在已检索的实例（预测为阳性）中，实际阳性实例（相对于假阳性）的百分比。换言之，所选项目有多少是阳性？
召回率	召回率显示了在相关实例总数中实际阳性的百分比（实际阳性）。换言之，阳性项目有多少已选定？

批量预测

批量预测功能可以一次性运行一组观察。这非常适合于没有实时要求的预测分析。

下表定义了与批量预测相关的术语。

期限	定义
输出位置	批量预测结果存储在 S3 存储桶输出位置。
清单文件	此文件将每个输入数据文件与其关联的批量预测结果相关联。它存储在 S3 存储桶输出位置。

实时预测

实时预测适用于具有低延迟要求的应用程序，例如交互式 Web、移动或桌面应用程序。任何 ML 模型都可通过低延迟实时预测 API 查询预测。

下表定义了与实时预测相关的术语。

期限	定义
实时预测 API	实时预测 API 接受请求负载中的单个输入观察并在响应中返回预测。
实时预测终端节点	要将使用 ML 模型与实时预测 API 配合使用，您需要创建实时预测终端节点。创建后，此终端节点包含可用来请求实时预测的 URL。

访问 Amazon Machine Learning

您可以使用以下任何方式访问 Amazon ML：

Amazon ML 控制台

要访问 Amazon ML 控制台，您可以登录 AWS 管理控制台并打开位于以下位置的 Amazon ML 控制台：https://console.aws.amazon.com/machinelearning/。

AWS CLI

有关如何安装和配置 AWS CLI 的信息，请参阅 AWS Command Line Interface 用户指南中的使用 AWS 命令行界面进行设置。

Amazon ML API

有关 Amazon ML API 的更多信息，请参阅 Amazon ML API 参考。

AWS 软件开发工具包

有关 AWS 软件开发工具包的更多信息，请参阅用于 Amazon Web Services 的工具。

区域和终端节点

Amazon Machine Learning (Amazon ML) 支持以下两个区域的实时预测终端节点：

区域名称	区域	终端节点	协议
美国东部（弗吉尼亚州北部）	us-east-1	machinelearning.us-east-1.amazonaws.com	HTTPS
Europe (Ireland)	eu-west-1	machinelearning.eu-west-1.amazonaws.com	HTTPS

您可在任何区域托管数据集、训练和评估模型以及触发预测。

我们建议您将您的所有资源保留在同一区域中。如果您的输入数据与您的 Amazon ML 资源位于不同区域，您将会产生跨区域数据传输费用。您可以从任何区域调用实时预测终端节点，但从不含某个终端节点的区域调用该终端节点可能会影响实时预测延迟。

Amazon ML 的定价

使用 AWS 服务时，可以按实际用量付费。无最低费用，无预先承诺。

Amazon Machine Learning (Amazon ML) 将按小时对计算数据统计以及训练和评估模型所花的时间计费，随后您按照该程序为您的应用程序所生成的预测数量付费。对于实时预测，您也将基于模型大小按小时支付预留容量费用。

Amazon ML 仅估算 Amazon ML 控制台中的预测的成本。

有关 Amazon ML 定价的更多信息，请参阅 Amazon Machine Learning 定价。

主题

估算批量预测成本

当您使用“创建批量预测”向导请求使用 Amazon ML 模型进行批量预测时，Amazon ML 会估算这些预测的成本。计算估算成本的方法因可用的数据类型而异。

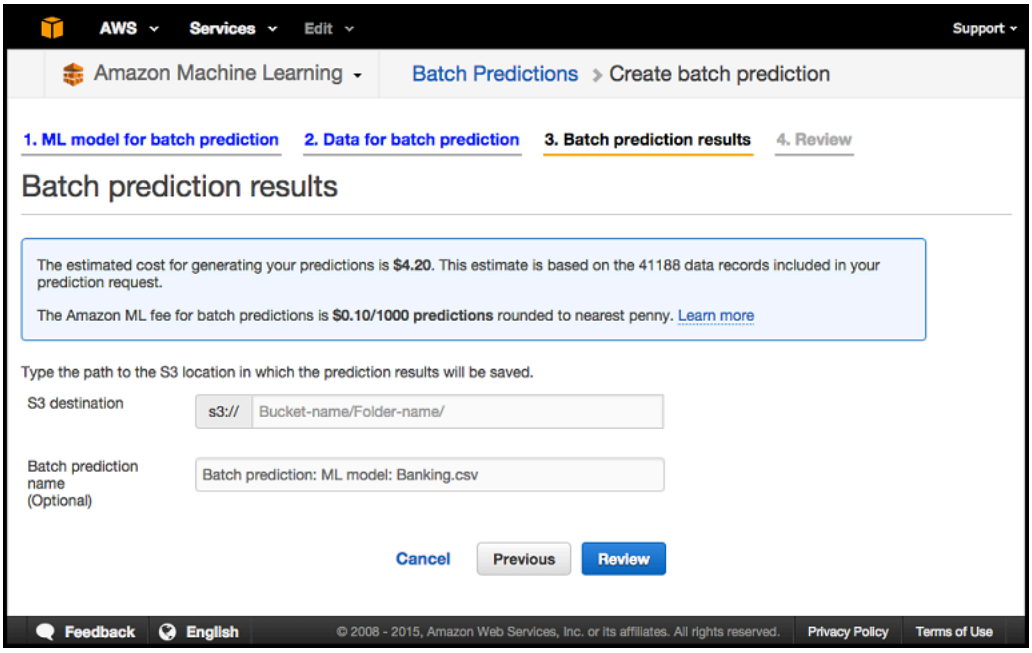
在数据统计可用时估算批量预测成本

在 Amazon ML 计算了用于请求预测的数据源的汇总统计数据时，获得的估算成本最准确。系统始终会为使用 Amazon ML 控制台创建的数据源计算这些统计数据。如果使用 CreateDataSourceFromS3、CreateDataSourceFromRedshift 或 CreateDataSourceFromRDS API 以编程方式创建数据源，API 用户必须将 `ComputeStatistics` 标记设置为 `True`。数据源必须处于 `READY` 状态才能使用统计数据。

Amazon ML 计算的其中一个统计数据是数据记录的数量。当数据记录数量可用时，Amazon ML 的“创建批量预测”向导会估算预测结果的数量，具体方法是用数据记录的数量乘以批量预测的费用。

您的实际成本可能由于以下原因而与此估算成本有所不同：

- 部分数据记录可能处理失败。对于使用失败的数据记录提供的预测，您不会支付任何费用。
- 估算时未考虑 AWS 预设的服务抵扣金额或应用的其他调整。



在只有数据大小可用时估算批量预测成本

当您请求的批量预测和请求数据源的数据统计均不可用时，Amazon ML 会根据以下项估算成本：

- 在数据源验证期间计算并保存的总数据大小
- 数据记录的平均大小，Amazon ML 通过读取和分析数据文件的前 100MB 来估算该大小

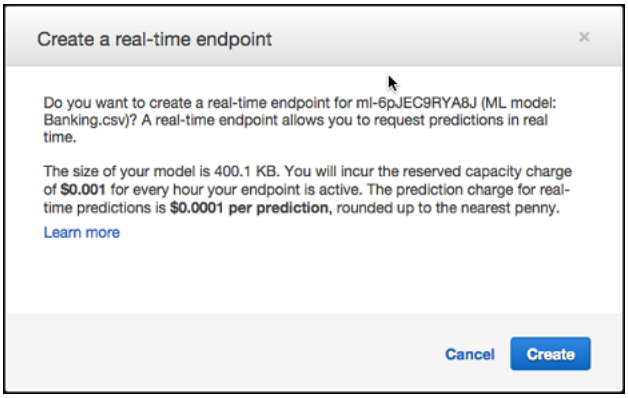
为了估算批量预测的成本，Amazon ML 会用总数据大小除以数据记录的平均大小。这种成本预测方法的准确性不如数据记录数量可用时使用的方法，因为数据文件的第一个记录可能无法准确体现平均记录大小。

在数据统计或数据大小都不可用时估算批量预测成本

当数据统计或数据大小都不可用时，Amazon ML 无法估算批量预测成本。当您用于请求批量预测的数据源尚未通过 Amazon ML 的验证时，通常会出现这种情况。当您创建了基于 Amazon Redshift (Amazon Redshift) 或 Amazon Relational Database Service (Amazon RDS) 查询的数据源，并且数据传输尚未完成时，或者数据源创建请求排在您账户中的其他操作之后，可能形成上述条件。在这种情况下，Amazon ML 控制台会通知您有关批量预测费用的信息。您可以选择继续处理批量预测请求而不估算成本，或在用于预测的数据源处于 INPROGRESS 或 READY 状态后取消向导并返回。

估算实时预测成本

当您使用 Amazon ML 控制台创建实时预测终端节点时，系统将向您显示估算的预留容量费用，用于预测处理的预留终端节点将持续收取此项费用。根据服务定价页面的说明，此项费用因模型大小而异。您还将了解标准 Amazon ML 实时预测费用。



机器学习概念

机器学习 (ML) 可以帮您利用历史数据制定更好的业务决策。ML 算法在数据中发现模式，并使用这些模式构建数学模型。然后，您可以使用这些模型对未来数据做出预测。例如，机器学习模型可能应用于根据客户过去的行为来预测客户购买某个特定产品的可能性。

主题

使用 Amazon Machine Learning 解决业务问题

您可以使用 Amazon Machine Learning 对您已有实际答案示例的问题应用机器学习。例如，如果您要使用 Amazon Machine Learning 预测电子邮件是否为垃圾邮件，您需要收集正确标记为垃圾邮件或非垃圾邮件的电子邮件示例。然后，您可以使用机器学习从这些电子邮件进行泛化，预测新电子邮件有多大可能为垃圾邮件。这种从已经标记了实际答案的数据进行学习的方法，称为受监管机器学习。

您可以为这些特定机器学习任务使用受监管 ML 方法：二进制分类 (预测两个可能结果之一)、多类别分类 (预测两个以上结果之一) 以及回归 (预测数值)。

二进制分类问题的示例：

- 客户会购买此产品还是不会购买此产品？
- 电子邮件是否为垃圾邮件？
- 此产品是一本书还是农场动物？
- 此评论是客户还是机器人写的？

多类别分类问题的示例：

- 此产品是书、电影还是服装？
- 这是浪漫喜剧、纪录片还是惊悚片？
- 此客户最感兴趣什么类别的产品？

回归分类问题的示例：

- 西雅图明天的温度是多少？
- 此产品将销售多少件？
- 此客户还有多少天停止使用该应用程序？
- 这套房屋将以什么价格出售？

何时使用机器学习

请务必记住，ML 并不是适用于所有问题类型的解决方案。在一些特定案例中，无需使用 ML 技术即可开发可靠的解决方案。例如，如果您可以使用简单规则、计算或预先确定的步骤来确定目标值，而这些方法可以通过编程完成而不需要任何数据驱动的学习，则无需 ML。

机器学习可用于以下情况：

- 无法编码规则：许多人工任务（例如识别电子邮件为垃圾邮件还是非垃圾邮件）无法使用基于规则的简单（确定性）解决方案妥善解决。影响答案的因素可能会有很多。如果规则取决于太多因素，并且其中众多规则重叠或者需要非常精细地调整，这直接就使得难于通过人力来准确编码规则。您可以使用 ML 有效解决这个问题。
- 无法扩展：您也许可以手动识别几百封电子邮件并确定是否为垃圾邮件。但是，在面对数百万封电子邮件时，此任务变得庞杂乏味。ML 解决方案可以有效处理大规模问题。

构建机器学习应用程序

构建 ML 应用程序是涉及到一系列步骤的迭代过程。要构建 ML 应用程序，请执行以下常规步骤：

1. 在所要观察的对象以及您希望模型预测的答案方面，为核心 ML 问题确定框架。
2. 收集、清除和准备数据，以使其适合 ML 模型训练算法使用。可视化和分析数据来运行健全性检查以验证数据的质量和了解数据。
3. 通常，原始数据（输入变量）和答案（目标）以不能用于训练高度预测性模型的方式表示。因此，您通常应尝试从原始变量构造预测性更高的输入表示形式或特征。
4. 将生成的特征提供给学习算法用于构建模型，并根据从模型构建中给出的数据来评估模型的质量。
5. 使用模型生成新数据实例的目标答案的预测。

系统地阐述问题

机器学习的第一步是确定您希望预测的内容，这称为标签答案或目标答案。假设一个您要制造产品的场景，但您制造每个产品的决策取决于潜在的销售量。在此场景中，您希望预测每个产品将被购买的次数（预测的销售量）。可以使用机器学习以多种方法定义此问题。根据您的使用案例或业务需求选择如何定义问题。

您是否希望预测您的客户将对每个产品进行的购买数（这种情况下目标是数字，您要解决回归问题）？或者，您是否只想预测哪些产品将获得 10 次以上的购买（在这种情况下，目标为二进制，您要解决二进制分类问题）？

重要的是避免过度复杂化问题并制定满足您需求的最简单解决方案。但是，避免丢失信息（特别是历史答案中的信息）也很重要。在这里，将过去的实际销售数量转换为二进制变量“over 10”与“fewer”时，会丢失有价值的信息。投入一些时间来决定预测哪个目标对您最有意义，可以帮助您避免构建无法回答您问题的模型。

收集标记的数据

ML 问题从数据开始 - 最好是您已知道其目标答案的大量数据（示例或观察）。已知其目标答案的数据称为标记的数据。在受监管 ML 中，算法教育自身从我们提供的标记的示例进行学习。

您的数据中的每个示例/观察必须包含两个元素：

- 目标 - 您要预测的答案。您向 ML 算法提供标记为目标（正确答案）的数据以从中学习。然后，您将使用经过训练的 ML 模型，对您不知道目标答案的数据来预测此答案。
- 变量/特征 - 这些示例属性可用于识别要预测目标答案的模式。

例如，对于电子邮件分类问题，目标是指示电子邮件是否为垃圾邮件的标签。变量示例是电子邮件发件人、电子邮件正文中的文本、主题行中的文本、电子邮件的发送时间以及发件人和收件人之间是否存在以前的通信信息。

通常，数据并不是使用已标记的形式提供。收集并准备变量和目标通常是解决 ML 问题最重要的步骤。示例数据应为在您使用模型进行预测时具有代表性的数据。例如，如果您希望预测电子邮件是否为垃圾邮件，您必须为机器学习算法收集阳性（垃圾电子邮件）和阴性（非垃圾电子邮件），这样才能查找可以区别两种类型电子邮件的模式。

在您具有标记的数据之后，您可能需要将它转换为您的算法或软件可接受的格式。例如，要使用 Amazon ML，您需要将数据转换为逗号分隔 (CSV) 格式，每个示例组成 CSV 文件的一行，每列包含一个输入变量，并且有一列包含目标答案。

分析数据

在您将标记的数据提供给 ML 算法之前，最佳实践是检查您的数据以发现问题，并获得有关您所使用的数据的见解。您的模型的预测能力与您提供的数据相关。

分析数据时，您应记住以下注意事项：

- 变量和目标数据摘要 - 了解您的变量获取的值以及哪些值在数据中是主要的，这非常重要。您可以让您希望解决的问题的主题专家来运行这些摘要。向自己或主题专家提问：数据是否符合您的预期？ 您是否可能有数据收集问题？ 您的目标中的一个类别是否比另一个类别更频繁？ 是否有比预期更多的缺失值或无效数据？
- 变量-目标关联 - 了解各个变量和目标类别之间的相关性会有帮助，因为高相关性表示变量和目标类别之间有关系。一般而言，您希望包括具有高相关性的变量，因为这些是具有更高预测能力（信号）的变量，忽略低相关性的变量，因为它们可能不相关。

在 Amazon ML 中，您可以通过创建数据源和检查生成的数据报告来分析数据。

特征处理

在通过数据摘要和可视化了解您的数据之后，您可能希望进一步转换变量以使其更有意义。这称为特征处理。例如，假如您有一个变量，用于捕获发生事件时的日期和时间。此日期和时间永远不会再次出现，因此对于预测您的目标不会有用。但是，如果此变量转换表示一天中几点、一周中日期和月份的特征，这些变量会非常有用，用于了解在特定小时、工作日或月份发生事件的趋势。这种特征处理用于形成可以从中学习的更可概括的数据点，提供对预测模型的显著改进。

其他常见特征处理的示例：

- 使用更有意义的值替换缺失或无效的数据（例如，如果您知道某个产品类型变量的缺失值实际上意味着图书，您随后可以使用图书的值替换该产品类型中的所有缺失值）。处理缺失值的一种常用策略是使用平均值或中值替换缺失值。重要的是在选择用于替换缺失值的策略之前了解您的数据。
- 将一个变量与另一个变量构成笛卡尔积。例如，如果您有两个变量，例如人口密度（urban、suburban、rural）和州（Washington、Oregon、California），通过将这两个变量的笛卡尔积构成一个特征（urban_Washington、suburban_Washington、rural_Washington、urban_Oregon、suburban_Oregon、ru可以从该特征中获得有用信息。
- 非线性转换，例如分箱数值变量转换为分类。在许多情况下，数值特征与目标之间的关系并非线性（特征值不随目标单增或单）。在这种情况下，将数值特征分箱到表示数值特征不同范围的分类特

征中可能会很有用。然后，每个分类特征（分箱）可以建模为具有与目标的自身线性关系。例如，假设您知道连续数值特征 age 与购买某书的可能性并非线性相关。您可以将寿命分箱到可以更准确捕获与目标关系的分类特征。数值变量的最佳分箱数量取决于变量的特性及其与目标的关系，最好通过试验来确定。Amazon ML 建议，基于建议配方中的数据统计信息确定数值特征的最佳分箱数量。有关建议配方 的详细信息，请参阅开发人员指南。

- 域特有的特征（例如，您有长度、宽度和高度作为单独的变量；您可以创建一个新的体积特征作为这三个变量的积）。
- 变量特有的特征。一些变量类型，例如文本特征、捕获网页结构的特征或者句子结构的特征具有通用的处理方式，可以帮助提取结构和上下文。例如，从文本“the fox jumped over the fence”构成 n 元 可以使用一元 表示：the、fox、jumped、over、fence，或者使用二元：the fox、fox jumped、jumped over、over the、the fence。

包括更多相关特征有助于提高预测能力。显然，并不总是有可能预先知道具有“信号”或预测影响力的特征。因此，最好包括所有可能与目标标签相关的特征，并让模型训练算法选取具有最强相关性的特征。在 Amazon ML 中，创建模型时可以在配方中指定特征处理。有关可用特征处理器的列表，请参阅开发人员指南。

将数据拆分为训练数据和评估数据

ML 的基本目标是在用于训练模型的数据实例之外归纳。我们希望评估模型来估算其模式针对未用于训练模型的数据的归纳质量。但是，由于未来的实例具有未知的目标值，并且我们无法立即检查预测对未来实例的准确性，我们需要使用一些现在已知答案的数据来用作未来数据的代理。使用已用于训练的相同数据评估模型并没有用处，因为它会奖励可以“记住”训练数据的模型，而不是通过它进行归纳。

一种常见策略是获取所有可用的标签数据，将它拆分为训练和评估子集，通常比例为 70-80% 的数据用于训练，20-30% 用于评估。ML 系统使用训练数据训练模型来查看模式，并使用评估数据来评估训练模型的预测质量。ML 系统使用多种指标，将评估数据集的预测与真实值（称为基本实际情况）进行比较来评估预测性能。通常情况下，您可以针对评估子集使用“最佳”模型来预测您不知道目标答案的未来实例。

Amazon ML 会将通过 Amazon ML 控制台发送用于训练模型的数据拆分 70% 用于训练，30% 用于评估。默认情况下，Amazon ML 将前 70% 的输入数据按照在源数据中的显示顺序用于训练数据源，将剩余的 30% 数据用于评估数据源。Amazon ML 还允许您随机选择源数据的 70% 用于训练，而不是使用前 70%，并使用此随机子集的补充进行评估。您可以使用 Amazon ML API 来指定自定义拆分比率并提供在 Amazon ML 外部拆分的训练和评估数据。Amazon ML 还提供了拆分数据的策略。有关拆分策略的更多信息，请参阅拆分数据。

训练模型

现在，您已准备好提供 ML 算法（即，学习算法）与训练数据。算法从将变量映射到目标的训练数据模式学习，并输出捕获这些关系的模型。然后可以使用 ML 模型获取您不知道目标答案的新数据的预测。

线性模型

有大量的 ML 模型可用。Amazon ML 学习一种类型的 ML 模型：线性模型。术语“线性模型”意味着模型被指定为线性特征的组合。根据训练数据，学习过程会计算每个特征的一个权重，用于形成一个可以预测或估算目标值的模型。例如，如果您的目标是客户购买的保险金额，您的变量是年龄和收入，一个简单的线性模型将如下所示：

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

学习算法

学习算法的任务是了解模型的权重。权重描述了模型学习的模式反映数据中实际关系的可能性。学习算法包含损失函数和优化技术。损失是 ML 模型提供的目标估算与目标不精确相等时产生的惩罚。损失函数将此惩罚量化为单个值。优化技术旨在最大程度地减少损失。在 Amazon Machine Learning 中，我们使用三个损失函数，每个函数对应一种类型的预测问题。Amazon ML 中使用的优化技术是在线随机梯度下降 (SGD)。SGD 对训练数据进行连续扫描，在每次扫描中，一次一个示例地更新特征权重，其目标是达到能最大程度减少损失的最佳权重。

Amazon ML 使用以下学习算法：

- 对于二进制分类，Amazon ML 使用逻辑回归（逻辑损失函数 + SGD）。
- 对于多类别分类，Amazon ML 使用多项逻辑回归（多项逻辑损失 + SGD）。
- 对于回归，Amazon ML 使用线性回归（平方值损失函数 + SGD）。

训练参数

Amazon ML 学习算法接受称为超级参数或训练参数的参数，使您可以控制生成模型的质量。Amazon ML 为各超级参数自动选择设置或提供静态默认值，具体做法取决于是哪个超级参数。尽管默认超级参数设置通常会生成有用的模型，但是您可以通过更改超级参数值来改进模型的预测性能。以下部分介绍与线性模型的学习算法关联的常见超级参数，例如由 Amazon ML 创建的线性模型。

学习速率

学习速率是在随机梯度下降 (SGD) 算法中使用的常量值。学习速率影响算法达到（收敛到）最佳权重的速度。SGD 算法对所看到的每个数据示例的线性模型的权重进行更新。这些更新的大小由学习速率

控制。太大的学习速率可能会妨碍权重达到最佳解决方案。太小的值会导致算法需要多次扫描才能达到最佳权重。

在 Amazon ML 中，学习速率是根据您的数据自动选择的。

模型大小

如果您有许多输入特征，数据中的可能模式的数量会导致大型模型。大型模型具有实际影响，例如在训练和生成预测时需要更多的 RAM 来存储模型。在 Amazon ML 中，您可以通过使用 L1 正则化来减小模型的大小，或者通过指定最大大小来专门限制模型大小。请注意，如果您将模型大小减小太多，可能会降低模型的预测能力。

有关默认模型大小的更多信息，请参阅训练参数：类型和默认值。有关正则化的更多信息，请参阅正则化。

扫描次数

SGD 算法连续扫描训练数据。`Number of passes` 参数控制算法扫描训练数据的次数。扫描次数较多会得到数据拟合更好的模型（如果学习速率不是太大），但随着扫描次数的增加，好处将减少。对于较小的数据集，您可以大幅提高扫描次数，这使得学习算法可以有效地更加紧密地拟合数据。对于特大型数据集，一次扫描可能已足够。

有关默认扫描次数的信息，请参阅训练参数：类型和默认值。

数据随机排序

在 Amazon ML 中，您必须对数据随机排序，因为 SGD 算法受训练数据中行的顺序影响。将训练数据随机排序会得到更好的 ML 模型，因为它有助于避免 SGD 算法针对它看到的第一类数据优化解决方案，而不是针对完整范围的数据进行优化。随机排序会打乱数据的顺序，这样 SGD 算法就不会在太多连续观察中遇到同一种类型的数据。如果它在许多连续权重更新中只看到一种类型的数据，该算法可能无法更正新数据类型的模型权重，因为更新可能太大。此外，当数据不是随机提供时，算法很难快速找到针对所有数据类型的最佳解决方案，在一些情况下，算法可能永远找不到最佳解决方案。将训练数据随机排序有助于算法在最佳解决方案上更快地收敛。

例如，假如您要训练 ML 模型以预测产品类型，而您的训练数据包括电影、玩具和视频游戏产品类型。如果您在将数据上传到 Amazon S3 前将按产品类型列对数据进行排序，则算法会看到按产品类型字母顺序排列的数据。该算法将先查看所有电影数据，然后您的 ML 模型开始学习电影的模式。随后在模型遇到玩具数据时，该算法的每个更新都会向玩具产品类型拟合模型，即使这些更新会让拟合电影的模式降级也是如此。这种从电影到玩具类型的突然转变，可能会生成不了解如何准确预测产品类型的模型。

有关默认随机排序类型的信息，请参阅训练参数：类型和默认值。

正则化

正则化通过惩罚极端权重值来帮助防止线性模型过度拟合训练数据示例（即记住模式而不是归纳模式）。L1 正则化具有减少模型中使用的特征数量的效果，其方法是将具有很小权重的特征的权重推向零。因此，L1 正则化会生成稀疏模型并降低模型中的噪音量。L2 正则化会生成较小的总体权重值，可在输入特征的相关性高的情况下稳定权重。您可以使用 `Regularization type` 和 `Regularization amount` 参数控制应用的 L1 或 L2 正则化的量。极大的正则化值会导致所有特征具有零权重，阻止模型学习模式。

有关默认正则化值的信息，请参阅 训练参数：类型和默认值。

评估模型准确度

ML 模型的目标是学习模式，可以针对未见过的数据很好地归纳，而不只是记住在训练过程中查看的数据。在您拥有模型之后，务必使用没有用于训练模型的未见过的示例，检查模型在其上是否表现良好。为进行此操作，您可使用模型预测评估数据集的答案（保存数据），然后将预测目标与实际答案（基本实际情况）对比。

ML 中使用一系列指标来衡量模型的预测准确度。准确度指标的选择取决于 ML 任务。务必检查这些指标以确定您的模型是否运行良好。

二元分类

许多二进制分类算法的实际输出是预测分数。这些分数指示系统的给定观察属于正类的确定性。作为此分数的使用者，为了决定观察应分类为正还是负，需要选取分类阈值（截断值），并与分数进行对比，以此来解释分数。然后，任何分数高于阈值的观察将视为正类，分数低于阈值的观察预测为负类。

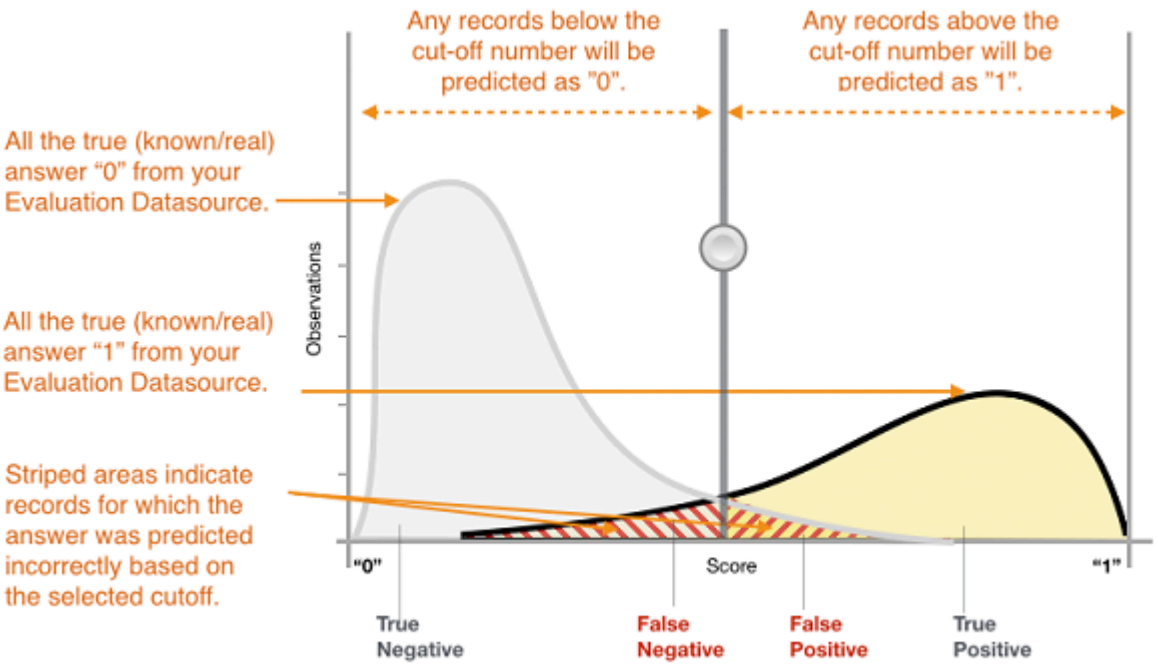


图 1：二进制分类模型的分数分布

现在，预测根据实际已知答案和预测答案分类为四组：正确正向预测（真阳性）、正确负向预测（真阴性）、错误正向预测（假阳性）和错误负向预测（假阴性）。

二进制分类准确性指标量化两种类型的正确预测和两种类型的错误预测。典型指标是准确性 (ACC)、精度、假阳性比率和 F1 度量。每个指标度量预测模型的不同方面。准确度 (ACC) 衡量正确预测的比率。精度衡量实际正向示例与预测为正向示例的比率。查全率衡量预测有多少实际正向示例预测为正向示例。F1 衡量精度和查全率的调和平均数。

AUC 是不同类型的指标。它衡量模型为正向示例预测出相比负向示例更高分数的能力。由于 AUC 独立于所选阈值，因此您可以从 AUC 指标感受到模型的预测性能，无需选取阈值。

根据您的业务问题，您可能会对在这些指标的特定部分中表现良好的模型更感兴趣。例如，两个业务应用程序可能对其 ML 模型具有迥然不同的需求：

- 一个应用程序可能需要严格保证正向预测实际是正向的（高精度），并能够承受将一些正向示例错误分类为负向（中等查全率）。
- 另一个应用程序可能需要尽可能多地预测正向示例（高查全率），并可以接受将一些负向示例错误分类为正向（中等精度）。

在 Amazon ML 中，观察得到的预测分数在范围 [0,1] 中。用于做出将示例分类为 0 或 1 的决策的分数阈值默认情况下设置为 0.5。Amazon ML 允许您查看选择不同分数阈值的含义，并允许您选取符合业务需求的合适阈值。

多类别分类

与二进制分类问题的处理不同，您不需要选择分数阈值以进行预测。预测的答案是预测分数最高的类（即标签）。在某些情况下，您可能希望仅当预测具有高分数才使用预测的答案。在这种情况下，您可以根据您是否接受答案来选择预测分数的阈值。

多类别中使用的典型指标与二进制分类案例中使用的指标相同。通过在将所有其他类别分组为属于第二个类别之后，将其作为二进制分类问题来处理，为每个类别计算指标。然后，在所有类别上对二进制指标取平均值以获取宏平均（相同处理每个类别）或加权平均（按类别频率加权）指标。在 Amazon ML 中，宏平均 F1 度量用于评估多类别分类器的预测成功性。

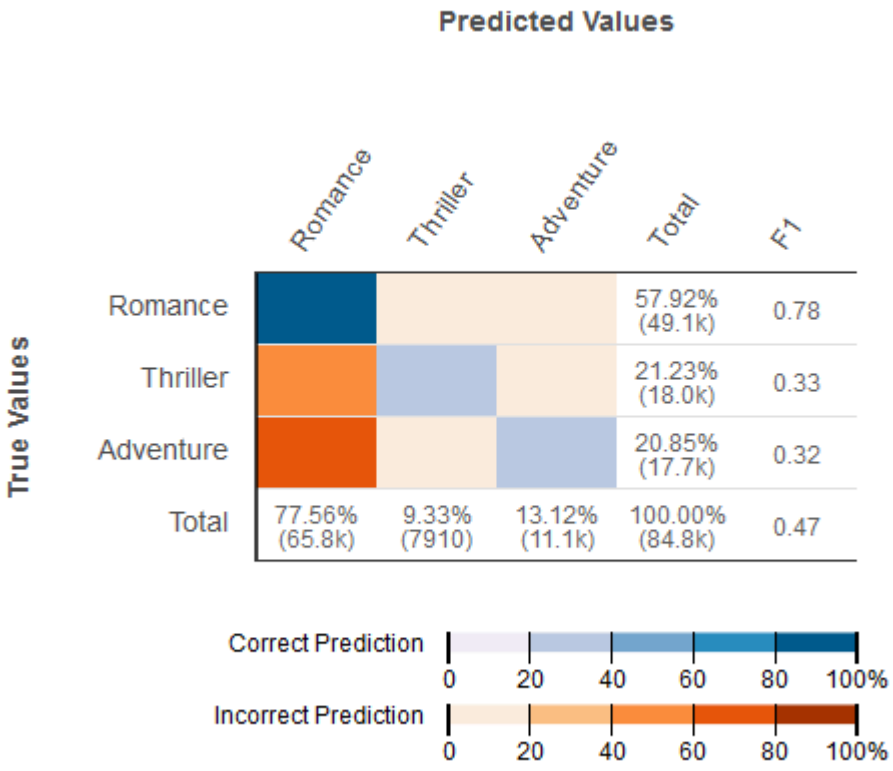


图 2：多类别分类模型的混淆矩阵

查看多类别问题的混淆矩阵会非常有帮助。混淆矩阵是一个表，其中显示了评估数据中的各个类以及正确预测和不正确预测的数量或百分比。

回归

对于回归任务，典型的准确性指标是均方根误差 (RMSE) 和平均绝对百分比误差 (MAPE)。这些指标测量预测数值目标与实际数值答案（基本实际情况）之间的差距。在 Amazon ML 中，RMSE 指标分数用于评估回归模型的预测准确性。

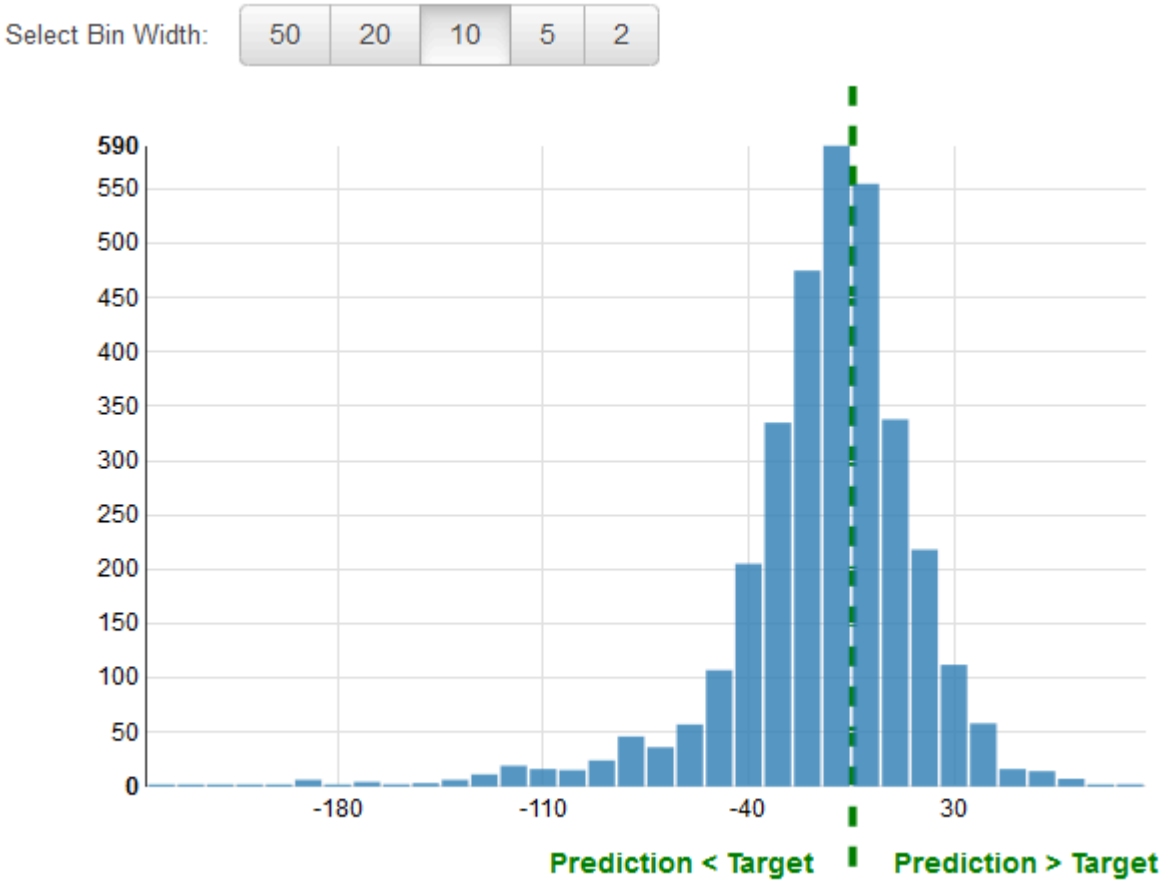


图 3：回归模型的残差分布

对于回归问题，常见的做法是检查残差。评估数据中某个观察的残差是真实目标与预测目标之间的差值。残差表示模型无法预测的目标部分。正残差表示模型低估了目标（实际目标大于预测目标）。负残差表示高估（实际目标小于预测目标）。评估数据残差的直方图在呈钟形分布并且中心在零上时，指示模型以随机方式产生错误，不会系统性地高于或低于预测目标值的任何特定范围。如果残差未构成以零为中心的钟形曲线，这种情况表示模型的预测中存在结构错误。向模型添加更多变量可能会帮助模型捕获当前模型未捕获的模式。

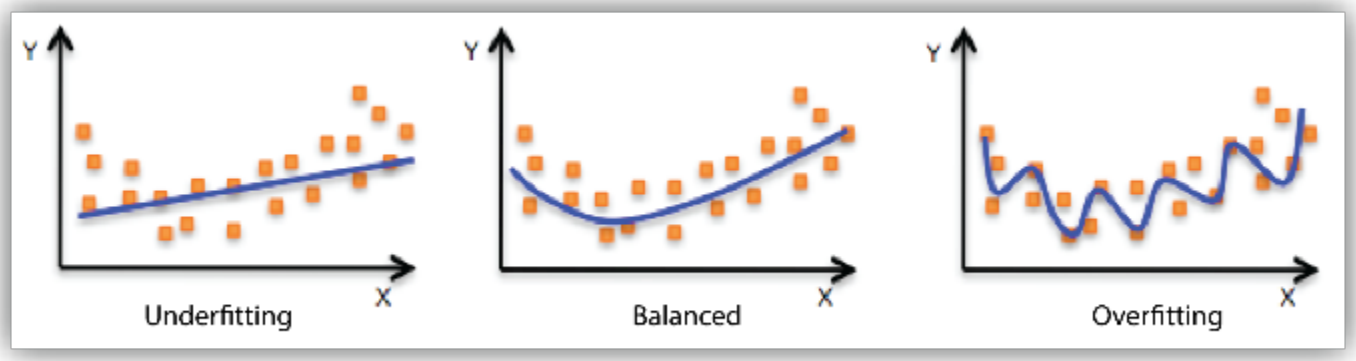
改进模型准确度

获取与您需求匹配的 ML 模型通常涉及到迭代此 ML 进程并尝试一些变化。您可能不会在第一次迭代时获得非常有预测性的模型，或者您可能希望改进模型以获得更好的预测。为了提高性能，您可以迭代以下步骤：

1. 收集数据：增加训练示例数
2. 特征处理：添加更多变量和更好的特征处理
3. 模型参数优化：考虑您的学习算法使用的训练参数的替代值

模型拟合：欠拟合与过度拟合

了解模型拟合对于了解模型准确性欠佳至关重要。这些了解将引导您采取纠正步骤。我们可以通过查看训练数据和评估数据的预测错误来确定预测模型是欠拟合还是过度拟合。



模型在训练数据上性能糟糕时，您的模型欠拟合。这是因为模型无法捕获输入示例（通常称为 X）与目标值（通常称为 Y）之间的关系。当您看到模型在训练数据上表现良好但在评估数据上表现不好时，表示您的模型过度拟合。这是因为模型记住了曾看到的数据，但无法归纳为未看过的示例。

训练数据的性能欠佳可能是因为模型太简单（输入特征表述性不够）而无法很好地描述目标。可以通过提升模型灵活性来改进性能。要提高模型的灵活性，请尝试以下操作：

- 添加新的域特有特征和更多特征笛卡尔积，并更改特征处理所用的类型（例如，增加 n 元大小）
- 减少使用的正则化数量

如果您的模型过度拟合训练数据，合理的做法是采取措施来降低模型的灵活性。要减少模型的灵活性，请尝试以下操作：

- 特征选择：考虑使用更少的特征组合，减少 n 元大小，以及减少数字属性分箱的数量。
- 增加使用的正则化数量。

训练和测试数据的准确性可能很糟糕，因为学习算法没有足够的数据来学习。您可以通过以下操作提高性能：

- 增加训练数据示例的数量。
- 增加现有训练数据的扫描次数。

使用模型来进行预测

现在您已拥有良好执行的 ML 模型，您可以用它来进行预测。在 Amazon Machine Learning 中，有两种方法来使用模型进行预测：

批量预测

当您想要一次性为一组观察生成预测，然后对特定百分比或特定数量的观察采取操作时，批量预测非常有用。通常情况下，您对于此类应用程序没有低延迟要求。例如，当您想要决定将哪些客户作为某个产品广告活动目标的一部分时，您可以获得所有客户的预测分数，排序模型预测来确定哪些客户最有可能购买，然后可以定位最可能购买客户的前 5%。

在线预测

在线预测场景是您希望在低延迟环境中让每个示例独立于其他示例一对一生成预测的情况。例如，您可以使用预测来立即做出某个特定交易是否可能为欺诈性交易的决定。

在新数据上重新训练模型

要让模型预测得更准确，进行预测所用的数据必须具有与训练模型所用的数据相似的分布。由于预期数据分布会随着时间发生偏差，所以部署模型并不是一次性的练习，而是连续过程。一种很好的做法是，连续监视传入数据，在发现数据分布与原始训练数据分布有显著偏差时使用较新的数据重新训练模型。如果监控数据以检测数据分布更改的开销太大，一种更简单的策略是定期训练模型，例如，每天、每周或每月。要在 Amazon ML 中重新训练模型，您需要根据新的训练数据创建新模型。

Amazon Machine Learning 流程

下表描述了如何使用 Amazon ML 控制台执行本文档中概述的 ML 流程。

ML 流程	Amazon ML 任务
分析您的数据	要在 Amazon ML 中分析您的数据，请创建数据源并查看数据洞察页面。
将数据拆分为训练数据源和评估数据源	Amazon ML 可以拆分数据源以将 70% 的数据用于模型训练，将 30% 的数据用于评估模型的预测性能。 在您使用“创建机器学习模型”向导和默认设置时，Amazon ML 会为您拆分数据。 如果您使用“创建机器学习模型”向导以及自定义设置，并且选择评估 ML 模型，您会看到一个选项，该选项允许 Amazon ML 为您拆分数据并对 30% 的数据运行评估。
将训练数据乱序化	在您使用“创建机器学习模型”向导和默认设置时，Amazon ML 为您乱序数据。您也可以在将数据导入 Amazon ML 之前将其乱序。
处理特征	将训练数据一起置入优化格式用于学习和归纳的处理称为特征转换。在您使用“创建机器学习模型”向导和默认设置时，Amazon ML 将推荐您数据的特征处理设置。 要指定特征处理设置，请使用“创建 ML 模型”向导的自定义选项并提供特征处理配方。
训练模型	使用“创建机器学习模型”向导在 Amazon ML 中创建模型时，Amazon ML 会训练您的模型。
选择模型参数	在 Amazon ML 中，您可以优化影响模型预测性能的四个参数：模型大小，扫描次数，乱序类型以及正则化。在使用“创建 ML 模型”向导创建 ML 模型并选择自定义选项时，您可以设置这些参数。
评估模型性能	使用“Create Evaluation”向导可评估您模型的预测性能。
特征选择	Amazon ML 学习算法可以删除对学习过程没有多大帮助的特征。要指示您希望删除这些特征，请在创建 ML 模型时选择 `L1 regularization` 参数。

ML 流程	Amazon ML 任务
为预测精度设置分数阈值	检查评估报告中不同分数阈值下模型的预测性能，然后根据您的业务应用设置分数阈值。分数阈值确定模型如何定义预测匹配。调整数字以控制错误肯定和错误否定。
使用模型	使用您的模型，通过“Create Batch Prediction”向导获取一批观察的预测。 或者，使用 `Predict` API 启用 ML 模型来处理实时预测，按需获得个别观察的预测。

设置 Amazon Machine Learning

首次使用 Amazon Machine Learning 之前，您需要有一个 AWS 账户。如果您没有账户，请注册 AWS。

注册 AWS

在注册 Amazon Web Services (AWS) 时，您的 AWS 账户会自动注册 AWS 中的所有服务，包括 Amazon ML。您只需为使用的服务付费。如果您已有 AWS 账户，请跳过此步骤。如果您还没有 AWS 账户，请使用以下步骤创建。

如需注册 AWS 账户

1. 转到 http://aws.amazon.com 并选择注册。
2. 按照屏幕上的说明进行操作。

作为注册流程的一部分，您会收到一个电话，需要您使用电话键盘输入一个 PIN 码。

教程：使用 Amazon ML 预测对营销方案的响应

使用 Amazon Machine Learning (Amazon ML)，您可以生成和训练预测模型，并将您的应用程序托管在可扩展的云解决方案中。在本教程中，我们将向您展示如何使用 Amazon ML 控制台创建数据源、生成机器学习 (ML) 模型，然后使用模型生成您可用于您应用程序的预测。

我们的示例练习演示如何确定目标营销活动的潜在客户，不过您可以应用相同的准则来创建和使用各种 ML 模型。为完成示例练习，您将使用来自加利福尼亚大学欧文分校 (UCI) 机器学习存储库公开提供的银行和营销数据集。这些数据集包含有关客户的一般信息，以及有关客户如何响应之前营销联系人的信息。您将使用此数据来确定哪些客户最有可能订阅您的新产品，即银行定期存款（也称为存折存款 (CD)）。

> **Warning**
>
> 本教程不包含在 AWS 免费套餐中。有关 Amazon ML 定价的更多信息，请参阅 Amazon Machine Learning 定价。

先决条件

要执行教程中的操作，您需要有 AWS 账户。如果您没有 AWS 账户，请参阅设置 Amazon Machine Learning。

步骤

步骤 1：准备数据

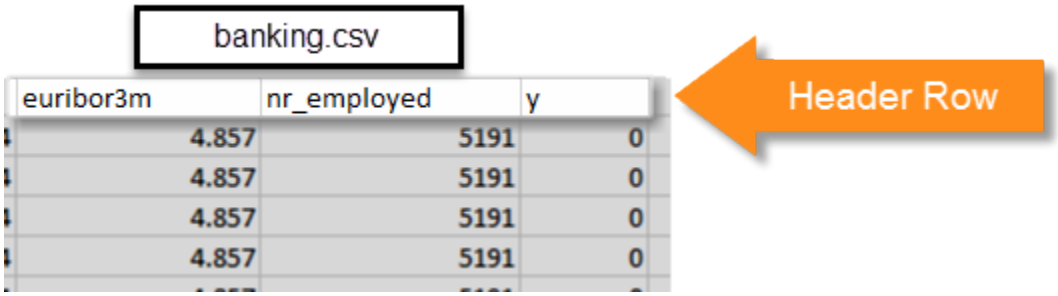
在机器学习中，您通常会获取数据并确保它经过了正确格式化，然后再开始训练过程。出于本教程的目的，我们从 UCI 机器学习存储库获取示例数据集，将其格式化以遵循 Amazon ML 指南，并使其可供

您下载。按照本主题中的以下过程操作，从我们的 Amazon Simple Storage Service (Amazon S3) 存储位置下载数据集，并将其上传到您自己的 S3 存储桶。

有关 Amazon ML 格式化要求，请参阅[了解 Amazon ML 的数据格式]。

下载数据集

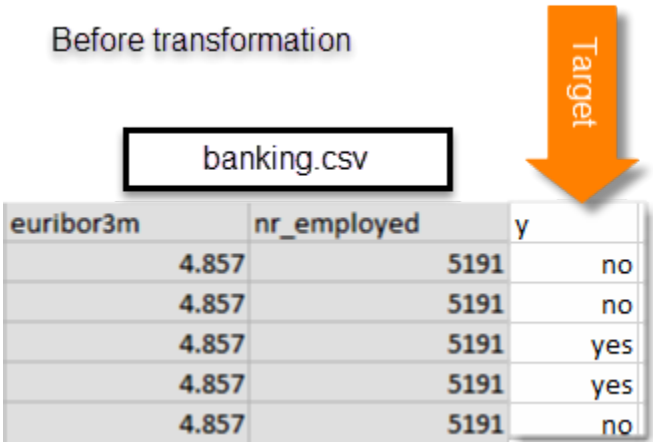
1. 单击 banking.zip，下载包含客户历史记录数据的文件，这些客户购买的产品与您的银行定期存款类似。解压缩该文件夹并将 banking.csv 文件保存到您的计算机上。
2. 单击 banking-batch.zip，下载您将用来预测潜在客户是否会响应您方案的文件。解压缩该文件夹并将 banking-batch.csv 文件保存到您的计算机上。
3. 打开 `banking.csv`。您将看到数据的行和列。标题行 包含各列的属性名称。属性 是指定的唯一属性，描述各客户的具体特征；例如 nr_employed 指示客户的雇佣状态。各行表示各个客户的相关观察的集合。



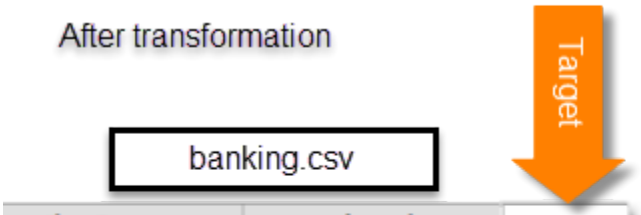
您希望 ML 模型回答问题“此客户是否会订阅我的新产品？”。在 `banking.csv` 数据集中，此问题的答案是属性 y，该属性包含值 1（表示“是”）或 0（表示“否”）。您希望 Amazon ML 用来学习如何进行预测的属性称为目标属性。

> **Note**
>
> 属性 y 是一个二进制属性。它只包含两个值之一，在这种情况下为 0 或 1。在原始 UCI 数据集中，y 属性为 Yes 或 No。我们已经为您编辑了原始数据集。属性 y 的所有表示 yes 的值现在是 1，所有表示 no 的值现在是 0。如果使用自己的数据，您可以为二进制属性使用其他值。有关有效值的更多信息，请参阅[使用 AttributeType 字段]。

以下示例显示我们将属性 y 中的值更改为二进制属性 0 和 1 前后的数据。



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



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

banking-batch.csv 文件不包含 y 属性。在创建了 ML 模型之后，您将使用该模型来预测该文件中各个记录的 y。

接下来，上传 banking.csv 和 banking-batch.csv 文件到 Amazon S3。

将文件上传到 Amazon S3 位置

1. 登录到 AWS Management Console，然后通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 在所有存储桶列表中，创建存储桶或者选择您希望将文件上传到的位置。
3. 在导航栏中，选择上传。
4. 选择 Add Files。
5. 在对话框中，导航到桌面，选择 banking.csv 和 banking-batch.csv，然后选择打开。

现在您已准备就绪，可创建训练数据源。

步骤 2：创建训练数据源

在将 `banking.csv` 数据集上传到 Amazon Simple Storage Service (Amazon S3) 位置之后，您可以用它来创建训练数据源。数据源是 Amazon Machine Learning (Amazon ML) 对象，包含输入数据的位置以及有关输入数据的重要元数据。Amazon ML 将数据源用于 ML 模型训练和评估等操作。

要创建数据源，请提供以下信息：

- 您数据的 Amazon S3 位置以及数据访问权限
- 架构，其中包含数据中各属性的名称及其类型（数值、文本、分类或二进制）
- 属性的名称，该属性包含您希望 Amazon ML 学习进行预测的答案，即目标属性

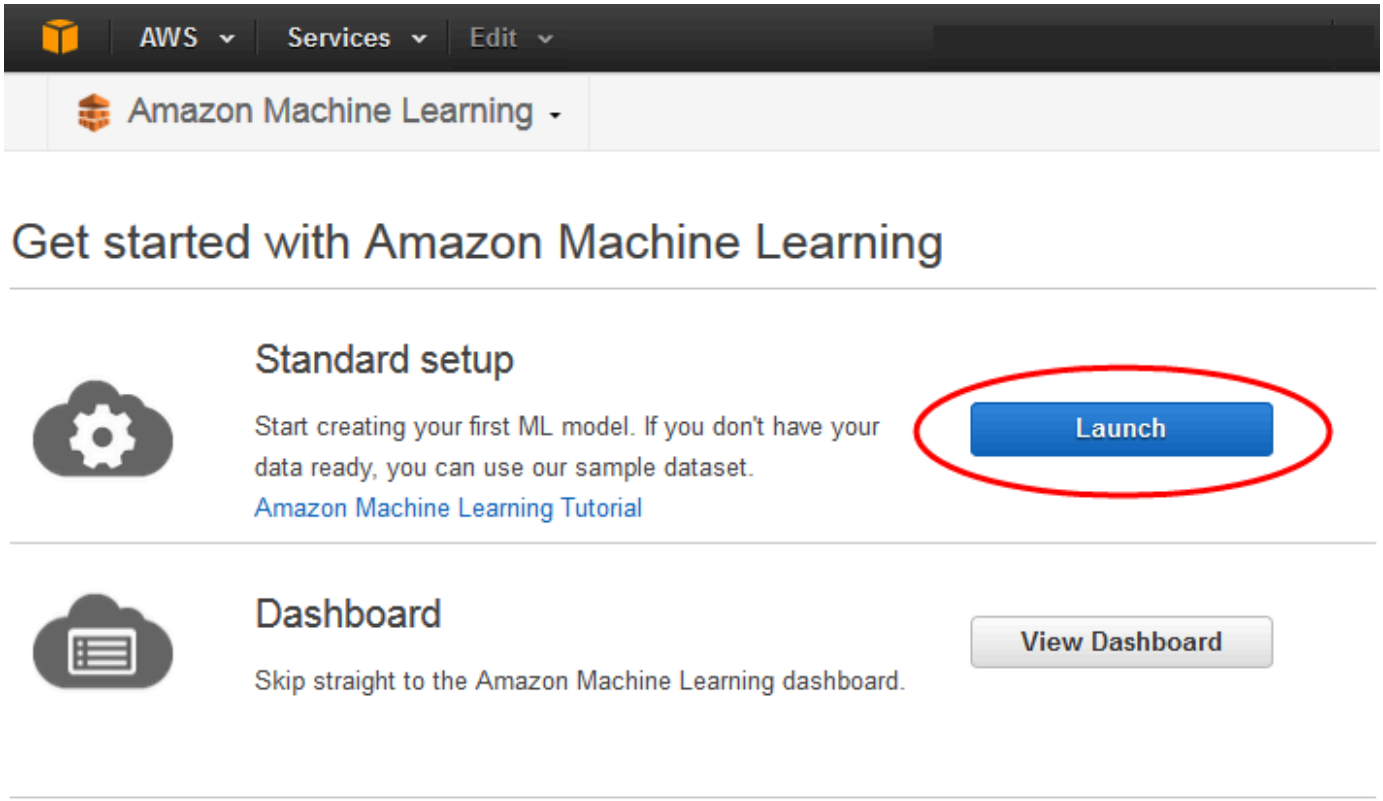
> **Note**
>
> 数据源并不实际存储您的数据，只是引用它。避免移动或更改在 Amazon S3 中存储的文件。否则，Amazon ML 无法访问它们来创建 ML 模型、生成评估或生成预测。

创建训练数据源

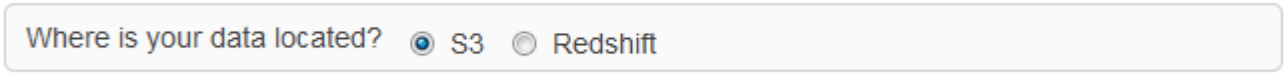
1. 打开 Amazon Machine Learning 控制台，网址为 https://console.aws.amazon.com/machinelearning/。
2. 选择开始使用。

   > **Note**
   >
   > 本教程假定您是首次使用 Amazon ML。如果您以前使用过 Amazon ML，则可以使用 Amazon ML 控制面板上的新建...下拉列表来创建新的数据源。

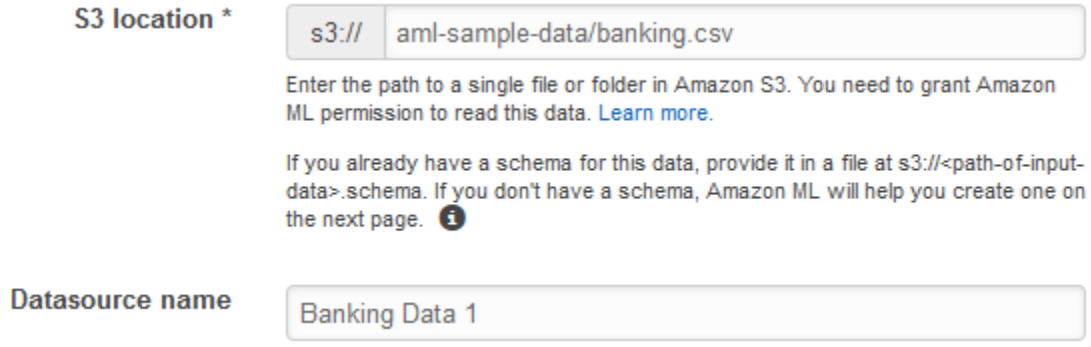
3. 在 Amazon Machine Learning 入门页面上，选择启动。



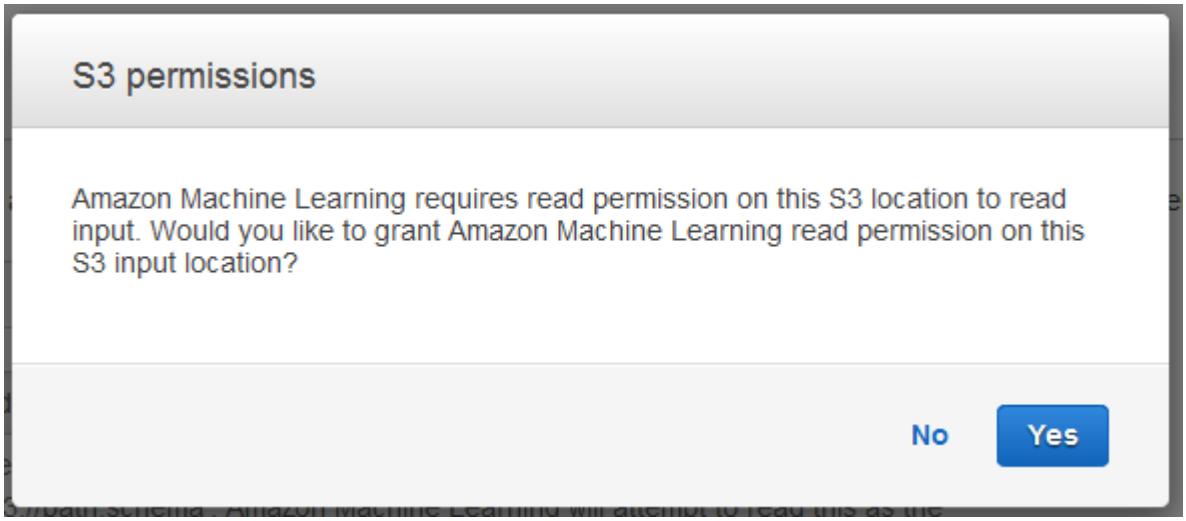
4. 在输入数据页面上，对于您的数据位于何处?，确保选择了 S3。



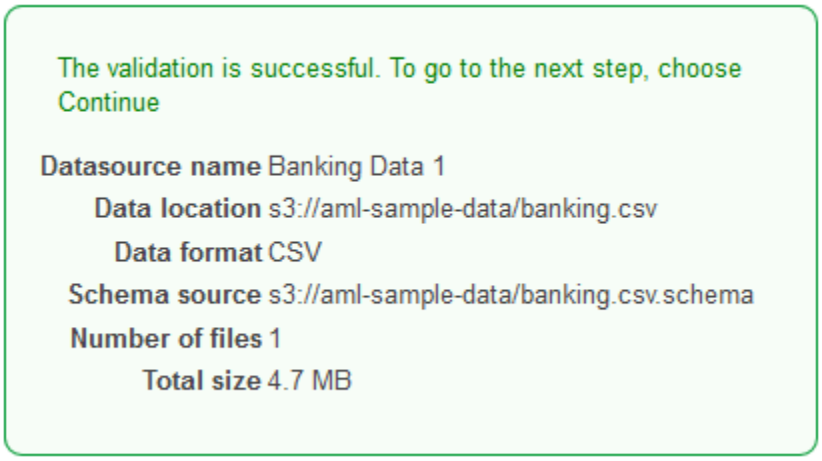
5. 对于S3 位置，键入来自“步骤 1：准备数据”中的 `banking.csv` 文件的完整位置。例如：*`your-bucket`***`/banking.csv`**。Amazon ML 会为您添加 s3:// 到存储桶名称前。
6. 为数据源名称 键入 **`Banking Data 1`**。



7. 选择验证。
8. 在 S3 权限对话框中，选择是。



9. 如果 Amazon ML 可以访问和读取 S3 位置中的数据文件，您将看到类似以下内容的页面。检查属性，然后选择继续。



接下来，建立架构。架构是 Amazon ML 解释 ML 模型的输入数据时需要的信息，包括属性名、为属性分配的数据类型以及特殊属性的名称。有两种方法可以向 Amazon ML 提供架构：

- 在上传您的 Amazon S3 数据时提供单独的架构文件。
- 允许 Amazon ML 推断属性类型并为您创建架构。

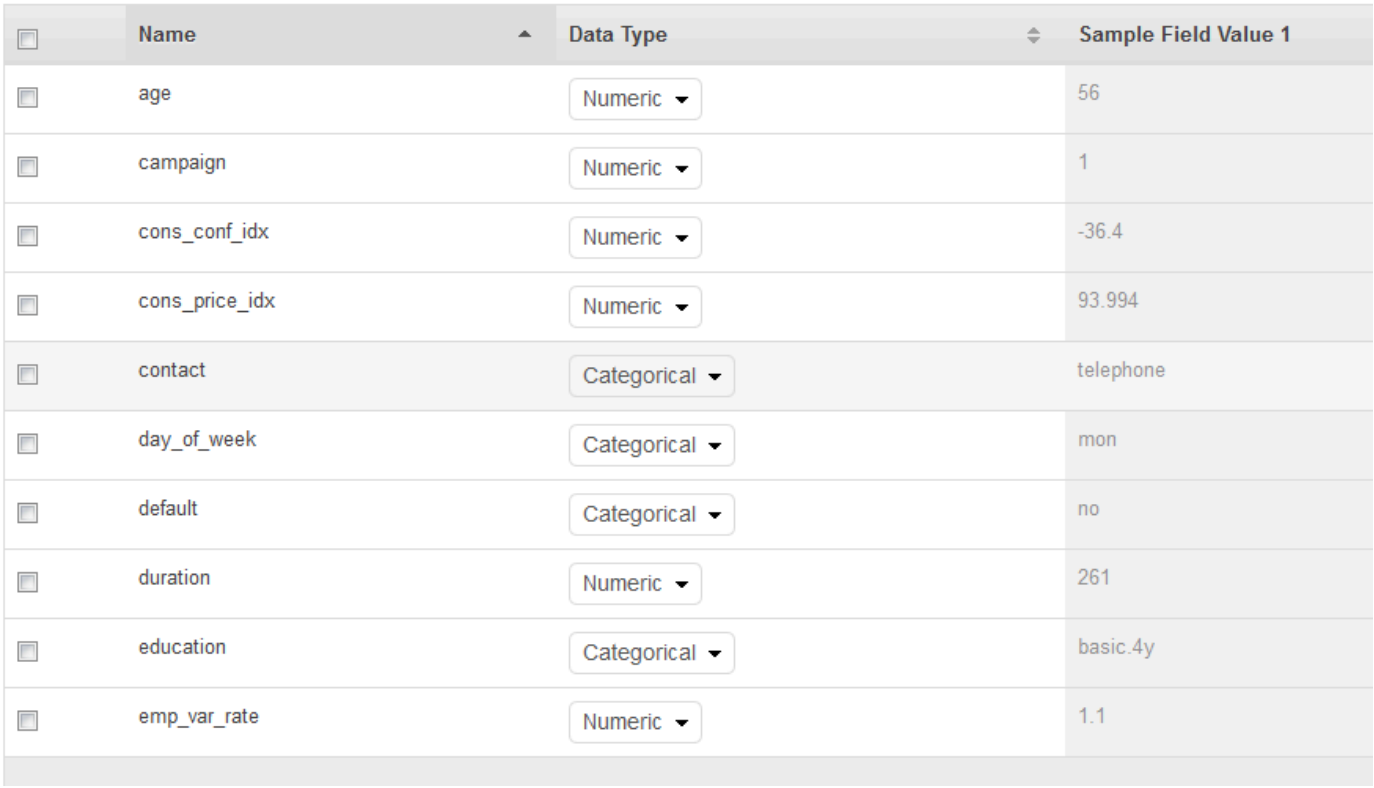
在本教程中，我们将要求 Amazon ML 推断架构。

有关创建单独架构文件的信息，请参阅为 Amazon ML 创建数据架构。

允许 Amazon ML 推断架构

1. 在架构页面上，Amazon ML 显示所推断的架构。检查 Amazon ML 为属性推断的数据类型。非常重要的一点是，向属性分配了正确的数据类型，以帮助 Amazon ML 正确提取数据并对属性实现正确的特征处理。

   - 只能有两种可能状态（例如 yes 或 no）的属性应标记为二进制。
   - 用于表示类别的数字或字符串属性应标记为 Categorical。
   - 对于数值数量的属性，如果其顺序有意义，则应标记为 Numeric。
   - 对于字符串属性，如果您希望将其视为空格分隔单词的字符串，则应标记为 Text。



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

2. 在本教程中，Amazon ML 能正确识别所有属性的数据类型，因此选择继续。

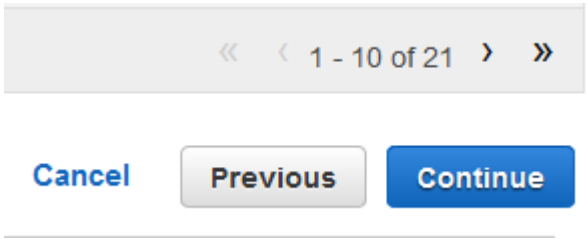
接下来，选择目标属性。

请记住，目标是 ML 模型必须学习预测的属性。属性 y 指示某个人过去是否订阅了营销活动：1（是）或 0（否）。

> **Note**
>
> 仅当您使用数据源来训练和评估 ML 模型时，才选择目标属性。

选择 y 作为目标属性

1. 在表的右下角中，选择单箭头以继续到表的下一页，其中显示了名为 y 的属性。



2. 在目标列中，选择 y。

   Amazon ML 确认已选择 y 作为目标。
3. 选择继续。
4. 在行 ID 页面上，对您的数据是否包含标识符?，确保已选择默认设置否。
5. 选择审核，然后选择继续。

现在您有一个训练数据源，您已准备好创建模型。

步骤 3：创建 ML 模型

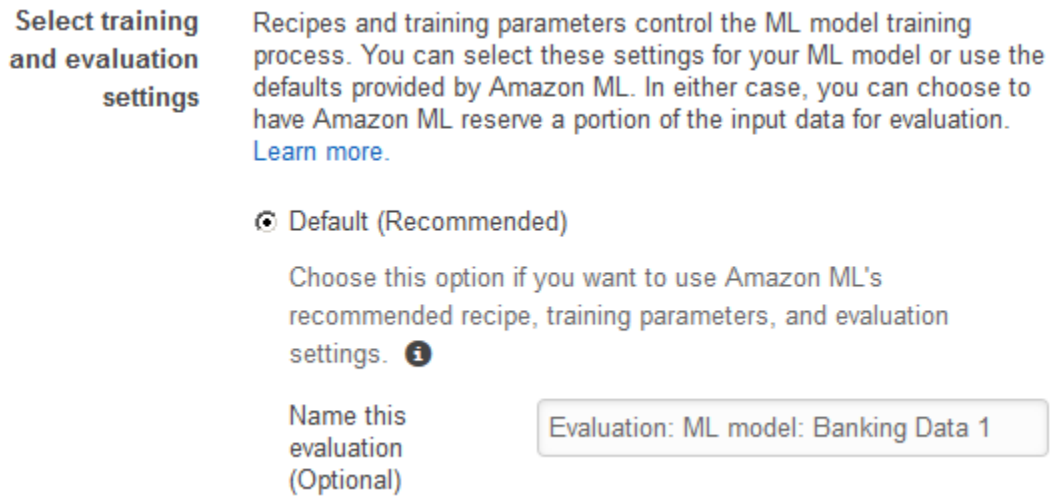
在创建训练数据源之后，您可以用它来创建 ML 模型、训练模型，然后评估结果。ML 模型是 Amazon ML 在训练期间从数据中发现的模式的集合。您可以使用模型创建预测。

创建 ML 模型

1. 由于“入门”向导创建了训练数据源和模型，Amazon Machine Learning (Amazon ML) 会自动使用您刚创建的训练数据源，并将您直接转到机器学习模型设置页面。在 ML 模型设置页面上，确保 ML 模型名称中显示了默认值 **`ML model: Banking Data 1`**。

   使用友好的名称，例如默认值，帮助您轻松识别和管理 ML 模型。

2. 对于训练和评估设置，请确保选择默认。



3. 对于为此评估命名，请接受默认值 **`Evaluation: ML model: Banking Data 1`**。
4. 选择审核，检查您的设置，然后选择完成。

   选择完成之后，Amazon ML 将您的模型添加到处理队列中。Amazon ML 创建您的模型时，它会应用默认值并应用以下操作：

   - 将训练数据源拆分为两个部分：一个包含 70% 的数据，另一个包含剩余的 30%
   - 在包含 70% 输入数据的部分上训练 ML 模型
   - 使用剩余的 30% 输入数据评估模型

   当您的模型在队列中时，Amazon ML 将状态报告为待处理。Amazon ML 创建您的模型时，它会将状态报告为正在进行。完成所有操作后，它会将状态报告为已完成。等待评估完成，然后再继续操作。

现在，您已准备就绪，可查看您的模型的性能和设置截断值分数。

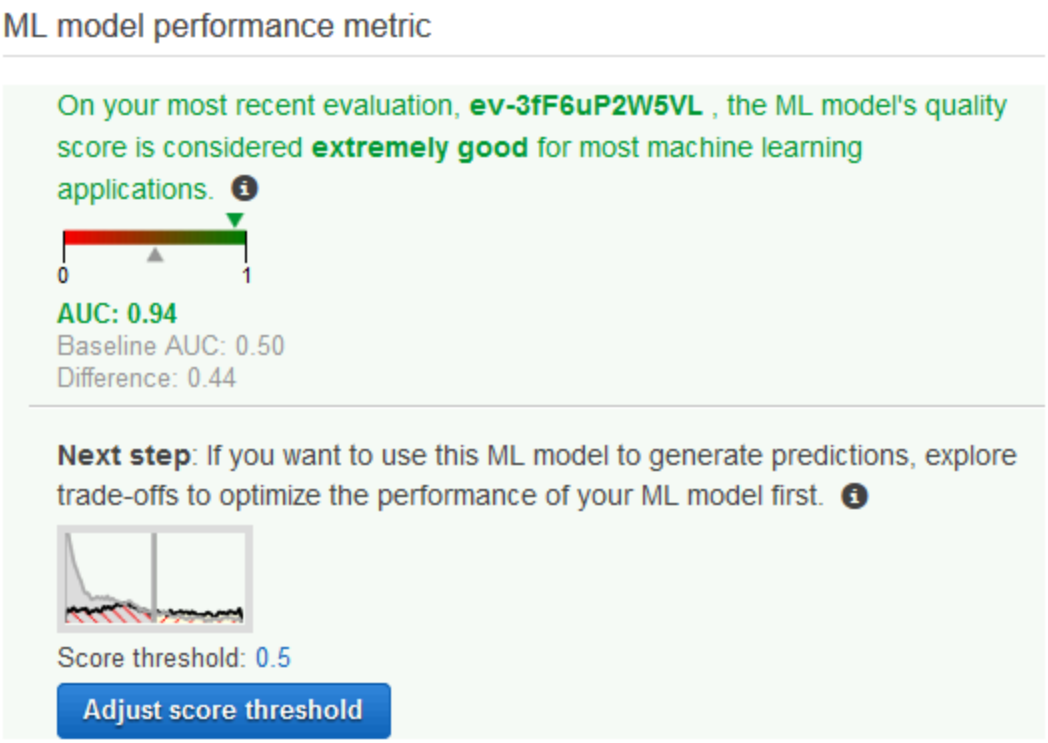
有关训练和评估模型的更多信息，请参阅训练 ML 模型和evaluate an ML model。

步骤 4：查看 ML 模型的预测性能和设置分数阈值

现在，您已经创建了 ML 模型，Amazon Machine Learning (Amazon ML) 也对其进行了评估，我们来看看该模型是否足以投入使用。在评估期间，Amazon ML 计算了称为曲线下面积 (AUC) 指标的行业标准质量指标，用于表示 ML 模型的性能质量。Amazon ML 还会解析 AUC 指标，让您了解 ML 模型的质量是否能够满足大多数机器学习应用程序的需求。（在衡量 ML 模型准确度中了解有关 AUC 的更多信息。）让我们看一看 AUC 指标，然后调整分数阈值或截断值，以优化您的模型的预测性能。

查看您的 ML 模型的 AUC 指标

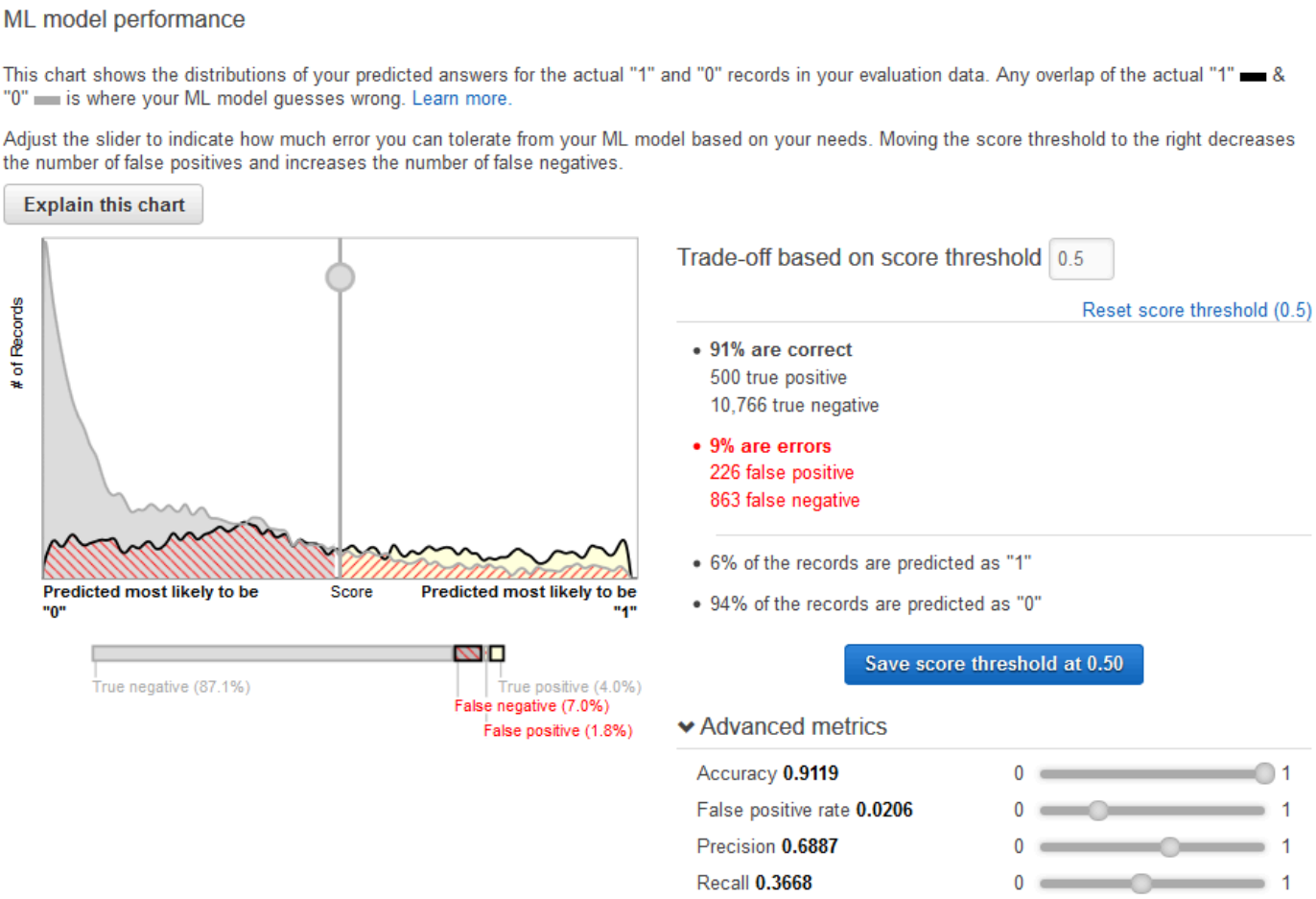
1. 在 ML 模型摘要页面上的 ML 模型报告导航窗格中，依次选择评估、评估: ML 模型: 银行模型 1 和摘要。
2. 在评估摘要页面上，查看评估摘要，包括模型的 AUC 性能指标。



ML 模型为预测数据源中的每个记录生成数字预测分数，然后应用一个阈值将这些分数转化为二进制标签 0（表示“否”）或 1（表示“是”）。通过更改分数阈值，您可以调整 ML 模型如何分配这些标签。现在，设置分数阈值。

为 ML 模型设置分数阈值

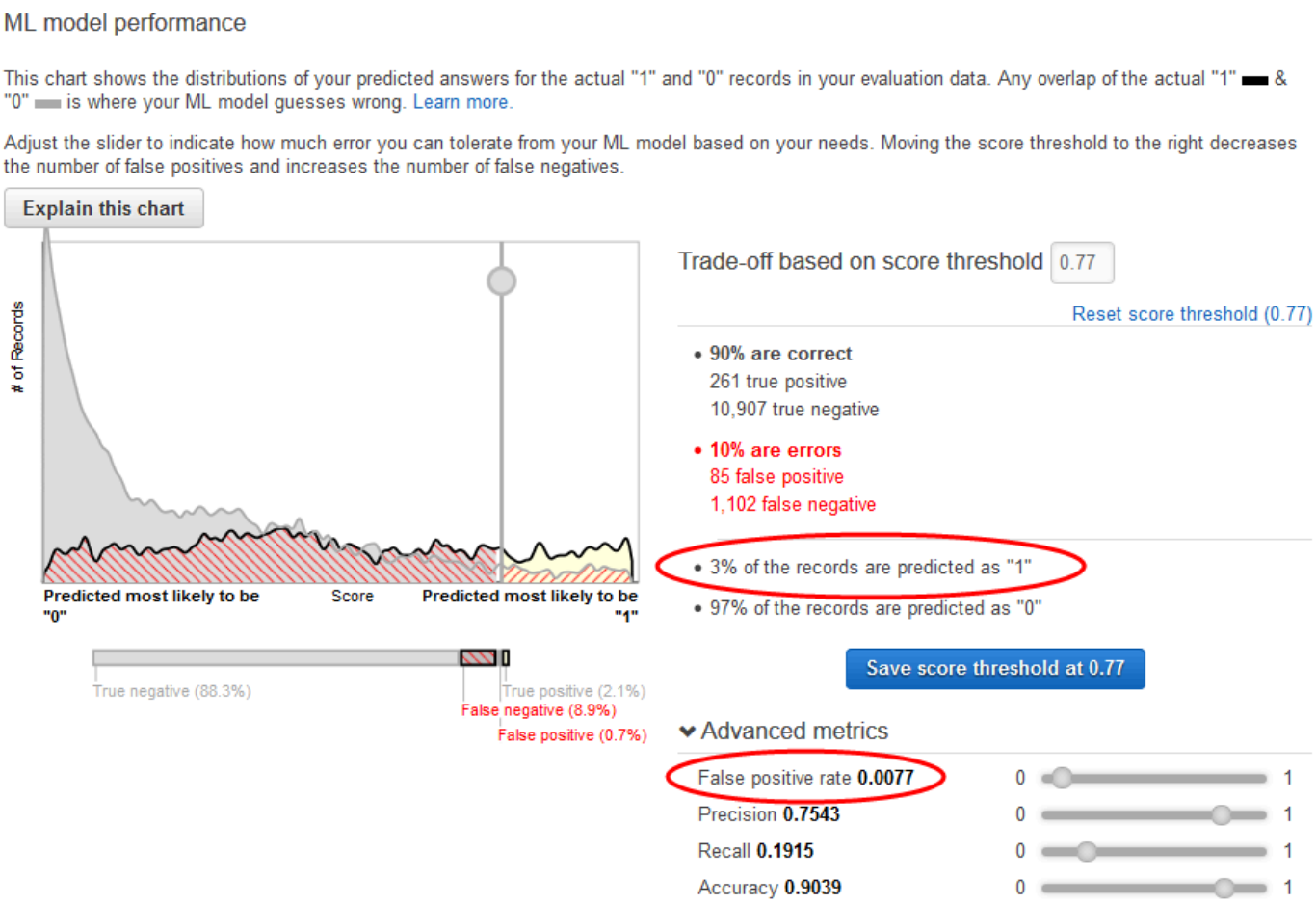
1. 在评估摘要页面上，选择调节分数阈值。



通过调整分数阈值，您可以优化您的 ML 模型的性能指标。调整此值会更改置信度级别，因为必须先对模型进行预测，然后才能将预测视为阳性。此外，它还会更改您在预测中可以容忍的假阴性和假阳性的数量。

您可以通过增加分数阈值来控制模型视为阳性预测的截断值，直到它仅将具有最大真阳性可能性的预测视为阳性。您也可以减少分数阈值，直到不再有任何假阴性。选择您的截断值，以反映您的业务需求。在本教程中，每个假阳性都会花费活动资金，因此我们需要较高的真阳性与假阳性比率。

2. 假设您希望将目标放在可能订阅产品的前 3% 的客户。滑动垂直选择器，将分数阈值设置为与 %3 的记录的预测值为“1”相对应的值。



请注意此分数阈值对 ML 模型性能的影响：假阳性错误比率为 0.007。我们假定该假阳性比率可接受。

3. 选择将分数阈值保存在 0.77。

每次使用此 ML 模型进行预测时，它将分数超过 0.77 的记录预测为“1”，将其余记录预测为“0”。

要了解有关分数阈值的更多信息，请参阅二元分类。

现在您已准备就绪，可以使用您的模型创建预测。

步骤 5：使用 ML 模型生成预测

Amazon Machine Learning (Amazon ML) 可以生成两种类型的预测 — 批量和实时。

实时预测是 Amazon ML 按需生成的单个观察的预测。实时预测适用于移动应用程序、网站和其他需要以交互方式使用结果的应用程序。

批量预测是对一组观察的预测集。Amazon ML 在批量预测中将记录放在一起处理，因此处理可能需要一些时间。对于应用程序，如果需要的是对一组观察进行的预测或者无需交互使用结果的预测。

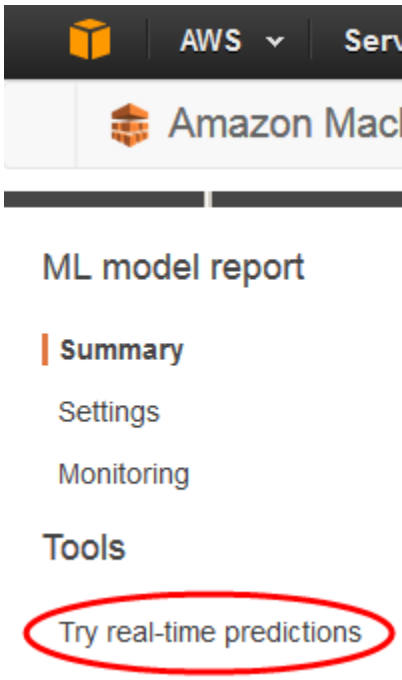
在本教程中，您将生成实时预测，预测一位潜在客户是否将订阅新产品。您还会为一大批潜在客户生成预测。对于批量预测，您将使用您在banking-batch.csv中上传的 步骤 1：准备数据 文件。

我们从实时预测开始。

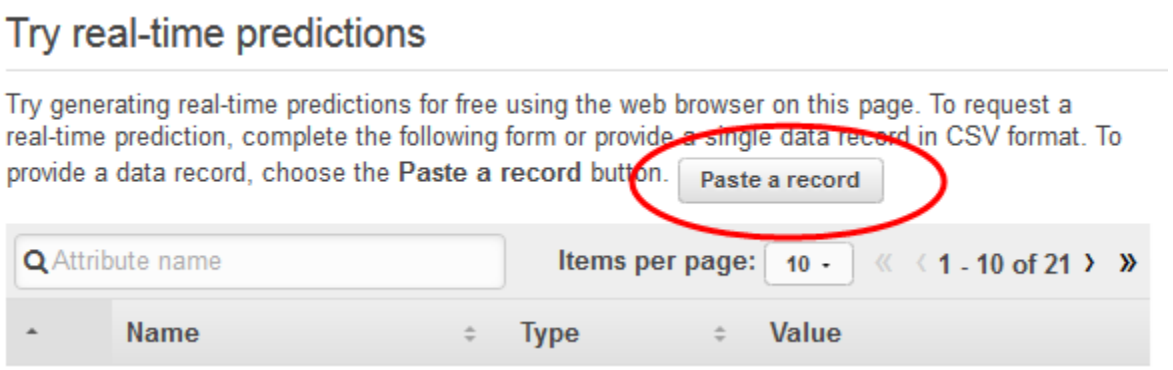
> **Note**
>
> 对于需要实时预测的应用程序，您必须为 ML 模型创建实时终端节点。在实时终端节点可用时，您会产生费用。在您承诺使用实时预测并开始产生与之相关的费用之前，您可以尝试在 Web 浏览器中使用实时预测功能，无需创建实时终端节点。这就是我们将在本教程中完成的操作。

尝试实时预测

1. 在 ML 模型报告导航窗格中，选择尝试实时预测。



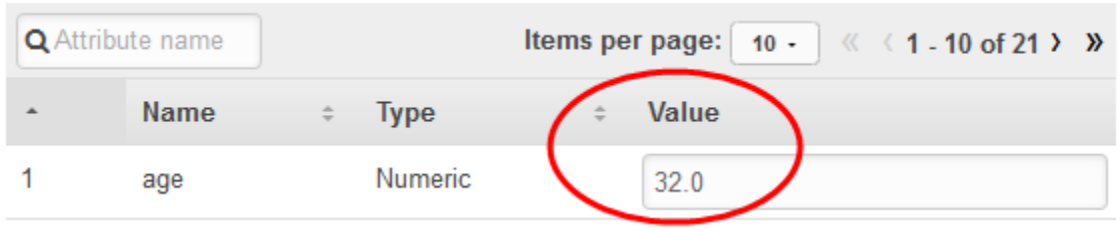
2. 选择粘贴记录。



3. 在粘贴记录对话框中，粘贴以下观察：

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

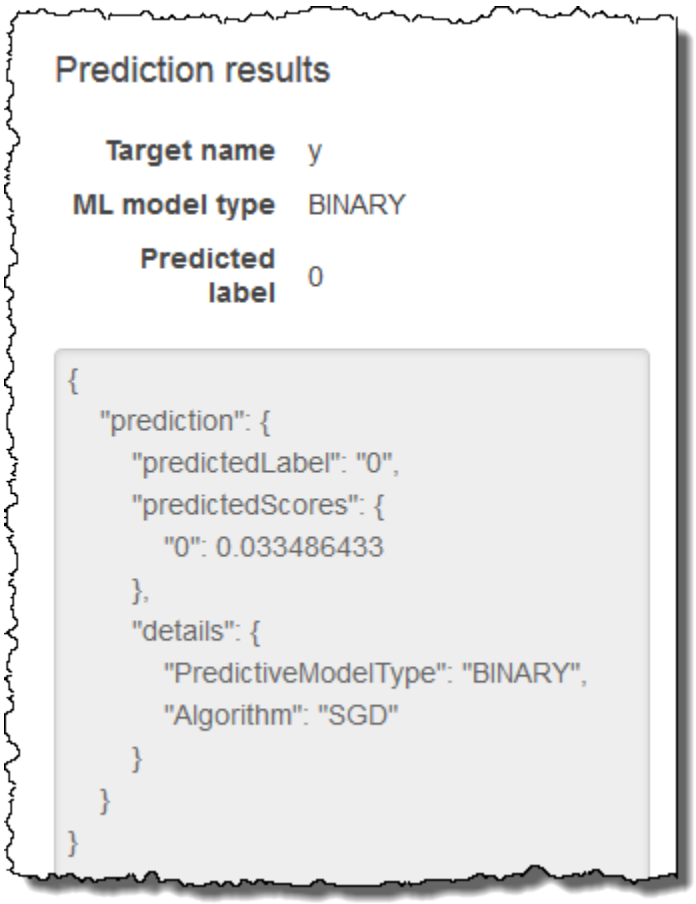
4. 在粘贴记录对话框中，选择提交以确认您希望为此观察生成预测。Amazon ML 填充实时预测表单中的值。



> **Note**
>
> 您还可以通过键入单个值来填充值字段。不论选择哪种方法，您应提供未用于训练模型的观察。

5. 在页面底部，选择创建预测。

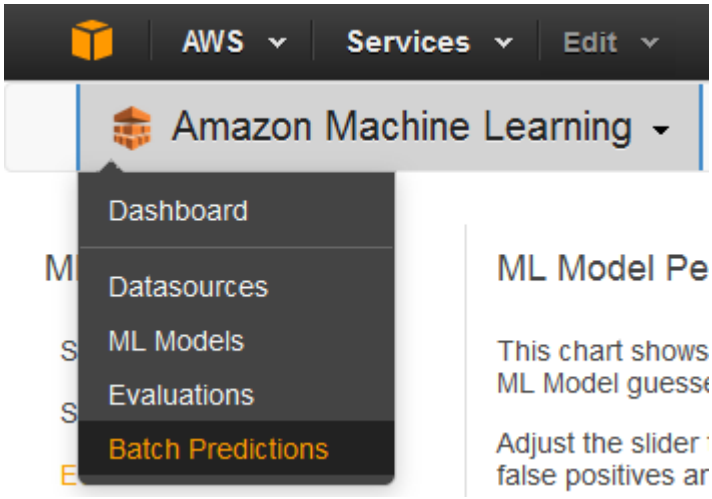
预测显示在右侧的预测结果窗格中。此预测具有为 0 的预测标签，这意味着此潜在客户不太可能响应市场活动。预测标签 为 1 意味着可能有可能响应市场活动。



现在，创建批量预测。您向 Amazon ML 提供所用 ML 模型的名称；希望为其生成预测的输入数据（Amazon ML 将从此数据创建批量预测数据源）的 Amazon Simple Storage Service (Amazon S3) 位置；用于存储结果的 Amazon S3 位置。

创建批量预测

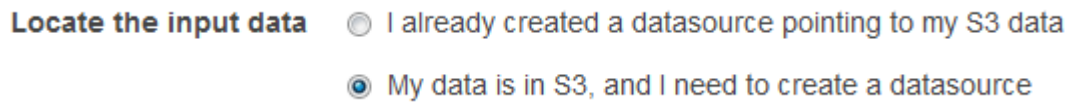
1. 选择 Amazon Machine Learning，然后选择批量预测。



2. 选择创建新批量预测。
3. 在用于批量预测的 ML 模型页面上，选择ML 模型: 银行数据 1。

   Amazon ML 显示 ML 模型名称、ID、创建时间以及关联的数据源 ID。
4. 选择继续。
5. 要生成预测，您需要向 Amazon ML 提供所要进行预测的数据。这称为输入数据。首先，将输入数据移入数据源以便 Amazon ML 访问。

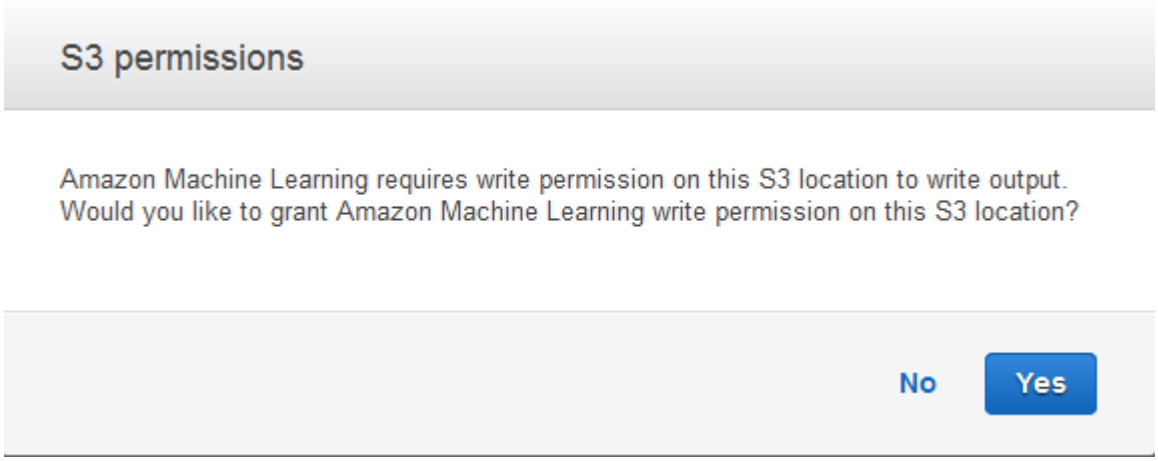
   对于找到输入数据，选择我的数据在 S3 中，并且我需要创建数据源。

6. 为数据源名称 键入 **`Banking Data 2`**。
7. 对于 SS3 位置，键入 `banking-batch.csv` 文件的完整位置：*`your-bucket`***`/banking-batch.csv`**。
8. 对于您的 CSV 中的第一行是否包含列名?，选择是。
9. 选择验证。

   Amazon ML 验证您数据的位置。
10. 选择继续。
11. 对于 S3 目标，键入您在“步骤 1：准备您的数据”中上传文件到 Amazon S3 位置的名称。Amazon ML 将预测结果上传到这里。

12. 对于批量预测名称，接受默认值 **Batch prediction: ML model: Banking Data 1**。Amazon ML 根据将用于创建预测的模型选择默认名称。在本教程中，模型和预测根据训练数据源 Banking Data 1 命名。
13. 选择审核。
14. 在 S3 权限对话框中，选择是。

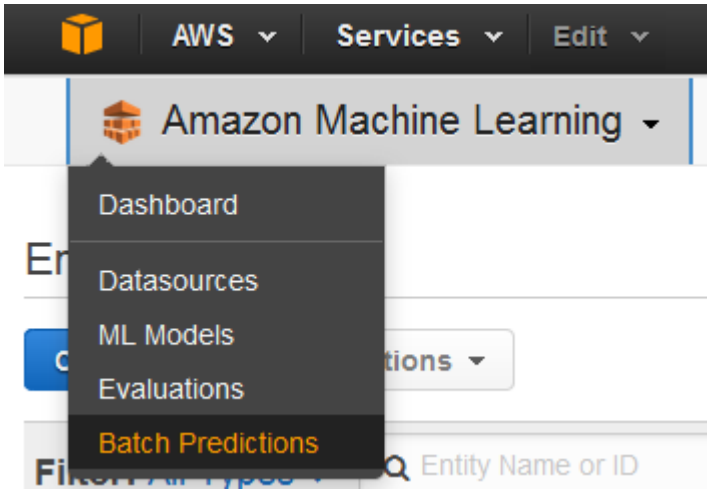


15. 在审核页面上，选择完成。

批量预测请求发送到 Amazon ML 并进入队列中。Amazon ML 处理批量预测所用的时间取决于您数据源的大小以及 ML 模型的复杂性。Amazon ML 处理请求时，它将状态报告为正在进行。批量预测完成后，请求的状态更改为完成。现在，您可以查看结果。

查看预测

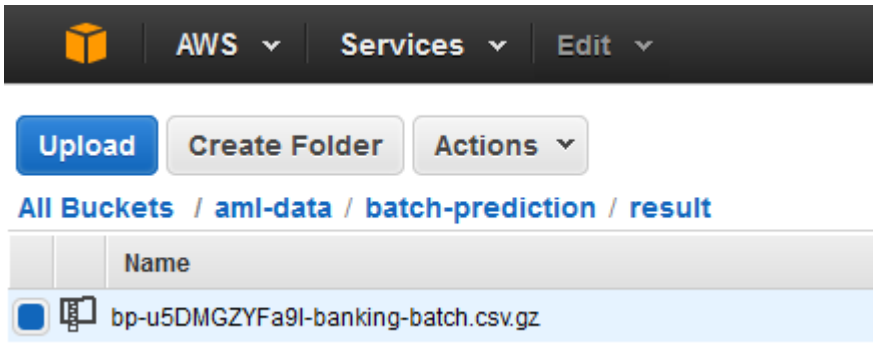
1. 选择 Amazon Machine Learning，然后选择批量预测。



2. 在预测列表中，选择批量预测: ML 模型: 银行数据 1。此时显示批量预测信息页面。

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. 要查看批量预测的结果，请转到 Amazon S3 控制台（网址为 https://console.aws.amazon.com/s3/）并导航到输出 S3 URL 字段中引用的 Amazon S3 位置。从该处导航到结果文件夹，其名称将类似于 `s3://aml-data/batch-prediction/result`。



预测存储在压缩的 .gzip 文件中，扩展名为 .gz。

4. 下载预测文件到您的桌面，解压缩，然后打开它。

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

该文件有两列 bestAnswer 和 score，数据源中的每个观察为一行。bestAnswer 列中的结果是基于您在步骤 4：查看 ML 模型的预测性能和设置分数阈值中设置的评分阈值 0.77。大于 0.77 的

score 会导致 bestAnswer 为 1，这是正向响应或预测，小于 0.77 的 score 会导致 bestAnswer 为 0，这是负向响应或预测。

以下示例基于 0.77 的分数阈值显示正向和负向预测。

正向预测：

bestAnswer	score
1	0.8228876

在此示例中，bestAnswer 的值为 1，score 的值为 0.8228876。bestAnswer 的值为 1 是因为 score 大于分数阈值 0.77。bestAnswer 为 1 指示客户希望购买产品，因此，视为正向预测。

负向预测：

bestAnswer	score
0	0.7695356

在此示例中，bestAnswer 的值为 0，因为 score 值为 0.7695356，这低于分数阈值 0.77。bestAnswer 为 0 表示客户不太可能购买您的产品，因此视为负向预测。

批处理结果中的每一列均与您的批处理输入中的一行相对应（您的数据源中的一个观察）。

分析预测之后，您可以执行定向市场营销活动；例如，向预测分数为 1 的所有人发送宣传材料。

在创建、查看并使用了自己的模型之后，现在清除您创建的数据和 AWS 资源以避免产生不必要的费用并保持工作区整洁。

步骤 6：清除

为避免产生额外的 Amazon Simple Storage Service (Amazon S3) 费用，删除存储在 Amazon Simple Storage Service (Amazon S3) 中的数据。其他未使用的 Amazon ML 资源不收取费用，但是我们建议您删除它们以保持工作区清洁。

删除存储在 Amazon S3 中的输入数据

1. 通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 导航到存储 `banking.csv` 和 `banking-batch.csv` 文件的 Amazon S3 位置。
3. 选择 `banking.csv`、`banking-batch.csv` 和 `.writePermissionCheck.tmp` 文件。
4. 选择 Actions，然后选择 Delete。

5. 当系统提示您确认时，选择 OK。

尽管保留 Amazon ML 运行批量预测的记录或者在教程中创建的数据源、模型和评估不会产生任何费用，但是我们建议您删除它们以防止工作区的杂乱。

删除批量预测

1. 导航到您存储批量预测的输出的 Amazon S3 位置。
2. 选择 `batch-prediction` 文件夹。
3. 选择 Actions，然后选择 Delete。
4. 当系统提示您确认时，选择 OK。

删除 Amazon ML 资源

1. 在 Amazon ML 控制面板中，选择以下资源。
   - `Banking Data 1` 数据源
   - `Banking Data 1_[percentBegin=0, percentEnd=70, strategy=sequential]` 数据源
   - `Banking Data 1_[percentBegin=70, percentEnd=100, strategy=sequential]` 数据源
   - `Banking Data 2` 数据源
   - `ML model: Banking Data 1` ML 模型
   - `Evaluation: ML model: Banking Data 1` 评估
2. 选择 Actions，然后选择 Delete。
3. 在对话框中，选择删除以删除所有选定资源。

现在您已成功完成了教程。要继续使用控制台来创建数据源、模型和预测，请参阅 Amazon Machine Learning 开发人员指南。要了解如何使用 API，请参阅 Amazon Machine Learning API 引用。

创建和使用数据源

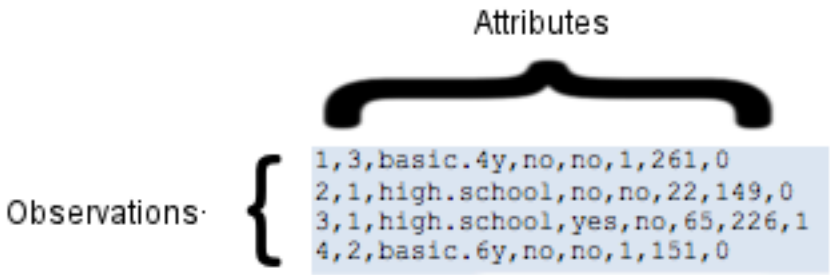
您可以使用 Amazon ML 数据源来训练 ML 模型、评估 ML 模型并使用 ML 模型生成批量预测。数据源对象包含有关输入数据的元数据。在您创建数据源时，Amazon ML 读取输入数据、计算其属性的描述性统计信息，并存储统计信息、架构和其他信息作为数据源对象的一部分。创建数据源之后，您可以使用 Amazon ML 数据洞察来探究输入数据的统计属性，并可以使用数据源来训练 ML 模型。

> **Note**
>
> 此部分假设您已熟悉 Amazon Machine Learning 概念。

主题

了解 Amazon ML 的数据格式

输入数据是您用于创建数据源的数据。您必须使用逗号分隔值 (.csv) 的格式保存输入数据。.csv 文件中的每一行是一个数据记录或观察。.csv 文件中的每一列包含观察的一个属性。例如，下图显示了一个 .csv 文件的内容，其中有四个观察，每个观察位于自己的行中。每个观察包含八个属性，以逗号分隔。这些属性展示了由观察呈现的每个人的以下信息：customerId，jobId，education，housing，loan，campaign，duration，willRespondToCampaign。



属性

Amazon ML 需要每个属性的名称。您可以通过以下方法指定属性名称：

- 在您用作输入数据的 .csv 文件的第一行（也称为标头行）中包括属性名称
- 在与输入数据处于相同 S3 存储桶的单独架构文件中包含属性名称

有关使用架构文件的更多信息，请参阅创建数据架构。

以下 .csv 文件示例在标头行中包括属性的名称。

```
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
```

输入文件格式要求

包含输入数据的 .csv 文件必须满足以下要求：

- 必须为使用 ASCII、Unicode 或 EBCDIC 等字符集的纯文本。
- 由观察组成，每行一个观察。
- 对于每个观察，属性值必须以逗号分隔。
- 如果属性值包含逗号（分隔符），整个属性值必须以双引号括起。
- 每个观察必须以行尾字符终止，这是一个特殊字符或字符序列，指示行结尾。

- 属性值不能包含行尾字符，即使属性值以双引号括起。
- 每个观察必须具有相同数量的属性和属性序列。
- 每个观察必须小于等于 100KB。在处理期间，Amazon ML 拒绝任何大于 100KB 的观察。如果 Amazon ML 拒绝的观察超过了 1 万个，它会拒绝整个 .csv 文件。

使用多个文件作为亚马逊机器学习的数据输入

您可以将输入以单个文件或文件集合的形式提供给 Amazon ML 学习。集合必须满足这些条件：

- 所有文件必须具有相同数据架构。
- 所有文件必须驻留在同一 Amazon Simple Storage Service (Amazon S3) 前缀中，并且您为集合提供的路径必须以正斜杠（“/”）字符结尾。

例如，如果您的数据文件名为 input1.csv、input2.csv 和 input3.csv，并且 S3 存储桶名称为 s3://examplebucket，则文件路径类似于下文：

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

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

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

您可以提供以下 S3 位置作为 Amazon ML 的输入：

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

CSV 格式的行尾字符

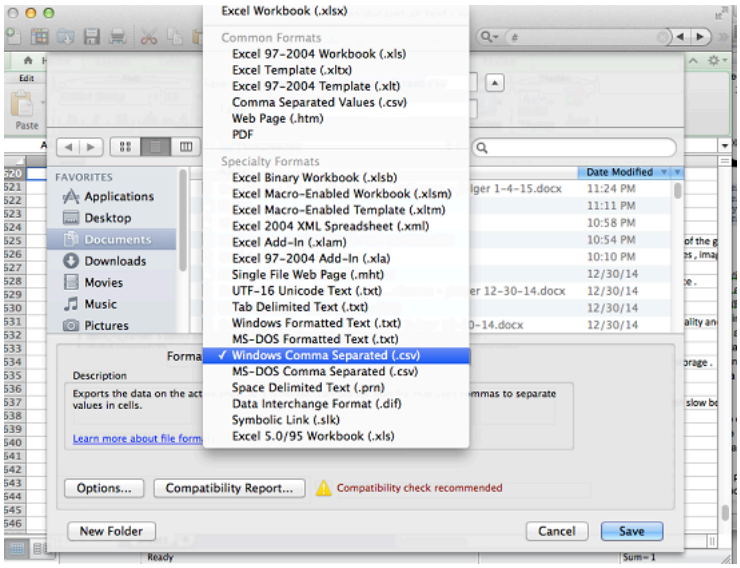
当您创建 .csv 文件时，每个观察将由特殊的行尾字符终止。此字符不可见，但在您按 Enter 或 Return 键时自动包括在每个观察的结尾。表示行尾的特殊字符取决于您的操作系统。Unix 系统，例如 Linux 或 OS X，使用由“\n”指示的换行符（ASCII 代码十进制 10，十六进制 0x0a）。Microsoft Windows 使用名为回车符和换行符，使用“\r\n”指示（ASCII 代码十进制 13 和 10，十六进制 0x0d 和 0x0a）。

如果您希望使用 OS X 和 Microsoft Excel 创建自己的 .csv 文件，请执行以下步骤。确保选择了正确的格式。

使用 OS X 和 Excel 时保存 .csv 文件

1. 保存 .csv 文件时，选择格式，然后选择 Windows Comma Separated (.csv)。

2. 选择保存。



> **Important**
>
> 请勿使用以逗号分隔值 (.csv) 或 MS-DOS 逗号分隔 (.csv) 格式保存 .csv 文件，因为 Amazon ML 无法读取这些格式。

为 Amazon ML 创建数据架构

架构 包含输入数据中的所有属性及其相应的数据类型。Amazon ML 可以通过架构了解数据源中的数据。Amazon ML 使用架构中的信息来读取和解释输入数据、计算统计数据、应用正确的属性转换以及优化其学习算法。如果您未提供架构，Amazon ML 会通过数据推断出一个架构。

示例架构

要让 Amazon ML 正确读取输入数据并生成准确的预测结果，必须为每个属性分配正确的数据类型。我们通过一个例子来了解如何将数据类型分配给属性以及如何在架构中包含属性和数据类型。我们将调用示例“Customer Campaign”，因为我们要预测哪些客户将响应我们的电子邮件营销活动。我们的输入文件是 .csv 文件，其中包含九列：

```
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
```

这是此数据的架构：

```
{
    "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"
        }

    ]
}
```

在此示例的架构文件中，rowId 的值为 customerId：

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

属性 willRespondToCampaign 被定义为目标属性：

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

customerId 属性和 CATEGORICAL 数据类型与第一列关联，jobId 属性和 CATEGORICAL 数据类型与第二列关联，jobDescription 属性和 TEXT 数据类型与第三列关联，education 属性和 CATEGORICAL 数据类型与第四列关联，依此类推。第九列与 willRespondToCampaign 属性和 BINARY 数据类型关联，此属性还被定义为目标属性。

使用 targetAttributeName 字段

targetAttributeName 值是要预测的属性的名称。您在创建或评估模型时必须分配 targetAttributeName。

在您训练或评估 ML 模型时，targetAttributeName 会标识包含目标属性的“正确”答案的输入数据中属性的名称。Amazon ML 使用包含正确答案的目标来发现模式和生成 ML 模型。

在您评估您的模型时，Amazon ML 使用目标来检查您的预测结果的准确性。在您创建和评估完 ML 模型之后，可以通过 ML 模型利用数据和未分配的 targetAttributeName 生成预测。

您可在 Amazon ML 控制台中（创建数据源时）或在架构文件中定义目标属性。如果您要创建自己的架构文件，请使用以下语法来定义目标属性：

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

在此示例中，exampleAttributeTarget 是输入文件中属性的名称，该属性是目标属性。

使用 rowID 字段

row ID 是与输入数据中的属性相关联的可选标志。如果指定该标志，标记为 row ID 的属性则会包含在预测输出中。借助此属性，您可以更轻松地将预测与对应的观察进行关联。一个很好的例子是，row ID 是客户 ID 或类似的唯一属性。

> **Note**
>
> 行 ID 仅供您参考。Amazon ML 在训练 ML 模型时不使用此属性。选择一个属性作为行 ID，则在训练 ML 模型时不会使用此属性。

您可在 Amazon ML 控制台中（创建数据源时）或在架构文件中定义 row ID。如果您要创建自己的架构文件，请使用以下语法来定义 row ID：

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

在上述示例中，exampleRow 是在输入文件中被定义为行 ID 的属性的名称。

生成批量预测时，您可能会得到以下输出：

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

在此示例中，RowID 表示属性 customerId。例如，预测 customerId 55 响应电子邮件营销活动的置信度低 (0.46317)，而预测 customerId 102 响应电子邮件营销活动的置信度 (0.89625) 高。

使用 AttributeType 字段

在 Amazon ML 中，有四个用于属性的数据类型：

二进制

为只包含两种可能的状态（例如 BINARY 或 yes）的属性选择 no。

例如，属性 isNew 用于跟踪某个人是否是新客户，该属性可能包含 true 值（用于表示这个人是新客户）和 false 值（用于表示此人不是新客户）。

有效的负值为 0、n、no、f 和 false。

有效的正值为 1、y、yes、t 和 true。

Amazon ML 会忽略二进制输入的大小写和去掉两边的空格。例如，" FaLSe " 是有效的二进制值。您可以在同一个数据源中混合使用二进制值，例如使用 true、no 和 1。Amazon ML 只为二进制属性输出 0 和 1。

分类

为使用的唯一字符串值数量有限的属性选择 CATEGORICAL。例如，用户 ID、月份和邮政编码都是分类值。分类属性将被视为单个字符串，不会进一步令牌化。

数值

为将数量作为值的属性选择 NUMERIC。

例如，温度、重量和点击率都是数值。

并非所有以数字为值的属性都为数值型。诸如月份中的天和 ID 之类的分类属性通常也用数字表示。要被视为数值，一个数字必须可与另一个数字进行比较。例如，通过客户 ID 664727 无法了解有关客户 ID 124552 的任何信息，但您可以了解权重为 10 的属性比权重为 5 的属性影响更大。月份的天不是数值，因为一个月的第一天可能出现在另一个月的第二天之前或之后。

> **Note**
>
> 使用 Amazon ML 创建架构时，它会将 Numeric 数据类型分配给所有使用数字的属性。如果使用 Amazon ML 创建架构，请检查是否有不正确的分配，并将这些属性设置为 CATEGORICAL。

文本

为值为单词字符串的属性选择 TEXT。读取文本属性时，Amazon ML 将其转换为用空格分隔的令牌。

例如，email subject 将变为 email 和 subject，email-subject here 将变为 email-subject 和 here。

如果训练架构中变量的数据类型与该变量在评估架构中的数据类型不一致，Amazon ML 会更改评估数据类型以与训练数据类型保持一致。例如，如果训练数据架构将数据类型 TEXT 分配给变量 age，但

评估架构将数据类型 NUMERIC 分配给 age，Amazon ML 会将评估数据中的 age 看作 TEXT 变量，而不是 NUMERIC。

有关与每种数据类型相关的统计信息，请参阅描述性统计信息。

为 Amazon ML 提供架构

每个数据源都需要一个架构。您可以从两种方式中选择所需方式来为 Amazon ML 提供架构：

- 允许 Amazon ML 推断输入数据文件中每个属性的数据类型并自动为您创建架构。
- 在上传 Amazon Simple Storage Service (Amazon S3) 数据时提供架构文件。

允许 Amazon ML 创建您的架构

使用 Amazon ML 控制台创建数据源时，Amazon ML 将根据变量的值使用简单规则来创建您的架构。我们强烈建议您检查 Amazon ML 创建的架构，并纠正不正确的数据类型。

提供架构

创建架构文件后，您需要使其对 Amazon ML 可用。您有两种选择：

1. 使用 Amazon ML 控制台提供架构。

   使用控制台创建您的数据源，并通过在输入数据文件的文件名中附加 .schema 扩展名来包含架构文件。例如，如果输入数据的 Amazon Simple Storage Service (Amazon S3) URI 为 s3://my-bucket-name/data/input.csv，架构的 URI 将为 s3://my-bucket-name/data/input.csv.schema。Amazon ML 会自动定位您提供的架构文件，而不是尝试通过您的数据推断出架构。

   要将文件目录作为 Amazon ML 的数据输入，请将 .schema 扩展名附加到您的目录路径中。例如，如果您的数据文件驻留在位置 s3://examplebucket/path/to/data/，架构的 URI 将为 s3://examplebucket/path/to/data/.schema。
2. 使用 Amazon ML API 提供架构。

   如果您计划调用 Amazon ML API 来创建数据源，则可将架构文件上传到 Amazon S3 中，然后将 URI 提供给 CreateDataSourceFromS3 API 的 DataSchemaLocationS3 属性中的文件。有关更多信息，请参阅 CreateDataSourceFromS3。

   您可以直接提供 CreateDataSource* APIs 负载中的架构，而不是先将其保存到 Amazon S3。为此，您可以将完整的架构字符串放入 DataSchema、CreateDataSourceFromS3 或

CreateDataSourceFromRDS API 的 CreateDataSourceFromRedshift 属性中。有关更多信息，请参阅 Amazon Machine Learning API 参考。

拆分数据

ML 模型的基本目标是在用于训练模型的数据实例之外，准确预测未来的数据实例。使用 ML 模型进行决策时，我们需要评估模型的预测性能。要使用 ML 模型未读取的数据评估其预测的质量，可以保留或拆分一部分我们已知答案的数据，用作未来数据，评估 ML 模型预测该数据的正确答案的质量如何。您可以将数据源拆分一部分作为训练数据源，一部分作为评估数据源。

Amazon ML 提供了三个选项用于拆分数据：

- 预拆分数据 - 在将数据上传到 Amazon Simple Storage Service (Amazon S3) 之前，您可以将数据拆分为两个数据输入位置，使用它们创建两个单独的数据源。
- Amazon ML 顺序拆分 - 您可以让 Amazon ML 在创建训练数据源和评估数据源时按顺序拆分数据。
- Amazon ML 随机拆分 - 您可以让 Amazon ML 在创建训练数据源和评估数据源时使用做种随机方法拆分数据。

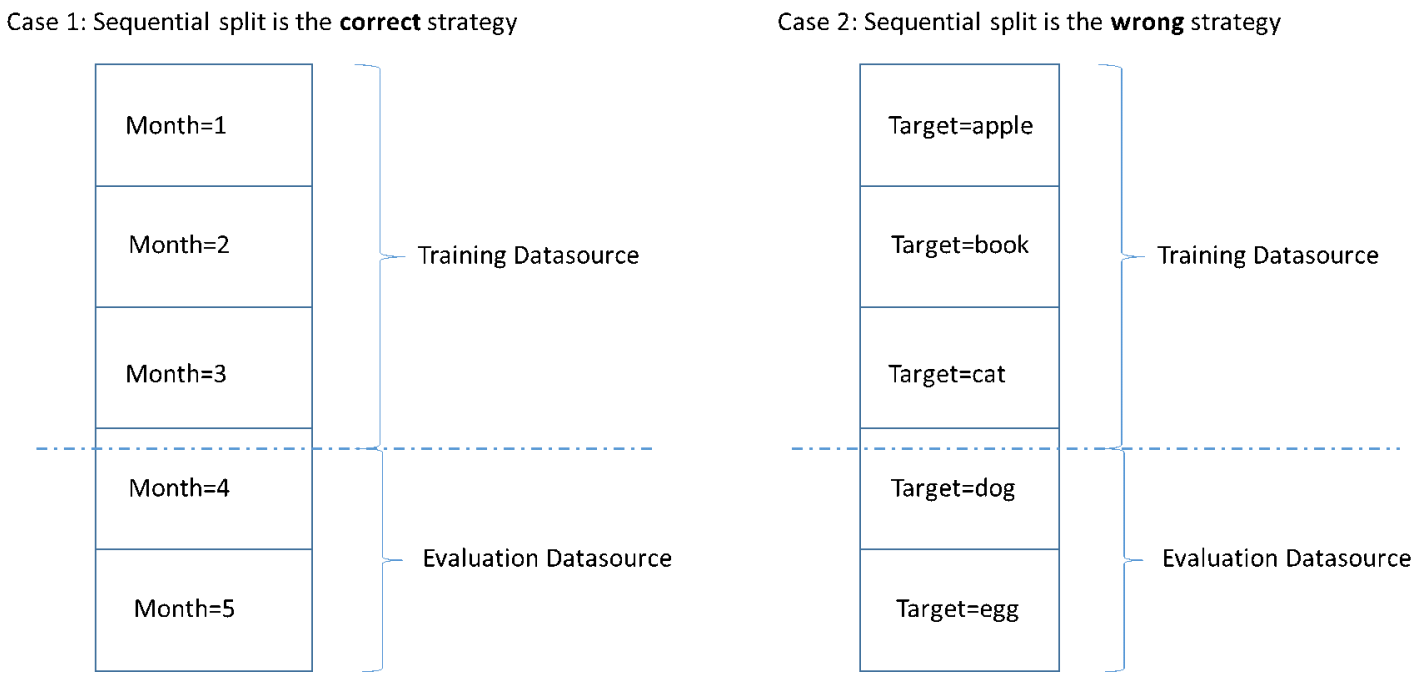
预拆分数据

如果您希望明确控制训练数据源和评估数据源中的数据，请将数据拆分到单独的数据位置，并为输入位置和评估位置创建单独的数据源。

按顺序拆分数据

为训练和评估拆分输入数据的一种简单方法是选择数据中未重叠的子集，同时保留数据记录的顺序。如果您希望针对特定日期或特定时间范围内的数据评估 ML 模型，这种方法非常有用。例如，假设您有过去五个月的客户参与数据，并希望使用此历史数据来预测下个月的客户参与情况。使用时间范围的开头进行训练并使用时间范围的结尾进行评估，相比从整个数据范围中提取记录数据，这种方法可以更准确地估算模型质量。

下图显示了您什么时候应使用顺序拆分策略，什么时候应使用随机策略。

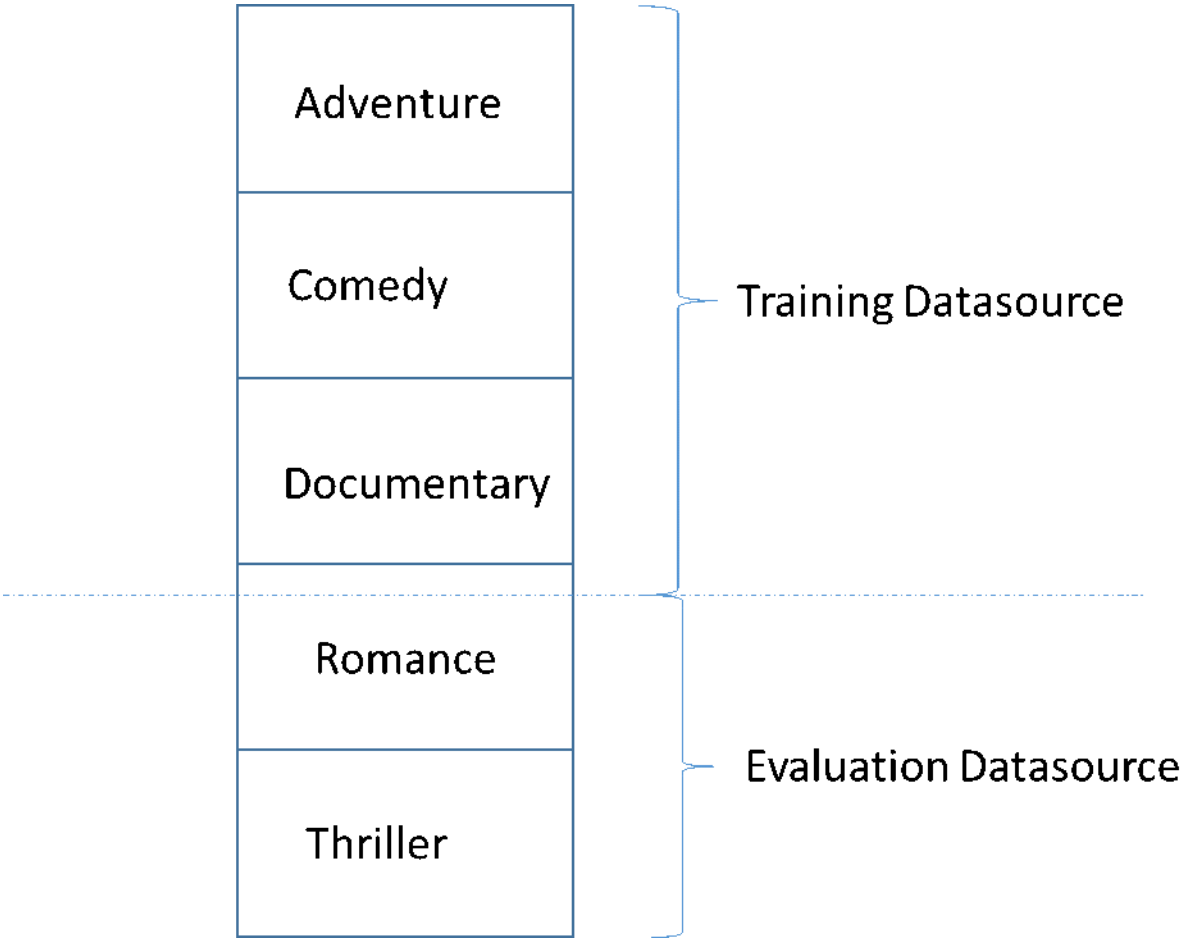


创建数据源时，您可以选择顺序拆分数据源，Amazon ML 使用前 70% 的数据进行训练，使用剩余的 30% 数据进行评估。这是使用 Amazon ML 控制台拆分数据时的默认方法。

随机拆分数据

将输入数据随机拆分为训练数据源和评估数据源，可确保这两种数据源中的数据分布相似。在您不需要保留输入数据的顺序时，选择此选项。

Amazon ML 使用做种的伪随机数字生成方法拆分数据。种子一部分基于输入字符串值，一部分基于数据本身的内容。默认情况下，Amazon ML 控制台使用输入数据的 S3 位置作为字符串。API 用户可以提供自定义字符串。这意味着只要 S3 存储桶和数据不变，Amazon ML 每次都使用相同的方法拆分数据。要更改 Amazon ML 拆分数据的方法，可以使用 `CreateDatasourceFromS3`、`CreateDatasourceFromRedshift` 或 `CreateDatasourceFromRDS` API 并为种子字符串提供值。使用这些 API 创建单独的数据源用于训练和评估时，必须为两个数据源使用相同的种子字符串值，并为一个数据源使用补充标记，确保训练数据和评估数据之间没有重叠。



开发高质量 ML 模型中的一个常见陷阱是，评估 ML 模型所用数据与训练所用数据不相似。例如，假设您使用 ML 预测电影的类型，并且您的训练数据包含冒险片、喜剧片和纪录片类型的电影。但是，您的评估数据只包含来自爱情片和惊悚片类型的数据。这种情况下，ML 模型未学习任何关于爱情片和惊悚片类型的信息，评估过程未评估模型的冒险片、喜剧片和纪录片模式学习模式效果如何。因此，类型信息无用，ML 模型的所有类型预测质量受损。模型和评估太过不同（描述性统计信息具有极大的差别）而无用。当输入数据按数据集中的一列排序然后按顺序拆分时，会出现这种情况。

如果您的训练数据源和评估数据源具有不同的数据分布，您可在模型评估中看到评估警报。有关评估警报的更多信息，请参阅评估警报。

如果您已将输入数据随机化（例如使用以下方法：在 Amazon S3 中随机乱序输入数据，或者在创建数据源时使用 Amazon Redshift SQL 查询的 `random()` 函数或 MySQL SQL 查询的 `rand()` 函数），则无需在 Amazon ML 中使用随机拆分。在这些情况下，您可以依靠顺序拆分选项，使用类似的分布创建训练数据源和评估数据源。

数据洞察

Amazon ML 计算输入数据的描述性统计信息，您可以使用这些信息来了解自己的数据。

描述性统计信息

Amazon ML 针对不同属性类型计算以下描述性统计信息：

数值：

- 分布直方图
- 无效值的数量
- 最小值、中值、平均值和最大值

二进制和分类：

- 计数（每个类别的不同值）
- 值分布直方图
- 最常用值
- 唯一值计数
- true 值的百分比（仅限二进制）
- 最突出单词
- 最常用单词

文本：

- 属性的名称
- 与目标的相关性（如果已设置目标）
- 总单词数
- 唯一单词数
- 一行中单词数的范围
- 单词长度范围
- 最突出单词

在 Amazon ML 控制台上访问数据洞察

在 Amazon ML 控制台上，您可以选择任意数据源的名称或 ID 以查看其数据洞察页面。此页面提供一些指标和可视化内容，让您了解与数据源关联的输入数据，包括以下信息：

- 数据摘要
- 目标分布
- 缺失值
- 无效值
- 按数据类型列出的变量摘要统计信息
- 按数据类型列出的变量分布

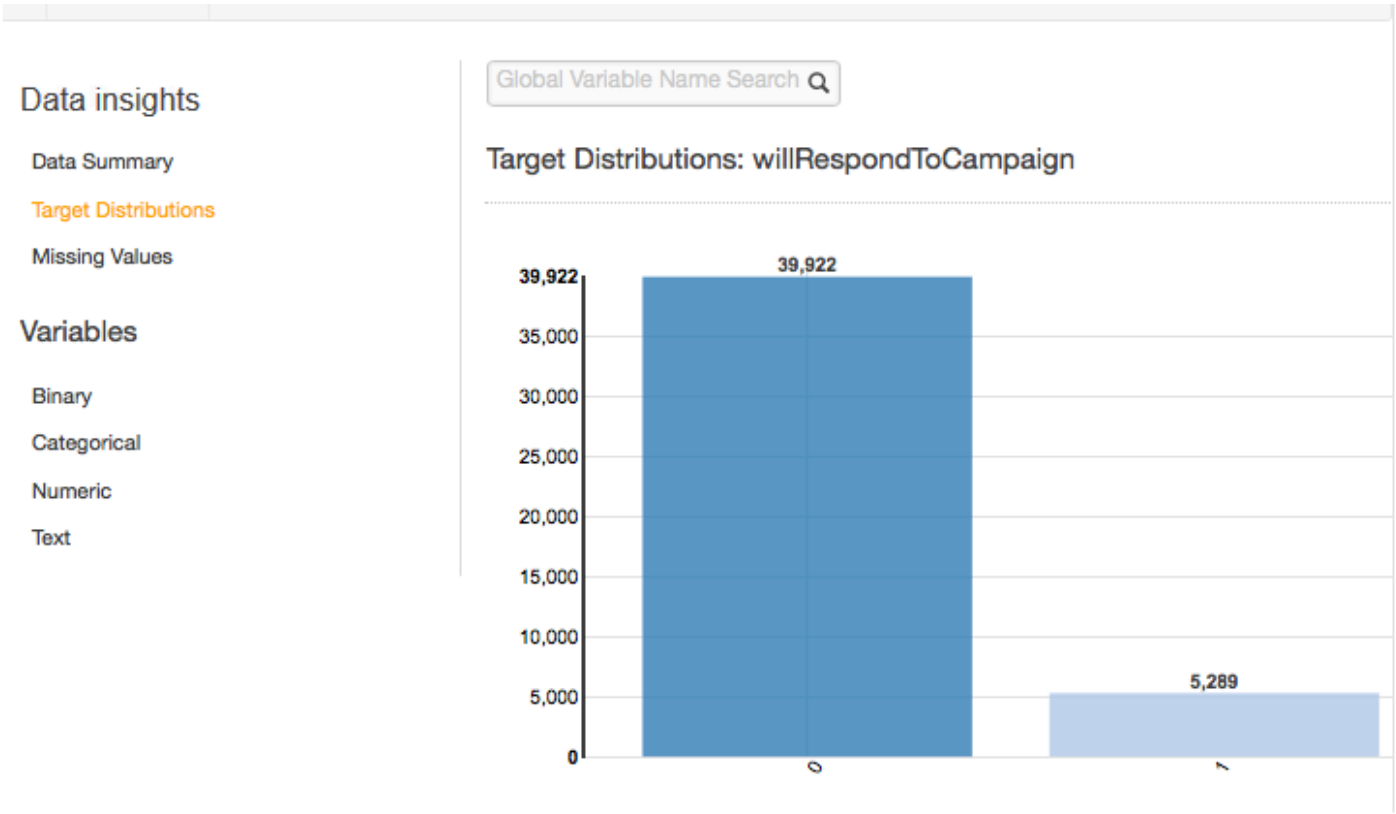
以下各部分更详细地描述了各指标和可视化内容。

数据摘要

数据源的数据摘要报告显示摘要信息，包括数据源 ID、名称、其完成位置、当前状态、目标属性、输入数据信息（S3 存储桶位置、数据格式、处理的记录数和处理期间遇到的错误记录数）以及按数据类型列出的变量数。

目标分布

目标分布报告显示数据源的目标属性的分布。在以下示例中，有 39922 次观察的 willRespondToCampaign 目标属性等于 0。这是未响应电子邮件营销活动的客户数。有 5289 次观察的 willRespondToCampaign 等于 1。这是响应电子邮件营销活动的客户数。



缺失值

缺失值报告列出输入数据中缺失值的属性。只有具有数字数据类型的属性才能具有缺失值。由于缺失值会影响 ML 模型的训练质量，我们建议尽可能提供缺失值。

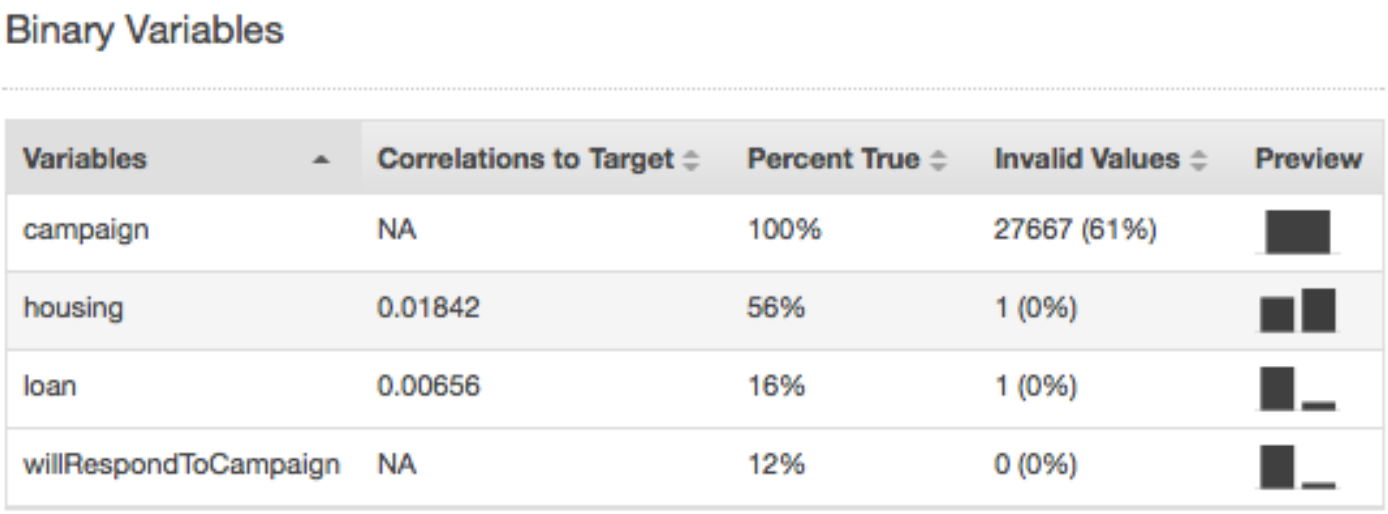
在 ML 模型训练期间，如果目标属性缺失，Amazon ML 会拒绝对应的记录。如果记录中存在目标属性，但其他数字属性的值缺失，则 Amazon ML 会忽略缺失值。在这种情况下，Amazon ML 创建替代属性并将其设置为 1，指示此属性缺失。这使得 Amazon ML 可以根据缺失值的出现来了解模式。

无效值

只有数字和二进制数据类型会出现无效的值。您可以通过在数据类型报告中查看变量的摘要统计信息来查找无效值。在以下示例中，持续时间数字属性中有一个无效值，二进制数据类型中有两个无效值（一个在 housing 属性中，一个在 loan 属性中）。

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	



变量-目标相关性

创建数据源之后，Amazon ML 可以评估数据源，并确定变量与目标之间的相关性或影响。例如，产品的价格可能会对是否为最佳卖家产生巨大影响，而产品尺寸带来的预测能力则可能很小。

通常的最佳实践是在训练数据中包括尽可能多的变量。但是，引入预测能力很小的多个变量会带来干扰，从而会对 ML 模型的质量和准确性产生负面影响。

在训练模型时，您可以移除影响较小的变量，从而提升模型的预测性能。您可以在配方中定义哪些变量可供机器学习过程使用，这是 Amazon ML 的转换机制。要了解有关配方的更多信息，请参阅机器学习的数据转换。

按数据类型列出的属性摘要统计信息

在数据洞察报告中，您可以按以下数据类型查看属性摘要统计信息：

- 二进制
- 分类
- 数值
- 文本

二进制数据类型的摘要统计信息显示所有二进制属性。Correlations to target 列显示在目标列与属性列之间共享的信息。Percent true 列显示值为 1 的观察的百分比。Invalid values 列显示无效值的数量以及各属性的无效值的百分比。Preview 列提供各属性的分布图的链接。

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%)	

分类数据类型的摘要统计信息显示所有分类属性以及唯一值的数量、最常用的值以及最不常用的值。Preview 列提供各属性的分布图的链接。

Categorical Variables

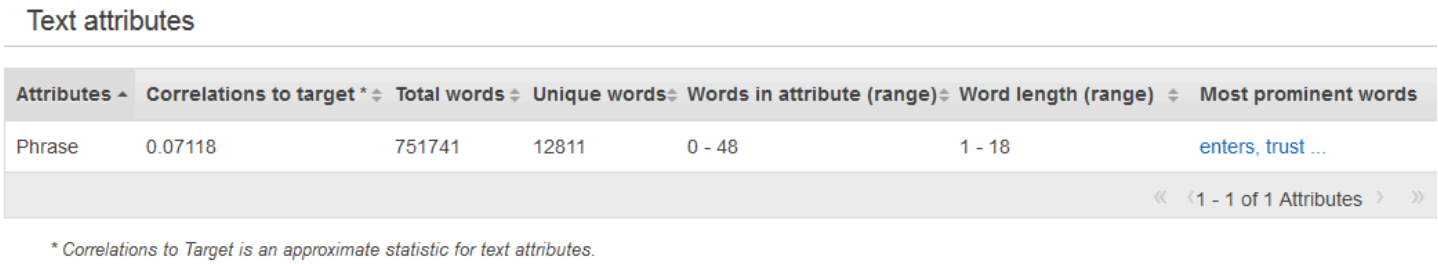
Variables	Correlations to Target	Unique Values	Most Frequent	Least Frequent	Preview
campaign	0.00433	49	1	39	
customerId	NA	45211	45211	1	
education	0.00355	5	secondary		
housing	0.01846	4	1		
jobId	0.00671	13	blue-collar		
willRespondToCampaign	NA	3	0		

数字数据类型的摘要统计信息显示所有数字属性以及缺失值的数量、无效值、值范围、平均值和中值。Preview 列提供各属性的分布图的链接。

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	

文本数据类型的摘要统计信息显示所有文本属性、该属性中的单词总数、该属性中的唯一单词数、属性中的单词数范围、单词长度的范围以及最突出单词。Preview 列提供各属性的分布图的链接。



下一个示例显示了名为“review”的文本变量的文本数据类型统计信息，它有四条记录。

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

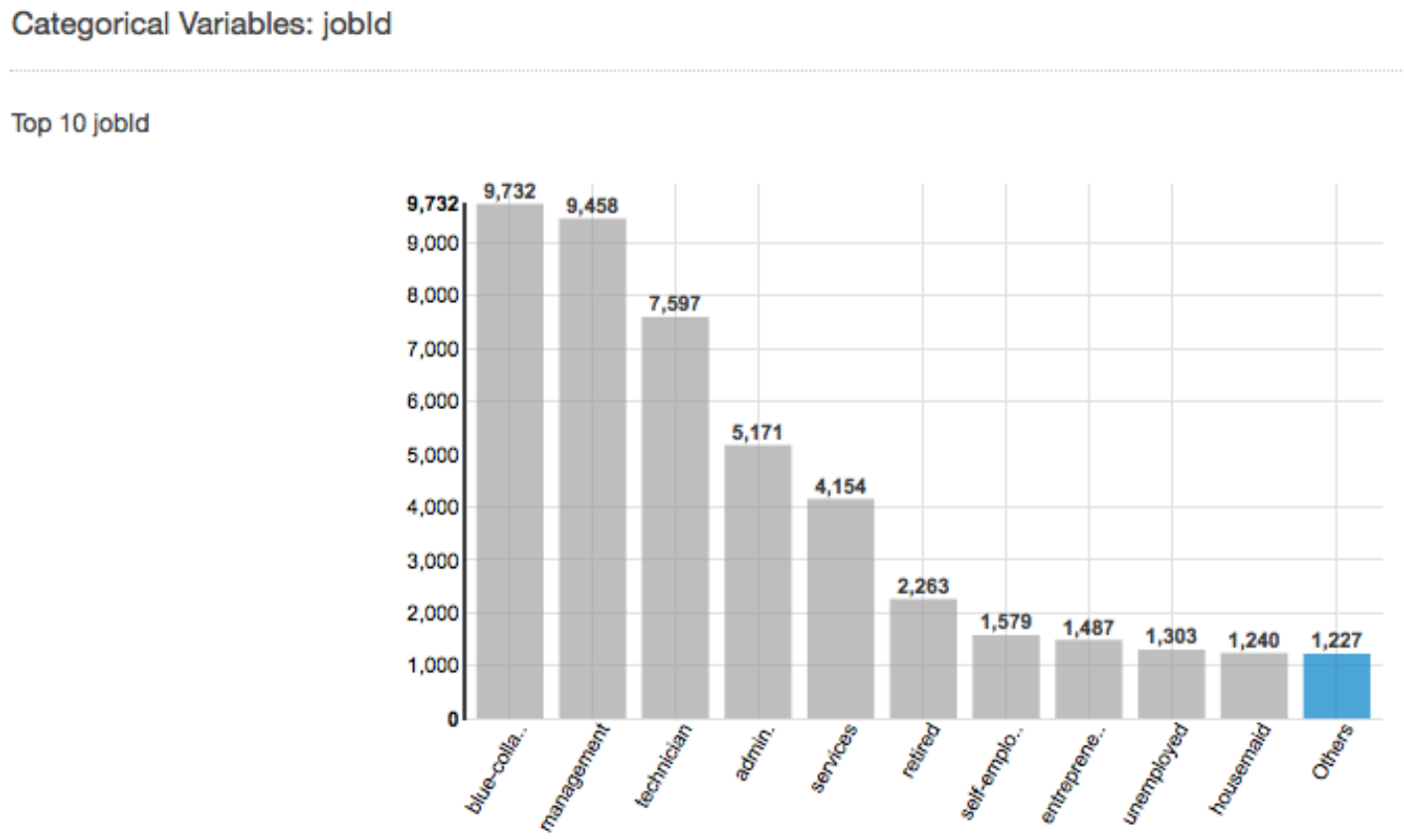
此示例的各列将显示以下信息。

- Attributes 列显示变量的名称。在本示例中，此列将显示“review”。
- 只有指定了目标，才存在 Correlations to target 列。相关性用于衡量此属性提供的有关目标的信息量。相关性越高，此属性揭示的有关目标的信息也越多。相关性按照文本属性与目标的简化表示形式之间的交互信息来衡量。
- Total words 列显示通过令牌化各个记录生成的单词（以空格分隔单词）数量。在本示例中，此列将显示“12”。
- Unique words 列显示属性的唯一单词数。在本示例中，此列将显示“10”。
- Words in attribute (range) 列显示属性的单行中的单词数。在本示例中，此列将显示“0-6”。
- Word length (range) 列显示单词中字符数的范围。在本示例中，此列将显示“2-11”。
- Most prominent words 列显示在属性中显示的单词排名列表。如果存在目标属性，则单词按照与目标的相关性排名，也就是相关性最高的单词列在最前。如果数据中没有目标，则单词按照其熵排名。

了解分类和二进制属性的分布

您可单击与某个分类属性或二进制属性关联的 Preview 链接来查看属性的分布，以及输入文件中属性的各分类值的示例数据。

例如，以下屏幕截图显示分类属性 jobId 的分布。分布显示了前 10 个分类值，所有其他值归为“other”组。它对前 10 的各个分类值与输入文件中包含相应值的观察数进行排名，并提供查看输入数据文件中示例观察的链接。

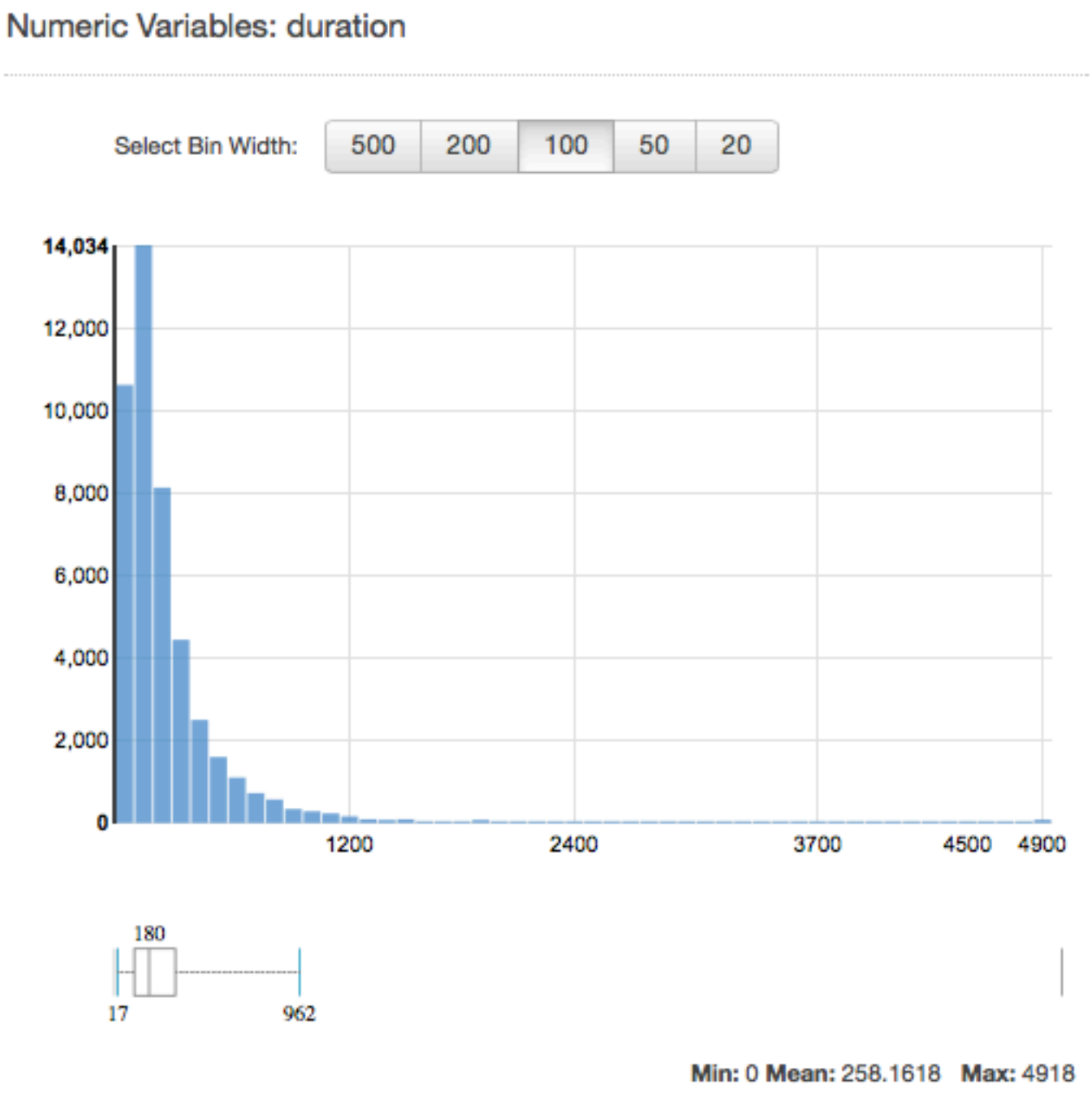


Ranking	Category	Count	
1	blue-collar	9732	Sample data
2	management	9458	Sample data
3	technician	7597	Sample data

了解数字属性的分布

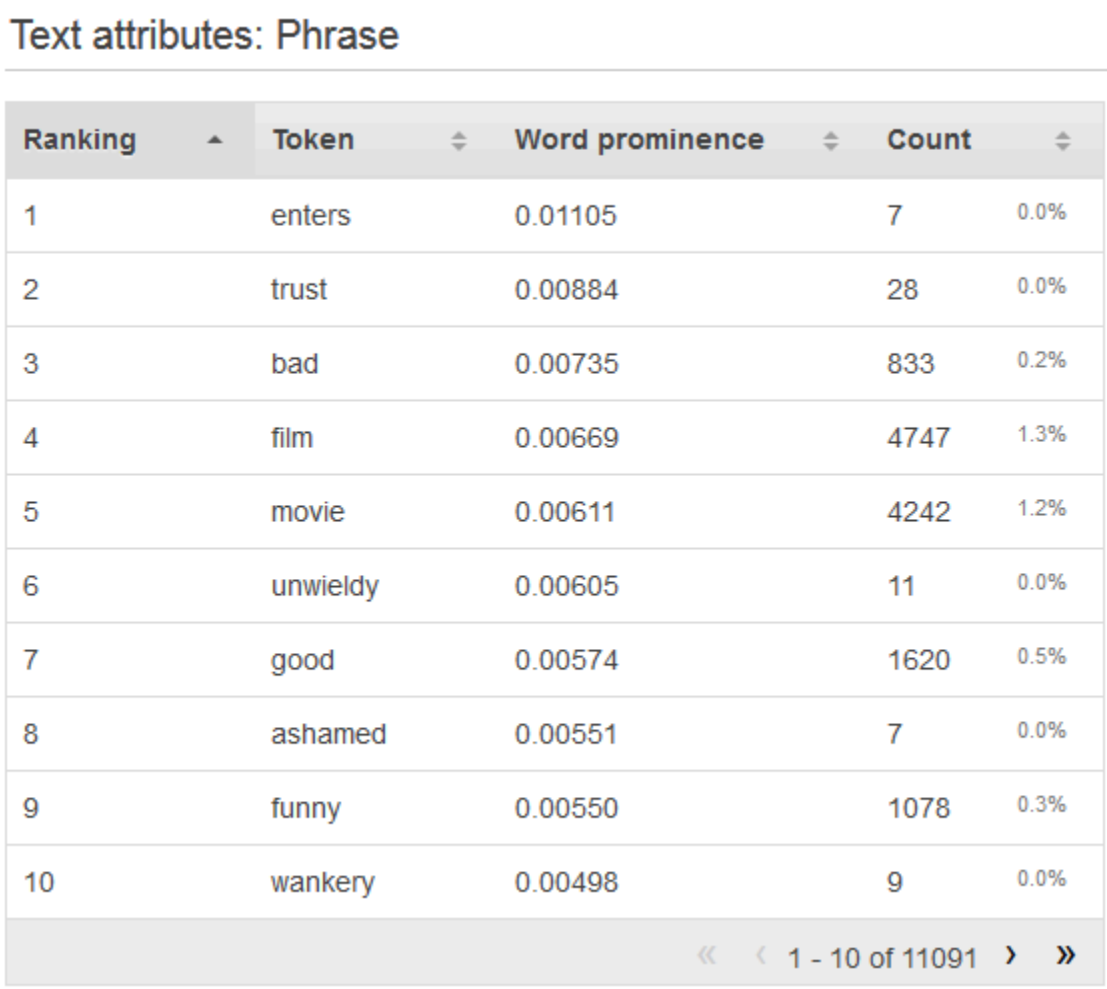
要查看数字属性的分布，请单击属性的 Preview 链接。查看数字属性的分布时，您可以选择 500、200、100、50 或 20 的箱大小。箱大小越大，将显示的条形图的数量就越少。此外，较大的箱大小会导致较粗糙的分布分辨率。与之相对，将存储桶大小设置为 20 会提升所示分布的分辨率。

图中还将显示最小值、平均值和最大值，如以下屏幕截图中所示。



了解文本属性的分布

要查看文本属性的分布，请单击属性的 Preview 链接。查看文本属性的分布时，您会看到以下信息。



排名

文本令牌按照所传达的信息量排名，从最具信息性到最不具信息性。

令牌

令牌显示统计信息行所相关的输入文本中的单词。

单词突出部分

如果存在目标属性，单词按照与目标的相关性排名，这样相关性最高的单词列在最前。如果数据中不存在目标，则单词按照其熵（即它们可以传达的信息量）排名。

计数

计数显示其中出现了令牌的输入记录的数量。

计数百分比

计数百分比显示其中出现了令牌的输入数据行的百分比。

将 Amazon S3 与 Amazon ML 结合使用

Amazon Simple Storage Service (Amazon S3) 是一种面向 Internet 的存储服务。您可以通过 Amazon S3 随时在 Web 上的任何位置存储和检索的任意大小的数据。Amazon ML 将 Amazon S3 作为执行以下任务的主数据存储库：

- 访问您的输入文件以创建用于训练和评估 ML 模型的数据源对象。
- 访问您的输入文件以生成批量预测。
- 在使用您的 ML 模型生成批量预测时，将预测文件输出到您指定的 S3 存储桶中。
- 将 Amazon Redshift 或 Amazon Relational Database Service (Amazon RDS) 存储的数据复制到 .csv 文件中，然后将其上传到 Amazon S3。

要让 Amazon ML 执行这些任务，您必须授予 Amazon ML 访问您的 Amazon S3 数据的权限。

> **Note**
>
> 您无法将批量预测文件输出到仅接受服务器端加密文件的 S3 存储桶。通过确认在请求中没有 `Deny` 标头时，策略对 `s3:PutObject` 操作不会有 `s3:x-amz-server-side-encryption` 效果，确保您的存储桶策略允许上传未加密的文件。有关 S3 服务器端加密存储桶策略的更多信息，请参阅 Amazon Simple Storage Service 用户指南中的使用服务器端加密保护数据。

将您的数据上传到 Amazon S3

您必须将输入数据上传到 Amazon Simple Storage Service (Amazon S3)，因为 Amazon ML 会从 Amazon S3 位置读取数据。您可以将您的数据直接上传到 Amazon S3（例如，从您的计算机上传），也可以让 Amazon ML 将您存储在 Amazon Redshift 或 Amazon Relational Database Service (RDS) 中的数据复制到 .csv 文件，然后将文件上传到 Amazon S3。

有关从 Amazon Redshift 或 Amazon RDS 复制数据的更多信息，请分别参阅将 Amazon Redshift 与 Amazon ML 结合使用或将 Amazon RDS 与 Amazon ML 结合使用。

本节的剩余部分介绍如何将输入数据从您的计算机直接上传到 Amazon S3。开始本节的操作过程之前，您需要将数据保存在 .csv 文件中。有关如何正确设置 .csv 文件的格式以便 Amazon ML 使用的信息，请参阅了解 Amazon ML 的数据格式。

将您的数据从计算机上传到 Amazon S3

1. 登录 AWS 管理控制台，并通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3。
2. 创建存储桶或选择现有的存储桶。
   a. 要创建存储桶，请选择创建存储桶。为您的存储桶命名，选择区域（您可以选择任何可用区域），然后选择创建。有关更多信息，请参阅 Amazon Simple Storage 入门指南 中的创建存储桶。
   b. 要使用现有的存储桶，通过在所有存储桶列表中选择存储桶来搜索存储桶。当存储桶名称出现后，选择该存储桶，然后选择上传。
3. 在上传对话框中，选择添加文件。
4. 导航到包含输入数据 .csv 文件的文件夹，然后选择打开。

权限

要授权 Amazon ML 访问您的一个 S3 存储桶，您必须编辑存储桶策略。

有关授权 Amazon ML 从您在 Amazon S3 中的存储桶中读取数据的信息，请参阅为 Amazon ML 授予从 Amazon S3 读取您的数据的权限。

有关授权 Amazon ML 将批量预测结果输出到您在 Amazon S3 中的存储桶的信息，请参阅向 Amazon ML 授予输出预测到 Amazon S3 的权限。

有关管理 Amazon S3 资源访问权限的信息，请参阅 Amazon S3 开发人员指南。

根据 Amazon Redshift 中的数据创建 Amazon ML 数据源

如果您有数据存储在 Amazon Redshift 中，可以使用 Amazon Machine Learning (Amazon ML) 控制台中的创建数据源向导创建数据源对象。利用 Amazon Redshift 数据创建数据源时，您可以指定包含您数据的集群和 SQL 查询来检索您的数据。Amazon ML 通过对集群调用 Amazon Redshift `Unload` 命令来执行查询。Amazon ML 将结果存储在您选择的 Amazon Simple Storage Service (Amazon S3) 位置，然后使用 Amazon S3 中存储的数据创建数据源。数据源、Amazon Redshift 集群和 S3 存储桶必须全部位于同一区域。

> **Note**
>
> Amazon ML 不支持利用私有 VPC 中的 Amazon Redshift 集群创建数据源。集群必须有公有 IP 地址。

主题

“Create Datasource”向导的必需参数

要允许 Amazon ML 连接到您的 Amazon Redshift 数据库并代表您读取数据，您必须提供以下内容：

- Amazon Redshift `ClusterIdentifier`
- Amazon Redshift 数据库名称
- Amazon Redshift 数据库凭证（用户名和密码）
- Amazon ML Amazon Redshift AWS Identity and Access Management (IAM) 角色
- Amazon Redshift SQL 查询
- （可选）Amazon ML 架构的位置
- Amazon S3 暂存位置（Amazon ML 在创建数据源之前将数据放在此位置）

此外，您需要确保创建 Amazon Redshift 数据源（无论是通过控制台还是使用 `CreateDatasourceFromRedshift` 操作进行创建）的 IAM 用户或角色拥有 `iam:PassRole` 权限。

Amazon Redshift**`ClusterIdentifier`**

使用此区分大小写的参数启用 Amazon ML 以查找并连接到您的集群。您可以从 Amazon Redshift 控制台获取集群标识符（名称）。有关集群的更多信息，请参阅 Amazon Redshift 集群。

Amazon Redshift 数据库名称

使用此参数告诉 Amazon ML，Amazon Redshift 集群中的哪个数据库包含您要用作数据源的数据。

Amazon Redshift 数据库凭证

使用这些参数来指定将在其上下文中执行安全查询的 Amazon Redshift 数据库用户的用户名和密码。

> **Note**
>
> Amazon ML 需要 Amazon Redshift 用户名和密码才能连接到您的 Amazon Redshift 数据库。在将数据卸载到 Amazon S3 之后，Amazon ML 永远不会重复使用您的密码，也不会进行存储。

Amazon ML Amazon Redshift 角色

使用此参数可指定 IAM 角色的名称，Amazon ML 应使用该角色配置 Amazon Redshift 集群的安全组以及 Amazon S3 暂存位置的存储桶策略。

如果您没有可访问 Amazon Redshift 的 IAM 角色，Amazon ML 可以为您创建角色。当 Amazon ML 创建角色时，它会创建客户管理型策略并将其附加到 IAM 角色。Amazon ML 创建的策略授予 Amazon ML 权限以仅访问您指定的集群。

如果您已有一个 IAM 角色访问 Amazon Redshift，您可以键入该角色的 ARN，或者从下拉列表中选择该角色。具有 Amazon Redshift 访问权限的 IAM 角色在下拉菜单顶部列出。

IAM 角色必须具有以下内容：

```
{
    "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/*" }
        }
    }]
```

```
}
```

有关客户管理型策略的更多信息，请参阅 IAM 用户指南中的客户管理型策略。

Amazon Redshift SQL 查询

使用此参数可指定 SQL SELECT 查询，Amazon ML 对您的 Amazon Redshift 数据库执行该查询以选择数据。Amazon ML 使用 Amazon Redshift UNLOAD 操作将您的查询结果安全地复制到 Amazon S3 位置。

> **Note**
>
> 当输入记录处于随机顺序（乱序）时，Amazon ML 最适合。您可以通过使用 Amazon Redshift random() 函数将您的 Amazon Redshift SQL 查询结果轻松地随机化。例如，假设原始查询如下：
>
> ```
> "SELECT col1, col2, … FROM training_table"
> ```
>
> 您可以通过如下所示更新查询来嵌入随机乱序：
>
> ```
> "SELECT col1, col2, … FROM training_table ORDER BY random()"
> ```

架构位置（可选）

使用此参数可以为 Amazon ML 将导出的 Amazon Redshift 数据指定架构的 Amazon S3 路径。

如果您没有提供数据源的架构，Amazon ML 控制台会根据 Amazon Redshift SQL 查询的数据架构自动创建 Amazon ML 架构。Amazon ML 架构比 Amazon Redshift 架构拥有更少的数据类型，因此并不是一对一转换。Amazon ML 控制台使用以下转换方案将 Amazon Redshift 数据类型转换为 Amazon ML 数据类型。

Amazon Redshift 数据类型	Amazon Redshift 别名	Amazon ML 数据类型
SMALLINT	INT2	NUMERIC
INTEGER	INT、INT4	NUMERIC
BIGINT	INT8	NUMERIC

Amazon Redshift 数据类型	Amazon Redshift 别名	Amazon ML 数据类型
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
DATE		TEXT
TIMESTAMP	TIMESTAMP WITHOUT TIME ZONE	TEXT

要转换为 Amazon ML `Binary` 数据类型，您数据中的 Amazon Redshift 布尔值必须是支持的 Amazon ML 二进制值。如果您的布尔数据类型具有不支持的值，Amazon ML 会将其转换为最可能的具体数据类型。例如，如果 Amazon Redshift 布尔值具有值 `0`、`1` 和 `2`，Amazon ML 会将布尔值转换为 `Numeric` 数据类型。有关支持的二进制值的更多信息，请参阅 使用 AttributeType 字段。

如果 Amazon ML 无法指出数据类型，则默认为 `Text`。

Amazon ML 转换架构后，您可以在“创建数据源”向导中查看和更正分配的 Amazon ML 数据类型，并在 Amazon ML 创建数据源之前修改架构。

Amazon S3 暂存位置

使用此参数可指定 Amazon S3 暂存位置的名称，将 Amazon Redshift SQL 查询结果存储在该暂存位置。创建数据源之后，Amazon ML 使用暂存位置中的数据而不是返回到 Amazon Redshift。

> **Note**
>
> 由于 Amazon ML 代入 Amazon ML Amazon Redshift 角色定义的 IAM 角色，因此 Amazon ML 有权访问指定 Amazon S3 暂存位置中的任何对象。因此，建议您在 Amazon S3

暂存位置中仅存储那些不包含敏感信息的文件。例如，如果您的根存储桶是 s3://mybucket/，我们建议您创建一个位置，在其中仅存储您希望 Amazon ML 访问的文件，如 s3://mybucket/AmazonMLInput/。

利用 Amazon Redshift 数据创建数据源（控制台）

Amazon ML 控制台提供两种方式来使用 Amazon Redshift 数据创建数据源。您可以通过完成“创建数据源”向导创建数据源，或者，如果您已经利用 Amazon Redshift 数据创建数据源，您可以复制原始数据源并修改其设置。复制数据源可以轻松创建多个相似的数据源。

有关使用 API 创建数据源的信息，请参阅 CreateDataSourceFromRedshift。

有关以下步骤中的参数的更多信息，请参阅 “Create Datasource”向导的必需参数。

主题

创建数据源（控制台）

要将数据从 Amazon Redshift 卸载到 Amazon ML 数据源，请使用“创建数据源”向导。

利用 Amazon Redshift 中的数据创建数据源

1. 打开 Amazon Machine Learning 控制台，网址为 https://console.aws.amazon.com/machinelearning/。
2. 在 Amazon ML 控制面板上的实体下，选择新建...，然后选择数据源。
3. 在输入数据页面上，选择 Amazon Redshift。
4. 在“创建数据源”向导中，对于集群标识符，请键入您的集群的名称。
5. 对于数据库名称，请键入 Amazon Redshift 数据库的名称。
6. 对于数据库用户名，请键入数据库用户名。
7. 对于数据库密码，请键入数据库密码。
8. 对于IAM 角色，请选择您的 IAM 角色。如果您还没有角色，请选择创建新的角色。Amazon ML 会为您创建一个 IAM Amazon Redshift 角色。

9. 要测试您的 Amazon Redshift 设置，请选择测试访问（在 IAM 角色旁边）。如果 Amazon ML 无法使用提供的设置连接到 Amazon Redshift，则您无法继续创建数据源。有关问题排查帮助，请参阅纠正错误。
10. 对于 SQL 查询，键入您的 SQL 查询。
11. 对于架构位置，请选择您是否希望 Amazon ML 为您创建架构。如果您已经自己创建了架构，请键入您的架构文件的 Amazon S3 路径。
12. 对于 Amazon S3 暂存位置，请键入存储桶的 Amazon S3 路径，您希望 Amazon ML 将所卸载数据从 Amazon Redshift 放入该存储桶。
13. （可选）对于数据源名称，请键入您数据源的名称。
14. 选择验证。Amazon ML 将验证它是否能连接到您的 Amazon Redshift 数据库。
15. 在架构页面上，检查所有属性的数据类型并根据需要进行纠正。
16. 选择继续。
17. 如果您希望使用此数据源创建或评估 ML 模型，则对于是否计划使用此数据集创建或评估 ML 模型?，请选择是。如果您选择是，请选择目标行。有关目标的信息，请参阅使用 targetAttributeName 字段。

    如果您希望使用此数据源以及您已创建的模型来创建预测，请选择否。
18. 选择继续。
19. 对于您的数据是否包含标识符?，如果您的数据不包含行标识符，请选择否。

    如果您的数据包含行标识符，请选择是。有关行标识符的信息，请参阅使用 rowID 字段。
20. 选择审核。
21. 在审核页上，检查您的设置，然后选择完成。

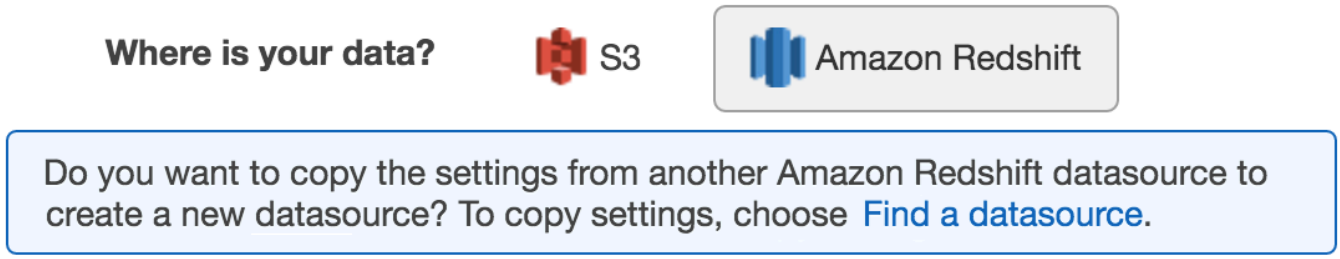
创建数据源后，您可以使用它create an ML model。创建模型后，您可以使用数据源evaluate an ML model或generate predictions。

复制数据源（控制台）

当您想要创建一个与现有数据源类似的数据源时，您可以使用 Amazon ML 控制台复制原始数据源并修改其设置。例如，您可以选择从现有的数据源开始，然后修改数据架构，以更准确地匹配您的数据；更改用于从 Amazon Redshift 卸载数据的 SQL 查询；或者指定不同的 AWS Identity and Access Management (IAM) 用户来访问 Amazon Redshift 集群。

复制和修改 Amazon Redshift 数据源

1. 打开 Amazon Machine Learning 控制台，网址为 https://console.aws.amazon.com/machinelearning/。
2. 在 Amazon ML 控制面板上的实体下，选择新建...，然后选择数据源。
3. 在输入数据页面上，对于您的数据位于何处?，选择 Amazon Redshift。如果您已经利用 Amazon Redshift 数据创建了一个数据源，您可以选择从其他数据源复制设置。



   如果您还没有利用 Amazon Redshift 数据创建数据源，则不会显示此选项。
4. 选择查找数据源。
5. 选择要复制的数据源，然后选择复制设置。Amazon ML 使用原始数据源的设置自动填充大部分数据源设置。它不会从原始数据源复制数据库密码、架构位置或数据源名称。
6. 修改您希望更改的任何自动填充设置。例如，如果您希望更改 Amazon ML 从 Amazon Redshift 卸载的数据，请更改 SQL 查询。
7. 对于数据库密码，请键入数据库密码。Amazon ML 不会存储或重新使用您的密码，因此，您必须始终提供密码。
8. （可选）对于架构位置，Amazon ML 预先为您选择我希望 Amazon ML 生成一个推荐的架构。如果您已经创建架构，请选择我希望使用已在 Amazon S3 中创建和存储的架构，然后键入 Amazon S3 中架构文件的路径。
9. （可选）对于数据源名称，请键入您数据源的名称。否则，Amazon ML 会为您生成新的数据源名称。
10. 选择验证。Amazon ML 将验证它是否能连接到您的 Amazon Redshift 数据库。
11. （可选）如果 Amazon ML 为您推断了架构，请在架构页面上检查所有属性的数据类型，并根据需要进行更正。
12. 选择继续。
13. 如果您希望使用此数据源创建或评估 ML 模型，则对于是否计划使用此数据集创建或评估 ML 模型?，请选择是。如果您选择是，请选择目标行。有关目标的信息，请参阅使用 targetAttributeName 字段。

如果您希望使用此数据源以及您已创建的模型来创建预测，请选择否。

14. 选择继续。
15. 对于您的数据是否包含标识符?，如果您的数据不包含行标识符，请选择否。

    如果您的数据包含行标识符，请选择是，然后选择您要用作标识符的行。有关行标识符的信息，请参阅使用 rowID 字段。
16. 选择审核。
17. 检查设置，然后选择完成。

创建数据源后，您可以使用它create an ML model。创建模型后，您可以使用数据源evaluate an ML model或generate predictions。

Amazon Redshift 问题排查

在您创建 Amazon Redshift 数据源、ML 模型和评估时，Amazon Machine Learning (Amazon ML) 在 Amazon ML 控制台中报告您 Amazon ML 对象的状态。如果 Amazon ML 返回错误消息，请使用以下信息和资源来排查问题。

有关 Amazon ML 的一般问题的答案，请参阅 Amazon Machine Learning 常见问题。您还可以在 Amazon Machine Learning 论坛上搜索答案和发布问题。

主题

- 纠正错误
- 联系 AWS Support

纠正错误

角色的格式无效。请提供有效的 IAM 角色。例如，arn:aws:iam::YourAccountID:role/YourRedshiftRole。

原因

您的 IAM 角色的 Amazon 资源名称 (ARN) 格式不正确。

解决方案

在“Create Datasource”向导中，更正您角色的 ARN。有关格式化角色 ARN 的信息，请参阅 IAM 用户指南中的 IAM ARN。IAM 角色 ARN 的区域可选。

角色无效。Amazon ML 无法代入 <角色 ARN> IAM 角色。请提供有效的 IAM 角色并使其可供 Amazon ML 访问。

原因

您的角色未设置为允许 Amazon ML 代入它。

解决方案

在 IAM 控制台中，编辑您的角色使其具有信任策略，允许 Amazon ML 代入附加到其上的角色。

此 <用户 ARN> 用户未授权传递 <角色 ARN> IAM 角色。

原因

您的 IAM 用户没有允许传递角色到 Amazon ML 的权限策略。

解决方案

附加权限策略到您的 IAM 用户，以允许您传递角色到 Amazon ML。您可在 IAM 控制台中将权限策略附加到您的 IAM 用户。

不允许跨账户传递 IAM 角色。IAM 角色必须属于此账户。

原因

您不能传递属于其他 IAM 账户的角色。

解决方案

登录您创建角色时使用的 AWS 账户。您可在 IAM 控制台中查看您的 IAM 角色。

指定的角色无权执行操作。提供一个角色，该角色具有策略向 Amazon ML 提供了所需权限。

原因

您的 IAM 角色无权执行请求的操作。

解决方案

在 IAM 控制台中编辑附加到您角色的权限策略以提供所需的权限。

Amazon ML 无法在该 Amazon Redshift 集群上使用指定的 IAM 角色配置安全组。

原因

您的 IAM 角色没有所需的权限来配置 Amazon Redshift 安全集群。

解决方案

在 IAM 控制台中编辑附加到您角色的权限策略以提供所需的权限。

Amazon ML 尝试在集群上配置安全组时出错。请稍后重试。

原因

Amazon ML 尝试连接到您的 Amazon Redshift 集群时遇到问题。

解决方案

确保您在“Create Datasource”向导中提供的 IAM 角色具有全部必需权限。

集群 ID 格式无效。集群 ID 必须以字母开头，并且必须只包含字母数字字符和连字符。其中不能包含两个连续的连字符，也不能以连字符结束。

原因

您的 Amazon Redshift 集群 ID 格式不正确。

解决方案

在“Create Datasource”向导中，更正您的集群 ID，使其仅包含字母数字字符和连字符，并且不包含两个连续的连字符或以连字符结束。

没有 <Amazon Redshift 集群名称> 集群，或者集群与您的 Amazon ML 服务不在相同区域。指定与此 Amazon ML 位于相同区域中的集群。

原因

由于您的 Amazon Redshift 集群不在您创建 Amazon ML 数据源的区域中，Amazon ML 找不到该集群。

解决方案

确保 Amazon Redshift 控制台集群页面中存在您的集群，您在 Amazon Redshift 集群所在的区域中创建了数据源，并且在“创建数据源”向导中指定了集群 ID。

Amazon ML 无法读取您的 Amazon Redshift 集群中的数据。提供正确的 Amazon Redshift 集群 ID。

原因

Amazon ML 无法读取您指定的 Amazon Redshift 集群中的数据。

解决方案

在“创建数据源”向导中，指定正确的 Amazon Redshift 集群 ID，确保您在与 Amazon Redshift 集群相同的区域中创建了数据源，并且您的集群在 Amazon Redshift 集群页面上列出。

<Amazon Redshift 集群名称> 集群不可公开访问。

原因

Amazon ML 无法访问您的集群，因为该集群不可公开访问，并且没有公共 IP 地址。

解决方案

请使集群可公开访问并向其提供公共 IP 地址。有关如何使集群可公开访问的信息，请参阅 Amazon Redshift 管理指南中的修改集群。

<Redshift> 集群状态对 Amazon ML 不可用。使用 Amazon Redshift 控制台查看和解决此集群状态问题。集群状态必须为“Available”。

原因

Amazon ML 无法查看集群状态。

解决方案

确保您的集群可用。有关检查集群状态的信息，请参阅 Amazon Redshift 管理指南中的获取集群状态概览。有关重启集群以使其可用的信息，请参阅 Amazon Redshift 管理指南中的重启集群。

此集群中没有 <数据库名称> 数据库。确保数据库名称正确或者指定其他集群和数据库。

原因

Amazon ML 在指定集群中找不到指定的数据库。

解决方案

确保在“Create Datasource”向导中输入的数据库名称正确，或者指定正确的集群和数据库名称。

Amazon ML 无法访问您的数据库。为数据库用户 <用户名> 提供有效的密码。

原因

您在“创建数据源”向导中提供用于允许 Amazon ML 访问 Amazon Redshift 数据库的密码不正确。

解决方案

为您的 Amazon Redshift 数据库用户提供正确的密码。

Amazon ML 尝试验证查询时出错。

原因

您的 SQL 查询有问题。

解决方案

确保您的查询是有效 SQL。

执行 SQL 查询时出错。验证数据库名称和提供的查询。根本原因：{serverMessage}。

原因

Amazon Redshift 无法运行查询。

解决方案

确保您在“Create Datasource”向导中指定了正确的数据库名称并且查询是有效 SQL。

执行 SQL 查询时出错。根本原因：{serverMessage}。

原因

Amazon Redshift 找不到指定的表。

解决方案

确保您的 Amazon Redshift 集群数据库中存在您在“创建数据源”向导中指定的表，并且您输入了正确的集群 ID、数据库名称和 SQL 查询。

联系 AWS Support

如果您拥有 AWS Premium Support，则可在 AWS Support 中心创建技术支持案例。

使用来自 Amazon RDS 数据库的数据创建 Amazon ML 数据源

Amazon ML 允许您从存储在 Amazon Relational Database Service (Amazon RDS) 内 MySQL 数据库中的数据创建数据源对象。执行此操作时，Amazon ML 创建执行您所指定 SQL 查询的 AWS Data Pipeline 对象，并将输出放在您选择的 S3 存储桶上。Amazon ML 使用该数据来创建数据源。

> **Note**
>
> Amazon ML 仅支持 VPC 中的 MySQL 数据库。

在 Amazon ML 读取您的输入数据之前，您必须将该数据导出到 Amazon Simple Storage Service (Amazon S3)。您可以设置 Amazon ML 使用 API 为您执行导出。（RDS 限制为 API，并且不可从控制台使用。）

要让 Amazon ML 连接到 Amazon RDS 中您的 MySQL 数据库并代表您读取数据，您需要提供以下内容：

- RDS 数据库实例标识符
- MySQL 数据库名称
- 用于创建、激活和执行数据管道的 AWS Identity and Access Management (IAM) 角色
- 数据库用户凭证：
  - 用户名称
  - 密码
- AWS Data Pipeline 安全信息：
  - IAM 资源角色
  - IAM 服务角色
- Amazon RDS 安全信息：
  - 子网 ID
  - 安全组 ID
- 指定您创建数据源所用数据的 SQL 查询
- 用于存储查询结果的 S3 输出位置（存储桶）
- （可选）数据架构文件的位置

此外，您需要确保使用 CreateDataSourceFromRDS 操作创建 Amazon RDS 数据源的 IAM 用户或角色具有 `iam:PassRole` 权限。有关更多信息，请参阅使用 IAM 控制对 Amazon ML 资源的访问。

主题

RDS 数据库实例标识符

RDS 数据库实例标识符是您提供的唯一名称，用于标识 Amazon ML 在与 Amazon RDS 交互时应使用的数据库实例。您可以在 Amazon RDS 控制台找到 RDS 数据库实例标识符。

MySQL 数据库名称

MySQL 数据库名称指定 RDS 数据库实例中 MySQL 数据库的名称。

数据库用户凭证

要连接到 RDS 数据库实例，您必须提供数据库用户的用户名和密码，该用户必须有足够权限执行您提供的 SQL 查询。

AWS Data Pipeline 安全信息

要启用安全 AWS Data Pipeline 访问，您必须提供 IAM 资源角色和 IAM 服务角色的名称。

EC2 实例代入资源角色以将数据从 Amazon RDS 复制到 Amazon S3。创建此资源角色的最简单方法是使用 `DataPipelineDefaultResourceRole` 模板，并列出 **`machinelearning.aws.com`** 作为可信服务。有关模板的更多信息，请参阅 AWS Data Pipeline 开发人员指南中的设置 IAM 角色。

如果您要创建自己的角色，则该角色必须包含以下内容：

```
{
    "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 代入服务角色以监视将数据从 Amazon RDS 复制到 Amazon S3 的进度。创建此资源角色的最简单方法是使用 DataPipelineDefaultRole 模板，并列出 machinelearning.aws.com 作为可信服务。有关模板的更多信息，请参阅 AWS Data Pipeline 开发人员指南中的设置 IAM 角色。

Amazon RDS 安全信息

要启用安全 Amazon RDS 访问，您需要提供 VPC Subnet ID 和 RDS Security Group IDs。您还需要为 Subnet ID 参数所指向的 VPC 子网设置相应的传入规则，并提供具有此权限的安全组的 ID。

MySQL SQL 查询

MySQL SQL Query 参数指定您要在 MySQL 数据库上执行的 SQL SELECT 查询。查询的结果复制到您指定的 S3 输出位置（存储桶）。

> **Note**
>
> 当输入记录处于随机顺序时（乱序），机器学习技术最适合。您可以使用 rand() 函数将您的 MySQL SQL 查询结果轻松地乱序。例如，假设原始查询如下：
> "SELECT col1, col2, … FROM training_table"
> 您可以如下所示更新查询来添加随机乱序：
> "SELECT col1, col2, … FROM training_table ORDER BY rand()"

S3 输出位置

S3 Output Location 参数指定 MySQL SQL 查询结果所输出到的“暂存”Amazon S3 位置的名称。

> **Note**
>
> 您需要确保从 Amazon RDS 导出数据后，Amazon ML 有权从此位置读取数据。有关设置这些权限的信息，请参阅“向 Amazon ML 授予权限以从 Amazon S3 读取数据”。

训练 ML 模型

ML 模型的训练过程涉及到提供 ML 算法（即学习算法）以及用于进行学习的训练数据。术语 ML 模型是指由训练过程创建的模型项目。

训练数据必须包含正确的答案，这称为目标 或目标属性。学习算法在训练数据中查找将输入数据属性映射到目标（您希望预测的答案）的模式，然后输出捕获这些模式的 ML 模型。

您可以使用 ML 模型获取针对您不知道目标的新数据的预测。例如，假设您要训练 ML 模型来预测一封电子邮件是否为垃圾邮件。您向 Amazon ML 提供的训练数据中应包含您知道目标的电子邮件（即，有标签说明某封电子邮件是否为垃圾邮件）。Amazon ML 将使用此数据来训练 ML 模型，得到一个模型，该模型可以尝试预测新电子邮件是否为垃圾邮件。

有关 ML 模型和 ML 算法的一般信息，请参阅机器学习概念。

主题

ML 模型的类型

Amazon ML 支持三种类型的 ML 模型：二进制分类、多类别分类和回归。您应根据所要预测的目标类型来选择模型的类型。

二进制分类模型

用于二进制分类问题的 ML 模型预测二进制结果（两个可能的类别之一）。为训练二进制分类模型，Amazon ML 使用称为逻辑回归的行业标准学习算法。

二进制分类问题的示例

- “电子邮件是否为垃圾邮件？”
- “客户是否会购买此产品？”
- “此产品是一本书还是农场动物？”
- “此评论是客户还是机器人写的？”

多类别分类模型

用于多类别分类问题的 ML 模型允许您为多个类别生成预测（预测两个以上结果之一）。为训练多类别模型，Amazon ML 使用称为多项逻辑回归的行业标准学习算法。

多类别问题的示例

- “此产品是书、电影还是服装？”
- “这是浪漫喜剧、纪录片还是惊悚片？”
- “此客户最感兴趣什么类别的产品？”

回归模型

用于回归问题的 ML 模型预测数字值。为训练回归模型，Amazon ML 使用称为线性回归的行业标准学习算法。

回归问题的示例

- “西雅图明天的温度是多少？”
- “此产品将销售多少件？”
- “这套房屋将以什么价格出售？”

训练过程

要训练 ML 模型，您需要指定以下内容：

- 输入训练数据源
- 包含要预测的目标的数据属性的名称
- 必需的数据转换说明
- 控制学习算法的训练参数

在训练过程中，Amazon ML 根据您在训练数据源中指定的目标类型，自动为您选择正确的学习算法。

训练参数

通常，机器学习算法接受可用于控制训练过程的特定属性和所生成的 ML 模型的特定属性的参数。在 Amazon Machine Learning 中，这些参数被称为训练参数。您可以使用 Amazon ML 控制台、API 或命

令行界面 (CLI) 设置这些参数。如果您没有设置任何参数，Amazon ML 将使用已知非常适用于许多机器学习任务的默认值。

您可以指定以下训练参数的值：

- 最大模型大小
- 训练数据的最大扫描次数
- 类型随机排序
- 正则化类型
- 正则化数量

在 Amazon ML 控制台中，训练参数使用默认设置。默认设置足以处理大多数 ML 问题，但您可以选择其他值来优化性能。系统已根据您的数据为您配置了一些其他训练参数（例如学习速率）。

以下部分提供了有关训练参数的更多信息。

最大模型大小

最大模型大小是 Amazon ML 在 ML 模型训练期间创建的模式的总大小（以字节为单位）。

默认情况下，Amazon ML 会创建一个 100MB 模型。您可以通过指定不同的大小指示 Amazon ML 创建一个更小或更大的模型。有关可用大小范围的信息，请参阅ML 模型的类型

如果 Amazon ML 无法找到足够的模式来填充模型大小，则会创建一个较小的模型。例如，如果您指定的最大模型大小为 100MB，但 Amazon ML 找到的模式总大小只有 50MB，生成的模型将为 50MB。如果 Amazon ML 找到的模式多于将填充的指定大小，它将通过修剪对学习模型的质量影响最小的模式来强制实施最大值截断。

通过选择模型大小，您可以控制模型的预测质量与使用成本之间的平衡。较小模型可能会导致 Amazon ML 为了满足最大大小限制而删除许多模式，从而影响预测质量。另一方面，较大模型在查询实时预测时需要花费更多成本。

> **Note**
>
> 如果您使用 ML 模型生成实时预测，将可能会产生少量容量预留费用，具体费用取决于该模型的大小。有关更多信息，请参阅Amazon ML 的定价。

较大的输入数据集不一定会生成较大的模型，因为模型用于存储模式，而不是输入数据；如果模式少并且简单，生成的模型也小。如果输入数据包含大量原始属性（输入列）或派生特征（Amazon ML 数据

转换的输出），则系统在训练过程中可能会发现和存储更多模式。最好通过几个实验来选择适用于您的数据和问题的正确模型大小。Amazon ML 模型训练日志（可从控制台或通过 API 下载）包含在训练期间修剪了多少个模型（如果有）的消息，您可以通过这些消息估算潜在的预测命中质量。

数据的最大扫描次数

为了获得最佳结果，Amazon ML 可能需要多次扫描数据进行训练来发现模式。默认情况下，Amazon ML 的扫描次数为 10 次，但您可以通过设置该数字更改默认值，最多为 100 次。Amazon ML 会在扫描数据时跟踪模式（模型收敛）的质量，并在无法发现更多数据点或模式时自动停止训练。例如，如果您将扫描次数设置为 20，但 Amazon ML 发现第 15 次扫描结束后找不到任何新模式，则将在第 15 次扫描后停止训练。

一般来说，仅包含少量观察数据的数据集通常需要更多的数据扫描次数，才能获得较高的模型质量。较大数据集通常包含很多类似的数据点，因而无需扫描很多次。选择更多数据扫描次数将带来两方面的影响：模型训练时间较长，花费的成本更多。

将训练数据的类型随机排序

在 Amazon ML 中，您必须将训练数据随机排序。随机排序会打乱数据的顺序，这样 SGD 算法就不会在太多连续观察中遇到同一种类型的数据。例如，如果您要训练 ML 模型来预测产品类型，您的训练数据包含电影、玩具和视频游戏等产品类型，如果您在上传数据之前已按产品类型列对数据进行排序，该算法会按产品类型的字母顺序查看数据。该算法将先查看所有电影数据，然后您的 ML 模型开始学习电影的模式。随后在模型遇到玩具数据时，该算法的每个更新都会向玩具产品类型拟合模型，即使这些更新会让拟合电影的模式降级也是如此。这种从电影到玩具类型的突然转变，可能会生成不了解如何准确预测产品类型的模型。

您必须将训练数据随机排序，即使您在将输入数据源拆分为训练和评估部分时选择了随机拆分选项也应如此。随机拆分策略将为每个数据源选择随机数据子集，但它不会更改数据源中行的顺序。有关拆分数据的更多信息，请参阅拆分数据。

使用控制台创建 ML 模型时，Amazon ML 默认使用伪随机乱序方法来随机排列数据。无论请求扫描多少次，Amazon ML 在训练 ML 模型之前只会对数据进行一次随机排序。如果您在将数据提供给 Amazon ML 之前对数据进行了随机排序，并且不希望 Amazon ML 对您的数据重新随机排序，则可将随机类型设置为 none。例如，如果您将 .csv 文件中的记录进行随机排序后再将其上传到 Amazon S3 中，通过 Amazon RDS 创建数据源时在 MySQL SQL 查询中使用了 `rand()` 函数，或通过 Amazon Redshift 创建数据源时在 Amazon Redshift SQL 查询中使用了 `random()` 函数，将随机类型设置为 none 不会影响 ML 模型的预测准确性。只对数据进行一次随机排序可减少运行时间和创建 ML 模型的成本。

> **Important**
>
> 使用 Amazon ML API 创建 ML 模型时，Amazon ML 在默认情况下不会对您的数据随机排序。如果您使用 API 而不是控制台创建 ML 模型，我们强烈建议您通过将 `sgd.shuffleType` 参数设置为 `auto` 来对数据随机排序。

正则化类型和数量

复杂 ML 模型（包含许多输入属性）的预测性能在数据包含太多模式时会受到影响。随着模式的数量增加，模型学到并非预期数据项目的可能性也会增加，而不是学习真正的数据模式。在这种情况下，模型在训练数据上表现非常出色，但不能对新数据实现很好的泛化。这种现象被称为过度拟合 训练数据。

正则化有助于防止线性模型通过惩罚极端权重值来过度拟合训练数据示例。L1 正则化会将本来权重极小的特征的权重推向零，从而减少模型中使用的特征数。L1 正则化会生成稀疏模型并降低模型中的噪音数量。L2 正则化会生成较小的总体权重值，可在特征的相关性高的情况下稳定权重。您可以使用 `Regularization amount` 参数控制 L1 或 L2 正则化的数量。指定极大的 `Regularization amount` 值会导致所有特征的权重值为零。

选择和优化最佳正则化值是机器学习研究中的一个活跃主题。您可能会受益于选择中等数量的 L2 正则化，这是 Amazon ML 控制台的默认设置。高级用户可以在三种类型的正则化（无、L1 或 L2）和数量之间选择。有关正则化的更多信息，请转到正则化（数学运算）。

训练参数：类型和默认值

下表列出了 Amazon ML 训练参数及每个参数的默认值和允许范围。

训练参数	类型	默认值	描述
maxMLModelSizeInBytes	整数	100,000,000 字节 (100 MiB)	允许的范围：100,000 (100 KiB) 到 2,147,483,648 (2 GiB) 根据输入数据，模型大小可能会影响性能。
sgd.maxPasses	整数	10	允许的范围：1-100
sgd.shuffleType	字符串	自动	允许的值：`auto` 或 `none`

训练参数	类型	默认值	描述
sgd.l1Regularizationamount	双	0（默认不使用 L1）	允许的范围：0 到 MAX_DOUBLE 已发现介于 1E-4 到 1E-8 之间的 L1 值可以生成好结果。较大的值生成的模型可能并不是很有用。 您不能同时设置 L1 和 L2。必须从中选择一项。
sgd.l2Regularizationamount	双	1E-6（默认情况下，L2 与此正则化数量配合使用）	允许的范围：0 到 MAX_DOUBLE 已发现介于 1E-2 到 1E-6 之间的 L2 值可以生成好结果。较大的值生成的模型可能并不是很有用。 您不能同时设置 L1 和 L2。必须从中选择一项。

创建 ML 模型

创建数据源之后，您就可以创建 ML 模型。如果您使用 Amazon Machine Learning 控制台创建模型，则可选择使用默认设置或者通过应用自定义选项自定义您的模型。

自定义选项包括：

- 评估设置：您可以选择让 Amazon ML 预留部分输入数据来评估 ML 模型的预测质量。有关评估的信息，请参阅评估 ML 模型。
- 配方：配方会告诉 Amazon ML 哪些属性和属性转换可用于模型训练。有关 Amazon ML 配方的信息，请参阅使用数据配方进行功能转换。
- 训练参数：参数用于控制训练流程和生成的 ML 模型的特定属性。有关训练参数的更多信息，请参阅训练参数。

要选择这些设置或为其指定值，请在使用“创建 ML 模型”向导时选择自定义选项。如果您希望 Amazon ML 应用默认设置，请选择默认。

在您创建 ML 模型时，Amazon ML 会根据目标属性的属性类型来选择将使用的学习算法类型。（目标属性是包含“正确”答案的属性。）如果您的目标属性是“二进制”，Amazon ML 会创建一个使用逻辑回归算法的二进制分类模型。如果您的目标属性是“分类”，Amazon ML 会创建一个使用多项逻辑回归算法的多类别模型。如果您的目标属性是“数字”，Amazon ML 会创建一个使用线性回归算法的回归模型。

主题

先决条件

使用 Amazon ML 控制台创建 ML 模型之前，您需要创建两个数据源，一个用于训练模型，另一个用于评估模型。如果您尚未创建这两个数据源，请参阅教程中的步骤 2：创建训练数据源。

使用默认选项创建 ML 模型

如果您希望 Amazon ML 应用默认设置，请选择默认选项：

- 将输入数据拆分为使用第一个 70% 的数据进行训练，使用其余 30% 的数据进行评估
- 根据在训练数据源（占 70% 的输入数据源）上收集的统计信息建议配方
- 选择默认训练参数

选择默认选项

1. 在 Amazon ML 控制台中，选择 Amazon Machine Learning，然后选择机器学习模型。
2. 在ML 模型摘要页面上选择创建新 ML 模型。
3. 在输入数据页面上，确保已选择我已创建指向 S3 数据的数据源。
4. 在表中选择您的数据源，然后选择继续。
5. 在ML 模型设置页面上，为ML 模型名称键入您的 ML 模型名称。
6. 对于训练和评估设置，请确保选择默认。
7. 对于 Name this evaluation，请键入评估名称，然后选择查看。Amazon ML 会跳过向导的其余步骤，转到查看页面。
8. 检查您的数据，删除从不希望应用到模型和评估中的数据源复制的任何标签，然后选择完成。

使用自定义选项创建 ML 模型

通过自定义您的 ML 模型，您可以：

- 提供您自己的配方。有关如何提供您自己的配方的信息，请参阅配方格式参考。
- 选择训练参数。有关训练参数的更多信息，请参阅训练参数。
- 选择除 70/30 的默认比率之外的训练/评估拆分比率或提供另一个您已准备好进行评估的数据源。有关拆分策略的信息，请参阅拆分数据。

您还可以选择其中任何设置的默认值。

如果您已经使用默认选项创建了模型并且希望改进模型的预测性能，请使用自定义选项创建包含一些自定义设置的新模型。例如，您可以将更多特征转换添加到配方中或增加训练参数的通过次数。

使用自定义选项创建模型

1. 在 Amazon ML 控制台中，选择 Amazon Machine Learning，然后选择机器学习模型。
2. 在ML 模型摘要页面上选择创建新 ML 模型。
3. 如果您已经创建了数据源，请在输入数据页面上，选择我已创建指向我的 S3 数据的数据源。在表中选择您的数据源，然后选择继续。

   如果您需要创建数据源，请选择我的数据在 S3 中，并且我需要创建数据源，然后选择继续。您将重定向到创建数据源向导。指定您的数据在 S3 还是 Redshift 中，然后选择验证。完成创建数据源的过程。

   创建了数据源之后，系统会将您重定向到创建 ML 模型向导的下一个步骤。
4. 在ML 模型设置页面上，为ML 模型名称键入您的 ML 模型名称。
5. 在选择训练和评估设置中，选择自定义，然后选择继续。
6. 在配方页面上，您可以 customize a recipe。如果您不想自定义配方，Amazon ML 会为您建议一个配方。选择继续。
7. 在高级设置页面上，指定最大 ML 模型大小、传递的最大数据量、将训练数据的类型随机排序、正则化类型和正则化数量。如果您未指定这些参数，Amazon ML 会使用默认的训练参数。

   有关这些参数及其默认值的更多信息，请参阅训练参数。

   选择继续。
8. 在评估页面上，指定是否要立即评估 ML 模型。如果您不想立即评估 ML 模型，请选择审核。

如果您希望立即评估 ML 模型：

a. 对于为此评估命名，键入评估的名称。

b. 对于选择评估数据，选择您是否希望 Amazon ML 预留一部分输入数据进行评估，如果是，选择您希望如何拆分数据源，如果不是，请提供其他数据源进行评估。

c. 选择审核。

9. 在审核页面上，编辑您的选择，删除从不希望应用到模型和评估中的数据源复制的任何标签，然后选择完成。

创建了模型之后，请参阅步骤 4：查看 ML 模型的预测性能和设置分数阈值。

用于机器学习的数据转换

机器学习模型的质量好坏与训练所用的数据相关。好的训练数据的一个关键特性是，它是以一种针对学习和归纳进行优化的方式提供。在业内，这种将数据一起置入此优化格式的过程称为特征转换。

主题

特征转换的重要性

请考虑任务是确定信用卡交易是否是欺诈行为的机器学习模型。根据您的应用背景知识和数据分析结果，您可能需要确定哪些数据字段 (或特征) 务必要包含在输入数据中。例如，交易金额、商户名称、地址和信用卡所有者的地址都必须提供给模型学习过程。另一方面，随机生成的交易 ID 不含信息 (如果我们知道该 ID 确实是随机的)，因此没有用处。

确定要包含哪些字段之后，您可以转换这些特征，以帮助学习过程。通过转换为输入数据添加背景经验，可让机器学习模型从中受益。例如，以下商户地址将用字符串表示：

"123 Main Street, Seattle, WA 98101"

地址本身的表达能力有限，它只对与确切地址相关的学习模式有用。但是，将其分成各组成部分可创建其他特征，例如“地址”(123 Main Street)、“城市”(Seattle)、“州”(WA) 和“邮政编码”(98101)。现在，学习算法可以将多个离散的交易分组在一起并发现更广泛的模式，可能某些商户邮政编码遭遇欺诈行为的几率比其他商户更多。

有关特征转换方法和过程的更多信息，请参阅机器学习概念。

使用数据配方进行特征转换

在使用 Amazon ML 创建 ML 模型之前，有两种方法可用于转换特征：您可在将输入数据提供给 Amazon ML 之前直接进行转换，或者使用 Amazon ML 的内置转换。您可以使用 Amazon ML 配方，这是针对常见转换的预先格式化指令。利用配方，您可以执行以下操作：

- 从内置常用机器学习转换列表中选择，并将这些转换应用到单独的变量或一组变量
- 选择使哪些变量和转换可供机器学习过程使用

使用 Amazon ML 配方提供了多种优势。Amazon ML 为您执行数据转换，因此您不需要自行实施。此外，这些转换非常快，因为 Amazon ML 在读取输入数据时应用转换并将结果提供给学习过程，没有将结果保存到磁盘的中间步骤。

配方格式参考

Amazon ML 配方包含将您的数据转换为机器学习过程的一部分的指令。配方使用类似 JSON 的语法定义，但它们包含除常规 JSON 限制之外的其他限制。配方包含以下部分，这些部分必须按如下所示的顺序显示：

- 组 可对多个变量进行分组，以便应用转换。例如，您可以为与网页（标题、正文）上自由文本部分相关的所有变量创建分组，然后立即对所有这些部分执行转换。
- 分配 可以创建在处理过程中可重复使用的中间命名变量。
- 输出 可定义学习过程中将使用哪些变量以及对这些变量应用哪些转换（如果有）。

组

您可以定义变量分组以集中转换这些分组中的所有变量，或者在不转换这些变量的情况下将其用于机器学习。默认情况下，Amazon ML 会为您创建以下分组：

ALL_TEXT、ALL_NUMERIC、ALL_CATEGORICAL、ALL_BINARY - 特定于类型的分组，基于在数据源架构中定义的变量。

> **Note**
>
> 您不能创建包含 `ALL_INPUTS` 的分组。

这些变量无需定义即可在配方的输出部分使用。您也可以通过在现有分组中增加或减少变量来创建自定义组，或直接从变量集创建。在以下示例中，我们将演示所有三种方法以及用于分配分组的语法：

```
"groups": {

"Custom_Group": "group(var1, var2)",
```

```
"All_Categorical_plus_one_other": "group(ALL_CATEGORICAL, var2)"

}
```

组名必须以字母字符开头，长度可在 1 至 64 个字符之间。如果组名不以字母字符开头或者包含特殊字符 (, ' " \t \r \n ( ) \)，则需用引号将名称括起来才能包含在配方中。

分配

为了便于使用和读取，您可以将一个或多个转换分配给中间变量。例如，如果您有名为 email_subject 的文本变量，并且为其应用了小写转换，则可将生成的变量命名为 email_subject_lowercase，这样便于您在配方中的其他位置跟踪该变量。分配也可以链接在一起，便于您按指定的顺序应用多个转换。以下示例显示了配方语法中的单个分配和链式分配：

```
"assignments": {

"email_subject_lowercase": "lowercase(email_subject)",

"email_subject_lowercase_ngram":"ngram(lowercase(email_subject), 2)"

}
```

中间变量名称必须以字母字符开头，长度可在 1 至 64 个字符之间。如果该名称不以字母字符开头或者包含特殊字符 (, ' " \t \r \n ( ) \)，则需用引号将名称括起来才能包含在配方中。

输出

输出部分可控制在学习过程中使用哪些输入变量以及为其应用哪些转换。输出部分为空或不存在即出现错误，因为未将数据传递到学习过程。

最简单的输出部分只包含预定义的 ALL_INPUTS 分组，用于指示 Amazon ML 使用在用于学习的数据源中定义的所有变量：

```
"outputs": [

"ALL_INPUTS"

]
```

输出部分还可以通过指示 Amazon ML 使用其他预定义分组中的所有变量来引用这些分组：

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

输出部分还可以引用自定义组。在以下示例中，只有在上一示例中的分组分配部分中定义的一个自定义组将用于机器学习。所有其他变量都将删除：

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

输出部分还可以引用分配部分的定义变量分配：

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

输入变量或转换可直接在输出部分中定义：

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

输出需要明确指定预计将用于学习过程的所有变量和转换变量。例如，假如您在输出中包含 var1 和 var2 的笛卡尔积。如果您希望也包含原始变量 var1 和 var2，则需在输出部分添加原始变量：

```
"outputs": [

"cartesian(var1,var2)",

"var1",

"var2"

]
```

输出时可以通过添加与变量相关的注释文本来包含注释，以便阅读：

```
"outputs": [

"quantile_bin(age, 10) //quantile bin age",

"age // explicitly include the original numeric variable along with the
binned version"

]
```

您可以在输出部分中混合搭配所有这些方法。

> **Note**
>
> 添加配方时不允许在 Amazon ML 控制台中添加注释。

完整配方示例

以下示例引用了前面示例中介绍的几个内置数据处理器：

```
{

"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"

]

}
```

建议配方

如果您在 Amazon ML 中创建新的数据源，并为该数据源计算了统计信息，则 Amazon ML 还将创建建议配方，此配方可用于从数据源创建新的 ML 模型。建议的数据源基于数据和数据中存在的目标属性，并为创建和微调 ML 模型提供有用的起点。

要在 Amazon ML 控制台中使用建议的配方，请从新建下拉列表中选择数据源或数据源和 ML 模型。对于 ML 模型设置，您可在创建机器学习模型向导的机器学习模型设置步骤中选择训练和评估设置“默认”或“自定义”。如果您选择“Default”选项，Amazon ML 将自动使用建议的配方。如果您选择“Custom”选项，下一步中的配方编辑器将显示建议的配方，您可以进行验证，或者根据需要进行修改。

> **Note**
>
> Amazon ML 允许您创建数据源，然后在统计数据计算完成之前，立即使用该数据源来创建 ML 模型。在这种情况下，您无法查看“Custom”选项中的建议配方，不过您仍可以继续完成该步骤，让 Amazon ML 将默认配方用于模型训练。

要在 Amazon ML API 上使用建议的配方，您可以在“Recipe”和“RecipeUri API”参数中传递空字符串。无法使用 Amazon ML API 检索建议的配方。

数据转换参考

主题

N 元转换

N 元转换获取文本变量作为输入，并生成与滑动 (用户可配置) n 个单词的窗口对应的字符串，在这一过程中生成输出。例如，请考虑文本字符串“I really enjoyed reading this book”。

指定窗口大小为 1 的 N 元转换仅向您提供该字符串中的所有单个单词：

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

指定窗口大小为 2 的 N 元转换将返回所有 2 个单词组合以及所有单个单词组合。

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

指定窗口大小为 3 的 N 元转换将添加 3 个单词组合到此列表中，得到以下内容：

```
{"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"}
```

您可以请求大小范围从 2 到 10 个单词的 N 元。对于数据架构中其类型标记为文本的所有输入，隐式生成大小为 1 的 N 元，因此您无需要求它们。最后请记住，N 元是通过按照空格字符拆分输入数据来生成的。这意味着，例如，标点符号将视为单词令牌的一部分：为字符串“red, green, blue”使用窗口 2 生成 N 元将得到 {“red,” , “green,” , “blue,” , “red, green” , “green, blue”}。如果您不希望有标点，可以使用标点删除程序处理器 (本文档的稍后部分中介绍) 来删除标点符号。

为变量 var1 计算窗口大小 3 的 N 元：

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

正交稀疏二元 (OSB) 转换

OSB 转换用于帮助进行文本字符串分析，是 2 元转换的替代 (窗口大小为 2 的 N 元)。OSB 通过以下方式生成：滑动文本上方的大小为 n 个词的窗口并输出包含窗口中的第一个词的每个单词对。

为了生成各个 OSB，其组成单词使用“_”(下划线) 字符联接，每个跳过的令牌由添加到 OSB 中的另一个下划线来指示。因此，OSB 不仅对在窗口中发现的令牌进行编码，而且还对在相同窗口中跳过的令牌数的指示进行编码。

为了说明，请考虑字符串“The quick brown fox jumps over the lazy dog”以及 OSB 大小为 4。以下示例中显示了 6 个窗口，窗口的大小为 4 个单词，最后较短的两个窗口来自字符串结尾，并显示了为每个窗口生成的 OSB：

窗口，{生成的 OSB}

```
"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}
```

正交稀疏二元是 N 元的替代方法，在某些情况下可能会合适。如果您的数据具有大型文本字段 (10 个或更多个单词)，可以通过试验来了解哪种方式更好。请注意，什么构成大型文本字段可能会取决于具体情况。但是，如果文本字段较大，则根据经验表明，由于特殊的跳过 字符（下划线），OSB 会以独特的方式表示文本。

您可以在输入文本变量上为 OSB 转换请求大小为 2 到 10 的窗口。

为变量 var1 计算窗口大小 5 的 OSB：

"osb(var1, 5)"

小写转换

小写转换处理器将文本输入转换为小写。例如，假设输入为“The Quick Brown Fox Jumps Over the Lazy Dog”，处理器将输出“the quick brown fox jumps over the lazy dog”。

为变量 var1 应用小写转换：

"lowercase(var1)"

删除标点转换

Amazon ML 隐式拆分数据架构中标记为文本的输入，以空格为依据。字符串中的标点要么与邻近的单词令牌放在一起，要么完整作为单独的令牌，具体取决于周围的空格。如果不希望有标点，可以使用标

点删除程序转换从生成的特征中删除标点符号。例如，假设字符串为“Welcome to AML - please fasten your seat-belts!”，将隐式生成以下一组令牌：

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

对此字符串应用标点删除程序处理器会得到以下一组结果：

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

请注意，只删除前缀和后缀标点符号。不删除显示在令牌中间的标点，例如“seat-belts”中的连字符。

为变量 var1 应用标点删除：

"no_punct(var1)"

分位数分箱转换

分位数分箱处理器采用两种输入：一个数值变量和一个称作分箱数 的参数，并输出一个分类变量。其目的是通过将观察值分组在一起，发现变量分布中的非线性情况。

在许多情况下，数值变量与目标之间的关系并非线性 (数值变量值不随目标单增或单减)。在这种情况下，将数值特征分箱到表示数值特征不同范围的分类特征中可能会很有用。然后，每个分类特征值 (分箱) 可以建模为具有与目标的自身线性关系。例如，假设您知道连续数值特征 account_age 与购买某书的可能性并非线性相关。您可以将寿命分箱到可以更准确捕获与目标关系的分类特征。

分位数分箱处理器可用于指示 Amazon ML 根据所有输入值的 age 变量分布，建立相同大小的 n 个分箱，然后使用包含分箱的文本令牌替换每个数字。数值变量的最佳分箱数量取决于变量的特性及其与目标的关系，最好通过试验来确定。Amazon ML 建议，基于建议配方中的数据统计信息确定数值特征的最佳分箱数量。

您可以请求为任意数值输入变量计算 5 到 1000 之间的分位数分箱。

以下示例显示了如何在数值变量 var1 中计算和使用 50 个分箱：

"quantile_bin(var1, 50)"

标准化转换

标准化转换器将数值变量标准化为 0 以及与 1 的差的平均值。如果数值变量有非常大范围的差别，而最高量级的变量会主导 ML 模型且不论特征是否提供目标的信息，那么数值变量的标准化可以帮助学习过程。

要将此转换应用到变量 var1，请将此项添加到配方：

normalize(var1)

此转换器还可以获取用户定义的数值变量组或所有数值变量 (ALL_NUMERIC) 的预定义组作为输入：

normalize(ALL_NUMERIC)

注意

系统并不 强制要求为数值变量使用标准化处理器。

笛卡尔积转换

笛卡尔积转换生成多个文本或分类输入变量的排列。此转换在对变量之间的交互存在疑问时使用。例如，考虑教程中的银行营销数据集：使用 Amazon ML 预测对营销方案的响应。使用此数据集，我们希望根据经济和人口统计信息，预测某人是否会积极回应银行推广活动。我们可能会怀疑个人的工作类型某种程度上很重要 (就职于特定领域与获得可支配收入方面可能存在关联)，而获得的最高教育水平也会很重要。我们还可能有更深的直觉，这两个变量之间的交互存在很强的信号，例如，推广活动可能尤其适合获得了大学学位并且是企业家的客户。

笛卡尔积转换获取分类变量或文本作为输入，生成捕获了这些输入变量之间交互的新特征。具体而言，对于每个训练示例，它将创建特征组合，然后将其作为独立特色添加。例如，假设我们简化输入行，如下所示：

target, education, job

0, university.degree, technician

0, high.school, services

1, university.degree, admin

如果我们指定笛卡尔转换要应用到分类变量 education 和 job 字段，生成的特征 education_job_interaction 将如下所示：

target, education_job_interaction

0, university.degree_technician

0, high.school_services

1, university.degree_admin

笛卡尔转换在处理令牌的序列时功能更为强大，适合在其参数之一是隐式或显式拆分为多个令牌的文本变量。例如，考虑是否将某本书分类为教科书的情况。直观上，我们会认为书名中会有某些信息告诉我们这是否为教科书 (特定单词可能会更频繁出现在教科书名中)，我们还会认为书本的装订会提供一些预测性的信息 (教科书更可能使用硬书皮)，但实际情况是，将书名中的一些单词与装订组合起来具有最好的预测性。对于真实应用示例，下表显示了对输入变量 binding 和 title 应用笛卡尔处理器的结果：

Textb	标题	Binding	no_punct(Title) 和 Binding 的笛卡尔积
1	Economics : Principles, Problems, Policies	Hardcover	{"Economics_Hardcover", "Principles_Hardcover", "Problems_Hardcover", "Policies_Hardcover"}
0	The Invisible Heart: An Economics Romance	Softcover	{"The_Softcover", "Invisible_Softcover", "Heart_Softcover", "An_Softcover", "Economics_Softcover", "Romance_Softcover"}
0	Fun With Problems	Softcover	{"Fun_Softcover", "With_Softcover", "Problems_Softcover"}

以下示例显示如何将笛卡尔转换器应用到 var1 和 var2：

cartesian(var1, var2)

数据重新排列

数据重新排列功能使您可以仅基于所指向的一部分输入数据创建数据源。例如，当您在 Amazon ML 控制台中使用创建机器学习模型向导创建 ML 模型并选择了默认评估选项时，Amazon ML 自动保留 30% 的数据用于 ML 模型评估，使用另外的 70% 进行训练。此功能由 Amazon ML 的数据排列功能启用。

如果您在使用 Amazon ML API 创建数据源，您可以指定新数据源将基于哪一部分的输入数据。要执行此操作，您可将 `DataRearrangement` 参数中的指令传递到 `CreateDataSourceFromS3`、`CreateDataSourceFromRedshift` 或 `CreateDataSourceFromRDS` API。DataRearrangement 字符串的内容是 JSON 字符串，包含您数据的开头和结尾位置，以百分比、一个补充标记和一个拆分策略表示。例如，以下 DataRearrangement 字符串指定将使用数据的前 70% 创建数据源：

```
{
    "splitting": {
```

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

DataRearrangement 参数

要更改 Amazon ML 创建数据源的方式，请使用以下参数。

PercentBegin（可选）

使用 percentBegin 指示数据源的数据开始位置。如果您未包括 percentBegin 和 percentEnd，Amazon ML 将在创建数据源时包括所有数据。

有效值为 0 到 100（含）。

PercentEnd（可选）

使用 percentEnd 指示数据源的数据结束位置。如果您未包括 percentBegin 和 percentEnd，Amazon ML 将在创建数据源时包括所有数据。

有效值为 0 到 100（含）。

Complement（可选）

complement 参数告知 Amazon ML 使用未包括在 percentBegin 到 percentEnd 范围中的数据来创建数据源。如果您需要为训练和评估创建补充数据源，complement 参数非常有用。要创建补充数据源，请为 percentBegin 和 percentEnd 使用相同值，并包括 complement 参数。

例如，以下两个数据源不共享任何数据，并可用于训练和评估模型。第一个数据源具有 25% 的数据，第二个具有 75% 的数据。

用于评估的数据源：

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

用于训练的数据源：

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

有效值为 `true` 和 `false`。

Strategy（可选）

要更改 Amazon ML 如何为数据源拆分数据，请使用 `strategy` 参数。

`strategy` 参数的默认值为 `sequential`，这意味着 Amazon ML 按照记录在输入数据中的显示顺序，获取 `percentBegin` 和 `percentEnd` 参数之间的所有数据记录用于数据源。

以下两行 `DataRearrangement` 是按顺序排序的训练和评估数据源示例：

用于评估的数据源：`{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential"}}`

用于训练的数据源：`{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential", "complement":"true"}}`

要从随机选择的数据创建数据源，请将 `strategy` 参数设置为 `random` 并提供用作随机数据拆分种子值的字符串（例如，您可以使用数据的 S3 路径作为随机种子字符串）。如果您选择随机拆分策略，Amazon ML 会向每个数据行分配一个伪随机编号，然后选择分配编号在 `percentBegin` 和 `percentEnd` 之间的行。伪随机编号使用字节偏移值作为种子进行分配，因此更改数据会导致不同的拆分。保留所有现有排序。随机拆分策略可以确保训练和评估中的变量具有类似分布。输入数据可能会有隐式排序顺序时，会导致训练和评估数据源包含不相似的数据记录，这种情况下该策略会非常有用。

以下两行 `DataRearrangement` 是按非顺序排序的训练和评估数据源示例：

用于评估的数据源：

```
{
    "splitting":{
        "percentBegin":70,
```

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

用于训练的数据源：

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

有效值为 `sequential` 和 `random`。

（可选）Strategy:RandomSeed

Amazon ML 使用 randomSeed 拆分数据。API 的默认种子是空字符串。要为随机拆分策略指定种子，请传入字符串。有关随机种子的更多信息，请参阅 Amazon Machine Learning 开发人员指南中的 随机拆分数据。

有关演示如何使用 Amazon ML 进行交叉验证的示例代码，请转到 Github 机器学习示例。

评估 ML 模型

您应该始终评估模型，以确定该模型能够针对新数据和未来数据出色地预测目标。由于未来实例具有未知的目标值，您需要针对已知目标答案的数据检查 ML 模型的准确性指标，并将这项评估作为未来数据的预测准确性的代理。

要正确评估模型，您需要留存来自训练数据源并已用目标（基本实际情况）进行标记的数据示例。使用已用于训练的相同数据评估 ML 模型的预测准确性并没有用，因为它会奖励可以“记住”训练数据的模型，而不是通过它进行归纳。完成 ML 模型的训练之后，即可向模型发送已知目标值的留存观察。然后，您可以将 ML 模型返回的预测结果与已知目标值进行比较。最后，计算汇总指标，您可以通过该指标了解预测值和实际值的一致程度。

在 Amazon ML 中，您可以通过创建评估来评估 ML 模型。要为 ML 模型创建评估，您需要一个待评估的 ML 模型，还需要未用于训练的已标记数据。首先，创建一个用于评估的数据源，具体方法是使用留存数据创建 Amazon ML 数据源。评估所用数据必须与训练所用数据具有相同的架构，并且包含目标变量的实际值。

如果您的所有数据都位于单个文件或目录中，您可以使用 Amazon ML 控制台来拆分数据。“Create ML Model”向导的默认路径会拆分输入数据源，并将前 70% 的数据用于训练数据源，将剩余 30% 的数据用于评估数据源。您也可以使用“创建 ML 模型”向导中的自定义选项来自定义拆分比率，您可以在向导中选择 70% 的随机样本用于训练，剩余 30% 的数据用于评估。要进一步指定自定义拆分比率，请使用创建数据源 API 中的数据重新排列字符串。有了评估数据源和 ML 模型之后，就可以创建评估并审查评估的结果。

主题

ML 模型洞察

评估 ML 模型时，Amazon ML 提供了行业标准的指标以及一系列洞察，可用于检查模型的预测准确度。在 Amazon ML 中，评估结果包含以下内容：

- 预测准确度指标，报告模型的整体成功情况
- 可视化，用于帮助在预测准确度指标之外，探索模型的准确度
- 检查设置分数阈值影响的能力 (仅适用于二进制分类)
- 针对检查评估有效性标准的提醒

指标和可视化的选择取决于所评估的 ML 模型的类型。务必查看这些可视化内容以确定您的模型表现是否足够好，是否符合您的业务需求。

二进制模型洞察

解释预测

许多二进制分类算法的实际输出是预测分数。该分数指示给定观察属于正类的系统确定性（实际目标值为 1）。Amazon ML 中的二进制分类模型输出一个介于 0 和 1 之间的分数。作为此分数的使用者，为了决定观察应分类为 1 还是 0，需要选取分类阈值或者选取截断值，并与分数进行对比，以此来解释分数。当目标等于 1 时，将预测其分数高于截断值的任何观察；当目标等于 0 时，将预测其分数低于截断值的观察。

在 Amazon ML 中，默认的分数截断值为 0.5。您可以选择更新此截断值以满足您的业务需求。您可以在控制台中使用可视化内容来了解截断值的选择将对您的应用程序造成怎样的影响。

衡量 ML 模型准确度

Amazon ML 为二进制分类模型提供行业标准的准确度指标，称为（受试者操作特征）曲线下面积 (AUC)。AUC 衡量模型为正面示例预测出相比负面示例更高分数的能力。由于它独立于分数截断值，因此您可以从 AUC 指标感受到模型的预测准确度，无需选取阈值。

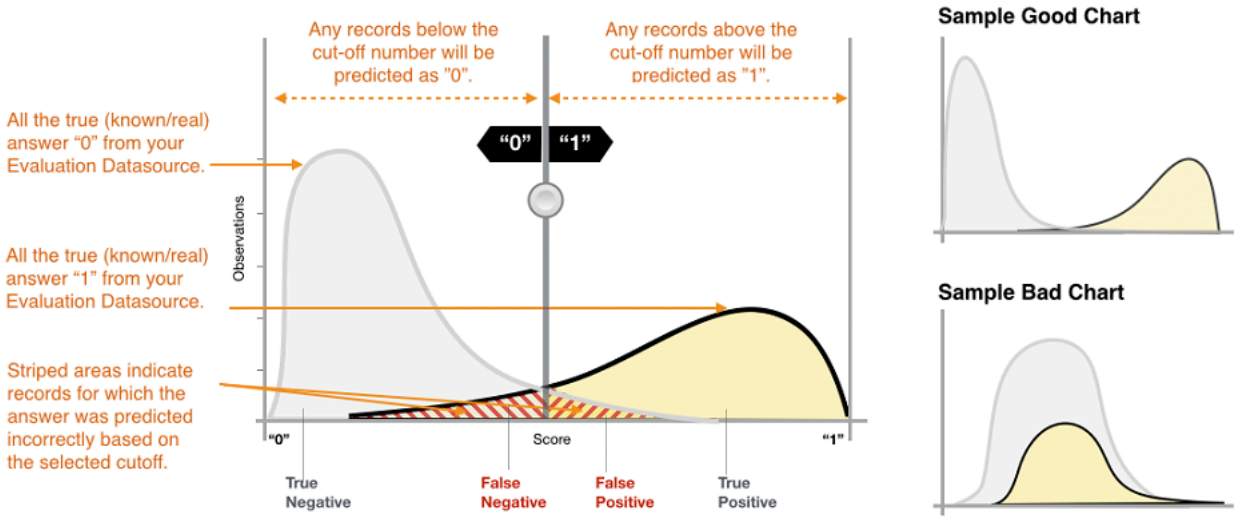
AUC 指标返回从 0 到 1 的数值。接近 1 的 AUC 值指示高度准确的 ML 模型。接近 0.5 的值指示 ML 模型比随便猜测好不了多少。值接近 0 的情况很少见，这通常表示数据有问题。基本上，接近 0 的 AUC 表示 ML 模型已学习了正确的模式，但使用它们来预测会得到与实际颠倒的结果（将“0”预测为“1”或者将“1”预测为“0”）。有关 AUC 的更多信息，请转到 Wikipedia 上的受试者操作特征页面。

二进制模型的基线 AUC 指标为 0.5。这是随机预测 1 或 0 答案的假想 ML 模型的值。您的二进制 ML 模型要想有价值，表现得应该比此值要好才行。

使用性能可视化

要探究 ML 模型的准确度，您可以查看 Amazon ML 控制台的评估页面上的图表。此页显示两个直方图：a) 实际正例分数的直方图（目标为 1）以及 b) 实际负例分数的直方图（目标为 0）。

具有良好预测准确度的 ML 模型对于实际的 1 将预测较高的分数，对实际的 0 将预测较低的分数。完美的模型将在 x 轴两端具有两个直方图，实际正例全部得到高分，实际负例全部得到低分。但是，ML 模型会出错，一般的图表会显示两个直方图在某些分数重叠。性能极差的模型会无法区分正类和负类，这两个类的直方图重叠最多。



利用可视化内容可以确定属于两种正确预测类型的预测数以及属于两种不正确预测类型的预测数。

正确预测

- 真正 (TP)：Amazon ML 预测的值为 1，真正的值为 1。
- 真负 (TN)：Amazon ML 预测的值为 0，真正的值为 0。

错误预测

- 假正 (FP)：Amazon ML 预测的值为 1，但真正的值为 0。
- 假负 (FN)：Amazon ML 预测的值为 0，但真正的值为 1。

> **Note**
>
> TP、TN、FP 和 FN 的数量取决于所选分数阈值，对这些类型中的任意一种进行优化都意味着需要对其他类型作出让步。TP 数高通常会导致 FP 数高和 TN 数低。

调整分数截断值

ML 模型的工作方式是生成数字预测分数，然后应用截断值将这些分数转换为二进制 0/1 标签。通过更改分数截断值，您可以在模型出错时调整其行为。在 Amazon ML 控制台的评估页面上，您可以查看不同分数截断值的影响，并可以保存希望用于模型的分数截断值。

在您调整分数截断值阈值时，请观察两种错误类型之间的平衡。将截断值向左侧移动将会获得更多“真正”数，但同时“假正”数也会增加。将其向右侧移动将获得较少的“假正”错误数，但同时也会失去一些“真正”数。对于您的预测应用程序，您需要通过选择合适的截断值分数来确定更能容忍哪种类型的错误。

查看高级指标

Amazon ML 提供以下额外指标来测量 ML 模型的预测准确度：准确度、精度、召回率和假正率。

准确度

准确度 (ACC) 衡量正确预测的比率。范围为 0 至 1。值越大说明预测准确度越高：

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

精度

精度衡量实际正例与预测为正例的比率。范围为 0 至 1。值越大说明预测准确度越高：

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

召回率

召回率衡量预测为正例的实际正例的比率。范围为 0 至 1。值越大说明预测准确度越高：

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

假正率

假正率 (FPR) 衡量错误警报率或预测为正例的实际负例的比率。范围为 0 至 1。值越小说明预测准确度越高：

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

根据您的业务问题，您可能会对在这些指标的特定部分中表现良好的模型更感兴趣。例如，两个业务应用程序可能对其 ML 模型有着迥然不同的要求：

- 一个应用程序可能需要严格保证正面预测实际为正面（高精度），并能够承受将一些正面示例错误分类为负面示例（中等召回率）。
- 另一个应用程序可能需要尽可能多地正确预测正面示例（高召回率），并可以接受将一些负面示例错误分类为正面示例（中等精度）。

Amazon ML 允许您选择与之前任意高级指标的特定值相对应的分数截断值。它还显示由于优化任意一项指标而导致作出的让步。例如，如果您选择一个与高精度对应的截断值，则通常不得不接受较低的召回率。

> **Note**
>
> 您必须为其保存分数截断值，这样才能用于分类 ML 模型未来的任何预测。

多类别模型洞察

解释预测

多类别分类算法的实际输出是一组预测分数。这些分数指示模型的给定观察属于各个类别的确定性。与二进制分类问题不同，您不需要选择分数截断值以进行预测。预测的答案是预测分数最高的类别（例如，label）。

衡量 ML 模型准确度

多类别中使用的典型指标与对所有类别进行平均化之后二进制分类案例中使用的指标相同。在 Amazon ML 中，宏平均 F1 分数用于评估多类别指标的预测准确性。

宏平均 F1 分数

F1 分数是考虑二进制指标精度和重新调用的二进制分类指标。它是精度和重新调用之间的调和平均数。范围为 0 至 1。值越大说明预测准确度越高：

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

宏平均 F1 分数是多类别案例中所有类别的 F1 分数的未加权平均数。它不会考虑类别在评估数据集中出现的频率。较大的值表示更好的预测准确度。以下示例显示了评估数据源中的 K 类别：

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

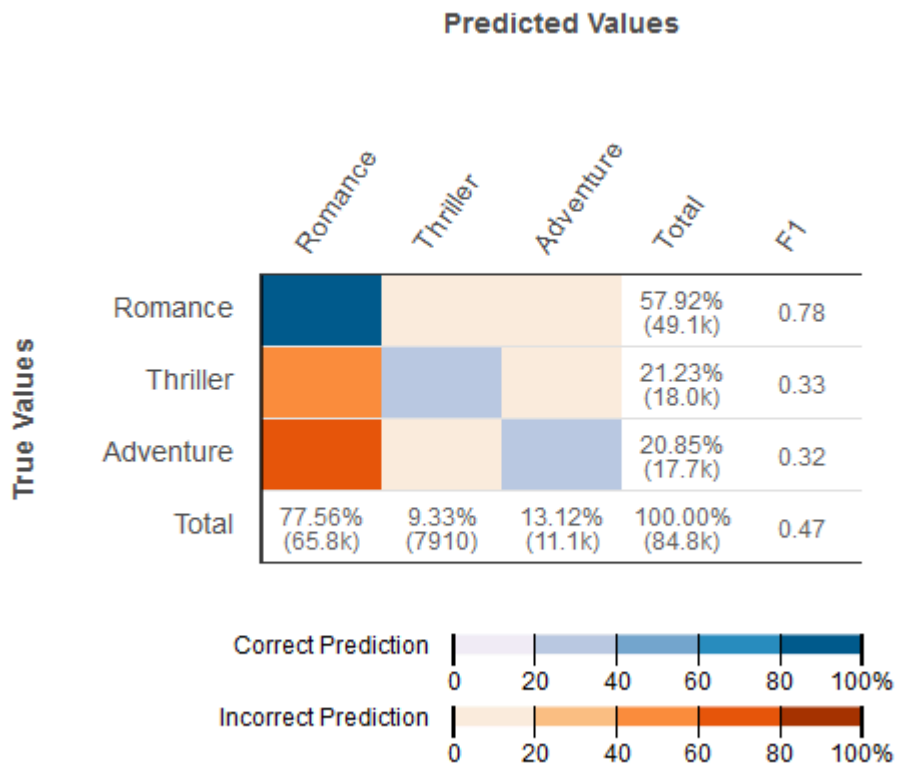
基准宏平均 F1 分数

Amazon ML 提供了多类别模型的基准指标。它是假想多类别模型的宏平均 F1 分数，该模型始终将最常见的类别预测为答案。例如，如果您在预测电影流派时，您训练数据中最常见的流派是爱情片，则基准模型会始终将流派预测为爱情片。您可以根据此基准来比较您的 ML 模型，以验证您的 ML 模型是否优于预测此常量答案的 ML 模型。

使用性能可视化

Amazon ML 提供混淆矩阵来直观地呈现多类别分类预测模型的准确性。混淆矩阵通过比较观察的预测类别及其真正的类别，在表中列出了各类别正确和错误预测的数量或百分比。

例如，如果您尝试将电影归入某个流派，预测模型可能预测其流派（类别）是爱情片。但其真正的流派可能是惊悚片。当您评估多类别分类 ML 模型的准确度时，Amazon ML 会识别错误分类并在混淆矩阵中显示结果，如下图所示。



系统会在混淆矩阵中显示以下信息：

- 每个类别的正确和错误预测的数量：混淆矩阵中的每一行都对应一个真正类别的指标。例如，第一行显示了实际流派为爱情片的电影，多类别 ML 模型正确预测了 80% 以上的案例。它将不到 20% 的案例的流派错误地预测为惊悚片，并将不到 20% 的案例的流派错误地预测为冒险片。
- 类域 F1 分数：最后一个列显示每个类别的 F1 分数。
- 评估数据中的真正类别频率：第二列至最后一列显示，在评估数据集中，评估数据中有 57.92% 的观察为爱情片，21.23% 为惊悚片，20.85% 为冒险片。
- 评估数据的预测类别频率：最后一行显示预测中每个类的频率。77.56% 的观察被预测为爱情片，9.33% 被预测为惊悚片，13.12% 被预测为冒险片。

Amazon ML 控制台提供了可见显示，最多可在混淆矩阵中包含 10 个类别，这些类别将按照评估数据中最常见类别到最不常见类别的顺序列出。如果您的评估数据包含 10 个以上的类别，将会在混淆矩阵中看到前 9 个最常见的类别，所有其他类别都折叠为名为“others”的类别。借助 Amazon ML，您还能通过多类别可视化页面上的链接下载完整混淆矩阵。

回归模型洞察

解释预测

回归 ML 模型的输出是数值，是模型对目标的预测。例如，如果您要预测房价，模型的预测可能为 254013 这样的值。

> **Note**
>
> 预测的范围可能与训练数据中目标的范围不同。例如，假设您预测房价，训练数据中目标的值范围是 0 到 450000。预测目标无需位于同样的范围，并可以为任意正值 (大于 450000) 或负值 (小于零)。请务必计划如何解决预测值在您的应用程序可接受范围之外的情况。

衡量 ML 模型准确度

对于回归任务，Amazon ML 使用行业标准的均方根误差 (RMSE) 指标。该指标衡量预测数值目标与实际数值答案 (基本实际情况) 之间的差距。RMSE 的值越小，模型的预测精度就越高。预测完全正确的模型的 RMSE 为 0。以下示例显示包含 N 条记录的评估数据：

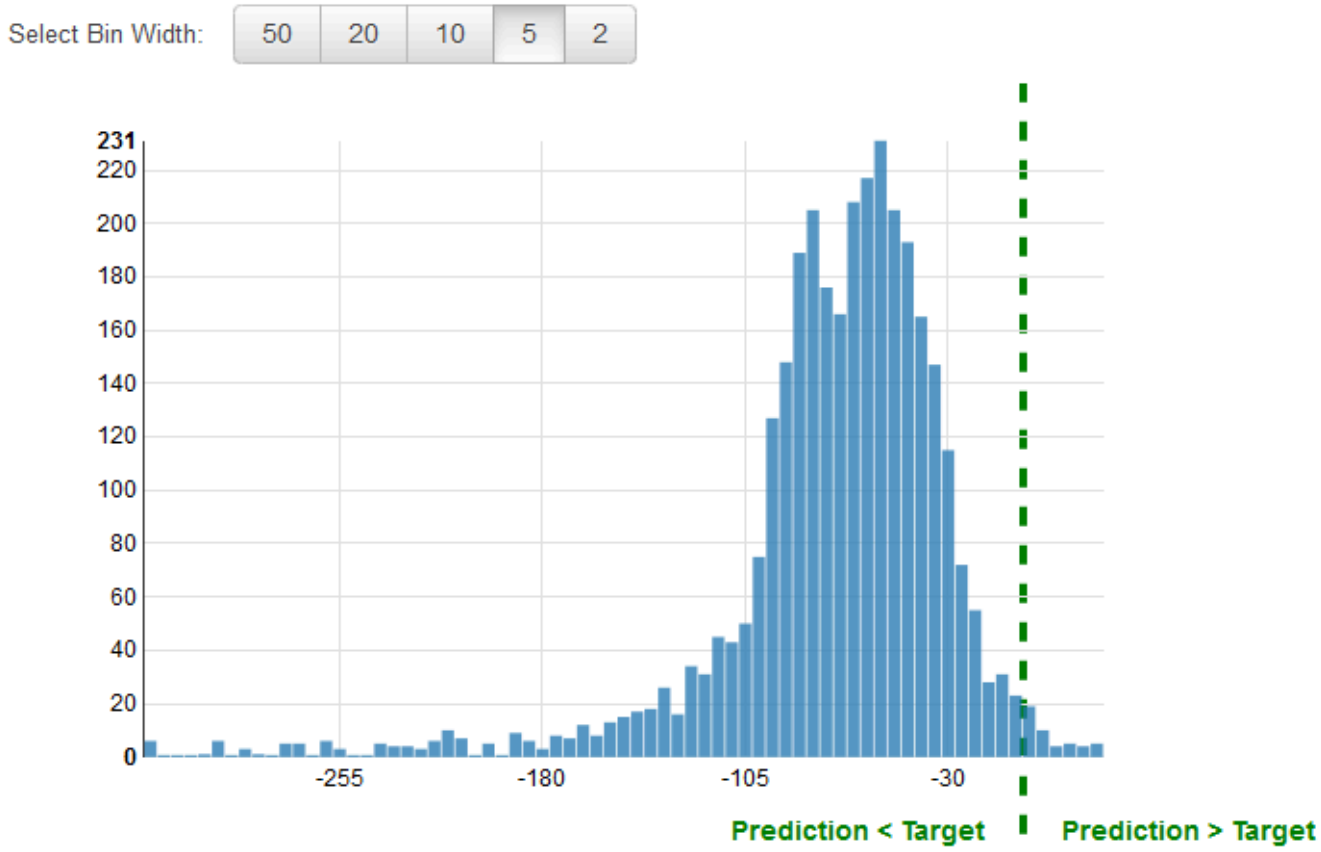
$$RMSE = \sqrt{1/N \sum_{i=1}^{N} (actual\ target - predicted\ target)^2}$$

基准 RMSE

Amazon ML 提供了回归模型的基准指标。这是假设回归模型的 RMSE，该模型始终预测目标的平均值作为答案。例如，如果您预测房产买家的年龄，并且训练数据中所有观察的平均年龄为 35 岁，则基准模型始终将答案预测为 35 岁。您可以根据此基准来比较您的 ML 模型，以验证您的 ML 模型是否优于预测此常量答案的 ML 模型。

使用性能可视化

对于回归问题，常见的做法是检查残差。评估数据中某个观察的残差是真实目标与预测目标之间的差值。残差表示模型无法预测的目标部分。正残差表示模型低估了目标 (实际目标大于预测目标)。负残差表示高估 (实际目标小于预测目标)。评估数据残差的直方图在呈钟形分布并且中心在零上时，指示模型以随机方式产生错误，不会系统性地高于或低于预测目标值的任何特定范围。如果残差未构成以零为中心的钟形曲线，这种情况表示模型的预测中存在结构错误。向模型添加更多变量可能会帮助模型捕获当前模型未捕获的模式。下图显示了不以零为中心的残差。



防止过度拟合

创建和训练 ML 模型时，目的是选择能够提供最佳预测的模型，这意味着要选择具有最佳设置（ML 模型设置或超级参数）的模型。在 Amazon Machine Learning 中，您可以设置四个超级参数：扫描次数、正则化、模型大小和随机类型。不过，如果您选择可为评估数据生成“最佳”预测性能的模型参数设置，您可能会过度拟合模型。当模型记住了在训练数据源和评估数据源中出现的模式，但未能在数据中归纳这些模式时，会出现过度拟合。这种情况通常在训练数据包含所有用于评估的数据时出现。过度拟合的模型在评估阶段表现很好，但无法对未见过的数据提供准确的预测。

为避免选择过度拟合模型作为最佳模型，您可以预留更多数据，以验证 ML 模型的性能。例如，您可以按照以下比例拆分数据：60% 的数据用于训练，20% 的数据用于评估，还有 20% 的数据用于验证。选择非常适合评估数据的模型参数之后，您可以使用验证数据再次进行评估，查看 ML 模型在运行验证数据时的表现。如果模型运行验证数据时的表现满足您的预期，就表示此模型没有过度拟合数据。

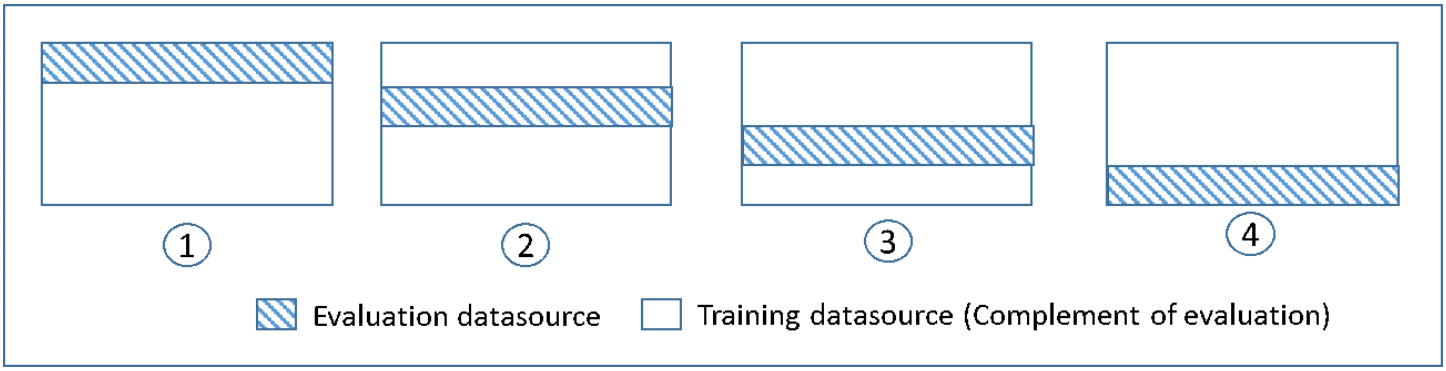
使用第三个数据集进行验证有助于您选择合适的 ML 模型参数来防止过度拟合。但是，将训练过程使用的数据预留用于评估和验证会减少可用于训练的数据。当数据集较小时，这个问题尤其明显，因为使用尽可能多的数据进行训练始终是最好的。要解决这个问题，您可以执行交叉验证。有关交叉验证的信息，请参阅交叉验证。

交叉验证

交叉验证是一种评估 ML 模型的方法，具体方法是通过使用可用输入数据子集训练多个 ML 模型并使用补充数据子集对其进行评估。使用交叉验证来检测过度拟合，即无法泛化模式。

在 Amazon ML 中，您可以使用 K 折交叉验证方法执行交叉验证。在 K 折交叉验证中，您将输入数据拆分为 k 个数据子集（也称为折叠）。 您在除某个子集 (k-1) 之外的所有子集上训练 ML 模型，然后在训练时未使用的子集上评估该模型。此过程将重复执行 k 次，每次使用一个预留用于评估的不同子集（训练中不包含的子集）。

下图显示了在 4 折交叉验证过程中创建和训练的四个模型中为每个模型生成的训练子集和补充评估的示例。模型一使用第一个 25% 的数据进行评估，使用其余 75% 的数据进行训练。模型二使用第二个 25%（25% 到 50%）的子集进行评估，使用其余三个数据子集进行训练，依此类推。



每个模型使用补充数据源进行训练和评估，包含评估数据源中的数据，并且仅限于训练数据源中不包含的所有数据。您可以使用 DataRearrangement、createDatasourceFromS3 和 createDatasourceFromRedShift API 中的 createDatasourceFromRDS 参数为这些子集中的每个子集创建数据源。在 DataRearrangement 参数中，通过指定每个分段的开始和结束位置来指定数据源将包含的数据的子集。要创建补充 4K 折交叉验证所需的补充数据源，请按照以下示例所示指定 DataRearrangement 参数：

模型一：

用于评估的数据源：

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

用于训练的数据源：

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

模型二：

用于评估的数据源：

```
{"splitting":{"percentBegin":25, "percentEnd":50}}
```

用于训练的数据源：

```
{"splitting":{"percentBegin":25, "percentEnd":50, "complement":"true"}}
```

模型三：

用于评估的数据源：

```
{"splitting":{"percentBegin":50, "percentEnd":75}}
```

用于训练的数据源：

```
{"splitting":{"percentBegin":50, "percentEnd":75, "complement":"true"}}
```

模型四：

用于评估的数据源：

```
{"splitting":{"percentBegin":75, "percentEnd":100}}
```

用于训练的数据源：

```
{"splitting":{"percentBegin":75, "percentEnd":100, "complement":"true"}}
```

执行 4 折交叉验证会生成四个模型、四个训练模型的数据源、四个评估模型的数据源和四个评估，每个模型一个。Amazon ML 为每个评估生成模型性能指标。例如，在适用于二进制分类问题的 4 折交叉验证中，每个评估都会报告曲线下面积 (AUC) 指标。您可以通过计算四个 AUC 指标的平均值获得整体性能测量值。有关 AUC 指标的信息，请参阅衡量 ML 模型准确度。

有关显示如何创建交叉验证和计算模型分数平均值的示例代码，请参阅 Amazon ML 示例代码。

调整您的模型

对模型执行交叉验证后，如果模型的表现未达到您的标准，您可以调整下一个模型的设置。有关过度拟合的更多信息，请参阅模型拟合：欠拟合与过度拟合。有关正则化的更多信息，请参阅正则化。有关更改正则化设置的更多信息，请参阅使用自定义选项创建 ML 模型。

评估警报

Amazon ML 提供见解，帮助验证您是否已正确评估模型。如果评估不满足任何验证标准，Amazon ML 控制台会显示已违反的验证标准来提醒您，如下所示。

- 已对留存数据完成 ML 模型的评估

  如果您对训练和评估使用同一个数据源，Amazon ML 会发出警报。如果您使用 Amazon ML 拆分数据，您将符合此有效性标准。如果您不使用 Amazon ML 来拆分数据，请务必使用训练数据源以外的数据源评估您的 ML 模型。
- 已将足量数据用于预测模型评估

  如果评估数据中的观察数/记录数少于训练数据源中观察数的 10%，Amazon ML 会发出警报。要正确地评估您的模型，必须提供足够大量的数据样本，这一点很重要。此标准提供了一个检查，让您知道您使用的数据是否太少。您的 ML 模型评估所需的数据量是主观决定的。在此处选择 10% 是作为缺乏更好措施时的权宜之计。
- 架构已匹配

  如果训练和评估数据源的架构不相同，Amazon ML 会发出警报。如果您的某些属性在评估数据源中不存在，或者如果您有其他属性，Amazon ML 会显示此警报。
- 评估文件中的所有记录已用于预测模型性能评估

  请务必了解用于评估的所有记录是否实际用于评估相应模型。如果评估数据源中的某些记录无效，并且未包含在准确性指标计算中，Amazon ML 会向您发出警报。例如，如果在评估数据源过程中，一些观察的目标变量缺失，Amazon ML 将无法检查针对这些观察的 ML 模型预测是否正确。在这种情况下，记录与缺失的目标值将被视为无效。
- 目标变量的分布

  Amazon ML 向您展示来自训练和评估数据源的目标属性的分布，以便您可以查看这两个数据源中的目标分布是否相似。对于利用训练数据建立的模型，如果其目标分布不同于评估数据的目标分布，则评估质量可能会受到影响，因为计算评估时依据的数据具有完全不同的统计数据。最好让训练数据和评估数据具有相似的数据分布，并让这些数据集尽可能地模拟在进行预测时模型将遇到的数据。

如果此警报触发，请尝试使用随机拆分策略将数据拆分为训练数据源和评估数据源。在极少数情况下，该警报可能会错误地提醒您目标分布还是有区别，即使您随机拆分了数据。Amazon ML 使用近似统计数据来评估数据分布，有时会错误地触发此警报。

生成和解释预测

Amazon ML 提供两种机制来生成预测：异步（分批次）和同步（一次一个）。

当您有大量观察并希望一次性获取这些观察的预测时，使用异步预测（即批量预测）。此过程使用数据源作为输入，将预测输出到所选 S3 存储桶中存储的 .csv 文件。您需要等待直至批量预测过程完成，然后才能访问预测结果。Amazon ML 在一批文件中可以处理的最大数据源大小为 1TB（大约 1 亿条记录）。如果您的数据源大于 1TB，则任务将失败，Amazon ML 会返回错误代码。为防止出现这种情况，请将数据拆分成几批。如果您的记录一般会更长，则在处理 1 亿条记录之前就会达到 1 TB 的限制。在这种情况下，我们建议您联系 AWS Support 以增加批量预测的任务大小。

在您希望马上获得预测时，使用同步（即实时预测）。实时预测 API 接受序列化为 JSON 字符串的单个输入观察，并在 API 响应中同步返回预测和关联的元数据。您可以同时调用多个 API 以并行获取同步预测。有关实时预测 API 吞吐量限制的更多信息，请参阅 Amazon ML API 参考中的实时预测限制。

主题

创建批量预测

要创建批量预测，您可以使用 Amazon Machine Learning (Amazon ML) 控制台或 API 创建 `BatchPrediction` 对象。`BatchPrediction` 对象描述 Amazon ML 使用您的 ML 模型和一组输入观察生成的预测集。在您创建 `BatchPrediction` 对象时，Amazon ML 启动计算预测的同步工作流。

对于您为下面两种数据源使用的架构必须相同：获取批量预测时使用的数据源，以及在训练为预测而查询的 ML 模型时使用的数据源。其中的一个例外是，批量预测的数据源无需包括目标属性，因为 Amazon ML 预测目标。如果您提供目标属性，Amazon ML 会忽略其值。

创建批量预测（控制台）

要使用 Amazon ML 控制台创建批量预测，请使用“创建批量预测”向导。

创建批量预测（控制台）

1. 通过以下链接登录 AWS Management Console 并打开 Amazon Machine Learning 控制台：https://console.aws.amazon.com/machinelearning/。
2. 在 Amazon ML 控制面板上的对象下，选择新建...，然后选择批量预测。
3. 选择您要用于创建批量预测的 Amazon ML 模型。
4. 要确认您希望使用此模型，请选择继续。
5. 选择您要为其创建预测的数据源。数据源必须与您模型具有相同的架构，尽管它无需包括目标属性。
6. 选择继续。
7. 对于S3 目标，键入您 S3 存储桶的名称。
8. 选择审核。
9. 检查您的设置，然后选择创建批量预测。

创建批量预测 (API)

要使用 Amazon ML API 创建 `BatchPrediction` 对象，您必须提供以下参数：

Datasource ID

指向要预测的观察的数据源的 ID。例如，如果您希望预测名为 `s3://examplebucket/input.csv` 的文件中的数据，则应创建指向该数据文件的数据源对象，然后将数据源的 ID 与此参数一起传入。

BatchPrediction ID

要分配到批量预测的 ID。

ML Model ID

Amazon ML 应在其中查询预测的 ML 模型的 ID。

Output Uri

S3 存储桶的 URI，在其中存储预测的输出。Amazon ML 必须有权将数据写入此存储桶。

`OutputUri` 参数必须引用以正斜杠（“/”）字符结尾的 S3 路径，如下例中所示：

s3://examplebucket/examplepath/

有关配置 S3 权限的信息，请参阅 向 Amazon ML 授予将预测输出到 Amazon S3 的权限。

（可选）BatchPrediction Name

（可选）批量预测的人类可读名称。

查看批量预测指标

Amazon Machine Learning (Amazon ML) 创建批量预测后，它会提供两个指标：`Records seen` 和 `Records failed to process`。`Records seen` 说明 Amazon ML 在运行您的批量预测时查看了多少条记录。`Records failed to process` 说明 Amazon ML 无法处理多少条记录。

要允许 Amazon ML 处理失败的记录，请检查用于创建您的数据源的数据中的记录格式，并确保所有必需的属性均已存在并且所有数据均正确。修复您的数据后，您可以重新创建批量预测，或者使用失败的记录创建一个新的数据源，然后使用新数据源创建新的批量预测。

查看批量预测指标（控制台）

要查看 Amazon ML 控制台中的指标，请打开批量预测摘要页面并查看已处理信息部分。

查看批量预测指标和详细信息 (API)

您可以使用 Amazon ML API 来检索 `BatchPrediction` 对象的详细信息，包括记录指标。Amazon ML 提供以下批量预测 API 调用：

- `CreateBatchPrediction`
- `UpdateBatchPrediction`
- `DeleteBatchPrediction`
- `GetBatchPrediction`
- `DescribeBatchPredictions`

有关更多信息，请参阅 Amazon ML API 参考。

读取批量预测输出文件

请执行以下步骤检索批量预测输出文件：

1. 找到批量预测清单文件。
2. 读取清单文件以确定输出文件的位置。

3. 检索包含预测结果的输出文件。
4. 解释输出文件的内容。所含内容因用于生成预测的 ML 模型的类型而异。

以下各部分详细介绍了这些步骤。

找到批量预测清单文件

批量预测的清单文件包含将输入文件映射到预测输出文件的信息。

要查找清单文件，请先从您在创建批量预测对象时指定的输出位置开始查找。您可以使用 Amazon ML API 或访问 https://console.aws.amazon.com/machinelearning/ 查询已完成的批量预测对象，以检索此文件的 S3 位置。

清单文件所在的输出位置的路径中包含附加到输出位置的静态字符串 `/batch-prediction/` 和清单文件的名称，即附加了扩展名 `.manifest` 的批量预测的 ID。

例如，如果创建包含 ID `bp-example` 的批量预测对象，并将 S3 位置 `s3://examplebucket/output/` 指定为输出位置，则将在此处找到您的清单文件：

`s3://examplebucket/output/batch-prediction/bp-example.manifest`

读取清单文件

.manifest 文件的内容以 JSON 映射的格式进行编码，其中的键是表示 S3 输入数据文件名称的字符串，值是表示与其关联的批量预测结果文件的字符串。每个输入/输出文件对都包含一个映射行。继续使用我们的示例，如果创建 `BatchPrediction` 对象时使用的输入包含一个名为 data.csv 的文件，该文件位于 `s3://examplebucket/input/` 中，您可能会看到如下所示的映射字符串：

```
{"s3://examplebucket/input/data.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data.csv.gz"}
```

如果创建 `BatchPrediction` 对象时使用的输入包含三个分别名为 data1.csv、data2.csv 和 data3.csv 的文件，这些文件都存储在 S3 位置 `s3://examplebucket/input/` 中，您可能会看到如下所示的映射字符串：

```
{"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"}
```

检索批量预测输出文件

您可以下载通过清单映射获取的每个批量预测文件，然后在本地处理映射该文件。文件格式为 CSV，使用 gzip 算法进行压缩。相应输入文件中的每个输入观察在该文件中对应一行。

要联接预测结果和批量预测的输入文件，您可以对两个文件执行简单的按记录逐个合并。批量预测的输出文件包含的记录数量始终与预测输入文件的记录数量相同，并且采用相同的顺序。如果输入观察处理失败，无法生成任何预测，批量预测的输出文件将在相应的位置包含一个空白行。

解释二进制分类 ML 模型的批量预测文件的内容

二进制分类模型的批量预测文件的列命名为 bestAnswer 和 score。

bestAnswer 列包含的预测标签（“1”或“0”）是通过对截断值分数评估预测分数得到的。有关截断值分数的更多信息，请参阅调整分数截断值。您可以使用 Amazon ML API 或 Amazon ML 控制台上的模型评估功能为 ML 模型设置截断值分数。如果未设置截断值分数，Amazon ML 会使用默认值 0.5。

score 列包含 ML 模型为此预测分配的原始预测分数。Amazon ML 使用逻辑回归模型，因此该分数尝试为观察的概率建模，该概率对应于 true（“1”）值。请注意，score 采用科学计数法报告，因此在以下示例的第一行中，值 8.7642E-3 等于 0.0087642。

例如，如果 ML 模型的截断值分数是 0.75，二进制分类模型的批量预测输出文件的内容可能如下所示：

```
bestAnswer,score

0,8.7642E-3

1,7.899012E-1


0,6.323061E-3

0,2.143189E-2
```

```
1,8.944209E-1
```

输入文件中第二个和第五个观察收到的预测分数高于 0.75，因此这些观察的 bestAnswer 列显示值“1”，而其他观察的值为“0”。

解释多类别分类 ML 模型的批量预测文件的内容

多类别模型的批量预测文件包含在训练数据中找到的每个类对应的列。列名显示在批量预测文件的标头行中。

当您请求多类别模型进行预测时，Amazon ML 会为输入文件中每个观察计算几个预测分数，每个分数对应于在输入数据集定义的每个类。这等同于询问“此观察属于此类而不是任何其他类的概率（测量值位于 0 和 1 之间）是多少？” 每个分数可以解释为“观察属于此类的概率”。由于预测分数为观察属于一个类或另一个类的潜在概率建模，因此每行所有预测分数的总和为 1。您需要先选择一个类作为模型的预测类。通常，您应选择概率最高的类作为最佳答案。

例如，请考虑尝试在 1 星至 5 星的范围内预测客户的产品评级。如果类命名为 `1_star`、`2_stars`、`3_stars`、`4_stars` 和 `5_stars`，多类别预测输出文件可能如下所示：

```
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
```

在本示例中，第一个观察具有 `3_stars` 类的最高预测分数（预测分数 = 4.77781E-1），因此您可以理解为此结果表示 `3_stars` 类是此项观察的最佳答案。请注意，预测分数采用科学计数法报告，因此预测分数 4.77781E-1 等于 0.477781。

在某些情况下，您可能不想选择概率最高的类。例如，您可能想创建低于您不考虑作为最佳答案的类的最小阈值，即使它包含最高预测分数。假设您要按流派为电影分类，并且要让预测分数至少为 5E-1，才声明该流派将为您的最佳答案。获得的预测分数为 3E-1 是喜剧片，2.5E 是戏剧，2.5E-1 是纪录片，2E-1 是动作片。在这种情况下，ML 模型会预测喜剧是您最有可能的选择，但您决定不选择它作为最佳答案。由于没有任何预测分数超出您的基准预测分数 5E-1，因此您确定此项预测不足以自信地预测流派，从而决定选择其他预测。您的应用程序可能会将此电影的流派字段视为“未知”。

解释回归 ML 模型的批量预测文件的内容

回归模型的批量预测文件包含名为分数的单个列。此列包含输入数据中每个观察的原始数字预测。这些值采用科学计数法报告，因此以下示例的第一行中分数的值 -1.526385E1 等于 -15.26835。

此示例显示了在回归模型上执行的批量预测的输出文件：

```
score

-1.526385E1

-6.188034E0

-1.271108E1

-2.200578E1

8.359159E0
```

请求实时预测

实时预测是对 Amazon Machine Learning (Amazon ML) 的同步调用。Amazon ML 在收到请求时进行预测，并立即返回响应。实时预测通常用于实现交互式 Web、移动或桌面应用程序中的预测功能。您可以使用低延迟 `Predict` API，查询使用 Amazon ML 创建的 ML 模型进行实时预测。`Predict` 操作接受请求负载中的单个输入观察并在响应中同步返回预测。这使其有别于批量预测 API，后者使用指向输入观察位置的 Amazon ML 数据源对象的 ID 进行调用，并异步返回 URI，指向包含所有这些观察的预测的文件。Amazon ML 响应大多数实时预测请求的时间不超过 100 毫秒。

您可以在 Amazon ML 控制台中尝试实时预测而不产生任何费用。如果您随后决定使用实时预测，您必须首先为生成实时预测而创建终端节点。您可以在 Amazon ML 控制台中或者使用 `CreateRealtimeEndpoint` API 执行此操作。在您有终端节点之后，使用实时预测 API 来生成实时预测。

> **Note**
>
> 在您为模型创建实时终端节点之后，您将开始产生基于模型大小的容量预留费用。有关更多信息，请参阅 定价。如果您在控制台中创建实时终端节点，控制台会显示终端节点将持续产生的估计费用明细。要在您不再需要从模型获取实时预测时停止产生费用，请使用控制台或 `DeleteRealtimeEndpoint` 操作删除实时终端节点。

有关 `Predict` 请求和响应的示例，请参阅 Amazon Machine Learning API 参考中的预测。要查看使用您模型的确切响应格式的示例，请参阅试用实时预测。

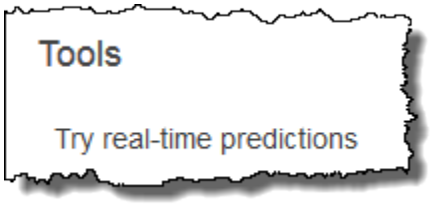
主题

试用实时预测

为了帮助您决定是否启用实时预测，Amazon ML 允许您尝试针对单个数据记录生成预测，而不会产生与设置实时预测终端节点相关的额外费用。要试用实时预测，您必须拥有 ML 模型。要创建更大规模的实时预测，请使用 Amazon Machine Learning API 参考中的预测 API。

尝试实时预测

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏的 Amazon Machine Learning 下拉菜单中，选择 ML 模型。
3. 从教程中选择您试用实时预测所要使用的模型，例如 `Subscription propensity model`。
4. 在 ML 模型报告页面的预测下，选择摘要，然后选择尝试实时预测。



   Amazon ML 显示变量列表，这些变量组成了 Amazon ML 训练您模型时使用的数据记录。
5. 您可以继续在表单的各个字段中输入数据，或者以 CSV 格式粘贴单个数据记录到文本框中。

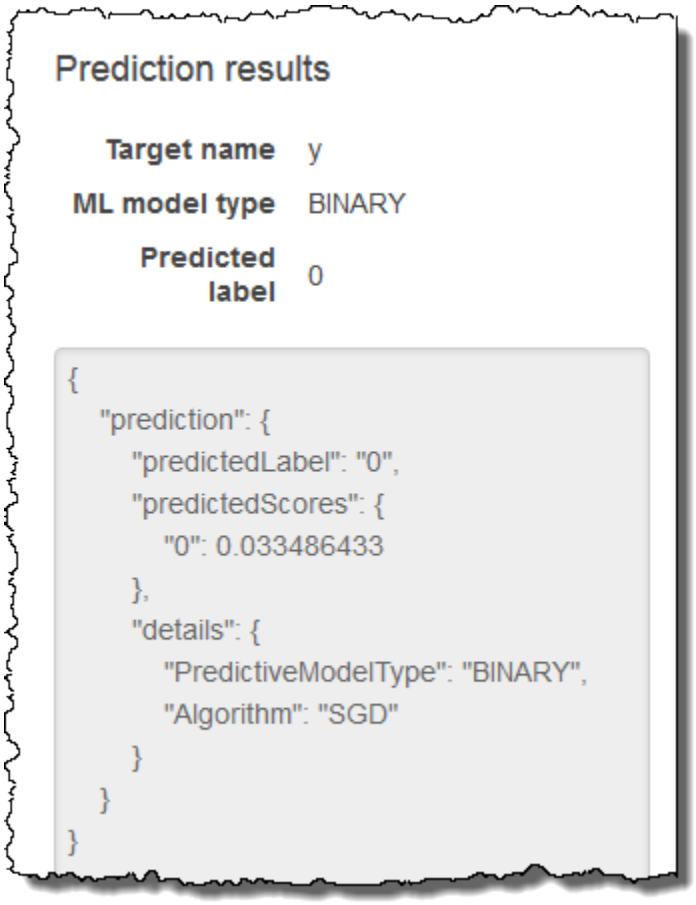
   要使用表单，对于各个值字段，输入您希望用于测试实时预测的数据。如果您输入的数据记录不包含一个或多个数据属性的值，请将条目字段留空。

要提供数据记录，请选择粘贴记录。将 CSV 格式的一行数据粘贴到文本字段中，然后选择提交。Amazon ML 自动为您填充值字段。

> **Note**
>
> 数据记录中的数据必须与训练数据具有相同的列数，并且按相同顺序排列。唯一例外是您应省略目标值。如果您包括目标值，Amazon ML 将忽略它。

6. 在页面底部，选择创建预测。Amazon ML 立即返回预测。

在预测结果窗格中，您可以看到 `Predict` API 调用返回的预测对象，以及 ML 模式类型、目标变量的名称以及预测的类别或值。有关解释结果的更多信息，请参阅解释二进制分类 ML 模型的批量预测文件的内容。



创建实时终端节点

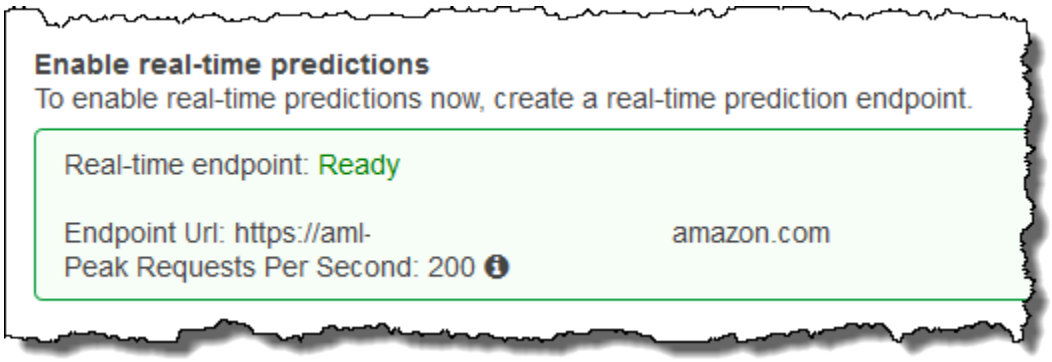
要生成实时预测，您需要创建实时终端节点。要创建实时终端节点，您必须已经有要用于生成实时预测的 ML 模型。您可以使用 Amazon ML 控制台或者调用 `CreateRealtimeEndpoint` API 来创建实时终端节点。有关使用 `CreateRealtimeEndpoint` API 的更多信息，请参阅《Amazon Machine Learning API 参考》中的 https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_CreateRealtimeEndpoint.html。

创建实时终端节点

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏的 Amazon Machine Learning 下拉菜单中，选择 ML 模型。
3. 选择要生成实时预测的模型。
4. 在 ML 模型摘要页面上的预测下，选择创建实时终端节点。

   此时会显示一个对话框，说明如何为实时预测定价。
5. 选择创建。实时终端节点请求发送到 Amazon ML 并进入队列中。实时终端节点的状态为正在更新。

6. 实时终端节点就绪之后，状态更改为准备就绪，并且 Amazon ML 显示终端节点 URL。使用终端节点 URL 可通过 `Predict` API 创建实时预测请求。有关使用 `Predict` API 的更多信息，请参阅《Amazon Machine Learning API 参考》中的 https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html。



查找实时预测终端节点（控制台）

要使用 Amazon ML 控制台查找 ML 模型的终端节点 URL，请导航到模型的 ML 模型摘要页面。

查找实时终端节点 URL

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏的 Amazon Machine Learning 下拉菜单中，选择 ML 模型。
3. 选择要生成实时预测的模型。
4. 在 ML 模型摘要页面上，向下滚动直至您看到预测部分。
5. 模型的终端节点 URL 在实时预测中列出。使用 URL 作为您实时预测调用的终端节点 Url。有关如何使用终端节点生成预测的信息，请参阅《Amazon Machine Learning API 参考》中的 https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html。

查找实时预测终端节点 (API)

在您使用 `CreateRealtimeEndpoint` 操作创建实时终端节点时，在响应中向您返回终端节点的 URL 和状态。如果您使用控制台创建实时终端节点，或者如果您希望检索以前创建的终端节点的 URL 和状态，请使用您要用来查询实时预测的模型的 ID 调用 `GetMLModel` 操作。终端节点信息包含在响应的 `EndpointInfo` 部分中。对于关联了实时终端节点的模型，`EndpointInfo` 可能类似于：

```
"EndpointInfo":{
    "CreatedAt": 1427864874.227,
    "EndpointStatus": "READY",
    "EndpointUrl": "https://endpointUrl",
    "PeakRequestsPerSecond": 200
```

```
}
```

没有实时终端节点的模型将返回以下内容：

```
EndpointInfo":{
    "EndpointStatus": "NONE",
    "PeakRequestsPerSecond": 0
}
```

创建实时预测请求

示例 Predict 请求有效负载可能类似于下面这样：

```
{
    "MLModelId": "model-id",
    "Record":{
        "key1": "value1",
        "key2": "value2"
    },
    "PredictEndpoint": "https://endpointUrl"
}
```

PredictEndpoint 字段必须对应于 EndpointInfo 结构的 EndpointUrl 字段。Amazon ML 使用此字段将请求路由到实时预测队列中的相应服务器。

MLModelId 是以前训练的模型（带有实时终端节点）的标识符。

Record 是变量名到变量值的映射。每一对表示一个观察。Record 映射包含对 Amazon ML 模型的输入。这类似于训练数据集中无目标变量的单行数据。无论在训练数据中使用哪种值类型，Record 都包含字符串到字符串映射。

> **Note**
>
> 您可以忽略您没有值的变量，不过这可能会减少预测的准确性。您包括的变量越多，模型就越准确。

Predict 请求返回的响应格式取决于进行预测时查询的模型类型。在所有情况下，details 字段包含有关预测请求的信息，特别是包括带有模型类型的 PredictiveModelType 字段。

以下示例显示二进制模型的响应：

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "BINARY"
        },
        "predictedLabel": "0",
        "predictedScores":{
            "0": 0.47380468249320984
        }
    }
}
```

请注意包含预测标签的 predictedLabel 字段，在本例中为 0。Amazon ML 将预测分数与分类截断值进行比较来计算预测标签。

- 您可以检查 GetMLModel 操作的响应中的 ScoreThreshold 字段，或者在 Amazon ML 控制台中查看模型信息，来获取当前与 ML 模型关联的分类截断值。如果您未设置分数阈值，Amazon ML 会使用默认值 0.5。
- 您可以通过检查 predictedScores 映射来获取二进制分类模型的确切预测分数。在此映射中，预测标签与确切的预测分数成对使用。

有关二进制预测的更多信息，请参阅解释预测。

以下示例显示来自递归模型的响应。请注意在 predictedValue 字段中找到的预测数值：

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "REGRESSION"
        },
        "predictedValue": 15.508452415466309
    }
}
```

以下示例显示多类别模型的响应：

```
{
    "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
        }
    }
}
```

预测标签/分类与二进制分类模型类似，可在 `predictedLabel` 字段中找到。您可以通过查看 `predictedScores` 映射，进一步了解预测与各分类相关联的强度。此映射中某个分类的分数越高，预测与该分类的相关性就越强，最高值最终被选择作为 `predictedLabel`。

有关多分类预测的更多信息，请参阅多类别模型洞察。

删除实时终端节点

当您完成实时预测时，请删除实时终端节点以避免产生额外的费用。在您删除终端节点之后，立即停止产生费用。

删除实时终端节点

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏的 Amazon Machine Learning 下拉菜单中，选择 ML 模型。
3. 选择不再需要实时预测的模型。
4. 在 ML 模型报告页面上的预测下，选择摘要。
5. 选择删除实时终端节点。
6. 在删除实时终端节点对话框中，选择删除。

管理 Amazon ML 对象

Amazon ML 提供了四种您可以通过 Amazon ML 控制台或 Amazon ML API 管理的对象：

- 数据源
- ML 模型
- 评估
- 批量预测

在构建机器学习应用程序的生命周期中，每个对象起到不同的作用，并且每个对象有仅适用于该对象的特定属性和功能。尽管存在这些差异，您仍可以使用相似的方式管理对象。例如，您可以使用几乎相同的流程来列出对象、检索其说明以及更新或删除它们。

以下各部分介绍对所有四个对象通用的管理操作，并说明了所有区别。

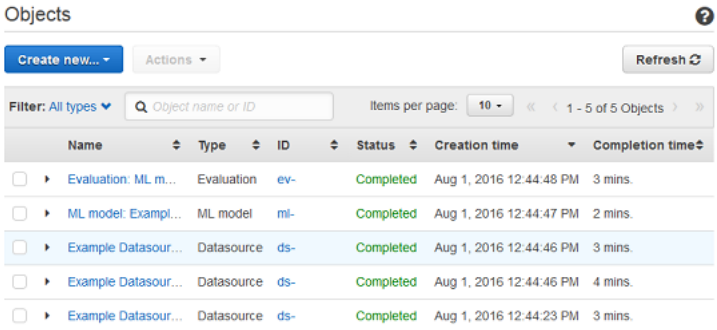
主题

列出对象

如需 Amazon Machine Learning (Amazon ML) 数据源、ML 模型、评估和批量预测的深入信息，请列出它们。对于每个对象，您将看到其名称、类型、ID、状态代码和创建时间。您还可以查看具体对象类型特有的详细信息。例如，您可以查看某个数据源的数据洞察。

列出对象（控制台）

要查看您创建的后 1000 个对象的列表，请在 Amazon ML 控制台中打开对象控制面板。要显示对象控制面板，请登录 Amazon ML 控制台。



要查看有关对象的详细信息（包括该对象类型特有的详细信息），请选择该对象的名称或 ID。例如，要查看某个数据源的数据洞察，请选择数据源名称。

对象控制面板上的各列显示有关各对象的下列信息。

名称

对象的名称。

Type

对象的类型。有效值包括数据源、ML 模型、评估和批量预测。

> **Note**
>
> 要查看某个模型是否设置为支持实时预测，请通过选择名称或模型 ID 转到 ML 模型摘要页面。

ID

对象的 ID。

状态

对象的状态。值包括待处理、正在进行、已完成和已失败。如果状态为已失败，请检查您的数据并重试。

Creation time

Amazon ML 完成创建此对象的日期和时间。

Completion time

Amazon ML 创建此对象所用的时间长度。您可以使用模型的完成时间来估算新模型的训练时间。

Datasource ID

数据源的 ID，适用于使用数据源创建的对象（例如模型和评估）。如果您删除了数据源，则不能再使用以该数据源创建的 ML 模型来创建预测。

通过选择任意列标题旁的双三角形图标可以按该列排序。

列出对象 (API)

在 Amazon ML API 中，您可以使用以下操作按类型列出对象：

- `DescribeDataSources`
- `DescribeMLModels`
- `DescribeEvaluations`
- `DescribeBatchPredictions`

每个操作均包括对一长列对象进行筛选、排序和分页。通过 API 可以访问的对象数没有限制。要限制列表的大小，请使用 `Limit` 参数，其最大值可以为 100。

API 对 `Describe*` 命令的响应包括分页令牌 (`nextPageToken`)（如果适用），以及各对象的简要说明。对于控制台中显示的各对象类型，对象说明提供的信息都是相同的，包括某个对象类型特有的详细信息。

> **Note**
>
> 即使响应包括的对象数少于指定的限制，它也可能包括指示有更多结果可用的 `nextPageToken`。甚至包含 0 个项目的响应也可能包含 `nextPageToken`。

有关更多信息，请参阅 Amazon ML API 参考。

检索对象描述

您可以通过控制台或 API 查看任何对象的详细描述。

通过控制台查看详细描述

要在控制台上查看描述，请导航到特定对象类型 (数据源、ML 模型、评估或批量预测) 的列表。接下来，找到与对象对应的表中的行，可通过浏览对象列表或搜索其名称或 ID 来查找。

通过 API 查看详细描述

每个对象类型都包含可检索 Amazon ML 对象的完整详细信息的操作：

- GetDataSource
- GetMLModel
- GetEvaluation
- GetBatchPrediction

每个操作有且仅有两个参数：对象 ID 和名为 Verbose 的布尔值标记。将 Verbose 设置为 true 进行调用将包含对象的额外详细信息，导致出现较长的延迟和更大的响应。要了解设置 Verbose 标记将包含哪些字段，请参阅 Amazon ML API 参考。

更新对象

每个对象类型都有一个更新 Amazon ML 对象详细信息的操作（请参阅 Amazon ML API 参考）：

- UpdateDataSource
- UpdateMLModel
- UpdateEvaluation
- UpdateBatchPrediction

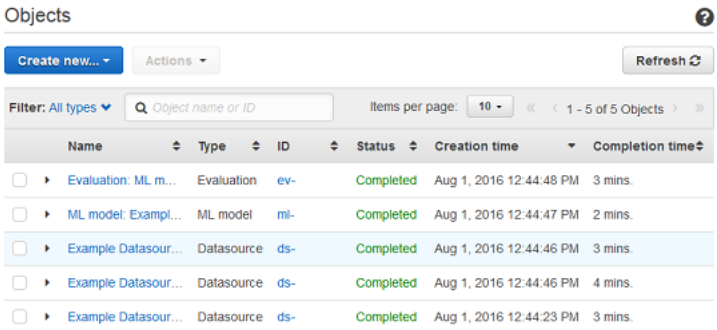
每个操作需要对象 ID 来指定要更新的对象。您可以更新所有对象的名称。您无法更新数据源、评估以及批量预测的对象的任何其他属性。对于 ML 模型，只要 ML 模型没有关联的实时预测终端节点，您就可以更新 ScoreThreshold 字段。

删除对象

当您不再需要您的数据源、ML 模型、评估和批量预测时，您可以删除它们。尽管在完成后，您可以保留除批量预测之外的 Amazon ML 对象而没有任何额外费用，不过删除这些对象有助于保持您的工作区

整洁和易于管理。您可以使用 Amazon Machine Learning (Amazon ML) 控制台或 API 删除单个或多个对象。

> **Warning**
>
> 当您删除 Amazon ML 对象时，其效果是即时、永久和不可逆转的。



删除对象（控制台）

您可以使用 Amazon ML 控制台删除对象，包括模型。删除模型时使用的过程取决于您是否使用模型来生成实时预测。要删除用于生成实时预测的模型，请首先删除实时终端节点。

删除 Amazon ML 对象（控制台）

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台（网址为 https://console.aws.amazon.com/machinelearning/）。
2. 选择您要删除的 Amazon ML 对象。要选择多个对象，请使用 SHIFT 键。要取消选择所有选定对象，请使用

   或

   按钮。
3. 对于操作，选择删除。
4. 在对话框中，选择删除以删除模型。

删除具有实时终端节点的 Amazon ML 模型（控制台）

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台（网址为 https://console.aws.amazon.com/machinelearning/）。
2. 选择要删除的模型。
3. 对于操作，选择删除实时终端节点。
4. 选择删除以删除终端节点。
5. 再次选择模型。
6. 对于操作，选择删除。
7. 选择删除以删除模型。

删除对象 (API)

您可以使用以下 API 调用删除 Amazon ML 对象：

- `DeleteDataSource` - 获取参数 `DataSourceId`。
- `DeleteMLModel` - 获取参数 `MLModelId`。
- `DeleteEvaluation` - 获取参数 `EvaluationId`。
- `DeleteBatchPrediction` - 获取参数 `BatchPredictionId`。

有关更多信息，请参阅 Amazon Machine Learning API 参考。

使用 Amazon CloudWatch 指标监控 Amazon ML

Amazon ML 会自动向 Amazon CloudWatch 发送指标，这样您可以收集并分析 ML 模型的使用情况统计数据。例如，要跟踪批量预测和实时预测，您可以根据 RequestMode 维度监视 PredictCount 指标。系统将每五分钟自动收集这些指标并发送到 Amazon CloudWatch。您可以使用 Amazon CloudWatch 控制台、AWS CLI 或 AWS 软件开发工具包来监控这些指标。

通过 CloudWatch 报告 Amazon ML 指标无需任何费用。如果您在指标上设置了警报，则会向您收取标准 CloudWatch 费率。

有关信息，请参阅《Amazon CloudWatch 开发人员指南》中 Amazon CloudWatch 名称空间、维度和指标参考的 Amazon ML 指标列表。

使用 AWS CloudTrail 记录 Amazon ML API 调用

Amazon Machine Learning (Amazon ML) 与 AWS CloudTrail 集成，后者是在 Amazon ML 中提供用户、角色或 AWS 服务所采取操作的记录的服务。CloudTrail 将 Amazon ML 的所有 API 调用作为事件捕获。捕获调用中包括通过 Amazon ML 控制台的调用和对 Amazon ML API 操作的代码调用。如果您创建跟踪，则可以使 CloudTrail 事件持续传送到 Amazon S3 存储桶（包括 Amazon ML 的事件）。如果您不配置跟踪，则仍可在 CloudTrail 控制台中的 Event history（事件历史记录）中查看最新事件。使用 CloudTrail 收集的信息，您可以确定向 Amazon ML 发出了什么请求、发出请求的 IP 地址、何人发出的请求、请求的发出时间以及其他详细信息。

要了解有关 CloudTrail 的更多信息（包括如何对其进行配置和启用），请参阅 AWS CloudTrail 用户指南。

CloudTrail 中的 Amazon ML 信息

在您创建 AWS 账户时，将在该账户上启用 CloudTrail。当 Amazon ML 中发生受支持的事件活动时，该活动将记录在 CloudTrail 事件中，并与其他 AWS 服务事件一同保存在事件历史记录中。您可以在 AWS 账户中查看、搜索和下载最新事件。有关更多信息，请参阅使用 CloudTrail 事件历史记录查看事件。

要持续记录 AWS 账户中的事件（包括 Amazon ML 的事件），请创建跟踪。通过跟踪，CloudTrail 可将日志文件传送至 Amazon S3 存储桶。默认情况下，在控制台中创建跟踪记录时，此跟踪记录应用于所有亚马逊云科技区域。此跟踪在 AWS 分区中记录所有区域中的事件，并将日志文件传送至您指定的 Amazon S3 存储桶。此外，您可以配置其他 AWS 服务，进一步分析在 CloudTrail 日志中收集的事件数据并采取行动。有关更多信息，请参阅下列内容：

- 创建跟踪概览
- CloudTrail 支持的服务和集成
- 为 CloudTrail 配置 Amazon SNS 通知
- 从多个区域接收 CloudTrail 日志文件和从多个账户接收 CloudTrail 日志文件

Amazon ML 支持将以下操作记录为 CloudTrail 日志文件中的事件：

- AddTags
- CreateBatchPrediction
- CreateDataSourceFromRDS

- CreateDataSourceFromRedshift
- CreateDataSourceFromS3
- CreateEvaluation
- CreateMLModel
- CreateRealtimeEndpoint
- DeleteBatchPrediction
- DeleteDataSource
- DeleteEvaluation
- DeleteMLModel
- DeleteRealtimeEndpoint
- DeleteTags
- DescribeTags
- UpdateBatchPrediction
- UpdateDataSource
- UpdateEvaluation
- UpdateMLModel

以下 Amazon ML 操作使用包含凭证的请求参数。在这些请求发送到 CloudTrail 之前，使用三个星号替换凭证（“***”）：

- CreateDataSourceFromRDS
- CreateDataSourceFromRedshift

使用 Amazon ML 控制台执行以下 Amazon ML 操作时，属性 `ComputeStatistics` 不包括在 CloudTrail 日志的 `RequestParameters` 组件中：

- CreateDataSourceFromRedshift
- CreateDataSourceFromS3

每个事件或日志条目都包含有关生成请求的人员信息。身份信息可帮助您确定以下内容：

- 请求是使用根用户凭证还是 AWS Identity and Access Management (IAM) 用户凭证发出的。
- 请求是使用角色还是联合身份用户的临时安全凭证发出的。

- 请求是否由其它 AWS 服务发出。

有关更多信息，请参阅 CloudTrail userIdentity 元素。

示例：Amazon ML 日志文件条目

跟踪是一种配置，可用于将事件作为日志文件传送到您指定的 Amazon S3 桶。CloudTrail 日志文件包含一个或多个日志条目。一个事件表示来自任何源的一个请求，包括有关所请求的操作、操作的日期和时间、请求参数等方面的信息。CloudTrail 日志文件不是公用 API 调用的有序堆栈跟踪，因此它们不会按任何特定顺序显示。

下面的示例显示了一个 CloudTrail 日志条目，该条目说明了操作。

```
{
    "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"
        }
    ]
}
```

标记您的 Amazon ML 对象

通过向您的 Amazon Machine Learning (Amazon ML)对象分配带有标签的元数据，整理和管理这些对象。标签是您为对象定义的键值对。

除了使用标签来整理和管理 Amazon ML 对象之外，您还可以使用它们来分类和跟踪您的 AWS 成本。当您将标签应用于 AWS 对象（包括 ML 模型）时，您的 AWS 成本分配报告将包括按标签汇总的使用率和成本。通过应用代表业务类别（例如成本中心、应用程序名称或所有者）的标签，您可以整理多种服务的成本。有关更多信息，请参阅 AWS Billing 用户指南中的对自定义账单报告使用成本分配标签。

目录

有关标签的基本知识

使用标签分类对象可以轻松管理它们。例如，您可以按用途、所有者或环境分类对象。然后，您可以定义一组标签来帮助您按所有者和关联应用程序跟踪模型。下面是几个示例：

- 项目：项目名称
- 所有者：名称
- 用途：营销预测
- 应用程序：应用程序名称
- 环境：生产

使用 Amazon ML 控制台或 API 可完成以下任务：

- 向对象添加标签
- 查看对象的标签
- 编辑对象的标签
- 删除对象的标签

默认情况下，应用到 Amazon ML 对象的标签复制到使用该对象创建的对象。例如，如果 Amazon Simple Storage Service (Amazon S3) 数据源具有“Marketing cost: Targeted marketing campaign”标签，则使用该数据源创建的模型也将具有“Marketing cost: Targeted marketing campaign”，对该模型的评估也是如此。这样，您就可以使用标签跟踪相关对象，如用于营销活动的所有对象。如果标签源之间有冲突，例如某个模型带有标签“Marketing cost: Targeted marketing campaign”，而数据源带有标签“Marketing cost: Target marketing customers”，Amazon ML 应用来自模型的标签。

标签限制

以下限制适用于标签。

基本限制:

- 每个对象的最大标签数为 50。
- 标签键和值区分大小写。
- 无法更改或编辑已删除对象的标签。

标签键限制:

- 每个标签键必须是唯一的。如果您添加的标签具有已使用的键，则您的新标签将覆盖该对象的现有键值对。
- 标签键不能以 `aws:` 开头，因为此前缀将预留以供 AWS 使用。AWS 将代表您创建以此前缀开头的标签，但您不能编辑或删除这些标签。
- 标签键的长度必须介于 1 和 128 个 Unicode 字符之间。
- 标签键必须包含以下字符：Unicode 字母、数字、空格和以下特殊字符：`_ . / = + - @`。

标签值限制:

- 标签值的长度必须介于 0 和 255 个 Unicode 字符之间。
- 标签值可以为空。另外，它们必须包含以下字符：Unicode 字母、数字、空格和以下任意特殊字符：`_ . / = + - @`。

标记 Amazon ML 对象（控制台）

您可以使用 Amazon ML 控制台查看、添加、编辑和删除标签。

查看对象的标签（控制台）

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏中，展开区域选择器并选择一个区域。
3. 在对象页面上，选择对象。
4. 滚动到所选对象的标签部分。该对象的标签在这一部分的底部列出。

向对象添加标签（控制台）

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏中，展开区域选择器并选择一个区域。
3. 在对象页面上，选择对象。
4. 滚动到所选对象的标签部分。该对象的标签在这一部分的底部列出。
5. 选择 Add or edit tags。
6. 在添加标签下的密钥字段中指定标签键，（可选）在值字段中指定标签值，然后选择应用更改。

   如果未启用应用更改按钮，则您指定的标签键或标签值不满足标签限制。有关更多信息，请参阅标签限制。
7. 要在标签部分的列表中查看您的新标签，请刷新页面。

编辑标签（控制台）

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏中，展开区域选择器并选择一个区域。
3. 在对象页面上，选择对象。
4. 滚动到所选对象的标签部分。该对象的标签在这一部分的底部列出。
5. 选择 Add or edit tags。
6. 在已应用标签下，编辑值字段中的标签值，然后选择应用更改。

   如果未启用应用更改按钮，则您指定的标签值不满足标签限制。有关更多信息，请参阅标签限制。
7. 要在标签部分的列表中查看您更新后的标签，请刷新页面。

删除对象的标签（控制台）

1. 登录 AWS Management Console 并打开 Amazon Machine Learning 控制台 (https://console.aws.amazon.com/machinelearning/)。
2. 在导航栏中，展开区域选择器并选择一个区域。
3. 在对象页面上，选择对象。
4. 滚动到所选对象的标签部分。该对象的标签在这一部分的底部列出。
5. 选择 Add or edit tags。
6. 在已应用标签下，选择要删除的标签，然后选择应用更改。

标记 Amazon ML 对象 (API)

您可以使用 Amazon ML API 添加、列出和删除标签。有关示例，请参阅以下文档：

AddTags

为指定对象添加或编辑标签。

DescribeTags

列出指定对象的标签。

DeleteTags

删除指定对象的标签。

Amazon Machine Learning 参考

主题

为 Amazon ML 授予从 Amazon S3 读取您的数据的权限

要在 Amazon S3 中使用您的输入数据创建数据源对象，您必须在存储输入数据的 S3 位置为 Amazon ML 授予以下权限：

- S3 存储桶和前缀的 GetObject 权限。
- S3 存储桶的 ListBucket 权限。与其他操作不同，必须授予存储桶范围（而不是前缀）的 ListBucket 权限。但是，您可以使用 Condition 子句限定特定前缀的权限范围。

如果您使用 Amazon ML 控制台创建数据源，这些权限可添加到您的存储桶。系统将提示您确认是否要添加这些权限以完成向导中的步骤。以下策略示例介绍了如何为 Amazon ML 授予从示例位置 s3://*examplebucket*/*exampleprefix* 读取数据的权限，而将 ListBucket 权限的范围限定为只限 *exampleprefix* 输入路径。

```
{
    "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:*" }
        }
    }]
}
```

要将此策略应用到您的数据，您必须编辑与存储数据的 S3 存储桶关联的策略语句。

为 S3 存储桶编辑权限策略（使用旧控制台）

1. 登录到 AWS Management Console，然后通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 选择数据所在的存储桶的名称。
3. 请选择属性。
4. 选择编辑存储桶策略。
5. 如上所示输入策略，进行自定义以符合您的需求，然后选择保存。
6. 选择保存。

为 S3 存储桶编辑权限策略（使用新控制台）

1. 登录到 AWS Management Console，然后通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 选择存储桶名称，然后选择权限。
3. 请选择桶策略。

4. 如上所示输入策略，进行自定义以符合您的需求。
5. 选择保存。

向 Amazon ML 授予将预测输出到 Amazon S3 的权限

要将批量预测操作的结果输出到 Amazon S3，您必须向 Amazon ML 授予输出位置的以下权限，该位置是作为“创建批量预测”操作的输入提供的：

- S3 存储桶和前缀上的 GetObject 权限。
- S3 存储桶和前缀上的 PutObject 权限。
- S3 存储桶和前缀上的 PutObjectAcl。
  - Amazon ML 需要此权限来确保可在创建对象之后将标准 ACL bucket-owner-full-control 权限授予您的 AWS 账户。
- S3 存储桶的 ListBucket 权限。与其他操作不同，必须授予存储桶范围（而不是前缀）的 ListBucket 权限。不过，您可以使用 Condition 子句限定特定前缀的权限范围。

如果您使用 Amazon ML 控制台创建批量预测请求，这些权限可添加到您的存储桶。在您完成向导中的步骤时，系统将提示您确认是否要添加它们。

以下示例策略介绍如何授予 Amazon ML 权限以将数据写入示例位置 s3://examplebucket/exampleprefix，同时将 ListBucket 权限仅限定为 exampleprefix 输入路径，并授予 Amazon ML 在输出前缀上设置放置对象 ACL 的权限。

```
{
    "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:*" }
        }
    }]
}
```

要将此策略应用到您的数据，您必须编辑与存储数据的 S3 存储桶关联的策略语句。

为 S3 存储桶编辑权限策略（使用旧控制台）

1. 登录到 AWS Management Console，然后通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 选择数据所在的存储桶的名称。
3. 请选择属性。
4. 选择编辑存储桶策略。
5. 如上所示输入策略，进行自定义以符合您的需求，然后选择保存。
6. 选择保存。

为 S3 存储桶编辑权限策略（使用新控制台）

1. 登录到 AWS Management Console，然后通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 选择存储桶名称，然后选择权限。
3. 请选择桶策略。
4. 如上所示输入策略，进行自定义以符合您的需求。
5. 选择保存。

使用 IAM 控制对 Amazon ML 资源的访问

AWS Identity and Access Management (IAM) 使您能够安全地控制用户对 AWS 服务和资源的访问权限。使用 IAM 可以创建和管理 AWS 用户、组和角色，您还可以使用权限来允许或拒绝对 AWS 资源的访问。借助 IAM 和 Amazon Machine Learning (Amazon ML)，可以控制组织中的用户是否可以使用特定 AWS 资源以及他们是否可以使用特定 Amazon ML API 操作来执行任务。

IAM 让您能够：

- 在您的 AWS 账户下创建用户和组。
- 为您的 AWS 账户下的每个用户分配唯一的安全凭证
- 控制每个用户使用 AWS 资源执行任务的权限
- 轻松地在您 AWS 账户中与用户共享您的 AWS 资源
- 创建 AWS 账户角色并管理其权限，以定义可以代入这些角色的用户或服务
- 在 IAM 中创建角色和管理权限，控制代入该角色的实体或 AWS 服务可执行的操作。您也可以定义由哪个实体承担该角色。

如果您的组织已有 IAM 身份，您可以使用它们来授予使用 AWS 资源执行任务的权限。

有关 IAM 的更多信息，请参阅 IAM 用户指南。

IAM 策略语法

IAM 策略是包含一个或多个语句的 JSON 文档。每个语句具有以下结构：

```
{
```

```
    "Statement":[{
        "Effect":"effect",
        "Action":"action",
        "Resource":"arn",
        "Condition":{
            "condition operator":{
                "key":"value"
            }
        }
    }]
}
```

策略语句包含以下元素：

- Effect：控制使用您在语句后面指定的资源和 API 操作的权限。有效值为 `Allow` 和 `Deny`。在默认情况下，IAM 用户没有使用资源和 API 操作的许可，因此，所有请求均会被拒绝。显式 `Allow` 将覆盖默认值。显式 `Deny` 将覆盖任意 `Allows`。
- Action：您对其授予或拒绝权限的特定 API 操作。
- Resource：受操作影响的资源。要在语句中指定资源，您可使用其 Amazon 资源名称 (ARN)。
- 条件（可选）：控制您的策略何时生效。

为简化 IAM 策略的创建和管理，您可以使用 AWS 策略生成器和 IAM 策略模拟器。

为 Amazon ML 指定 IAM 策略操作

在 IAM 策略语句中，您可以为支持 IAM 的任何服务指定 API 操作。在您为 Amazon ML API 操作创建策略语句时，请附加 `machinelearning:` 到 API 操作的名称，如下例中所示：

- `machinelearning:CreateDataSourceFromS3`
- `machinelearning:DescribeDataSources`
- `machinelearning:DeleteDataSource`
- `machinelearning:GetDataSource`

要在单个语句中指定多项操作，请使用逗号将它们隔开：

```
"Action": ["machinelearning:action1", "machinelearning:action2"]
```

您也可以使用通配符指定多项操作。例如，您可以指定名称以单词“Get”开头的所有操作：

```
"Action": "machinelearning:Get*"
```

要指定所有 Amazon ML 操作，请使用 * 通配符：

```
"Action": "machinelearning:*"
```

有关 Amazon ML API 操作的完整列表，请参阅 Amazon Machine Learning API 参考。

在 IAM 策略中指定 Amazon ML 资源的 ARN

IAM 策略语句应用到一个或多个资源。您可以按 ARN 为策略指定资源。

要为 Amazon ML 资源指定 ARN，请使用以下格式：

"资源": `arn:aws:machinelearning:region:account:resource-type/identifier`

以下示例显示如何指定通用 ARN。

数据源 ID：`my-s3-datasource-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:datasource/my-s3-datasource-id
```

ML 模型 ID：`my-ml-model-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/my-ml-model-id
```

批量预测 ID：`my-batchprediction-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/my-batchprediction-
id
```

评估 ID：`my-evaluation-id`

```
"Resource": arn:aws:machinelearning:<region>:<your-account-id>:evaluation/my-
evaluation-id
```

Amazon ML 的策略示例

示例 1：允许用户读取机器学习资源元数据

以下策略允许用户或组通过在指定资源上执行 DescribeDataSources、DescribeMLModels、DescribeBatchPredictions、DescribeEvaluations、GetDataSou 和 GetEvaluation 操作来读取数据源、ML 模型、批量预测和评估的元数据。Describe * 操作权限无法限定为特定资源。

```
{
    "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": [
            "*"
        ]
    }]
}
```

示例 2：允许用户创建机器学习资源

以下策略允许用户或组通过 CreateDataSourceFromS3、CreateDataSourceFromRedshift、CreateDataSourceFromRDS、Cr 和 CreateEvaluation 操作，创建机器学习数据源、ML 模型、批量预测和评估。您不能将这些操作的权限限制为特定资源。

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:CreateDataSourceFrom*",
            "machinelearning:CreateMLModel",
            "machinelearning:CreateBatchPrediction",
            "machinelearning:CreateEvaluation"
        ],
        "Resource": [
            "*"
        ]
    }]
}
```

示例 3：允许用户创建和删除实时终端节点以及对 ML 模型执行实时预测

以下策略允许用户或组对该模型执行 CreateRealtimeEndpoint、DeleteRealtimeEndpoint 和 Predict 操作，以创建和删除实时终端节点以及为特定 ML 模型执行实时预测。

```
{
    "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"
        ]
    }]
}
```

示例 4：允许用户更新和删除特定资源

以下策略通过向用户或组授予执行 UpdateDataSource、UpdateMLModel、UpdateBatchPrediction、UpdateEvaluation、DeleteDa 和 DeleteEvaluation 操作的权限，允许用户或组在您的 AWS 账户中更新和删除特定资源。

```
{
    "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"
        ]
    }]
}
```

示例 5：允许任意 Amazon ML 操作

以下策略允许用户或组使用任意 Amazon ML 操作。由于此策略会授予对您的机器学习资源的全部访问权限，您应该将其限制为仅对管理员可用。

```
{
    "Version": "2012-10-17",
    "Statement": [{
        "Effect": "Allow",
        "Action": [
            "machinelearning:*"
        ],
        "Resource": [
            "*"
        ]
```

```
    }]
}
```

跨服务混淆代理问题防范

混淆代理问题是一个安全性问题，即不具有操作执行权限的实体可能会迫使具有更高权限的实体执行该操作。在 AWS 中，跨服务模拟可能会导致混淆代理问题。一个服务（呼叫服务）调用另一项服务（所谓的服务）时，可能会发生跨服务模拟。可以操纵调用服务以使用其权限对另一个客户的资源进行操作，否则该服务不应有访问权限。为了防止这种情况，AWS 提供可帮助您保护所有服务的服务委托人数据的工具，这些服务委托人有权限访问账户中的资源。

我们建议在资源策略中使用 aws:SourceArn 或 aws:SourceAccount 全局条件上下文键，以限制 Amazon Machine Learning 为其他服务提供的资源访问权限。如果 aws:SourceArn 值不包含账户 ID，例如 Amazon S3 存储桶 ARN，您必须使用两个全局条件上下文密钥来限制权限。如果同时使用全局条件上下文密钥和包含账户 ID 的 aws:SourceArn 值，则 aws:SourceAccount 值和 aws:SourceArn 值中的账户在同一策略语句中使用时，必须使用相同的账户 ID。如果您只希望将一个资源与跨服务访问相关联，请使用 aws:SourceArn。如果您想允许该账户中的任何资源与跨服务使用操作相关联，请使用 aws:SourceAccount。

防范混淆代理问题最有效的方法是使用 aws:SourceArn 全局条件上下文键和资源的完整 ARN。如果不知道资源的完整 ARN，或者正在指定多个资源，请针对 ARN 未知部分使用带有通配符 (*) 的 aws:SourceArn 全局上下文条件键。例如，arn:aws:*servicename*:*:*123456789012*:*。

以下示例演示从 Amazon S3 存储桶读取数据时如何使用 Amazon ML 中的 aws:SourceArn 和 aws:SourceAccount 全局条件上下文键来防范混淆代理问题。

```
{
    "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:*" }
        }
    }]
}
```

异步操作的依赖项管理

Amazon ML 中的批量操作需要依靠其他操作才能成功完成。为了管理这些依赖项，Amazon ML 会识别具有依赖项的请求，并确保这些操作已经完成。如果这些操作未完成，Amazon ML 将搁置初始请求，直到这些操作依赖的操作已完成。

批量操作之间存在一些依赖关系。例如，您必须先创建可以训练 ML 模型的数据源，然后才能创建 ML 模型。如果没有可用的数据源，Amazon ML 无法训练 ML 模型。

但是，Amazon ML 支持异步操作的依赖项管理。例如，您不必等到数据统计计算完成之后，即可发送对数据源训练 ML 模型的请求。相反，只要创建了数据源，您就可以发送使用数据源训练 ML 模型的请求。Amazon ML 会等到数据源统计计算完成之后，才实际开始执行训练操作。createMLModel 请求将排入队列，直到统计计算完成；一旦计算完成，Amazon ML 会立即尝试执行 createMLModel 操作。同样，您也可以为尚未完成训练的 ML 模型发送批量预测和评估请求。

下表显示了继续执行其他 AmazonML 操作的要求

目的…	您必须具有……
创建 ML 模型 (createMLModel)	包含已计算数据统计的数据源
创建批量预测 (createBatchPrediction)	数据源 ML 模型

目的…	您必须具有……
创建批量评估 (createBatchEvaluation)	数据源 ML 模型

检查请求状态

提交请求后，您可以使用 Amazon Machine Learning (Amazon ML) API 检查其状态。例如，如果您提交 `createMLModel` 请求，可使用 `describeMLModel` 调用检查其状态。Amazon ML 将使用以下状态之一进行响应。

状态	定义
PENDING	Amazon ML 正在验证请求。 或 Amazon ML 正在等待计算资源变为可用，然后再运行请求。当您的账户已经超过并行运行的批量操作请求的最大数量时，可能会出现这种情况。如果出现这种情况，当其他正在运行的请求完成或取消时，状态会转换为 InProgress。 或 Amazon ML 正在等待批量操作，您的请求取决于此操作的完成状态。
INPROGRESS	您的请求仍在运行。
COMPLETED	请求已完成，您可以使用（ML 模型和数据源）或查看（批量预测和评估）对象。
FAILED	您提供的数据存在问题，或者您已经取消该操作。例如，如果您尝试为未能完成的数据源计算数据统计，可能会收到 Invalid 或 Failed 状态消息。错误消息解释了操作未成功完成的原因。
DELETED	对象已删除。

Amazon ML 还提供了有关对象的信息，例如，当 Amazon ML 完成对象的创建时。有关更多信息，请参阅列出对象。

系统限制

为了提供强大、可靠的服务，Amazon ML 为您对系统提出的请求实施了一些限制。大多数 ML 问题都轻松适应这些限制条件。但是，如果您在使用 Amazon ML 时发现这些限制条件让您受限，您可以联系 AWS 客户服务，请求提高限制上限。例如，您可以同时运行的任务数量可能限制为五个。如果您发现这个限制导致经常有任务需要排队等待资源，则表明为您的账户提高限制上限很可能非常有意义。

下表显示了 Amazon ML 中每个账户的默认限制。并非所有这些限制都能由 AWS 客户服务提高限制上限。

限制类型	系统限制
每个观察数据的大小	100 KB
训练数据的大小 *	100 GB
批量预测输入的大小	1TB
批量预测输入的大小（记录数）	100 百万
数据文件（架构）中的变量数	1000
配方复杂性（处理的输出变量数）	10000
各个实时预测终端节点的 TPS	200
所有实时预测终端节点的 TPS 总和	10000
所有实时预测终端节点的 RAM 总和	10GB
同时运行的作业数	25
任何作业的最长运行时间	7 天
多类别 ML 模型的类别数	100
ML 模型大小	最小为 1 MB，最大为 2 GB

限制类型	系统限制
每个对象的标签数量	50

- 您的数据文件大小受限，以确保作业能够及时完成。已运行七天以上的作业将自动终止，导致 FAILED 状态。

所有对象的名称和 ID

Amazon ML 中的每个对象都必须有一个标识符，也就是 ID。Amazon ML 控制台会为您生成 ID 值，但如果您使用的是 API，您必须自行生成。在您的 AWS 账户中，在类型相同的所有 Amazon ML 对象中，每个 ID 都必须唯一。即，两个评估不能有同一 ID。评估和数据源可以有相同 ID，但不建议这样做。

我们建议您为对象使用随机生成的标识符，使用短字符串作为前缀来标识其类型。例如，当 Amazon ML 控制台生成数据源时，它会向数据源随机分配唯一 ID，例如“ds-zScWIuWiOxF”。此 ID 随机度足够高，可以避免任何单个用户发生冲突，并且也紧凑易读。为方便起见，使用“ds-”前缀进行明确，但该前缀并不是必需的。如果您不确定使用什么 ID 字符串，我们建议您使用十六进制的 UUID 值（类似于 28b1e915-57e5-4e6c-a7bd-6fb4e729cb23），这在任何现代编程环境中都可以直接获取。

ID 字符串可以包含 ASCII 字母、数字、连字符和下划线，最长可达 64 个字符。可以将元数据编码为 ID 字符串，这种做法可能会比较方便。但不建议这样做，因为对象在创建之后，其 ID 不能更改。

对象名称为您提供了一种简单的方式，将用户友好的元数据与各个对象关联。您可在创建对象之后更新名称。这使得对象名称可以反映您的 ML 工作流的一些方面。例如，您可能最初将 ML 模型命名为“experiment #3”，以后会将模型命名为“final production model”。名称可以是任何字符串，但不得超过 1024 个字符。

对象生命周期

您使用 Amazon ML 创建的任何数据源、ML 模型、评估或批量预测对象，在创建之后至少两年内可供您使用。Amazon ML 可能会自动删除超过两年未访问或未使用的对象。

资源

下列相关资源在您使用此服务的过程中会有所帮助。

- Amazon ML 产品信息 – 在一个中心位置捕获所有 Amazon ML 相关的产品信息。
- Amazon ML 常见问题 – 涵盖了开发人员对此产品提出的一些最热门的问题。
- Amazon ML 示例代码 – 使用亚马逊 ML 的示例应用程序。您开始时可以使用示例代码创建自己的 ML 应用程序。
- Amazon ML API 参考 – 详细介绍 Amazon ML 的所有 API 操作。它还提供了所支持 Web 服务协议的示例请求和响应。
- AWS 开发人员资源中心 - 提供中央起点，用于查找相关的文档、代码示例、发布说明和其他信息，以帮助您通过 AWS 构建创新型应用程序。
- AWS 培训和课程 – 指向基于角色的专业课程和自主进度动手实验室的链接，这些课程和实验室旨在帮助您增强 AWS 技能并获得实践经验。
- AWS 开发人员工具 – 指向开发人员工具和资源的链接，其中提供了文档、代码示例、发行说明和有助于您利用 AWS 构建创新应用程序的其他信息。
- AWS Support 中心 – 用于创建和管理 AWS Support 案例的中心。还包括指向其他有用资源的链接，如论坛、技术常见问题、服务运行状况和 AWS Trusted Advisor。
- AWS Support – 提供有关 AWS Support 信息的主要网页，是一种一对一的快速响应支持渠道，可帮助您在云中构建和运行应用程序。
- 联系我们 – 用于查询有关 AWS 账单、您的账户、事件、滥用和其他问题的中央联系点。
- AWS 网站条款 – 有关我们的版权和商标、您的账户、许可和网站访问以及其他主题的详细信息。

文档历史记录

下表介绍了对此版 Amazon Machine Learning (Amazon ML) 的文档所做的重要更改。

- API 版本：2015-04-09
- 文档上次更新时间：2016-08-02

更改	描述	更改日期
添加了指标	此版 Amazon ML 为 Amazon ML 对象添加了新指标。 有关更多信息，请参阅列出对象。	2016 年 8 月 2 日
删除多个对象	此版 Amazon ML 添加了删除多个 Amazon ML 对象的功能。 有关更多信息，请参阅删除对象。	2016 年 7 月 20 日
添加了标签	此版 Amazon ML 添加了为 Amazon ML 对象应用标签的功能。 有关更多信息，请参阅标记您的 Amazon ML 对象。	2016 年 6 月 23 日
复制 Amazon Redshift 数据源	此版 Amazon ML 添加了将 Amazon Redshift 数据源设置复制到新 Amazon Redshift 数据源的功能。 有关复制 Amazon Redshift 数据源设置的更多信息，请参阅复制数据源（控制台）。	2016 年 8 月 11 日
添加了随机排序	此版 Amazon ML 添加了为输入数据随机排序的功能。 有关使用将类型随机排序参数的更多信息，请参阅将训练数据的类型随机排序。	2016 年 8 月 5 日
改进了使用 Amazon Redshift 创建数据源	此版 Amazon ML 添加了在控制台中创建 Amazon ML 数据源后测试 Amazon Redshift 设置以验证连接是否有效的功能。有关更多信息，请参阅利用 Amazon Redshift 数据创建数据源（控制台）。	2016 年 3 月 21 日
改进了 Amazon	此版 Amazon ML 改进了 Amazon Redshift 数据架构向 Amazon ML 数据架构的转换。	2016 年 2 月 9 日

更改	描述	更改日期
Redshift 数据架构转换	有关将 Amazon Redshift 与 Amazon ML 结合使用的更多信息，请参阅根据 Amazon Redshift 中的数据创建 Amazon ML 数据源。	
添加了 CloudTrail 日志	此版 Amazon ML 添加了使用 AWS CloudTrail (CloudTrail) 记录请求的功能。 有关使用 CloudTrail 记录的更多信息，请参阅使用 AWS CloudTrail 记录 Amazon ML API 调用。	2015 年 12 月 10 日
添加了其他的 DataRearrangement 选项	此版 Amazon ML 添加了随机拆分输入数据和创建补充数据源的功能。 有关使用 `DataRearrangement` 参数的更多信息，请参阅数据重新排列。有关如何使用新选项进行交叉验证的信息，请参阅交叉验证。	2015 年 12 月 3 日
试用实时预测	此版 Amazon ML 添加了在服务控制台上试用实时预测的功能。 有关试用实时预测的更多信息，请参阅 Amazon Machine Learning 开发人员指南中的请求实时预测。	2015 年 11 月 19 日
新地区	此版 Amazon ML 添加了对欧洲（爱尔兰）地区的支持。 有关欧洲（爱尔兰）地区 Amazon ML 的更多信息，请参阅 Amazon Machine Learning 开发人员指南中的区域和终端节点。	2015 年 8 月 20 日
首次发布	本指南是 Amazon ML 开发人员指南的第一个版本。	2015 年 8 月 9 日