第十一届“泰迪杯”
数据挖掘挑战赛

优秀作品

作品名称：产品订单需求预测分析

荣获奖项：一等奖并获网宿创新奖

作品单位：华南师范大学

作品成员：黄洁华 刘雨岚 陈可心

产品订单需求预测分析

摘要

对产品订单需求的分析和预测一直是企业优化生产，降低风险的一项重要手段。只有充分了解产品需求才能更好的进行生产活动；只有提高需求预测的准确性，才能对未来的运营计划做决策参考。因此，我们将基于国内某大型制造企业在 2015 年 9 月 1 日至 2018 年 12 月 20 日面向经销商的出货数据，对产品订单需求量进行特征分析及预测分析。

由于历史数据可能在收集和传输时出现一定的错误，因此先要对出货数据的进行预处理，并对已有特征的有效信息进行处理与提取，构建新的特征，得到最终可以进行分析的订单数据。

针对问题 1，首先通过**逐步回归法**进行特征筛选，除促销日外其他数据特征都与需求量有关。对产品订单数据特征进行深入分析，①通过产品价格对产品需求量的影响探究发现不同类型产品的价格对需求量的影响不同，对于大部分产品，产品价格越低，需求量越高，但整体上需求量会在一个稳定的价格区间内波动；②不同的销售区域内产品的需求量以及平均每订单产品需求量有一定差异，这可能取决于销售区域与制造企业的距离，距离远会增加产品的运输成本；③整体上在节假日上的日需求量会小于非节假日的日需求量，但不同类型的节假日对产品需求量的影响程度不同，其中春节对需求量的影响最大；④在一些大型促销活动如 618、双 11，经销商需在促销活动开始前提前一个月或两个月订购产品，避免因企业出货晚使得促销活动无法顺利进行，这也导致了需求量的季节性波动；⑤经销商倾向于在气温适宜的春秋季节订购较多的产品，在天气炎热的夏季产品需求量最低，而冬季气温低下，又逢春节，其产品需求量仅高于夏季；⑥一般来说线下实体经销商的产品需求量远高于线上电商平台的需求量，只有在大型电商促销活动如双 11 的影响下，线上的需求量才会短暂高于线下；⑦ 一般产品需求量在月头时较低，在月末时较高，需求量会呈现月周期性波动。基于上述对需求量的特征分析，再深入探究不同品类间产品需求量在这些特征影响下的相同点与不同点。

针对问题 2，为了从不同时间粒度对每个产品未来三个月的需求量进行预测，将每个产品的缺失日期补全，对应需求量填充为 0，并按天、周、月粒度进行处理并进行特征筛选。以 2015 年 9 月-2018 年 9 月的历史数据作为训练集，以 2018 年 10 月-2018 年 12 月的历史数据作为测试集，基于产品需求量的数据特征以及月周期性，构建**随机森林模型**、**XGBoost 模型**、**Prophet 模型**，比较在测试集上的预测效果，得到在相同时间粒度下，对于 MAE 和 RMSE：Prophet 模型＞随机森林模型＞XGBoost 模型。对于 R^2：Prophet 模型＜随机森林回归模型＜XGBoost 模型。**XGBoost** 的预测效果要优于其他两种模型。在这三种预测模型中，R^2 的值都是月粒度＞周粒度＞天粒度，因此，三种时间粒度下预测精度的大小为：月粒度＞周粒度＞天粒度，由此可见，对天粒度的产品需求量进行预测的准确率是最低的，对**月粒度**预测精度会更高，这刚好符合企业和经销商对产品需求量的预测需求，因为为了方便生产，企业只需要大致了解未来一个月的产品需求量以制定生产计划，由于产品需求量具有月周期性，且天粒度的需求量具有随机性，所以企业对需求的预测并不需具体到天。

最后，我们对月数据进行特征工程处理，加入**滞后**、**差分**、**拓展窗口**、**滑动窗口**特征，对 XGBoost 模型进行改进。使用改进的模型对月粒度数据进行预测，改进的 XGBoost 模型在测试集上的 MAE 和 RMSE 均小于原来的模型，且 R^2 提高到 0.57，说明改进的 XGBoost 模型的预测精度得到了进一步的提升，并基于改进后的 XGBoost 模型对 2619 种产品进行预测，新产品使用对应产品细类历史数据的时间序列均值，预测结果见 result1.xlsx。

关键词：逐步回归法 时间序列 特征工程 随机森林 XGBoost 模型 Prophet 模型

目录

1 绪论

1.1 问题背景

近年来，由于复杂多变的市场环境，使得企业的外部环境变得越来越不确定，一些制造企业的产品供应面临较大风险。因此，在产品价格确定的情况下，销售量的确定和产品需求的稳定性就显得十分重要，因为会直接影响到企业的利润和收益。了解客户需求是企业供应链上不可或缺的环节，只有了解市场需求，企业的业务计划才能更有效地实施，也能更好地响应市场需求。

了解产品需求的重要方式是需求预测。需求预测作为企业供应链的第一道防线，对不同行业的企业起着重要作用，尤其是在降低商业活动风险方面。需求预测是基于历史数据对未来需求进行预测得出的有理论依据的结论，有利于公司管理层对未来的销售和运营计划做决策参考，也有助于采购计划和安排生产计划的制定，减少业务波动的影响。

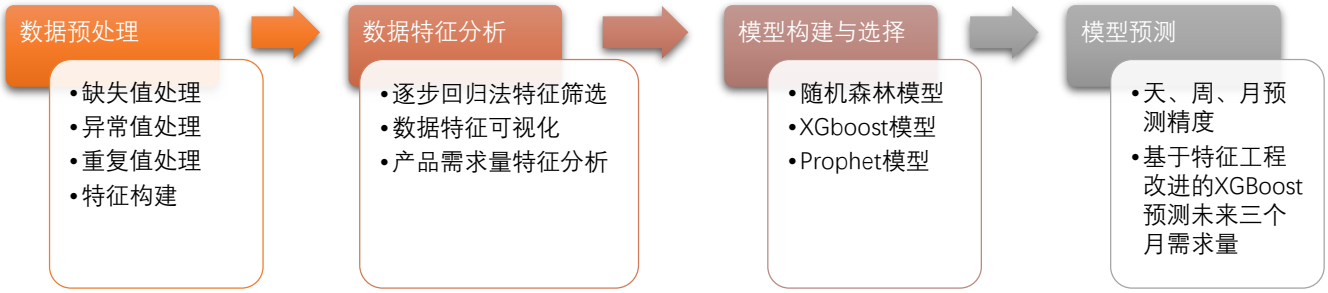
采取适当的预测方法把握市场需求信息并将其作为决策依据已经成为企业获取竞争优势的有效手段。然而，由于需求预测受多种因素影响，如价格、时间、促销等，其分析的复杂性，导致预测的准确率普遍降低，使得公司内部产生库存和资金积压或不足等问题，增加企业的库存成本，因此，如何使用更优秀的预测方法，提高预测的准确性，也是公司亟待解决的问题。

1.2 问题重述

问题一：基于国内某大型制造企业在 2015 年 9 月 1 日至 2018 年 12 月 20 日面向经销商的出货数据，深入分析产品需求量的特征及其影响因素。

问题二：基于问题一分析，建立数学模型，分别按天、周、月的时间粒度预测不同产品未来 3 个月的月需求量，并分析不同的时间粒度对预测的精确度的影响。

1.3 主要工作



2 问题分析

针对问题一，基于附件的训练数据，对数据进行预处理，通过逐步回归法探究产品需求量与各项数据特征如产品价格、产品销售区域、节假日、促销日以及季节因素的关系，对数据特征进行筛选。并通过数据可视化，分析这些特征对产品需求量的影响，研究在不同销售方式以及不同时间段下产品需求量的特性。接着，基于以上分析，将产品根据大类编码和小类编码进

行分类，通过在对比不同品类在价格、销售区域、节假日、促销活动和季节的影响变化，以及不同种类在不同销售方式以及不同时间段时的特性，分析不同品类之间产品需求量的共同点和不同点。

针对问题二，将预处理后的数据处理为按天、周、月三种时间粒度的数据，并将每个产品缺失日期进行补充，并将对应产品需求量填充为 0。基于问题 1 的数据特征分析筛选特征，我们将建立 ARIMA、SARIMA、Prophet 时间序列模型和随机森林、XGBoost 机器学习模型，以 2015 年 9 月-2018 年 9 月的历史数据作为训练集，以 2018 年 10 月-2018 年 12 月的历史数据作为测试集，分别按天、周、月三种时间粒度的数据集进行训练和测试，并进行调参和预测效果的评估，进行不同模型之间预测效果的比较与选择，并分析不同的预测粒度对预测精度产生的影响。

最后，选择预测效果最佳的模型进行特征工程处理，加入滞后、差分、拓展窗口、滑动窗口特征，对模型进行改进，从而提高预测精度。通过改进的模型，选择最高预测精度的时间粒度对模型进行训练，进而对文件中 2619 个产品进行预测，其中有 432 个无任何数据的新产品，我们将通过相同的产品细类的历史数据的均值作为这些新产品历史数据的替代，最后将 2619 个产品预测结果保存到文件 result1.xlsx 中。

3 模型假设

（1）假设附件中提供的数据都是真实可靠的；
（2）假设每个产品在没有订单数据的日期上的需求量为 0；
（3）假设 2019 年 1-3 月无突发事件对产品需求量产生影响。

4 数据预处理

对产品需求量的特征和影响因素分析以及对需求量的预测都是基于某企业历史出货数据进行的，而这些历史数据在采集和传输时可能存在一定的误差，因此需要分别对数据的缺失值、异常值和重复值进行清洗和处理，并且将数据中的日期格式进行转换，经过处理的数据集不仅更方便进行特征和影响因素分析，也对提升预测的真实准确性具有重要的作用。

4.1 缺失值、异常值和重复值的处理

4.1.1 对缺失值的处理

由于不同产品在不同日期的需求量不同，某些产品的需求量为 0 是较为正常的现象，因此只需考虑每天是否有产品需求即可。

使用 Python 中 pandas 库下 Dataframe 类的 isnull().sum()方法，查看数据集中是否存在缺失值，发现已有训练数据集中并不存在缺失值。但对数据进一步探索发现，2015 年 10 月 1 日到 10 月 3 日，2016 年 1 月 1 日，2016 年 2 月 4 日到 2 月 14 日、2017 年 1 月 24 日到 2 月 2 日、2018 年 2 月 12 到 2 月 21 日这几天中数据是缺失的，说明在这些时段内该企业可能没有向经销商出售产品。考虑到这些数据缺失的时段正值国庆、元旦和春节假期，没有销售产品也符合正常逻辑，因此，我们将缺失值填充为 0。

4.1.2 对异常值进行处理

通过 Python 中 pandas 库下 Dataframe 类的 describe()方法进行统计分析，数据整体情况如表 4.1 所示，可以发现价格和需求量的最小值和最大值的差异较大，两者的最大值与 75%分位数的差异也很大。

表 4.1 价格和需求量整体情况

	date	item_price	ord_qty
count		597694	597694
mean		1076.2416	91.6505
std		1167.5111	199.8433
min	2015/9/1	1	1
25%		598	10
50%		883	29
75%		1291	101
max	2018/12/20	260014	16308

将出货数据中的价格和需求量进行可视化，如图 4.1 所示，存在少数离群值。

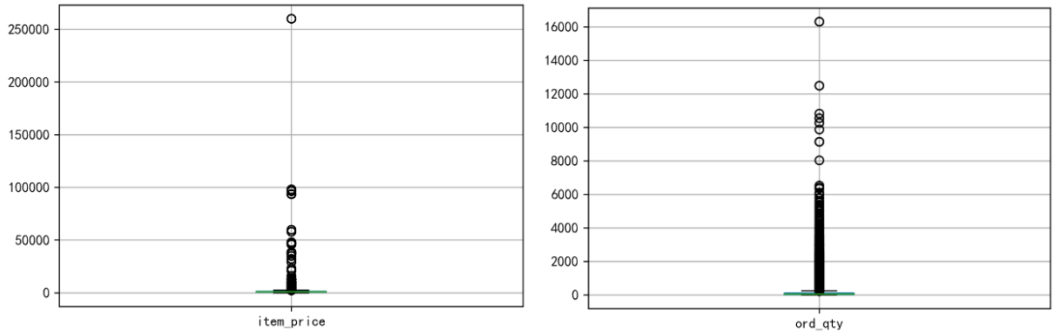


图 4.1 价格和需求量图示

产品价格方面，部分离群值超过了 250000。通过数据对比，如表 4.2 所示，发现离群值都在同一产品，且在其他时间产品价格数量级仅为 10^2。根据消费者行为理论分析，一般来说经销商不可能在产品价格比三个月前翻了 300 倍的情况下还选择购买产品，这会大大提高自己的成本，还会面临消费者不再对其进行消费的风险。

表 4.2 异常数据与正常数据对比

order_date	sales_region_code	item_code	first_cate_code	second_cate_code	sales_chan_name	item_price	ord_qty
2018/3/14	103	21801	301	405	offline	849	155
2018/6/16	103	21801	301	405	offline	852	53
2018/9/30	103	21801	301	405	online	260014	9
2018/10/1	103	21801	301	405	online	260006	15
2018/10/22	103	21801	301	405	offline	501.7	11

因此，把离群值替换成最近的正常值，需求量对应的价格分布结果如图 4.2 所示：

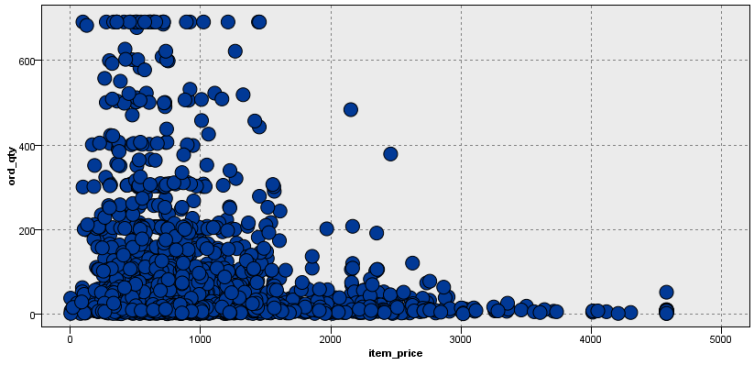


图 4.2 异常值处理后的价格分布

而产品的需求量数值差异较大可能与其他因素有关，这也是我们正亟待解决的问题，因此不考虑将离群值作为异常值进行处理。

4.1.3 对重复值进行处理

使用 Python 中 pandas 库下 Dataframe 类的 duplicated().sum()方法，对出货数据进行重复值分析，发现在 597694 条原始数据中存在重复数据 312 条，用 Dataframe 类的 drop_duplicates()方法，将重复值删除，剩余数据为 597382 条。

4.2 产品价格和需求量的处理

（1）对于相同产品在同一日期同一销售区域同一销售方式下有不同的价格的情况，利用 Python 中的 groupby 函数和 agg 函数，取所有价格的平均值作为该产品在这一天的最终价格；

（2）对于相同产品在同一日期同一销售区域同一销售方式下有多个需求量的情况，利用 Python 中 groupby 函数和 agg 函数，取所有需求量的总和作为该产品在这一天的最终需求量；

通过以上两个步骤，即可得到每个产品在同一天同一销售区域内的唯一价格和需求量。如下表 4.3 所示，编码 20027 的产品在 2017 年 1 月 3 日有多个价格和需求量，因此在当天该产品在 101 销售区域的最终价格为 955，最终需求量为 166。

表 4.3 价格和需求量处理示例

date	sales_region_code	item_code	first_cate_code	second_cate_code	sales_chan_name	item_price	ord_qty
2017/1/3	101	20027	301	405	offline	952	23
2017/1/3	101	20027	301	405	offline	959	19
2017/1/3	101	20027	301	405	offline	954	124

4.3 已有特征的处理

4.3.1 销售渠道名称的处理

由于数据中销售渠道的名称为“offline”和“online”，无法被算法识别，因此需要将两类标签映射到整数值进行编码。'offline'用 0 编码，'online'用 1 编码。

4.3.2 出货日期的处理

将 Dataframe 中的’order_date’列日期数据的 object 格式转为时间戳格式，方便后续日期有关特征的提取。

4.4 新特征的构建

为了更多地提取原数据集特点的有效信息，方便后续对训练数据的分析以及预测模型的建立，我们基于已有的数据特征以及自身经验构建新的衍生数据特征，将原始数据的有效信息充分挖掘，所构造的新特征如下表 4.4 所示：

表 4.4 新特征描述

特征列名	意义	描述
year	年	2015-2018 分别表示 2015 年至 2018 年
month	月	1-12 分别表示 1-12 月
day	日	1-31 分别表示 1-31 号
weekday	星期	1-7 分别表示周一至周日
season	季节	1-4 分别表示春夏秋冬
month_period	月份时间段	1、2、3 分别表示月头、月中、月末
holiday_type	节假日类型	分别代表各个节假日
is_holiday	是否节假日	1 表示节假日，0 表示非节假日
sales_promotion_type	促销日类型	分别代表各个促销日
is_sales_promotion	是否有促销	1 表示促销日，0 表示非促销日

根据原数据信息及生活经验，将构造的新特征数据进行填充。其中，在季节特征中，春季为 3、4、5 月，夏季为 6、7、8 月，秋季为 9、10、11 月，冬季为 12、1、2 月。

月份时间段的特征中，月头为 1-10 号，月中为 11-20 号，月末为 21-31 号。

节假日类型的特征中，仅考虑国家传统节假日（暂不考虑特殊节日），包括：元旦、春节、清明节、劳动节、端午节、中秋节和国庆节。用 0 表示非节假日，1-7 分别表示各个节假日类型，具体编码方式及各节假日具体日期如下表 4.5（其中各个节假日的具体日期源于国务院发布的中国法定节假日公休安排的通知）

表 4.5 节假日特征描述

	类型编码	2015	2016	2017	2018
元旦	1		1.1-1.3	12.31-1.2	12.30-1.1
春节	2		2.7-2.13	1.27-2.2	2.15-2.21
清明	3		4.3-4.5	4.2-4.4	4.5-4.7
劳动	4		5.1-5.3	4.29-5.1	4.29-5.1
端午	5		6.9-6.11	5.28-5.30	6.16-6.18
中秋	6	9.26、9.27	9.15-9.17	10.4	9.22-9.24
国庆	7	10.1-10.7	10.1-10.7	10.1-10.3, 10.5-10.8	10.1-10.7

促销日类型的特征中，用 0 表示非促销日，1-6 分别表示各个促销日类型，各个促销日的编码及各促销日具体日期如表 4.6 所示。

表 4.6 促销日特征描述

促销日	情人节	妇女节	618	国庆	双十一	双十二
日期	2.14	3.8	6.18	10.1	11.11	12.12
编码	1	2	3	4	5	6

基于上述新特征描述，对数据进行处理和补充，其中前五条数据如图 4.3 所示。

	sales_chan_name	item_price	ord_qty	year	month	day	weekday	season	month_period	holiday_type	is_holiday	sales_promotion_type	is_sales_promotion
0		1012.0	12	2015	9	1	2	3	1	0	0	0	0
0		1114.0	19	2015	9	1	2	3	1	0	0	0	0
0		2996.0	18	2015	9	2	3	3	1	0	0	0	0
0		99.0	502	2015	9	2	3	3	1	0	0	0	0
0		164.0	308	2015	9	2	3	3	1	0	0	0	0

图 4.3 新特征数据示例

5 问题 1.1：产品需求量特征分析

产品需求量是取决于经销商向该大型制造企业的订货数量。一般来说，经销商具有独立的经营机构，通过从企业进货后进行商品销售，从而获得经营利润，在经营活动过程不受或很少受供货商限制，与供货商责权对等。因此，经销商与大型制造企业一般会签订长期合作协议，经销商需要确定产品在未来一段时间内的大致价格范围，制造企业需要确定经销商在未来一段时间内的订单需求，从而使双方互利互赢。因此，产品需求量会在一定程度下呈现周期性波动，这是经销商和制造企业基于市场需求共同作用的结果。

接下来，我们将基于预处理后的数据，对产品需求量进行特征分析。

5.1 逐步回归法—特征的筛选

为了研究数据特征对产品的需求量是否有显著的线性关系，我们使用逐步回归法进行线性回归。逐步回归法通过剔除变量中不太重要又和其他变量高度相关的变量，降低多重共线性程度，因此我们以需求量为因变量 y，使用逐步回归法进行特征变量的筛选。

逐步回归的基本思想是有进有出。具体做法是将变量一个一个地引入，每引入一个自变量后，对已选入的变量要进行逐个检验,当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。引入一个变量或从回归方程中剔除一个变量,为逐步回归的一步,每一步都要进行 F 检验,并对已经选入的解释变量逐个进行 t 检验，以确保每次引入新的变量之前回归方程中只包含显著的变量。这个过程反复进行,直到既无显著的自变量选入回归方程,也无不显著的自变量从回归方程中剔除为止。弥补了前进法和后退法各自的缺陷,保证了最后所得的回归子集是最优回归子集。

使用 SPSS 进行逐步回归，一共得到 14 个回归模型，最终的模型 14 的自变量有 14 个。回归结果如表 5.1-表 5.4 所示。

表 5.1 模型摘要

模型摘要				
模型	R	R 方	调整后 R 方	标准估算的错误
14	.197	.039	.039	289.836

表 5.2 方差分析表

模型	平方和	自由度	均方	F	显著性
回归	1380149516.993	14	98582108.357	1173.530	.000
残差	34006652356.890	404818	84004.793		
总计	35386801873.883	404832			

表 5.3 回归系数表

变量	未标准化系数		标准化系数	t	显著性
	B	标准错误	Beta		
(常量)	20789.628	1023.064		20.321	.000
item_price	-.027	.000	-.108	-68.732	.000
first_cate_code	13.178	.233	.089	56.630	.000
sales_region_code	-21.005	.360	-.110	-58.274	.000
sales_chan_name	64.984	1.246	.099	52.145	.000
second_cate_code	-3.912	.146	-.042	-26.864	.000
year	-10.416	.505	-.033	-20.635	.000
season	10.206	.448	.038	22.758	.000
weekday	3.425	.239	.022	14.336	.000
month	-1.666	.142	-.020	-11.719	.000
item_code	.004	.001	.008	5.009	.000
holiday_type	-8.787	1.414	-.033	-6.214	.000
is_holiday	45.338	8.158	.030	5.558	.000
day	1.231	.157	.035	7.817	.000
month_period	-12.472	1.654	-.034	-7.541	.000

表 5.4 排除的变量

	输入 Beta	t	显著性
sales_promotion_type	-.001	-.613	.540
is_sales_promotion	-.001	-.559	.576

表 5.2 方差分析表显示 p 值小于 0.05，拒绝原假设，可认为在显著性水平 α =0.05 下，需求量 y 至少与自变量中的一个变量有显著的线性关系，即回归方程是显著的。

表 5.3 中各个变量的的 p 值小于 0.05，说明显著性水平 α =0.05 下，需求量与变量 item_price、first_cate_code、sales_region_code、sales_chan_name、second_cate_code、year、season、weekday、month、item_code、holiday_type、is_holiday、day、month_period 有显著的线性关系。需求量与促销类型和是否促销无显著的线性关系，这可能是因为经销商不是在促销当日才订货，而通常是在促销前一段时间就已订货，以备促销日时能够销售出更多的产品。

5.2 产品价格对需求量的影响

为了观察所有产品的不同价格下需求量的总体情况，利用 SPSS Modeler 绘制所有产品的不同价格下需求量的散点图，如图 5.1 所示。

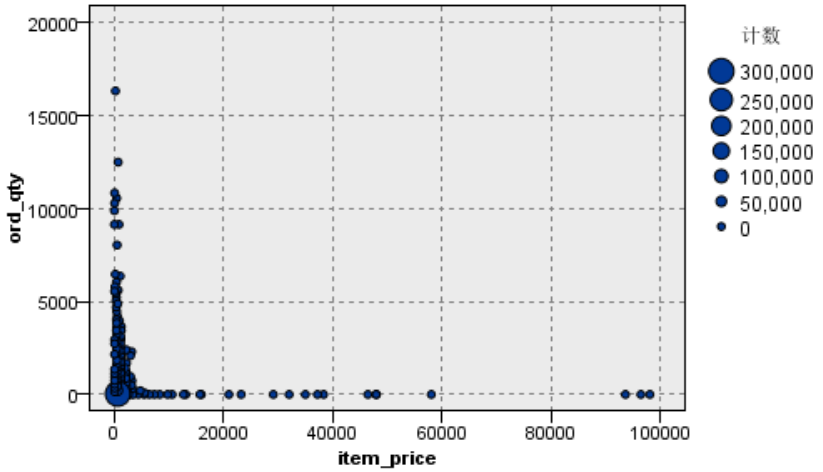


图 5.1 不同价格下的需求量

为了使结果更加清晰，利用 Python 将价格分段，从 0-100000 将价格分成 10 个区间，再用 groupby 方法分组统计各个价格范围下的平均订单需求量，并绘制柱状图，如图 5.2 所示可观察到所有产品的不同价格范围内需求量的对比。

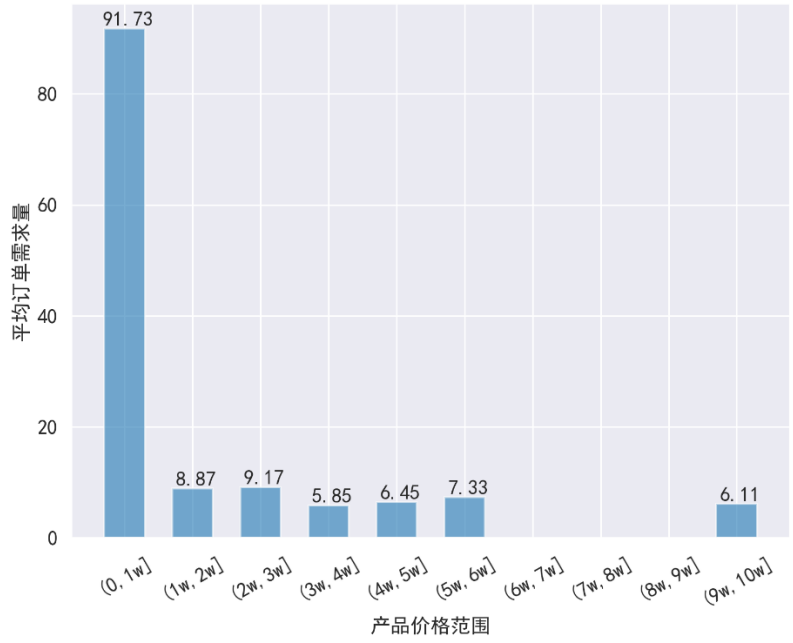


图 5.2 不同价格区间的需求量

由图上可得出，需求量基本集中在价格较低的区间内，说明价格越低，产品的需求量越多。

经济学上称购买产品时所能接受的最高价格为保留价格，不同的经销商的保留价格并不相同，但是价格上升的影响却是一致的。当产品价格升高甚至超过了保留价格时，经销商们可能会选择同个品类中的其他相应的替代品或者减少直接减少订单以期望价格降低再进行采购，从而使得的需求量降低。

上述是在没有将产品分类情况下做大致分析，但由于所有产品中包含了多个种类，为了更具可比性，进一步将产品分为八大类，研究同一类产品下，不同价格对需求量的影响。

由于每个产品大类下的产品很多，仅选取了八个大类下订单数量最多的产品进行分析，分别是产品编码为 20003、21715、20717、22040、22046、20271、21081、21061，并绘制这八类产品的散点图。

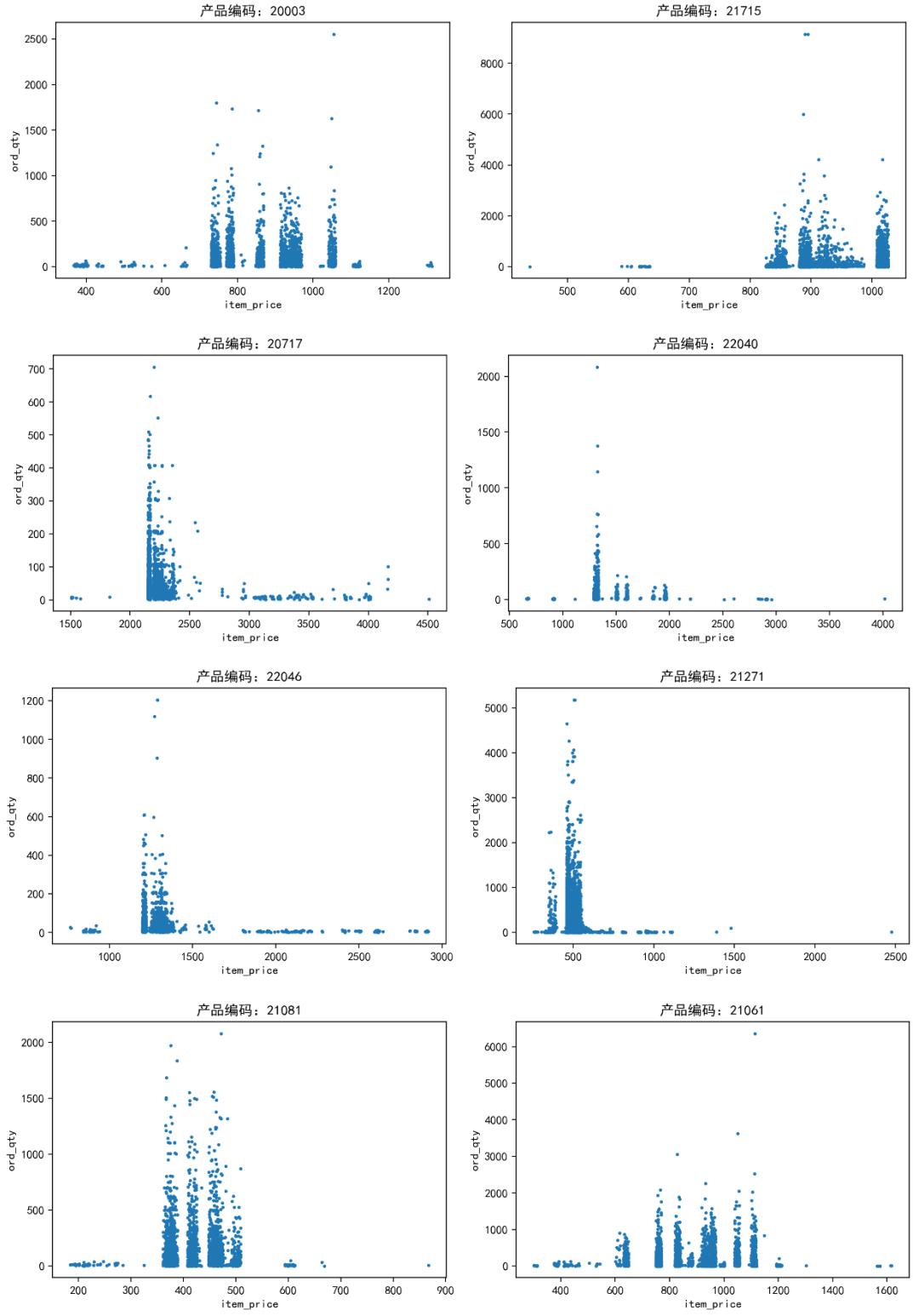


图 5.3 八类产品价格下的需求量

一般来说，每个产品都有最高价格，且不同的产品有不同的最高价格，超过最高价格的产品说明其价格已经远超过产品的价值，这时该产品的需求量就会减少。当然，需求量也受其他

因素的影响。从图中可知，当产品价格较低，其产品的需求量也可能降低，这时影响需求量的主要因素不再是价格，在产品销售淡季时，即使价格再低，其需求量也不会升高。

从图中可以发现，这些产品基本上会一个确定的价格区间上波动，在这个区间里就包含了能使需求和供给相等的均衡价格。不同产品有不同的均衡价格，一般来说，低于或高于均衡价格时产品的需求量都会减少，所以企业应该合理制定产品价格以保证需求量的稳定。

5.3 产品销售区域对需求量的影响

为了探究产品销售区域对产品需求量的影响，利用 Python 的 groupby 方法分组统计不同区域的订单数、总订单需求量、平均订单需求量，如表 5.5 所示。

表 5.5 不同销售区域的订单情况

	订单数	总订单需求量	平均订单需求量
101	82901	12400949	149.587448
102	99406	13966622	140.500795
103	75375	11519878	152.834202
104	14955	2387342	159.635038
105	132196	14495521	109.651737

根据上述结果，绘制不同区域的总订单需求量占比饼图，如图 5.4 所示，并绘制平均订单需求量的柱状图，如图 5.5 所示。

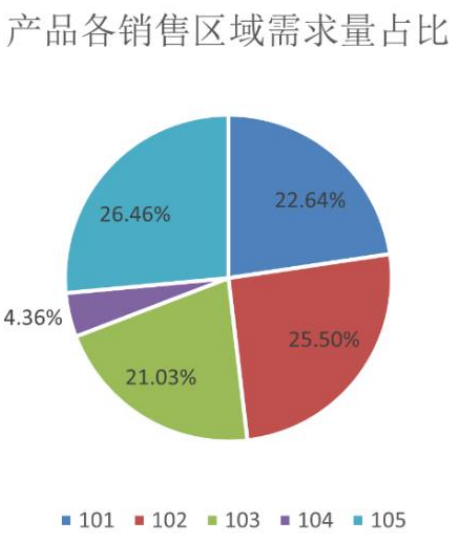


图 5.4 不同区域总订单需求量占比

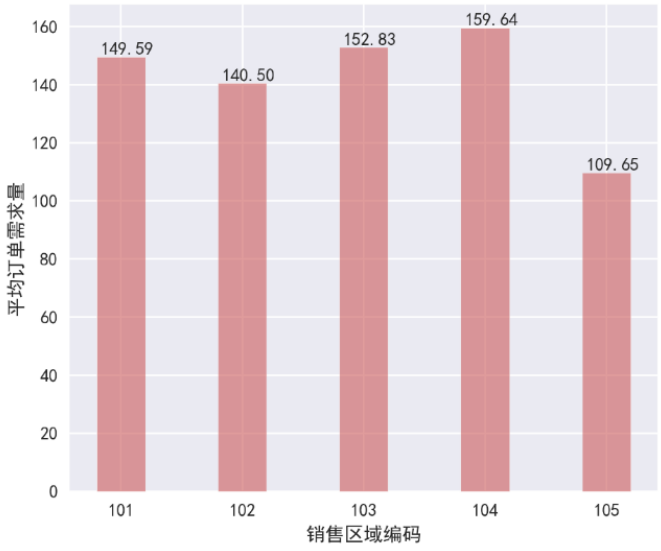


图 5.5 不同区域平均订单需求量

由上述结果可明显看出，104 销售区域的总订单需求量占比最小，105 销售区域的总订单需求量最大，说明不同的销售区域内产品的需求量是不同的。

而 104 的平均订单需求量最高，105 的平均订单需求量最低，说明 104 销售区域内的经销商实行的是多量少次取货，105 销售区域内的经销商采用少量多次取货，这可能和销售区域与该大型企业的距离有关。104 销售区域的经销商距离该企业较远，运输成本较高，相对较近的 105 销售区域所付出的成本更高，所以需求量相对更低，且每一次取货都尽可能提高订单的需求量。这说明了销售区域与产品的需求有关，如果销售区域距离供货地较远，使得经销商的运输成本增加，导致产品的实际价值增大，需求量则降低。

101、102、103 销售区域总产品需求量和平均订单数差异不大，推测可能这三个销售区域的需求市场以及与该大型企业的距离大致相同。

为了更清晰地体现不同销售区域下订单需求量的差异，绘制了不同销售区域的月需求量折线图，如图 5.6 所示。

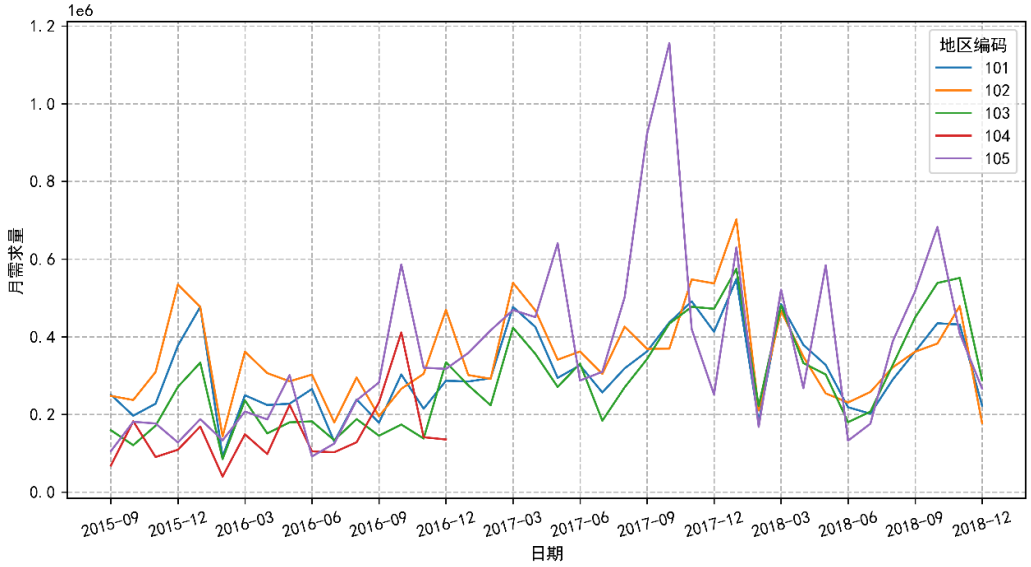


图 5.6 不同销售区域的月需求量变化

由上图可知，104 销售区域在 2017 年就停止向该企业进行订货。而在 2015 年-2016 年的月需求量也相对较少，在 2016 年秋季时期需求量达到最高峰，推测可能该销售区域的经销商在合作结束前夕向该企业采购了较多产品，通过囤积产品以应对交易结束时暂时的产品供应。

而其他销售区域，105 销售区域的月需求量波动幅度较大，且在一年中订单需求量的高峰一般处在春季和秋季，最高峰在 2017 年的秋季。推测在该销售区域中，经销商可能会在订购大批量产品后，在一段时间内进行售卖，并少批量地进货以维持库存，直到产品库存达到再订货点时，再进行第二次大批量采购，如此循环，呈现一定程度上的周期性波动。

销售区域 101、102、103 的月需求波动幅度较小，虽然订单需求量有一定差异，但三个销售区域上订单需求量随时间的变化方向基本是一致的，这也符合上述对三个销售区域内市场需求基本一致的猜想。

综上，产品销售区域对需求量是有影响的，不同销售区域的需求量不同，推测这可能与销售区域消费者的消费习惯与销售区域与供货企业的距离有关。

5.4 节假日对需求量的影响

为了探究节假日对产品需求量的影响，根据国家法定节假日定义 2015 年 9 月 1 日-2018 年 12 月 20 日的所有假期，包括元旦、春节、清明节、劳动节、端午节、中秋节、国庆节。这些节日假期称为节假日，其他普通日期设为非节假日。绘制 2015 年 9 月 1 日-2018 年 12 月 20 日每日订单需求量的变化如图 5.7 所示，其中节假日用红星标出。

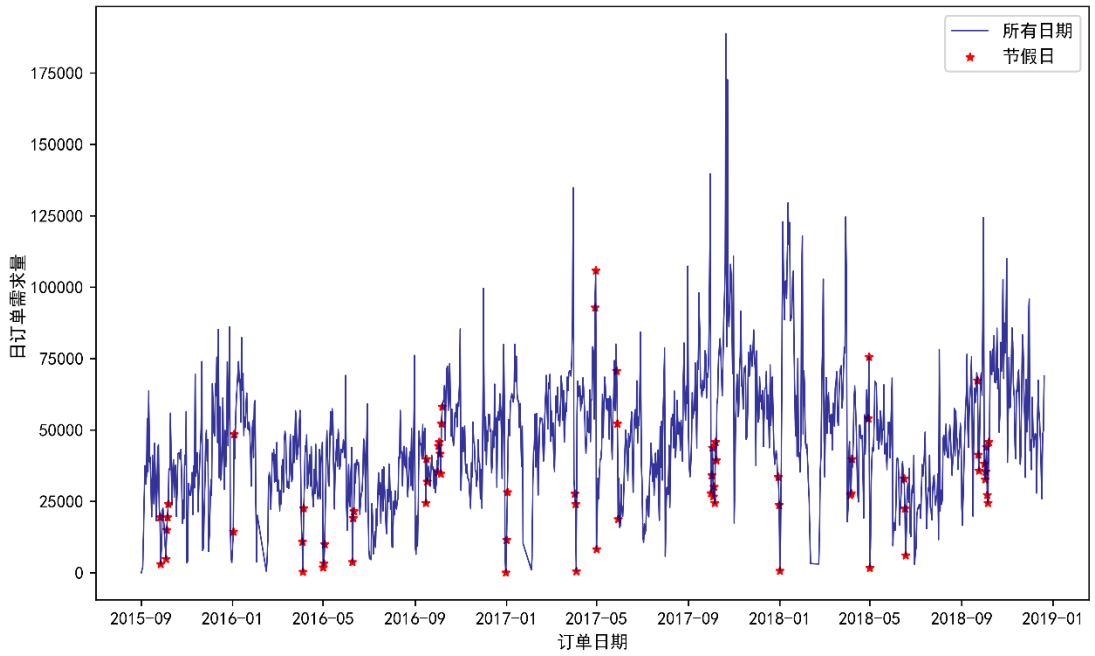


图 5.7 节假日与非节假日每日订单需求量变化

在 4.1 对缺失值进行处理时就发现，春节长假期间并没有订单需求量，这可能是因为企业员工或经销商春节停工，买卖双方没有进行任何交易导致的。由图中也可明显看出，除了部分节假日的需求量不受影响外，大部分节假日的产品需求量都处在较低位置，说明节假日可能会导致产品需求量的降低。

由于节假日的天数比非节假日少，节假日订单总数明显少于非节假日，因此不考虑节假日与非节假日总订单需求量的对比，只分析平均每订单需求量和平均日订单需求量，如表 5.6 所示。

表 5.6 节假日与非节假日的订单情况

	订单数	天数	平均每订单需求量	平均日订单需求量
非节假日	389244	1103	135.302193	47747.567543
节假日	15589	90	135.014754	30503.550725

根据上述结果，绘制节假日和非节假日的平均日订单需求量的柱状图，如图 5.8 所示。

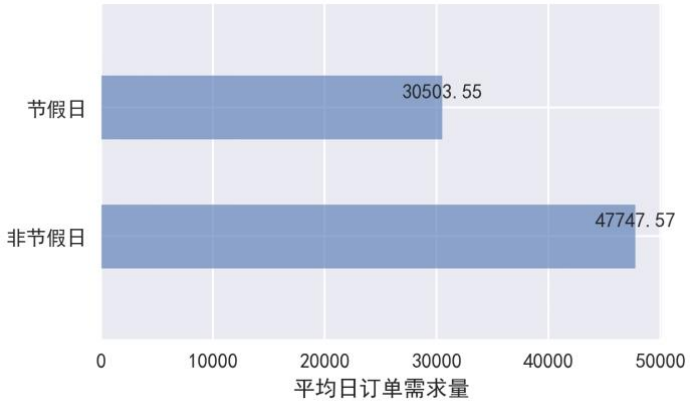


图 5.8 节假日与非节假日平均日订单需求量

通过上图可明显看出，节假日的平均日需求量相对非节假日的需求量明显降低，说明节假日对产品的需求量有显著影响。

为了进一步探究哪种节假日对产品需求量的影响较大，我们统计不同类型节假日的订单需求量情况，如表 5.7 所示。

表 5.7 各类型节假日的订单情况

节假日	订单数	天数	平均每订单需求量	平均日订单需求量
元旦	760	8	211.927632	20133.125000
春节	0	21	0	0
清明	1282	9	141.525741	20159.555556
劳动	1888	9	187.463453	39325.666667
端午	2135	9	116.219204	27569.777778
中秋	3253	9	89.150938	32223.111111
国庆	6271	25	138.762079	34807.080000

基于上述结果，绘制不同类型节假日平均日订单需求量的柱状图，如图 5.9 所示。

图 5.9 不同类型节假日平均日订单需求量

由上图可知，不同类型节假日对产品需求量的影响不同，但相对非节假日来说节假日的产品的需求量都有明显降低。元旦和清明假期的产品需求量下降较为明显，推测可能因为元旦是一年中的开头，清明节又需要扫墓，活动较多，工人停工，产品需求量减少。而劳动节与国庆节相对来说无固定活动要求，影响相对较小。

综上可以得出，节假日对产品的需求量有影响，一般来说，在节假日的产品需求量相比非节假日会降低。

5.5 促销活动对需求量的影响

一般来说，经销商向该大型企业的订货需求量并不代表经销商就能在当天立即收到所订购的产品，而是需要经历企业生产制造的时间，所以在一些大型促销活动如 618、双 11、双 12 等来临之际，经销商通常会提前一段时间向企业进行订货。众所周知，一般大型促销活动并不仅仅在当天进行促销，一般促销期会持续一周或两周时间，经销商需要在促销活动开始前就收到产品，方便在促销活动开始后进行销售。

基于以上考虑，我们选择观察在促销日前 1 个月以及促销日后两周这一段时间内订单需求量的变化。通过分别绘制 6.18、11.11、12.12 促销日在 2015 年、2016 年、2017 年、2018 年需求量的变化，如下图 5.10-图 5.12 所示。

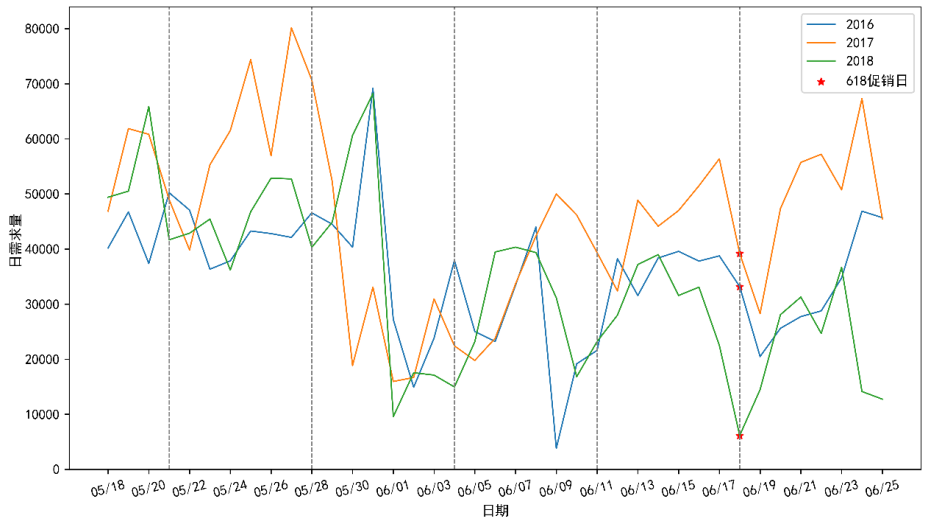


图 5.10 促销日 618 前后需求量变化

通常来说，618 促销活动会在 5 月 29 日开始预热，在 5 月 31 日 20 点开启预售，相当于提前两周开始促销。由图可以发现，在 618 前一个月即 5 月 18 日后两周，每年的产品需求量基本会高于 618 时候的订单需求量，2017 年在 5 月 26 日的订单需求量达到顶峰，2016 年和 2018 年的订到需求量则在 5 月 31 日达到顶峰，说明经销商会提前一个月开始向供货企业订购产品，陆续两周内都会保持较高的需求量，以备在 618 促销活动期间应对产品消费量的增加。

图 5.11 促销日 11.11 前后需求量变化

双 11 是全年中折扣力度最大的一次促销活动，每年双 11 的销售额不断增长，在 2018 年交易总额度已经超过了 2000 亿。双 11 会在 11 月 1 日正式开始，在 11 月 11 日迎来销售的高潮期。由上图可看出，在 10 月 11 日到 10 月 31 日时，产品的需求量会高于双 11 活动期间产品的需求量，明显在 2017 年 10 月 21 日和 23 时的产品需求量最高，且 10 月 31 日到 11 月 1 日时，每年产品的需求量都会显著降低，且下降幅度基本相同，说明每年经销商都会在双 11 促销活动前一段时间就开始增加订单需求量，在 11 月 1 日有充足的产品应对销售量的增加，订货量就开始减少，在后来开始保持稳定波动又低于前两周的订单需求，保证企业持续稳定的供货。

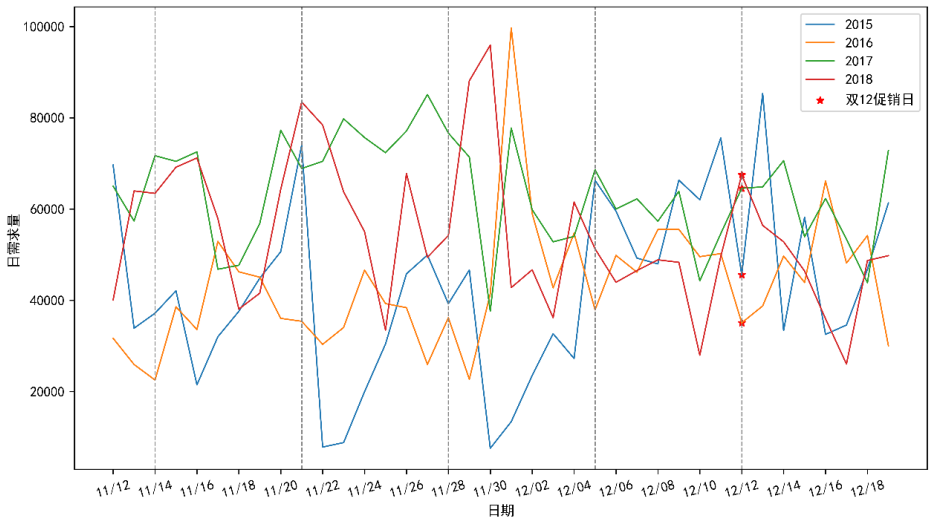


图 5.12 促销日 12.12 前后需求量变化

双 12 相对于前两个促销活动来说相对活动优惠力度比较小，但也算是一年中比较大型促销活动了。双 12 一般从 12 月 5 日开始预热，到 12 月 14 日活动结束。观察上图，2015 年产品需求量在双 12 促销活动前后的订单需求量差别不大，在 12 月 4 日到 12 月 14 日的活动期间反而保持了较高的产品需求量，而在 12 月 5 日前的需求量波动比较大，需求量的最高峰在 11 月 21 日；在 2016 年，需求量的最高峰在 12 月 1 日，但先前的订单需求量与双 12 促销期间的订单需求了差别不大，推测可能由于不久前双 11 的促销力度大，使得双 12 消费开始疲软。而在 2017 年和 2018 年，双 12 前一个月的订单量与双 12 活动有明显对比，说明在 2017 年和 2018 年经销商会开始批量订购产品以确保双 12 促销活动时产品的供应，推测可能是因为双 12 促销活动开始逐年被重视起来。

综上可知，促销活动能够促进消费，使产品的订单需求量增加，经销商们一般会提前开始订货，导致促销日前夕期间的订单需求量高于普通时期。

5.6 季节因素对需求量的影响

为了探究季节因素对需求量的影响，通过绘制产品需求在各季节的分布情况，初步了解各季节需求量的变化，如图 5.13 所示。

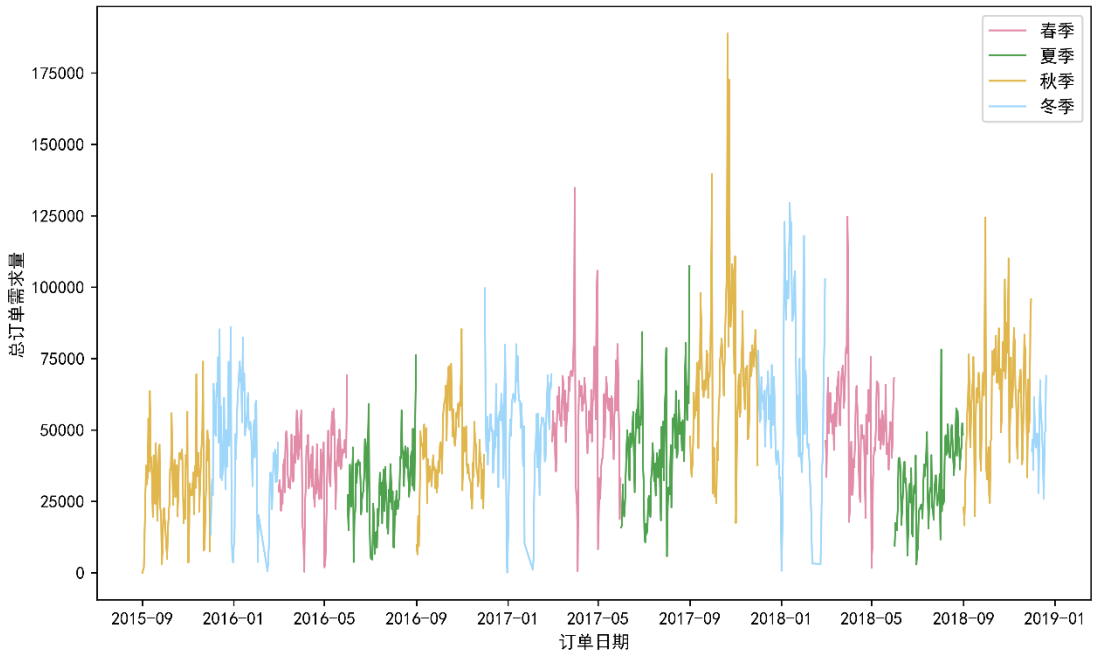


图 5.13 各季节总需求的整体情况

由上图可发现，同一年中秋季的需求量相对其他季节来说基本偏高，在夏季时期产品的需求量较低。且产品需求量在冬季的波动幅度较大，这可能是冬季期间的春节假期，企业无订单需求导致的波动。通过观察 2016-2018 年订单需求量的变化，订单需求量会整体呈现季节性的周期波动，且波动幅度越来越明显。在从春季过渡到夏季时，订单需求量整体降低；从夏季到秋季时，订单需求量开始增加；从秋季到冬季时，又开始缓慢减少直到第二年的春天。

由于 2015 年只有秋季和冬季，2018 年的冬季不完整，因此只考虑分析不同季节的总订单需求量和订单数。利用 Python 计算不同季节的平均订单需求量和平均季度总订单需求量，如表 5.8 所示。

表 5.8 不同季节的订单情况

季节	订单数	平均订单需求量	平均季度总订单需求量
春季	100792	131.896480	4431370.00
夏季	86611	109.878099	3172217.33
秋季	128604	144.213244	4636600.00
冬季	88826	151.004773	4162701.72

将计算结果可视化，绘制各季节的平均季度总订单需求量和各年份不同季节的季度总订单需求量的分组的柱状图，如图 5.14-图 5.15 所示。

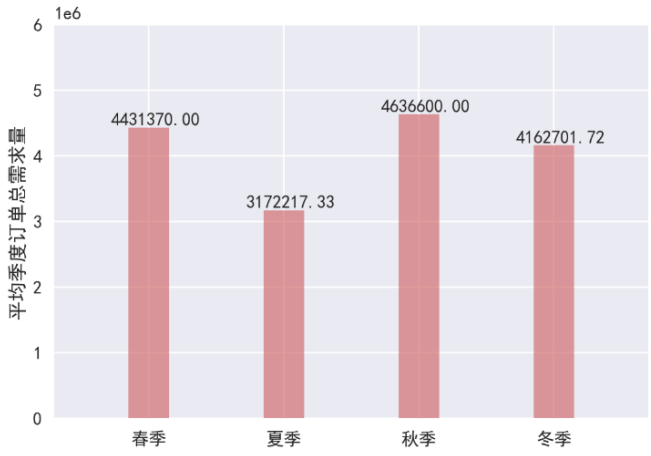


图 5.14 平均季度总订单需求量

图 5.15 各年份不同季节的季度总订单需求量

由图可以看出，不同季节平均季度的需求量不同，且各年份不同季节的产品需求量分布也不相同。总的来说，秋季的需求量较高，春季次之，而夏季的需求量较低。推测这可能与季节温度有关，夏季节气温较高，冬季气温较低，可能会影响劳动和运输效率，从而导致经销商向企业订货的需求量较少，且冬季受春节长假影响，需求量相对较低。而春秋季气温适宜，比较适合各种活动，产品的需求量相对会比较高。

为了进一步探究季节因素对需求量的影响，我们引入了季节指数。季节指数刻画了序列在一个年度内各月份或季度的典型季节特征。由于历史数据 order_train1.csv 的数据时间从 2015 年 9 月 1 日到 2018 年 12 月 20 日，月份数据分布不均，因此采用每日平均需求量进行计算。

$$某月季节指数 = \frac{某月份的产品每日平均需求量}{全年产品每日平均需求量}$$

通过计算产品各月份需求量的季节指数并绘制折线图，季节指数图如图 5.16 所示。

图 5.16 产品需求量的季节指数

由图 5.10 可知 2 月份和 7 月份的季节指数最低，该季节指数在 2-3 月、7-10 月时递增，3-7 月、1-2 月时逐渐递减，可得到该公司产品需求量在春秋季较高，夏季较低。结合公司实际经营情况，每年 2 月份的春节是法定节假日，因此产品的需求量在 2 月份骤减，随着春节后生产经营活动的正常开展，3 月需求量较 2 月需求量有了大幅的上升。

5.7 不同销售方式的需求量的特性

产品的销售方式有线下(offline)和线上(online)两种，首先绘制不同销售方式下产品需求量的二维点图，如图 5.17 所示。

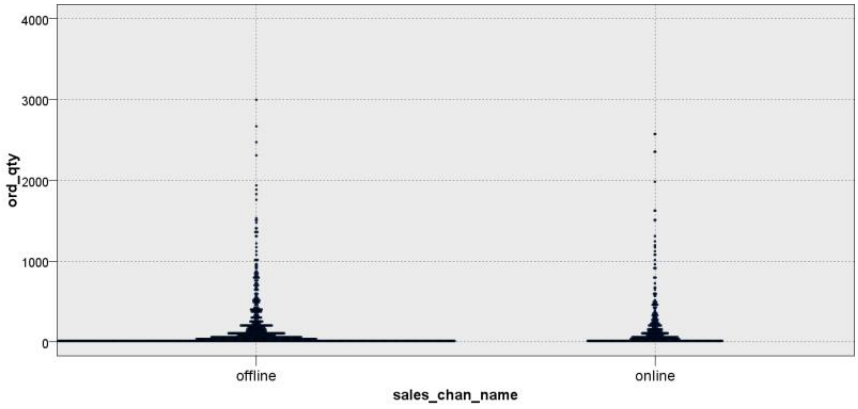


图 5.17 不同销售方式产品需求量

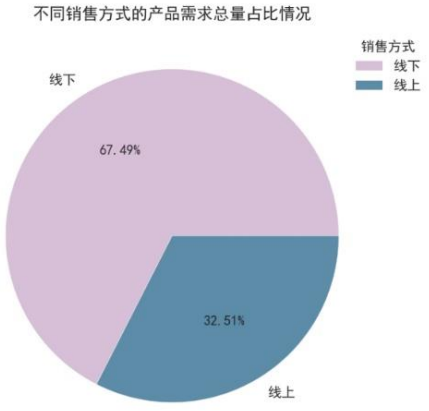
从图中可看出，许多点构成的直线说明在该订单量下的该种销售方式的数据点较多，即订单数较多，说明需求量较低的订单数较多。且需求量较高的点大多集中在线下销售方式，线下销售方式比线上的总体订单数多，说明该大型企业的主要销售目标还是线下实体经销商，淘宝、京东等电商平台的需求量相对较少。

为了使结果更加清晰，通过利用 Python 的 groupby 方法分组统计不同销售方式的订单数、总订单需求量、平均订单需求量，如表 5.9 所示。

表 5.9 不同销售方式的订单情况

	订单数	总订单需求量	平均订单需求量
线下	290594	36966407	127.209808
线上	114239	17803905	155.847872

根据上述结果，绘制不同销售方式的总订单需求量占比饼图，如图所示，并绘制平均订单需求量的柱状图，如图 5.18-5.19 所示。



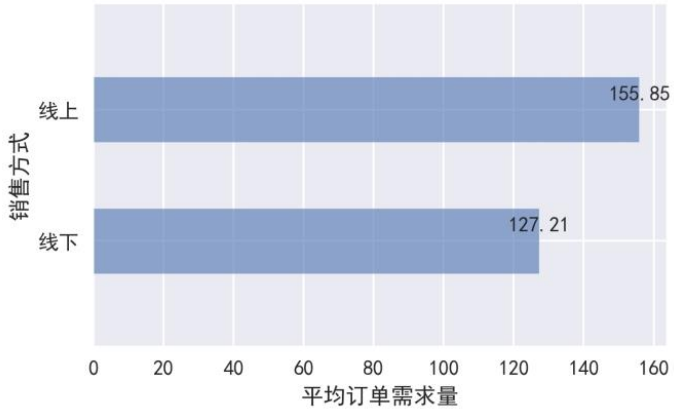


图 5.18 不同销售方式的总订单需求量占比　　图 5.19 不同销售方式平均每个订单需求量

由上图可知，线下销售渠道的总需求量较高，而线上平均每个订单的需求量最高的。

一般来说，线下采购可以直接接触到产品，方便检验产品的质量，增强买卖双方交易的信心，因此会选择在线下进行交易。且线下经销商对产品需求量较多，一般进行批量交易的单个产品价格会比线上电商平台的价格低，这也可能导致线下需求量较多的原因。

当然，线上电商平台的销售比较方便，虽然相对线下来说无法明确得知产品的质量，却也是消费者在没有空闲时间的情况下最优选择，但由于淘宝和京东电商平台上产品是通过快递进行运输，可能会面临暴力运输或中途丢件的风险，虽然较为方便，能够减少时间成本，但其相对线下方式来说不适合大批量采购，所以线下的订单需求量会比线上的多。

为了进一步探究不同销售方式下产品需求量的特性，通过绘制两种销售方式月需求量随时间变化，如图 5.20 所示。

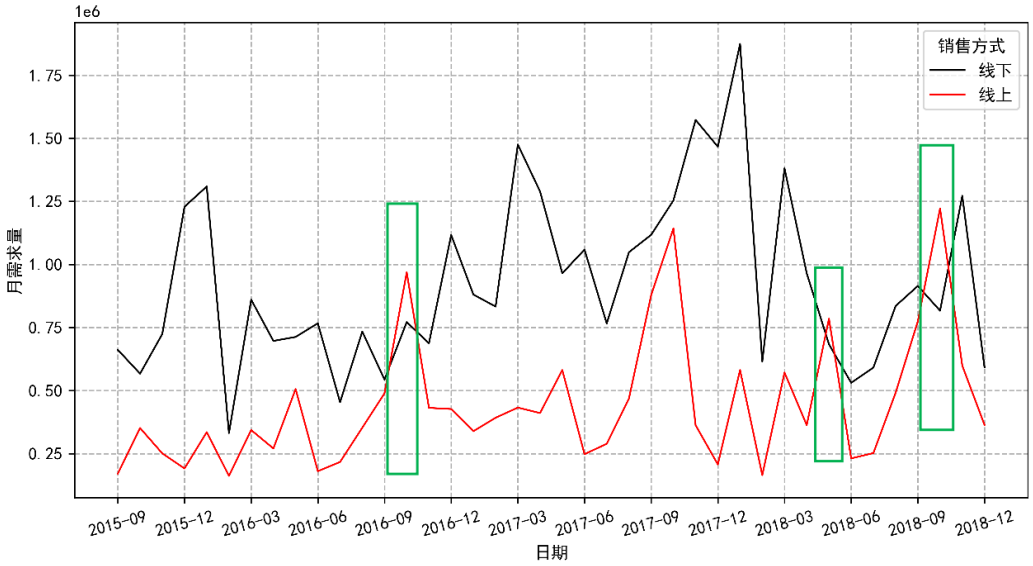


图 5.20 不同销售方式下月需求量变化

由该图可知，在一般情况下，线下实体经销商的需求量都会高于线上电商平台的需求量，只有在三段时间内线上的需求量大于线下，分别在 2016 年 10 月、2018 年 5 月、2018 年的 10 月期间。

基于上述促销日对产品需求量的影响分析，推测可能是因为促销日 618 和双 11，电商平台上的经销商会提前一个月进行大批量订货。因为一些大型促销节日一般是在电商平台如淘宝、京东上进行的，通过商品打折、跨店盲满减等促销方式刺激线上消费，而双 11 一般是一年中折扣力度最大的促销活动，所以对应的线上销售的订单需求量在这一时期会高于线下销售渠道，因为在大型促销活动线上的折扣力度会比实体经营的折扣力度大。

因此，不同的销售方式下的订单需求量有其各自的特点，一般来说线下的订单需求量会高于线上，但也不排除像 618、双 11 等大型促销活动的刺激下线上电商平台的消费，导致线上的订单需求量会高于线下。

5.8 不同时间段产品需求量的特性

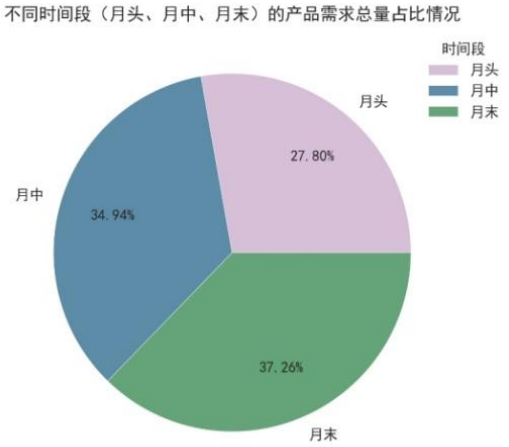
利用 Python 的 groupby 方法分组统计不同时间段（月头、月中、月末）的订单数、总订单需求量、平均订单需求量，如表 5.10 所示。

表 5.10 不同时间段的订单情况

时间段	订单数	总订单需求量	平均订单需求量
月头	111316	15226626	136.787398

月中	142433	19137303	134.360036
月末	151084	20406383	135.066473

根据上述结果，绘制不同时间段（月头、月中、月末）的总订单需求量占比饼图，如图所示，并绘制平均订单需求量的柱状图，如图 5.21-5.22 所示。



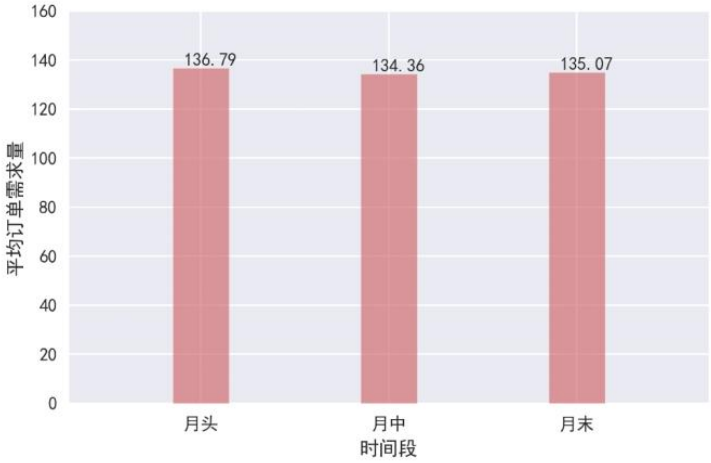


图 5.21 不同时间段总订单需求量占比　　图 5.22 不同时间段平均每个订单需求量

由上图可知，不同时间段平均订单的需求量差别不大，不管任何时间段，经销商每一次订单的需求量都是大致相同的。但对总需求量来说，月头的需求量是最少的，而月末的需求量是最多的，这说明虽然每次订单的需求量大致相同，但是经销商往往会倾向于在月末的时候多订一些订单，使得产品的需求量增大，而在月末产品的囤积，也可能是导致月初时订单需求量减少的原因。

为了使需求量在不同时间段的变化更加明显，将所有相同日的需求量进行加总，绘制不同日的总订单需求量折线图，如图 5.23 所示：

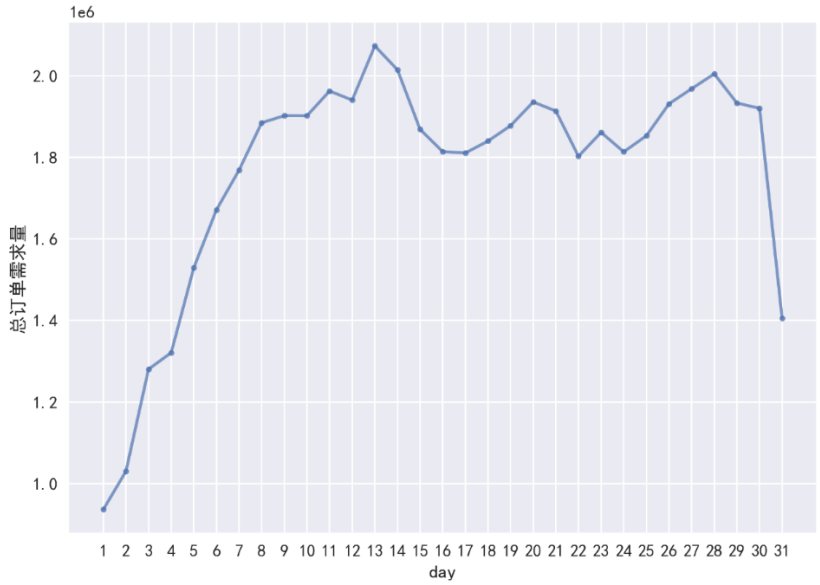


图 5.23 不同日的总订单需求量

从上图可明显看出，月初时总需求量呈现上升趋势，在月中时达到稳定后会随着时间波动，在月末时会出现小幅度增长。因为部分月份没有 31 号，所以 31 号的总订单需求量显著减少。

因此，不同时间段产品的需求量不同，在月初时需求量较少，在月末时需求量较多，月中的需求量处于两者之间。

为了进一步探究每个月不同时间段产品需求量的特性，通过绘制三个时间段对应的总需求量随时间的变化折线图，如图 5.24 所示。

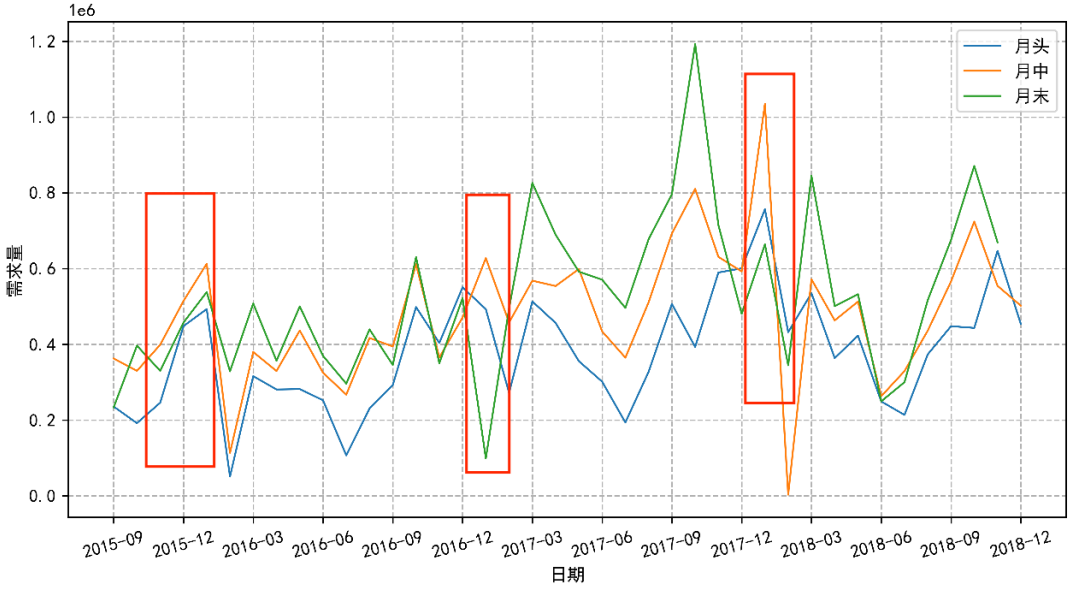


图 5.24 三个时间段月需求量的变化

可以观察到，一般每个月的订单需求量在月末最高，在月头最低。红框中部分是比较特殊的情况，即月中的需求量会高于月末的需求量，可以发现这三个特殊情况主要出现在一年中的开头，2015 年末也出现月中的产品需求量大于月末，推测可能是年末产品一般不会大量囤积到下一年，所以在月末会减少对产品的需求。而在年开头，由于产品存货不足，经销商会提前进行批量订购，不会再等到月末。且春节假期一般在 1 月份末或 2 月份，企业没有订单供货，因此月中的需求量也会偏高。

综上可知，除了年末与年初的特殊情况，一般来说产品的需求量在月末的最多的，在月初的最少的。

5.9 产品需求量随时间变化规律

基于上述对产品需求量影响因素分析可知，产品的需求量受多方面因素影响，因此产品需求量的有变化是合理的。从经销商与制造企业的长期交易方面说，经销商会预估未来一段时间的产品需求并通过签订协议来向制造企业确定未来相应时间产品的价格，虽然产品的需求量在不同时间段、不同季节的需求量会有所波动，且波动大小不等，但这种波动会呈现出一定的周期。根据 5.6 季节因素对需求量的影响分析可知，产品需求量会呈现季节性波动，因此我们将探究其随时间变化可能存在的其他规律。

因此，以 2016 年、2017 年产品需求量变化为例，通过绘制了 2016 年和 2017 年每年的产品日需求量的折线图，如图 5.25、图 5.26 所示，分析产品需求量随时间变化规律。

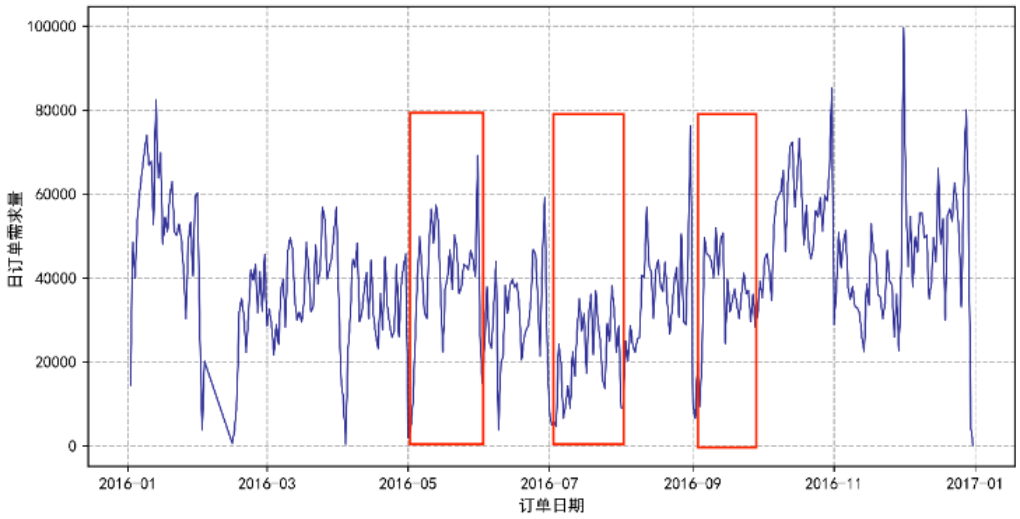


图 5.25 2016 年日产品需求量变化

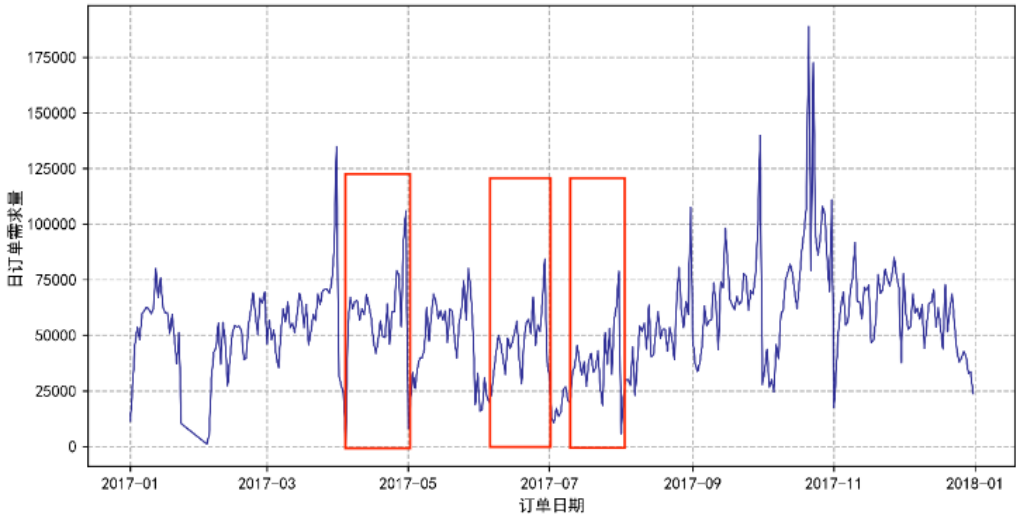


图 5.26 2017 年日产品需求量变化

由上图可观察，产品的需求量大概在每一个月就会出现一次与附近时间区域内最低的波谷，且波谷基本出现在月与月的交界处。由上述对不同时间段产品需求量的特性分析可知，产品需求量一般在月初最低，在月末是最高，因此产品的需求量呈现月周期波动是合理的。且通过对 2016 年与 2017 年日产品需求量的大致走向，也可发现在整体上春秋两季的产品需求量会高于冬夏两季，呈现了相同的季节波动方向，说明产品需求量也存在季节周期性波动。

综上可知，产品需求量会存在月周期以及季节周期性的波动变化。

5.10 不同品类间产品需求量的对比

利用 Python 的 groupby 方法分组统计不同品类的订单数、总订单需求量、平均订单需求量，如表 5.11 所示。

表 5.11 不同品类间的订单情况

产品大类编码	产品细类编码	订单数	总订单需求量	平均订单需求量
301	405	14505	1685723	109.391451
302	408	67358	6221334	92.362214

303	401	38510	3605644	93.628772
	406	3096	47022	15.187984
	410	4122	81654	19.809316
	411	984	13705	13.927846
304	409	8477	618444	72.955527
305	412	57982	6324256	109.072747
306	402	7338	2530770	344.885527
	407	132367	22017667	166.338037
307	403	34268	5728696	167.173340
308	404	35826	5994397	167.319740

根据上述结果，绘制不同大类产品的总订单需求量占比饼图，如下图所示，并绘制不同细类产品平均订单需求量的柱状图，如下图 5.27、图 5.28 所示。

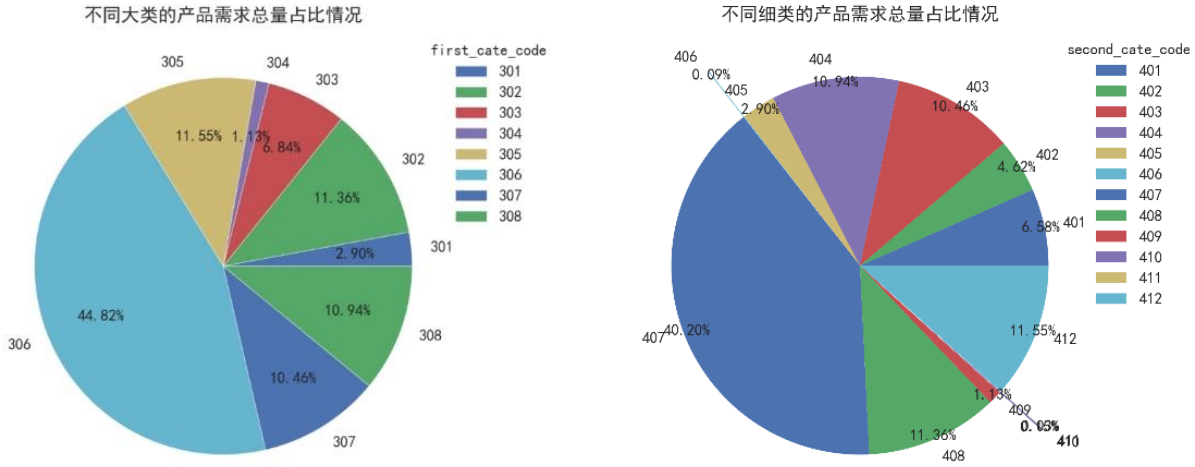


图 5.27 不同大类与细类产品总订单需求占比

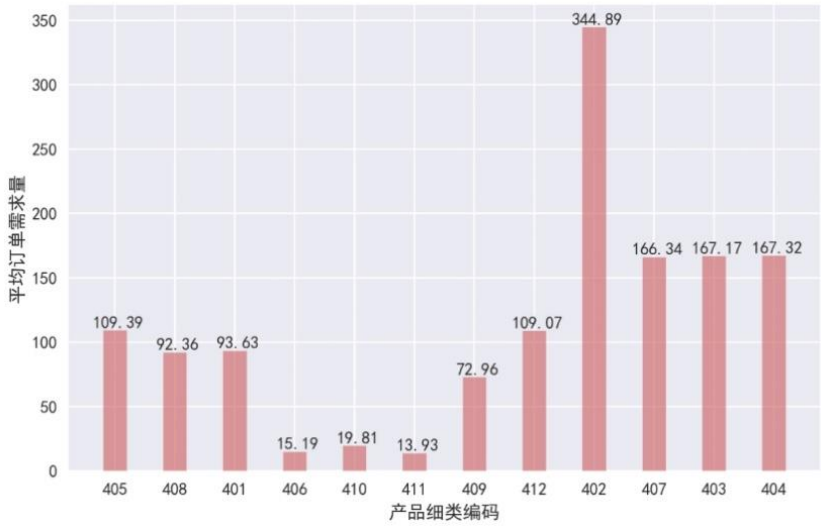


图 5.28 不同细类产品的平均订单需求量

由图可看出，不同产品大类的需求量中，产品大类 306 的需求量是最大的，产品大类 304 的需求量是最小的。

在不同产品细类的需求量中，产品 406、产品 410 和产品 411 的需求量是最小的，这三者都属于产品大类 303。在这一类别中，产品 401 的不管是订单数、需求总量还是平均每个订单的需求量都超过这三类产品的总和，说明在 303 大类产品中，经销商们更倾向于订购产品 401 作为前三者的替代品，而产品 406、产品 410 和产品 411 在同品类竞争中优势相对不明显，因此需求量较低。

而产品 407 的需求量是最大的，刚好属于需求量最大的产品大类 306，且其需求量远远超过同大类中的产品 402，说明产品 407 可能是必需品且在同类竞争中取得了较大的优势，使得经销商大多订购产品 407，需求量大大增加。

为了进一步探究不同品类间需求量的相同点和不同点，我们基于以上对需求量的影响因素和特征，分析不同品类需求量特征的异同。

5.10.1 不同品类在不同价格下需求量的对比

通过绘制不同产品大类在不同价格下的需求量散点图，如图 5.29 所示。

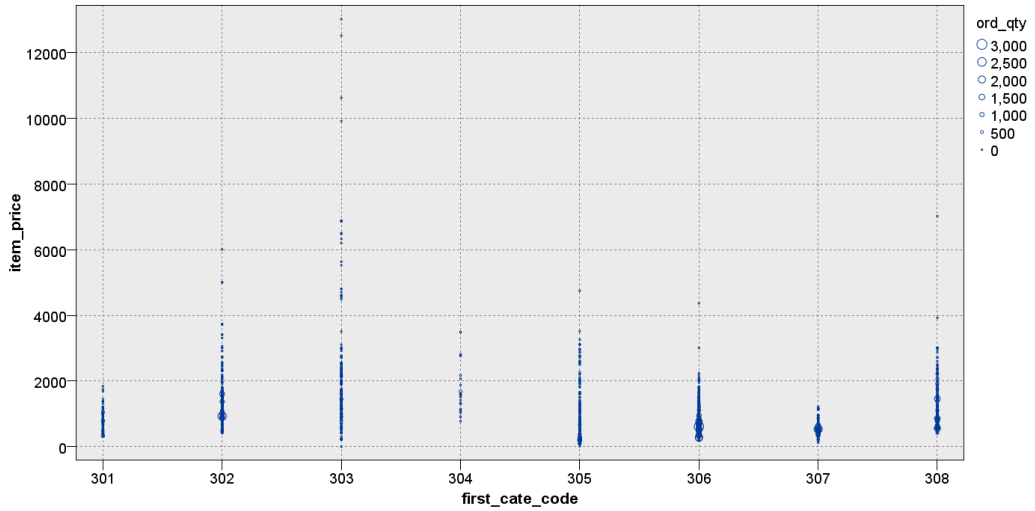


图 5.29 不同品类在不同价格下的需求量

由上图中可知，不同品类的订单需求量都集中在价格较低的区域，说明不同产品的需求量都遵循在一定程度上，价格越低，需求量越高。但是，不同产品集中的价格区域不同，这取决于产品本身的价值，且有些品类间产品的价格跨度较大，如产品大类 303，其中有产品的价格处于 12000 以上；而有些品类的产品价格较为集中，如产品大类 308。因此，不同品类间受价格的影响变化是一致的，但产品价格的波动区间不一致，不同品类有不同的均衡价格。

5.10.2 不同品类在不同销售区域下需求量的对比

基于上述对产品销售区域对需求量的影响分析，将不同品类在不同销售区域下的总需求量可视化，绘制柱状图，如图 5.30 所示。

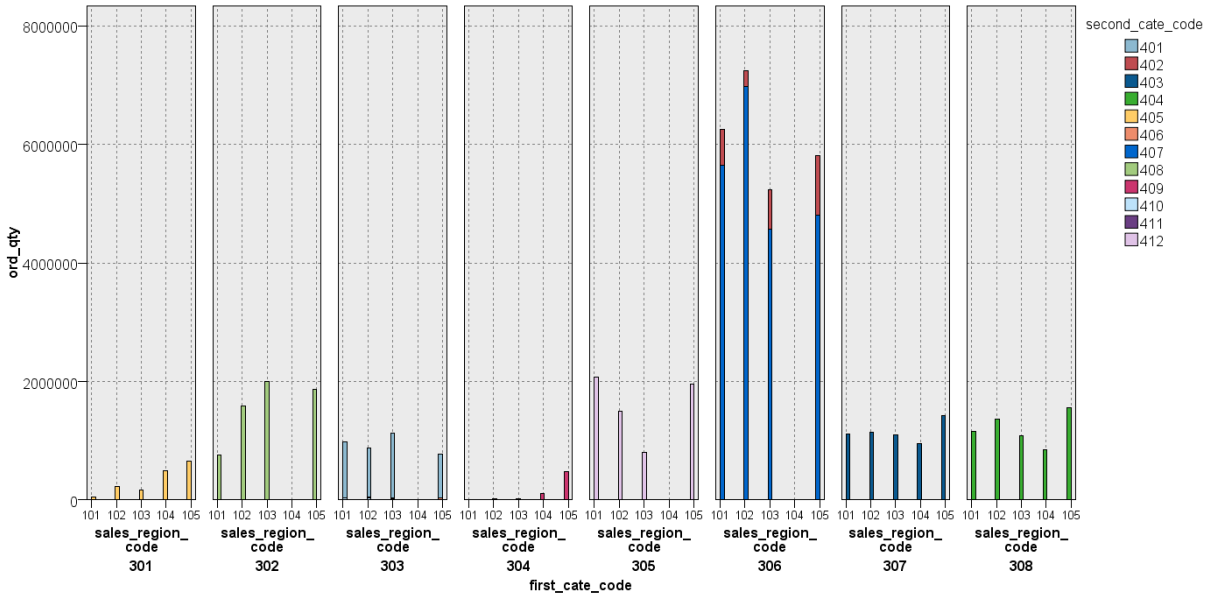


图 5.30 不同品类在不同销售区域下的需求量

由图可知，不同品类的需求量不同，其在不同销售区域的需求量也不相同。可以观察到：

①仅有品类 301、307 和 308 在各个销售区域都有订单需求；

②相对其他销售区域，品类 306 在 102 销售区域的需求量是最高的；

③在 104 销售区域内，只有品类 301、304、307 和 308 有订单需求，其他品类的需求量几乎为 0，这可能是因为其他品类在 104 销售区域内并不是必需的，或者有其他与 104 区域更近的供货商能够承担其他品类的需求量。

④品类 304 在各个销售区域的需求量都是最低的；

为了使结果更加清晰，我们基于不同品类在不同销售区域下的订单数和平均订单需求量进行分析。八个产品大类在不同销售区域下的订单数对比，如图 5.31 所示，蓝色直线越长表明订单数越多。

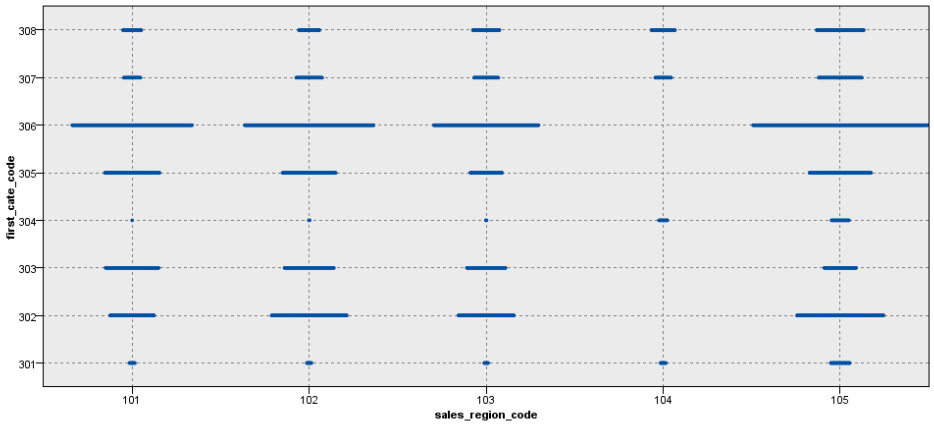


图 5.31 不同品类在不同销售区域下的订单数

从上图可知，不同类别产品的订单数量在不同的销售区域内分布不同。相对其他销售区域，各个品类在销售区域 105 的订单数都是最多的，且各个品类在销售区域 104 的订单数是最少的。

综上可知，不同品类在不同销售区域下的需求量不相同，每个销售区域都有比较倾向的品类，且有的品类并不是当前销售区域必需的，因此对应的订单需求量较少。

5.10.3 不同品类在节假日需求量的对比

基于上述节假日对需求量的影响分析，将不同品类在节假日与非节假日下平均日订单需求量可视化，绘制柱状图，如图 5.32 所示。

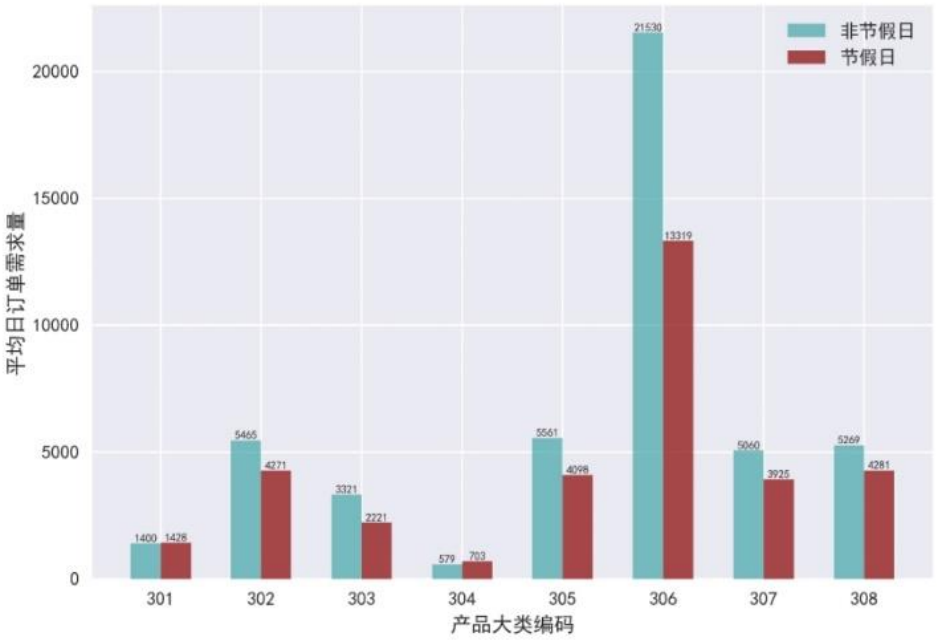


图 5.32 不同品类在节假日和非节假日下的需求量

由上图可知，大部分品类在非节假日时的平均日订单需求量都高于节假日，这符合 5.3 节假日对需求量影响的一般规律。但也有部分产品例外，如品类 301 与品类 304，节假日似乎对这两类产品的需求量没有影响。因此，不同品类之间的产品需求量受节假日影响的情况不同，有些品类的产品需求量会受节假日影响，需求量降低，如品类 306 等，而有些品类的产品需求量不受节假日影响，其在节假日时的产品需求量与非节假日的需求量差异不明显，甚至在节假日的平均日需求量会高于非节假日。

综上，不同品类之间的产品需求量受节假日的影响程度不同，某些品类的产品需求量甚至不受节假日影响。

5.10.4 不同品类在不同季节下需求量的对比

基于上述季节因素对需求量的影响分析，将不同品类在不同季节下的总需求量可视化，绘制柱状图，如图 5.33 所示。

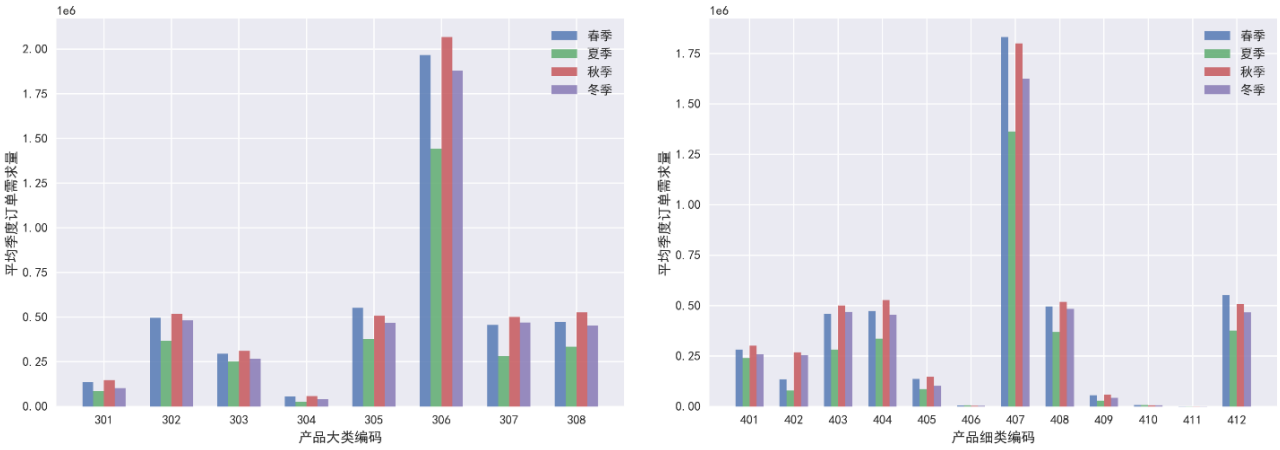


图 5.33 不同品类在不同季节下的需求量

由上图可看出，不同品类虽然订单需求量不同，但是其需求量基本都在秋季最高，在夏季最低，与上述季节因素对产品需求量影响的分析结果是一致的。不同品类之间产品的需求量一般都在夏季达到最低，推测可能因为气温较高，影响人们外出购物的欲望，使得线下实体经销商不得不降低订货需求量；而在春秋季时气温适宜，所以订单需求量相对夏季和冬季较高。

综上，不同品类之间的产品需求量受季节因素的影响结果是相同的。

5.10.5 不同品类在不同销售方式下需求量的对比

基于上述不同销售方式产品需求量的特性分析，我们将探究不同品类的产品需求量是否满足这一特性。将不同品类在不同销售方式下的总需求量可视化，绘制柱状图，如图 5.34 所示。图中 0 表示“线下”，1 表示“线上”。

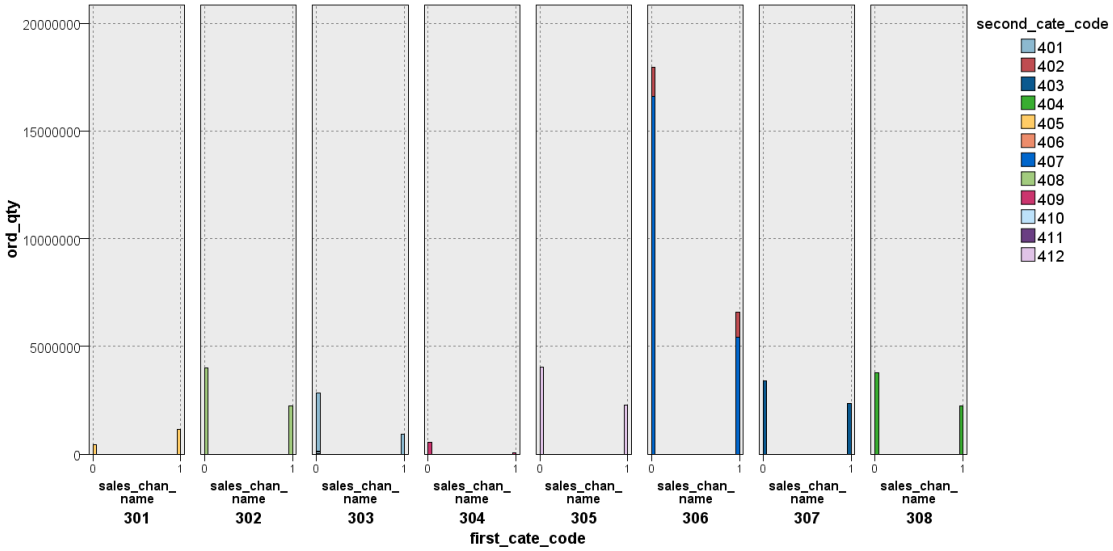


图 5.34 不同品类在不同销售方式下的需求量

由上图可以看出，除了产品大类 301，其他品类在线下实体经销商的产品需求量高于线上淘宝、京东等电商平台的需求量，满足在 5.6 不同销售方式产品需求量的特性分析。

只有品类 301 是线上电商平台的订单需求量大于线下实体经销商的需求量，推测出现这种情况可能是因为品类 301 的需求量较小。

绘制不同品类在不同销售区域的平均订单需求量的热力图，如图 5.35 所示。由图可知，品类 301、302、303 和 305 线上平均订单需求量都高于线下，而品类 301 的线上总需求量又高于线下需求量，说明品类 301 的需求量主要是由线上电商平台承担的。

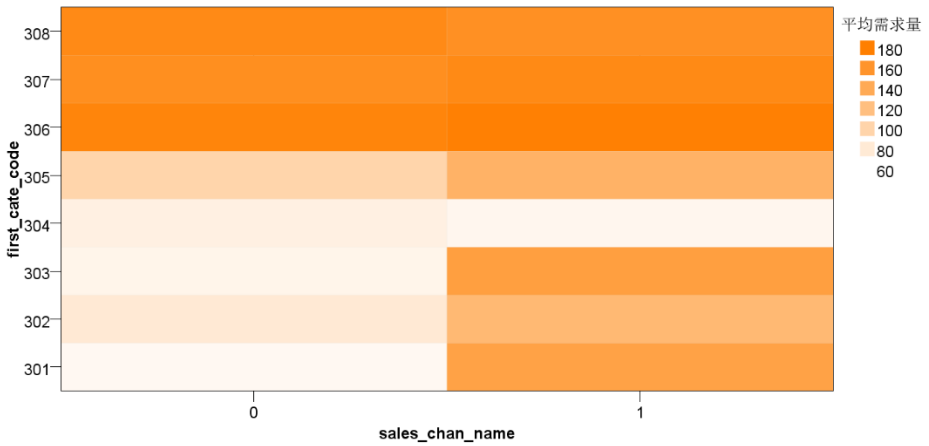


图 5.35 不同品类在不同销售区域的平均订单需求量

综上可知，大部分品类的需求量都符合不同销售方式的需求量特性，但也有个别产品的在线上的总需求量高于线下，因此，不同品类在不同销售方式下的产品需求量特性也有一定的差异。

5.10.6 不同品类在不同时间段下需求量的对比

基于上述不同时间段产品需求量的特性分析，我们将探究不同品类的产品需求量是否满足这一特性。将不同品类在不同时间段下的总需求量可视化，绘制柱状图，如图 5.36 所示。图中 1 表示“月初”，2 表示“月中”，3 表示“月末”。

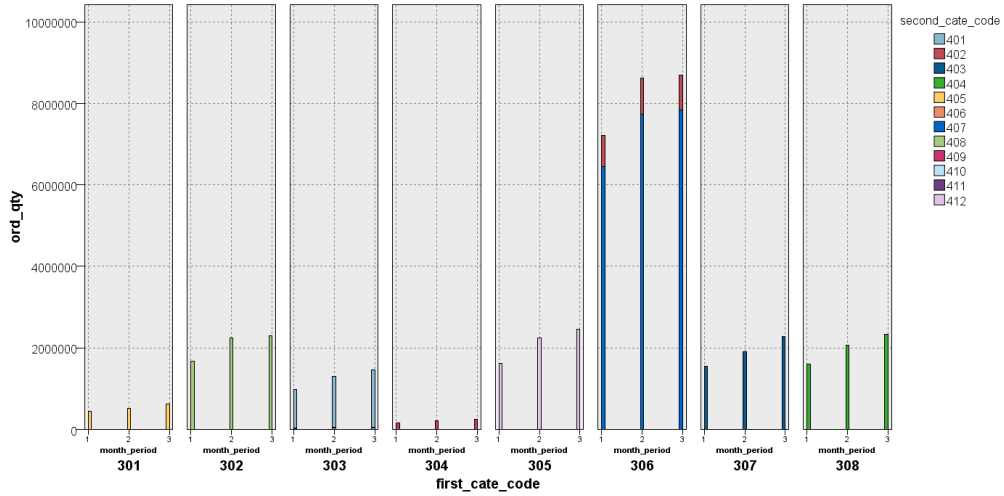


图 5.36 不同品类在不同时间段下的需求量

由上图可知，不同品类的产品需求量在月末时达到最高，在月初时的产品需求量都比其他两个时间段的需求量低，这符合 5.7 不同时间段产品需求量的特性。因此，不同品类间产品的需求量都满足不同时间段产品需求量的特性，即产品在月初时需求量较低，在月末时的需求量较高，呈现出产品需求量随时间增加而增加的周期性变化。

综上，不同品类的产品需求量都满足不同时间段产品需求量的特性。

5.10.7 相同点和不同点总结

基于上述分析，我们从中得到了不同品类之间产品需求量的相同点和不同点。

(1) 相同点：

① 不同品类间产品需求量受产品价格的影响相同，每个品类需求量最多所对应的价格都较低，且产品价格与需求量相互作用，维持两者的稳定波动；

② 不同品类间产品需求量受季节因素的影响相同，每个品类的产品需求量都在春秋两季较高，在冬夏两季较低；

③ 不同品类间在不同时间段下产品需求量的特性是相同的，每个品类的产品需求量都在月初时较低，在月末时较高。

(2) 不同点：

① 不同产品大类间的订单需求量差异较大，且相同大类下不同细类产品的订单需求量也有比较大的差异；

② 不同品类间产品需求量受不同销售区域的影响不同，某些产品甚至在一些销售区域的需求量为 0；

③ 不同品类间产品需求量受节假日的影响不同，某些品类的需求量并不受节假日影响；

④ 不同品类间在不同销售方式下产品需求量的特性不同，大部分品类线下实体经销商的产品需求量会大于线上电商平台如淘宝、京东的需求量，但也有个别产品的在线上的总需求量会高于线下。

6 问题 1.2：销售额与产品价格、需求量、产品品类的关系

为了对出货数据进行深入研究，且由问题 1.1 逐步回归法可知产品需求量与价格的相关性的最强的，但不同产品的需求量较为集中的产品价格区间却不同，对于有的产品来说，价格高但是需求量也高，而有些产品价格高需求量反而会显著降低，这与产品的价值有关。一般来说，产品价格高，说明产品的价值也高。因此，可以根据价格的高低将产品大致分为高端品、中端品和低端品。

销售额是连接产品需求量和价格关系的桥梁，销售额=产品需求量×价格，因此，我们考虑对产品的销售额进行分析。销售额不仅可以估算出企业在当期营业运作中所实现的实际经济效益多少,而且可以衡量企业经营水平的高低。从公司的角度上看，销量的稳步提升，才能帮助企业占领更多的市场份额，攫取更多的营业收入，维持企业的可持续发展。

为了研究销售额与产品价格、需求量、产品品类的关系，使用 SPSS Modeler 对产品价格、每笔订单的需求量、每笔订单的销售额、细类编号这 4 个特征对产品进行 K-Means 聚类，通过分析聚类结果的特征，

6.1 K-Means 聚类原理

K-Means 算法如下：

对于给定的一个包含 n 个12维数据点的数据集 $X = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$，其中 $\mathrm{x}_i \in R^{\mathrm{d}}$，以及要生成的数据子集的数目K, K-Means 聚类算法将数据对象组织为 k 个划分$C = \{c_k, i = 1,2, \cdots, K\}$。每个划分代表一个类$c_k$，每个类$c_k$有一个类别中心$\mu_{\mathrm{i}}$。选取欧氏距离作为相似性和距离判断准则，计算该类内各点到聚类中心 μ_{i} 的距离平方和

$$J(\mathrm{c}_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

聚类目标是使各类总的距离平方和 $J(\mathrm{C}) = \sum_{k=1}^{K} J(c_k)$ 最小。

$$J(\mathrm{C}) = \sum_{k=1}^{K} J(c_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ki} \|x_i - \mu_k\|^2$$

其中，$\mathrm{d}_{ki} = \begin{cases} 1, \text{若 } x_i \in c_i \\ 0, \text{若 } x_i \notin c_i \end{cases}$

显然，根据最小二乘法和拉格朗日原理，聚类中心 μ_k 应该取为类别 c_k 类各数据点的平均值。K-means 聚类算法首先随机从数据集中选取 K 个点作为初始聚类中心，然后计算各个样本到聚类中的距离，把样本归到离它最近的那个聚类中心所在的类。计算新形成的每一个聚类的数据对象的平均值来得到新的聚类中心，如果相邻两次的聚类中心没有任何变化，说明样本调整结束，聚类准则函数已经收敛。

6.2 聚类结果与分析

聚类结果如图 6.1 所示。聚类的最优个数为 5；轮廓系数为 0.784，接近 1，聚类效果良好。从聚类中心及各类的总销售额可以看出中低端产品占销售额的主要部分，高端产品价格虽高但需求量较少，因此所占销售份额也较少。从产品细类来看，407、404 所占销售份额较大，410 所占销售额份较少。

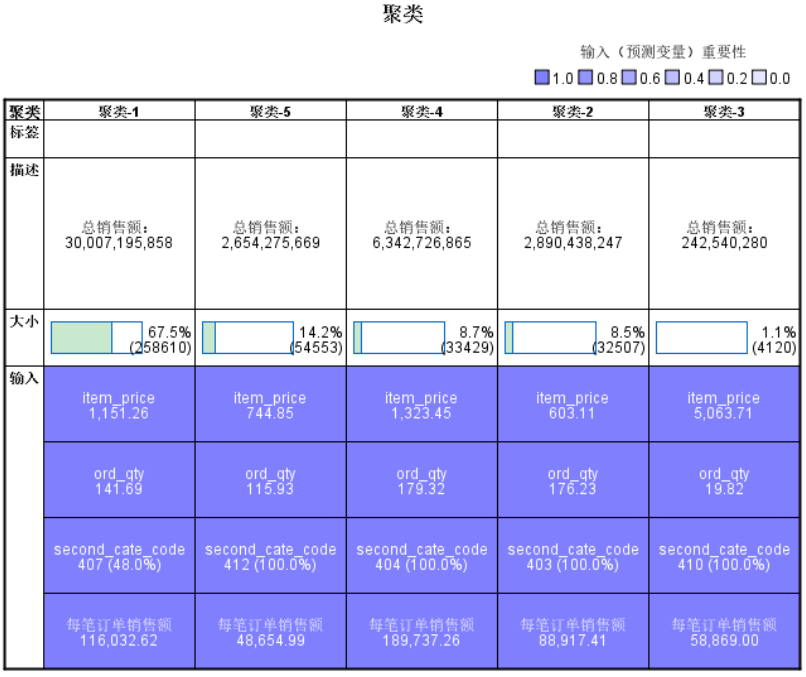


图 6.1 K-Means 聚类结果

7 问题 2：产品需求预测

7.1 数据处理

在预测附件产品未来三月的需求量前，我们要对数据进行进一步的处理。首先，我们观察到，附件产品的销售区域没有 104 编码的区域，因此我们在原有的训练数据集中剔除掉销售区域编码为 104 的所有数据。以减小训练集规模并结合附件产品信息精准预测需求量。

其次，考虑到题目所给的训练数据集中，每种产品在 2015 年 9 月 1 日至 2018 年 12 月 20 日期间的每日的数据并不全，原因在于并不是每种产品每天都有需求，因此我们将每个产品缺失日期进行补充，并将需求量数据填充为 0。

基于上述的分析，我们在预处理后的训练数据集的基础上，根据问题 1 的分析对特征进行筛选，并对训练数据集进行时间补全以及划分天、周、月粒度的训练集数据。详细的数据处理流程如下。

7.1.1 特征筛选

由问题 1 的逐步回归法可知，产品需求量与是否为促销日没有关系，因为经销商一般会在促销活动开始前一段时间就订购产品，对促销日需求量的影响不大，因此我们将特征 sales_promotion_type 剔除，最终选择了 order_date、sales_region_code、item_code、first_cate_code、second_cate_code、year、month、day、weekday、season、month_period、holiday_type、item_price 共 13 维的特征。

7.1.2 天粒度的训练数据集

首先，按天的时间粒度对训练数据进行聚合。由于题目要求是要分别预测 101、102、103、105 这四个销售区域未来三个月的需求量，由于在 4 个销售区域下同个产品在一天内可能有多个订单，且价格不同，所以我们取同一销售区域下同种产品在同一天的平均价格作为这一产品在这一天在该销售区域的唯一价格，并将对应的需求量求和，而销售方式取该日在该地区内的产品订单中销售方式最多的类型，最终得到了每个产品在同一天同一销售区域内的唯一数据。

对于含有缺失的时间的数据，我们为每种销售区域下每种产品的时间补全，时间范围为 2015 年 9 月 1 日至 2018 年 12 月 20 日（共 1207 天）。将缺失的日期的产品需求量填充为 0，而每种销售区域的每种产品的销售区域、产品编码、大类编码和细类编码都是相同的，从而对缺失的数据进行对应补充。其次，再次为填充好时间和需求量的数据进行特征提取，提取了特征 year、month、day、weekday、season、month_period、holiday_type、sales_promotion_type，对于产品价格，考虑将每个产品缺失日期的产品价格填充上对应产品历史数据的价格平均值，由于产品的销售方式无法确定，故而删除了 sales_chan_name 这一数据特征。

最终处理好的数据即是以天为粒度的训练集数据。统计得到剔除 104 销售区域数据后的 4 个销售区域下所有产品的组合共 5109 种产品，补全所有时间的数据后得到的新数据集规模为，6166563 行×14 列。6166563 行即为 5109 种产品×1207 天。14 列即为 order_date、sales_region_code、item_code、first_cate_code、second_cate_code、year、month、day、weekday、season、month_period、holiday_type、item_price 共 13 维的特征加上一列需求量 ord_qty，处理后的前五条数据如图 7.1 所示。

order_date	sales_region_code	item_code	first_cate_code	second_cate_code	year	month	day	weekday	season	month_period	holiday_type	item_price	ord_qty
2015-09-01	101	21984	306	407	2015	9	1	2	3	1	0	809.450000	0
2015-09-02	101	21984	306	407	2015	9	2	3	3	1	0	809.450000	0
2015-09-03	101	21984	306	407	2015	9	3	4	3	1	0	809.450000	0
2015-09-04	101	21984	306	407	2015	9	4	5	3	1	0	809.450000	0
2015-09-05	101	21984	306	407	2015	9	5	6	3	1	0	809.450000	0

图 7.1 天粒度数据集示例

7.1.3 周粒度的训练数据集

按周的时间粒度对训练数据进行聚合，将同一销售区域同一产品同一周的订单价格取平均数，并将对应的需求量求和，销售方式取该周在该销售区域内该种产品订单中销售方式最多的类型，得到同一销售区域同一产品在同一周的唯一数据，并且保留 sales_region_code、item_code、first_cate_code、second_cate_code、year、item_price 共 6 维的特征，新增了 week_of_year 这一特征。处理后的前五条数据如图 7.2 所示。

sales_region_code	item_code	first_cate_code	second_cate_code	year	week_of_year	item_price	ord_qty
101	20001	302	408	2015	36	703.500000	0
101	20001	302	408	2015	37	703.500000	0
101	20001	302	408	2015	38	703.500000	0
101	20001	302	408	2015	39	703.500000	0
101	20001	302	408	2015	40	703.500000	0

图 7.2 周粒度数据集示例

7.1.4 月粒度的训练数据集

按月的时间粒度对训练数据进行聚合，将同一地区同一产品同一年月的订单价格取平均数，需求量求和，销售方式取该月在该销售区域内该种产品订单中销售方式最多的类型。得到同一地区同一产品同一月的唯一数据。并且保留 sales_region_code、item_code、first_cate_code、second_cate_code、year、month、season、item_price 共 8 维的特征。处理后的前五条数据如图 7.3 所示。

sales_region_code	item_code	first_cate_code	second_cate_code	year	month	season	item_price	ord_qty
101	20001	302	408	2015	9	3	703.200000	0
101	20001	302	408	2015	10	3	703.200000	0
101	20001	302	408	2015	11	3	703.200000	0
101	20001	302	408	2015	12	4	703.200000	0
101	20001	302	408	2016	1	4	703.200000	0

图 7.3 月粒度数据集示例

基于上述数据处理，接下来我们将分别按天、周、月的时间粒度对未来三个月的产品需求量进行预测，并分析不同时间粒度对预测精度的影响。

7.2 模型建立、调参及评估

对附件产品未来三月进行需求量预测，该问题为时间序列预测问题。附件中的产品有 2619 种，若用传统的时间序列模型，需要为每个产品分别建立模型，即 2619 个模型。此外，附件

的产品中有部分是新产品，没有历史数据，因此我们考虑使用该新产品的同品类以及同销售区域的产品对该新产品进行分析。

常见的时间序列模型有 ARIMA、SARIMA、Prophet 等。但这些传统的时间序列模型只通过日期和需求量进行训练及预测，丢失了数据集中含有的销售方式、价格等特征信息。且需要构建 2619 个模型并分别调参，效率较低，并且很多产品在训练集中的 1207 天中需求量为 0 的占比很大，对模型的预测效果造成一定影响。此外，除了传统的时间序列模型，还可以将时间特征提取出来，并结合其他有用的特征，用机器学习模型进行预测，如随机森林、XGBoost 等。用机器学习的方法则可以结合时间特征以及销售方式、价格等特征进行预测，且可以将所有产品的数据放在一起训练以及预测，以产品编码和销售区域作为不同产品的标识特征，不需要建立两千多个模型，从而大大提高了预测效率。

本文尝试了 SARIMA、Prophet、随机森林、XGBoost 模型，并进行了比较和选择，最终决定选择 XGBoost 模型进行未来三月的需求量预测。我们一开始尝试的是 ARIMA 和 SARIMA 模型，但预测效果很差，因此便不详细介绍这两种方法。

本文以 2015 年 9 月-2018 年 9 月的历史数据作为训练集，以 2018 年 10 月-2018 年 12 月的历史数据作为测试集。利用 Python 分别建立 Prophet 模型、随机森林回归模型、XGBoost 模型，以天、周、月的不同粒度数据集进行训练和测试，并进行调参和预测效果的评估，进行不同模型之间的比较与选择，并分析不同的预测粒度对预测精度产生的影响。

7.2.1 Prophet 预测模型

由问题 1 分析可知，产品需求量一般有月周期和季节周期，而 Prophet 刚好适用于各种具有潜在特殊特征的预测问题，如有明显内在规律的、周期性的数据，如季节性变化、节假日趋势等，并对时间序列趋势变化点的检测、季节性、节假日以及突发事件具有更好的拟合效果。

7.2.1.1 Prophet 模型原理

Prophet 模型是 Facebook 在 2017 年 2 月开源的一款基于 Python 和 R 语言的数据预测框架。Prophet 是基于时间序列分解和机器学习拟合而设计的一种预测时间序列的加法回归模型，由趋势项、周期项、节假日效应、残差项四部分组成，与传统的 ARIMA 不同，它整合了建模、评估两大块，采用了时间序列分解的方法，实现了时间序列模型的快速迭代优化，基本形式如下：

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

$g(t)$表示趋势项，表示对时间序列中的非周期变化趋势的响应；$s(t)$表示周期项，也叫季节项，一般默认以周或年为单位；$h(t)$表示节假日项，表示当天是否是节假日，体现了节假日等持续一天或几天的无规律变化；$\varepsilon(t)$表示误差项，通常服从正态分布，表示模型未预测到的波动。

趋势项$g(t)$是一个基于分段逻辑回归增长模型：

$$g(t) = \frac{C(t)}{1 + \exp\left(-\left(k + a(t)^t\delta \cdot \left(t - (m + a(t)^T\gamma)\right)\right)\right.}$$

其中，

$$\begin{aligned} a(t) &= (a_1(t), \dots, a_s(t))^T \\ \delta &= (\delta_1, \dots \delta_s)^T \\ \gamma &= (\gamma_1, \dots \gamma_s)^T \end{aligned}$$

$C(t)$表示模型容量，k表示增长率，δ表示增长率的变化量。随着t的增加，$g(t)$趋于$C(t)$。在使用 Prophet 的 growth='logistic'时需要提前设置$C(t)$的取值。

周期（季节）项$s(t)$在模型原理中通过傅里叶级数来近似表达周期性分量：

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

其中，T表示某个固定的周期，$2n$表示期望在模型中使用的个数。对于以年为周期的序列$(T=365.25), N=10,$ 参数可以形成向量

$$\beta=(a_1,b_1,\ldots,a_N,b_N)^T$$

节日项$h(t)$将每个节假日在不同时刻下的影响视作独立模型，为每个独立模型设置了一个虚拟变量，该模型可表示为：

$$h(t)=\sum_{i=1}^{L}K_i\, l\ \ (t\in D_i)$$

$$Z(t)=[\ \ l(t\in D_1),\cdots,l(t\in D_L)\ \ ]$$

$$h(t)=Z(t)_K, K\sim Normal(0,v)$$

其中，K_i表示窗口期中的节假日对预测值的影响；D_i表示第i个虚拟变量，若时间变量 t 属于虚拟变量，则虚拟变量D_i值为 1，否则为 0；i表示节假日，D_i表示窗口期中包含的时间t。

7.2.1.2 模型的构建与调参

第一问的分析结果显示，节假日、促销日、季节性对产品的需求量有较大的影响。Prophet 模型在预测时间序列数据时，能较好的适应节假日和促销日的趋势变化点，是一个由周期项、趋势项、节假日项和误差项组成的加法模型。它对原始数据建模，需要 ds 时间列和 y 数值列，周期性是模型的重要参数之一，采用标准傅里叶级数，可以调整傅里叶级数的项数 n，值越大项数越多，模型拟合越好，但也容易过拟合，默认有年、周和天的周期性；seasonality_prior_scale 控制季节性的灵活性，较大的值可以使得季节性适应较大的波动，较小的值可以减小季节性的幅度；holiday 是特殊日期的时间，设定中国一些主要节假日、促销日及其影响范围；holidays_prior_scale 控制节假日效果的灵活性，值越大，表示节假日对模型的影响越大；seasonality_mode 表示季节的模型方式，本文默认加法，所有参数通过实验调整。

以 2015 年 9 月到 2018 年 9 月的数据作为训练集，2018 年 10 月到 2018 年 12 月共 3 个月的数据作为测试集，判断模型效果。使用测试集均方根误差作为调参指标，其中，调参范围和节假日信息以及促销日信息如下表 7.2-表 7.3。

表 7.1 **调参范围**

参数名称	调参范围
年周期性 yearly_seasonality	[10,35]
周周期性 weekly_seasonality	[3,7]
特殊日期信息 holidays	手动设置
节假日影响力 holidays_prior_scale	[0.01,10]
季节影响力 seasonality_prior_scale	[0.01,20]
拐点控制参数 changepoint_prior_scale	[0.001,0.5]

表 7.2 节假日信息列表

节假日	日期	lower_window	upper_window
元旦	2016-01-01；2017-01-01；2018-01-01；2019-01-01	0	1
春节	2016-02-08；2017-01-28；2018-02-16；2019-02-05	-1	6
清明节	2016-04-04；2017-04-04；2018-04-05	0	1
五一	2016-05-01；2017-05-01；2018-05-01	0	5
中秋	2015-09-15；2016-09-15；2017-09-04；2018-09-24	0	3
国庆	2015-10-01；2016-10-01；2017-10-01；2018-10-01	0	7

表 7.3 促销日信息列表

促销日	日期	lower_window	upper_window
618	2016-06-18；2017-06-18；2018-06-18；	-14	7
双 11	2015-11-11；2016-11-11；2017-11-11；2018-11-11	-14	7
双 12	2015-12-12；2016-12-12；2017-12-12；2018-12-12	-14	7

7.2.1.3 模型预测效果评估

使用调参所得到的最优参数，对测试集上的所有产品进行预测，以 MAE、RMSE、R2 对 Prophet 模型的预测效果进行评估，如表 7.4 所示。

表 7.4 Prophet 模型预测效果

	MAE	RMSE	R^2
天粒度	17.55	100.43	0.18
周粒度	100.73	562.20	0.32
月粒度	557.68	1647.98	0.45

由天、周、月粒度的评估结果可得，对于 RMSE 和 MAE：天粒度<周粒度<月粒度。而 R^2：天粒度<周粒度<月粒度。但由于不同时间粒度对应的产品需求量的量级不同，因此不能只通过 MAE 和 RMSE 来衡量不同预测粒度的预测精度。再比较 R^2，天和周粒度的 R^2 都较小，月粒度的 R^2 更接近 1，结合例子中真实值和预测值散点图，可看出，月预测粒度的预测精度更好。

预测粒度为天和周的预测效果不如以月时间粒度的预测效果，说明不同的预测粒度时产品需求量的预测精度不同，时间跨度大的时间粒度的预测效果会更好，这可能与产品需求量的月周期性有关，而天和周粒度的产品需求量可能存在随机性，所以预测效果不如月粒度好。

7.2.1.4 模型预测结果

考虑到预测的产品数量较多，采用 python 将 Prophet 时间序列模型封装成函数求解。用最优参数建模，以 2015 年 9 月到 2018 年 12 月的数据作为训练集，对 result1.xlsx 中的产品预测 2019 年 1 月到 3 月的需求量。

由于预测产品数量较多，下面以 result1.xlsx 中产品编号：21271，销售区域：101 为例，对时间粒度为月、周、天的需求量预测结果进行分析。

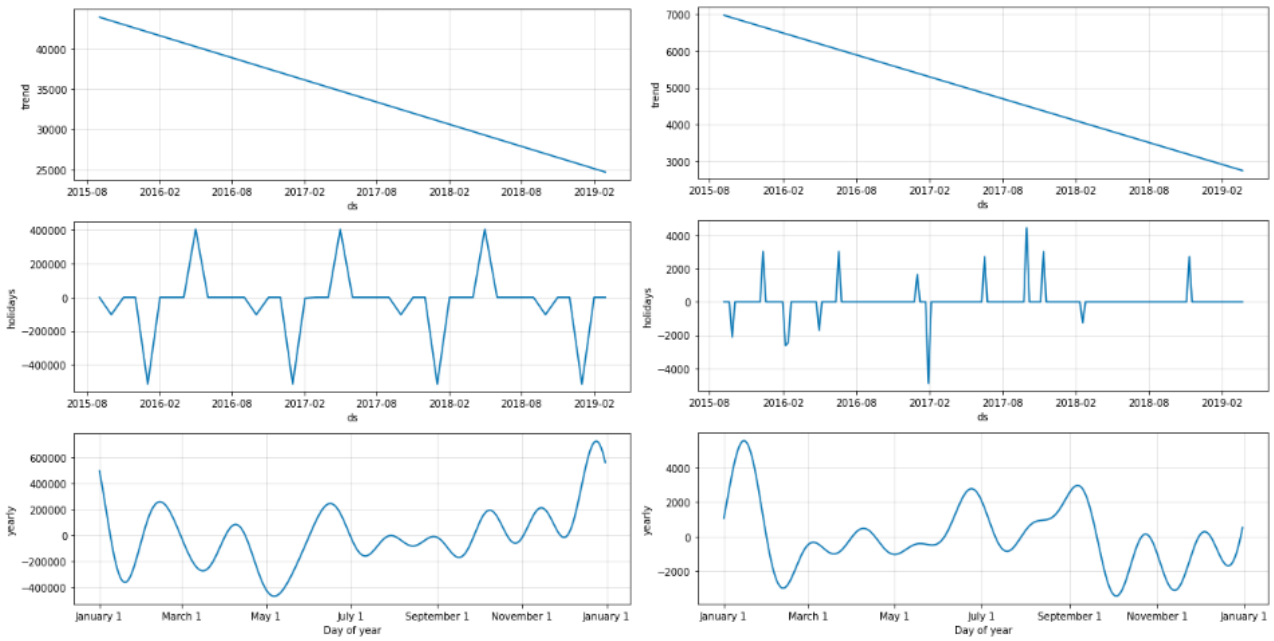


图 7.4 月、周预测成分分析图

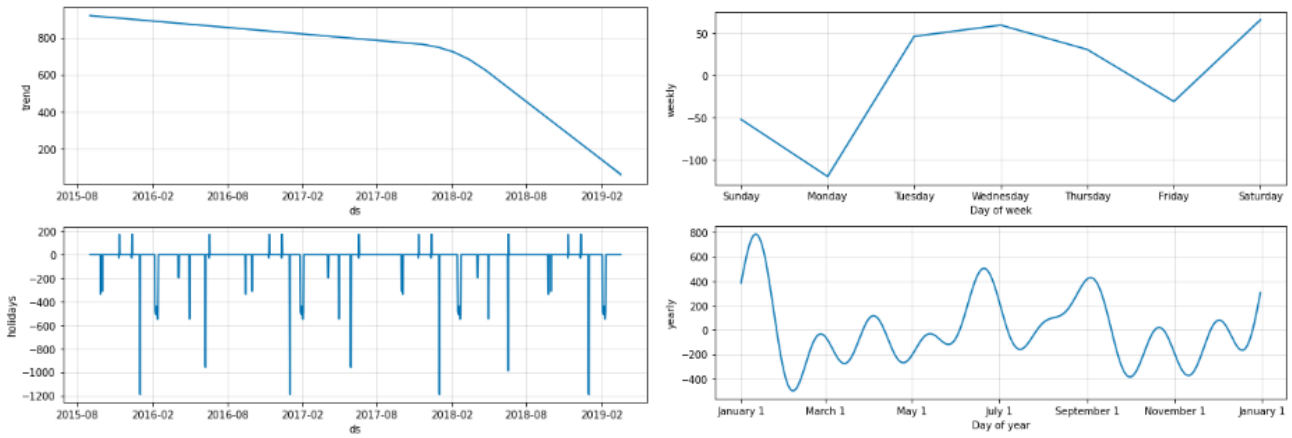


图 7.5 天预测成分分析图

图 7.4、图 7.5 给出了模型对数据各成分单独分析的结果，从上到下依次是 2015-2018 年该商品需求量被 Prophet 加法模型分解的长期趋势（trend）、节假日趋势（holiday）、周趋势（weekly）、年度趋势（yearly）。长期趋势体现该商品需求量逐年下降；节假日趋势中，几乎所有节假日和促销日都是负向影响，其中春节的影响最大，由于节假日经销商、工厂停工休息，人们选择出游等，导致需求降低，由于经销商一般在促销日前几个月就开始订货，因此促销日临近的几天需求量反而会较平日需求量有所下降；年度趋势体现订单量在每年的春秋季较高，夏季较低；当模型时间粒度为天时，可分析周趋势，周趋势显示，星期一需求量最低。

该模型预测结果如图 7.6、图 7.7 所示。

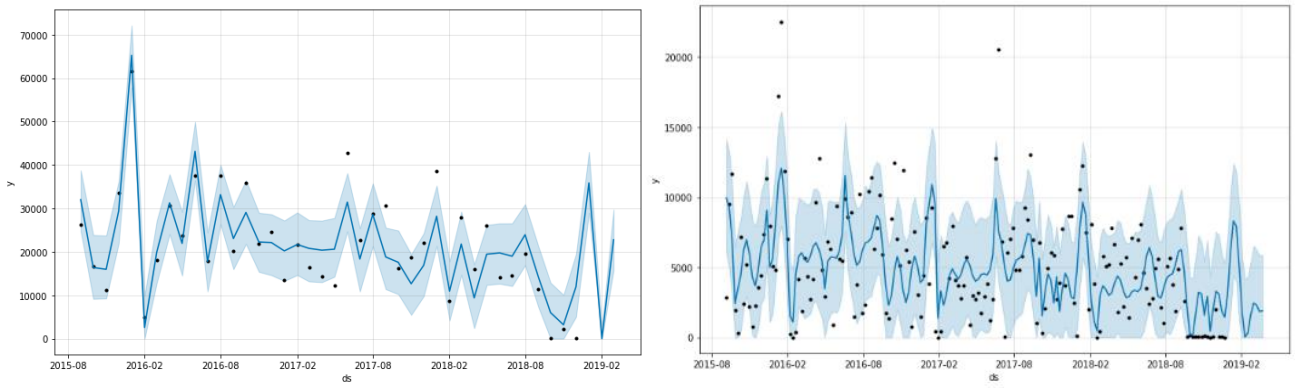


图 7.6 月、周预测结果

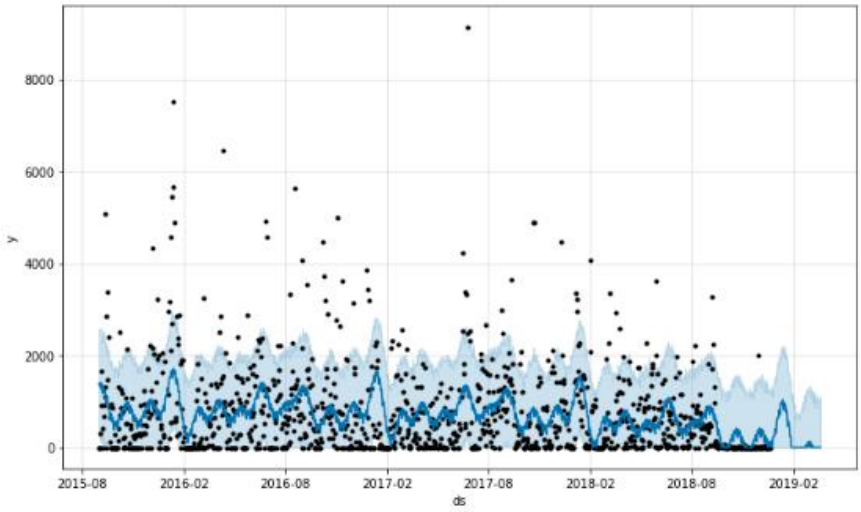


图 7.7 日预测结果

由图可知，时间预测粒度为月的预测结果显示 2019 年 1 月销售量为 35871，2019 年 2 月销售量为 0，2019 年 3 月销售量为 22780。由于受 2 月份春节放假和需求量逐渐下降的长期趋势的影响，1 月份到 3 月份需求量先增后减的趋势符合现实。由于周、天的需求量随机性较大，因此时间粒度为周、天的预测结果不是很理想。

7.2.2 随机森林回归模型

Prophet 时间序列模型只是针对时间序列上产品需求量的预测，没有考虑各种数据特征的影响，因此我们利用随机森林回归模型，基于多种数据特征对产品需求量进行集成学习。

7.2.2.1 随机森林回归原理

随机森林（Random Forest）是由 Leo Breiman 提出一种基于 Bagging 和决策树的集成学习算法，集成学习的大致思路是训练多个弱模型打包起来组成一个强模型，强模型的性能要比单个弱模型好很多，其中的弱模型可以是决策树、SVM 等模型。在随机森林中，弱模型选用决策树，其核心思想是一个由多颗随机生成的决策树组成的森林，每一个数据输入后，由各个不相关的决策树做分类或者回归，并投票决定该数据该如何分类或者回归。

单独的决策树算法往往会在数据比较复杂时的效果较差，其由于过拟合的缺点，使得模型不具有普遍性和工程上的应用能力。为了弥补决策树的不足，随机森林引入了随机采样的概念，即在森林中的决策树在训练中得到的数据都是全局样本中的一部分，在很大程度上改善了决策树容易过拟合的问题，不易受噪声和异常值干扰，并且可以通过算法本身进行特征选择，不需

要对数据进行规范化，相比于 SVM、ANN 等算法，工程化更简单，容易并行化，对高维数据也能有较好的拟合效果。

常见的决策树算法有三种：ID3、C4.5 和 CART，这三种算法都是采用自上而下的贪婪算法来构建树状结构，在每个节点处选择一个最优特征进行分裂，递归建树直到满足终止条件。随机森林的算法有分类和回归两种情况，若用于分类，则决策树使用 ID3 和 C4.5；若用于回归，则一般使用 CART 算法。CART 算法生成的是二叉决策树，其本质是对特征空间进行二元划分，并且能够对标称属性（nominal attribute）与连续属性（continuous attribute）进行分裂。

CART 回归树在分裂时选择特征和划分点的原则是最小均方差(MSE)，即对于任意划分特征A，对应的任意划分点s两边划分成的数据集D_1和D_2，求出使D_1和D_2各自集合的均方差最小，同时D_1和D_2的均方差之和最小所对应的特征和特征值划分点，表达式为：

$$\min_{A,S}\left[\min_{c_1}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2\right]$$

其中，c_1为D_1数据集的样本输出均值，c_2为D_2数据集的样本输出均值。

在训练阶段，为了使用多个不同的子模型来增加最终模型预测结果的鲁棒性和稳定性，随机森林使用 Bootstrap 采样从输入训练数据集中采集多个不同的子训练数据集来依次训练多个不同决策树。它从原始训练样本集中有放回地随机抽取训练样本生成训练子集，并训练得到单个弱学习器，在随机森林回归模型中该弱学习器为回归树，重复这一过程生成多棵回归树组成随机森林，并由所有树的预测值的平均值决定最终预测结果。Bootstrap 是对输入训练样本集合 D 进行 N 次放回重复抽样得到样本集 D_b，每次抽样每个样本被抽取的概率为，进行 N 次抽样，被选中的概率为

$$\left(1-(1-\frac{1}{N})^N\right)$$

当 N 足够大时，即

$$1-\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N=1-\frac{1}{e}\cong 0.633$$

这意味着原始训练样本集 D 中每个样本会以 63%的概率被选中，用于训练随机森林回归模型。剩余 37%的样本则为袋外数据，OOB 估计就是使用袋外数据来评估模型的拟合效果。对于每一棵回归树，都可以计算得到一个 OOB 误差估计，通过平均随机森林中所有回归树的 OOB 误差估计，可以获得随机森林的泛化误差估计。

在预测阶段，随机森林将内部多个决策树的预测结果取平均得到最终的结果。针对某一输入样本，从二叉决策树的根节点起，判断当前节点是否为叶子节点，如果是则返回叶子节点的预测值即当前叶子中样本目标变量的平均值，如果不是则进入下一步；根据当前节点的切分变量的和切分值，将样本中对应变量的值与节点的切分值对比。如果样本变量值小于等于当前节点切分值，则访问当前节点的左子节点；如果样本变量值大于当前节点切分值，则访问当前节点的右子节点；循环步骤 2，直到访问到叶子节点，并返回叶子节点的预测值。

综上可知，建立随机森林回归模型的基本步骤如下：

(1) 抽样：从训练数据集 S 中，通过有放回的 Boostrasp 抽样，生成 K 组数据集，每组数据集分为被抽中数据与未被抽中数据 2 种，每组数据集会通过训练产生一个决策树。

(2)生长：通过训练数据对每个决策树进行训练。在每次分节点时，从 M 个属性中随机选取 m 个特征，依据 Gini 指标选取最优特征进行分支充分生长，直到无法再生长为止，不进行剪枝。

(3) 利用袋外数据检验模型的精度，由于袋外数据未参与建模，其能一定程度上检验模型效果与泛化能力。通过袋外数据的预测误差，确定算法中最佳决策树数目重新进行建模。

(4) 利用确定的模型对新数据集进行预测，所有决策树的预测结果的平均即为最终的输出结果。

随机森林的最大优势是每个决策树均利用所有样本中的一部分，并只抽取其中一部分属性进行建模。这种做法能极大的提高模型的多样性，最小化了各棵决策树的相关性。依照集成学习理论来说，基学习器的多样性越强，其泛化能力就越高[2]。

7.2.2.2 模型的构建与调参

利用 Python 建立随机森林模型，采用网格搜索的方式进行调参优化。网格搜索是在限定指定的参数范围，按步长依次调整参数，利用调整的参数训练学习器，从所有的参数中找到在验证集上精度最高的参数。

固定随机种子为 0，以均方根误差（RMSE）为评估指标，设置参数范围进行网格搜索来调参优化。具体的模型调参及最优参数见表 7.5。

表 7.5 随机森林模型调参及最优参数

参数	调参范围	最优参数		
		天粒度	周粒度	月粒度
max_depth	4-12	10	12	12
n_estimators	50-800	800	800	100
min_samples_leaf	1-20	12	8	2
min_samples_split	1-20	6	2	6

7.2.2.3 模型预测效果评估

利用网格搜索得到的最优参数构建模型，用测试集的数据训练模型，并在测试集上进行预测。以 MAE、RMSE、R^2 对模型的预测效果进行评估，如表 7.6 所示。

表 7.6 随机森林模型的预测效果

	MAE	RMSE	R^2
天粒度	18.46	88.13	0.15
周粒度	108.38	403.68	0.16
月粒度	422.39	1360.93	0.28

在所有产品的不同粒度的测试集上的预测值和真实值对比散点图如下图所示。

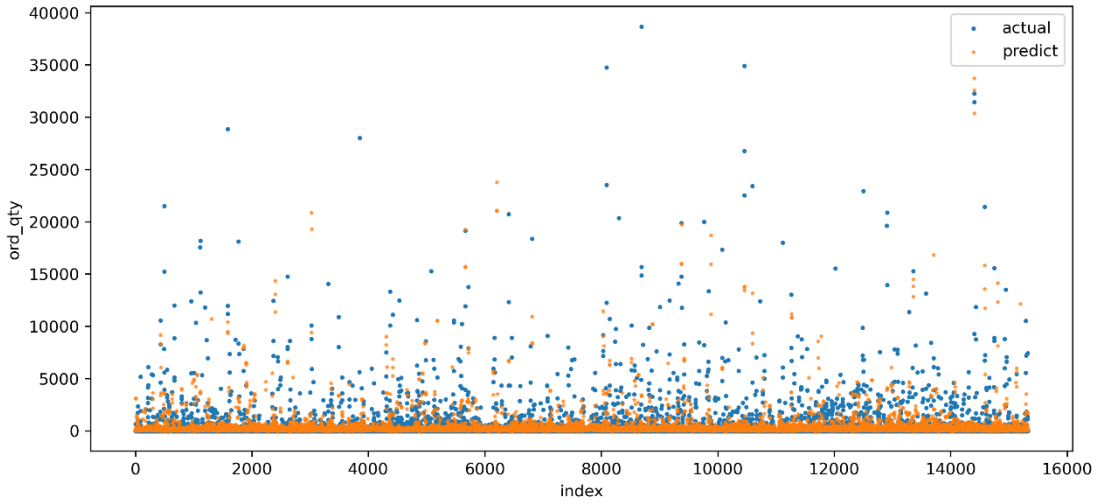


图 7.8 月粒度预测值与真实值对比

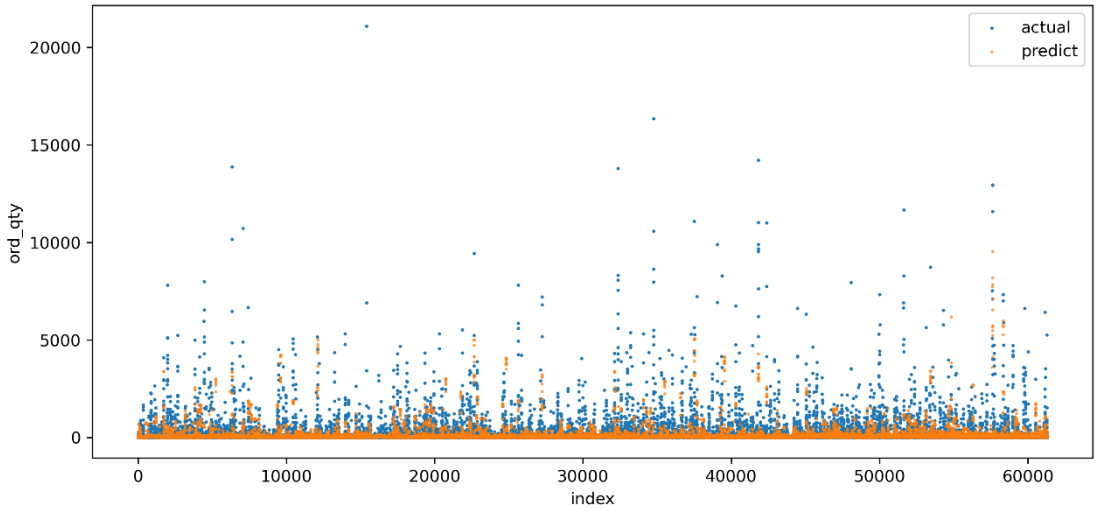


图 7.9 周粒度预测值与真实值对比

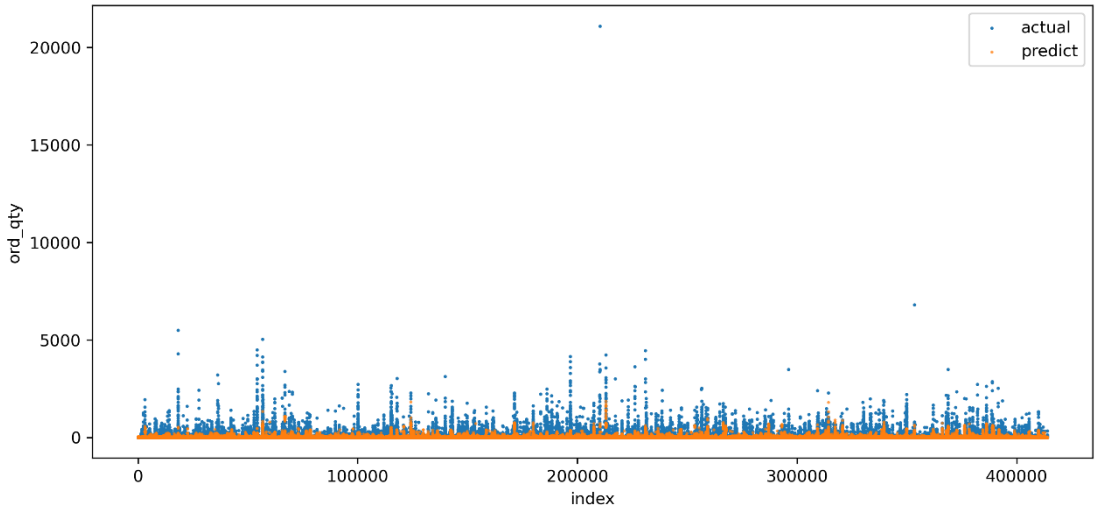


图 7.10 天粒度预测值与真实值对比

由天、周、月粒度的评估结果可得，对于 RMSE 和 MAE：天粒度<周粒度<月粒度。而 R^2：天粒度<周粒度<月粒度。再结合不同粒度的测试集上的预测值和真实值的对比图，天、周、月的需求量的整体水平是不同的，天粒度的数据集需求量整体水平在 0-5000 之间，月的整体需求量水平在 5000-10000 之间，天、周、月的需求量整体水平依次增高，因此造成整体水平越低的 MAE 和 RMSE 较小，因此不能只通过 MAE 和 RMSE 来衡量不同预测粒度的预测精度。再比较 R^2，天和周粒度的 R^2 都较小，月粒度的 R^2 更接近 1，且结合真实值和预测值散点图，可看出，月预测粒度的预测精度更好。

根据月粒度的模型，得到特征权重如下图 7.11 所示：

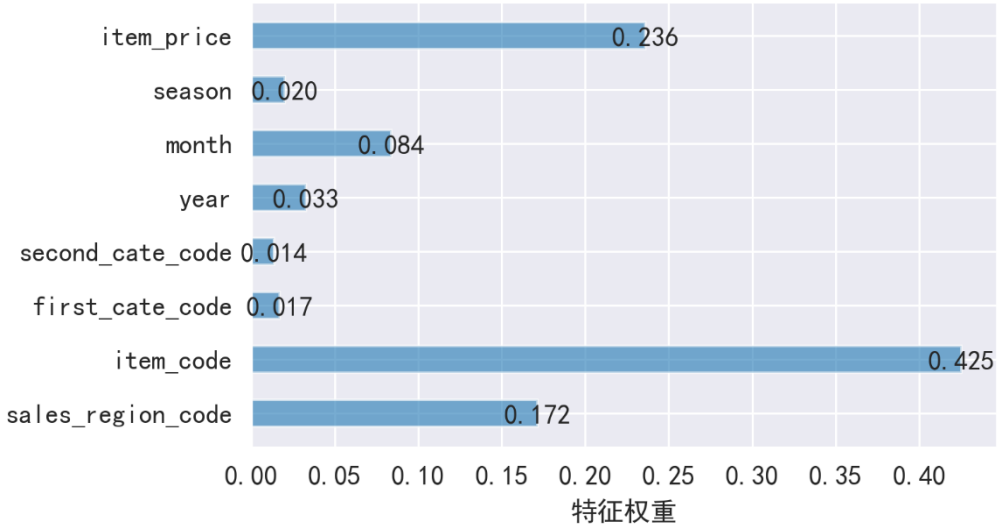


图 7.11 随机森林月粒度特征权重

由随机森林回归模型得出的各特征权重中，产品类别的特征权重最高，产品价格次之，销售区域的权重也相对较高，说明不同类别、不同价格、不同销售区域间的产品需求量显著不同，刚好符合问题 1 对产品需求量特征的分析，产品需求量受产品价格波动的影响，且不同销售区域下的产品需求量不同，但也有部分销售区域的需求量差异较小，所以该特征权重相对其他两个特征较小。而产品类别的需求量几乎占特征权重的一半，说明在该模型中，产品类别是影响产品需求量的最关键因素，不同产品的需求量有较明显的差异。

7.2.3 XGBoost 预测模型

采用 XGBoost 算法可以针对未来三个月产品需求量预测问题构建有监督训练的预测模型。由于产品需求量的数据量规模较大，且数据特征较多，而 XGBoost 模型不仅训练速度快，并综合考虑多种影响因素以及数据特征，同时还具有多项避免模型训练过拟合的措施，在进行产品需求量预测时具有一定的优势。

7.2.3.1 XGBoost 模型原理

XGBoost（Extreme Gradient Boost）模型是 Chen 等提出的一种特殊的基于梯度提升决策树改进的有监督学习算法，该算法可以解决分类、回归、排序等问题，是 GBDT 方法里的完全加强版本。XGBoost 本质上还是基于树结构并结合集成学习的一种算法，其基础树结构为分类回归树，在某些方面和传统的梯度提升树 GBDT 算法类似，都是以 CART 作为基分类器，并通过计算目标函数的负梯度，让子模型以负梯度作为目标值进行训练，从而保证模型训练的优化方向，实现多颗子模型融合的高精度集成模型。

XGBoost 的核心算法思想是：

（1）不断的添加树，不断地进行特征分裂来生长一棵树，每次添加一棵树，其实是学习一个新函数$f(x)$，去拟合上次预测的残差。

（2）基于训练完成得到 K 棵树的树形结构，要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点上，每个叶子节点就对应一个分数，同时每个叶子结点本身含有一个权重分数。

（3）该样本的最终预测值是由将每棵树所在的叶子结点对应的分数加权的总和。

假设训练采用的数据集有n个样本，m个特征，则定义为$D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m, y_i \in R)$，其中$x_i$表示第$i$个样本，$y_i$表示第$i$个样本的类别标签，模型共包含 K 棵树，则 XGBoost 模型的定义如下：

$$\hat{y_i} = F_K(x_i) = F_{K-1}(x_i) + f_K(x_i)$$

其中，$\hat{y_i}$表示预测的样本标签，$F_{K-1}(x_i)$表示 XGBoost 的前$K-1$棵决策树预测值之和，$f_K(x_i)$表示第 K 棵决策树。

在训练 XGBoost 模型之前，需要先设置一个目标函数作为该算法的优化方向，使得树群的预测值$\hat{y_i}$尽可能接近真实值y_i。 因此，XGBoost 的目标函数定义如下：

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant$$

传统 GBDT 与 XGBoost 两者最大区别就在于对目标函数的定义，XGBoost 的目标函数包括损失函数和正则项两部分，其中$constant$为常数项。

（1）损失函数l：用于评估模型拟合数据的程度，该函数必须是可微分的凸函数。通常用其一阶导数指出梯度下降的方向，XGBoost 还计算了它的二阶导数，进一步考虑了梯度的变化趋势，拟合更快，精度更高。

（2）正则项$\Omega(f_t)$：用来控制模型的复杂程度，防止过拟合。叶子结点越多，模型越大，不仅运算时间长，超过一定限度后还会过拟合，导致分类效果的下降。XGBoost 的正则项是一个惩罚机制，叶子结点的数量越多，惩罚力度越大，从而限制他们的数量。

XGBoost 和传统 GBDT 都属于集成学习算法，都采用回归树作为基学习器，使其具有很强的学习能力。与 GBDT 算法还有一些不同的是，XGBoost 算法采取很多手段防止模型训练过拟合：①在损失函数里引入了正则项，用于限制模型的复杂程度，使模型更加简单直接；②通过借鉴随机森林，支持列抽样的方式，通过确定用于每个基学习器的样本数据量来降低模型的计算复杂度。而且传统的 GBDT 在优化时只用到一阶导数，而 XGBoost 对损失函数进行二阶泰勒展开，还可以支持自定义的损失函数，只要能满足二阶连续可导函数都可作为损失函数。

因此，相对于传统 GBDT 模型，XGBoost 模型具有善于捕捉复杂数据之间的依赖关系，从大规模的数据中获取有效模型的优点，且训练速度快，对于复杂特征关系的分析具有一定的优势。

7.2.3.2 模型的构建与调参

利用 Python 建立 XGBoost 模型，采用网格搜索的方式进行调参优化。网格搜索是在限定指定的参数范围，按步长依次调整参数，利用调整的参数训练学习器，从所有的参数中找到在验证集上精度最高的参数。

固定随机种子为 0，以均方根误差（RMSE）为评估指标，设置参数范围进行网格搜索来调参优化。具体的模型调参及最优参数见表 7.7。

表 7.7 XGBoost 模型调参及最优参数

参数	调参范围	最优参数		
		天粒度	周粒度	月粒度
max_depth	4-10	6	10	10
n_estimators	50-1000	1000	200	1000
learning_rate	0.01-0.4	0.2	0.1	0.05
gamma	0-0.4	0.3	0.2	0.1
reg_alpha	0.0001-100	0.001	0.0001	0.01
reg_lambda	0.0001-100	1	0.001	0.001
subsample	0.6-0.9	0.9	0.6	0.7

7.2.3.3 模型预测效果评估

利用网格搜索得到的最优参数构建模型，用测试集的数据训练模型，并在测试集上进行预测。以 MAE、RMSE、R^2 对模型的预测效果进行评估，如表 7.8 所示。

表 7.8 XGBoost 模型的预测效果

	MAE	RMSE	R^2
天粒度	18.46	86.51	0.18
周粒度	107.42	377.59	0.29
月粒度	356.63	1189.87	0.38

在所有产品的不同粒度的测试集上的预测值和真实值对比散点图如下图所示。

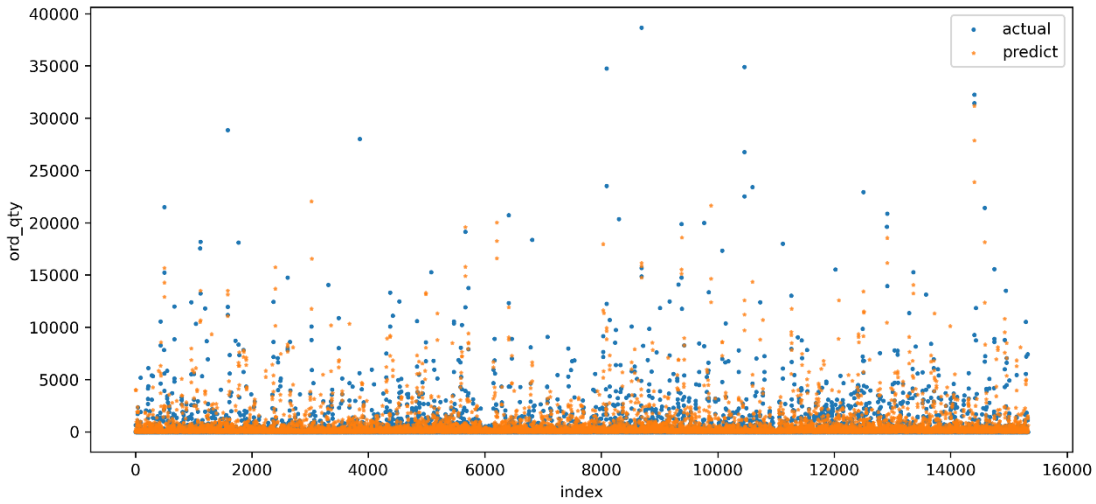


图 7.12 月粒度预测值与真实值对比

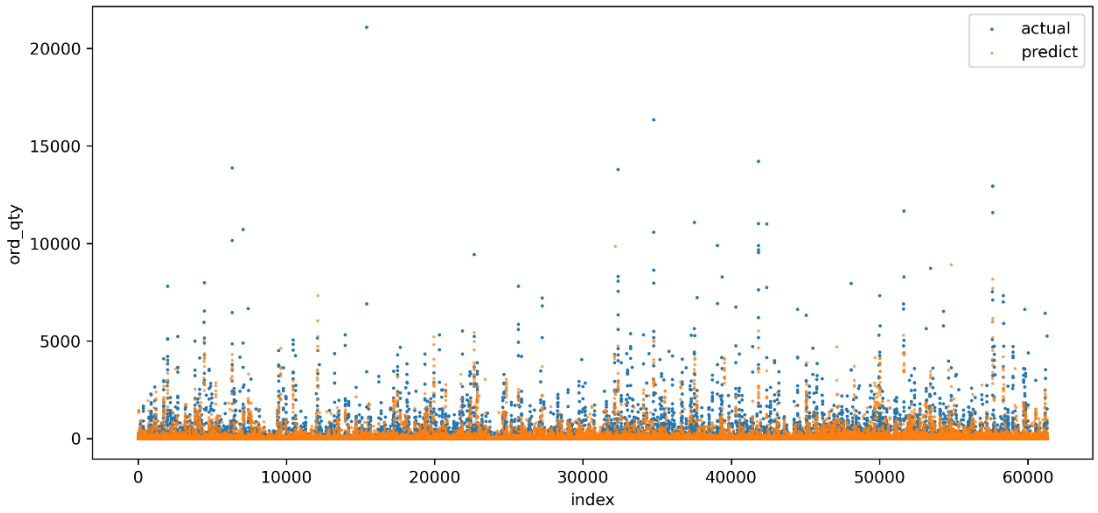


图 7.13 周粒度预测值与真实值对比

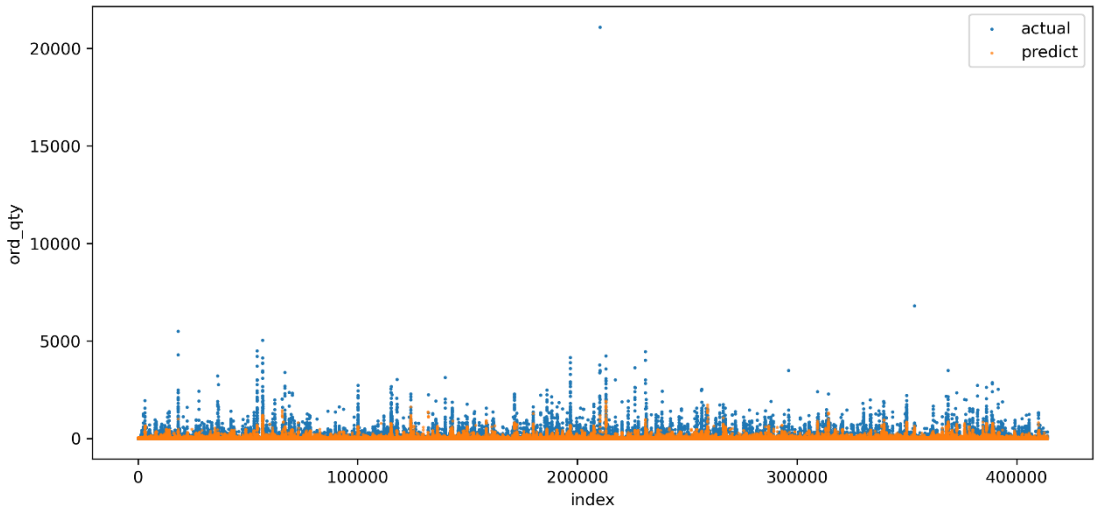


图 7.14 天粒度预测值与真实值对比

由天、周、月粒度的评估结果可得，对于 RMSE 和 MAE：天粒度<周粒度<月粒度。而 R^2：天粒度<周粒度<月粒度。再结合测试集上的预测值和真实值的对比图，天、周、月的需求量整体水平依次增高，造成整体水平越低的 MAE 和 RMSE 较小，因此不能只通过 MAE 和 RMSE 来衡量不同预测粒度的预测精度。再比较 R^2，天和周粒度的 R^2 都较小，月粒度的 R^2 更接近 1，且结合真实值和预测值散点图，可看出，月预测粒度的预测精度更好。

根据月粒度的模型，得到特征权重如图 7.15 所示：

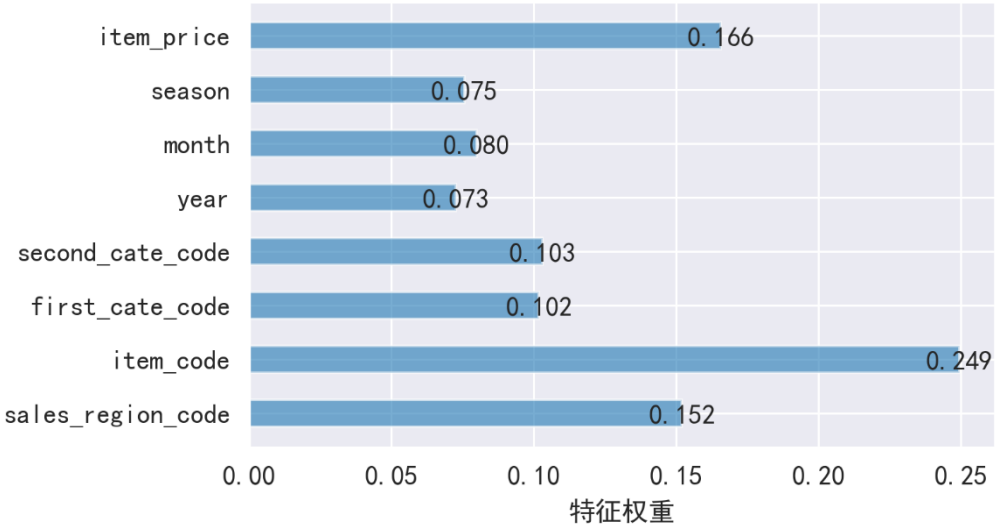


图 7.15 月粒度模型特征权重

在 XGBoost 模型得到的各特征权重中，产品类别所占权重最高，产品价格次之，年的特征权重最小。与随机森林回归模型相比，各特征权重分布较为均匀，说明 XGBoost 能够较好的运用问题 1 的各个特征进行分析建模。产品在线下实体经销商的需求量一般都会高于线上电商平台的需求量，所以不同的产品销售渠道会对产品需求量的影响；且相对随机森林回归模型，产品大类和产品细类的特征权重更高，说明在 XGBoost 模型中产品大类和产品细类特征的有效性大于比年、季、月的特征，这与问题 1 逐步回归法的结果基本一致，说明该模型效果较为合理。

7.3 三种模型不同预测粒度效果对比

表 7.9 三种模型不同预测粒度效果

	Prophet			随机森林回归			XGBoost		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
天粒度	17.55	100.43	0.12	18.46	88.13	0.15	18.46	86.51	0.18
周粒度	100.73	562.20	0.16	108.38	403.68	0.16	107.42	377.59	0.29
月粒度	557.68	1647.98	0.25	422.39	1360.93	0.28	356.63	1189.87	0.38

7.3.1 三种模型预测效果的比较

观察表 7.9 可以看出，在相同时间粒度下，对于 MAE 和 RMSE：Prophet 模型＞随机森林回归模型＞XGBoost 模型。对于 R^2：Prophet 模型＜随机森林回归模型＜XGBoost 模型，说明从整体上说，XGBoost 的拟合效果要优于其他两种模型。

综合 7.2 对预测结果的分析可知，Prophet 时间序列模型需要对 2619 个产品分别构建模型并调参，效率较低，可能对部分时间特征不明显的产品的预测效果较差，且运行速度是三种模型中最长的，但部分序列的信息还是没有被预测出来；随机森林回归模型和 XGBoost 模型能够对所有数据特征进行处理，能够最大程度上保留问题 1 所得出的数据特征，且效率相对较高，而 XGBoost 的预测效果比随机森林回归模型的预测效果好。因此，后面我们将基于上述算法基础对 XGBoost 模型进行改进。

7.3.2 不同预测粒度对预测精度的影响

根据 7.2 对三种模型的预测过程和预测结果的分析可知，从整体上看，月时间粒度对应的产品需求量的数量级比周和天粒度的大，所以月粒度的 MAE 和 RMSE 会大于周和月粒度的是较为合理的，所以应从 R^2 作比较，三种模型中 R^2 的值都是月粒度＞周粒度＞天粒度，因此，三种时间粒度下预测精度的大小为：月粒度＞周粒度＞天粒度，由此可见，对天粒度的产品需求量进行预测的准确率是最低的，对月粒度预测精度会更高。

基于问题 1 观察到产品需求量呈现月周期性，从经销商和制造企业的交易方式来看，为了方便企业调控生产，经销商一般会在提前一段时间订购未来所需要的产品，企业根据订单再进行生产出货，所以可合理推测对企业来说月的需求量会比周和天更重要，因为这决定企业在未来一个月内大致的生产量和业务收入。所以对产品需求量的预测来说，以月粒度进行预测的误差要较小，即预测准确率要高，才能让企业更有效地制定运营计划，而天粒度的产品需求量随机性比较强，除了促销活动的影响外，其他正常时间段内经销商每天的不同产品间需求量波动变化较为随机，经销商可能每日都会进货，但所订购的产品每日却各有不同，所以误差较大在所难免。因此，从实际应用上讲，企业从不同的时间粒度进行预测产生的误差的容忍度不同，所以更重要的是如何提高对每个产品需求量的月粒度预测。

综上可知，XGBoost 模型在月粒度上的预测精度是最高的，R^2 达到了 0.38，但还仍然存在着不小的误差，所以我们将进一步对 XGBoost 模型进行改进，以提高月粒度预测的准确率。

7.4 加入差分、滞后、窗口特征改进的 XGBoost

考虑到选择的 XGBoost 模型可以处理有空值的特征，我们进行特征工程处理，加入滞后、差分、时间拓展窗口、时间滑动窗口的特征，对原来构建的 XGBoost 模型进行改进，以进一步提高模型的预测效果。基于月粒度的历史数据，我们对改进的 XGBoost 进行调参，以最优参数构建模型来预测未来三月的需求量。

7.4.1 特征工程介绍与处理流程

由于时间序列是通过历史来预测未来，那么时间序列的历史数据，也就是当前时间点之前的信息就非常有用，通过构建时序值衍生的特征可以引入时间序列的趋势因素、季节性周期性因素以及一些不规则的变动，进一步提升机器学习对时间序列问题的预测效果。具体来说这部分特征可以分为四种：差分特征、滞后特征、滑动窗口特征和拓展窗口特征。我们对训练数据集进行特征工程，加入差分、滞后、窗口特征。具体介绍及特征构建介绍如下。

差分特征：对于时间序列数据，可以计算每个时间点的差分值，然后将这些差分值作为输入特征，这有助于模型学习时间序列的变化趋势和周期性。考虑到 t 月的需求量可能跟上月需求量的增长量相关，构建了一阶差分特征 diff _1。t 月的一阶差分特征 diff _1 即为 t-1 月的需求量减去 t-2 月的需求量。

滞后特征：根据第一问的分析，我们认为 t 月需求量跟 t-1 月的需求量、t-3 月的需求量、去年同月的数据是高度相关的，分别构建了 1ag_1、1ag_3、1ag_12 作为滞后特征。其中 t 月的 1ag_1、1ag_3、1ag_12 特征分别为 t-1、t-3、t-12 月的需求量。

滑动窗口特征：将时间序列数据按照滑动窗口的方式进行分组，然后统计每个窗口内的统计特征，这种类型的特征叫做滑动窗口统计。对于 t 月需求量，我们以前 3 个月作为窗口，取前 3 个月的统计值（平均值、标准差、最小值、最大值）作为特征，构建 roll_average、roll_std、roll_min、roll_max 作为滑动窗口特征。

拓展窗口特征：拓展窗口统计的数据是某产品整个时间序列全部的数据，并计算统计值。这种特征一般是用在多序列建模，不同的产品的需求量，有着不同的内在属性。因此对所有产

品，计算该产品历史数据的统计值（平均值、标准差、最小值、最大值）作为特征，构建 his_average、his_std、his_min、his_max 作为滑动窗口特征。

总共构建了 12 种时序值衍生特征，部分示例如图 7.16。

:ode	year	month	season	item_price	diff_1	lag_1	lag_3	lag_12	his_average	his_std	his_min	his_max	roll_average	roll_std	roll_min	roll_max
401	2016	9	3	1056.000000	0.0	0.0	0.0	0.0	178.175	335.516967	0	1503	0.000000	0.000000	0.0	0.0
401	2016	10	3	1056.500000	993.0	993.0	0.0	0.0	178.175	335.516967	0	1503	331.000000	573.308817	0.0	993.0
401	2016	11	3	1055.500000	510.0	1503.0	0.0	0.0	178.175	335.516967	0	1503	832.000000	764.325193	0.0	1503.0
401	2016	12	4	1054.000000	-892.0	611.0	993.0	0.0	178.175	335.516967	0	1503	1035.666667	447.528025	611.0	1503.0
401	2017	1	4	975.000000	-603.0	8.0	1503.0	0.0	178.175	335.516967	0	1503	707.333333	752.141166	8.0	1503.0

图 7.16 衍生特征示例

7.4.2 改进的 XGBoost 模型构建及效果评估

利用 Python 建立 XGBoost 模型，以加入差分、滞后、窗口特征的新的数据集进行训练，采用网格搜索的方式进行调参优化。固定随机种子为 0，以均方根误差（RMSE）为评估指标，设置参数范围进行网格搜索来调参优化。具体的模型调参及最优参数见表 7.10。

表 7.10 改进的 XGBoost 模型调参及最优参数

参数	调参范围	最优参数
max_depth	4-10	4
n_estimators	50-1000	300
learning_rate	0.01-0.4	0.03
gamma	0-0.4	0.3
reg_alpha	0.0001-100	100
reg_lambda	0.0001-100	100
subsample	0.6-0.9	0.6

利用网格搜索得到的最优参数构建模型，用月粒度的测试集数据训练模型，并在测试集上进行预测。以 MAE、RMSE、R^2 对模型的预测效果进行评估，改进后的模型与原 XGBoost 模型的预测效果对比如表 7.11 所示。

表 7.11 改进前后预测效果对比

	MAE	RMSE	R^2
改进的 XGBoost	240.58	989.36	0.57
原 XGBoost	356.63	1189.87	0.38

改进的 XGBoost 模型在测试集上的 MAE 和 RMSE 均小于原来构建的模型，而 R^2 大于原来构建的模型，说明改进的 XGBoost 模型的预测精度得到了进一步的提升。

在所有产品的测试集上的预测值和真实值对比散点图如下图所示。

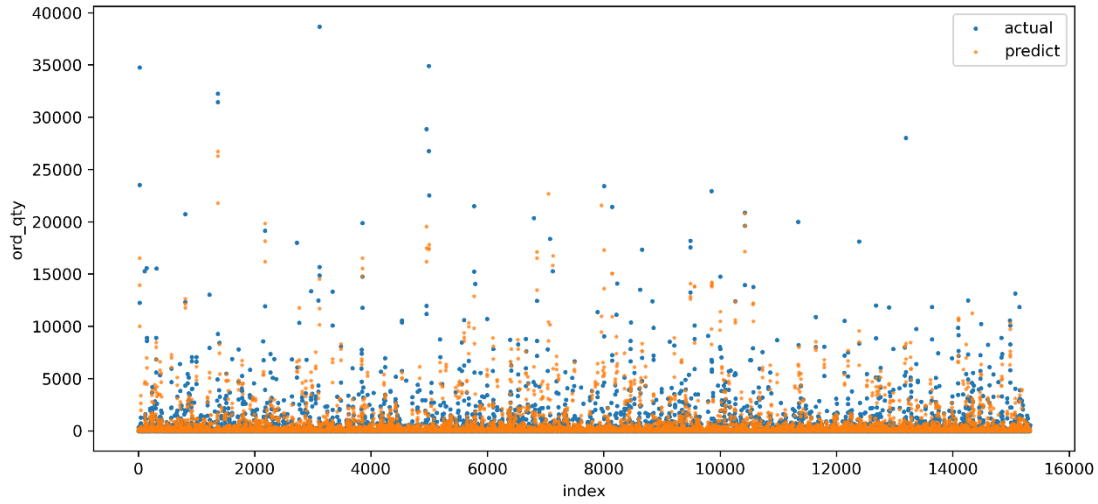


图 7.17 改进后模型预测值和真实值对比

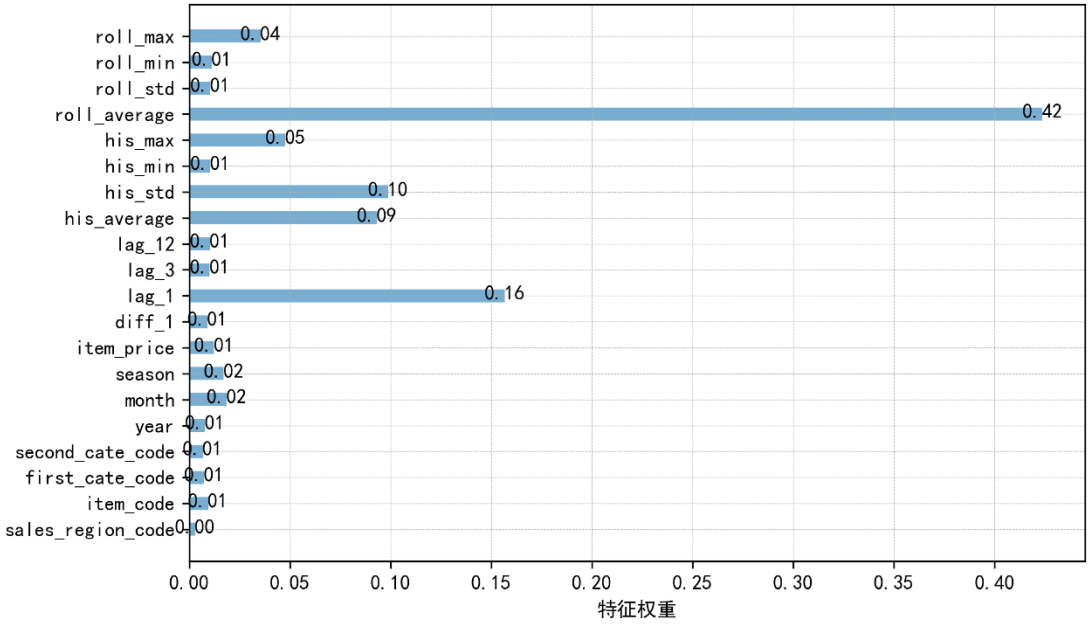


图 7.18 模型的特征权重

由特征权重可看出，新添加的滑动平均、滞后一期、扩展标准差特征分别是模型最重要的三个特征，说明新添加的特征工程相关特征对模型的预测影响显著，使模型的预测精度得到了进一步的提升。

7.5 基于改进的 XGBoost 的未来需求量预测

我们以前面调参得到的最优参数来构建改进的 XGBoost 模型，用所有的历史数据训练模型，从而预测未来三月的需求量。

考虑到 2018 年 12 月的数据只到 20 号的，在构建模型之前，我们先用天粒度构建的 XGBoost 预测 12 月 21 日至 12 月 31 日的需求量，得到 12 月的完整需求量数据，再用该月粒度的历史数据集训练模型。

在预测之前，我们要首先完成预测数据集的特征构建。我们根据历史数据构建 2019 年 1 月的特征，包括年、月、季节、价格和差分、滞后、窗口特征。对于价格特征，若该产品有历史数据，则用要预测的月份的该产品历史数据的平均价格；若该产品没用历史数据，是新产品，

则用要预测的月份的该销售区域的该品类的所有产品的历史数据的平均价格。对于差分、滞后、窗口特征，则根据产品的历史数据构建，而对于无历史数据的新产品，使用该销售区域该品类细的需求量均值作为历史数据的替代。处理好的特征示例如图 7.19。

sales_region_code	item_code	first_cate_code	second_cate_code	year	month	season	item_price	diff_1	lag_1	lag_3	lag_12	his_average	his_std	his_min	his_max	roll_average	roll_std	roll_min	roll_max
101	20002	303	406	2019	1	4	2194.278736	20	56	19	0	14.825	31.286	0	124	37	18.52026	19	56
101	20003	301	405	2019	1	4	788.4485294	85	85	661	534	260.4	394.389	0	1748	248.66667	359.6114	0	661
101	20006	307	403	2019	1	4	590.6867089	94	119	112	515	97.55	156.702	0	662	85.333333	52.36729	25	119
101	20011	303	401	2019	1	4	1642.068842	-107	77.364	177.44	213.11	144.785494	52.2592	67.7963	331.9321	146.3642	59.85388	77.364198	184.29012
101	20014	307	403	2019	1	4	369.112027	0	0	123	8161	2020.65	3569.58	0	15055	41	71.01408	0	123

图 7.19 新产品的数据特征示例

将构建好的特征的数据放入训练好的模型中进行预测，得到2019年1月的需求量预测值。于是再根据 2019 年 1 月份的需求量预测值，构建 2019 年 2 月份的差分、滞后、窗口特征，再将构建好的数据放入模型进行预测，以此类推，最终得到要预测产品的 2019 年 1、2、3 月的需求量预测值。详细预测情况见“result1.xlsx”，其中前 10 种产品的预测值如图 7.20 所示。

sales_region_code	item_code	2019 年 1 月预测需求量	2019 年 2月预测需求量	2019 年 3月预测需求量
101	20002	85.443954	76.63936	73.22213
101	20003	242.44052	251.59131	268.39874
101	20006	103.322266	94.00359	84.31859
101	20011	465.75214	437.82065	448.49545
101	20014	358.19894	750.6803	1474.5554
101	20016	809.42975	1262.7228	1489.5121
101	20018	1561.0542	2329.9753	2768.8848
101	20020	43.2361	55.31169	98.76452
101	20021	20.255087	20.251757	20.251757
101	20024	304.38306	186.2465	185.4738

图 7.20 预测结果示例

8 结论

8.1 研究的优势

（1）本文在已有特征的基础上构建了新的特征，在对产品需求量特征分析时，通过多角度多方面对需求量的影响因素进行研究，由浅入深，从全部产品的总需求量的探究到细分到不同品类的数据特征的分析，从而得出较为准确的结果；

（2）本文在对未来三个月产品需求量的预测中，尝试了时间序列模型和机器学习模型，建立了三模型——prophet 模型、随机森林回归模型以及 XGboost 模型，通过对比三种模型的预测效果，选择更好的模型，使预测结果更加准确。

（3）对预测效果最好的 XGBoost 模型进行特征工程处理，加入滞后、差分、拓展窗口、滑动窗口特征，对模型进行改进，进一步提高 XGBoost 模型对时间序列数据的预测效果，使产品需求量的预测更加准确。

8.2 研究的不足与改进

因为时间和人力原因，本文只对产品需求量的部分数据特征作分析，后续可对产品需求量的各项数据特征从深度和广度进行延伸，更深入的了解产品需求量的特点；在建立预测模型上，只用到了 Prophet 时间序列模型、随机森林回归模型、XGBoost 模型，方式比较单一，模型的误差较大，后续可尝试使用两种组合的集成学习模型，对模型进行优化，从而提高预测的精确

度。而且由于产品需求量数据差异较大，后续可以再将产品分情况分别构建不同的模型，或构建能够根据产品需求量特性自动选择合适模型的自适应模型或混合模型。

8.3 总结

本文主要对制造企业的产品需求量进行深入分析，探究企业产品需求量的影响因素和数据特征，并基于历史数据通过多种方式对该企业未来三个月的产品需求量进行预测。

首先，首先通过逐步回归法进行特征筛选，除促销日外其他数据特征都与需求量有关。考虑到经销商向企业订购产品时一般不是当天订购当天收货，企业需要 时间和人力生产产品，因此，经销商一般会在促销活动之前就订购货物，因此促销日的产品需求量与非促销日无差异。接着，我们对产品订单数据进行深入分析，通过分析产品价格、产品销售区域、节假日、促销活动以及季节因素对产品需求量的影响，可以得出：

①通过产品价格对产品需求量的影响探究发现不同类型产品的价格对需求量的影响不同，对于大部分产品，产品价格越低，需求量越高，但整体上需求量会在一个稳定的价格区间内波动；

②在不同的销售区域内产品的需求量以及平均每订单产品需求量有一定差异，104 区域的产品需求量虽小，但平均每订单的产品需求量高，这可能取决于销售区域距离制造企业的距离，距离远的销售区域每一次订货会增加产品的运输成本，且销售区域 104 的经销商在 2017 年就与该企业停止交易，猜测该销售区域的经销商可能转而向其他供货商订货；

③从整体上看在节假日上的日需求量会小于非节假日的日需求量，但不同类型的节假日对产品需求量的影响大小不同，春节假期的产品需求量缺失，说明在此期间企业无出货数据，需求量为 0，而劳动节和国庆节对产品需求量的影响则相对较小；

④在一些大型促销活动如 618、双 11、双 12 都会持续大约两到三周时间且销售量可观，经销商需在促销活动开始前提前一两个月订购产品，避免因企业出货晚使得促销活动无法顺利进行，因此一般在促销活动开始前一段时间企业的订单需求量会有明显增多的现象；

⑤一般经销商倾向于在气温适宜的春秋季节订购较多的产品，特别是在秋季产品需求量是最高的，而在天气炎热的夏季产品需求量是最低的，而冬季气温低下，又逢春节，其产品需求量仅高于夏季，所以从整体上看，产品的需求量呈现季节性波动。

⑥企业产品的销售方式有线下实体经销商和线上淘宝、京东等电商平台两种。总的来说线下实体经销商的产品需求量远高于线上电商平台的需求量，只有在大型电商促销活动双 11 的影响下，线上的需求量才会短暂高于线下，这也反向说明了双 11 的促销力度最大，且促销活动能够提高产品需求量；

⑦ 一般产品需求量在月头时较低，在月末时较高，需求量会呈现月周期性波动。

基于上述对需求量的特征分析，再深入探究不同品类间产品需求量在这些特征影响下的相同点与不同点，相同点在于不同品类间产品需求量受产品价格、季节因素的影响相同，且在不同时间段下产品需求量的特性是相同的；不同点在于不同品类间产品需求量差异较大，产品需求量受不同销售区域、节假日的影响程度不同，且不同品类在不同销售方式下产品需求量的特性不同。

接着，为了对出货数据进行深入研究，我们使用 K-means 聚类探究产品销售额与产品需求量、产品价格和产品品类的关系，通过观察聚类后五种类型产品的特征，可以看出中低端产品占销售额的主要部分，高端产品价格虽高但需求量较少，因此所占销售份额也较少。

最后，为了从不同时间粒度对 2619 个产品未来三个月的需求量进行预测，将每个产品的缺失日期补全，对应需求量填充为 0，并按天、周、月粒度进行处理并进行特征筛选。以 2015

年 9 月-2018 年 9 月的历史数据作为训练集，以 2018 年 10 月-2018 年 12 月的历史数据作为测试集。由于产品种类多，且部分产品的时间特征并不明显，通过 ARIMA 和 SARIMA 预测的效果很差，所以不采取传统的时间序列模型进行预测，而是选则 XGBoost 模型 、随机森林回归模型、Prophet 模型进行预测。

①基于问题 1 产品需求量的季节性和节假日影响，选择适用于各种明显内在规律、周期性、具有突发事件点的预测问题 Prophet 模型，该模型预测效果天、周、月粒度的 R^2 分别为 0.12、0.16、0.25，天、周粒度的预测精确度比月粒度的差，且该模型的预测效果也比较差；

②基于产品需求量的数据特征分析，利用随机森林回归模型和 XGBoost 模型，能最大程度上保留问题 1 的所有数据特征。随机森林模型的预测效果天、周、月粒度的 R^2 分别为 0.15、0.16、0.28， XGBoost 模型的预测效果天、周、月粒度的 R^2 分别为 0.18、0.29、0.38，天、周粒度的预测精确度比月粒度的差，XGBoost 模型相对其他两个模型的预测效果较好。

从以上三种模型的预测结果可知，在这三种模型中，R^2 的值都是月粒度＞周粒度＞天粒度，因此，三种时间粒度下预测精度的大小为：月粒度＞周粒度＞天粒度，由此可见，对天粒度的产品需求量进行预测的准确率是最低的，对月粒度预测精度会更高。

最后，我们对月数据进行特征工程处理，加入滞后、差分、时间拓展窗口、时间滑动窗口的特征，对原来构建的预测效果最佳的 XGBoost 模型进行改进。改进的 XGBoost 模型在测试集上的 MAE 和 RMSE 均小于原来的模型，且 R^2 提高到 0.51，说明改进的 XGBoost 模型的预测精度得到了进一步的提升。由于在 2619 个产品中有 432 个新产品，没有任何历史数据，所以我们将相同的产品细类数据的均值作为新产品的历史数据，最后基于改进后的 XGBoost 模型对 2619 种产品进行预测。

参考文献

[1]葛娜,孙连英,石晓达等.Prophet-LSTM 组合模型的销售量预测研究[J].计算机科学,2019,46(S1):446-451.

[2]冯明刚,严伟,葛新民,朱林奇.利用随机森林回归算法预测总有机碳含量[J].矿物岩石地球化学通报,2018,37(03):475-481.

[3]吴文培. 基于 Prophet 模型优化及在区域用电量预测中的应用[D].河南大学,2020.

[4]张丹峰. 基于 LightGBM,XGBoost,ERT 混合模型的风机叶片结冰预测研究[D].上海师范大学,2018.

[5]汪撼宇.基于组合模型的燃煤电站电煤库存短期预测方法研究.2020.上海大学,MA thesis.

[6]潘杉. 基于时间序列的铁路客流量预测[D].华南理工大学,2017.