DeepL

订阅DeepL Pro以翻译大型文件。

欲了解更多信息，请访问www.DeepL.com/pro。

ResearchGate

有关该出版物的讨论、统计资料和作者简介，请访问： https://www.researchgate.net/publication/339362837

从 Kaggle 预测竞赛中汲取的经验

预印本 - 2020 年 2 月

DOI: 10.13140/RG.2.2.21579.75046

引文
1

阅读
15,742

2 位作者：

卡斯珀-索尔海姆-博伊尔

阿尔堡大学

9 篇著作 **264** 次引用

查看简介

Jens Peder Meldgaard

阿尔堡大学

3 篇著作 **235** 次引用

查看简介

本页下面的所有内容由 Casper Solheim Bojer 于 2020 年 2 月 19 日上传。

用户要求增强下载的文件。

从 Kaggle 预测竞赛中汲取的经验

作者：卡斯帕-索尔海姆-博伊尔和延斯-佩德-梅尔德加德

摘要

竞赛在预测领域发挥着不可估量的作用，最近的 M4 竞赛就是一个例证。该竞赛受到了学术界和从业人员的关注，并引发了关于商业预测数据代表性的讨论。然而，Kaggle 平台上以真实商业预测任务为特色的几项竞赛在很大程度上被学术界所忽视。我们相信，从这些比赛中汲取的经验对预测界大有裨益，并对 Kaggle 六次比赛的结果进行了回顾。我们发现，大多数 Kaggle 数据集的间歇性和熵值都高于 M 竞赛，而且全局集合模型往往优于局部单一模型。此外，我们还发现梯度提升决策树的性能很强，神经网络在预测方面的成功率越来越高，而且有多种技术可以使机器学习模型适应预测任务。

1 导言

预测涉及对未来的准确预测，是财务规划、库存管理和产能规划等许多业务规划流程的关键输入。因此，工业界和学术界都对开发能够进行准确可靠预测的方法产生了浓厚的兴趣，每年都会提出许多新方法。

预测竞赛是对各种时间序列的预测方法进行比较和经验评估，被广泛认为是预测界的标准，因为预测竞赛评估的是与现实预测环境一致的事前预测（Hyndman，2020 年）。

过去 50 年间，预测界举办了多次预测竞赛，其中以 Spyros Makridakis 和 Michèle Hibon 组织的 M 竞赛最受关注。最近的 M4 竞赛试图总结过去 20 年中开发的新方法，并回答以往竞赛未回答的一些问题。M4 竞赛的一个假设是，机器学习（ML）界最近的建模进展是否准确。这些模型比传统的时间序列模型复杂得多，主办方根据以往比赛的结果预测，这些模型不会比简单的方法更准确。M4 竞赛的结果支持了早期研究的说法，即组合模型比单一模型表现出更优越的性能，但有关简单模型性能优越的假设没有得到支持，因为前两个解决方案都使用了 ML 的复杂方法（Makridakis 等人，2020b）。

M4 竞赛结果与商业预测领域的相关性一直是一些讨论的主题，因为一些从业人员对竞赛数据集的代表性提出了质疑。他们认为，该数据集不能代表企业组织所面临的许多预测任务（Darin & Stellwagen, 2020; Fry & Brundage, 2019）。主要的批评意见涉及每周、每日和次每日级别的高频序列代表性不足，以及无法获取时间序列外部的有价值信息，例如外生变量和时间序列层次结构。组织者承认了这两点（Makridakis 等人，2020a），因此仍需进一步研究可获取外部信息的高频商业时间序列预测方法的相对性能。作为对批评意见的回应，新的竞赛涉及

上述问题以 M5 竞赛的形式宣布，该竞赛将在在线数据科学平台 Kaggle 上举行（M Open Forecasting Center，2019 年）。[1](M开放预测中心，2019年）。

然而，在 Kaggle 平台上已经完成了几项针对真实商业预测任务的预测竞赛，但预测界在很大程度上忽视了这些竞赛的结果。我们认为，这些竞赛为预测界提供了一个学习机会，竞赛的结果可能预示着 M5 竞赛的结果。

为了概述预测界可以从 Kaggle 预测竞赛中学到什么，我们首先确定了最近的六个相关预测竞赛。然后，我们从预测任务和时间序列特征的角度分析了竞赛数据集，并将其与 M3 和 M4 竞赛进行了比较。为确保 Kaggle 解决方案比简单、成熟的预测方法更有价值，我们对 Kaggle 解决方案进行了基准测试。随后，我们回顾了比赛中表现最出色的解决方案，并将我们的发现与 M4 比赛中的解决方案进行了对比。基于这些经验，我们为即将举行的 M5 竞赛的结果提出了假设。

本文接下来的内容安排如下。第 2 节介绍了预测竞赛的背景，重点是最近的 M4 竞赛；第 3 节介绍了所选的竞赛。随后，我们在第 4 节对六个竞赛数据集进行了分析，并在第 5 节对竞赛解决方案进行了基准测试。然后，我们将在第 6 节中对竞赛中表现最出色的解决方案进行回顾。最后，我们在第 6 节中总结和讨论了我们的发现，并对即将举行的 M5 竞赛的发现提出了假设，最后在第 7 节中总结了从 Kaggle 竞赛中学到的知识。

2 背景介绍

M-competitions 在预测界具有很大的影响力，因为它们将预测界的注意力集中在方法的经验准确性上，而不是模型的理论属性上。此外，比赛允许任何人参加，使具有不同偏好和技能的参赛者都能使用自己喜欢的模型。这样就能更公平地比较各种方法，并利用预测界现有的各种建模能力。读者可参阅 Hyndman（2020 年）的文章，了解前三届 M 竞赛的回顾，我们将重点关注最近一届竞赛。M4 竞赛是针对 Makridakis 等人（2020b）在前几届竞赛中提出的反馈意见而举办的：

- 包括每周、每天和每小时的高频数据、
- 要求预测区间以解决预测不确定性问题
- 强调可重复性、
- 将许多行之有效的方法作为基准、
- 将样本量增加到 100.000 个时间序列，以消除对研究结果统计意义的担忧。

参赛的时间序列主要来自商业领域，且仅限于连续时间序列。此外，时间序列不允许是间歇性的或有缺失值，而且每个频率都要求有三个以上的完整季节期（Spiliotis 等人，2020 年），只有周时间序列只要求有 80 个观测值（Makridakis 等人，2020 年b）。

[1]https://www.kaggle.com/

竞赛结果可分为四大主题：i) 复杂模型与简单模型；ii) 交叉学习；iii) 预测不确定性；iv) 集合，详情请参见 Makridakis 等人（2020b）。关于复杂模型与简单模型，竞赛发现复杂 ML 方法优于常用于时间序列预测的简单模型，因为前两个解决方案分别使用了神经网络和梯度提升决策树。值得注意的是，这些方法已被调整用于预测，因此不是开箱即用的 ML 模型，正如主办方所假设的那样，这些模型表现不佳。比赛还展示了交叉学习的优势，即在多个时间序列中学习时间序列模式。表现最好的两名选手都使用了在多个时间序列上估计的模型，这与每个时间序列一个模型的主流方法不同。比赛中最令人惊讶的发现之一是，获胜者对预测不确定性的估计非常准确。在预测领域，准确估计不确定性是一项长期挑战，因为大多数方法都低估了不确定性（Fildes & Ord，2007 年）。最后，比赛再次证实，预测方法的组合（在 ML 中称为集合）比单一方法产生的结果更准确。

比较 ML 方法和统计方法的预测性能是 M4 竞赛的目标之一。然而，这些术语对分类的有用性一直存在争议。Januschowski 等人（2020 年）认为，这种区分是无益的，因为它没有对方法进行清晰客观的分类。举例来说，M4 竞赛中排名前两位的方法使用了神经网络和梯度提升决策树（GBDT）等与 ML 术语相关的方法，以及指数平滑和其他经典时间序列预测方法等与统计术语相关的方法。作为替代方案，Januschowski 等人（2020 年）主张使用一套更全面的分类维度，包括全局模型与局部模型、线性模型与非线性模型、数据驱动与模型驱动、集合模型与单一模型，以及可解释性与预测性。

Kaggle 是一个在线数据科学平台，以商业问题、招聘和学术研究为目的举办数据科学竞赛。该平台拥有一个由来自不同背景的数据科学家组成的庞大社区，他们在竞赛中竞争，并通过分享知识和讨论潜在策略参与论坛。在以业务问题为重点的竞赛中，公司会提供相关预测任务的数据集，并通常会为表现最出色的选手提供现金奖励。与大多数由学术机构主办的预测竞赛不同，Kaggle 竞赛以公开排行榜的形式对提交的预测提供实时反馈，排行榜上会显示参赛者的排名及其分数。参赛者可以提交多个预测，从而促进学习，并做出更好的预测（Athanasopoulos & Hyndman，2011 年）。最终的比赛结果以私人排行榜的表现为基础，而私人排行榜的表现是在一个未见过的数据集上进行评估的，以防止对排行榜的过度拟合。

3 竞赛选择

我们最初检查了在线数据科学平台 Kaggle 的竞赛数据库，只保留了以预测为重点的竞赛供进一步考虑。之后，由于自 2014 年以来预测和 ML 领域发生了很多变化，我们只考虑了 2014 年及以后的比赛，从而减少了比赛库。这样，竞赛库就减少到了以下七项竞赛：

- *沃尔玛商店销售预测*[2]
- *罗斯曼商店销售*[3]
- *风雨中的沃尔玛销售*[4]
- *Grupo Bimbo 库存需求*[5]
- *维基百科网络流量时间序列预测*[6]
- *Corporación Favorita 杂货销售预测*[6]
- *招聘餐厅游客预测*[6]

在对竞赛数据集进行更彻底的审查后，我们排除了 *Grupo Bimbo 库存需求*竞赛的进一步审查。该数据集的每个时间序列每周最多只有七个观测值，许多时间序列只有一个观测值，因此不适合进行时间序列预测。

表 1 概述了被选中进行审查的六项竞赛及其预测任务的特点。六项竞赛中有四项来自零售领域，其余两项分别来自网络流量和餐饮领域。虽然零售领域的比赛属于同一领域，但它们在时间序列数量和聚合级别方面却有很大不同。*沃尔玛商店销售额*竞赛和 *Rossmann* 竞赛在业务层次上的聚合度都很高，序列数量相对较少，分别需要按商店/部门/周和商店/日预测美元销售额。另一方面，*Corporación Favorita* 竞赛和*沃尔玛风暴天气竞赛*的分类水平都很高，需要按产品/商店/天预测单位销售额。不过，它们在时间序列的数量上有所不同。

其余两个竞赛项目都需要对网页和餐馆的每日访问量进行预测，因此都需要进行分类预测，但其领域和时间序列的数量有所不同。*Recruit Restaurant* 数据集因其领域而包含了即将到来的预订数据，预计将包含对预测任务有用的信息。*维基百科*数据集是一个更传统的大规模预测任务，尽管它提供了对预测层次结构的访问。

[2]https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting
[3]https://www.kaggle.com/c/rossmann-store-sales
[4]https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather
[5]https://www.kaggle.com/c/grupo-bimbo-inventory-demand
[6]https://www.kaggle.com/c/web-traffic-time-series-forecasting

[7]https://www.kaggle.com/c/favorita-grocery-sales-forecasting
[8]https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting

表 1：部分 Kaggle 竞赛

竞赛	时间单位	预测股	#观察	#Timeseries	地平线	精度测量
沃尔玛商店销售额 (2014)	每周	销售额 部门	143	3331	1-39	WMAE
沃尔玛《暴风骤雨》（2015 年）	每日	按产品和商店分列的单位销售额	851-1011	255	1-7	RMSLE
罗斯曼（2015）	每日	按商店分列的销售额	942	1115	1-48	RMSPE
维基百科（2017）	每日	按页面和流量类型划分的浏览量	970	~145k	12-42	SMAPE
最受欢迎公司 (2018)	每日	单位销售额 产品和商店	1684	~210k	1-16	NWRMSLE
招聘餐厅（2018）	每日	餐厅访问量	478	821	1-39	RMSLE

因此，所考虑的竞赛就其特点而言是多种多样的，但与 M4 竞赛相比，所呈现的公司所面临的商业预测任务的子集要有限得多。尽管如此，这些竞赛在可用信息方面更能代表现实，因为外生变量和业务层次结构都可以在创建预测时加以利用。

4 竞赛数据集分析

分析已确定的 Kaggle 竞赛数据集的目的是将这些数据集与 M3 和 M4 竞赛相对比。Kaggle 竞赛中使用的数据集描述了已知公司预测任务的实际情况，因此我们知道这些数据集代表了特定的现实世界环境。为了将 Kaggle 竞赛数据集与 M3 和 M4 数据集相比较，我们利用 Kang 等人（2017 年）开发的方法在二维空间中表示单个时间序列，以便分析大规模时间序列数据集。此外，我们还可以讨论 M3 和 M4 竞赛的结果是否适用于与已确定的 Kaggle 竞赛类似的情况。

4.1 数据预处理

由于所有已确定的 Kaggle 竞赛都代表了一个特定的现实世界背景，因此第 2 节中讨论的 M4 竞赛在时间序列的长度和规律性方面的选择标准并不适用。因此，有必要进行一些初步预处理，以便对时间序列实例空间进行外推。所有预处理均使用 R 软件包 **data.table** (Dowle & Srinivasan, 2019) 和 **base** (R Core Team, 2019) 进行，可总结为五个步骤：

1. 将 NA 或负值设为零。
2. 删除所有零值的时间序列。
3. 如果有测试集，则只保留训练集和测试集中的时间序列。
4. 通过用零填充缺失的时间间隔，将不规则时间序列转换为规则时间序列。

5. 去掉前导零。

有关预处理步骤及其影响的完整说明，请参见附录 A。

4.2 竞争代表性

由于能够提供有关 M3 竞赛数据的有用信息，Kang 等人（2017 年）提出了一组特征 *F1*、*F2*、……、*F6*，可将任何长度的时间序列概括为特征向量 **F** = (*F1*、*F2*、*F3*、*F4*、*F5*、*F6*)：

1. 根据 Goerg（2013 年）的定义，*频谱熵*（*F1*）衡量 "可预测性"。
2. *趋势强度*（*F2*）衡量时间序列平均水平长期变化的影响。
3. *季节性强度*（*F3*）衡量季节性因素的影响。
4. *季节周期*（*F4*）解释了周期性模式的长度。
5. *一阶自相关*（*F5*）衡量时间序列与一步滞后序列之间的线性关系。
6. *最佳箱-柯克斯变换参数*（*F6*）用于衡量整个序列的方差是否近似恒定。

为了计算特征向量，我们使用了 R 软件包 **feats**（O'Hara-Wild 等人，2019 年），随后使用 R 软件包 **stats**（R Core Team，2019 年）中的 prcomp 算法应用主成分进行降维，将所有特征向量投影到二维空间，以便使用 R 软件包 **ggplot2**（Wickham，2016 年）轻松实现可视化。Spiliotis 等人（2020 年）在评估 M4 竞赛的代表性时也使用了类似的方法。在研究中，他们使用了 M4 竞赛作者定义的*季节周期（F4）*，即年、周和日时间序列的季节周期为 1，季度时间序列的季节周期为 4，月时间序列的季节周期为 12，小时时间序列的季节周期为 24。因此，由于估算算法要求季节周期大于 1，因此无法估算周序列和日序列的季节性。在我们的方法中，我们将周时间序列的原始季节周期替换为 52，将日时间序列的原始季节周期替换为 7。因此，我们假定周时间序列表现出每年（52 周）的季节性，而日时间序列表现出每周（7 天）的季节性，从而可以估计*季节性的强度（F3）*。此外，我们在降维过程中排除了*季节性时期（F4）*，因为它除了根据指定的季节性时期分离时间序列特征向量外，没有其他价值。有关改变和排除*季节期（F4）*的影响的完整说明，请参见附录 B。

图 1 描述了 M3、M4 和 Kaggle 竞赛的时间序列实例空间，每个六边形区域的数据密度以深灰色表示低密度，蓝色表示高密度。除 M4 图外，其他所有图都包含说明 M4 时间序列实例空间的浅灰色背景，而 M4 图则包含除 M4 竞赛外所有时间序列实例空间的浅灰色背景。图 1 显示了在业务层级中具有相似聚合层级的竞赛之间在时间序列定位方面的相似性。*Rossmann*、*Recruit Restaurant* 和*沃尔玛商店销售*比赛的聚合度都很高，在时间序列实例空间的同一区域内都有密度峰值。我们看到，与 M4 竞赛中的大多数时间序列相比，高度聚合的时间序列具有较低的*趋势*度和*一阶自相关度*，以及较高的*频谱熵*度。*Corporacíon Favorita*、*Walmart Stormy Weather* 和*维基百科*竞赛的聚合度都较低，但它们在时间序列实例空间中的位置相似性并不明显。在这里，我们看到 *Corporacíon Favorita* 和 *Walmart Stormy Weather* 非常相似，都表现出很高的*频谱熵*、较低的*趋势*和*一阶自相关性*，以及不同程度的*季节性*和λ。相反，*维基百科*比赛在较低的聚合水平上显示出明显高于其他比赛的*趋势*和*一阶自相关性*。

我们预计，M-竞赛和 Kaggle 竞赛之间的*熵*差异在一定程度上是由 M-竞赛中禁止间歇性的选择标准造成的，因为与 M-竞赛相反，所有 Kaggle 竞赛都有一些间歇性时间序列。明确地说，我们发现在 *Corporacíon Favorita*、*Walmart Stormy Weather*、*Recruit Restaurant* 和 *Rossmann* 比赛中，98% 以上的时间序列都表现出一定程度的间歇性。此外，我们还发现，在*沃尔玛商店销售*和*维基百科*比赛中，分别约有 16% 和 26% 的时间序列表现出间歇性。M3 和 M4 竞赛涵盖了时间序列实例空间的很大一部分，但如果能包含更多熵值和间歇性程度更高的时间序列，将会提高竞赛数据集的代表性。

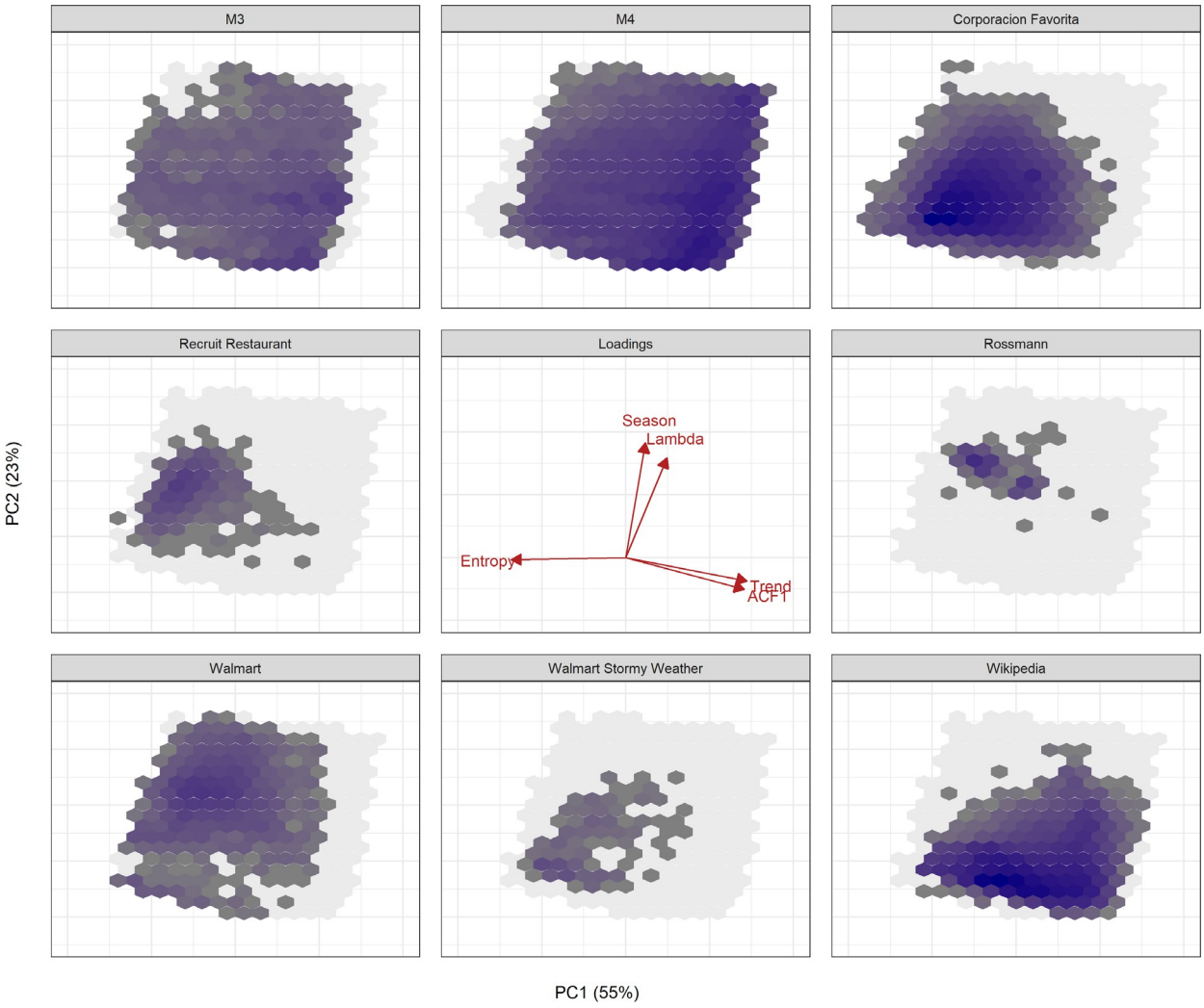

图 1：M- 和 Kaggle 竞赛的时间序列实例空间六边形图。每个六边形的颜色表示时间序列在实例空间特定区域的密度，蓝色表示高密度，深灰色表示低密度。此外，M4 竞赛的实例空间在所有图中都以浅灰色背景表示，但 M4 图除外，浅灰色背景表示除 M4 竞赛外所有竞赛的综合实例空间。

5　对 Kaggle 解决方案进行基准测试

我们这篇文章的基本前提是，从 Kaggle 竞赛中表现最出色的解决方案中学到的知识是有价值的。要做到这一点，这些解决方案至少应优于简单的和

这两种预测方法都是经过验证的时间序列预测方法。为了验证情况是否如此，我们将这些解决方案与天真预测法和季节性天真预测法这两种预测方法进行了比较。这些方法通常用于识别预测方法或流程是否增加了价值，例如预测增值分析（FVA）（Gilliland，2011 年）或 MASE 预测准确度测量（Hyndman & Koehler，2006 年）。我们之所以选择这两种方法，是因为它们既简单又对缺失数据具有鲁棒性，而缺失数据在所有 Kaggle 竞赛中都存在。其他经常使用的基准方法，如 Theta 模型和指数平滑模型，需要一个没有缺失数据的时间序列。因此，我们没有使用这些方法，因为这需要在预测前使用估算程序来填补缺失值。

为了建立基准，我们使用天真法和季节性天真法对所有比赛进行了预测。有些比赛要求对训练数据集中不存在的时间序列进行预测，因此我们不得不使用后备方法。作为后备方法，我们使用预测层次结构中下一级的平均值来进行一些简单的交叉学习。如果下一级的数据仍然缺失，我们就继续往上走，直到有数据为止。有关每个数据集所用程序的具体细节，请参见附录 C。我们使用相对误差来衡量解决方案与基准相比的性能，将每次竞赛中第 1 名和第 50 名解决方案的准确度与天真方法和季节性天真方法的准确度相除。表 2 显示了两种基准方法中较好的基准程序的结果。
总体而言，排名第一的解决方案比简单基准提高了 25% 以上。同样，排名第 50 位的解决方案都比简单基准高出至少 10%。在 *Rossmann* 和 *Corporación Favorita* 比赛中，性能改进尤为显著，第一名的解决方案分别将预测误差降低了 74% 和 60%。从这些基准中可以清楚地看出，Kaggle 解决方案的附加值都高于简单的时间序列基准，值得进一步关注。

表 2：Kaggle 竞赛第 1 名和第 50 名与 Naïve 和 Seasonal Naïve 预测方法的基准比较。

竞赛	方法	Rel.第 1 名	Rel.误差 第 50 位
最受欢迎公司	季节性天真	0.40	0.41
招聘餐厅	季节性天真	0.73	0.75
罗斯曼	季节性天真	0.26	0.29
沃尔玛商店销售额	季节性天真	0.73	0.89
沃尔玛暴风雨天气	季节性天真	0.70	0.75
维基百科	天真	0.73	0.82

6 比赛回顾

为了进行评测，我们阅读了 Kaggle 论坛上每场比赛的帖子。我们收集了参赛者发布的解决方案的所有信息，包括文字描述和代码。在每次比赛中，我们都会考虑对排名前 25 位的解决方案进行评审，以关注表现最出色的选手。此外，我们还检查了论坛中简单方法的应用情况，如历史平均值或经过验证的预测方法，以研究使用简单方法所获得的改进。表 3 显示了我们在审查过程中发现的每场比赛中前 25 名选手报告的解决方案，以及所使用方法的编码。空白格表示没有在论坛上描述其方法的参赛者。从表中可以明显看出，有相当一部分参赛选手没有报告他们的方法，尽管成绩最好的选手通常会报告。这是 Kaggle 竞赛学习的一个局限，我们将在第 7.3 节讨论其影响。对表 3 的分析还揭示

了另一种模式：两次沃尔玛竞赛的前 25 名主要是时间序列和统计模型，而后来的四次竞赛主要是梯度提升决策树和神经网络。在接下来的章节中，我们将分别介绍六项竞赛的详细评审结果。

表 3：在六项 Kaggle 竞赛中均进入前 25 名的已审核解决方案概览。空白格表示没有解决方案的描述。文本为解决方案所使用方法的编码逗号分隔列表，其中多个值表示使用了一个集合：**GBDT**：梯度提升决策树；**DTF**：决策树森林（随机森林和极随机化树）；**TS**：时间序列方法（如指数平滑法、ARM）；**DTF**：决策树森林（随机森林和极随机化树）；**TS**：时间序列方法（如指数平滑法、ARM）。**NN**：神经网络；**LM**：线性回归；**STAT**：其他统计方法（多项式回归、投影回归、无观测成分模型、主成分回归和奇异值分解）；**ML**：**其他 ML** 方法（支持向量机和 K 近邻）。

放置	沃尔玛商店销售额	沃尔玛风暴天气	罗斯曼	维基百科	公司最爱	招聘餐厅
1	STAT, TS	STAT, LM	GBDT	NN	GBDT, NN	GBDT, NN
2	TS、STAT、DTF、ML、LM		GBDT	GBDT, NN, LM	NN	
3	TS	GBDT	NN	NN	GBDT, NN	
4	TS		GBDT	NN	NN	
5	TS	STAT		STAT	GBDT, NN	GBDT, NN
6	TS	LM、DTF、ML、TS		NN	GBDT, NN	
7				NN		GBDT
8	LM			TS	GBDT, NN	GBDT
9	LM					
10	GBDT, LM		GBDT			GBDT
11	TS	GBDT、DTF、ML、LM		NN, TS		GBDT、DTF、TS
12					GBDT, NN	NN
13	TS				GBDT, NN	
14				NN, TS		
15					GBDT, NN	
16	GBDT			NN, TS	GBDT, NN	
17					GBDT	
18					GBDT	GBDT
19		LM		GBDT, TS		
20						
21						GBDT, NN, TS
22						
23			GBDT, ML			GBDT, NN
24						
25						GBDT, NN

6.1 沃尔玛商店销售额预测（2014 年）

沃尔玛商店销售额预测竞赛是最古老的竞赛。比赛要求参赛者按部门和门店预测 1 至 39 周的每周销售额（单位：美元）。参赛者可以访问 45 家门店和 81 个部门的 33 个月数据，以及门店元数据、节假日信息、促销指标、每周气温、燃料价格、消费价格指数 (CPI) 和失业率。

在这次竞赛中，对季节性和节假日的准确建模变得至关重要，表现优异的解决方案主要采用传统的时间序列预测方法，并略作调整。获胜者的主要创新点是在全球范围内学习季节性和节假日模式，并利用这些模式对数据进行去噪处理。

单个时间序列。获胜者通过对每类时间序列（部门）进行截断奇异值分解（SVD）来实现这一目标，并以此来重建单个时间序列。截断的作用是去除数据中的低信号变化，从而有效地过滤掉噪音。然后，使用 STL 分解法结合指数平滑法和 ARIMA 等局部预测方法对去噪后的时间序列进行预测。最后，将这些方法的预测结果与简单预测模型（如季节性天真模型、线性趋势和季节性模型以及历史平均值）组合在一起。虽然所有的集合都提高了他的预测准确度，但他的集合中由 SVD、STL 和指数平滑法组成的单一模型的准确度足以单独赢得比赛。

所有进入八强的选手都使用了一个重要的调整方法，那就是调整数据，将每年的节假日排成一行，这样就可以使用时间序列模型将这些节假日作为季节模式的一部分来建模。ML 模型本身在比赛中的表现并不理想，主要是作为同样包含时间序列模型的集合的一部分使用。获得第二名的解决方案就使用了其中的一个模型，该方案对每个部门都使用了 ARIMA、非观测成分模型、随机森林、K-近邻、线性时间序列回归和主成分回归的组合。因此，这些模型在全局与局部维度上处于中间位置。有趣的是，这个相对复杂的模型组合并没有击败第一名的简单解决方案。

比赛中可用的外生变量，包括气温、燃料价格、消费价格指数、失业率和减价信息，并没有被证明有助于做出准确的预测。虽然前 10 名选手中的一些人使用了这些变量，但前两名和第四名却没有使用，这表明这些变量的附加值很小。其他有趣的参赛作品还包括因简单而获得第三名的作品。该作品将节假日排在一起，并使用了去年最接近的两个星期的加权平均值，并根据时间序列的增长率和温暖天数进行了调整。结果表明，这个简单的解决方案只比最佳解决方案差 4%。至于标准时间序列基准，简单的天真方法是一个很好的基准，但在所有前 10 名选手中，仍然被击败了 20% 以上。

6.2 沃尔玛在暴风雨天气中的销售情况（2015 年）

沃尔玛 "暴风雨天气下的销售 "竞赛的形式与其他竞赛略有不同，其目标是预测极端天气对销售的影响。比赛的任务是对总共 255 个时间序列的产品和商店的日单位销售额进行预测。比赛形式与其他比赛不同，不要求预测未来一段时间。相反，要求对极端天气事件发生前后±3 天的时间段进行预测，这已从可用数据中删除。因此，这项任务并不是纯粹意义上的预测任务，因为预测时段之后的观测数据是可用的。为了构建这些预测，参赛者可以访问 44 家商店和 111 种产品的 28 个月数据（其中已删除了一些极端天气事件）以及大量天气信息。

沃尔玛 "暴风骤雨 "竞赛的获胜者使用了零售预测软件常用方法的一种变体，即首先估算基线销售额，然后使用带有外生变量的线性回归建立偏离基线的模型。该解决方案使用投影追踪回归，仅使用时间作为输入，按时间序列估算基线销售额，同时考虑到趋势和潜在的年度季节性。使用 Vowpal Wabbit 库（Vowpal Wabbit，2020 年）建立了带有交互作用的全局 L1 规则化线性回归模型，对偏离基线的情况进行建模。因此，与零售预测软件的典型方法的主要区别在于使用了比常用移动平均值更复杂的平滑器，以及全局而非局部回归模型。该模型

获胜者[9]根据外生变量，包括周末/周日、节假日及其交互作用建模，以及时间信息（年、月、日和趋势）和 "黑色星期五 "建模（包括滞后和先导效应）。不出所料，解决方案中使用了天气数据，以指标变量的形式对降水量和偏离正常气温的阈值效应进行建模。不过，优胜者在撰写解决方案时提到，使用天气信息对预测结果的帮助并不大，这一点也得到了其他排名靠前的参赛者的证实。这一发现有些出人意料，因为比赛的目的是预测极端天气对销售额的影响，而实际天气可以代替预测。

排名靠前的选手还使用了其他几种方法，例如第 5 名的解决方案主要使用了日期特征，使用了局部高斯过程回归。第 3、第 6 和第 11 名选手成功地使用了各种 ML 模型的集合，如使用 XGBoost 算法（Chen & Guestrin，2016 年）的 GBDT、随机森林、SVM 和线性回归等模型。不过，有趣的是，这些在后来的 Kaggle 竞赛中通常表现出色的复杂模型组合都没有优于冠军的简单方法。这次比赛还首次使用了 XGBoost，虽然获得了第三名，但其表现并不突出。另一方面，传统时间序列模型在报告的解决方案中使用不多。排名第六的解决方案是个例外，它使用了时间序列模型（如 ARIMA）作为集合的一部分。不过，ARIMA 本身的性能并不出众，其性能比公开排行榜上的优胜方案低 17%。

6.3 罗斯曼商店销售额（2015 年）

罗斯曼商店销售竞赛的特点是全球 ML 模型集合的崛起，更具体地说是 XGBoost 的崛起。这也是神经网络首次进入前三名。比赛要求参赛者预测 1 到 48 天内各商店以美元为单位的日销售额。参赛者可以访问 1115 家商店 31 个月的数据，以及商店的元数据、促销指标、节假日信息、天气信息和谷歌趋势统计数据。

比赛的获胜者主要通过调整 XGBoost 模型，使其在时间序列上表现出色，从而超越了其他参赛者。这种调整包括利用时间序列和外生变量构建许多特征，以及利用脊回归模型进行趋势调整，以解决 GBDT 无法推断趋势的问题。在特征方面的主要创新包括在不同层次、不同周日和促销期计算的统计数据及其滚动版本。例如，按产品划分的平均销售额、按产品划分的销售额移动平均值，以及按产品和促销状态划分的平均销售额。此外，事件计数器也很有用。这些计数器包括节假日或促销等活动之前、期间和之后的天数。该解决方案还包括降水量和最高气温等形式的天气信息，以及月、年、月日、年周和年日等季节性指标，以便准确估计多种季节性影响。许多 ML 模型要想取得好成绩，关键在于选择适当的特征和超参数，以在不过度拟合训练数据集的情况下最大限度地提高准确性。许多参赛者采用的策略是使用与预测范围相同长度的暂留数据集来评估模型质量，并决定超参数和选择特征。多个 XGBoost 模型的集合使用比最佳单一模型的性能提高了约 5%。通过在不同的数据子集上训练模型，在集合中引入了变化

[9]在本综述中，我们用外生变量来指比赛中提供的原始信息，用特征来指可能经过处理的模型输入。

通过使用直接预测和迭代预测，以及在模型中包含不同的特征子集，可以建立不同的模型。

大多数表现优异的选手使用全局 XGBoost 模型集合来创建预测，但也有少数选手将局部 XGBoost 模型作为其集合的一部分。所使用的特征通常与优胜者类似，都包含事件计数器和在不同层次结构中计算出的统计数据。因此，与事件（即节假日和促销）相关的外生变量对于在比赛中取得优异成绩至关重要。这可能就是为什么与季节性天真基准相比，性能有了显著提高的原因。两个最佳解决方案的显著特点是，它们使用了移动平均数或中位数形式的滚动统计作为特征，从而调整和利用了时间序列预测文献中的著名方法。

获得第 3 名的解决方案首次在 Kaggle 预测竞赛中成功使用了神经网络。所使用的神经网络是一个全局全连接神经网络，它使用了比赛中提供的外生变量以及节假日和促销活动的事件计数器。时间序列方面主要通过使用季节性指标来处理。季节性指标和分类元数据使用分类嵌入建模，学习分类的向量表示并由网络用于预测。该解决方案不包括自回归输入，这在预测神经网络中很常见。更多详情请读者参阅参赛者发布的论文（Guo & Berkhahn, 2016）。

得分最高的简单方法是第 26 位使用了包含传统时间序列模型的混合方法。首先，使用局部 ARIMA（包含和不包含外生变量）和指数平滑模型进行预测。之后，使用全局 XGBoost 模型预测基于工作日、事件计数器、谷歌趋势模式和天气信息的残差，以捕捉时间序列模型无法充分模拟的外生变量的影响。因此，传统的时间序列模型在比赛中表现不佳，只能与全局 ML 模型一起使用，或者使用移动平均值来构建特征。获胜者以 11% 的优势击败了时间序列混合模型，并以 31% 的优势击败了由商店、工作日、年份和晋升状态中位数组成的简单基准，这清楚地表明，更复杂的模型在此次竞赛中产生了更好的解决方案。

6.4 维基百科网络流量预测（2017 年）

维基百科网络流量预测竞赛的规模更上一层楼，需要对超过 145.000 个时间序列进行预测。比赛还展示了深度学习在预测方面的威力，深度学习赢得了比赛，并在前八名中占据了六个席位。比赛要求参赛者预测维基百科 12 至 42 天内的每日页面访问量。参赛者可以访问 32 个月的页面访问数据以及维基百科页面的元数据。

获胜的解决方案提出了一种既优雅又准确的深度学习方法，而无需大量的特征工程。[10]这也是使用梯度提升决策树的解决方案的典型特点。该解决方案由具有相同结构的全局循环神经网络集合组成。由于神经网络预测可能会因噪声数据而不稳定，因此采用了多种集合方法来降低预测的方差。使用不同的随机种子训练了三个模型，以抵消网络权重初始化的随机性。为了避免对训练迭代的确切次数产生敏感性，我们采用了两种方法。首先，在训练过程中保存模型检查点，并使用检查点预测的平均值。其次，使用神经网络权重的移动平均值

[10]特征工程是指从外生变量或时间序列本身构建特征的过程。

而不是最终权重，这也被称为随机权重平均法（SWA）（Izmailov et al.）神经网络使用的特征包括历史页面浏览量和分类变量，如代理、国家、网站以及星期。递归神经网络的一个弱点是难以模拟长期依赖关系，例如每年的季节性。获胜者找到了解决这一问题的方法，他将一个季度、半年和一年前的页面浏览量作为模型的输入。此外，他还加入了滞后 365 和滞后 90 的自相关函数值，以便更好地模拟年度季节性。为了便于对季节性和时间动态进行交叉学习，对序列进行了独立缩放，但使用了页面浏览量中位数的缩放措施，以便让模型学习任何潜在的与缩放相关的模式。在使用贝叶斯优化算法 SMAC3（Lindauer 等人，2020 年）进行自动超参数调整算法的同时，还使用了暂缓验证集来决定神经网络的超参数。有趣的是，获胜者报告说，最终性能对超参数相对不敏感，该算法找到了几个性能相似的模型。

其他表现出色的选手使用了不同的神经网络架构，包括递归神经网络（RNN）、卷积神经网络（CNN）和前馈神经网络，这表明几种不同的架构可以提供类似的性能。同样，顶尖选手也使用了不同程度的特征工程。第 4 名和第 6 名的解决方案使用了有限的特征工程，而第 2 名则使用了广泛的特征工程，包括将各种集合模型的预测结果作为另一个模型的输入（在 ML 中称为堆叠）。由此看来，多种架构都能发挥作用，复杂的特征工程并不是使用该数据集进行高性能神经网络预测的必要条件。虽然神经网络在比赛中占据了主导地位，但第 8 名的另一个简单得多的解决方案也值得一提。这位选手使用了一种分段方法，其中包括卡尔曼滤波器来预测高信号序列，以及一种使用不同窗口移动中位数的稳健方法来预测低信号序列。该方案是前几名中唯一使用传统时间序列模型的方案，虽然表现出色，但仍比获胜方案差 6%。

6.5 Corporación Favorita 食品杂货销售预测（2018 年）

Corporación Favorita 食品杂货销售预测竞赛很好地展示了 Kaggle 社区如何借鉴和改进以往竞赛的解决方案，因为 *Rossmann* 竞赛中使用的梯度提升方法和*维基百科*竞赛中使用的神经网络方法都被顶尖选手大量采用。比赛要求参赛者对 210,000 多个时间序列按商店和产品预测 1 到 16 天的日单位销售额。参赛者可以访问 54 家商店和 3901 种产品的 55 个月数据，以及商店和产品的元数据、促销指标、节假日信息和油价。

获胜者使用了相对复杂的模型组合，包括梯度提升模型和神经网络模型。与以前的比赛相比，这次比赛的一个变化是使用了新的梯度提升库LightGBM（Ke等人，2017年），其速度明显更快，更容易尝试不同的特征和参数。解决方案中的一项创新是在每个预测期训练一个模型，而不是在所有预测期训练一个模型，以便让模型学习哪些信息对每个预测期有用。虽然取得了良好的效果，但这种方法需要 16 个模型，而不是 1 个模型。这种方法用于 LightGBM 模型以及与其他两个模型组合的前馈神经网络。这些模型包括另一个经过所有水平训练的 LightGBM 模型和在*维基百科*竞赛中获得第 6 名的 CNN 架构。前馈神经网络和 GBDT 模型中使用的特征与维基百科竞赛中成功使用的特征基本相似

*罗斯曼*竞赛。这些特征主要是按商店、项目、类别及其组合等各种因素分组的滚动统计数据。使用的统计数据包括中心性和传播度量，以及指数移动平均值。

有趣的是，获胜者只在模型中使用了最近的数据，而根据验证数据集的表现，选择放弃较早的观测数据。因此，尽管可以获得多个季节的数据，但最终的模型使用了不到一个完整季节的数据进行模型拟合，其形式为 1 个月、3 个月或 5 个月的数据。其他排名靠前的选手也倾向于采用这种方法，如排名第 5 和第 6 的选手。为什么这种方法在忽略年度季节性的情况下仍能奏效，一个可能的解释是数据中存在的趋势，以及只有 16 天的短预测期。

比赛中排名靠前的选手并没有采用简单的方法，他们都采用了类似的建模方法，包括基于滚动统计的特征工程 LightGBM、受维基百科比赛成功架构启发的神经网络或两者的组合。这些解决方案之间的主要区别在于特征工程和架构的细节，或所使用的验证方法。

虽然在之前的大多数比赛中都采用了保持策略来防止过度拟合，但也有几位参赛者尝试了其他验证方法。其中一个例子是第四名的解决方案，该方案保留了一定比例的时间序列，因此完全依靠交叉学习进行性能评估。另一种有趣的验证方法是将分组 K-Fold 交叉验证和时间序列交叉验证结合使用，前者用于估算参数，后者用于估算模型性能。在分组 K-Fold 交叉验证中，每个时间序列被限制为一个折叠，以避免跨折叠的信息泄漏，因此也纯粹依赖交叉学习。时间序列交叉验证使用两个连续 16 天的保留数据集来估计模型性能。尽管多种验证方法在预测中的成功应用很有趣，但由于前三名解决方案都使用了暂缓验证方法，因此暂缓验证方法似乎仍然足够。

6.6 招聘餐厅游客预测（2018 年）

Recruit 餐厅游客预测竞赛证实了之前使用的方法（如使用滚动统计的梯度提升决策树）以及在一定程度上使用的神经网络在不同领域取得了成功。比赛要求参赛者按餐厅预测 1 至 39 天内的每日餐厅访问量。参赛者可以访问 821 家餐厅 15 个月的数据，以及餐厅的元数据、节假日信息和提前不同时间访问餐厅的预订信息。

比赛的获胜者是一个由四名参赛者组成的团队，他们使用了由基于 LightGBM、XGBoost 和前馈神经网络的模型的平均值组成的集合。所有模型都使用了基于滚动统计的特征以及餐厅预订的滞后值，这是与以往不同领域竞赛的主要区别。比赛中的另一个挑战是，测试集包括了 "黄金周 "假期，而选手们在训练数据集中只能获得一个较早的假期，这使得比赛的行为有了很大的不同。一些参赛者发现了一种对数据进行智能调整的方法，即把节假日视为周六，把节假日的前一天和后一天分别视为周五和周一，从而利用现有的少量数据对这些节假日进行更好的建模。赛后，多位排名靠前的选手对这一调整进行了评估，结果表明，这一调整一般都能显著提高性能。正如第一名的解决方案所示，使用这一技巧并不是赢得比赛的必要条件。不过，它强调了使用领域知识和手动调整数据来实现最佳性能的价值，这与沃尔玛商店销售竞赛的结果类似。

第 1 名和第 5 名的解决方案使用了神经网络，但总体上不如之前的比赛那么成功，主要用于增加集合的多样性。在*维基百科*和 *Corporación Favorita* 比赛中成功使用的递归神经网络和卷积神经网络变体的表现一般略逊于基于增强决策树的模型。第 21、23 和 25 名使用这些方法的准确率比第 1 名低约 2%。造成这种情况的潜在原因可能是数据集的大小，它比*维基百科*和 *Corporación Favorita* 数据集小 100 倍以上。有趣的是，卡尔曼滤波器在*维基百科*比赛中也取得了第 33 名的好成绩，与第 1 名的成绩差距仅为 2.4%，这说明在有外生变量的情况下，传统的时间序列模型仍然是可行的。

与前几届比赛一样，大多数参赛者都使用了保留数据集来验证模型性能，但令人惊讶的是，第 7 名和第 8 名参赛者都使用了标准的 K 折验证方法，忽略了数据的时间序列性质，从而获得了较高的名次。受 Corporación Favorita 竞赛创新的启发，一些参赛者训练了地平线特定模型，第 1 名和第 5 名提交的一个模型共需要 42 个模型。第 11 名参赛者采取了折中办法，每周训练一个模型，总共训练了 6 个模型，以便仍能模拟一些潜在的地平线特定效应。然而，并不是所有前几名的参赛者都使用了地平线特定模型，这表明与模型数量的增长相比，该方法的性能改进可能并不大。

7 讨论

在一般建模策略方面，我们的回顾支持了 M4 竞赛中关于集合模型与单一模型以及全局模型与局部模型的研究结果。集合模型在所有比赛中都获胜了，因此这一结论在不同领域和预测任务中依然有效。所有比赛的获胜者也都使用了全局模型，尽管有时是与局部模型结合使用，这凸显了时间序列交叉学习的优势。全局模型和本地模型在*沃尔玛商店销售比赛*等基准赛中的表现差异表明，进入业务层次结构所带来的交叉学习优势甚至高于 M4 比赛中发现的优势。

在一些竞赛中，获取等级制度以外的外生变量带来了很大的好处，而在另一些竞赛中，获取这些变量带来的好处非常小，甚至没有任何好处。有关促销、节假日和活动的信息在大多数竞赛中都非常有用。需要预测的变量，即天气和宏观经济变量，尽管可以提供实际值，但似乎并没有带来明显的好处。

在对六项比赛的审查中，我们没有发现一种方法在所有比赛中都占据主导地位。最早的两项比赛--*沃尔玛商店销售额和沃尔玛暴风雨天气*，分别是通过创新使用时间序列和统计方法赢得的。后来的四项比赛则由利用滚动和分组统计的梯度提升决策树或神经网络等非传统预测方法胜出。此外，在各项比赛中，表现最出色的解决方案的结构惊人地相似。因此，一个有趣的问题是，为什么梯度提升决策树或神经网络在前两次比赛中表现不佳？一个显而易见的原因是，这些方法在前两次比赛时还不成熟，甚至还没有开发出来。在*罗斯曼竞赛*之前，神经网络还没有被用于预测，而第一个成功的梯度提升决策树算法 XGBoost 也是在第一届沃尔玛竞赛之后才发布的。虽然 XGBoost 算法在*沃尔玛风暴天气竞赛*中得到了应用，但该方法仍是新方法，对时间序列领域的调整，如滚动统计和分组统计等，仍有待改进。

没有使用。因此，我们无法回答，如果今天举行这些比赛，时间序列和统计方法是否还能获胜。

一个更好的问题可能是，为什么时间序列和统计方法在最近四次竞赛中表现不佳？我们认为，原因在于最近四个竞赛数据集的特点更适合梯度提升决策树和神经网络。最新的四个数据集都具有间歇性的特点，并且包含与预测任务相关的外部信息，如等级信息和预测性外生变量，如节假日、活动、促销和预订等。另一方面，*沃尔玛商店销售*竞争是连续的，可以获取层次信息，外生变量包含的有用信息很少，这可能是由于高度聚合的缘故。总之，这为全局时间序列方法提供了理想的条件。*沃尔玛"暴风骤雨"*比赛的数据集最小，而且缺少业务层次信息，这限制了交叉学习的机会。此外，所需的预测期前后数据的可用性和较短的预测期使得这种情况非常适合统计平滑方法，如获胜者所采用的投影追求回归法。因此，我们发现，对于间歇性或包含相关外部信息的分类数据集，ML 方法优于时间序列方法和统计方法。这与谷歌（Fry & Brundage, 2019）和亚马逊（Salinas et al.）

至于梯度提升决策树和神经网络之间的差异，我们注意到，在*维基百科竞赛*中，神经网络的表现优于梯度提升决策树，因为*维基百科竞赛的*规模非常大，而且不包含任何有用的外生变量。在其他三项最新竞赛中，两种方法都名列前茅。神经网络是当前许多预测研究的主题，M4 竞赛的获胜者也使用了神经网络（Smyl，2020 年）。不过，我们还不知道有哪些研究将梯度提升决策树与 Kaggle 竞赛中的策略结合起来使用。

虽然 M4 竞赛的第二名解决方案使用了梯度提升决策树，但它被用作元学习器，结合了传统的时间序列预测方法（Montero-Manso 等人，2020 年）。

因此，考虑到梯度提升决策树在比赛中的出色实证表现以及其在预测方面的一些有用特性，进一步的研究应探讨如何将其用于预测。由于该方法以决策树为基础，因此它可以通过沿时间维度进行分区，学会有效处理样本中的层次变化。通过使用滚动和分组统计数据对业务层次进行编码，它可以通过对这些统计数据进行分区来交叉学习，从而汇集来自相似时间序列的信息。此外，需要优化的损失函数可以自定义为任何具有明确梯度和赫斯的函数，例如预测预测区间所需的量化损失。梯度提升决策树的主要弱点在于推断趋势。不过，Kaggle 的参赛者们已经开发出了解决这一问题的方法，例如，与建立趋势模型的线性回归进行集合。

在所有六项比赛中，我们发现都成功地使用了一个长度等于预测范围的暂留数据集来验证模型性能和防止过拟合。令人惊讶的是，当我们使用验证集对 ML 模型进行多次评估以选择特征和超参数时，并没有发现验证集存在严重的过度拟合现象。一个可能的解释是 Kaggle 平台提供的公开排行榜反馈，因为排行榜上的成绩下降会告知参赛者，他们很可能过度拟合了验证集。因此，这种方法原则上相当于在进行预测竞赛时将数据分成四份：

- 用于估计模型的训练集、
- 用于评估模型性能和进行模型诊断的验证集

- 第二个小型验证集只提供简要性能指标，以防止过度拟合。
- 用于评估样本外性能的最终测试集。

进一步的研究应评估这种方法与其他既定预测验证策略（如时间序列交叉验证）的比较。

7.1 实际适用性

对于更复杂的方法，人们经常提出的一个问题是其实际适用性，以及提高的准确性是否能证明增加的复杂性和计算要求是合理的（Gilliland，2020 年）。最近四次 Kaggle 竞赛的解决方案都采用了数据驱动的 ML 解决方案，需要训练多个复杂模型。因此，就成本和时间而言，它们比流行的时间序列基准更为昂贵。增加的计算成本和时间是否合理，最终将取决于预测所支持的决策成本结构，因此不能一概而论。Gilliland (2020) 认为，考虑到 ML 方法的复杂性和成本，ML 方法在 M4 竞赛中获得的准确性提高并不足以证明在实践中使用 ML 是合理的。在我们的回顾中，我们发现顶级解决方案通常比简单基准（如季节性天真方法）有相当大的改进，大约在 20% 到 74% 之间。因此，在每日和每周的业务预测任务中，应认真考虑使用能有效利用业务层次和外生变量的更复杂方法。不过，获胜方案不太可能在准确性和复杂性之间实现最佳权衡。*沃尔玛商店销售额*和 *Rossmann* 比赛的优胜者使用了集合方法，略微提高了准确性。不过，就模型管理复杂性和计算成本而言，他们的最佳单一模型在实际操作中可能更实用。

正如 Januschowski 等人（2020 年）所指出的，在讨论中还应考虑与运行预测系统相关的人工成本，如数据清理和预测调整。举例来说，零售业通常使用简单的时间序列方法，并依靠人工判断来估计促销活动的影响并调整预测，因为简单的方法无法对此进行准确建模（Fildes 等人，2019 年）。简单方法的计算成本较低，但也需要大量人力，而使用促销信息的更复杂方法可能只需要人力进行异常管理，例如在出现新价位或促销策略时。进一步的研究应着眼于评估这一权衡维度，例如，调查现实生活中使用简单和 ML 方法的预测系统中的人工使用情况。

7.2 M5 假设

即将举行的 M5 竞赛采用了沃尔玛慷慨提供的零售领域分层数据集。比赛要求在商店和产品层面对每天 40,000 多个时间序列进行预测，并为参赛者提供有关价格、促销、活动和产品等级的信息。因此，预测任务与 Corporación Favorita 竞赛非常相似。与上述竞赛的主要区别在于，除了预测准确性之外，还对预测的不确定性进行了评估。根据我们的研究，我们提出了以下假设：

- M5 竞赛中时间序列的实例空间表示将与 Corporación Favorita 竞赛相似，这意味着与以往的 M 级竞赛相比，熵更高、趋势和一阶自相关性更低。
- 获胜的方法将利用交叉学习，全局模型和混合模型将主导局部模型。

- 与 M4 竞争车型相比，获取等级信息将拉大本地车型与使用交叉学习技术车型之间的性能差距。
- 使用基于滚动统计和神经网络等特征工程的 GBDT 将在竞赛中表现出色，并在准确性和不确定性方面优于现有的时间序列基准。
- 为了提供预测区间，GBDT 和神经网络将通过使用自定义损失函数（如量化损失）或根据输出分布调整训练程序/架构来进行调整，这也是近期许多研究的主题（例如，GBDT 参见 Duan 等人 (2020)，神经网络参见 Salinas 等人 (2019)）。
- 与所有 Kaggle 和 M 竞赛的结果一致，方法组合将继续占据前几名的位置。我们预计这些方法组合将同时包含神经网络和 GBDT，并有可能与其他方法相结合。
- 排名靠前者将使用保留数据集或时间序列交叉验证，以避免过度拟合。
- 使用价格、促销、节假日和其他事件等外生变量将提高预测准确性，这与之前的零售研究（Fildes et al.
- 参赛者将开发创新战略来应对分层预测的挑战，我们期待新的神经网络架构和 GBDT 战略能够最佳地利用这些信息。

7.3 局限性

与学术预测竞赛相比，Kaggle 专注于为现实生活中的预测任务提供解决方案，但这也有一个缺点，那就是限制了研究人员从竞赛中得出的结论。比赛结束后无法访问测试集，这意味着无法测试解决方案性能之间的显著差异，也无法使用其他误差测量方法评估性能。此外，我们也无法分析不同解决方案在不同数据集子集上的表现，从而加深对各种方法优缺点的理解。

虽然 Kaggle 鼓励共享解决方案，但并不要求参赛者公开共享其解决方案或代码，而缺乏公开共享的解决方案会影响我们评审的有效性。在前 25 名未报告的解决方案中使用线性回归或局部时间序列模型等方法有可能会改变我们的结果。然而，我们发现本地时间序列在最近的四项比赛中表现出竞争力的可能性很小，我们的依据是数据集的间歇性和外生变量的影响。我们的基准测试结果也证明了这一点。我们还发现，由于不同求解方法的分享意愿不同而导致的系统性报告偏差不太可能存在。尽管存在这些不足之处，但我们仍然相信，通过关注各项竞赛的有效模式，并将研究结果与数据集特征联系起来，我们可以学到很多东西。进一步的研究应该在各种数据集上对我们的假设进行测试，而即将举行的 M5 竞赛必将成为一个很好的初步测试场所。

8 结论

根据我们对最近六次 Kaggle 预测比赛的分析和回顾，我们认为，在预测每日和每周商业时间序列方面，预测界有很多地方需要向 Kaggle 界学习。

在我们的分析中，我们发现 M4 竞赛数据集包含与 Kaggle 竞赛类似的时间序列，尽管具有这些特征的时间序列在 M4 竞赛数据集中所占比例较低。此外，Kaggle 数据集与 M4 竞赛的不同之处在于，它们提供了外部信息（如外生变量或业务层次结构）的访问权限，从而显著提高了预测准确性。

与 M4 竞赛的结果类似，我们发现全局集合模型的表现优于局部单一模型。与 M4 和之前的两次 Kaggle 竞赛相比，在最新的四次 Kaggle 竞赛中，传统的时间序列和统计方法的表现明显优于机器学习方法。我们认为，这可归因于机器学习方法利用外部信息进行交叉学习，并对外部因素的影响进行建模。此外，我们发现 Kaggle 竞赛中的顶级解决方案与 M4 竞赛中的前两个解决方案有相似之处，它们都依赖于梯度提升决策树或神经网络。不过，要想从机器学习方法中获得性能优势，必须对机器学习方法及其验证策略进行一些调整。

我们强烈鼓励预测界学习时间序列预测的机器学习策略，并参与其进一步发展。M5 竞赛为此提供了一个理想的机会，因为预测任务和数据集与本文回顾的一些 Kaggle 竞赛高度相似。因此，我们相信从本文讨论的 Kaggle 竞赛中学到的知识将预示着 M5 竞赛的结果。

参考资料

Athanasopoulos, G., & Hyndman, R. J. (2011)。预测竞赛中的反馈价值。*国际预测期刊》，27*（3），845-849。https://doi.org/10.1016/j.ijforecast.2011.03.002。

Chen, T., & Guestrin, C. (2016)。*XGBoost：可扩展的树状助推系统》*。785–794. https://doi.org/10.1145/2939672.2939785

Darin, S. G., & Stellwagen, E. (2020).预测 M4 比赛的每周数据：Forecast Pro 的制胜之道。*国际预测期刊》，36*（1），135-141。https://doi.org/10.1016/j.ijforecast.2019.03.018。

Dowle, M., & Srinivasan, A. (2019)：*data.*frame "*的扩展*。https://cran.r- project.org/package=data.table

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A. Y., & Schuler, A. (2020).NGBoost：Natural Gradient Boosting for Probabilistic Prediction.*ArXiv:1910.03225 Cs, Stat].* http://arxiv.org/abs/1910.03225

Fildes, R., Ma, S., & Kolassa, S. (2019)。零售预测：研究与实践。*国际预测期刊》*，S016920701930192。https://doi.org/10.1016/j.ijforecast.2019.06.004。

Fildes, R., & Ord, K. (2007).*Forecasting Competitions：Their Role in Improving Forecasting Practice and Research* (pp. 322-353).John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470996430.ch15

Fry, C., & Brundage, M. (2019).*M4 预测竞赛--从业者的视角》*。

Gilliland, M. (2011).增值分析：商业预测的有效性。载于《*分析》杂志*。http://analytics-magazine.org/value-added-analysis-business-forecasting-effectiveness/

Gilliland, M. (2020).机器学习方法在预测中的增值。*国际期刊*

https://doi.org/10.1016/j.ijforecast.2019.04.016

Goerg, G. (2013)。可预测成分分析。*机器学习国际会议*，64-72。Guo, C., & Berkhahn, F. (2016)。分类变量的实体嵌入。*ArXiv:1604.06737 Cs]*.
http://arxiv.org/abs/1604.06737

Hyndman, R. J. (2020).预测竞赛简史。*International Journal of Forecasting*, *36*(1), 7-14.
https://doi.org/10.1016/j.ijforecast.2019.03.015

Hyndman, R. J., & Koehler, A. B. (2006).另一种预测准确性测量方法。*International Journal of Forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2019)。平均权重带来更宽的最佳值和更好的泛化。*ArXiv:1803.05407 Cs, Stat*]. http://arxiv.org/abs/1803.05407

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020)。预测方法的分类标准。*国际预测期刊》*，*36*（1），167-177。
https://doi.org/10.1016/j.ijforecast.2019.05.008。

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017)。使用时间序列实例空间可视化预测算法性能。*国际预测期刊》*，*33*（2），345-358。

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y.(2017).*Lightgbm：高效梯度提升决策树*。3146-3154.

Lindauer, M., Eggensperger, K., Feurer, M., Falkner, S., Biedenkapp, A., & Hutter, F. (2020)。*SMAC v3：Python 中的算法配置*。https://github.com/automl/SMAC3

M 开放预测中心。(2019).https://mofc.unic.ac.cy/m5-competition/ 。
访问日期：2020 年 2 月 19 日

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a)。Responses to discussions and commentaries. *国际预测期刊》*，*36*（1），217-223。https://doi.org/10.1016/j.ijforecast.2019.05.002。

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020b).M4 竞赛：100,000 个时间序列和 61 种预测方法。*International Journal of Forecasting*, *36*(1), 54-74.
https://doi.org/10.1016/j.ijforecast.2019.04.014

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020).FFORMA：基于特征的预测模型平均。*International Journal of Forecasting*, *36*(1), 86-92.
https://doi.org/10.1016/j.ijforecast.2019.02.011

O'Hara-Wild, M., Hyndman, R., & Wang, E. (2019).*盛宴：时间序列的特征提取与统计*。https://cran.r-project.org/package=feasts。

R 核心团队。(2019).*R：统计计算的语言和环境*。R Foundation for Statistical Computing. https://www.r-project.org/

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2019).DeepAR：利用自回归递归网络进行概率预测。*International Journal of Forecasting.* https://doi.org/10.1016/j.ijforecast.2019.07.001

Smyl, S. (2020).用于时间序列预测的指数平滑和递归神经网络混合方法。*International Journal of Forecasting*, *36*(1), 75-85. https://doi.org/10.1016/j.ijforecast.2019.03.017

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020)。预测竞赛数据是否代表现实？*国际预测期刊*》，*36*（1），37-53。

Vowpal Wabbit.(2020).*Vowpal* Wabbit. https://vowpalwabbit.org/ 。访问日期：2020 年 2 月 19 日。

Wickham, H. (2016)：*Elegant Graphics for Data Analysis*.Springer-Verlag New York. https://ggplot2.tidyverse.org

附录 A：Kaggle 竞赛数据集的预处理

事先应用了几个预定义的滤波器，以实现一些所需的特性，主要涉及 M4 序列的长度以及每个频率和域的比例（Spiliotis 等人，2019 年）。[11].例如，少于 10 个观测值或三个周期的序列以及缺失值的序列都被排除在外（Spiliotis 等人，2019 年）。[12].相反，Kaggle 竞赛数据集中的时间序列没有应用预定义的过滤器，因此需要进行预处理，使序列适合分析。所有预处理均使用 R 软件包 **data.table**[13]和 **base**[14]软件包进行，可归纳为五个步骤：

1. 将 NA 或负值设为零。
2. 删除所有零值的时间序列。
3. 如果有测试集，则只保留在训练集和测试集中都出现的组别
4. 从不规则时间序列转换为规则时间序列。
5. 删除前导零和尾零

一些 Kaggle 数据集包含 NA 值或负值，由于不知道如何处理这些值，因此决定用零代替，从而与最高分作品中使用的方法保持一致。下一步是检查时间序列的所有值是否都为零，如果都为零，则从进一步分析中剔除。在*沃尔玛 "暴风骤雨 "*竞赛中，剔除所有零序列尤为重要，该竞赛最初包括 4,995 个时间序列，其中 4,740 个全部为零，因此只保留了 255 个时间序列。

除 *Recruit Restaurant* 和*维基百科*外，所有比赛都有测试集，其中包括评估期间的分组和特征。因此，为了限制分析的范围，我们决定对训练集进行子集化，只包含测试集中也存在的分组。*沃尔玛 "暴风雪 "*比赛是一个特殊的案例，因为该比赛的目标是预测重大天气事件发生前后 3 天的单位销售额。与其他 Kaggle 竞赛不同的是，其他竞赛的测试集都是按照训练期的时间顺序排列的，而本次竞赛的天气事件则是在整个训练期内零星发生的，而且时间长短不一。因此，在训练集中，测试集中所有产品/商店/日期的单位销售额都被设为 NA，然后使用 R 软件包 **imputeTS** 中基于插值的季节性分割估算算法进行估算。[15].为确保估算的正确性，使用 R 软件包 **ggplot2** 对所有估算序列进行了可视化和人工检查。[16].

除了组别之外，所有 Kaggle 竞赛都是根据索引（即日期）组织的。因此，我们可以分析时间序列的规律性，即在整个时间序列中，观测值之间的间隔是否一致。分析表明，一些不规则的现象通过填补缺失的日期和将预测单位设为零而得到缓解。例如，在 *Rossmann* 竞争中，当商店关闭时，观测值缺失。在这里，最一致的不规则性是区分周日休息和不休息的商店，这意味着一周的长度可以是六天或七天。因此，缺失的周日用单位销售额设为零来填补

---

[11]预测竞赛数据是否代表现实?

[12]预测竞赛数据是否代表现实?

[13]https://cran.r-project.org/web/packages/data.table/index.html

[14] https://stat.ethz.ch/R-manual/R-devel/library/base/html/base-package.html
[15] https://cran.r-project.org/web/packages/imputeTS/index.html
[16] https://cran.r-project.org/web/packages/ggplot2/index.html

以纠正这种不规则现象，从而提供一致的 7 天周长。最后，Kaggle 竞赛中的多个时间序列要么以连续几个零开始，这些都被删除了。

9 附录 B：更改和排除*季节性时段 (F4)*。

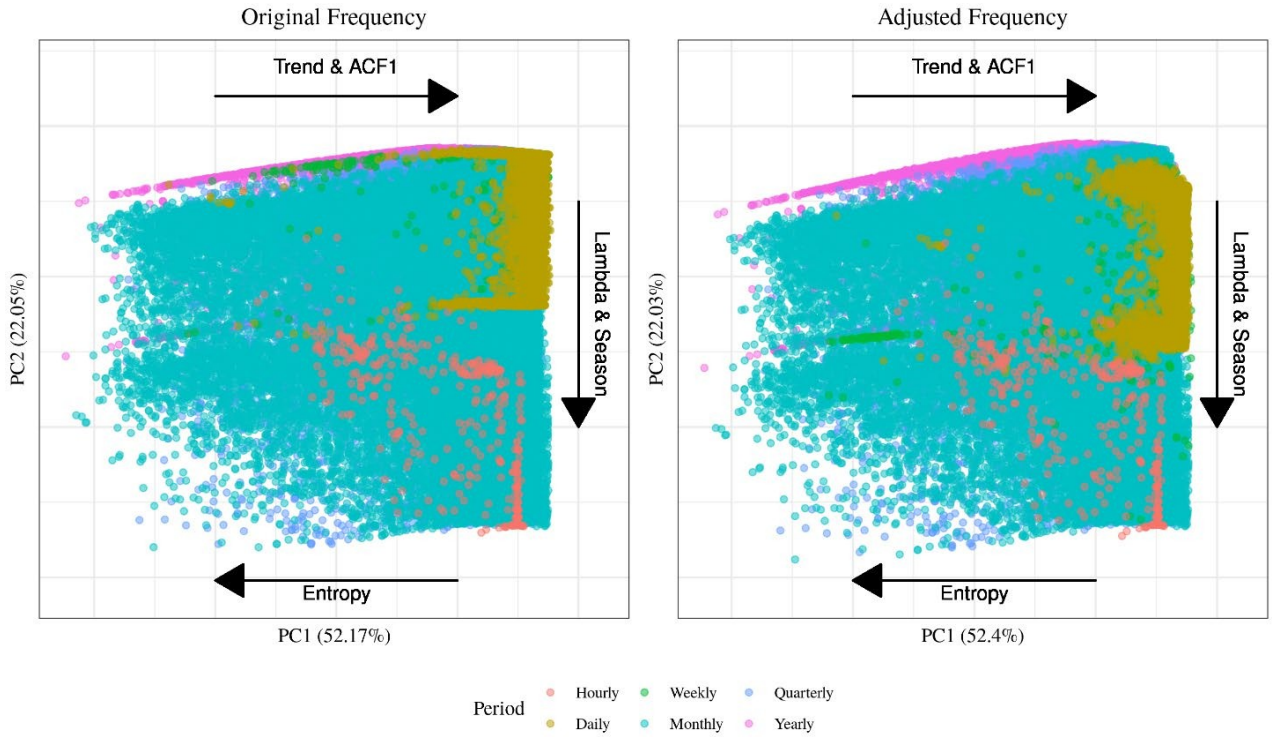

图 1：带有原始频率和调整频率的 M4 特征。

附录 C：基准制定程序详情

查看出版统计