基于 GBDT-Prophet 的产品需求量预测研究

摘 要

如何准确高效地预测产品需求是企业长久以来重点关注的问题。针对本次赛题给出的时间序列数据中显著存在的季节性因素，本文利用 **Prophet 模型**提取历史销售数据的**季节趋势特征**，并融合 XGBoost、LightGBM 和 CatBoost 三种**梯度提升树算法**（Gradient Boosting Decision Tree，GBDT），对产品进行**商品分层**处理后，最终提出**基于 GBDT-Prophet 的产品需求量预测模型**，采用 WMAPE 及 Tweedie 损失函数作为模型评价指标，通过网格搜索调整参数对模型进行优化。

针对赛题要求，**本文主要进行了两部分的工作**：一是对产品的历史销售数据进行探索性分析，得到各因素对产品订单需求量的影响程度及各因素影响下产品需求的特性；二是通过建立基于 GBDT-Prophet 的产品需求量预测模型对产品未来三个月的月需求量进行预测。

针对问题一所进行数据探索和数据分析：本文利用 Python 的 Matplotlib、Seaborn 等可视化库，绘制折线图、散点图、箱线图、气泡图等对示例数据进行可视化处理和分析，得出各因素对产品需求的影响及不同因素下的产品需求特性。

针对问题二（1）所进行数据处理和特征工程中：本文依据对训练数据集的探索性分析以及对预测集的产品分布分析所得到的结论，根据数据集中的重要规律进行相应预处理和特征量化，对产品进行**商品分层**这一特殊处理，这也是**本文的创新点之一**；特征工程部分，本文充分利用数据探索阶段的结论，并依据时序数据的特点引入了滞后特征、趋势特征等重要特征。

针对问题二（2）所进行的产品需求预测模型建立和求解过程中：本文通过基于 GBDT-Prophet 的产品需求量预测模型，利用 Tweedie 损失函数进行评估。建立**新品预测模型**以及**误差分析**等处理使模型的准确性得到较大提升，**这是本文的创新点之二**。其中建立新品模型后测试集 Tweedie deviance 提升 24.68%，误差分析并将结果覆盖提交后测试集 Tweedie deviance 提升 21.10%。

本文通过**模型融合**并对不同预测精度进行对比，最终**采用随机森林算法进行模型融合并按月精度对月需求量进行预测的效果最佳，其 Tweedie deviance 结果为 11.85**，其拟合效果优于单一预测模型。

关键词：需求预测；Prophet；梯度提升树；误差分析；多模型融合；

目 录

第一章 绪论

1.1 问题背景

近年来企业外部复杂多变的环境以及不断加剧的市场竞争，让企业供应链面临着较多难题。这些难题的核心，是企业应该如何应对供应链上各个环节的不确定性。在现代供应链理论中，需求预测是企业供应链应对不确定性的第一道防线，也是所有供应链规划的基础。供应链中所有推动流程根据对顾客需求的预测进行，而所有拉动流程根据对市场需求的响应进行，无论是推动式供应链还是拉动式供应链，其首要的工作都是要求对未来的商品需求量进行尽可能精准的预测。

需求预测的研究背景可以追溯到 20 世纪 60 年代，当时许多企业开始利用计算机处理销售数据以预测未来需求。随着计算机技术的发展和数据采集能力的提高，需求预测的方法和技术也在不断发展和改进。信息化的浪潮为企业的管理与决策带来智能化的推动和指导[1]。基于历史销售数据依靠智能化手段进行有价值的信息提取，并挖掘出潜在的规律和模式[2]，进而对产品需求进行预测已成为各企业提高其竞争力的重要方式。

1.2 研究意义

需求预测在企业的经营和决策中都具有重要的意义和价值：

（1）**优化资源配置**。需求预测基于历史数据预判未来并得出结论，帮助公司管理层对未来的销售及运营计划、资金预算等作出决策参考，优化物流、员工等资源安排，提高生产效率和效益。

（2）**制定销售策略**。需求预测可以帮助企业及时地掌握市场变化趋势，同时根据预测结果针对未来的市场需求制定出合理、有效的销售策略以及产品定价策略，提高盈利水平。

（3）**降低运营成本**。如果企业高层的重要决策只能依据经验而行，往往会导致缺乏对市场需求的了解，产生库存和资金的积压或不足等问题，增加企业库存成本。准确的需求预测有助于企业在获取利润与降低损失之间取得平衡，降低经营成本，从而降低企业运营风险，有利于企业实现长期而稳健的经营和发展。

（4）**提高企业竞争力**。准确的需求预测有助于企业快速响应市场需求，同时提高产品质量和服务水平，有效提升客户满意度，进而提高企业自身的竞争力。

理论和应用实践的不断发展使得需求预测的准确性在逐步提高。需求预测大多数以时间序列进行，其最初采用的定量研究方法是传统的基于统计学的预测方法；随着人工智能技术不断革新，基于机器学习的预测方法开始被广泛运用于商品的需求预测研究之中[3]。但单一的预测模型总有各自的适应条件和优缺点，同时，产品需求往往受到多种不定因素的影响，导致需求预测的准确性普遍较低。为了有效改善需求预测的结果，需要利用更加优秀的算法解决问题。大量实证研究表明，将多种模型加以适当组合进行预测，能够充分发挥每个模型的优势，从而有效应对时间序列不平稳的难题[4]，达到提升模型整体性能的效果。**本研究通过 Prophet 模型提取季节性趋势作为特征，使用 XGBoost、LightGBM 和 CatBoost 三种梯度提升模型（GBDT）进行预测模型融合。**经过参数调整，最终达到最优预测效果。

1.3 问题重述

问题一：根据问题一的要求，需要进行针对产品历史订单数据的探索性分析。本次赛题的主要任务在于利用数学建模方法，通过分析历史订单数据并依据分析结果来预测产品的订单需求量。因此，首要任务是对训练样本进行探索性分析，运用有效的可视化方法来分析数据特征，以获取不同因素对产品需求的影响，并为问题二中的预测模型建立提供必要的支持。

问题二：根据问题二的要求，需要基于训练数据集建立数学模型，对预测数据中给出的产品，预测起未来 3 个月（即 2019 年 1 月、2 月、3 月）的月需求量。根据任务一中所探索到的训练数据的分布、趋势和周期性等特征，选择适合的数学模型，如回归模型、时间序列模型等，建立预测模型。然后使用建立的数学模型，分别按天、周、月的时间粒度对未来 3 个月的月需求量进行预测；评估不同预测粒度的预测结果，比较预测结果与实际结果的误差，并分析不同时间粒度对预测精度的影响。

1.4 关键流程

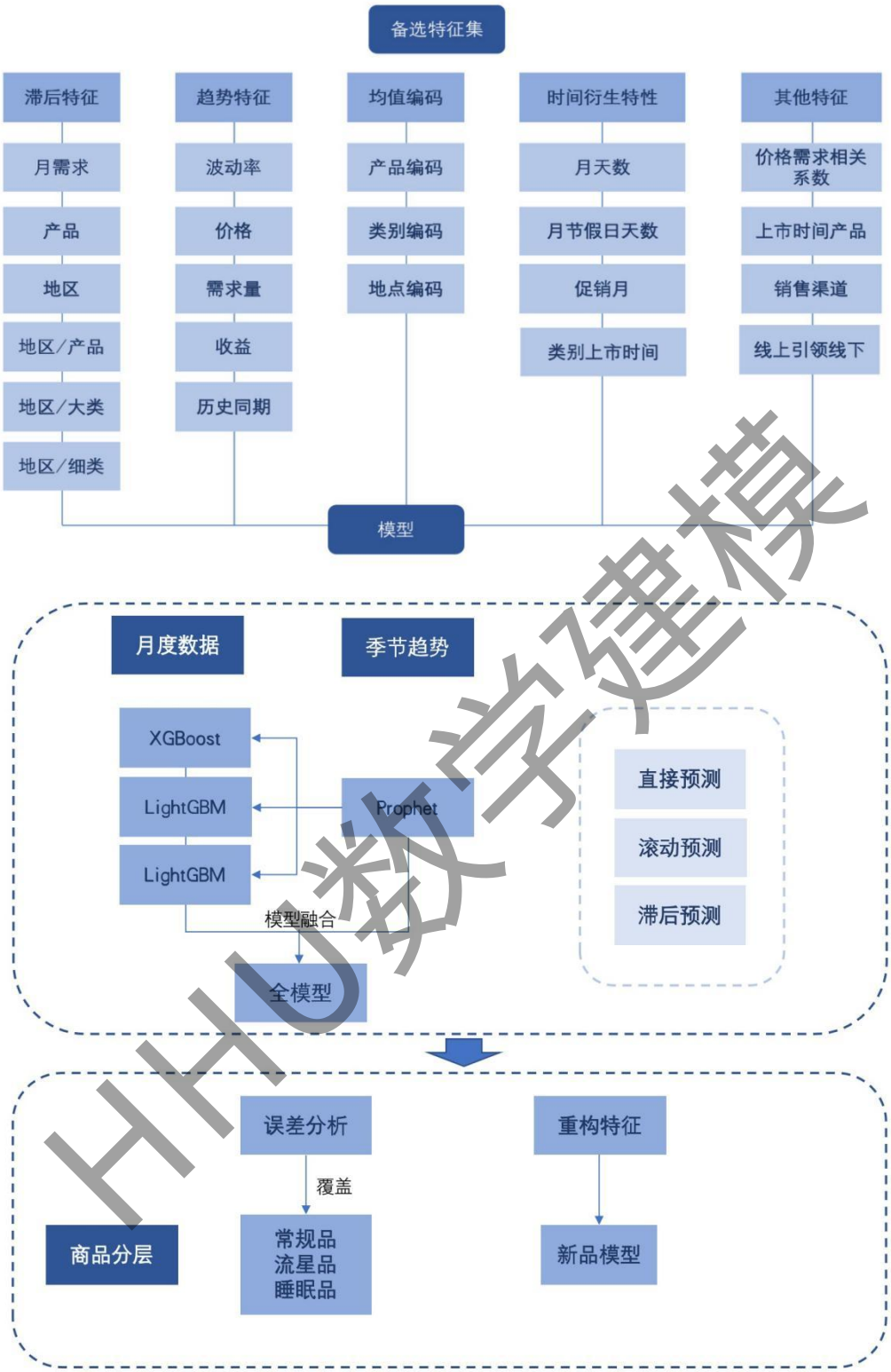


图 1 针对问题二的产品月需求预测模型建立与求解流程图

第二章 相关理论

2.1 机器学习

机器学习是研究如何使用计算机模拟或实现人类学习活动以获取新的知识或技能的科学[5]，是当前人工智能及模式识别领域的共同研究热点。机器学习模型通过对研究问题进行模型假设，然后输入海量训练数据对模型进行训练，使模型掌握历史数据所蕴含的潜在规律，并将规律应用至未知数据中进行预测和分析。常用于销量预测的机器学习方法包括人工神经网络、极限学习机、支持向量机和集成算法等。而针对时序特点，其中 XGBoost、LightGBM 和 CatBoost 是较为主流的应用方法。

2.1.1 XGBoost

极值梯度提升树（XGBoost）是梯度提升算法（GBDT）的一种高效且有效的实现。它的原理是不断地向集成模型中添加树，通过特征分裂来生长树模型，直到无法对模型再进行改进为止。每次添加树的本质都是在学习一个新函数去拟合之前所有树预测和的残差。XGBoost 在优化损失函数时运用二阶泰勒展开式来近似损失函数，求得模型最优解的效率更高，模型训练速度也更快；同时，它还在优化目标函数中显式地加入了正则项来限制模型复杂度，有效防止过拟合，从而提高模型的泛化能力。

XGBoost 在进行影响因素维度较多的销量预测时具有独特优势，且准确率更高[6]，适用于本文所要进行的基于历史需求数据的产品需求预测。

2.1.2 LightGBM

轻量级梯度提升机器学习（LightGBM）是根据 GBDT 算法衍生出来的另一种算法框架。LightGBM 通过引入基于梯度的 one-side 采样和互斥特征捆绑，同时在计算上主要采用直方图算法，使得该算法在处理具有高维特征的海量数据时，相较于同类型的 XGBoost 和 SGB 等树模型算法，在速度和内存消耗方等方面都占据明显优势。

2.1.3 CatBoost

CatBoost 是继 XGBoost、LightGBM 之后第三个基于 GBDT 改进的算法。与 XGBoost 和 LightGBM 不同，CatBoost 构建对称（平衡）树作为基模型，能够高

效、合理地处理类别型特征，解决梯度偏差以及预测偏差的问题，有效地减少了过拟合的发生，进而提高模型的准确性和泛化能力。

2.2 时间序列预测

2.2.1 Prophet 模型

Prophet 是 2017 年由 Facebook 提出的一种基于可加性模型解决大规模时间序列预测问题的实用方法。Prophet 模型将时间序列分解为趋势项、季节项、节假日项和误差项，对各项分别进行拟合后通过累加得到最终预测值。该模型拟合速度快，适应性、解释性强，对于时间序列中缺失数据、趋势变化以及突发事件等都具有更好的拟合效果，适用于处理内在规律较为明显的商业行为数据、具有强烈季节效应的时间序列数据以及有持续性历史趋势的数据[7]。

第三章 基于历史销售数据的探索性分析

探索性分析（Exploratory Data Analysis，EDA）是指通过绘制图表和进行统计分析等方式，探索数据中可能存在的结构和规律，而尽量不加入先验假设。在针对问题一提出的各项分析任务中，本部分将对示例数据进行探索性数据分析，主要包括数据准备、描述性统计和可视化等处理过程，以得出各因素对订单需求量的不同影响程度和特性等相关结论，为后续特征提取和构建模型打下良好基础。

3.1 数据准备

3.1.1 识别缺失值

利用示例数据创建 pandas 数据框，并使用 isnull()函数检查是否存在缺失值。经查验结果显示：数据集中没有缺失数据存在的情况。

3.1.2 数据描述

示例数据共计 597693 条，包含的数据类型有：float 类型、object 类型以及整型。查看示例数据中数值型特征的基本统计信息，如表 1 所示：

表 1 示例数据中数值变量描述统计结果

	item_price（产品价格）	ord_qty（订单需求量）
非缺失值数量	597694	597694
平均值	1076.242	91.651
标准差	1167.511	199.843
最小值	1	1
下四分位值	598	10
中位数	883	29
上四分位值	1291	101
最大值	260014	16308

3.2 数据探索和数据可视化

3.2.1 问题一（1）：不同价格对需求量的影响

以产品价格为 x 轴、订单需求量为 y 轴绘制散点图，可以观察到大部分产品的单价都在 50000 以内，订单需求量在 10000 以内。为了更方便地观察散点分布，将横轴设置为 0-20000，纵轴设置为 0-10000。

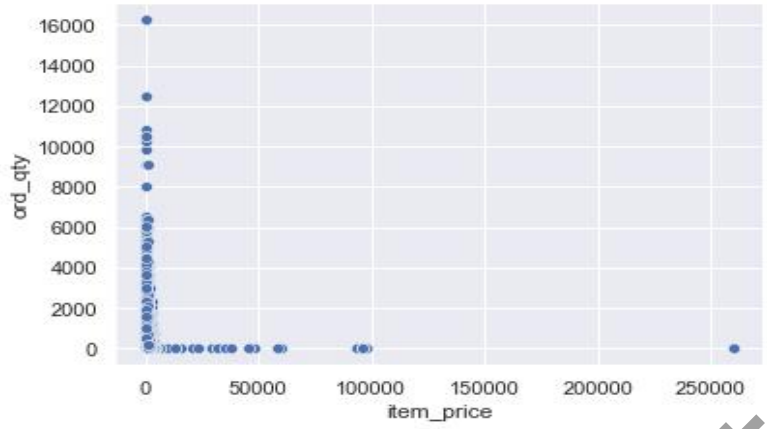


图 2 价格与需求量散点图

通过观察调整后的散点图，可以清晰地发现，订单需求量主要集中在中低价产品上，而高需求量订单比较罕见。此外，高价产品的需求量普遍较低。

为了探究不同产品价格对订单需求量的影响，我们根据产品价格进行了分组统计，并将产品价格分为四个区间：0~500 元、500~1000 元、1000~10000 元、10000~20000 元。接着计算了每个价格区间的平均需求量，并重新绘制了散点图，用于可视化展示不同价格区间的平均需求量，以探究产品价格对需求量的影响。

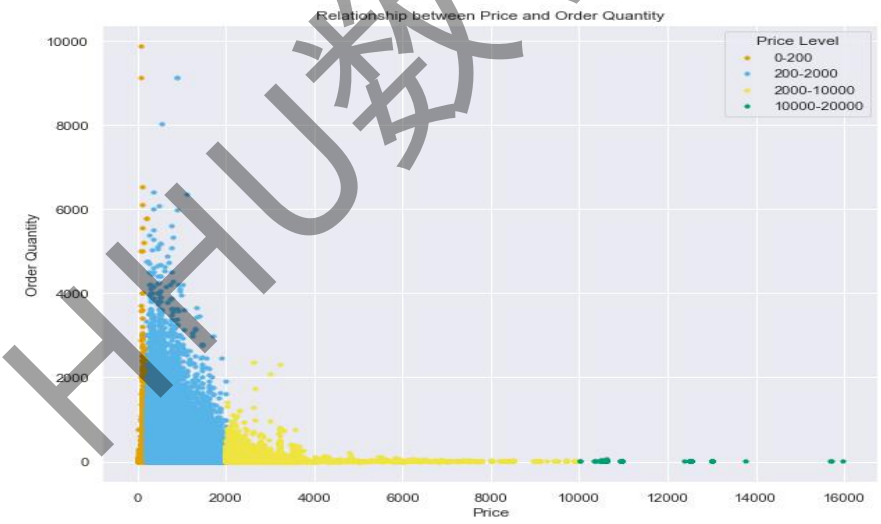


图 3 调整后坐标后的价格与需求量散点图

通过分析图表可以发现：订单需求量主要分布在价格区间 0~500 和 500~2000 之间，并且这两个价格区间内产品的订单需求量很容易达到较高水平，最高可达到 3500。当产品价格逐渐升高时，需求量呈现逐渐降低的趋势；尤其是在价格达到 4000 元后，价格对订单需求量的影响变得微弱。因此，这表明中低价产品具有更大的销售潜力，而高价产品的销量通常保持较低水平，不会以大批量销售

的方式出售。

3.2.2 问题一（2）：区域对需求量的影响及不同区域的产品需求量特性

提取各区域的订单需求量的平均值、中位数和标准差等描述统计值，并汇总每个区域的订单需求量，得到各销售区域的订单需求量柱状图，可以比较直观地看出：在 104 地区的需求量水平远低于其他地区；在 102 地区和 105 地区产品的需求较为可观。

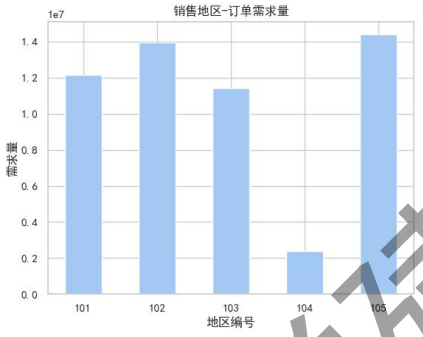


图 4 各销售区域需求量柱状图

通过观察各销售区域订单需求量的平均值、中位数和标准差等统计结果，可以发现所有销售区域（包括总需求水平极低的 104 地区）之间的订单需求差异并不显著。然而，对 104 地区的销售情况进行深入探究后发现，自 2017 年起该地区即没有销售记录。这表明：104 地区的低需求是由于停止销售所导致的，而非地域本身的影响。

表 2 各区域订单需求量统计值

	订单需求量		
销售区域编码	平均值	中位数	标准差
101	97.769776	33	213.431113
102	85.279410	33	154.004013
103	99.151032	32	210.253066
104	95.266050	34	190.182677
105	86.911657	21	202.177413

实际的销售数据通常不符合正态分布，因为销售数据受到很多非随机因素的影响，如季节性变化、市场趋势、市场份额、市场需求等因素。这些因素的影响使得销售数据呈现出一些非正态分布的特征，如偏态、厚尾、多峰等。所以在这里我们采用非参数检验。Kruskal-Wallis H 检验是一种非参数检验方法，用于比较三个或三个以上的独立样本的中位数是否相等。为探究不同销售区域之间的产品需求量是否具有显著性差异，我们利用 Kruskal-Wallis H 进行检验。

根据 Kruskal-Wallis H 检验的结果，p 值为 0。p 值小于 0.05 的显著性水平时则可以认为不同销售区域之间的订单需求量存在显著性差异。

表 3　Kruskal-Wallis H 独立样本中位数检验结果

总计	中位数	检验统计	自由度	渐进显著性（2-sided 检验）
597694	29	4281.047	4	0.000

Dunn 检验是一种多重比较方法，常用于在不同组之间进行两两差异比较。本文利用 Dunn 检验来比较不同销售区域之间订单需求量的差异性。检验结果以矩阵形式呈现，矩阵中每个单元格表示两个销售区域之间的差异显著性。在结果矩阵中，数值越小表示差异显著性越高，数值为 1 表示不存在差异，结果如图 5 所示：

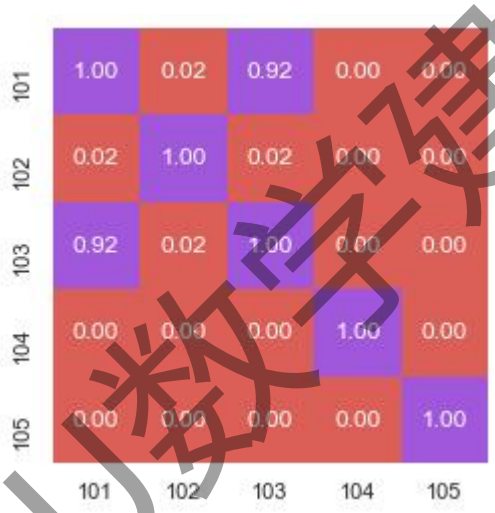


图 5 不同销售区域间订单需求量的 Dunn 检验结果

为分析进一步不同区域产品需求量的特性，提取各销售区域不同销售渠道的订单需求量绘制小提琴图，较为直观地观察到：101、102、103 地区产品销售基本是在线下进行的；而 104、105 地区采用了线上为主的销售方式。

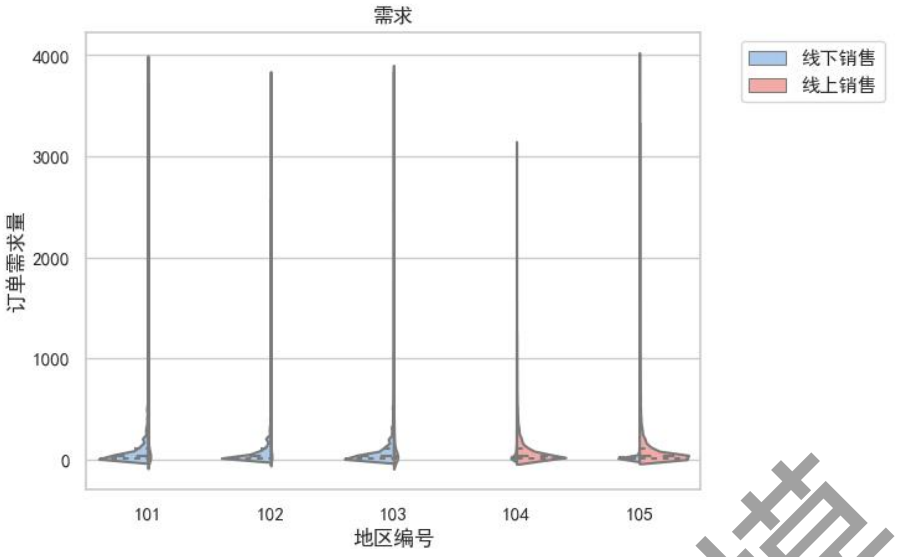


图 6 利用 sns.violinplot 函数绘制各区域不同销售渠道产品需求量小提琴图

3.2.3 问题一（3）：不同销售方式（线上/线下）的产品需求量特性

提取线上销售和线下销售的订单需求量进行汇总并计算平均值，结果显示：线下销售的产品总需求在总量中占据了更大的比重，但线下销售产品的平均需求量却不如线上销售的产品。这说明，该企业目前主要以传统线下销售为主要销售方式，同时辅之以线上销售模式。然而，考虑到线上销售的潜力极大，故该企业可在后期加强线上产品的比重，以期获得更高收益。

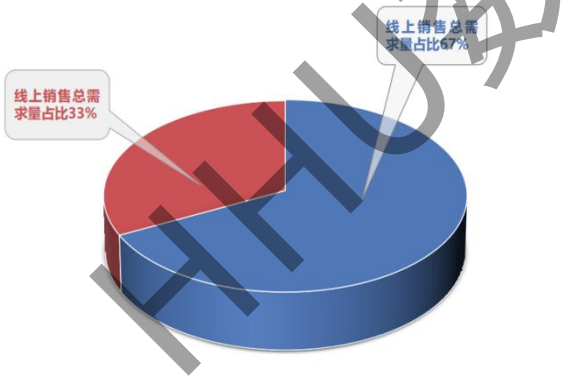


图 7 线下/线上订单总需求量占比饼状图

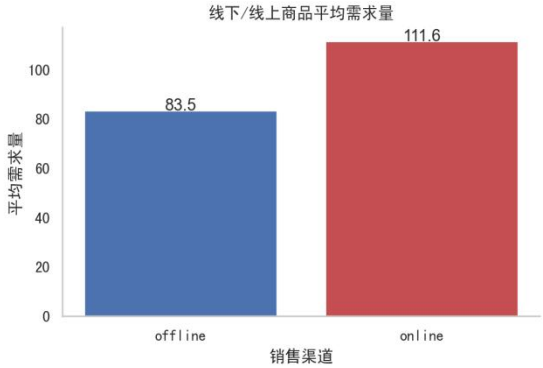


图 8 线下/线上订单平均需求量柱状图

图 9 线下/线上订单需求量随时间变化趋势图

从图 9 中观察线上、线下销售的产品需求量变化趋势，发现两种销售渠道的需求量均表现出一定范围内的波动，但总体而言，线下销售的表现优于线上销售。线上销售所呈现的产品需求折线图具有三个明显的高峰，这些高峰均出现在年底时期，暗示线上销售产品在年终时往往会迎来销售的高峰。

3.2.4 问题一（4）：不同品类之间产品需求量的不同点/共同点

根据各大类产品的总销量数据分析可得：306 大类产品的销售表现显著，其总需求量显著高于其他大类产品；经过对 306 大类产品销售数据的仔细检查，发现订单需求量中属于该类别的产品占据了总需求量的 36.7%。此外，306 大类产品销售记录中存在数量超过 10000 的大宗订单，这些订单很可能是推动该大类产品总需求水平领先的原因。综上所述，可以认为该大类产品是该企业重要且稳定的畅销产品。

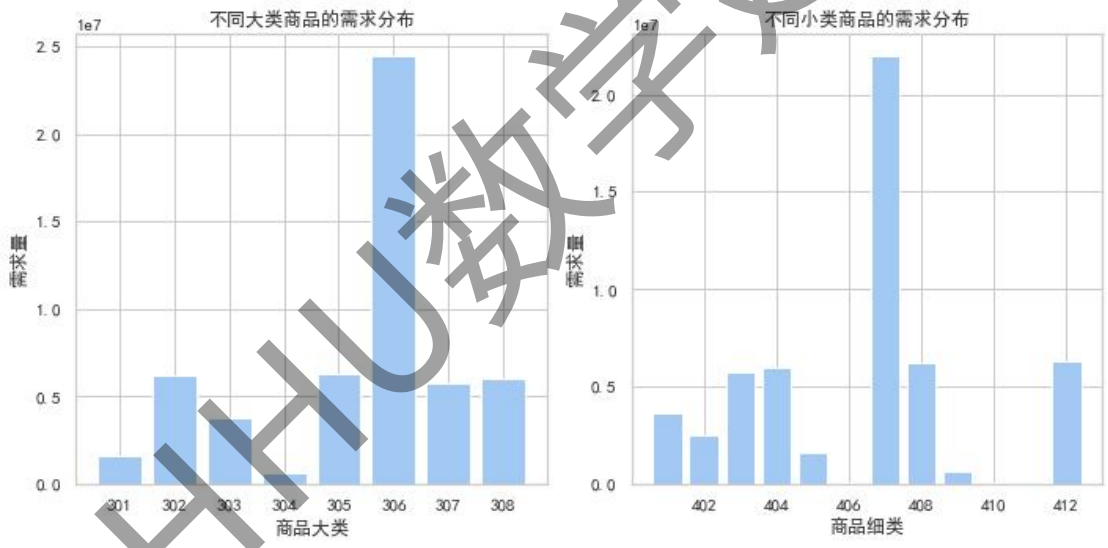
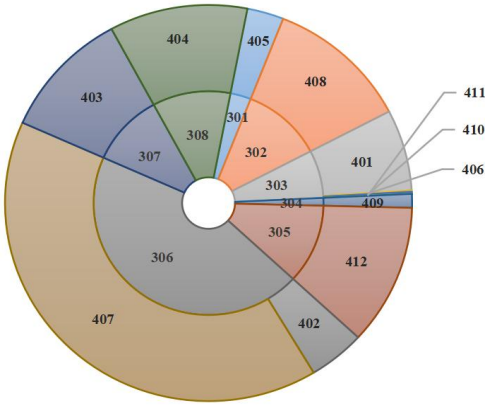


图 10 不同大类/细类产品的需求分布

根据细类产品的需求分布情况，可以得知 406、410 和 411 细类产品的销售表现不佳。根据不同类别产品需求量占比双环图可发现，这些细类产品均属于 303 大类，这也可解释其销售不佳的原因。

观察各大类、细类产品的订单需求量核密度分布图，可以发现所有类别（无论大类、小类）的产品的核密度图均呈现明显的偏态分布，且右侧尾部较长。每个类别都存在尖峰，说明各类别的订单需求量都较为集中。进一步观察细类，发现细类中 403、406 产品需求量的集中趋势非常明显。最后，观察各类别核密度图间的差异，发现它们之间都存在一定的差异。



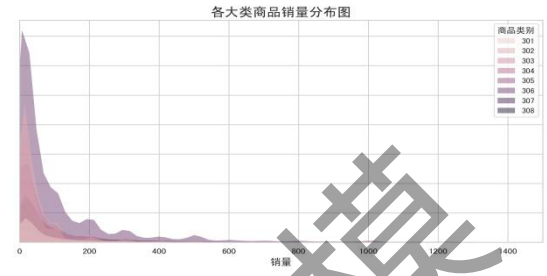


图 11 各大类产品订单需求量核密度分布图

图 12 各大类/细类产品需求量占比双环图

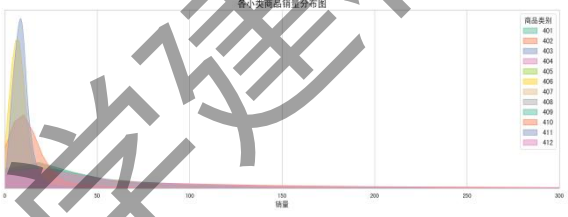


图 13 各细类产品订单需求量核密度分布图

绘制各大类产品、各细类产品的月需求量气泡图，用以展示和比较不同类别产品的订单需求量之间的联系和差异；气泡图中气泡的大小和颜色分别代表了不同维度的信息：气泡面积越大，表明订单需求量越高；气泡的不同颜色区分了不同类别（大类或细类）。通过观察可以确定大类中需求量最大的类别为 306，小类中需求量最大的类别为 401 和 404。

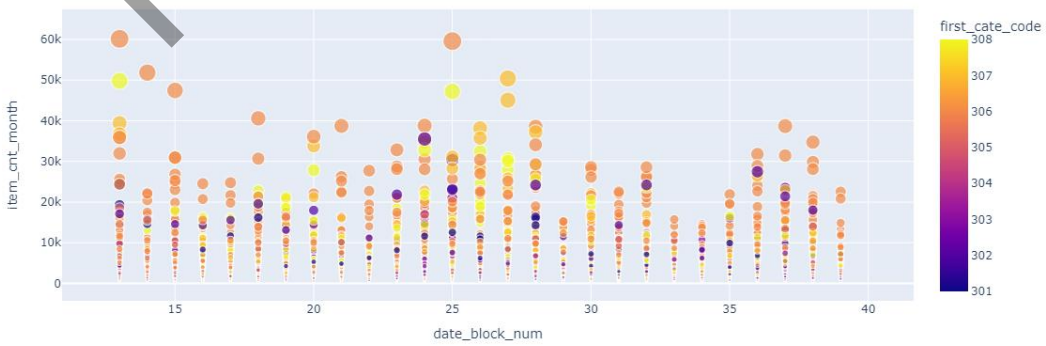


图 14 各大类产品月需求量气泡图

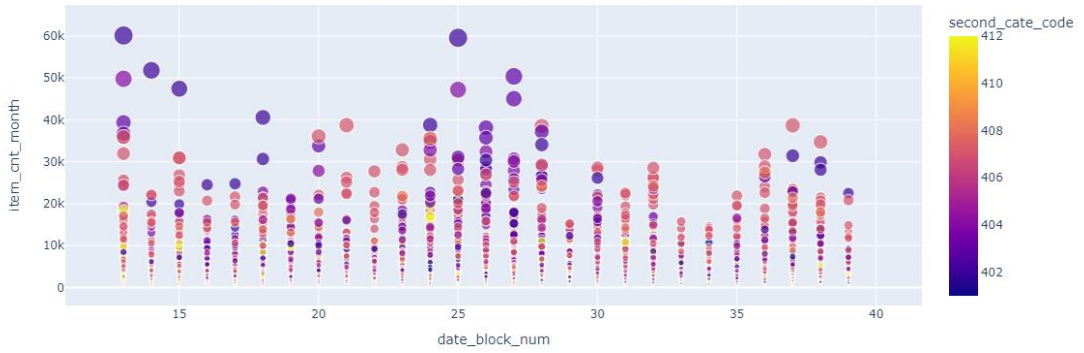


图 15 各大类产品月需求量气泡图

3.2.5 问题一（5）：不同时间段产品需求量的特性

为了研究不同时间段产品需求量的特性，本文将订单日期拆分为月初、月中、月末三个时间段，并提取每个时段的订单需求量，以此绘制 2015 年 9 月至 2018 年 12 月期间每个月订单需求量的趋势图。对比不同时间段之间的订单需求量，发现三个时间段之间的折线峰值出现的时点基本一致，并且在 2017 年以前波动趋势范围大致稳定，说明月初、月中和月末三个时间段的订单需求量随时间推移的整体变化趋势较为相似。然而，在 2017 年底到 2018 年初期间，订单需求量出现了较大程度的波动，可能是由于这段时间之内公司对销售策略作出了调整，导致订单需求量出现了不稳定的突变。此外，月初、月中和月末三个时间段之中，月末的产品需求量在多数情形下都高于月初和月中，其次是月中；月初的产品需求量总是要低一些，但其变动趋势范围较另外两个时间段而言更为稳定。

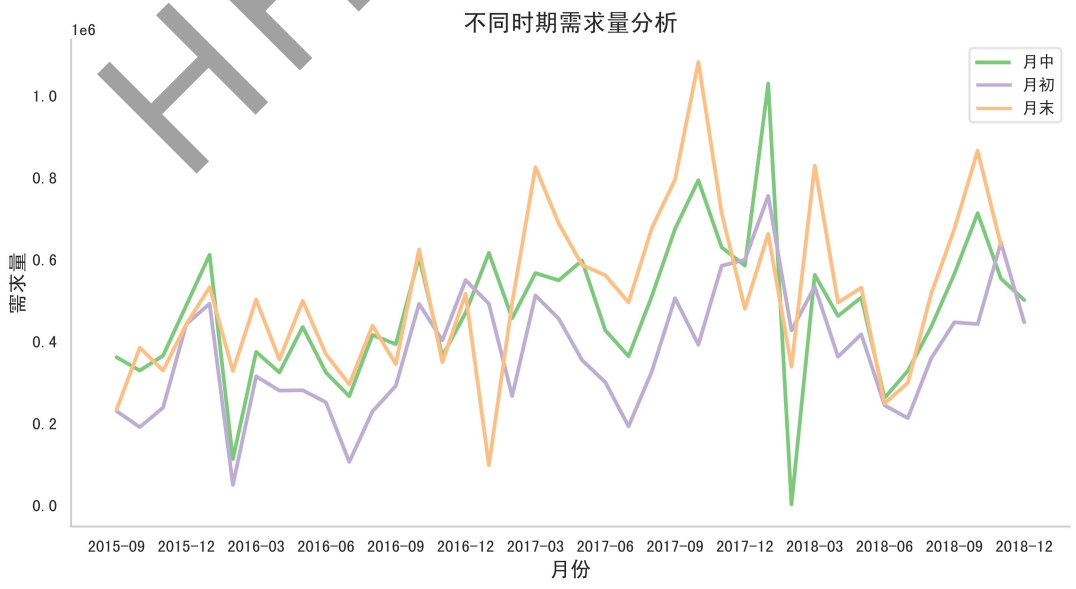


图 16 不同时段（月初、月中、月末）的产品需求量折线图

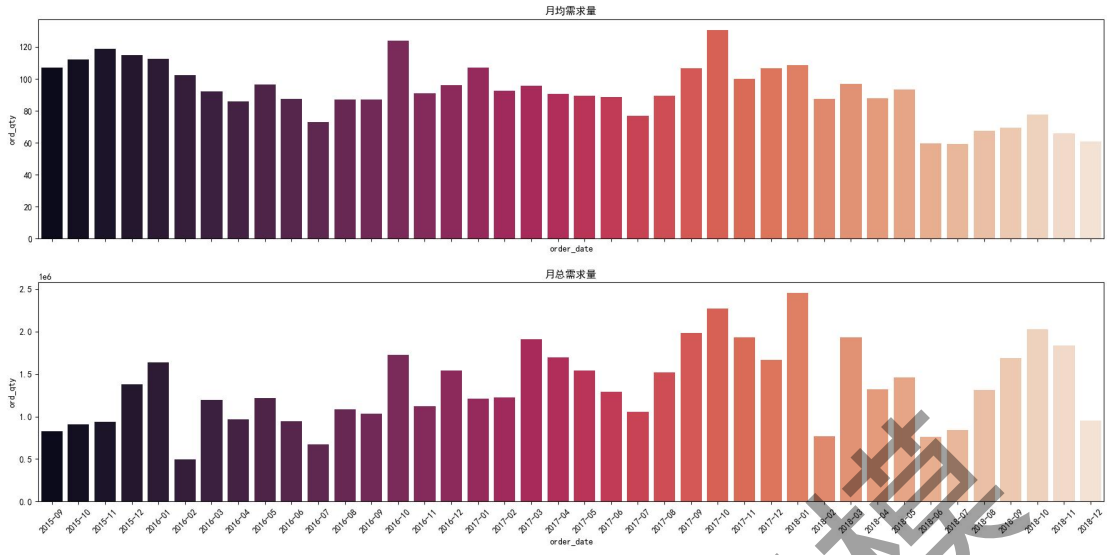


图 17 2015 年 9 月至 2018 年 12 月期间每个月的月均需求量与月总需求量柱状图

使用周为单位，对同一周内每一天的订单需求量进行分析。结果表明，在工作日销售情况相对稳定，但从周五开始到周末呈现出需求量上升的趋势，表明周末需求更高，且对其前后一天的需求也产生了正向影响。

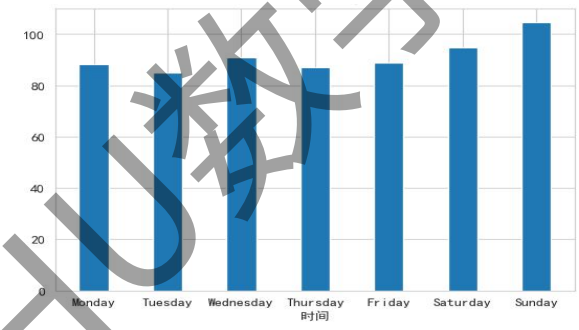


图 18 工作日/周末订单需求量柱状图

可以观察到，这些与时间相关的变量对于产品需求量具有显著影响，而不同产品在这些变量下的销售趋势也存在差异。因此，在制定后续的产品需求预测模型时，需要考虑到一周中的具体日期以及一年中的月份等时间因素的影响。

3.2.6 问题一（6）：节假日对产品需求量的影响

在研究节假日对产品需求量的影响时，本文不把周末双休日作为节假日的一部分。按照销售渠道进行划分，分别绘制 2015 年 9 月至 2018 年 12 月期间线上和线下产品需求量的折线图，并用红色五角星标注了节假日期间的订单。观察结果发现：除了在国庆假期期间线上销售的产品需求略有提高外，节假日对多数产品的需求并没有产生显著的提升效果。

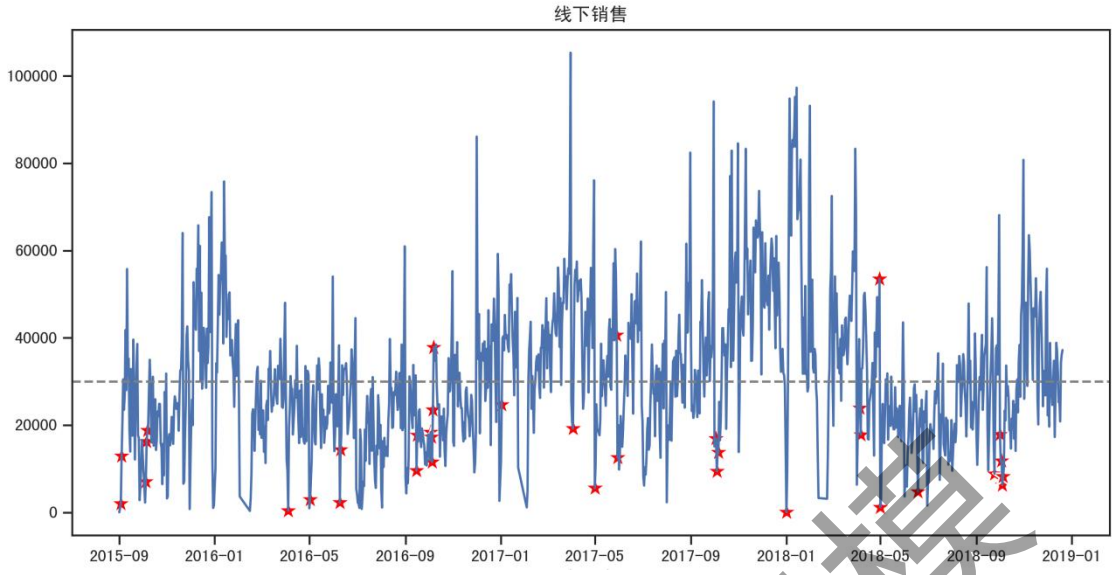


图 19 线下销售趋势

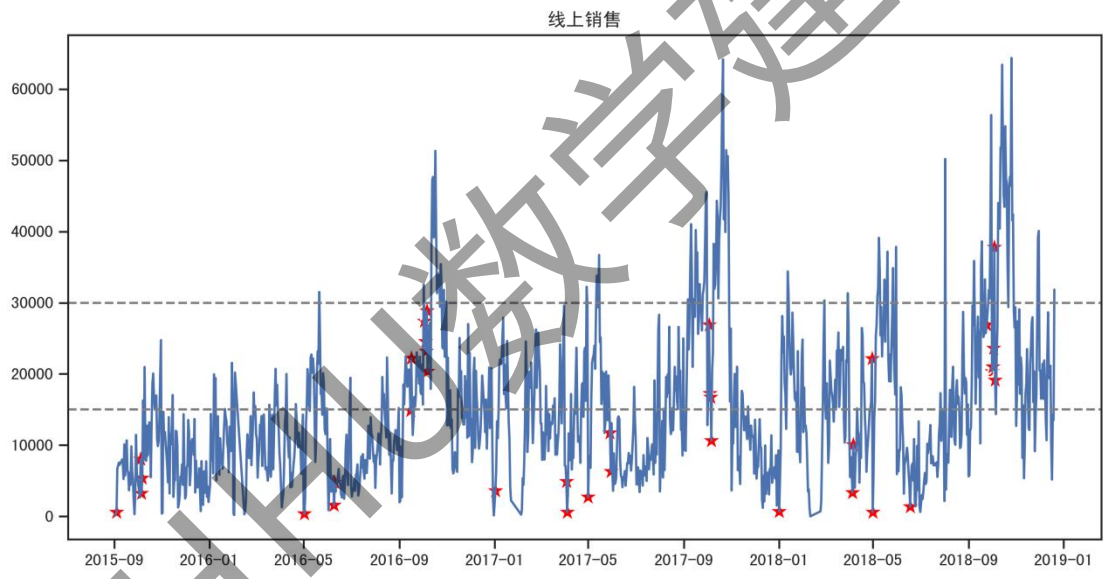


图 20 线上销售趋势

针对大类和细类分别统计节假日和非节假日期间的产品平均需求量。在大类层面，节假日对各大类产品的需求量都产生了一定程度的提升效应；进一步细分到各个细类，发现 401、406、410、411 细类产品的需求量受到节假日的影响微乎其微。根据细类产品需求量柱状图（详见附录）进行推测：这可能是因为这几个细类产品在平时销售情况就较为不佳，导致其需求量难以受到外部因素的影响而得到较大提升。

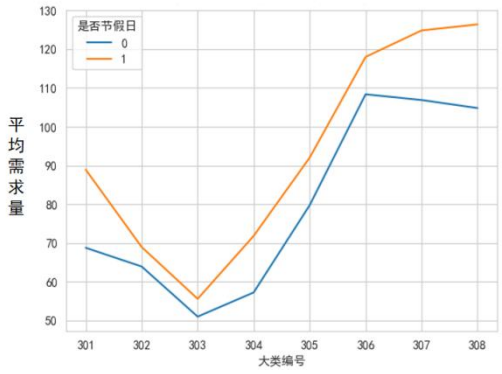


图 21 不同大类下节假日/非节假日平均需求量

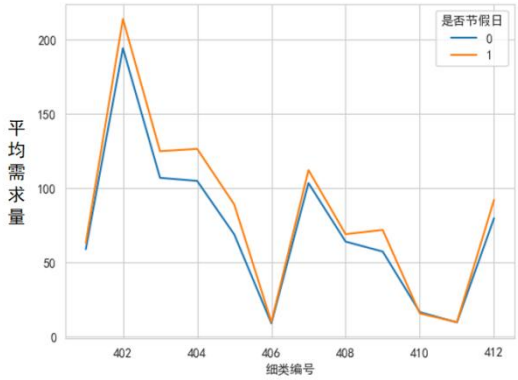


图 22 不同细类下节假日/非节假日平均需求量

3.2.7 问题一（7）：促销对产品需求量的影响

促销是一种重要的营销策略，其通过在特定时间内提供优惠购买条件或赠送礼品等方式，促进消费者的购买行为，从而提高产品需求。虽然促销通常以特定的日期如“6.18”或“双十一”为代表，但实际上大型促销活动并不仅限于单一的时间点，因为通常商家在促销日之前都会进行一系列预热活动，称为“预热期”。本文将促销定义为具有较长活动时间的常规促销，并选取“京东 6.18”和“淘宝双十一”两个典型的网络促销活动。在此基础上，将 6 月 4 日至 6 月 18 日和 11 月 1 日至 11 月 11 日期间定义为“促销日”。

由于网络促销活动可能会对同属于一个企业的产品的线下销售情况产生影响，为了更直观地观察促销活动对产品需求的影响，本文按照线上和线下两种销售渠道分别计算非促销日和促销日的平均需求量。结果表明：促销日的平均需求量低于非促销日，这表明促销日的产品需求情况并没有如预期一般优于非促销日。

表 4 促销日/非促销日下的线上/线下产品平均需求量

	非促销日	促销日
线上销售	113.795248	92.429559
线下销售	84.347899	81.495817

于是我们据此作出假设：并非所有产品都会受到促销的影响，或者说并非所有产品都适合通过促销刺激需求。可以通过提取促销期间订单需求量较非促销期间的增长率较高的产品编码及其销售区域、所属类别等信息来实现对促销产品更深入的观察。例如从图 25 可以观察到：受“6.18”促销影响最大的产品细类为 407，而受“双十一”促销影响最大的产品细类为 405；而 409、410、411 细类产品则未受到促销的影响。

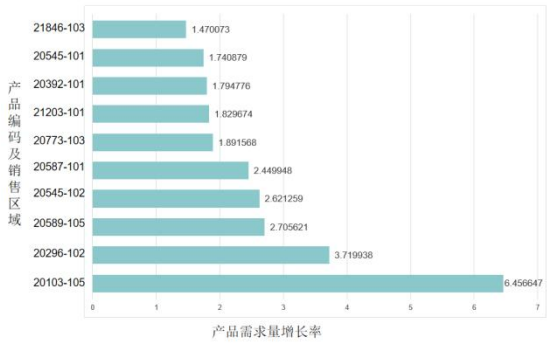


图 23 〝618〞期间产品需求量增长率 Top10

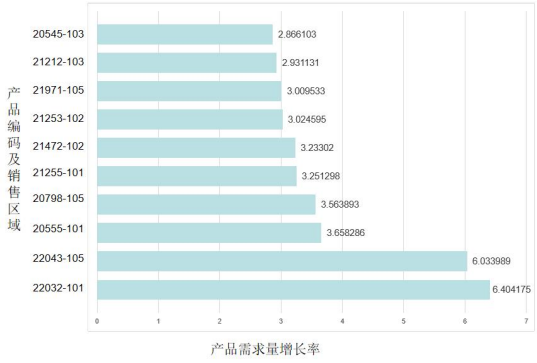


图 24 〝双十一〞期间产品需求量增长率 Top10

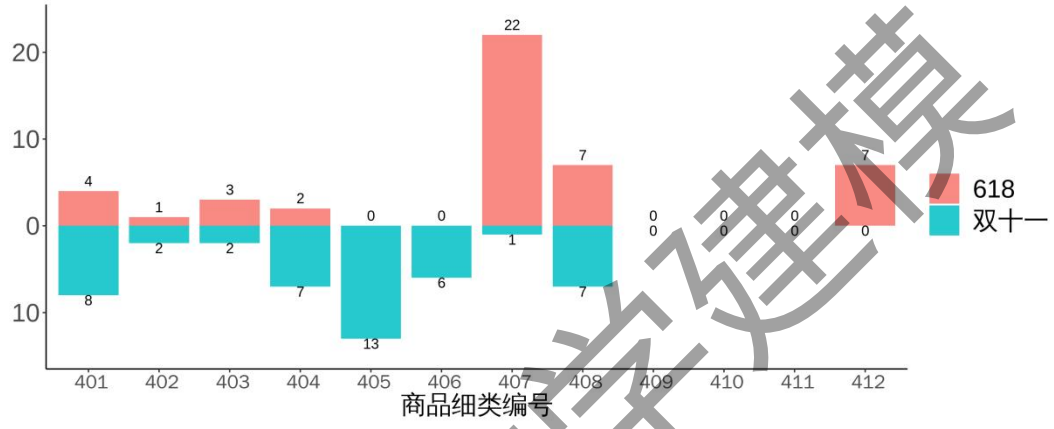


图 25 〝6.18〞和〝双十一〞期间 Top50 促销产品所属细类双向柱状图

3.2.8 问题一（8）：季节因素对产品需求量的影响

季节因素指的是季节性变化对产品需求所带来的影响。某些产品在特定季节的需求量可能会显著增加或减少，这被称为季节性需求。

季节因素指的是季节性变化给产品需求带来的影响，某些产品在特定季节的需求量可能会显著增加或减少，可以描述为这类产品存在季节性需求。了解季节性变化对产品需求的影响可以帮助企业制定更加合理、有效的市场营销策略，以提高销售额和市场份额。了解季节性变化对产品需求的影响，可以帮助企业制定更合理、有效的市场营销策略，以提高销售额和市场份额。

本文按照气象标准，将一年分为四个季节：3 月至 5 月为春季，6 月至 8 月为夏季，9 月至 11 月为秋季，12 月至次年 2 月为冬季，均采用公历。对比各个季节的产品需求量，可以发现：在四个季节中，夏季的产品需求最低，而冬季需求最高；一年当中春夏季节交替时，需求量会略有下降，而从秋季开始直到冬季，产品需求量都在逐步提升。

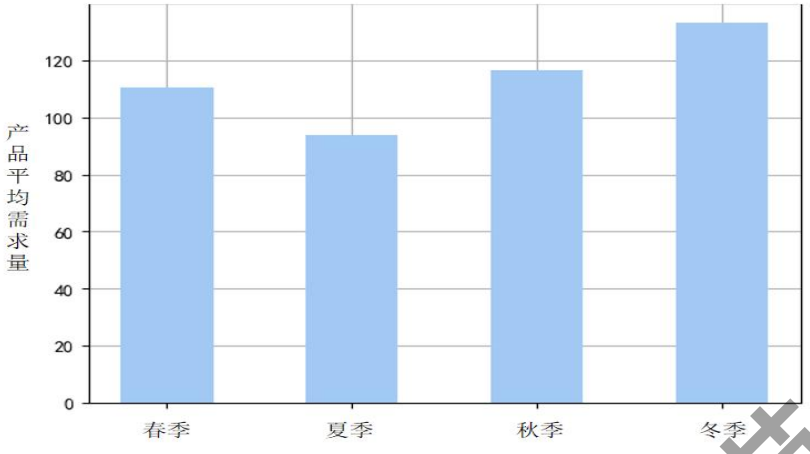


图 26 各个季节产品平均需求量柱状图

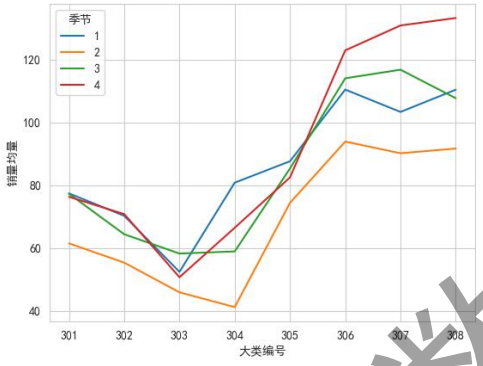


图 27 各个季节不同大类下产品需求趋势图

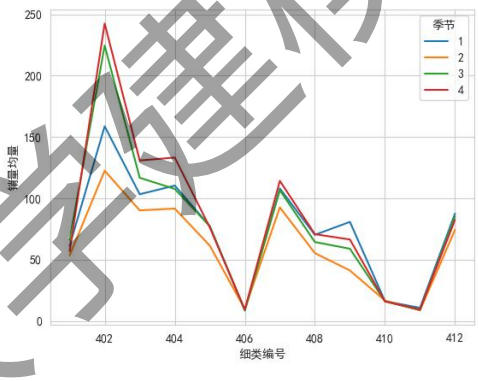


图 28 各个季节不同大类下产品需求趋势图

同样地，针对大类和细类分别绘制不同季节产品需求量的趋势图，观察发现：产品中存在着较多会受季节因素影响较大的产品，例如 304、306、307 大类产品，其需求量都出现了较大幅度的季节性波动；而从细类的角度进一步观察，这一结论更加明显：可以看到季节因素对 401、406、410、411 等细类产品的需求量几乎没有影响，但在 402 细类（隶属于 306 大类）产品上影响效果非常显著。此外还可以得出大部分产品都在冬季时需求最高这一结论。

第四章 数据预处理与特征工程

4.1 数据预处理

由前面分析中对示例数据的探索已知：该数据集中共有 597694 个样本，包括：1 个 float 类型特征，2 个 object 类型特征，5 个 int 类型特征。数据预处理阶段需要进行的工作有：对样本中可能影响模型训练的异常值进行判断和剔除；将非数值型数据转换为可供模型训练的数据格式。

4.1.1 异常值处理

（1）异常值检测

异常值是指在样本总体中偏离较远的观测值，其存在可能会降低数据的正态性和模型的拟合能力。因此，在建立模型之前，必须对高度异常的异常值进行识别和处理。

箱型图是用于清除异常值的常用手段之一。相比于 3sigma 准则，箱型图不要求数据服从正态分布，并且并且受异常值的干扰程度较小，能够直观地展示数据的实际形状。因此，使用箱型图检测异常值通常具有客观性。分别绘制订单需求量和产品价格的箱线图，观察发现：订单需求量和产品价格两组数据中均存在较多离群点，且均处于上边缘；订单需求量的离群点大部分位于 30000 以内，而产品价格的离群点多数位于 6000 以内。因此，初步将订单需求量大于 30000 或产品价格大于 6000 的数据视为异常值。

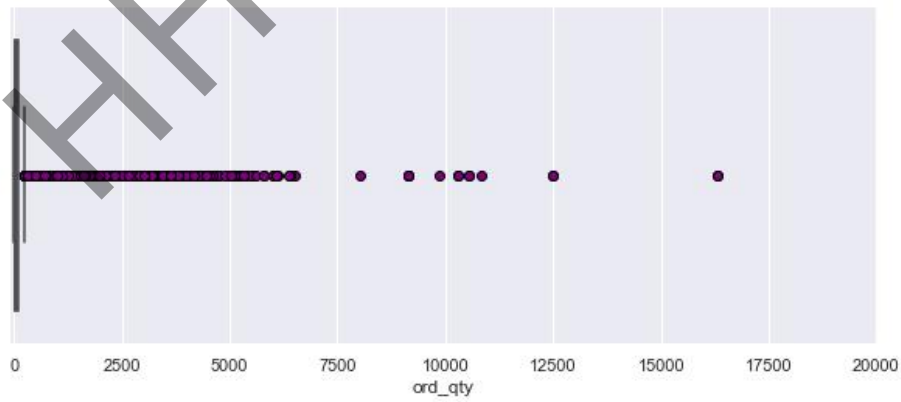


图 29 利用 sns.boxplot 函数绘制订单需求量箱线图

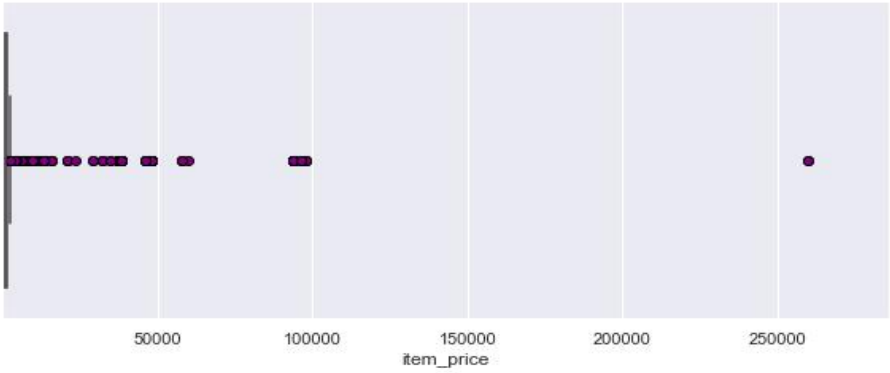


图 30 利用 sns.boxplot 函数绘制产品价格箱线图

（2）异常值分析和处理

将异常值与正常数据放在一起进行模型训练会影响模型的拟合效果，但如果只是简单地把所有异常值都剔除，也会导致这些异常值中可能包含的重要信息被直接忽略。所以异常值是否剔除需要根据具体情况进行判断。

将异常值全部查找出来进行逐个分析，检查每一条具有过高的需求量或过高的价格的样本数据是否属于合理的情况。结果显示：订单中的高需求量并不意味着这些订单就是错误数据，不能将这些订单数据全部当成异常值直接剔除掉。

对于这部分产品，我们先将其与测试集中的所有产品进行对照检查，如果这些可能异常的产品不存在于测试集中，则说明这部分数据对模型的训练并没有作用，可以直接剔除；如果存在于测试集中，则我们将对这些产品逐个建立单独的预测模型（预测结果见附录），以修正预测结果。在全模型构建阶段，保留产品价格 6000 以内的样本数据以及订单需求量在 30000 以内的样本数据。

异常值处理完毕后，该数据集共有 1753 种产品，产品类别分为 8 个大类，12 个细类，售往 5 个销售区域。其中有 332 种产品只售往一个地区，其他产品都在多个区域进行销售。

4.1.2 分类型数据处理

示例数据中存在的非数值型数据难以用于模型训练，需要对不同类型的数据进行相应处理，将所有数据转换为可供模型使用的格式。

（1）数据类型转换

基础数据中“订单日期”特征属于时间序列数据，以非结构化的格式存在而没有得到正确的排序，将其转换为结构化的日期时间数据类型，以便模型训练。

（2）真值转换

“销售渠道”特征格式为“offline/online”，由于分类数据无法直接应用于模型计算，通过真值转换利用 0、1 两个数值分别代表线上/线下两种销售渠道。

4.1.3 标签平滑处理

标签平滑处理（Label Smoothing）是一种常用的正则化技术，其作用是能够使数据更加符合正态分布，有助于提高模型的准确性和可靠性。常规的销售数据往往呈右偏长尾分布，即少数产品的订单需求量非常高但大多数的产品需求量都处于较低水平，这样的数据分布会导致训练数据中高需求水平的样本数量相对较少，可能导致模型会过度关注这一小部分高需求的样本而忽视大多数的产品；同时当订单需求量呈指数级别增长时，模型可能无法捕捉到这种关系，从而导致最终预测结果偏差较大。通过对订单需求量取对数，可以将右偏分布的订单需求量数据变换为更接近正态分布的数据，使得模型更容易捕捉到产品需求量的变化和趋势，从而提高模型准确性。

然而，对需求量取对数并不总是对预测模型有益的，有时候可能由于处理不当会降低模型的预测性能。本文尝试对产品订单需求量进行对数平滑处理，但利用评价指标对预测值进行评估后的结果显示：**对需求量取对数后，模型表现不佳**。这可能是因为：对需求量取对数后改变了数据的度量方式，从数量级变成了相对增长率，这可能导致了一些关键信息的丢失。

经文献阅读和实际尝试，本文发现对需求量进行对数平滑处理并用 RMSE 指标进行评价时，模型表现不佳；不对需求量进行平滑处理但使用 Tweedie 损失函数作为评价指标时，模型表现良好，所以本文通过利用 Tweedie 损失函数作为评价指标替代对数平滑处理。

4.2 数据集分析

在建立预测模型之前，对训练数据集和预测集进行深入分析有助于：

（1）理解数据分布：通过对训练数据集全面而深入的分析，可以帮助我们充分了解训练数据的特征和分布，以及数据中存在的问题和偏差。

（2）测试模型表现：通过对预测集进行分析可以帮助我们测试模型的表现，并识别在新数据上可能出现的问题。

总的来说，深入分析训练数据集和预测集有助于减少误差，提高模型的准确性和预测能力。

4.2.1 训练数据集分析

对训练数据按产品细类进行深入挖掘，发现各类产品具有很多潜在规律和特征，下面列出部分细类产品特征：

①403、404、405 细类产品最初完全依靠在线渠道销售，但自 2017 年起，开始增加线下销售方式。

②406 细类产品主要依靠线下销售，通常是小规模订单。2018 年 3 月，该类产品**从销售区域 105 迁移到其他三个销售区域**。

③407 细类产品的销售趋势呈多个小高峰，这表明该类产品具有季节性趋势。

④409 细类产品在 104 和 105 销售区域主要以线下形式销售；自 2018 年 8 月起，开始在线上销售；自 2018 年 5 月起，该类产品在其他三个销售区域上市。

⑤411 细类产品于 2017 年 11 月上市。

在对训练数据集进行探索性分析时，我们发现了一个引人注目的现象：自 2017 年起，地区 104 停止销售。因此，我们选择抽取在 104 地区有过销售记录的产品进行分析，并发现部分产品在 2017 年 1 月之前仅在 104 地区销售，但在 104 地区关停销售后，它们也并未停止销售，而是转移到其他区域进行销售。为验证这一假设，我们分析了在 104 地区销售过的所有产品在各个地区的销售情况。结果显示：原本在 104 地区销售的产品在 2017 年后部分集中地转移到了 105 地区进行销售，而部分则分散地转移到了 105 地区以外的其他销售区域。转移到 105 地区的产品具体情况如下：

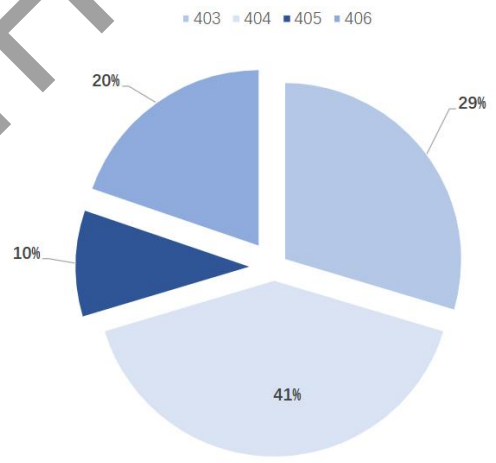


图 31　104 地区转移至 105 地区的产品信息

依据在训练数据集中观察和总结到的规律，我们对训练数据集作出以下处理：

（1）数据合成。

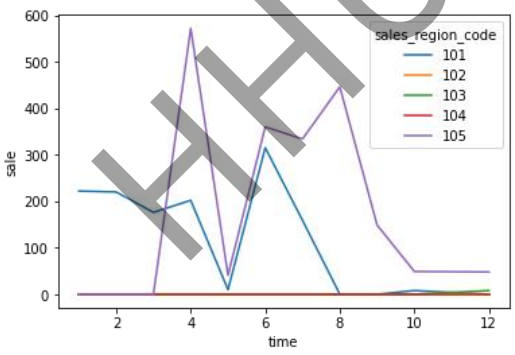
由于需要按月进行预测，所以本文对每个产品的订单需求量按销售区域和月份进行综合（暂时忽略销售渠道因素）。然而，产品在不同区域的销售情况存在差异，所以最终的合成结果中可能会存在缺失值。为了解决这个问题，我们对缺失数据用零值进行填补，并建立了一个包含销售区域、销售月份和产品等组合信息的数据集。值得注意的是，在前文对训练数据集的分析中，我们已发现销售区域编码为“104”的区域已经于 2017 年停止销售，同时“104”区域的产品大部分转移至“105”区域继续销售。因此在按月合成的过程中，将从“104”区域转至“105”区域的所有产品销售记录合并至“105”区域，然后删除“104”区域的相关数据。同时，由于 406 细类产品也在 2018 年 3 月发生了销售区域的转移，即从 105 地区迁往 101、102、103 地区，因此也需要对 406 细类产品的销售数据在相应的销售区域之间进行嫁接。

（2）商品分层

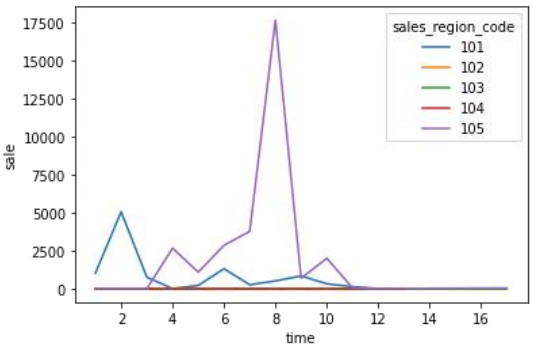
对产品的需求趋势曲线进行仔细检查后发现：不同产品的需求变化之间存在较为显著且复杂的差异，但其中也存在着一定的规律。通过对产品需求趋势变化的规律进行深入挖掘和总结归纳，本文将所有产品与地区进行组合然后对产品进行商品分层，共分为以下 4 类：

1.新品：直至第 36 个月（date_block_num）才开始出现在市场上的产品。

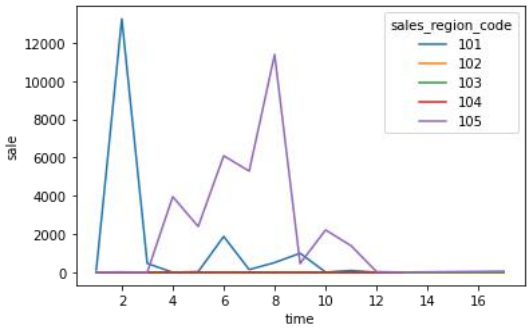
2.流星品：突然出现的商品；但销售时长不超过 5 个月，销量会急剧下降。



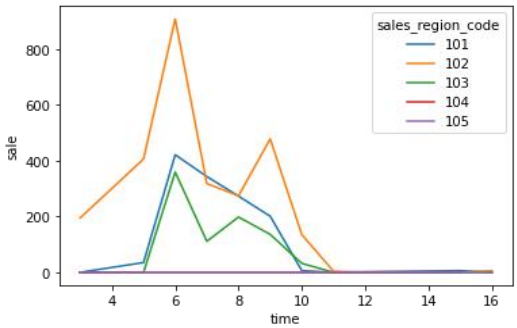
〝20677〞号产品



〝20132〞号产品



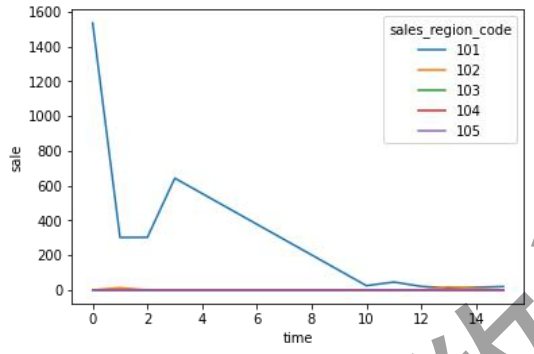
〝22032〞号产品



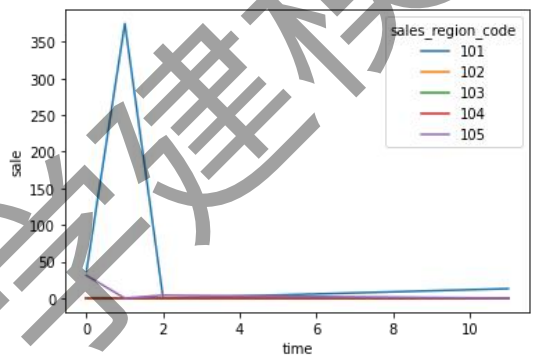
〝21136〞号产品

图 32 部分〝流星品〞月需求量趋势图

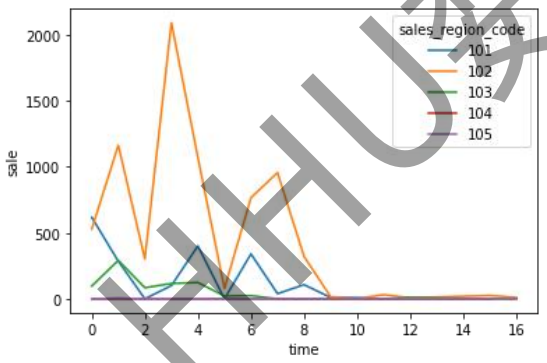
3.睡眠品：一直保持客观的销量，却在某个时间点之后销售量骤减，但究其原因并非季节性因素的产品。例如：



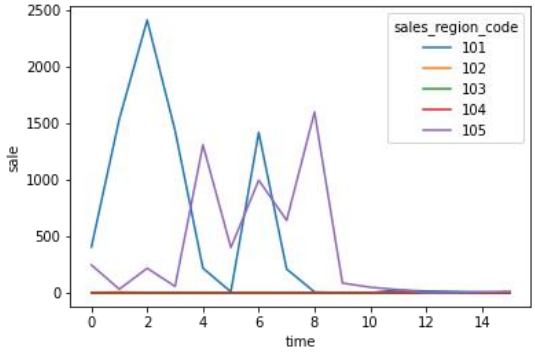
〝20910〞号产品



〝21104〞号产品



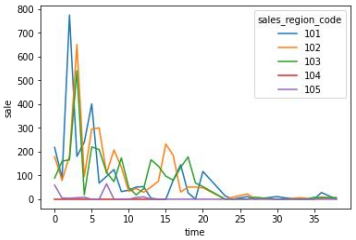
〝20302〞号产品

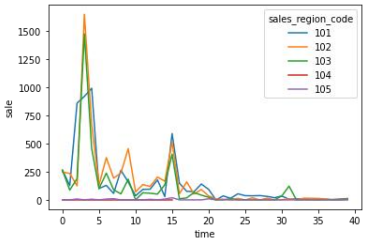


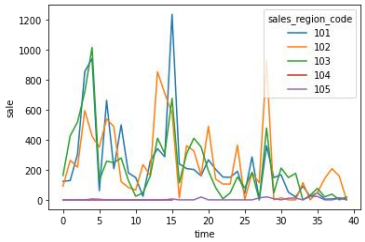
〝20374〞号产品

图 33 部分〝睡眠品〞月需求量趋势图

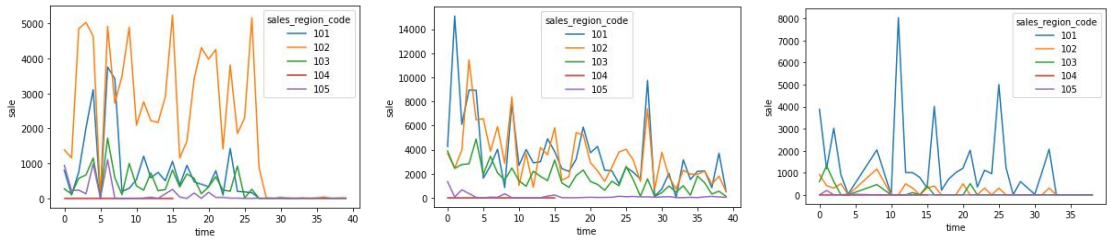
4.常规品：总有销量的产品；销售时长达 39 周以上或至少存在于市场中一年以上。







〝20082〞号产品　　〝20020〞号产品　　〝20140〞号产品



〝20189〞号产品　　〝20323〞号产品　　〝20342〞号产品

图 34 部分〝常规品〞月需求量趋势图

4.2.2 预测数据集分析

针对训练数据集分析中提出的商品分层，为避免不同类型的产品因采用同一模型进行预测所带来的较大误差，所要进行的必要措施是对预测数据集进行检查和分析。该分析旨在审查待预测产品的特征和分布，以确保预测模型的有效性和准确性。

表 5 地区/产品组合中不同商品分层的数量

产品总数	新品	流星品	睡眠品	常规品
2619	666	5	95	1853

表 6 地区/产品组合中的新品信息

细类编码	401	402	403	404	405	406
组合数量	57	7	121	116	9	2
细类编码	407	408	409	410	411	412
组合数量	115	100	28	9	0	102

统计并分析预测集中的产品后发现：绝大部分待预测的产品均属于常规品，少量属于流星品和睡眠品。需要特别注意的是，**新品在预测集中占据了较大的比例**；同时这些新品大部分隶属于 403、404、407、408 和 412 细类。说明该企业经常推出新品，以满足不断变化的市场需求和顾客偏好的变化。因此，为提高模型的预测准确度，本文将针对新品单独建立一个预测模型。

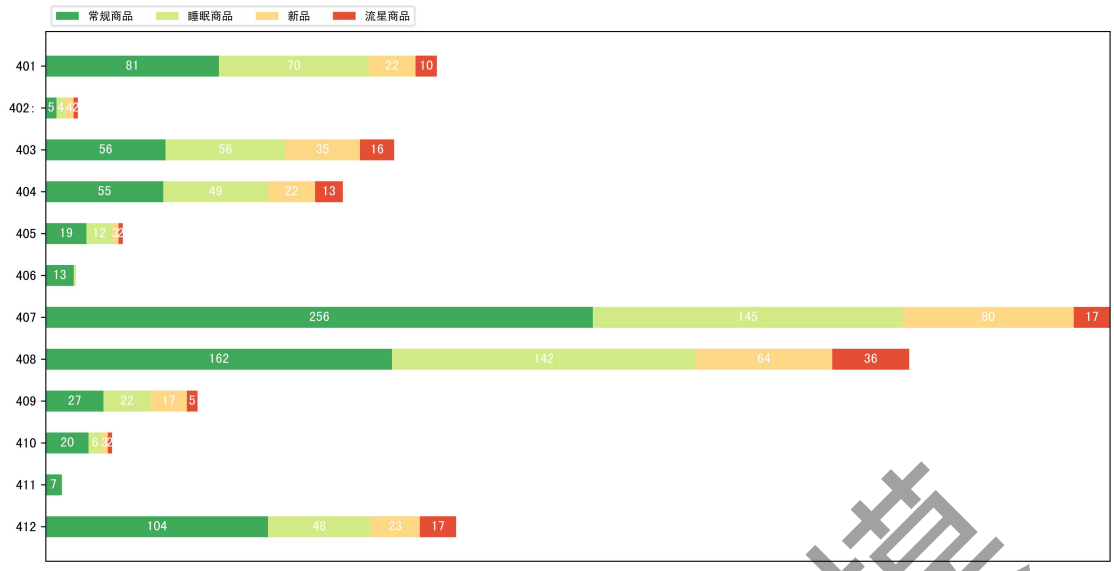


图 35 预测集中的商品分层堆叠图

4.3 特征工程

4.3.1 特征工程概述

特征工程是在通过建立机器学习模型进行需求预测时较为关键的步骤，是一个通过对不理想的原始样本数据采取一系列措施和方法进行有效处理后，得到可供后续机器学习模型使用的优化较好的特征数据组合的过程[8]。特征工程决定了模型训练结果的优劣。所提取的特征越重要，最终取得的预测效果往往越好。

根据任务要求，需要为模型建立充足的备选特征集。产品需求量的诸多影响因素中可能存在着一些随机干扰因素，但由于随机干扰因素具有不可测性，本文不单独抽取随机干扰因素作为特征，将在误差分析部分对其进行解释。下面本小节将重点介绍以下内容：特征工程的构造方法、具体的特征构造以及最终建立完成的备选特征集。

4.3.2 特征工程的构造方法

特征构造指的是从原始样本数据中构造新特征的处理过程，一般来说，特征以及特征的数据类型都要紧扣目标和需求进行选取。本文要对时间序列数据进行特征提取，可采取的特征构造方法有：

（1）选取基本特征

给定的训练数据集中其实已经具有多个维度的基本特征，可以对这些数据直接进行统计提取。对于非数值型的特征，常常采用均值编码（Mean Encoding）、标签编码（Label Encoding）和独热编码（One-hot Encoding）等编码技术将其转

换为可供模型使用的数值型特征。

（2）构造复杂特征

对于时间序列数据而言，可以采用多种方法基于基本特征构造一些复杂特征：

引入滞后特征：时间序列数据中的滞后特征是指用过去的数值作为预测特征。滞后特征的作用在于捕捉序列数据中的趋势和周期性变化，从而提高模型的预测准确性。为每个样本增加同一产品在上个月、半年前、一年前的销售字段，可以将其与历史数据建立一定的联系。滞后特征也可以看作是一种特殊的滑动窗口技术。

引入趋势特征：趋势特征可以反映出数据随时间变化的趋势和规律，因此可以帮助模型更好地理解数据。如果数据中存在一些随时间变化的规律或趋势，那么将这些特征引入模型中可以帮助模型更好地捕捉这些规律，从而提高模型的预测准确性和解释性。根据价格、订单需求量等基本特征可以引入丰富的趋势特征。

4.3.3 具体的特征构造

（1）滞后特征

计算各月份的平均需求量，然后提取前一个月的月平均需求量作为特征；

计算各产品每个月的月平均需求量；各销售区域每个月的月平均需求量；各产品每个月在某个销售区域的月需求量、月平均需求量；提取前 1、2、3、6、12 个月的相应值作为特征；

计算某产品所属细类、所属大类在某个销售区域的月平均需求量，提取前 1 个月的响应值作为特征。

（2）趋势特征

计算各产品每个月的平均价格，提取前 1 个月的价格变化趋势作为特征；

$$某月的价格趋势 = \frac{(前一个月的月平均价格 - 当前月的月平均价格)}{当前月的月平均价格}$$

计算各产品在每个月的订单需求量变化趋势，提取前 1 个月、前 1 年的同期的该值作为特征；

利用产品价格和订单需求量计算各产品在每个月的收益变化趋势，提取前 1 个月的该值作为特征；

提取产品前 6 个月订单需求量的标准差作为波动率特征。

（3）均值编码（Mean Encoding)

用产品编码、产品类别编码（包括大类、细类）、销售区域编码下的平均需求量来替换原始特征。

（4）时间衍生特征

提取某次订单所处月份的总天数作为特征；

若订单日期处于节假日期间，提取某次订单所处节假日的总天数作为特征；

若订单日期处于促销月期间，则标记该月为促销月。

（5）其他特征

计算产品价格与订单需求量的相关系数，将其作为特征；

提取产品距离上市的月数作为特征；

提取产品所属细类距离上市的月数作为特征；

提取订单的销售渠道作为特征。并在此基础上新增特征“线上引领线下”。这是因为在对不同渠道的销售数据进行观察时发现：同时经由线上、线下销售的产品，其线上订单的高峰总比线下订单高峰提前一个月左右，所以将该现象提取为一个特征。

4.3.4 备选特征集

最终建立完成的备选特征集如图所示：

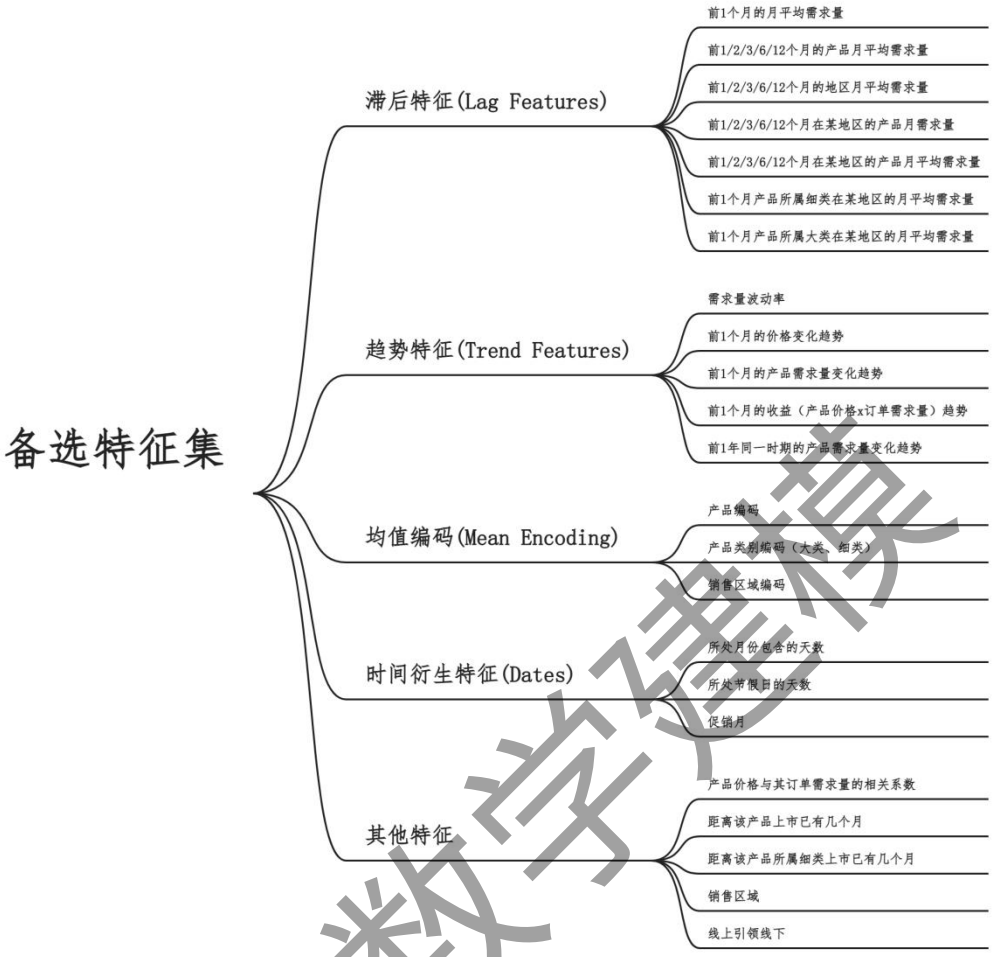


图 36 备选特征集

第五章 基于 GBDT-Prophet 的产品需求量预测模型

5.1 模型建立

本章针对问题二提出的要求旨在构建一个能够准确地预测未来三个月的产品月需求量的**基于 GBDT-Prophet 的产品需求量预测模型**。在这一部分当中，我们主要进行了以下工作：

①建立 Prophet 模型对训练数据中的季节趋势进行提取，并利用 Prophet 模型对未来三个月的月需求量按周和天的粒度进行预测；然后融合三种不同的机器学习算法（XGBoost、LightGBM、CatBoost），形成最终的全模型对未来三个月的月需求量按月进行预测。

②在模型建立时，我们进行了大量的特征工程以提高模型的性能；同时通过深入探索训练数据集中产品销售数据的潜在规律，对产品进行**商品分层**的特殊处理，并据此对预测集中的产品分布进行挖掘，最终选择针对产品中的“新品”和非新品分别构建特征并建立预测模型。

③在对模型进行评估时，我们尝试了两种评价指标：WMAPE 以及 Tweedie 损失函数；结果显示用 Tweedie 损失函数对模型进行评估的效果更好；

④对预测结果进行误差分析后，抽取预测偏差较大的产品进行分析，并对其单独建立一个预测模型进行预测，然后将预测值覆盖至原预测结果，提高模型的预测准确性。

⑤在对未来三个月的月需求量按月进行预测时，我们采用了直接预测、滚动预测以及滞后预测三种方式进行预测。最终根据预测结果的准确度选择采用融合滚动预测和滞后预测对未来三个月的月需求进行预测。

5.1.1 样本划分

由于时序问题的特殊性，应使用过去的数据预测未来，而不能用未来的数据预测过去；并且验证集应该尽量位于距离预测目标最近的时段。所以在训练初期，我们选择 2018 年 12 月的所有数据作为测试集，选择 2018 年 10 月、11 月两个月份的所有数据作为验证集，2018 年 9 月及以前的所有数据全部用作训练集，来进行特征工程和参数调优；在训练中期，将验证集调整为 2018 年 11 月和 12 月的数据，将训练集调整为 2018 年 10 月及以前的全部数据，将 2019 年 1 月需

预测的数据作为测试集进行模型融合和测试；而在提交最终结果时将 2019 年 1 月及以前的所有数据全部用作训练集。

表 7 不同训练阶段的样本划分

	训练前期	训练中期	最终
train 训练集	2018 年 10 月以前	2018 年 11 月以前	2019 年 1 月及以前
valid 验证集	2018 年 10 月、11 月	2018 年 11 月、12 月	
test 测试集	2018 年 12 月	2019 年 1 月	

5.1.2 模型框架

（1）Prophet 模型

Prophet 是一种由趋势项、季节项、节假日项和误差项组成的加法模型：

$$y(t) = g(t) + s(t) + h(t) + \varepsilon \quad ①$$

其中，$y(t)$是某一时刻的预测值；$g(t)$为趋势项，表示非周期的变化趋势；$s(t)$为季节项，或者称为周期项，一般以周或年为单位；$h(t)$即节假日项，表示时间序列中潜在的、具有非周期性的节假日对预测值的影响；ε为误差项，或称为剩余项，表示模型未预测到的波动，误差项 ε 服从高斯分布。

对于趋势函数 $g(t)$，有两种常用的构造方式：分段线性和分段逻辑回归。本文选用基于分段线性函数的模型，即线性可加模型（LAM）：

$$g(t) = (k + a(t)\,\delta) \cdot t + (m + a(t)^{T}\gamma) \quad ②$$

其中，k 表示增长率；δ表示增长率的变化量；m 表示偏移量；参数$\gamma = [\gamma_1, \gamma_2, \cdots, \gamma_S]^{T}$，$\gamma_S = -s_j\delta_j$，$S$ 是变点的数量，$s_j(1 \le j \le S)$表示变点的位置。

由于时间序列中可能包含天、周、月、年等不同周期类型的季节性趋势，因此在季节想 s（t）中采用傅里叶级数模拟时间序列的周期属性：

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right) \quad ③$$

其中，P 表示周期，例如 P=7 表示以周为周期；N 表示希望在模型中使用此周期的个数，N 值较大时可以拟合出更复杂的季节性函数，但同时也会带来更多的过拟合问题。

节假日项$h(t)$用于估计节假日效应，需要一个参数设定节假日的影响范围：

$$h(t) = Z(t)k = \sum_{i=1}^{L} k_i \bullet 1_{\{t \in D_i\}} \quad ④$$

其中，L 表示节假日的个数；k_i表示对应节假日的影响范围。

Prophet 模型适用性分析：GBDT 模型更适用于捕捉非周期性的趋势，例如数据中的非线性关系、交互作用等，因为它本身并没有对数据中季节性趋势的捕捉能力进行特别设计，因此它对季节性趋势的捕捉效果可能会受到一定的限制。如果数据中存在较为明显的季节性趋势，GBDT 可以在一定程度上学习到这些模式，但当数据中的季节性趋势非常强烈，且希望得到更加准确的预测结果时，需要建立专门针对时间序列数据的模型。相对于 GBDT，Prophet 更加专注于时间序列分析，可以更好地捕捉和分析季节性趋势。因此我们可以通过建立 Prophet 预测模型对训练数据中潜在的季节趋势进行提取。

（2）XGBoost 模型

XGBoost 模型通过弱学习器的迭代计算来改进模型的分类效果，逐步提升模型的分类准确率[9]。XGBoost 以分类回归树（CART）作为基分类器[10]，假设有 K 个数，则树集成模型为：

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i) \quad f_k \in \Gamma \quad ⑤$$

其中 Γ 为包含所有回归树的函数空间；f_k为函数空间 Γ 里的一个函数，对应第 k 棵独立树的结构 q 和叶子的权重ω。

XGBoost 模型采用前向分布加法算法，即通过在每一次迭代中添加增量函数$f_i(x_i)$来优化目标函数。XGBoost 的目标函数可以分解为两部分：第一部分是用于衡量模型预测误差的损失函数；第二部分是用于控制模型复杂度的正则项。通过迭代地最小化这两部分的和，可以得到最终的模型。其目标函数定义如下：

$$J^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-i)} + f_i(x_i)) + \Omega(f_i) + C \quad ⑥$$

l 为单个样本的损失函数，n 为样本数量，y_i为样本实际值，$\hat{y}_i$为样本预测值。

在每次迭代过程中，都需要对每个样本的残差进行计算，并拟合一棵新的决策树来预测这些残差。为了避免过拟合，可以使用 L1 和 L2 正则化来控制叶子节点的权重，以及子采样和列采样来控制每棵树的形状和大小，以此控制每棵树

的复杂度。最终，将所有树的输出加权求和，得到最终的预测值。

优点：可以自动捕捉输入特征之间的复杂关系，并能够处理高维稀疏数据；同时具有高准确性和良好的泛化能力，能够对产品需求量的时间序列进行准确的预测；对于噪声和异常值也具有一定的鲁棒性，能够在存在一定干扰的情况下仍然保持较高的预测精度。

缺点：对于输入数据的质量要求较高，需要对数据进行清洗、特征提取和特征选择等操作，否则会对模型的预测效果产生较大的影响；训练和预测过程的计算资源消耗大，对于参数的选择和调整也比较敏感，容易出现过拟合现象。

（3）LightGBM 模型

LightGBM 相较于 GBDT、XGBoost，有效地解决了处理海量数据的问题，在实际应用中常常能够取得出色的效果。

LightGBM 将 N 棵若回归树线性组合为强大的回归树，通过使用基于 Histogram 的决策树优化算法降低了时间的复杂度。直方图算法的基本思想是：将连续的浮点特征量离散为 k 个整数，，从而构成多个盒子，这一步称为装箱处理；在此基础上，构造一个宽度为 k 的直方图。在遍历数据时，将离散后的数值作为索引在直方图中累计统计量；遍历完一次数据后，得到所需统计量；然后根据直方图的离散值便利寻找最佳分割点。

同时，LightGBM 还采用了带深度限制的按叶子生长（Leaf-Wise）策略。该策略遍历一次数据可以同时分裂同一层的叶子，容易进行多线程优化，也好控制模型复杂度，不容易发生过拟合现象。但在样本量较小的时候，Leaf-Wise 策略容易造成过拟合，因此在 Leaf-wise 之上增加了一个最大深度的限制，以保证高效率的同时能够防止过拟合。

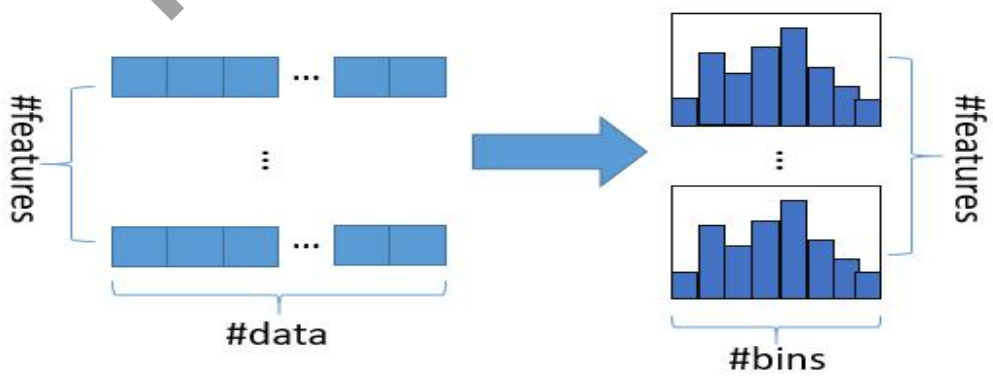


图 37 直方图算法图解

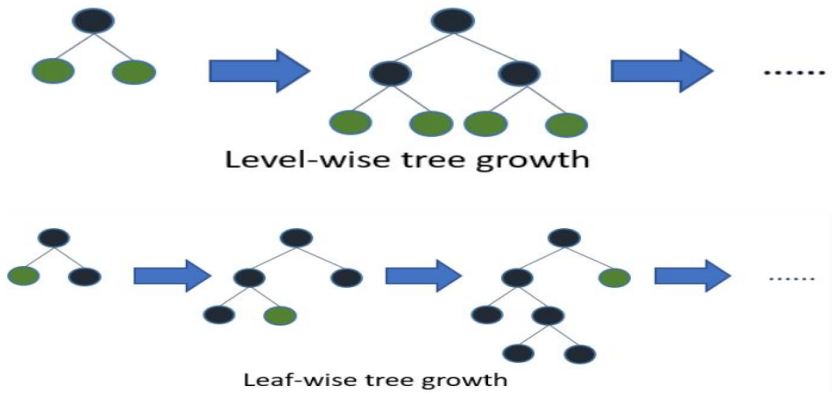


图 38 Leaf-Wise 策略图解

优点：通过使用优化算法有效地提高模型的训练速度和预测速度，特别是在大规模数据集上表现尤为突出；可以自动处理缺失值、离散值等问题，同时可以使用类别特征和数值特征，并支持多分类任务和回归任务等多种任务类型，能够得到较为准确的预测结果；对于数据噪声和异常值的处理能力较强，能够有效地避免过拟合和欠拟合的问题。

缺点：由于采用的是梯度提升算法，容易陷入局部最优解，如果不进行正确的调参和特征选择，容易导致模型过拟合；在训练过程中对异常值的处理能力较弱，如果数据集中存在异常值，可能会对模型的预测效果产生负面影响。

（4）CatBoost 模型

CatBoost 能够更好地处理 GBDT 特征中的类别特征（Categorical Features），常用于解决分类和回归问题。它在处理类别特征时，利用类别特征对应的标签平均值来替换。在决策树中，也将标签平均值作为节点分裂的标准。这种方法被称为 Greedy TBS，用公式表达为：

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n} [x_{j,k} = x_{i,k}] \bullet Y_i}{\sum_{j=1}^{n} [x_{j,k} = x_{i,k}]} \qquad ⑦$$

但是，由于特征在通常情况下比标签要包含更多的信息，如果强行用标签均值来表示特征，一旦训练数据集和测试数据集的数据结构和分布存在一定差异时就有可能会出现条件偏移问题。据此通过添加先验分布项对 Greedy TBS 作出改

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1} [x_{\sigma_j,k} = x_{\sigma_p,k}] \bullet Y_{\sigma_j} + a \bullet P}{\sum_{j=1}^{p-1} [x_{\sigma_j,k} = x_{\sigma_p,k}] + a} \qquad ⑧$$

进：

其中 P 是添加的先验项，a 是权重系数。添加先验项后可以减少噪声和低频率数据对于数据分布的影响。

优点：性能卓越，能够有效地处理具有大量特征的数据集；支持对类别型数据的直接处理，无需对类别型数据进行独热编码等处理，可以有效减少特征处理时间；不容易出现过拟合现象，模型的鲁棒性较好。

缺点：与其他 GBDT 模型相比训练时间较长；有很多可调参数，且不同的参数之间可能会产生相互影响，因此参数调整较为困难。

5.1.3 评价指标

本文选用 WMAPE 和 Tweedie 损失函数作为模型的评价指标。

WMAPE（Weighted Mean Absolute Percentage Error）加权平均绝对百分比误差：相较于 MAPE 和 WAPE 考虑了在产品之间或时间上可能存在的优先级差异，通过对优先项目进行加权使预测误差偏向优先项目，减小数据中极端值带

$$WMAPE = 100\% \times \frac{\sum_{t=1}^{n}(w_t|A_t - F_t|)}{\sum_{t=1}^{n}(w_t|A_t|)} \quad ⑨$$

来的误差波动。在实际销售预测中，多用 WMAPE 作为评价指标。

Tweedie 损失函数：在理解 Tweedie 函数之前需要理解一个概念：Tweedie 分布。Tweedie 分布是一类用于描述广义线性模型误差分布的统计模型。具有可以控制方差大小和分布形状的参数 p，当 p=1 时，Tweedie 分布可以等同于泊松分布；当 p=2 时，等同于伽马分布；当 1<p<2 时，Tweedie 是泊松分布和伽马分布的复合分布。**本文将 p 值设为 1.5。**

表 8　Tweedie 分布的参数 p 决定其分布所属的子函数族

Tweedie EDMs	p	V(μ)	φ
Normal	0	1	σ^2
Poisson	1	μ	1
Poisson-Gamma	1<p<2	μ^p	φ
Gama	2	μ^2	φ
Inverse Gaussian	3	μ^3	φ

Tweedie 分布允许各种不同类型的分布和参数组合，通常用于处理数据中存

在严重偏斜的情况。对于呈长尾分布的数据，利用基于 Tweedie 分布的损失函数进行建模其效果优于其他分布的损失函数。同时，Tweedie 分布可以很好地模拟数值为 0 的情况，由于示例数据中很多产品存在间歇性需求，即很多产品在某些时段的销量为 0，因此模型训练的损失函数采用 tweedie 效果更佳。

$$Tweedie = \frac{1}{n}\sum_{i=1}^{n} -(y_i \frac{\mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho}) \quad ⑩$$

5.2 模型训练

在这一部分我们将不同模型在训练集上试验，并对比验证集上的性能。

在这一部分当中，本文将先建立 Prophet 模型进行季节性趋势提取，并以日、周为精度进行预测，接着利用 XGBoost、LightGBM 和 CatBoost 分别建立一个预测模型后融合成一个全模型。本文将以订单需求量作为调整模型结构的基准，并将 WMAPE 和 Tweedie 指标函数作为评价模型的指标。

在模型建立之初，我们会给模型一个初始值（一般是默认值），并设定调参范围，然后对各个模型的超参数进行调整使达其到各自的最佳超参数设置，提高模型预测精度。最后再应用随机森林算法进行模型融合。

在对超参数进行调优时，本文选择网格搜索方法对模型中的各个超参数进行调整。网格搜索是一种重要的调参手段，也是应用最广泛的超参数搜索算法，其本质是一种穷举遍历算法[11]。它在给定超参数范围内，穷举出所有可能的参数值组合放入模型中进行训练，选取其中表现最好的超参数组合作为最终结果。

5.2.1 Prophet 模型（提取季节趋势/按周和日预测）

①Prophet 模型的重要控制参数如下：

表 9 Prophet 模型调参结果表

参数名称	初始值	最佳参数
changepoint_prior_scale	0.05	0.05
yearly_seasonality	TRUE	TRUE
growth	linear	linear
seasonality_prior_scale	addtive	multiplicative
Holidays_prior_scale	10	17

changepoint_prior_scale 指设定自动突变点选择的灵活性，值越大越容易出现

changepoint；yearly_seasonality 代表时间序列的时序周期性，增加预测模型的拟合程度；seasonality_prior_scale 和 holidays_prior_scales 则代表改变周期性影响因素的强度，值越大，周期性因素在预测值中的影响程度越大。growth 指增长趋势，分为“linear”与“logistic”，分别代表线性与非线性的增长，默认值为 linear。

②我们以 105 地区的 21715 号产品为例来解释 Prophet 模型的预测结果：

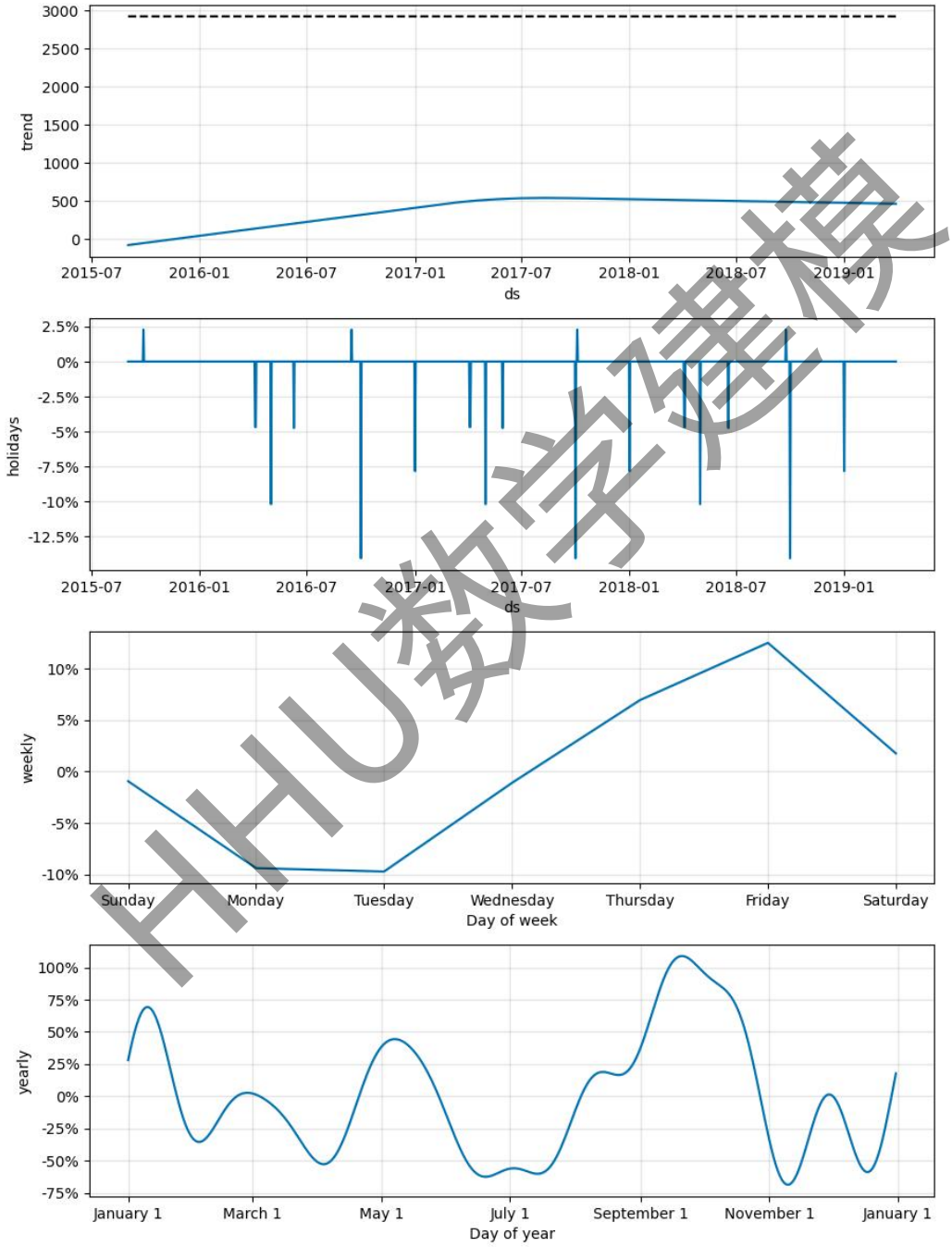


图 39 105 地区的 21715 号产品 Prophet 模型预测结果图

其中第一个小图呈现的是趋势组件（Trend Component），用于展示时间序列的长期趋势，通常是一个向上或向下的线性或非线性曲线。可以看到：21715

号产品的长期趋势呈先向上后趋于平稳的线性曲线。

第二个小图呈现的是节假日组件（Holidays），用于呈现在时间序列中存在的节假日，这些节假日会对时间序列的预测产生影响。

第三个小图呈现的是周期性组件（Weekly Component），用于展示时间序列中每周的波动，是一个以 7 天为周期的周期性波动。

第四个小图呈现的是周期性组件（Yearly Component），用于展示了时间序列中每年的波动，是一个周期为一年的周期性波动。

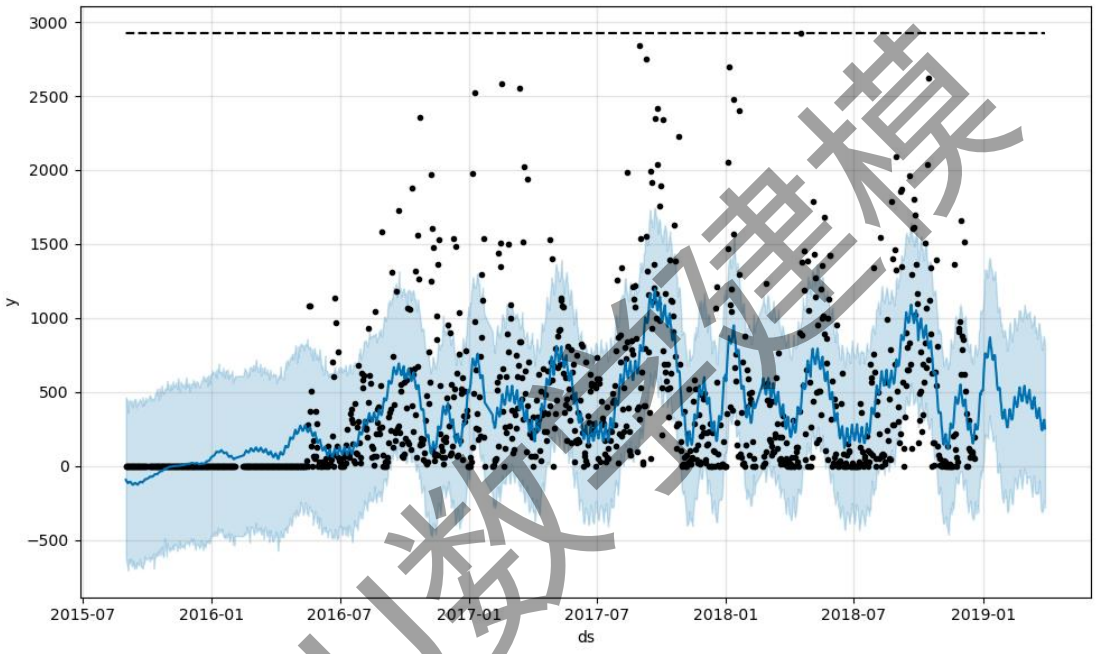


图 40 21715 号产品 Prophet 模型按天预测结果图

可以用 m.plot()函数用于绘制 Prophet 模型对未来数据的预测结果，生成一个包含许多子图的大图。图中的黑色点表示原始数据中的每个数据点；蓝色线表示模型的拟合结果；浅蓝色区域表示模型的置信区间：置信区间越宽，表示模型的不确定性越大；黑色线表示未来数据的预测结果。这些数据点都是 Prophet 模型基于历史数据所做出的预测。

利用 Prophet 按周和天对产品需求量进行预测后，我们对结果进行比较：

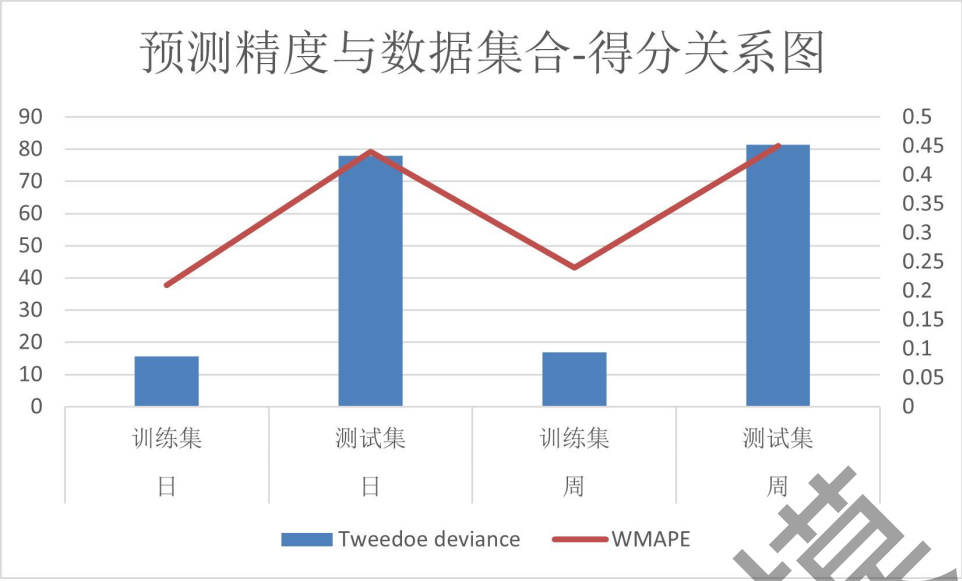


图 41 利用 Prophet 模型按周预测和按日预测的预测精度对比

对需求量按月预测将在之后利用全模型进行，并会对按月、按周、按日的预测结果进行对比分析。

5.2.2 XGBoost-LightGBM-CatBoost 预测模型（按月预测）

（1）XGBoost

①XGBoost 的重要学习控制参数如下：

Max_Depth：一个整数,表示子树的最大深度。该参数的取值越大，模型的学习越具体，越容易过拟合；

Eta：表示学习速率。在更新叶子节点权重时，乘以 Eta 的值，减少每次迭代的权重值，从而防止过拟合现象的发生。Eta 的值越小，算法越稳健，学习过程越仔细，无法收敛的可能性越小。

Subsaple:该参数用于控制每棵树的随机采样比例，即训练样本数量占整体样本数量的比例。这个参数值越小，算法越稳健。但当该值过小时，可能会发生欠拟合现象；

Colsample_bytree：该参数用于控制每棵树随机采样的列数（特征）占总体列数（特征）的比例；

Verbosity：该参数用于控制输出消息的详细程度。

②XGBoost 模型参数的初始化值如表 10 所示：

表 10 XGBoost 模型参数调优结果

参数名	Max_Depth	Eta	Subsaple	Colsample_bytree	Verbosity
最优值	4	0.1	0.6	0.9	0

经网格搜索优化发现：‘Max_Depth’取 4，‘Eta’取 0.1，‘Subsaple’取 0.6，‘Colsample_bytree’取 0.9，Verbosity 取 0 时，模型表现最佳。

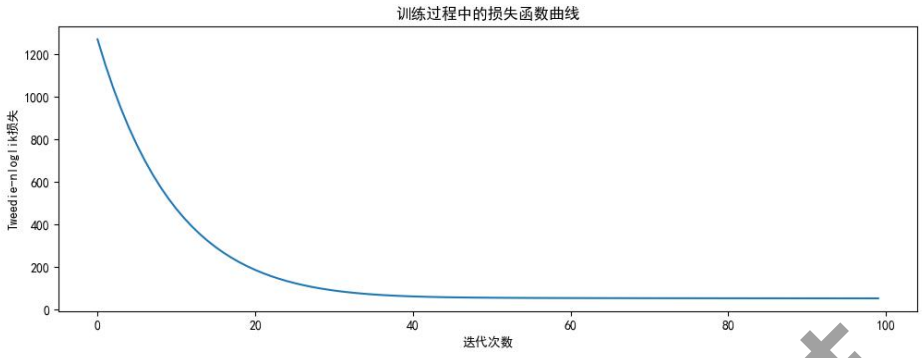


图 42 XGBoost 模型的损失函数曲线

③XGBoost 模型超参数优化后的模型提升效果如表 11 **所示：**

表 11 XGBoost 模型超参数优化后各评价指标提升结果

	WMAPE	Tweedie deviance
训练集调优前	0.384	9.37
训练集调优后	0.356	8.93
测试集调优前	0.594	14.24
测试集调优后	0.573	13.35
测试集调整后提升百分比	3.53%	6.25%

（2）LightGBM

①LightGBM 的重要学习控制参数如下：

Num_Leaves：该参数控制了每个决策树的叶子节点数。该值越大，模型的复杂度越高，但容易出现过拟合；该值越小，模型的复杂度越低，但也容易出现欠拟合。

Learning_Rate：一个浮点数，表示学习率；

②LightGBM 模型部分参数调优如表 12 **所示：**

表 12 LightGBM 模型部分参数调优结果

参数名	Learning_Rate	Max_Depth	Num_Leaves
最优值	0.034	30	140

经网格搜索优化发现：‘Learning_Rate’取 0.034，‘Max_Depth’取 30，‘Num_Leaves’取 140 时，模型表现最佳。

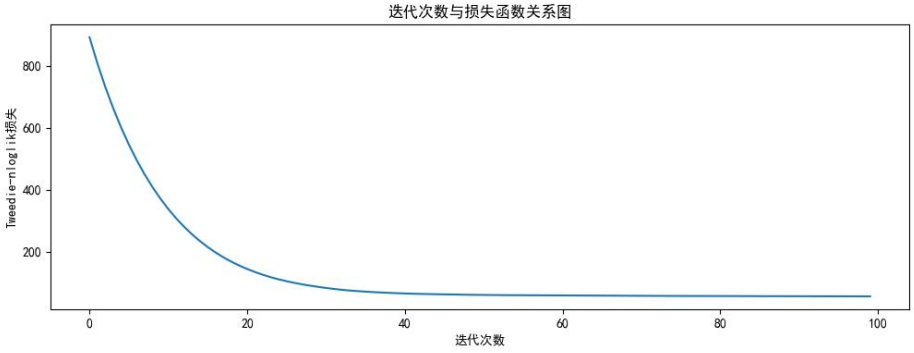


图 43 LightGBM 模型的损失函数曲线

③LightGBM 模型超参数优化后的模型提升效果如表 13 **所示：**

表 13 LightGBM 模型超参数优化后各评价指标提升结果

	WMAPE	Tweedie deviance
训练集调优前	0.366	7.90
训练集调优后	0.348	7.73
测试集调优前	0.528	15.27
测试集调优后	0.513	15.19
调整后提升百分比	2.84%	0.52%

（3）CatBoost

①CatBoost 的重要学习控制参数如下：

Iterations：该参数表示树的最大数量。默认值为 100；

Depth：该参数表示了树的深度；

Colsample_Bylevel：该参数表示训练过程中输出的度量值。

②CatBoost 模型部分参数的初始化值如表 14 **所示：**

表 14 CatBoost 模型部分参数调优结果

参数名	Learning_Rate	Depth	Subsample	Iterations	Colsample_Bylevel
最优值	0.1	4	0.7	100	1

经网格搜索优化发现：‘Learning_Rate’取 0.1，‘Depth’取 4，‘Subsample’取 0.7，‘Iterations’取 100，‘Colsample_Bylevel’取 1 时，模型表现最佳。

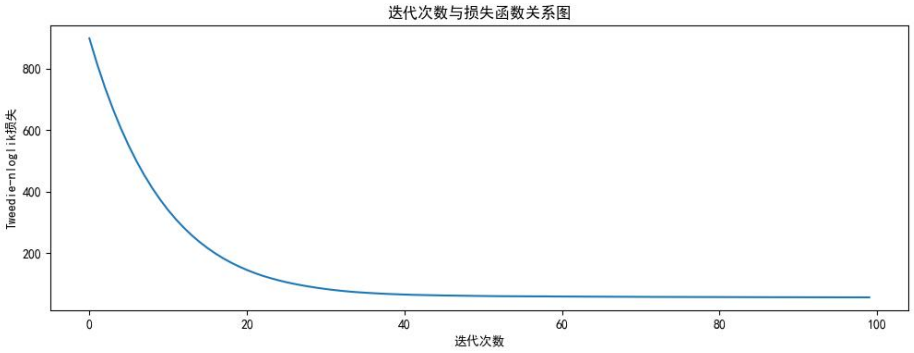


图 44 CatBoost 模型的损失函数曲线

③CatBoost 模型超参数优化后的模型提升效果如表 15 所示：

表 15 CatBoost **模型超参数优化后各评价指标提升结果**

	WMAPE	Tweedie deviance
训练集调优前	0.366	7.90
训练集调优后	0.348	7.73
测试集调优前	0.661	15.76
测试集调优后	0.634	15.58
调整后提升百分比	4.08%	1.14%

5.2.3 误差分析

以 LightGBM 为基础模型，在对预处理后的数据进行模型训练并在验证集上进行预测后，我们使用加权平均绝对百分误差（WMAPE）以及 Tweedie 损失函数计算了所有商品的需求量的预测误差。基于误差分析，我们对使模型出现高损失的订单进行了检查，并针对这些订单找到了相应的产品/地区组合，对这些预测误差较大的产品/地区组进行深入分析：结果发现预测偏差较大的 261 个组合中，有 17 个组合不在预测集范围内，说明这些产品既会对模型训练造成干扰，又没有训练价值，因此我们直接从训练集和验证集中剔除这些组合；剩余的 244 个组合存在于测试集中的产品，其规模较大，所以我们单独对其进行训练和预测，然后将预测结果覆盖提交到原先的预测模型中。在训练初期，使用误差分析和覆盖提交对模型的预测效果有显著的提升，测试集 Tweedie deviance 提升 21.10%。

5.2.4 特征筛选

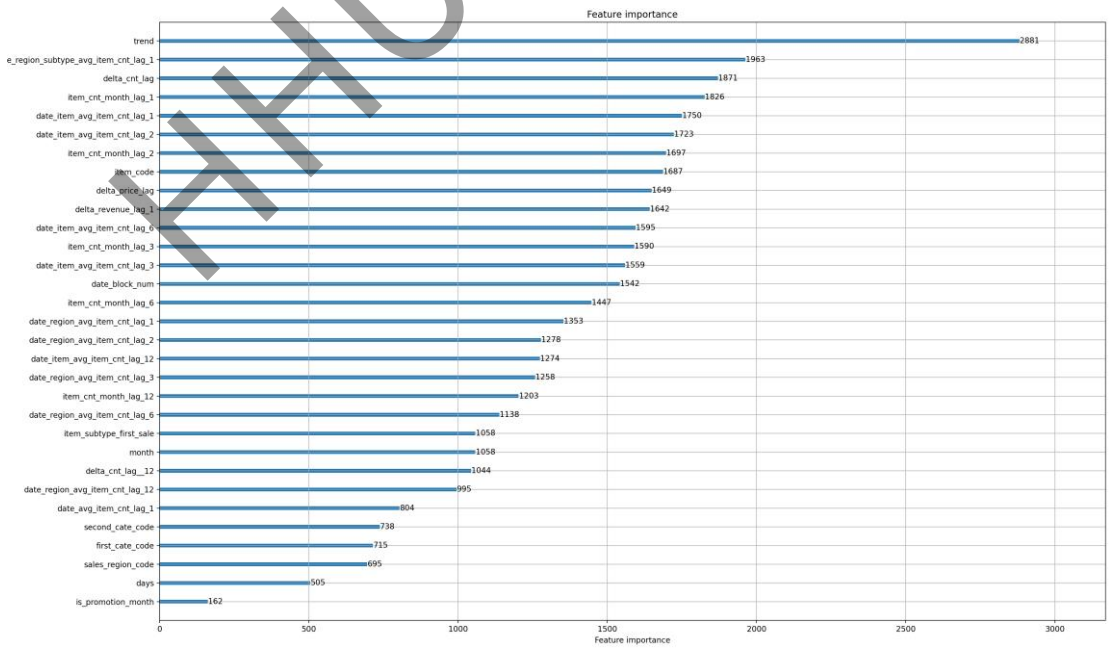


图 45 **全模型特征重要性图**

通过 lightgbm 的 plot_importance 函数绘制特征重要性图，可以发现排名靠前的特征分别是：trend(prophet 预测的季节性趋势)、一个月前的地区/细类商品平均月需求量、前几个月的销量增长趋势、一个月前商品需求量、两个月前商品需求量等等，说明了近期的价格和趋势特征是最重要的，也是我们模型预测准确的关键。

同时为了减少计算成本，避免过拟合和提高模型的可解释性，我们剔除排名最靠后的十个特征，重新训练后，验证集 WMAPE 提升 1.75%，Tweedie deviance 提升 2.63%

5.3 模型融合

本部分对新品重新建立预测模型，并对调参完成的多种模型利用随机森林算法进行模型融合，同时对是否融合 Prophet 模型的预测结果进行讨论。

5.3.1 新品模型

由于预测集中新品的数量较多，容易造成较大的预测偏差，所以本文改进已有模型，结合商品分层、滑动窗口算法并重新构造特征，构建了新品需求预测模型。

（1）特征方面。针对新品重新进行特征的构建，剔除了一些特征，比如商品历史同期销量、趋势。并新增了一些特征，包括同类商品第一次销售平均价格，同类商品第一次销售平均月销量、同类商品最近月销量、同类商品初期销售趋势、历史同期销售趋势、新品促销相关特征等。

（2）模型方面。定义距离发售月月数<5 个月的商品为新品，依靠滑动窗口算法，在构造好的结构化特征集中，筛选出每个月为新品的商品组合的数据。

通过 lightgbm 的 plot_importance 函数绘制该模型特征重要性图：

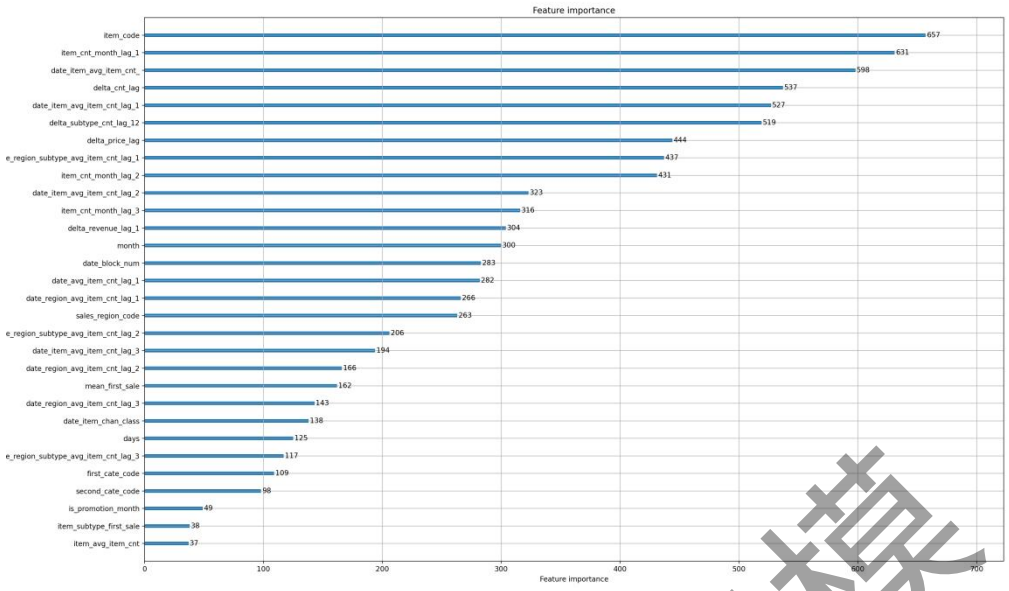


图 46 新品模型的特征重要性图

最后提取全模型中对新品预测的偏差和新品模型的预测偏差进行对比，如图所示：

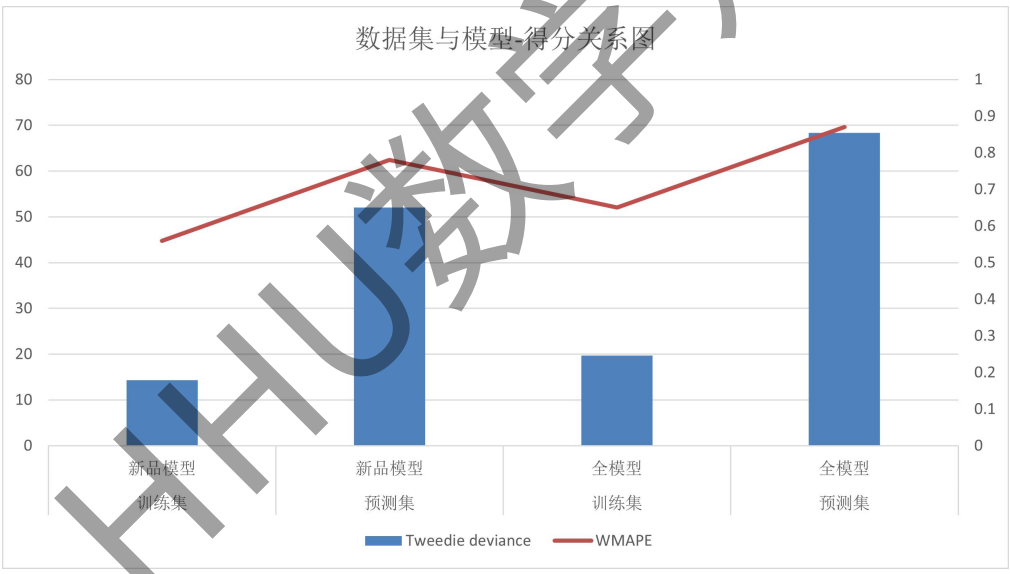


图 47 全模型与新品模型预测结果精度对比

经新品模型预测，新品测试集 Tweedie deviance 提升 24.68%

5.3.2 基于 GBDT-Prophet 的产品需求量预测模型

模型融合是指将多个不同的机器学习模型或深度学习模型集成在一起，以提高预测精度或性能。融合可以是简单的组合，例如投票或平均，也可以是更复杂的方式，例如堆叠或加权平均。为了进一步提升模型的预测精度、鲁棒性和可解释性，我们对四个基础模型， XGBoost、LightGBM、CatBoost 以及 Prophet 进行模型融合，以 Tweedie deviance 为指标，尝试的融合方法包括平均法、线性回

归、决策树和随机森林，同时对是否融合 Prophet 模型的预测结果进行讨论。

（1）利用平均法进行模型融合

在平均法中，将多个模型的预测结果简单求取平均值作为最终的预测结果。

（2）利用线性回归进行模型融合

机器学习模型融合可以通过多种方式实现，其中一种常见的方法是使用线性回归，其基本思想是通过对多个模型的预测结果进行线性组合，得到一个更好的预测结果。

（3）利用随机森林进行模型融合

本文采用随机森林算法进行模型融合。随机森林是一种集成学习方法，可以用于对多个决策树模型进行融合。利用随机森林进行模型融合的好处在于：它能够减少过拟合问题，并提高模型的鲁棒性和准确性；随机森林还能够处理高维度数据，对缺失数据也能较好地进行处理。

测试集 Tweedie deviance 结果如表 16 所示：

表 16 测试集 Tweedie deviance 结果

	平均法	线性回归	决策树	随机森林
融合 Prophet	38.36	28.49	15.04	12.72
不融合 Prophet	35.79	27.06	14.17	11.85

由表中可以看出，随机森林作为融合模型的预测效果最好，据此我们选择使用随机森林回归来组合基础模型的预测结果，同时不使用 Prophet 的预测结果，而是将 Prophet 预测的季节性趋势作为 GBDT 模型的特征。最终预测结果的损失值为 **11.85%**。

随机森林的模型参数设置如表 17 所示：

表 17 随机森林的模型参数

n_estimators	max_depth
80	4

5.3.3 预测结果

利用全模型对产品月需求量按月进行预测后，将按月、按周、按日进行预测的结果进行对比，由此可得出结论：按月进行预测的效果最好。

表 18 按月、周、日预测结果对比

预测精度	数据来源	Tweedie deviance	WMAPE
日	训练集	15.64	0.21
	测试集	77.87	0.44

周	训练集	16.93	0.24
	测试集	81.39	0.45
月	训练集	7.23	0.18
	测试集	11.85	0.41

第六章 总结

本文采用了基于 GBDT-Prophet 的产品需求量预测模型对未来三个月的产品月需求量进行预测。**该模型具有以下优点**:

1.本文依据对训练数据集的探索性分析以及对预测集的产品分布分析所得到的结论, 探索了该数据集中很多重要的规律, 并进行对应的预处理和特征量化, 对产品进行了商品分层这一特殊处理。

2.在建立最终的产品需求量预测模型时，我们通过能够灵敏地捕捉数据中季节性趋势的 Prophet 模型提取出训练数据中的季节性趋势作为全模型的特征，弥补了 GBDT 模型对季节性趋势捕捉效果的不足，从而提高预测精度。

3.通过对预测集中占比较大的“新品”单独构建了新品需求预测模型，降低了因新品历史数据过少、规律不明显所带来的预测偏差，新品测试集 Tweedie deviance 从 71.26 提升至 53.68，提升 24.68%.

4.本文对模型预测结果偏差较大的产品进行误差分析并将修正结果覆盖提交后，大大提升了预测准确度，Tweedie deviance 从 21.71 提升 17.13.，提升了 21.10%。

缺点:

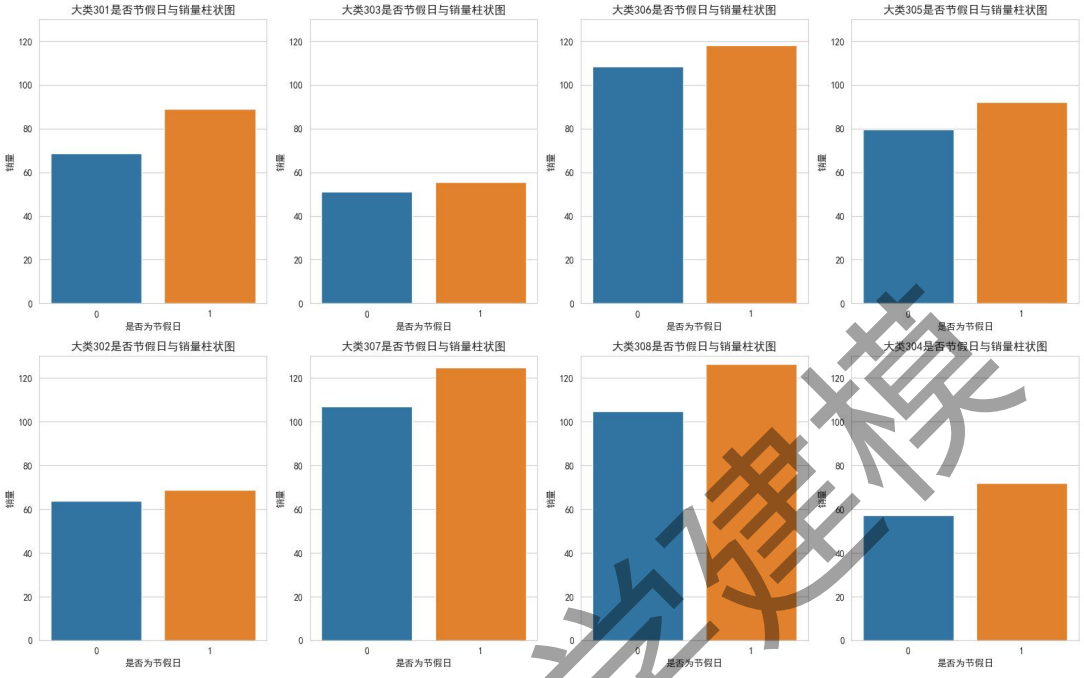
1.模型预测的结果较为保守，预测结果整体偏低，这是因为模型对于由促销因素造成的需求爆发这一现象的捕捉能力较弱；

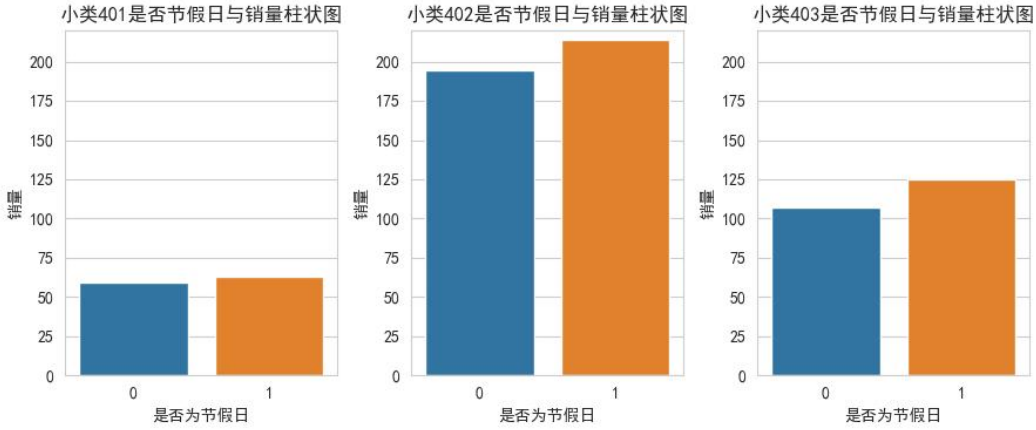
2.没有将产品进行进一步的分层。比如将需求曲线进行切割，然后使用时间序列相关性进行聚类，以此更好地找到商品之间的关系，提升预测效果。

3.直接对下一个月的产品需求量进行预测的效果较好，但是再往后预测的偏差较大。

附录

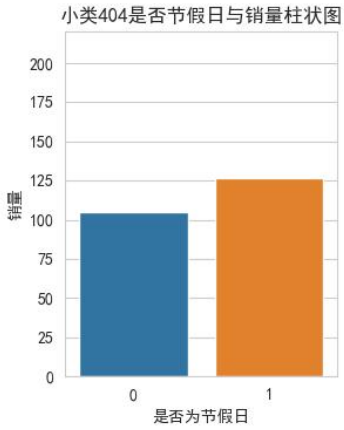
1.各大类、各细类产品在节假日/非节假日期间的订单需求量柱状图：

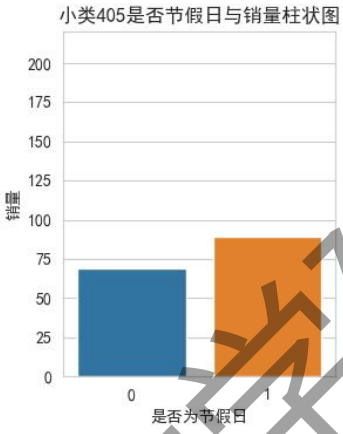


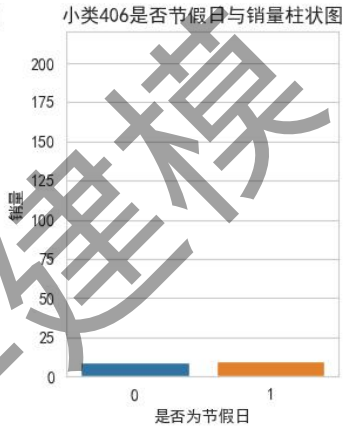

小类401是否节假日与销量柱状图
销量
是否为节假日
0
1

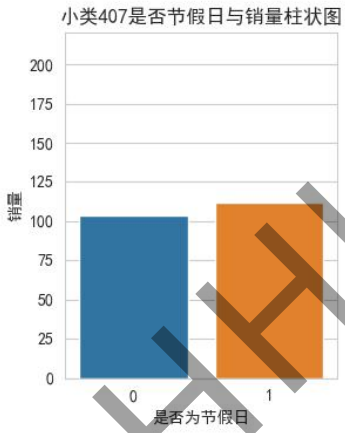
小类402是否节假日与销量柱状图
销量
是否为节假日
0
1

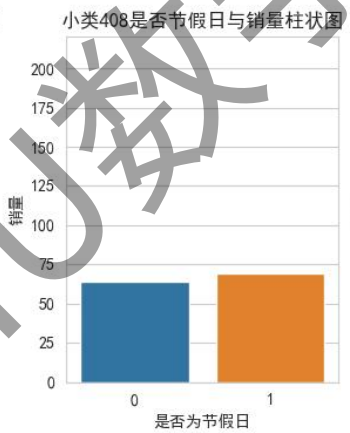
小类403是否节假日与销量柱状图
销量
是否为节假日
0
1

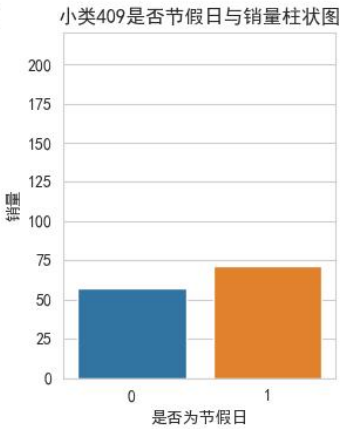

小类404是否节假日与销量柱状图
销量
是否为节假日
0
1

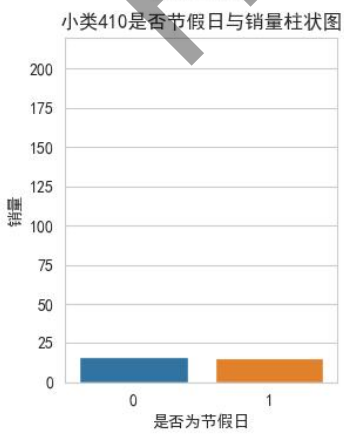

小类405是否节假日与销量柱状图
销量
是否为节假日
0
1

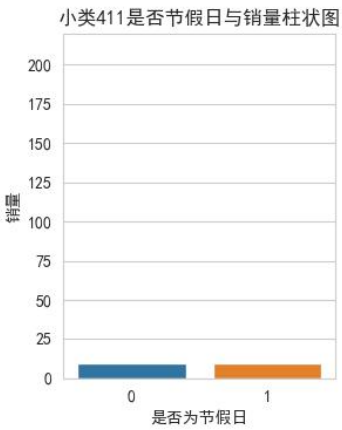

小类406是否节假日与销量柱状图
销量
是否为节假日
0
1

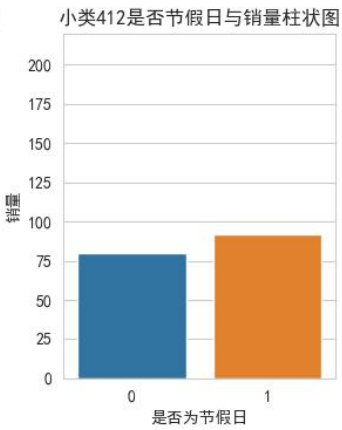

小类407是否节假日与销量柱状图
销量
是否为节假日
0
1


小类408是否节假日与销量柱状图
销量
是否为节假日
0
1


小类409是否节假日与销量柱状图
销量
是否为节假日
0
1


小类410是否节假日与销量柱状图
销量
是否为节假日
0
1


小类411是否节假日与销量柱状图
销量
是否为节假日
0
1


小类412是否节假日与销量柱状图
销量
是否为节假日
0
1

2.异常产品的预测结果:

参考文献

[1]周雨,段永瑞.基于聚类与机器学习的零售商品销量预测[J].计算机系统应用,2021,30(11):188-194.

[2]林木兴.需求不确定下的大规模数据高斯过程回归的商品销量预测模型研究[D].暨南大学,2020.

[3]赵钰逸,王春晓,吴桥.基于组合模型的中小电商商品短期需求预测[J].浙江万里学院学报,2023,36(01):7-13.

[4]吴庚奇,牛东晓,耿世平等.多价值链视角下基于深度学习算法的制造企业产品需求预测[J].科学技术与工程,2021,21(31):13413-13420.

[5]Trancy,林晓辰,nonameh. 机器学习- MBA 智库百科 (mbalib.com),[2023-4-24].

[6]王欢. 基于深度学习的短期电力负荷预测模型研究[D].云南大学,2021.

[7]邱连成. 基于 Prophet 和 LSTM 模型的建筑企业设备备件需求预测[D].内蒙古科技大学,2022.

[8]孙铭. 基于 LightGBM 的超市商品销量预测[D].大连理工大学,2021.

[9]弗朗索瓦·肖莱.Python 深度学习[M]. 北京: 人民邮电出版社, 2018.

[10]苌瑶,武建文,马速良,邵阳,林靖怡,梁传涛,杨宁.基于多特征评估与 XGBoost 的高压断路器故障诊断[J].高压电器,2023,59(04):1-9.

[12]潘少伟,王朝阳,张允等.基于长短期记忆神经网络补全测井曲线和混合优化 XGBoost 的岩性识别[J].中国石油大学学报(自然科学版),2022,46(03):62-71.