AWS 白皮书

使用 Amazon Forecast 的时间序列预测原理

使用 Amazon Forecast 的时间序列预测原理: AWS 白皮书

Table of Contents

使用 Amazon Forecast 的时间序列预测原理

发布日期：2021 年 9 月 1 日（文档历史记录）

现今，各公司使用各种工具从简单电子表格到复杂财务规划软件，试图准确预测未来的业务成果，如产品需求、资源需求和财务绩效。本白皮书介绍了预测、其术语、挑战和使用案例。本文档使用案例研究来强化预测概念、预测步骤，并提及 Amazon Forecast 如何帮助解决现实世界预测问题中的许多实际挑战。

概览

预测是指预测未来的科学技术。通过使用历史数据，企业可以了解趋势，确定可能发生的事情和时间，进而将这些信息纳入从产品需求到库存计划和人员配置等所有方面的未来计划。

考虑到预测的后果，精度很重要。如果预测值过高，客户可能会在产品和员工方面过度投资，导致投资浪费。如果预测值过低，客户可能会投资不足，导致原材料和库存短缺，从而造成糟糕的客户体验。

如今，企业试图使用从简单的电子表格到复杂的需求/财务规划软件的所有内容来生成预测，但由于两个原因，仍然无法实现高精度：

- 首先，传统预测很难纳入大量的历史数据，因而会错过在干扰中丢失的过往重要信号。
- 其次，传统预测很少包含相关但独立的数据，这些数据可以提供重要的背景信息（例如价格、假日/活动、缺货、营销促销等）。如果没有完整的历史数据和更广泛的背景，大多数预测都无法准确预测未来。

Amazon Forecast 是一项完全托管的服务，可以克服这些问题。Amazon Forecast 为当前的预测场景提供了最佳算法。它在适当的时候依靠现代化的机器学习（ML）和深度学习，提供高度准确的预测。Amazon Forecast 易于使用，不需要机器学习经验。该服务自动提供必要的基础设施，处理数据，并构建在 AWS 上托管并随时可以进行预测的自定义/私有机器学习模型。此外，随着机器学习技术的不断快速发展，Amazon Forecast 将其纳入其中，因此客户只需付出最少的额外努力，甚至无需付出任何额外努力，即可继续看到精度方面的改进。

您的架构是否完善？

当您在云中构建系统时，AWS Well-Architected Framework 可以帮助您了解所做决策的利弊。框架的六大支柱使您能够学习设计和操作可靠、安全、高效、经济实惠且可持续的系统的架构最佳实践。使用

AWS Management Console中免费提供的 AWS Well-Architected Tool，您可以通过回答面向每个支柱的一组问题，根据这些最佳实践来检查您的工作负载。

在本机器学习剖析中，我们重点介绍了如何在 AWS Cloud 中对机器学习工作负载进行设计、部署和架构设计。我们会将本剖析添加到 Well-Architected Framework 中描述的最佳实践。

有关云架构的更多专家指导和最佳实践（参考架构部署、图表和白皮书），请参阅 AWS Architecture Center。

关于预测

在本文档中，预测意味着预测时间序列的未来值：对一个问题的输入或输出具有时间序列的性质。

预测系统

预测系统包括一组不同的用户：

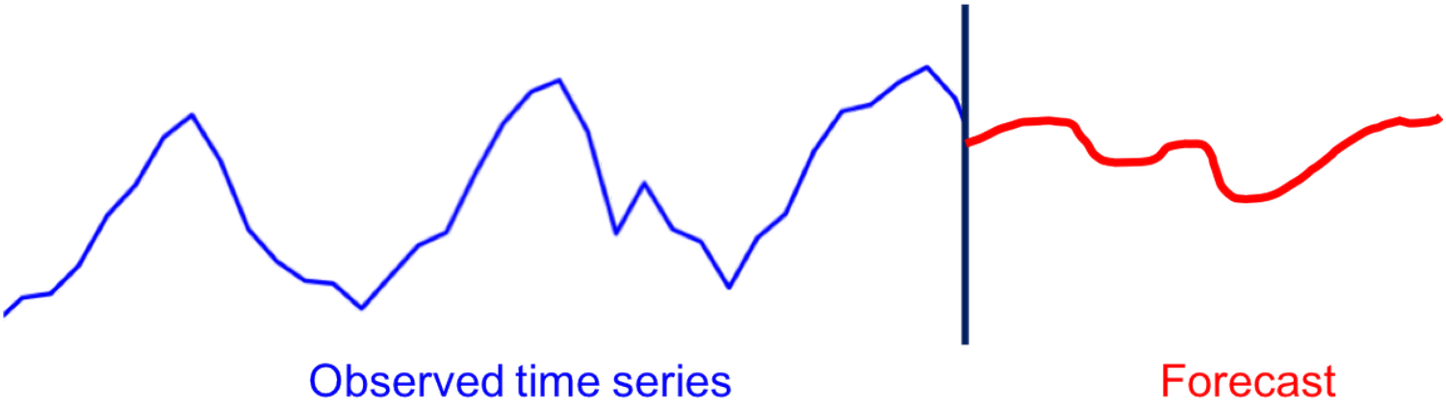
- 最终用户，他们查询特定产品的预测，并决定购买多少单位；这可能是一个人，也可能是自动化系统。
- 业务分析师/商业智能，他们负责为最终用户提供支持，运行和整理汇总报告。
- 数据科学家，他们以迭代方式分析需求模式和因果关系，并添加新功能，以对模型进行渐进式改进或改进预测模型。
- 工程师，他们负责建立数据收集的基础设施，并确保系统输入数据的可用性。

Amazon Forecast 减轻了软件工程师的工作，让数据科学能力有限的企业能够利用最先进的预测技术。具有数据科学能力的企业可以利用我们提供的诸多诊断功能，通过 Amazon Forecast 很好地解决预测问题。

预测问题发生在哪里？

预测问题发生在许多自然产生时间序列数据的领域。其中包括零售、医疗分析、容量规划、传感器网络监测、财务分析、社会活动挖掘和数据库系统。例如，在大多数支持数据驱动型决策的企业中，预测在自动化和优化运营流程方面发挥着关键作用。产品供求预测可用于优化库存管理、人员调度和拓扑规划，而且通常是供应链优化大多数环节的关键技术。

下图概述了一个预测问题，该问题基于呈现模式（在本例中为季节性）的观测时间序列，并在指定时间段内创建了预测。横轴表示时间，从过去（左）到未来（右）。纵轴表示测量单位。根据过去（蓝色）到垂直的黑线，确定未来（红色）是预测任务。



预测任务概述

尝试解决预测问题之前的注意事项

在解决预测问题之前，您需要了解几个最重要的问题，包括：

- 您需要解决预测问题吗？
- 您为什么要解决预测问题？

由于时间序列数据无处不在，很容易在任何地方发现预测问题。然而，这其中存在着一个关键问题，就是是否真的需要解决预测问题，或者说，是否可以在不牺牲业务有效决策的情况下完全规避这个问题。提出这个问题很重要，因为从科学上讲，预测是机器学习中最难的问题之一。

例如，思考一下在线零售商的产品推荐。此产品推荐问题可以描述为预测问题，即对于每对客户库存单位（SKU, stock keeping unit），您预测该特定客户将购买的特定商品的单位数量。这种问题表述有很多好处。一个好处是明确考虑了时间因素，因此您可以根据客户的购买模式推荐产品。

但是，产品推荐问题很少被描述为预测问题，因为解决这样的预测问题（例如，在客户 SKU 层面上信息的稀疏程度和问题的规模）比直接解决推荐问题困难得多。因此，在考虑预测应用程序时，重要的是要考虑预测的下游使用情况，以及是否有可能使用替代方法来解决这个问题。

在这些情况下，Amazon Personalize 可以提供帮助。Amazon Personalize 是一项机器学习服务，可让开发人员轻松使用其应用程序为客户提供个性化推荐。

在确定需要解决预测问题后，接下来要问的问题是，为什么要解决预测问题？ 在许多商业环境中，预测通常只是达到目的的一种手段。例如，对于零售环境中的需求预测，预测可用于制定库存管理决策。预测问题通常是决策问题的输入，而决策问题反过来又可以建模为优化问题。

此类决策问题的示例包括，要购买的单位数量或处理现有库存的最佳方法。其他业务预测问题包括，预测服务器容量或预测制造环境中对原材料/零件的需求。这些预测可用作其他过程的输入，要么用于上述决策问题，要么用于情景模拟，然后在没有明确模型的情况下用于规划。预测本身不是目的，但也有例外。例如，在财务预测中，预测直接用于积累金融储备或呈现给投资者。

要了解预测的目的，请考虑以下问题：

- 您应该预测未来多长时间？
- 您需要多久生成一次预测？
- 预测中是否有您应该深入研究的具体方面？

案例研究：电子商务企业的零售需求预测问题

为了更详细地说明预测概念，请以在线销售产品的电子商务企业为例。优化供应链中的决策（例如，库存管理）对于这种企业的核心竞争力至关重要，因为这有助于在适当的配送地点获得准确的产品数量。从本质上讲，这意味着有更多的选择，更短的配送时间和有竞争力的价格，从而提高客户满意度。供应链软件系统的关键输入是对需求的预测或对目录中每种产品的潜在销售额的预测。这一预测有助于做出重要的下游决策，其中关键包括：

- 宏观层面的规划（战略预测）：对于整个企业而言，总销售额/收入的预计增长是多少？ 在地理位置上，企业应该在哪里（更）活跃？ 应如何配备劳动力？
- 需求（或库存）预测：每个地点每种产品的预计销量是多少？
- 促销活动（战术预测）：促销活动应如何进行？ 产品是否应清算？

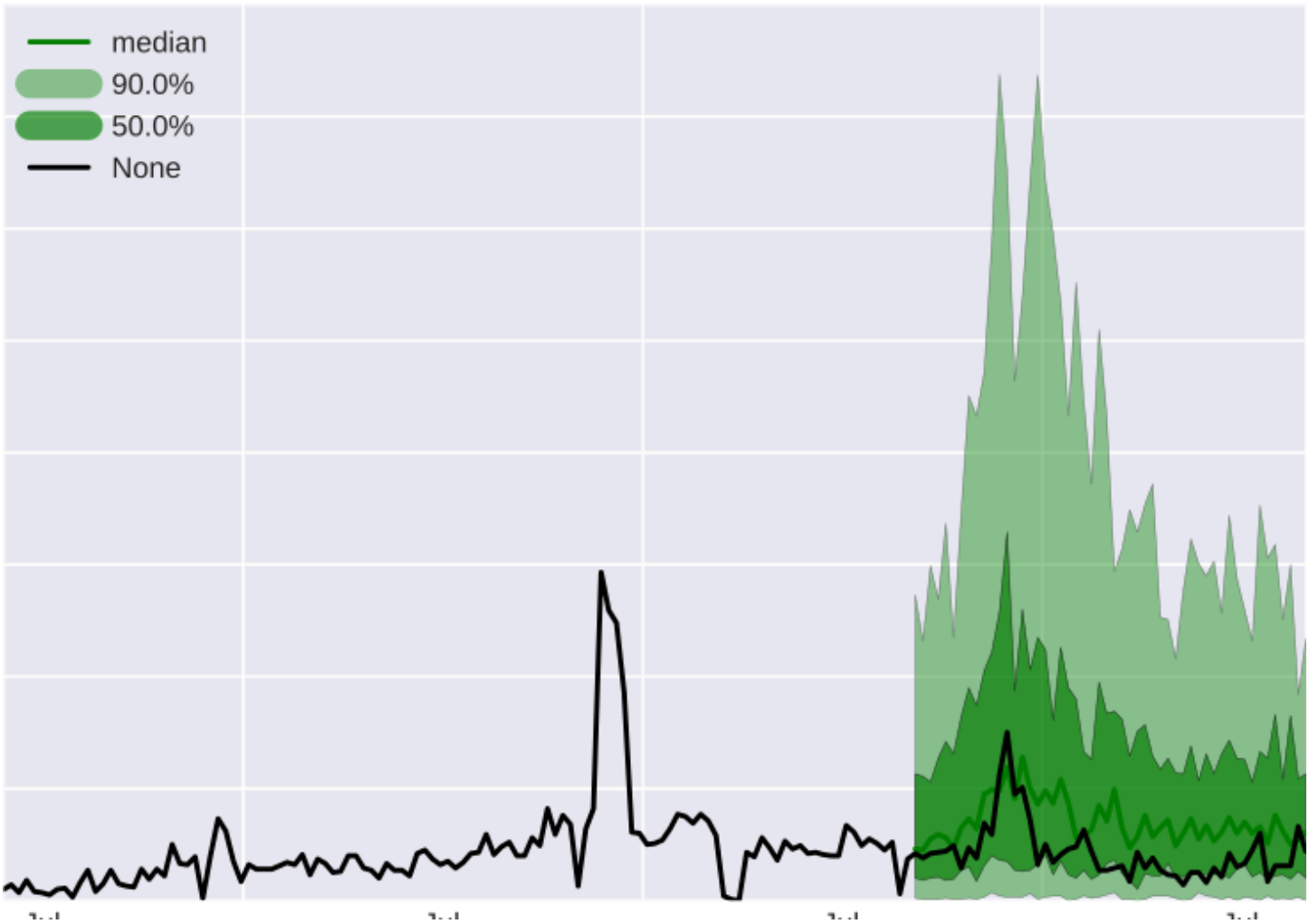
案例研究的其余部分侧重于第二个问题，这是运营预测问题系列的一部分（Januschowski & Kolassa，2019）。本文档关注的主要方面：数据、模型（预测器）、推论（预测）和产品化。

在本案例研究中，重要的是要记住，预测问题是达到目的的一种手段。尽管预测对企业至关重要，但下游供应链决策更为重要。在我们的案例研究中，这些决策是由依赖运筹学数学优化模型的自动购买系统做出的。这些系统试图将企业的预期成本降至最低。

关键词是预期，这意味着预测不仅应涵盖一种可能的未来，还应涵盖所有可能的未来，并根据特定结果的概率适当加权。为此，下游决策的关键推动因素是预测值的完整分布，而不仅仅是一个点预测。下图显示了概率预测（也称为密度预测）。请注意，您可以轻松地从这种概率预测中得出单点预测（最有可能的未来），但是从点预测得出概率预测要困难得多。

给定一个概率预测，您可以从中获得不同的统计数据，并对结果进行调整，以帮助您做出想要的决策。电子商务企业可能有许多关键产品，他们几乎不希望这些产品断货。在这种情况下，使用高分位数（例如，^{第} 90 个百分位数），这意味着 90% 的时间内，产品都是有现货的。对于其他产品，例如更容易找到替代品的产品（如铅笔），使用较低的百分位数可能更合适。

在 Amazon Forecast 中，您可以轻松地从概率预测中获得不同的分位数。



概率预测图解

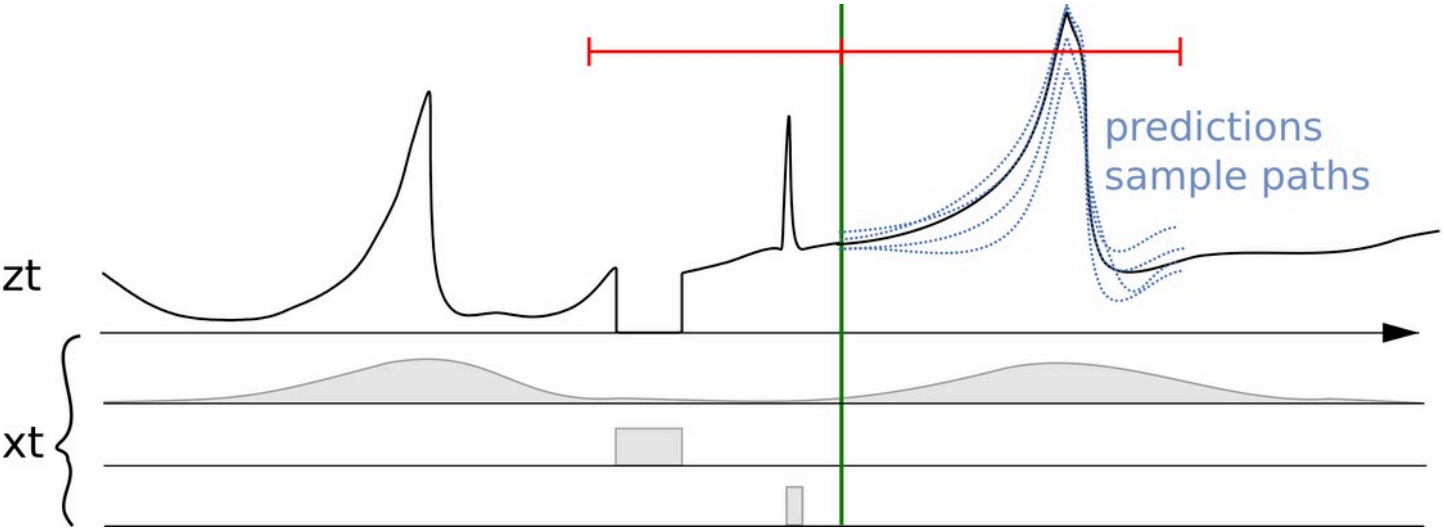
在上图中，黑线是实际值；深绿色线是预测分布的中值；深绿色阴影区域是您预计 50% 的值将落入的预测区间；浅绿色区域是您预计 90% 的实际值将落入的预测区间。

以下各节涵盖了解决这种企业的预测问题所涉及的步骤，包括：

- 数据收集和聚合（第 1 步）
- 数据准备（第 2 步）
- 创建预测器（第 3 步）
- 评估预测器（第 4 步）
- 自动生成预测（第 5 步）

第 1 步：收集和聚合数据

下图显示了预测问题的心理模型。目标是预测未来的时间序列 z_t，使用尽可能多的相关信息使预测尽可能准确。因此，第一步也是最重要的一步是收集尽可能多的正确数据。



时间序列 z_t 以及相关特征或协变量（x_t）和多个预测

在上图中，垂直线的右侧显示了多个预测。这些预测是概率预测分布的样本（或者相反，可以用来表示概率预测）。

零售企业需要记录的关键信息是：

- 交易销售数据 – 例如，库存单位（SKU, stock keeping unit）、地点、时间戳和销售单位。
- SKU 商品详情数据 – 商品的元数据。示例包括颜色、部门、尺寸等。
- 价格数据 – 带有时间戳的每件商品的价格时间序列。
- 促销信息数据 – 不同类型的促销，要么针对一组商品（类别），要么针对带有时间戳的个别商品。
- 库存信息数据 – 在每个时间单位内，SKU 是否有货或可购买与 SKU 是否缺货的信息。
- 位置数据 – 商品或销售在给定时间点的位置可以表示为字符串 `location_id` 或 `store_id` 或实际地理位置。地理位置可以是国家代码加上五位数的邮政编码或 `latitude_longitude` 坐标。位置被视为交易销售的一个“维度”。

在 Amazon Forecast 中，要预测的数量的历史数据称为目标时间序列（TTS, Target Time Series）。对于零售企业而言，TTS 是交易销售数据。与每笔销售交易完全同时已知的其他历史数据称为相关时间序列（RTS, Related Time Series）。对于零售企业而言，RTS 将包括价格、促销和库存变量。

请注意，库存信息很重要，因为这个问题关注的是预测需求，而不是销售，但企业仅记录销售额。当 SKU 缺货时，销售数量低于潜在需求，因此了解和记录此类缺货事件何时发生非常重要。

其他需要考虑的数据集包括网页访问量、搜索词详情、社交媒体和天气信息。拥有过去和未来的数据通常很重要，这样才能在模型中使用这些数据。这是许多预测模型和回测中的一个要求（如第 4 步：评估预测部分所述）。

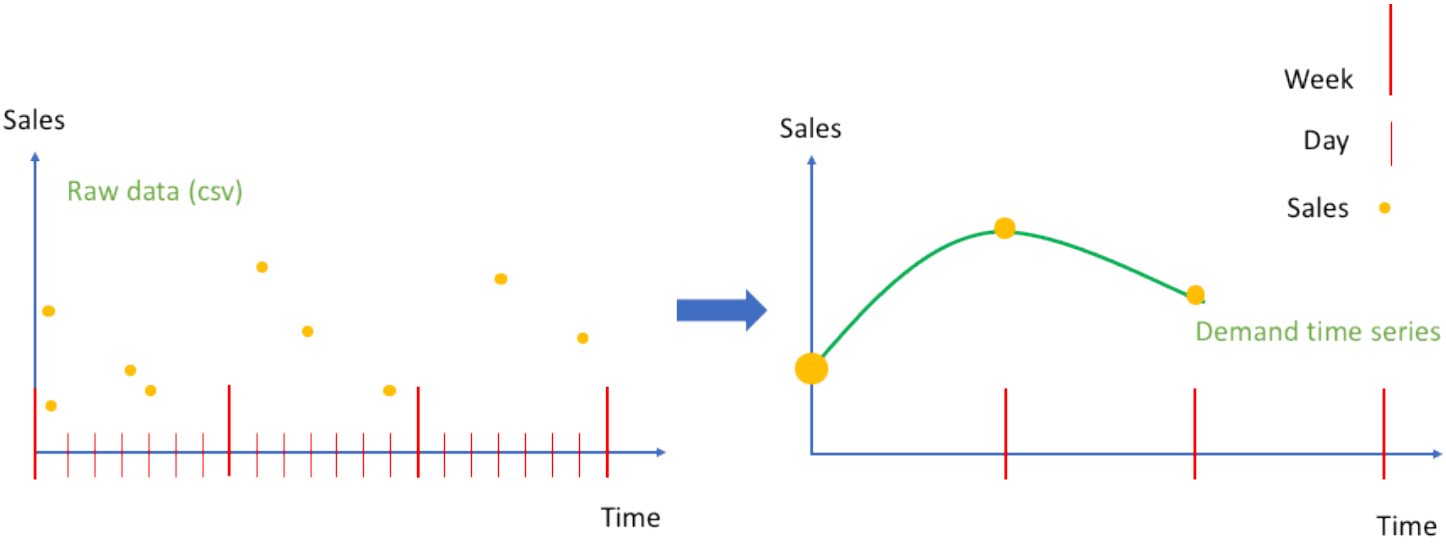
对于某些预测问题，原始数据的频率与预测问题的频率自然匹配。示例包括服务器容量的请求，该请求是按分钟采样的，此时您要按分钟频率进行预测。

数据通常以更精细的频率记录，或者只是在某个时间范围内的任意时间戳记录，但预测问题的粒度更粗。这种情况在零售案例研究中很常见，在零售案例研究中，销售数据通常记录为交易数据；例如，该格式由一个时间戳组成，其细粒度显示了销售发生的时间。在预测使用案例中，可能不需要这种低粒度，将这些数据聚合为每小时或每天的销售额可能更合适。在这里，聚合水平对应于下游问题；例如，库存管理或资源规划。

示例

在下图中，左图显示了可作为逗号分隔值（CSV, comma separated value）文件输入到 Amazon Forecast 的原始客户销售数据示例。在此示例中，销售数据是在更精细的每日时间网格上定义的，问题是预测未来更粗的时间网格上的每周需求。Amazon Forecast 在`create_predictor` API 调用中汇总给定一周的每日值。

结果是将原始数据转换为一组格式良好的时间序列，每周频率固定。右图使用默认的求和聚合方法说明了目标时间序列上的这种聚合。其他聚合方法包括求平均值、最大值、最小值或选择单个点（例如，第一个点）。必须选择聚合粒度和方法，使其与数据的业务使用案例最匹配。在此示例中，聚合值与每周聚合值保持一致。其他聚合方法可以由用户使用 `create_predictor` API 的 `FeaturizationConfig` 参数的 `FeaturizationMethodParameters` 键进行设置。



将原始销售数据作为事件（左）聚合到等间隔时间序列（右）

第 2 步：准备数据

获得原始数据后，您需要处理诸如数据缺失之类的复杂问题，确保为预测模型准备最能反映预期解释的数据。

如何处理缺失数据

在现实世界的预测问题中，常见的情况是原始数据中存在缺失值。时间序列中缺失值是指在指定频率下，每个时间点上对应的真实值无法用于进一步处理。将值标记为缺失的原因可能有多种。

缺失值可能是由于未进行交易或可能的测量错误（例如，因为监控某些数据的服务无法正常运行，或者无法正确进行测量）。在零售案例研究中，后者的主要示例是需求预测中的缺货情况，这意味着需求不等于当天的销售额。

在云计算场景中，当服务达到限制时（例如，某个 AWS 区域 中的 Amazon EC2 实例都很忙），也会出现类似的影响。另一个缺失值示例发生在产品或服务尚未推出或停产时。

缺失值也可以由特征处理组件插入，以确保时间序列的长度与填充长度相等。如果缺失值足够普遍，则会严重影响模型的准确性。

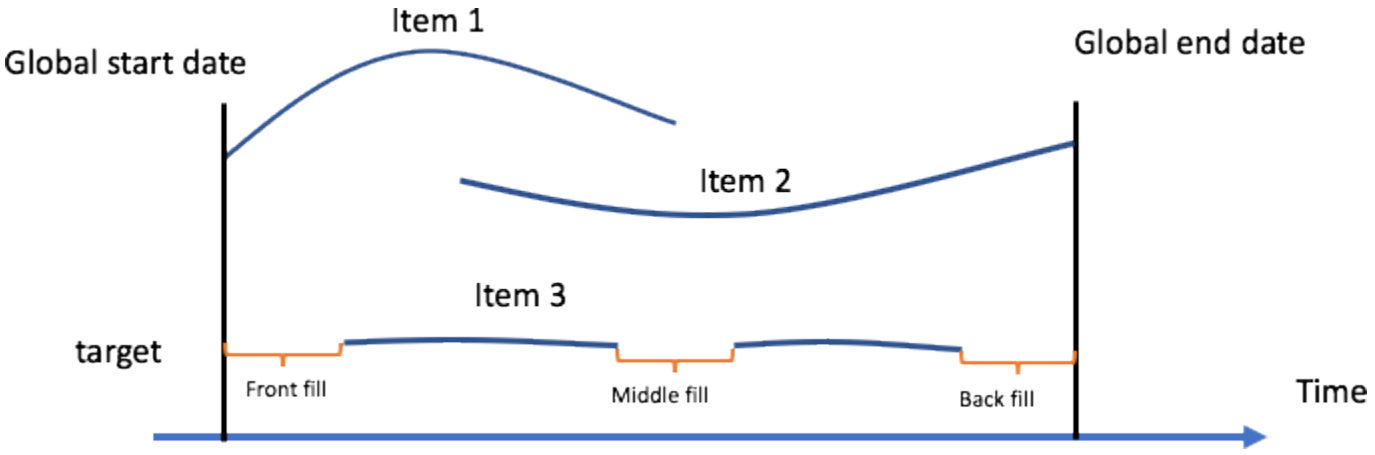
示例 1

填充是向数据集中缺少的条目添加标准化值的过程。在下图中，针对包含三个项目的数据集中的项目 2，说明了处理 Amazon Forecast 中缺失值的不同策略（前填充、中间填充、回填充和未来填充）。

Amazon Forecast 支持填充目标时间序列和相关时间序列。全局开始日期定义为数据集中所有项目的开始日期中的最早开始日期。在以下示例中，项目 1 的开始日期为全局开始日期。类似地，全局结束日期定义为所有项目的时间序列的最晚结束日期，该日期为项目 2 的结束日期。

前填充是指填充从特定时间序列开始到全局开始日期的每个值。在发布本文档时，Amazon Forecast 未开启任何前填充，允许所有时间序列从不同的时间点开始。中间填充表示在时间序列中间（例如，在项目的开始日期和结束日期之间）填充的值，回填充是指从该时间序列的最后日期到全局结束日期之间的填充。

对于目标时间序列，中间填充和回填充方法的默认填充逻辑为零。未来填充（仅适用于相关时间序列）是指填充项目的全局结束日期和客户指定的预测期间之间的任何缺失值。在 Prophet 和 DeepAR+ 中使用相关时间序列数据集时需要未来值，对于 CNN-QR 则是可选的。



Amazon Forecast 中的缺失值处理策略

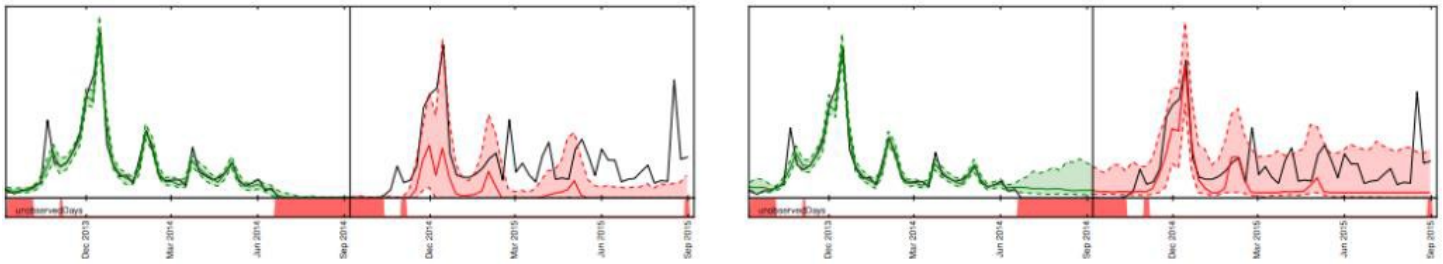
在上图中，全局开始日期表示所有项目的开始日期中的最早开始日期，全局结束日期表示所有项目的结束日期中的最晚结束日期。预测期间是指 Forecast 提供目标值预测的时段。

这是零售研究中常见的场景，表示可用项目的交易数据的销售额为零。这些值被视为真正的零，并在度量评估组件中使用。Amazon Forecast 使用户能够识别实际缺失的值，并将它们编码为非数字（NaN, not a number）以供算法处理。接下来，本白皮书将探讨这两种情况的不同之处，以及每种情况何时有用。

在零售案例研究中，零售商售出零件可用商品的信息不同于出售零件不可用商品的信息，无论是在商品存在以外的时期（例如，推出之前或弃用之后），还是在其存在期间内（例如，部分缺货或没有记录该时间范围内的销售数据）。默认的零填充适用于前一种情况。在后者中，即使相应的目标值通常为零，但标记为缺失的值还会传递额外的信息。最佳做法是保留缺失数据的信息，而不是丢弃这些信息。请参阅以下示例，了解为什么保留信息很重要。

Amazon Forecast 支持值、平均值、中值、最小值和最大值的其他填充逻辑。对于相关时间序列（例如，价格或促销），没有为中间填充、回填充或未来填充方法指定默认值，因为正确的缺失值逻辑因属性类型和使用案例而异。相关时间序列支持的填充逻辑包括零、值、平均值、中值、最小值和最大值。

要执行缺失值填充，请指定在调用 `CreatePredictor` 操作时要实施的填充类型。填充逻辑在 `FeaturizationMethod` 对象中指定。例如，要对不表示目标时间序列中不可用产品销量为零的值进行编码，请通过将填充类型设置为 NaN 将该值标记为真正缺失。与零填充不同，使用 NaN 编码的值被视为真正的缺失，并且不会在度量评估组件中使用



0 填充与 NaN 填充对同一项目预测的影响

在上图的左图中，垂直黑线左侧的值用 0 填充，从而导致预测偏差不足（在垂直黑线右侧）。在右图中，这些值被标记为 NaN，从而得出相应的预测。

示例 2

上图说明了正确处理线性状态空间模型（例如 ARIMA 或 ETS）缺失值的重要性。它绘制了对部分缺货商品的需求预测。训练区域在左图中以绿色显示，右侧面板中的预测范围以红色显示，真实目标以黑色显示。中位数、p10 和 p90 预测分别显示在红线和阴影区域中。底部显示用红色标记的缺货商品（占数据的 80%）。在左图中，缺货区域被忽略并用 0 填充。

这导致预测模型假设有很多零需要预测，因此预测值太低。在右图中，缺货区域被视为真正的缺失观测值，缺货区域的需求变得不确定。将缺货商品的缺失值适当标记为 NaN 后，您可以看到此图中的预测范围没有偏差。Amazon Forecast 填补了这些数据空白，使您可以轻松地正确处理缺失数据，而无需明确修改所有输入数据。

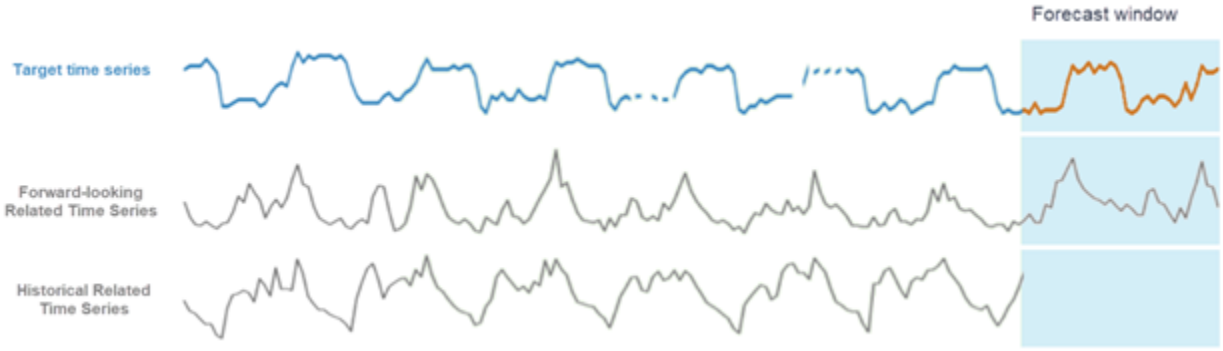
特征化和相关时间序列的概念

Amazon Forecast 使用户能够输入相关数据，以帮助提高某些支持的预测模型的准确性。这些数据可以有两种类型：相关时间序列或静态项目元数据。

> **Note**
>
> 元数据和相关数据在机器学习中被称为特征，在统计中被称为协变量。

相关时间序列是指与目标值有一定相关性的时间序列，它直观地解释了目标值，因此对目标值的预测具有一定的统计强度（有关示例，请参阅 Amazon Forecast：大规模预测时间序列）。与目标时间序列不同，相关时间序列是过去可能影响目标时间序列的已知值，并且可能在未来具有已知值。

在 Amazon Forecast 中，您可以添加两种类型的相关时间序列：历史时间序列和前瞻时间序列。历史相关时间序列包含预测期间之前的数据点，不包含未来预测期间内的任何数据点。前瞻相关时间序列包含预测期间之前和预测期间内的数据点。



在 Amazon Forecast 中使用相关时间序列的不同方法

示例 3

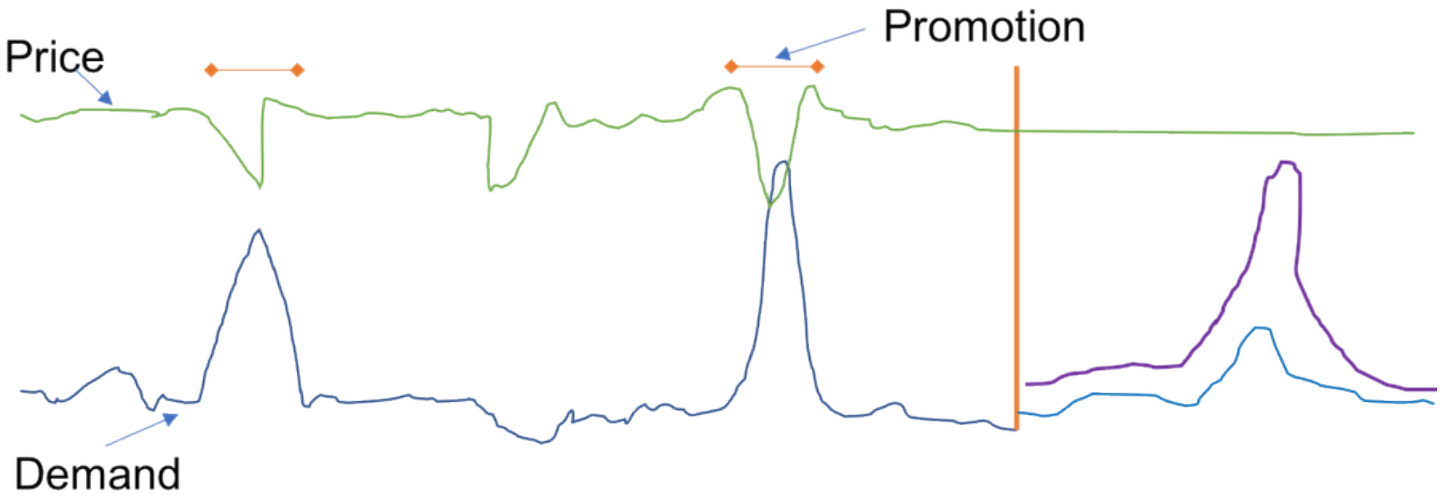
下图显示了如何使用相关时间序列来预测一本畅销书籍的未来需求的示例。蓝线代表目标时间序列中的需求。价格显示为绿线。垂直线表示预测开始日期，两个分位数的预测显示在垂直线的右侧。

此示例使用前瞻相关时间序列，该时间序列在预测粒度上与目标时间序列保持一致，并且在预测开始日期到预测开始日期加预测期间（预测结束日期）的范围内的未来所有（或大多数）时间已知。

下图还显示了价格是一个适合使用的特征，因为您可以看到价格下降与产品销量增加之间的相关性。相关时间序列可以通过单独的 CSV 文件提供给 Amazon Forecast，其中包含商品 SKU、时间戳和相关时间序列值（在本例中为价格）。

Amazon Forecast 支持目标时间序列的平均和求和等聚合方法，但不支持相关时间序列。例如，将每日价格与每周价格相加是没有意义的，对于每日促销也是如此。

Amazon Forecast 可以通过包含内置的特征化数据集，自动将天气和节假日信息合并到模型中（参见 `SupplementaryFeature`）。天气信息和节假日会显著影响零售需求。



特定商品的销售额（蓝色表示，垂直红线左侧）

项目元数据（也称为分类变量）是可以输入到 Amazon Forecast 的其他有用特征（有关示例，请参阅 Amazon Forecast：大规模预测时间序列）。分类变量和相关时间序列之间的主要区别在于，分类变量是静态的，它们不会随时间而变化。常见的零售示例包括商品颜色、图书类别，以及电视是否是智能电视的二进制指标。假设相似的 SKU 具有相似的销售额，则可以通过深度学习算法在了解库存单位（SKU, stock-keeping units）之间的相似之处时获取这些信息。由于此元数据不具有时间依赖性，因此商品元数据 CSV 文件中的每一行仅包含商品 SKU 和相应的类别标签或描述。

第 3 步：创建预测器

可以通过两种方式创建预测器：运行 AutoML 或者手动选择六种内置 Amazon Forecast 算法之一。如果运行 AutoML，则在撰写本文档时，Amazon Forecast 会自动测试六种内置算法，并选择在第 10、第 50（中位数）和第 90 个分位数上平均分位数损失最低的算法。

Amazon Forecast 提供四种局部模型：

- 自回归积分滑动平均模型（ARIMA）
- 指数平滑法（ETS）
- 非参数时间序列（NPTS）
- Prophet

局部模型是预测方法，将单个模型拟合到每个单独的时间序列（或特定的项目/维度组合），然后使用这些模型来推断未来的时间序列。

ARIMA 和 ETS 是 R 预测包中流行的局部模型的可扩展版本。NPTS 是 Amazon 开发的一种局部方法，与其他局部模型相比有着关键的区别。与简单的季节性预测器（通过重复最后一个值或相应季节性的值来提供点预测）不同，NPTS 生成概率预测。NPTS 使用固定时间索引，其中前一个索引（T - 1）或前一个季节（T - tau）是对时间步 T 的预测。该算法对集合 {0, ..., T - 1} 中的时间索引（t）随机抽样，生成当前时间步 T 的样本。NPTS 对包含许多零的间歇性（有时也称为稀疏）时间序列特别有效。Forecast 还包括贝叶斯结构时间序列模型 Prophet 的 Python 实施。

Amazon Forecast 提供两种全局深度学习算法：

- DeepAR+
- CNN-QR

全局模型在数据集中的整个时间序列集合上训练单个模型。当一组横截面单位上存在相似的时间序列时，这尤其有用。例如，对不同产品的需求、服务器负载和网页请求的时间序列分组。

通常，随着时间序列数量的增加，CNN-QR 和 DeepAR+ 的功效会提高。对于局部模型来说，情况并非总是如此。深度学习模型还可用于在几乎没有历史销售数据的情况下生成对新 SKU 的预测。这被称为“冷启动”预测。

	Neural Networks		Flexible Local Algorithms	Baseline Algorithms		
	CNN-QR	DeepAR+	Prophet	NPTS	ARIMA	ETS
Computationally intensive training process	High	High	Medium	Low	Low	Low
Accepts historical related time series*	✓	✗	✗	✗	✗	✗
Accepts forward-looking related time series*	✓	✓	✓	✗	✗	✗
Accepts item metadata (product color, brand, etc)	✓	✓	✗	✗	✗	✗
Suitable for sparse datasets	✓	✓	✗	✓	✗	✗
Performs Hyperparameter Optimization (HPO)	✓	✓	✗	✗	✗	✗
Allows overriding default hyperparameter values	✓	✓	✗	✓	✗	✗
Suitable for What-if analysis	✓	✓	✓	✗	✗	✗
Suitable for Cold Start scenarios (forecasting with little to no historical data)	✓	✓	✗	✗	✗	✗

比较 Amazon Forecast 中可用的算法

有关相关时间序列的更多信息，请参阅相关时间序列。

第 4 步：评估预测器

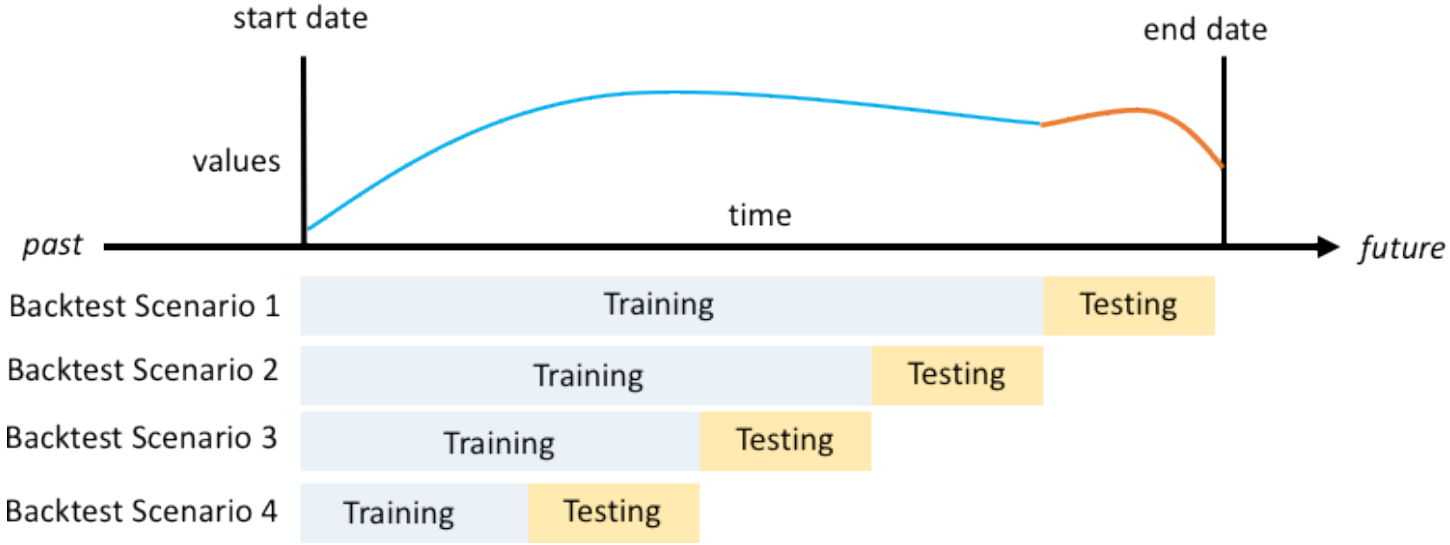
机器学习中的典型工作流程包括在训练集上训练一组模型或模型组合，并评测其在保留数据集上的精度。本部分讨论如何拆分历史数据，以及在时间序列预测中使用哪些指标来评估模型。对于预测，回溯测试技术是评测预测精度的主要工具。

回溯测试

正确的评估和回溯测试框架是使机器学习应用程序取得成功的最重要因素之一。您可以依靠对模型的成功回溯测试，获得对模型的未来预测能力的信心。此外，您可以通过超参数优化（HPO）调整模型，了解模型组合，并启用元学习和 AutoML。

就评估和回溯测试方法而言，时间序列预测特征时间使其与应用了机器学习的其他领域有所不同。在机器学习任务中，要评测回溯测试中的预测错误，通常需要按商品拆分数据集。例如，要在图像相关任务中进行交叉验证，您需要对一定比例的图片进行训练，然后使用其他一些图片进行测试和验证。在预测中，您需要主要按时间进行拆分（在较小程度上按商品进行拆分），以确保不会将训练集内的信息泄露到测试集或验证集，并且尽可能真实地模拟生产案例。

必须谨慎地按时间拆分，因为您不希望选择单个时间点，而是要选择多个时间点。否则，精度会过度依赖于拆分点所定义的预测开始日期。滚动预测评估，即针对多个时间点进行一系列拆分并输出平均结果，从而获得更稳健、更可靠的回溯测试结果。下图说明了四种不同的回溯测试拆分。



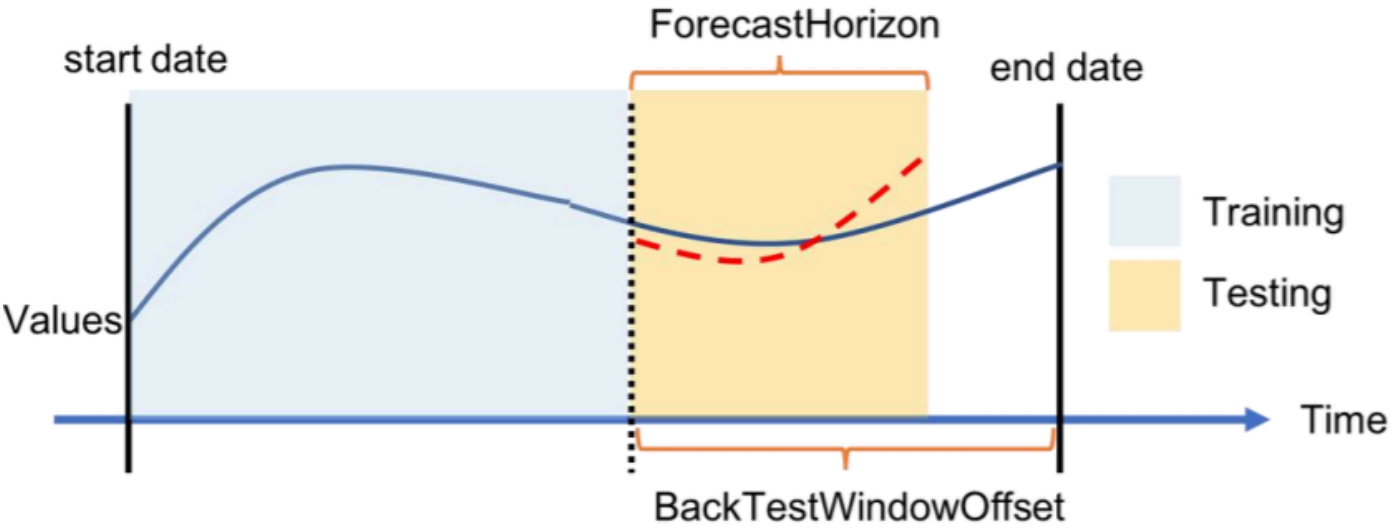
四种不同的回溯测试场景的示意图，这些场景的训练集大小不断增加，但测试大小恒定

在上图中，所有回溯测试场景都提供完整的数据，以便能够根据实际值评估预测值。

之所以需要多个回溯测试窗口，是因为现实世界中大多数时间序列内的变化通常并不平稳。案例研究中的电子商务业务位于北美，其大部分产品需求是由第 4 季度的峰值推动的，尤其是感恩节前后和圣诞节前的峰值。在第 4 季度的购物季中，时间序列的变异性高于今年其余时间。通过设置多个回溯测试窗口，您可以在更平衡的设置中评估预测模型。

对于每个回溯测试场景，下图显示了 Amazon Forecast 术语中的基本元素。Amazon Forecast 会自动将数据拆分成训练数据集和测试数据集。Amazon Forecast 可确定如何使用在 `create_predictor` API 中指定为参数的 `BackTestWindowOffset` 参数来拆分输入数据，或者使用该参数的默认值 `ForecastHorizon`。

在下图中，您可以看到更常见的前一种情况，即 `BackTestWindowOffset` 和 `ForecastHorizon` 参数不相等。`BackTestWindowOffset` 参数定义虚拟预测的开始日期，如下图中的垂直虚线所示。它可以用来回答以下假设问题：如果在这一天部署模型，预测会是什么？ `ForecastHorizon` 定义从虚拟预测开始日期到实际预测的时间步数。



Amazon Forecast 中单个回溯测试场景及其配置的示意图

Amazon Forecast 可以导出回溯测试期间生成的预测值和精度指标。导出的数据可用于评估特定时间点和分位数下的特定商品。

预测分位数和精度指标

预测分位数可以为预测提供上限和下限。例如，使用预测类型 0.1（P10）、0.5（P50）和 0.9（P90），可以提供一系列值，即 P50 预测周围的 80% 置信区间。通过在 P10、P50 和 P90 处生成预测，您能够预期 80% 的真实值会介于这些上下限之间。

本文进一步讨论了第 5 步中的分位数。

Amazon Forecast 在回溯测试期间使用加权分位数损失（wQL）、均方根误差（RMSE）和加权绝对百分比误差（WAPE）精度指标来评估预测器。

加权分位数损失（wQL）

加权分位数损失（wQL）误差指标可衡量模型在指定分位数下的预测精度。当预测不足和过度预测的代价不同时，它特别有用。设置 wQL 函数的权重（τ）会自动纳入对预测不足和过度预测的不同惩罚措施。

$$\mathrm{wQL}[\tau] = 2\frac{\sum_{i,t}[\tau \max(y_{i,t} - q_{i,t}^{(\tau)}, 0) + (1-\tau)\max(q_{i,t}^{(\tau)} - y_{i,t}, 0)]}{\sum_{i,t}|y_{i,t}|}$$

wQL 函数

其中：

- τ – 数据集内的分位数 {0.01，0.02，...，0.99}
- qi,t(τ) – 模型预测的 τ 分位数。
- yi,t – 点 (i,t) 处的观测值

加权绝对百分比误差（WAPE）

加权绝对百分比误差（WAPE）是衡量模型精度的常用指标。它测量预测值与观测值的总体偏差。

$$\mathrm{WAPE} = \frac{\sum_{i,t}|y_{i,t} - \hat{y}_{i,t}|}{\sum_{i,t}|y_{i,t}|}$$

WAPE

其中：

- yi,t – 点 (i,t) 处的观测值
- ŷi,t – 点 (i, t) 处的预测值

预测使用预测均值作为预测值 ŷi,t。

均方根误差（RMSE）

$$\mathrm{RMSE} = \sqrt{\frac{1}{nT}\sum_{i,t}(\hat{y}_{i,t} - y_{i,t})^2}$$

均方根误差（RMSE）是衡量模型精度的常用指标。与 WAPE 一样，它衡量估计值与观测值的总体偏差。

其中：

- yi,t – 点 (i,t) 处的观测值
- ŷi,t – 点 (i, t) 处的预测值
- nT – 测试集内的数据点数量

预测使用预测均值作为预测值 ŷi,t。计算预测器指标时，nT 是回溯测试窗口中的数据点数。

WAPE 和 RMSE 存在问题

在大多数情况下，可以在内部生成的点预测或通过其他预测工具生成的点预测应与 p50 分位数或预测均值相匹配。对于 WAPE 和 RMSE，Amazon Forecast 使用预测均值来表示预测值（yhat）。

对于 wQL[tau] 方程中的 tau = 0.5，两个权重相等，并且 wQL[0.5] 简化为用于点预测的常用加权绝对百分比误差（WAPE）：

$$\mathrm{wQL}[0.5] = 2\frac{\sum_{i,t} 0.5[\max(y_{i,t} - q_{i,t}^{(0.5)}, 0) + \max(q_{i,t}^{(0.5)} - y_{i,t}), 0]}{\sum_{i,t}|y_{i,t}|} = \frac{\sum_{i,t}|y_{i,t} - q_{i,t}^{(0.5)}|}{\sum_{i,t}|y_{i,t}|}$$

其中 yhat = q(0.5) 表示计算预测。在 wQL 公式中使用比例因子 2 来抵消 0.5 因子，以获得精确的 WAPE[median] 表达式。

请注意，WAPE 的上述定义不同于平均绝对百分比误差（MAPE）的常见解释。区别在于分母。WAPE 的上述定义避免了除以 0 的问题，这种问题在现实场景中经常出现，例如案例研究中的电子商务业务，经常在某一天售出 0 个单位的给定 SKU。

与 tau 的加权分位数损失指标不等于 0.5 不同，每个分位数的固有偏差无法通过像 WAPE 这样权重相等的计算来捕获。WAPE 还有其他缺点，包括它不对称，对于小数字，误差百分比过高，并且只是一个按点的指标。

RMSE 是 WAPE 中误差项的平方，也是其他机器学习应用程序中常见的误差指标。RMSE 指标会促成一个模型，其中各个误差具有一致的幅度，因为大的误差变化将按比例过大地增大 RMSE。由于平方误差的存在，因此在一个本来良好的预测中，一些未正确预测的值可能会增加 RMSE。此外，由于使用平方项，较小的误差项在 RMSE 中的权重小于在 WAPE 中的权重。

精度指标允许对预测进行定量评测。特别是对于大规模比较（总体而言，方法 A 比方法 B 好），这些至关重要。但是，用单个 SKU 的视觉效果来补充这一点通常很重要。

第 5 步：生成和使用预测进行决策

一旦您有了满足特定使用案例所需的精度阈值（通过回溯测试确定）的模型，最后一步就是部署模型和生成预测。要在 Amazon Forecast 中部署模型，您必须运行 Create_Forecast API。此操作托管通过对整个历史数据集进行训练而创建的模型（与 `Create_Predictor` 不同，它会将数据拆分为训练集和测试集）。然后，可以通过两种方式使用在预测范围内生成的模型预测：

- 您可以使用 `Query_Forecast` API 从 AWS CLI 或直接通过 AWS Management Console，查询特定商品的预测（通过指定商品或商品/维度的组合）。
- 您可以使用 `Create_Forecast_Export_Job` API 为所有分位数的所有商品和维度组合生成预测。此 API 会生成一个 CSV 文件，该文件安全地存储在您选择的 Amazon Simple Storage Service（Amazon S3）位置。然后，您可以使用 CSV 文件中的数据，将它们插入用于决策的下游系统。例如，您现有的供应链系统可以直接从 Amazon Forecast 中提取输出，帮助就特定 SKU 的制造做出决策。

概率预测

Amazon Forecast 可以在不同分位数处生成预测，这在预测不足和过度预测的成本不同时特别有用。与预测器训练阶段类似，可以为 p1 到 p99 之间的分位数生成概率预测。

默认情况下，Amazon Forecast 针对预测器训练期间使用的相同分位数生成预测。如果在预测器训练期间未指定分位数，则默认情况下将在 p10、p50 和 p90 处生成预测。

对于 p10 预测，预计真实值将在 10% 的时间内低于预测值，并且 wQL[0.1] 指标可用于评测其精度。这意味着在 90% 的时间内，P10 的预测被低估了，如果用来储备库存，90% 的时间内该商品会被售完。当存储空间不足，或者投资资本成本比较高时，P10 预测可能很有用。

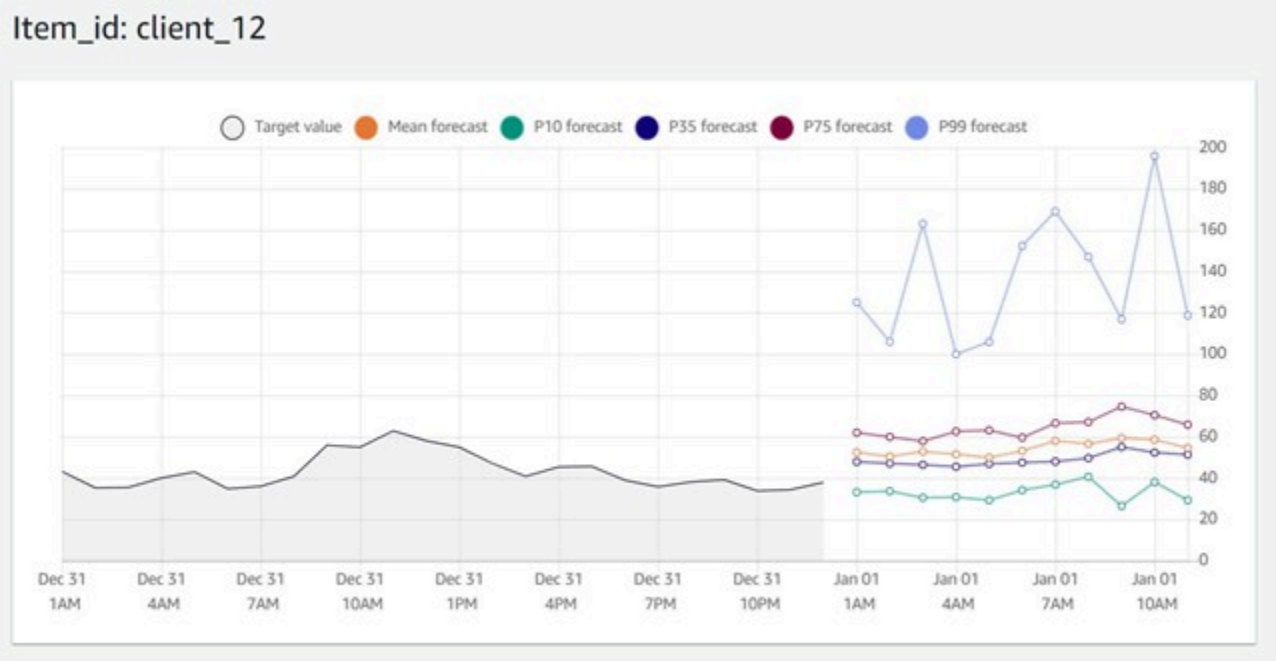
> **Note**
>
> 分位数预测的正式定义是 Pr（实际值 <= 分位数 q 处的预测）= q。从技术上讲，分位数是百分位数/100。统计学家倾向于说“P90 分位数层次”，因为这比“分位数 0.9”更容易说。例如，P90 分位数层次的预测意味着可以预计实际值在 90% 的时间内小于预测值。具体而言，如果 time=t1 和 quantile-level=0.9，预测值 = 30，这意味着如果您有 1000 次模拟，则在 time= t1 时的实际值预计将在 900 次模拟中小于 30，而对于 100 次模拟，实际值预计将超过 30。

另一方面，在 90% 的时间中，P90 预测是过度预测的，当不出售商品的机会成本极高或投资资本成本很低时，它很有用。对于杂货店来说，P90 预测可能用于牛奶或厕纸之类的东西，商店永远不想售完，也不介意总是有一些留在货架上。

对于 p50 预测（通常也称为中位数预测），预计真实值将在 50% 的时间内低于预测值，并且 wQL[0.5] 指标可用于评测其精度。如果库存积压并不太令人担忧，并且对给定商品的需求适度，此时 p50 分位数预测可能很有用。

可视化

Amazon Forecast 允许在AWS Management Console中本地绘制预测。此外，您还可以利用完整的 Python 数据科学堆栈（参见 Amazon Forecast 示例）。Amazon Forecast 允许通过 `ExportForecastJob` API 将预测导出为 CSV 文件，这允许用户在他们选择的分析工具中对预测进行可视化。



在 Amazon Forecast 控制台中对不同的分位数提供的可视化

预测工作流程和 API 摘要

下表将预测工作流程的每个步骤与相应的 Amazon Forecast API 相匹配。

表 1：预测步骤和 Amazon Forecast API

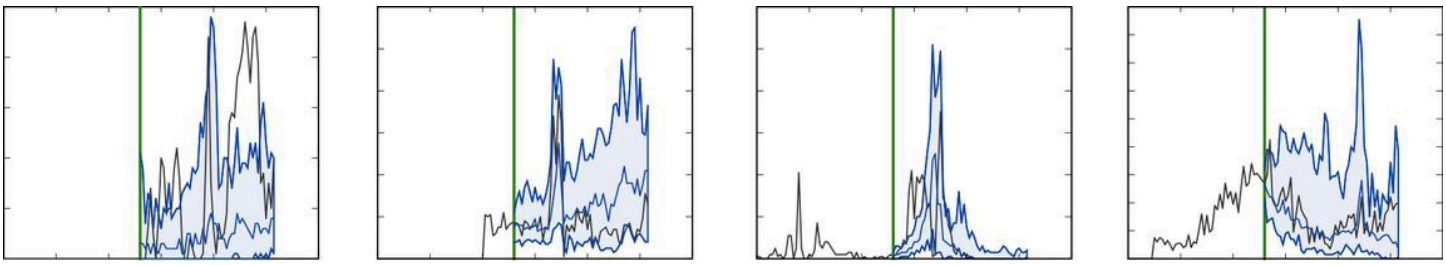
步骤	API	API 函数
第 1 步：收集和聚合数据 第 2 步：准备数据	`Create_Dataset_Group` ，`Create_Dataset` ，`Create_Dataset_Import_Job`	1. 为问题建立高级域（零售、指标等）。 2. 为不同的数据集（目标、相关、商品元数据）定义架构。 3. 将数据从 Amazon S3 导入 Amazon Forecast。
第 3 步：创建预测器 第 4 步：评估预测器	`Create_Predictor`	1. 执行 ETL。 2. 将数据拆分成训练集/测试集，并训练模型。 3. （可选）使用 `Create_predictor_backtest_Export_job` 将回溯测试结果导出为 CSV 以计算商品级指标。
第 5 步：生成和使用预测进行决策	`Create_Forecast`	1. 训练/托管模型。 2. 在预测范围内针对感兴趣的特定分位数（例如，介于 1 到 99 之间的任何整数，包括均值）生成预测。
	`Query_Forecast` 、`Create_Forecast_Export_Job`	使您能够使用由 `Create_Forecast` 创建的预测

在常见场景中使用 Amazon Forecast

您还可以根据外部变量（例如价格或促销）的变化生成不同的预测来进行假设分析。例如，在电子商务案例研究示例中，您可以根据您可能正在计划的促销活动创建不同的预测。您可以预测折扣为 10% 的产品需求，然后以 20% 的折扣预测产品需求，以了解为满足需求而需要库存的产品数量。要实现这一点，可以设置唯一的数据集组，并根据感兴趣的场景更新每个组中的相关时间序列。

此外，您还可以为没有先前历史记录的商品生成预测（有时称为冷启动问题）。这种方法需要使用 DeepAR+ 或 CNN-QR 以及元数据（例如商品元数据集）创建预测器，以生成新商品的预测。

下图显示了四个不同的 SKU 出现在现实世界运营预测问题中的示例。



在现实世界运营预测问题中出现的四个不同 SKU 的示例。

在上图中，观测到的实际值位于垂直线的左边，蓝色的预测值位于垂直线的右边，与黑色的实际值相对。请注意，每个人的 SKU 的历史记录（位于垂直线的左侧）并不表示绿线右侧的 SKU 的演变情况。

在生产中实施预测

完成端到端 Amazon Forecast 工作流程后，确定 `Create_Predictor` 和 `Create_Forecast` API 之间的关键区别以及应何时使用每个 API 至关重要。

前者主要在概念验证期间用于评估模型精度/指标，而后者用于在生产环境中生成预测。

投入生产后，`Create_Predictor` 无需在每次需要生成预测时运行，而只有在由于数据变化或作为预设节奏（例如，每两周或每月）的一部分而需要对模型进行重新训练时才需要运行。当使用新数据更新数据集时，只需要运行 `Create_Forecast`，即可为新的预测范围生成预测。

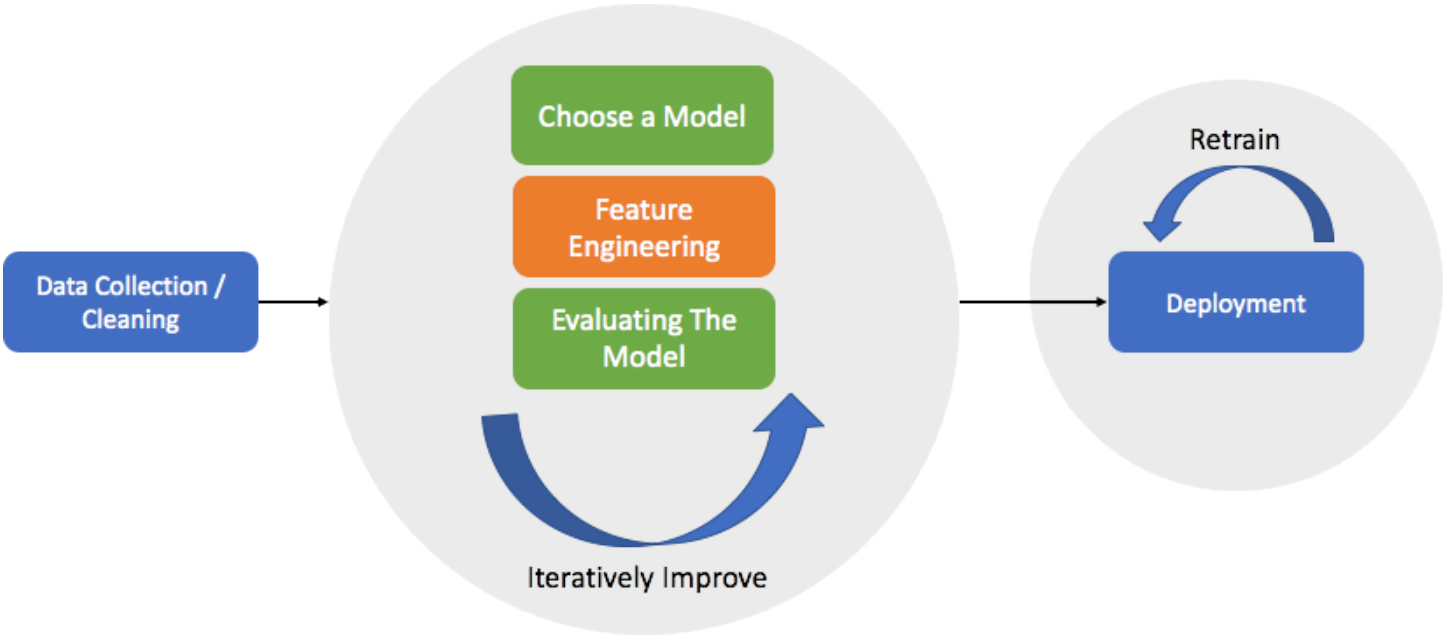
在生产中，您还需要自动执行数据集的导入和预测操作，以便滚动生成新的预测。如今，要实现这一点，可以通过组合使用 Amazon CloudWatch Events 日志、AWS Step Functions 和 AWS Lambda 函数来设置 cron 任务。反过来，cron 任务设置会自动调用 Amazon Forecast API，以进行导入/重新训练或生成预测。最后，需要管理您的资源并定期删除，这一点是关键，这样您就不会超过该服务规定的系统限制。请参阅这篇包括 Amazon Redshift 在内的博客文章，了解有关设置预定任务的更多信息。

总结

Januschowski 和 Kolassa（2019）对预测问题进行了分类，这些问题与企业需要做出的决策（包括战略、战术和运营决策）相一致。每个决策层都有相应的预测任务。

运营和战术预测问题的特点是包含大量数据，通常需要高度自动化。不同的预测方法可以解决这些问题。局部预测方法通常可以很好地解决战略预测问题，基于深度学习的方法可以很好地解决运营预测问题，对于介于两者之间的问题，可能需要进行一些实验。尽管本白皮书讨论的是运营预测问题，但 Amazon Forecast 对其提供的模型并不固执己见，它包括了用于解决战略以及运营和战术预测问题的模型。

运营预测问题解决过程可以分为从数据收集和准备到模型构建和部署的基本步骤。通常，将此过程视为迭代过程而不是线性过程是最有用的。例如，随着模型和使用案例被更好地理解，回到数据收集阶段可能是有意义的。模型开发本身也是高度迭代的。



将预测模型投入生产的简化开发流程。

贡献者

本文档的贡献者包括：

- Yuyang Wang，人工智能垂直服务高级机器学习科学家
- Danielle Robinson，人工智能垂直服务应用科学家
- Tim Januschowski，机器学习应用科学经理
- Namita Das，人工智能垂直服务高级产品经理
- Christy Bergman，高级人工智能/机器学习专家，解决方案架构师
- Kris Tonthat，人工智能/机器学习文档技术作家

延伸阅读

有关时间序列预测和深度学习方法的更多信息，请参阅：

- Amazon Forecast 文档
- Amazon Forecast 公开发布博客
- Amazon SageMaker 现已提供 DeepAR 算法（更准确）
- Amazon SageMaker DeepAR 现在支持缺失值、分类和时间序列特征以及广义频率
- Amazon Forecast 现在可以使用卷积神经网络（CNN）来训练预测模型，这使得速度提高 2 倍且准确性提高 30%
- Amazon Forecast 现在支持准确衡量个别商品
- 利用 Amazon Forecast 衡量预测模型准确性以优化业务目标
- Amazon Forecast Weather Index – 自动提供当地天气信息，以提高预测模型的准确性
- 关于时间序列预测模型的科学论文
- Amazon Forecast 示例 GitHub 页面
- AWS 架构中心

附录 A：常见问题解答

问：如何开始使用 Amazon Forecast？

1. 首先，您需要一个 AWS 账户。
2. 接下来，在 AWS Management Console 中打开 Forecast 服务，创建一个数据集组，然后将 `.csv` 文件导入到目标时间序列数据集（必需）。开始所需的最低数据是您想要预测的数量的历史数据，例如每个家庭每个时间戳的用电量。
3. 最后，通过运行 CreatePredictor 来创建模型，然后通过运行 CreateForecast 来生成结果。有关更多详细信息，请参阅入门文档页面。

另请参阅 GitHub 简介和最佳实践指南。

问：Amazon Forecast 适合我吗？

并非所有的机器学习问题都是预测问题。要问的第一个问题是：“我的业务问题是否在陈述中包含时间序列？”例如，您是否只需要未来某个特定时间和日期的特定值？ 预测不适合于一般的静态（特定日期/时间无关紧要）问题，例如欺诈检测或向用户推荐电影片名。对于静态问题，有更快的解决方案。

除了具有时间序列数据外，数据本身应该是“密集的”，并且具有较长的历史。下表对此进行了总结：

表 2 – 标准和 Amazon Forecast 算法类

标准	Amazon Forecast 算法类
包含多达 500 万个时间序列的大型数据集，具有相似的基础模式 + 季节效应 + 相关数据。每个时间序列都应该有很长的历史，如果试图捕捉年度事件，则最好超过两年，并且每个时间序列都应超过 300 个，理想情况下至少有 1000 个数据点。	Amazon Forecast 专有深度学习 DeepAR+、CNN-QR
包含 1-100 个时间序列的小型数据集，其中大多数时间序列有 300 多个数据点 + 季节效应 + 相关数据。	Prophet
包含 1-10 个时间序列的小型数据集，其中大多数时间序列有 300 多个数据点 + 季节效应。	ETS、ARIMA

标准	Amazon Forecast 算法类
包含 1-10 个时间序列的间歇数据集（稀疏，包含许多 0），其中大多数时间序列有 300 多个数据点。	Amazon Forecast 专有 NPTS
包含 1-10 个时间序列的小型数据集（常规或稀疏数据集），其中大多数时间序列的数据点少于 300 个。	对于 Amazon Forecast 来说，数据太小了。改用 Excel 中的 ETS 或传统的统计模型 ARIMA 和 Prophet。

最佳做法是第一次在预测器中使用 AutoML 模式训练数据。AutoML 将自动运行所有算法（DL 算法在 HPO 开启时运行），以了解哪种算法对您的数据最有效。

问：缺失的数据是什么情况？ 什么时候才能产生合理的预测？

可能是数据记录存在问题，或者数据的聚合水平过低或过高。一般规则是，预测长度不能超过训练数据的 1/3。

除了缺失的数据量外，另一个考虑因素是插补缺失的数据。您可以将所有 0 转换为空值，然后让 Amazon Forecast 完成自动插补缺失值的繁重工作。Amazon Forecast 将自动检测缺失值是由于新产品推出（冷启动）还是产品报废所致。您可以使用多个缺失值逻辑，包括值、中值、最小值、最大值、零、平均值和 nan（仅限目标时间序列）。请参阅空值填充语法文档。

- “frontfill”–（仅限 TTS）指的是新项目或冷启动项目，以及您想在项目开始有任何历史记录之前如何处理空值
- “middlefill”– 指的是时间序列值中间的空值
- “backfill”– 指的是生命周期已终止的项目，以及您想在项目停止销售后如何处理空值
- “futurefill”–（仅限 RTS）指的是训练数据结束后出现的空值

问：我的输入历史数据没有负值，但是我在需求预测中看到负值？ 为什么会发生这种情况？ 如何避免出现这种情况？

对于除 NPTS（基于非负数据训练）和 DeepAR（具有负二项式似然函数）之外的所有模型，不能保证会生成正数。解决方案是更改为上述模型之一，或者将预测值截断为非负值。

问：为什么准确度指标在分位数上有所不同？ 既然模型相同，错误不应该一样吗？

有关权重如何取决于分位数的更多说明，请参阅加权分位数损失（wQL）。

想象一下，所有预测都位于三个不同的分位数：p10、p50、p90。这三个预测本身都是随机变量。准确度是在每个分位数级别的实际值和预测值之间分别计算的。您可能会看到一个“wQL”表，即加权分位数损失，如下所示。wQL 值彼此之间没有确定性关系。（召回损失意味着错误，因此是无序的；然而，分位数预测是有序的）。因此，例如，p90 wQL 没有理由必须大于 p50 wQL。

表 3 – 预测分位数示例

	A	B，	C
1	P10 wQL	P50 wQL	P90 wQL
2	0.18647	0.50879	0.30428

问：如何提高预测准确性？

预测的准确性取决于在数量和质量上是否有正确的数据。如果准确性不令人满意，那么了解问题的可预测性（或数据的随机性/噪声/平稳程度）可能是有意义的。其他需要考虑的因素包括，评估不同的模型、超参数设置，以及使用相关的时间序列和项目元数据集整合其他特征。有关具体建议，请参阅 GitHub 上的这份最佳实践文档。

问：我有一个非常适合我的使用案例的算法，但 Amazon Forecast 中没有提供该算法。我应该怎么办？

Amazon Forecast 团队很乐意帮助您处理这个使用案例。通过电子邮件 <amazonforecast-poc@amazon.com> 联系 Amazon Forecast 的服务团队。

附录 B：参考资料

Tim Januschowski 和 Stephan Kolassa。业务预测问题的分类。预测：国际应用预测杂志。2019

David Salinas、Valentin Flunkert、Jan Gasthaus 和 Tim Januschowski。DeepAR：使用自回归循环网络进行概率预测。国际预测杂志。2019

Jan Gasthaus、Konstantinos Benidis、Yuyang Wang、Syama Sundar Rangapuram、David Salinas、Valentin Flunkert 和 Tim Januschowski。{使用样条分位数函数 RNN 进行概率预测。第 22 届国际人工智能与统计会议。2019

Tim Januschowski、Jan Gasthaus、Yuyang Wang、David Salinas、Valentin Flunkert、Michael Bohlke-Schneider 和 Laurent Callot。对预测方法进行分类的标准。国际预测杂志。2019（需要登录）

Tim Januschowski、Jan Gasthaus、Yuyang Wang、Syama Rangapuram 和 Laurent Callot。用于预测的深度学习。预测：国际应用预测杂志。2018

Tim Januschowski、Jan Gasthaus、Yuyang Wang、Syama Sundar Rangapuram 和 Laurent Callot。用于预测的深度学习：当前的趋势和挑战。预测：国际应用预测杂志。2018

Joos-Hendrik Bose、Valentin Flunkert、Jan Gasthaus、Tim Januschowski、Dustin Lange、David Salinas、Sebastian Schelter、Matthias Seeger 和 Yuyang Wang。大规模的概率需求预测。VLDB 投稿论文集。2017

文档历史记录

要获得有关此白皮书的更新通知，请订阅 RSS 源。

变更	说明	日期
已更新白皮书	更新。	2021 年 9 月 1 日
初次发布	首次发布白皮书。	2020 年 2 月 4 日

> **Note**
>
> 要订阅 RSS 更新，您必须为您正在使用的浏览器启用 RSS 插件。

声明

客户负责对本文档中的信息进行独立评估判断。本文档：（a）仅供参考，（b）代表 AWS 当前的产品服务和实践，如有变更，恕不另行通知，以及（c）不构成 AWS 及其联属公司、供应商或授权商的任何承诺或保证。AWS 产品或服务均“按原样”提供，没有任何明示或暗示的担保、声明或条件。AWS 对其客户的责任和义务由 AWS 协议规定，本文档与 AWS 和客户之间签订的任何协议无关，亦不影响任何此类协议。

AWS 词汇表

有关最新的 AWS 术语，请参阅 AWS 一般参考 中的 AWS 词汇表。