第十一届“泰迪杯”
数据挖掘挑战赛

优秀作品

作品名称：产品订单分析与需求预测

荣获奖项：特等奖并获泰迪杯

作品单位：武汉轻工大学

作品成员：郝青松 陆家驹 谭奥成

指导老师：卢冬晖

产品订单分析与需求预测

摘要

需求预测是企业供应链管理的基础，高效、准确的需求预测有助于采购计划和生产计划制定，减少业务波动影响。对于不规则时间序列数据，不同时间粒度趋势、波动、突变不同，因此以月、周、天为时间粒度，基于多个特征，采用不同建模方法预测订单需求量。

针对问题一，基于小样本、不规则时间序列数据进行预处理，BGCP 算法通过贝叶斯张量分解对不规则时间序列进行插补。剔除重复样本，使用隔离森林算法与四分位法确定异常样本，并分区域利用随机消除训练 BGCP 算法，得到各区域最优参，修正异常样本。

研究不同因素对产品订单需求量影响。观察核密度估计图、柱状图，发现当产品价格在 600～800 时，订单需求量最大，升降价格都会导致订单下降，说明该价位被普遍接受。根据方差分析，认为不同区域产品有显著差异。观察各区域箱型图，发现 101、102、105 区域订单需求量整体偏高，105 区域订单需求量稳定。观察箱型图、柱状图，发现线下订单需求量总和高于线上，比线上更稳定，但线上订单需求量中位数比线下高。观察各细类产品直方图、箱型图，发现 406、410、411 类产品需求量分布较为集中。402、403、404 类产品需求量中位数高。401、408、409 需求量分布均匀，订单需求量稳定。

研究时间趋势对产品订单需求量影响。根据时序图，可知产品 2 月、7 月总需求量最低，10 月总需求量达到峰值。月头需求量较少，月末需求量最大。观察柱状图、箱型图，发现节假日平均需求量比非节假日高，尤其是元旦假期间。观察箱型图，发现双十一促销活动比 618 产品平均需求量高，但 618 期间产品需求量更稳定。观察季节时序图发现每年秋季产品需求量最大，夏季最小，呈现明显季节趋势。各季节产品订单需求量中位数基本一致。

针对问题二，对历史数据进行特征工程，增加时间、节假日、促销、滞后等 15 个特征。考虑到各特征具有时序特性，且大多为非线性影响，使用 MMIFS 算法来量化各个特征与订单需求量间相关性，根据相关性强弱挑选出了 12 个特征进行训练。使用 CRU 模型、DeepAR 模型和 Prophet 模型来针对不同时间粒度需求量建模，以时间滚动交叉检验 MSE 作为模型评估标准。在以天为时间粒度预测上，CRU 模型表现最优，这得益于其时间增量和时间扭曲模块对不规则时间建模的优势以及 HuberLoss 损失函数。在测试集上，CRU 模型的 MSE 为 0.601。在以周为时间粒度上，DeepAR 模型对周期趋势较为敏感，预测表现最优，测试集上 MSE 为 0.475。在以月为时间粒度的小样本上 Prophet 模型使用时序分解模块，表现最优，测试集上 MSE 为 0.310。

本文针对数据特点，对比多种方法，选用表现最优模型，通过交叉检验评估验证了模型有效性，能够有效运用于其他不规则时序需求预测问题。

关键词： 不规则时间序列　BGCP 插补　CRU 模型　DeepAR 模型　Prophet 模型

目录

一、 引言

1.1 背景

近年来，企业面临着日益不稳定和复杂外部环境，导致供应链管理面临着诸多挑战。需求预测作为供应链管理的基础，对企业至关重要。需求预测有利于公司管理层制定销售及运营计划、目标，以及资金预算。此外，需求预测还有助于采购计划和生产计划制定，从而减少业务波动影响。

如果需求预测不准确，将导致公司管理层对市场预测不足，产生库存和资金积压或不足等问题，增加企业库存成本。由于需求预测受多种因素影响，预测准确率普遍较低，因此，需要通过优秀的算法来提高需求预测准确性，从而更好地管理企业供应链。

1.2 问题

(1) 对产品销售数据进行深入分析，包括但不限于：

- 产品价格对需求量影响
- 产品所在区域对需求量影响，以及不同区域产品需求量特性
- 不同销售方式产品需求量特性
- 不同品类间产品需求量异同点
- 节假日对产品需求量影响
- 促销对产品需求量影响
- 季节因素对产品需求量影响

(2) 建立数学模型，分别按天、周、月时间粒度，预测部分产品未来 3 个月需求量，并分析不同预测粒度对预测精度有何影响。

二、 问题分析

问题一分析：对数据进行预处理，剔除重复样本，使用隔离森林算法与四分位法确定异常样本，并使用 BGCP 算法修正异常样本。分别对价格、区域、销售方式、品类、节假日、促销、季节等因素进行分析。

问题二分析：首先根据数据构造特征工程，使用 MMIFS 算法对特性进行选择。然后分别使用 CRU 模型、DeepAR 模型和 Prophet 模型对天、周、月时间粒度产品订单需求量进行预测，根据时间滚动交叉检验 MSE 选择最优模型。最后对各时间粒度模型预测效果进行分析。

三、 数据预处理

3.1 缺失与重复样本处理

将产品销售数据导入 python 中，检查数据格式，将“订单日期”列转换为 datetime 格式。对数据进行缺失和重复检查，未发现缺失数据，但是有 312 个重复样本。保留首次出现的样本，并将其余重复样本剔除。

3.2 异常样本分析

绘制各产品价格与需求量箱线图，发现有部分产品存在明显异常，比如 103-21801-301-405 产品，如表3-1。

表 **3-1**　异常样本示例

订单日期	销售区域编码	产品编码	产品价格
2018-06-16	103	21801	852
2018-09-30	103	21801	260014
2018-10-01	103	21801	260006
2018-10-22	103	21801	501.7

因为产品价格与需求量可能出现粗大偏差，且各产品历史销售样本较少，定义异常点识别规则：

$$\begin{cases} c_i > (c_{q3} + \alpha * c_{iqr}) \\ c_i < (c_{q1} - \alpha * c_{iqr}) \end{cases}$$

其中，c_i 表示 c 产品 i 时刻价格；c_{q3} 表示 c 产品价格第 3 个四分位数；c_{q1} 表示 c 产品价格第 1 个四分位数；c_{iqr} 表示 c 产品价格四分位距；α 表示异常容忍阈值。

使用四分位数可以尽量避免因粗大误差导致各统计量不准确的情况。考虑到各产品可能存在销售旺季、淡季以及促销活动等导致产品价格或需求量出现波动，将 α 定为 2。

隔离森林[11](Isolation Forest) 是一种基于树结构的异常检测算法。该算法可以在线性复杂度下构建树结构，快速定位异常样本。同时隔离森林还能嵌入时序信息，适用于时间序列异常检测。将四分位法与隔离森林交集认定为异常样本。部分产品价格异常样本数量如表3-2。

表 **3-2**　部分产品价格异常样本数

产品编号	异常样本数量
101-20657-303-410	1
102-21052-303-401	1
103-20338-302-412	2
104-20340-308-404	2
105-21122-306-407	1

同样的，对产品需求量进行检查，部分产品需求异常样本数量如表3-3。

表 **3-3**　部分产品需求异常样本数

产品编号	异常样本数量
101-20657-303-410	3
102-20354-303-401	2
103-20481-306-407	2
104-20606-308-404	1
105-20205-306-407	1

3.3 BGCP 修正异常样本

BGCP[6](Bayesian Gaussian Process for Time Series Completion, BGCP) 算法是一种通过贝叶斯张量分解对不规则时间序列插补的机器学习算法，其优势在于：采用贝叶斯方法，能够更好地利用先验信息，从数据中学习时空模式；采用自适应混合高斯分布对数据噪声建模；蒙特卡罗马尔科夫链 (Markov Chain Monte Carlo, MCMC) 采样方法自适应学习不同分解维度和分解系数。

算法重要参数说明：

(1) 分解秩 (rank)：矩阵分解秩，用来拟合原始张量隐因子数量。秩越高，模型可以拟合更多细节和噪声，可能出现过拟合现象。秩越低，模型更倾向于捕捉数据潜在结构和特征，可能导致模型欠拟合。

(2) 燃烧期迭代次数 (burn_iter)：在 burn 阶段，模型不会记录样本，而是用于使模型适应数据分布。

(3) 吉布斯采样迭代次数 (gibbs_iter)：在 gibbs 采样阶段，模型从后验分布中抽取样本。

观察数据发现，各区域内各产品价格与需求量有一定趋同性，说明同区域内产品可以共享 BGCP 算法参数。将同区域内不含异常样本产品数据做随机消除，然后使用 BGCP 算法进行插补，以 RMSE 为评估函数。则区域内插补总损失为：

$$q_i = \sum_{j=1}^{n} c_j^i$$

其中，q_i 为区域 i 损失，c_j^i 为 i 区域内 j 产品插补损失。使用贝叶斯优化 (随机迭代：50 次，贝叶斯优化：100 次)，同时调整 rank、burn_iter、gibbts_iter 参数，最小化各区域插补损失。各区域参数及最小损失如表3-4。

表 **3-4**　各区域参数

销售区域编码	rank	burn_iter	gibbts_iter	RMSE
101	16	180	828	0.442
102	19	195	800	0.489
103	14	167	793	0.436
104	10	132	720	0.370
105	20	200	1000	0.674

对各区域异常样本进行修正，部分修正效果如图3-1。

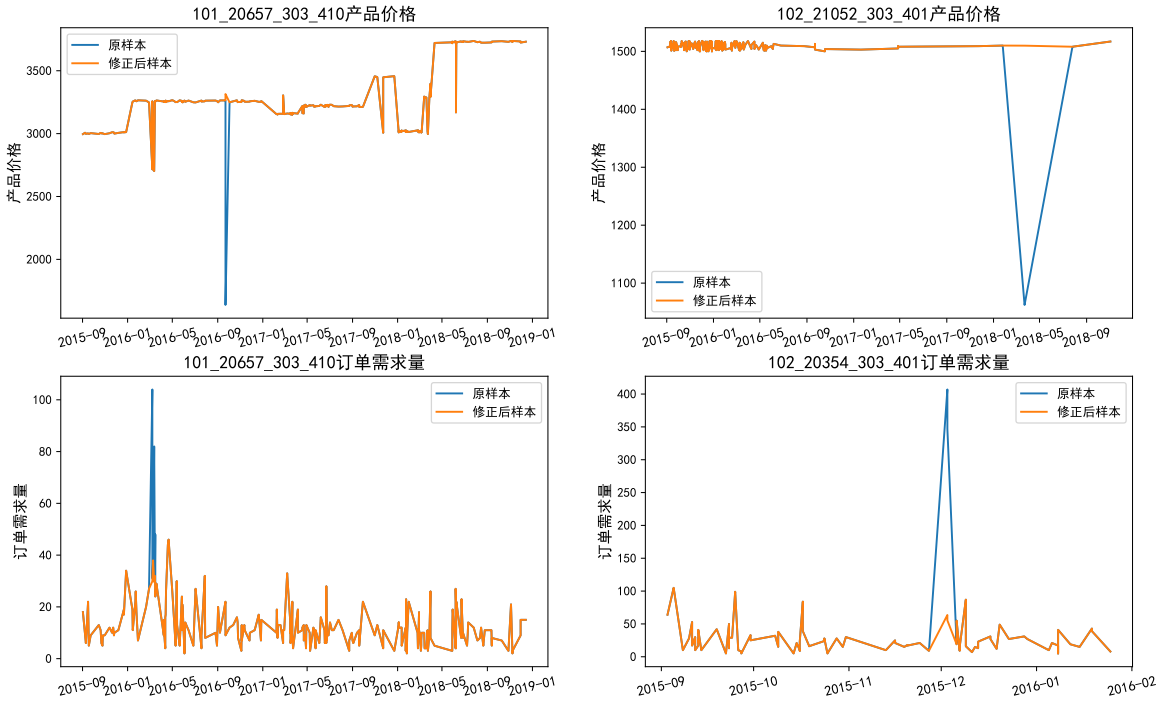

图 **3-1**　**BGCP** 修正部分产品效果图

四、 销售数据分析

为观察该企业产品价格与订单需求量分布情况，绘制核密度曲线图。因产品价格 >4000, 订单需求量 >400 样本仅占总样本 1%，并且会造成长尾效应。为方便观察，仅取产品价格 <4000，订单需求量 <400 样本。如图4-1

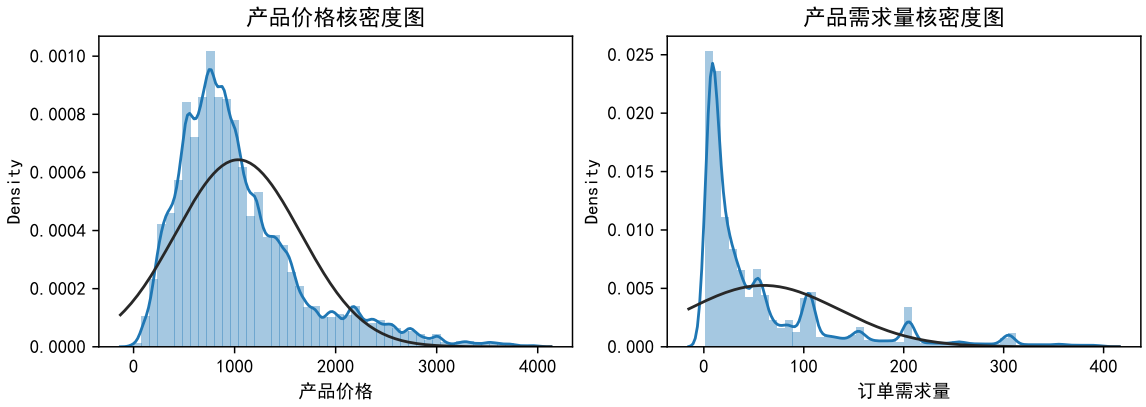

图 **4-1**　产品价格与需求量核密度估计图

观察可知，产品价格主要集中在 0~2000 价位内，而产品需求量主要集中在 0~100 范围内。

针对不同因素对产品订单需求量影响，绘制柱状图，观察各因素聚合特征。绘制圆

环图，观察各因素占比情况。绘制箱型图，观察各因素数值分布情况。绘制时序图，观察在不同时间粒度下各因素订单需求量趋势。

4.1 产品价格与需求量分析

按产品价格分组对订单需求量求和，将产品价格以 200 为间隔切分，观察各价格区间产品订单需求量，如图4-2

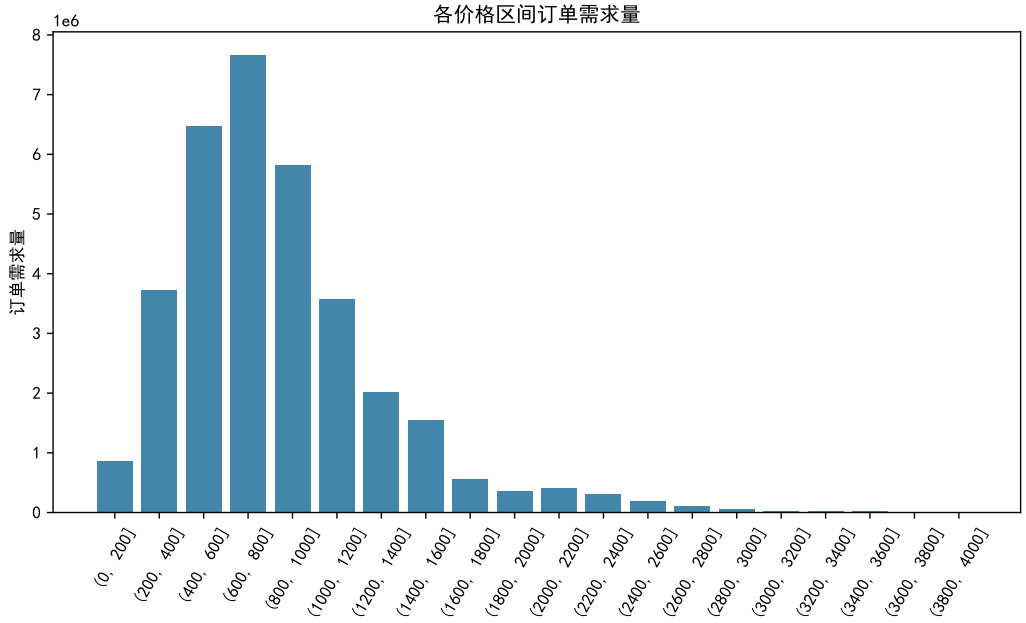

图 **4-2**　各区间订单需求柱状图

可以看到，当产品价格在 (600,800] 间需求量最大，说明该价位普遍被消费者认可。(0,200] 和 (1600,1800] 又因为太过便宜或者太贵导致订单量偏低。(400,600] 相较于 (600,800] 差距比 (800,1000] 相较于 (600,800] 差距小，说明消费者更偏向于小幅度降价。(200,400] 和 (1000,1200] 订单需求量基本相等，说明价格降低与升高都有阈值，超过该值消费者可能会更谨慎购买。

为进一步观察产品价格与需求量是否存在线性相关性，对两变量进行相关性分析。因产品价格与需求量均不满足正态分布，使用 Spearman 相关系数。两变量相关系数为-0.009，说明产品价格与需求量并无线性相关性。

为观察不同产品价格与需求量间关系，新增销售额列 (销售额 = 产品价格 × 订单需求量)。按产品编码分组对销售额进行求和，筛选出销售额排名前 4 产品进行分析，如图4-3。

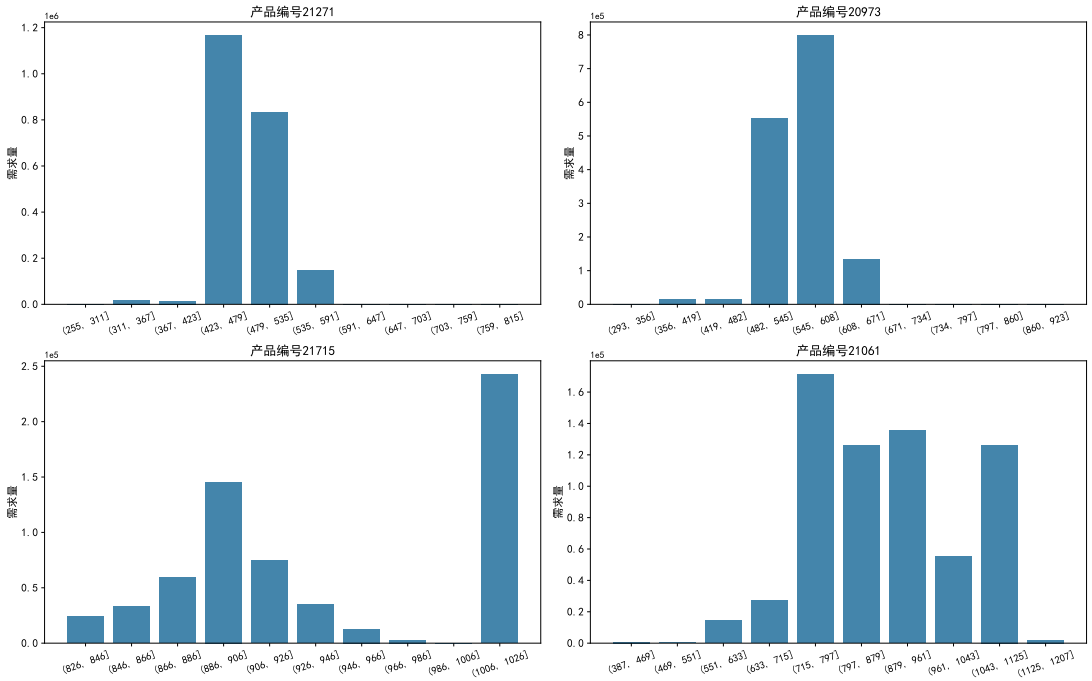

图 **4-3**　部分产品价格与需求量柱状图

可以看到，产品 21271 和产品 20973 具有相同需求分布特性，即需求量大部分分布在某 2 个区间。说明这两种产品市场需求较为明确，可能是某类长期产品。产品 21751 产品在价格为 1006~1026 时，需求量最高，而在价格为 826~986 之间，需求量呈正态分布。说明该产品在某一领域大量需求，而在零售市场则随价格变化产品需求随之变化。产品 21061 价格分布较为均匀，且各价格区间需求较为平衡。

4.2 产品区域与需求量分析

为探究不同区域产品需求量是否存在显著差异，以区域为分组变量，产品需求量为因变量进行单因素方差分析。

首先对数据进行正态性检验，由于样本数量较大，选择 Anderson-Darling 检验。根据样本与理论正态分布间距离，认为产品基本呈正态分布。Anderson-Darling 检验与方差分析结果，如表4-1

表 **4-1**　**Anderson-Darling** 检验

AD 检验	sum_sq	df	F	P 值
0.001	2.175×10^7	4	233.5605	0.000

可以看到，方差分析 P 值为 0，拒绝原假设，认为不同区域间产品需求量有显著差

异。

为探究差异性，按区域分组对产品需求量求和，绘制各区域订单需求量柱状图与圆环图，如图4-4。

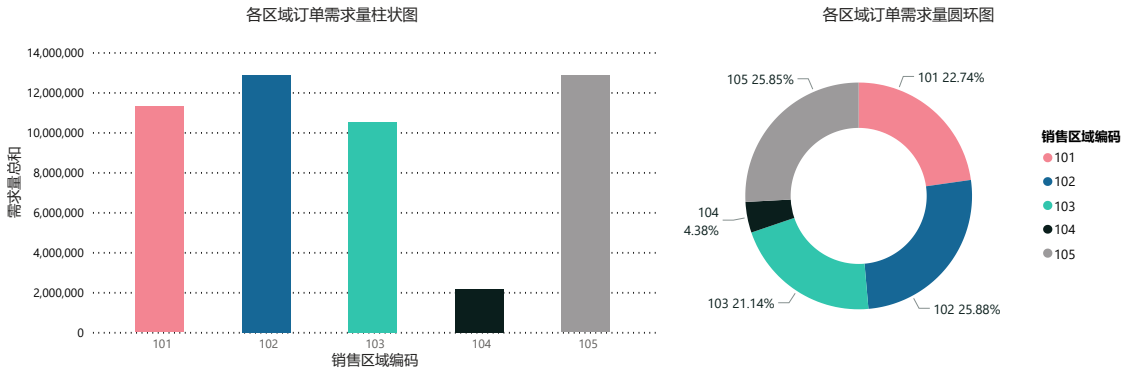

图 **4-4**　各区域订单需求量柱状、圆环图

观察可知，104 区域的订单需求量总和最少，且需求订单需求量占比最少，仅 4.38%，而其他地区订单需求量相对均衡，这可能与当地实体经销商数量、线上购物普及程度、地区经济因素、快递站点分布、当地居民平均年龄有关。

101、102、103、105 区域需求量占总需求量的 95.62%，说明这些地区为主要销售地区，可以考虑在这些地区投放更多产品，来增加收益。

为探究不同区域产品需求量均值、中位数等统计指标，绘制各区域订单需求量箱型图，如图4-5。

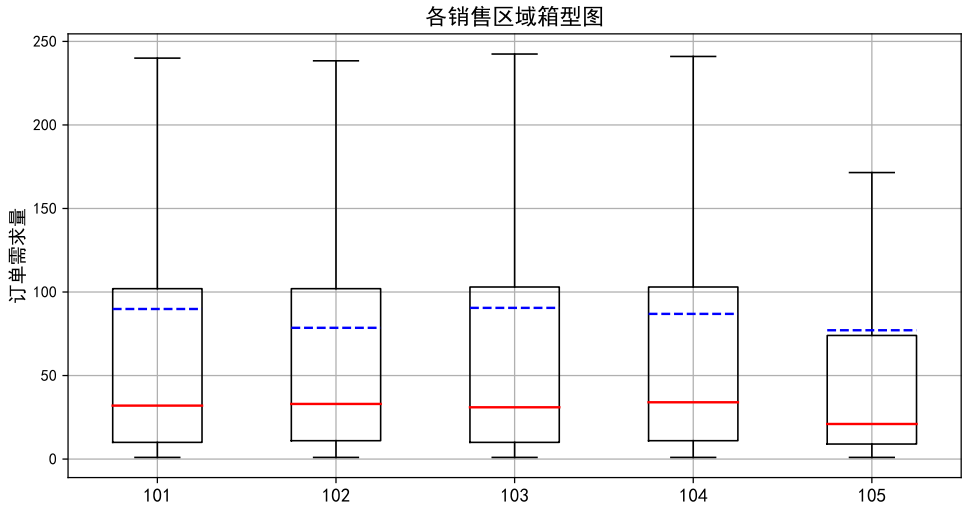

图 **4-5**　各区域订单需求量箱型图

可以看到，各区域订单需求量中位数基本相同，仅 105 区域中位数略微偏低，说明其他区域订单需求量在高端表现相较于 105 区域更好。101、103、104 区域平均值相较

于 102、105 区域高一些。101～104 区域产品需求量上边界比 105 区域高，说明 101～104 区域产品需求量存在较多波动和极端值，105 区域更稳定一些。

按区域和产品编码分组对产品需求量求和，分别筛选出各区域内产品需求量排名前 10 产品，绘制柱状图，如图4-6

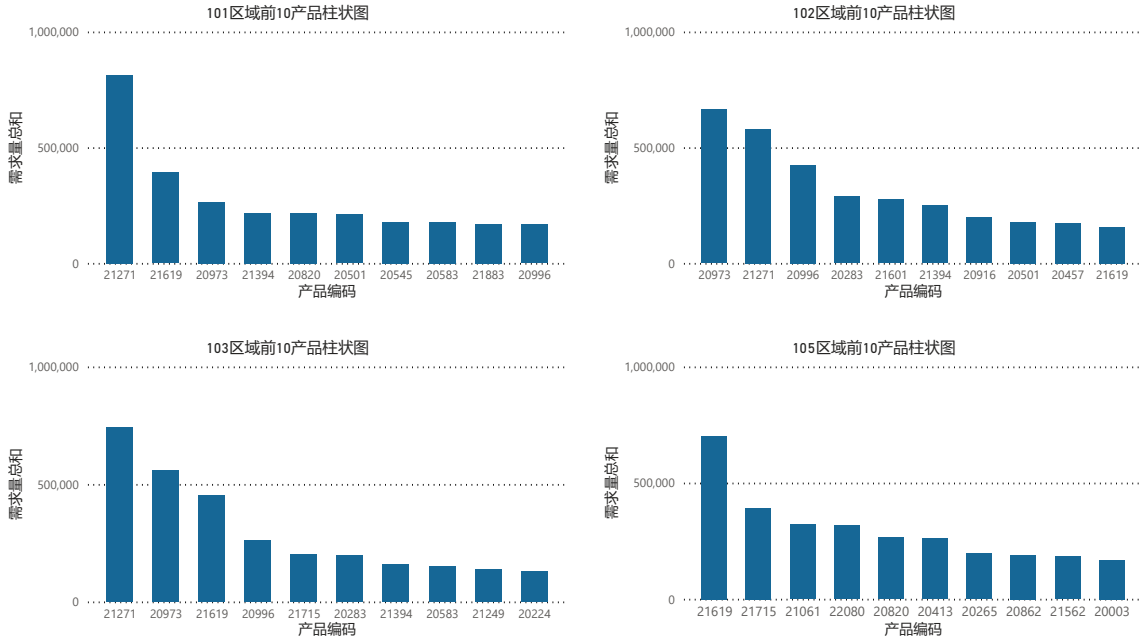

图 **4-6**　部分区域产品需求量柱状图

观察可知，101 区域与 103 区域需求量总和最多的都为 21271 产品，而 102 区域中需求量总和的第二也为 21271 产品，且其需求总量都超过了 500000，说明 21271 产品的需求量在这三个区域有较大的需求，可以增加 21271 产品在这三区域的供应，来增加收益。

20973 产品在 102 区需求总量最大，同时该产品在 101 区域与 103 区域有着一定的需求量，说明 20973 产品在这三个区域有一定的需求量，可以在 102 区域增加较多的 20973 产品的供应，而在 101 区域与 103 区域则适当增加 20973 产品的供应，来增加销售量与收益。

21619 产品在 105 区域、101 区域、103 区域都有一定的需求量，其中该产品在 105 区域的需求量总和的占比最高，说明该产品在 105 区域很受欢迎。而在 102 区域，21619 产品在 102 区域的占比较少。

各产品在各区域内占比不同，可以体现出各地方对于商品种类的需求有差异。这可能与各区域所属地理环境、人口年龄分布以及当地习俗有关。

分别以年、季度、月、周为时间粒度对订单需求量进行重采样，绘制各区域需求时序图，如图4-7

图 **4-7**　各区域不同时间粒度时序图 **(**从上到下依次为 **101**～**105)**

可以看到，观察以年为粒度的时序图，可以发现在 2017 年 12 月后，部分地区都呈现订单需求量下降的趋势。

观察以季为粒度的时序图，发现 101、102 以及 103 地区的订单需求量呈一致的趋势曲线，说明三地的市场环境、消费人群、地理因素等有相同的地方。

观察以月为粒度的时序图，可以看到，在 2015～2017 年 12 月前，产品的需求量总体是呈上升趋势，而在 2017 年 12 月需求量到达峰值，并快速下跌。

观察以日为粒度的时序图，发现 101 区域产品的需求量交其他地区波动大，市场占比不稳定。造成此类问题的原因可能是跟当地的市场竞争有关，出现了相同功用的产品来抢占市场。

4.3 销售方式与需求量分析

为探究不同销售方式产品需求量是否存在显著差异，将两者进行独立样本 t 检验，根据 t 值、p 值认为，不同销售方式产品需求量存在显著差异。

按销售渠道分组对产品需求量求和，绘制不同销售渠道订单需求量柱状图与圆环图，如图4-8。

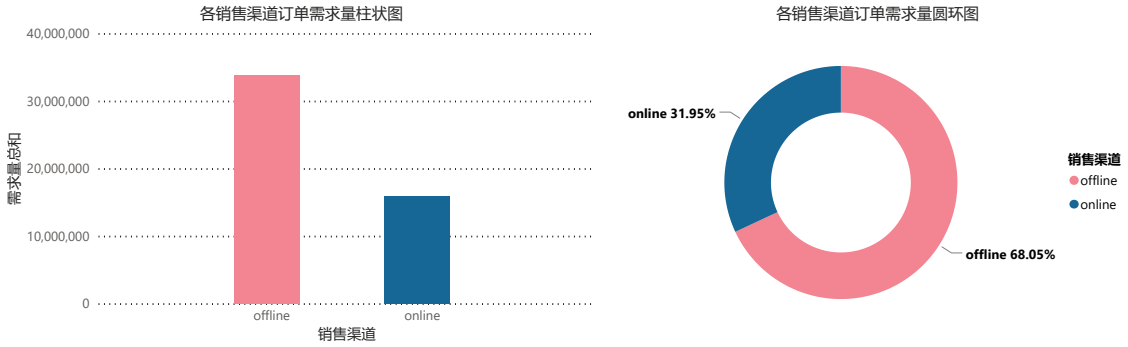

图 **4-8**　各销售渠道订单需求量柱状、圆环图

可以看到，offline 的订单需求量占比高于 online，达 68.05%，而 online 的订单需求量占总订单需求量的 31.95%，这与互联网的快速发展、网络购物的兴起、物流行业站点的快速扩张、社会节奏的加快、“互联网 +”政策的支持有关。

为观察不同销售方式产品需求量均值、中位数等统计指标，绘制各销售方式箱型图。如图4-9。

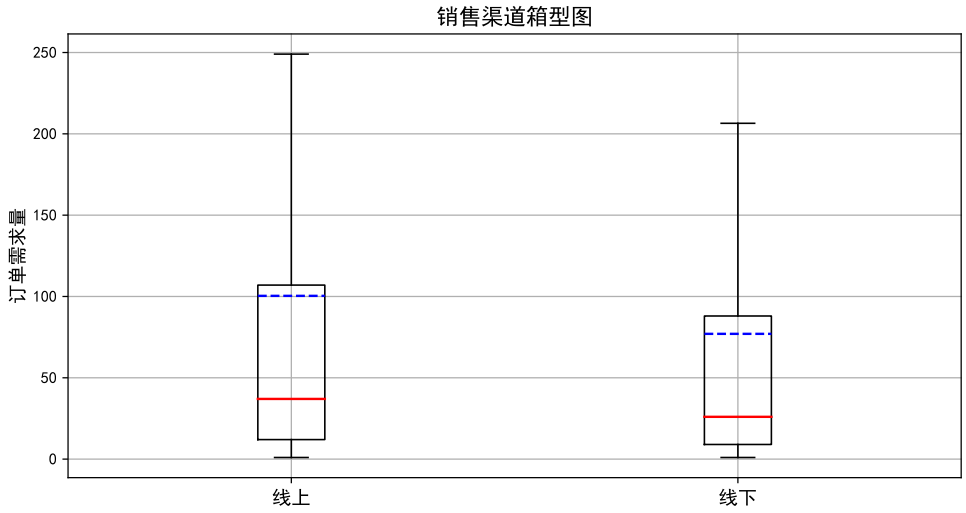

图 **4-9**　各销售方式箱型图

观察可知，线上中位数和平均值都比比线下高，说明线上订单需求量整体上比线下高，数据分布形态较为均匀，可能是由于互联网购物风靡造成。线下销售上边界比线上低，说明线下销售订单需求量更稳定些。

分别以年、季度、月、周为时间粒度对订单需求量重采样，绘制不同销售方式产品需求时序图，如图4-10。

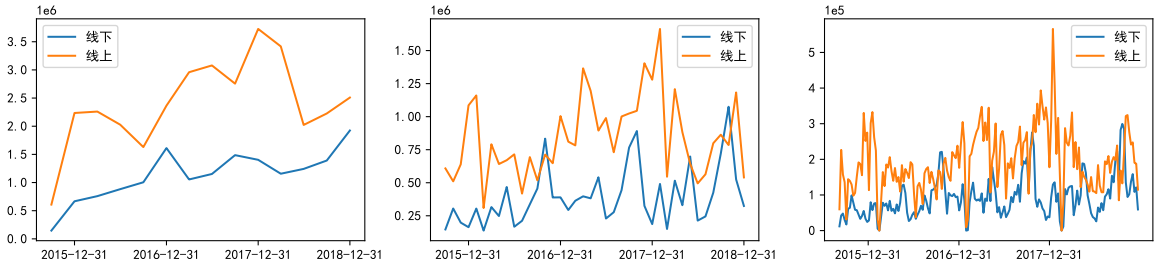

图 **4-10**　各销售方式产品需求季、月、日时序图

观察发现，线上与线下的产品需求量呈上升趋势，且线上的产品需求量高于线下。观察日时序图发现，线上的产品需求量波动范围较大，不稳定；线下的产品需求量波动范围小，较为稳定。

造成线上的产品需求量的波动大的原因：

① 与客户心理预期不符　网上购物看不到实体，不能辨别，只能通过图片、用户评论、咨询客服等途径来了解产品，购买后可能与顾客实际的需求不符而退货。

② 产品质量问题　产品可能存在质量问题，由于是线上购物，不能检测产品质量，在顾客签收后检查产品发现存在质量问题，会进行退货处理，造成需求波动。

③ 产品的运输风险　由于线上购物采用物流方式进行产品的运输，由于一辆物流车内运输的货物种类繁多，存在物件之间的碰撞，造成产品的损坏，使顾客退货，降低产品的需求量。

4.4 各类产品需求量特性

分别筛选出各产品细类样本，为观察各细类产品需求量均值、中位数等统计指标，绘制各产品细类箱型图，如图4-11。

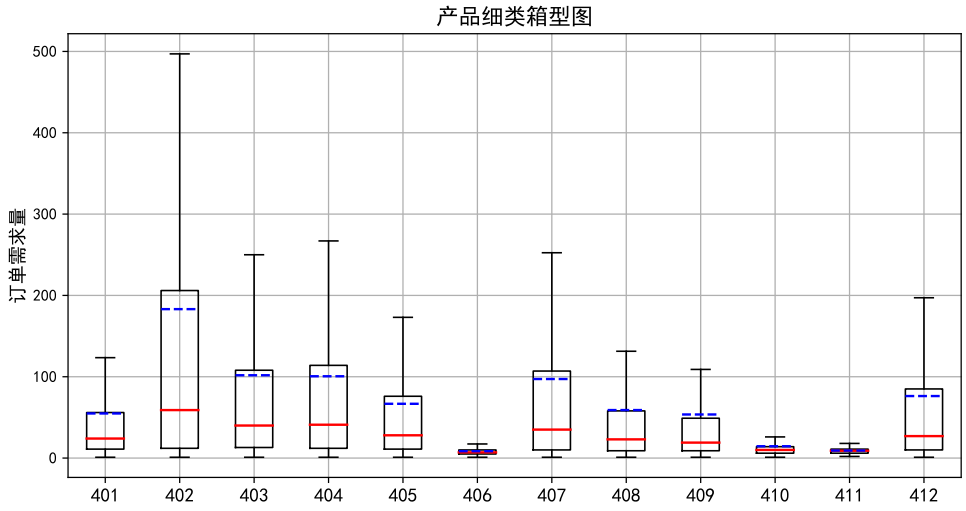

图 **4-11**　各产品细类箱型图

可以看到，406、410 和 411 产品订单需求量分布较为集中，各项指标基本相等。402、403、404 产品订单需求量中位数较高，说明这些企业为主要产品。401、408、409 产品分布均匀，订单需求量稳定。

为探究各细类产品需求量分布情况，绘制各细类产品订单需求量直方图，如图4-12。

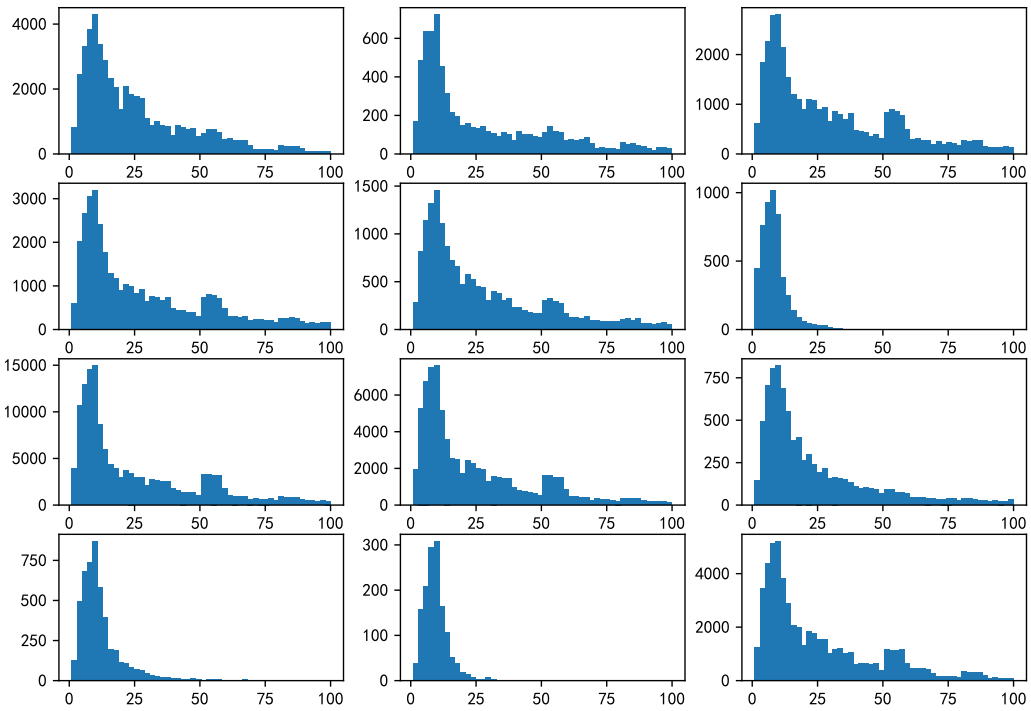

图 **4-12**　各细类产品订单需求量直方图 **(**从左到右依次为 **401**～**412)**

可以看到，各类产品需求量分布可以划分为三类：

① 以 403 产品的需求量为例，403 产品需求量存在两个峰值，在需求量为 10 附近存在最大峰值，且最高频率达 1000 次以上，在需求量为 50 附近存在峰值。符合此需求量分布的产品有：403，404，405，407，408，412。

② 以 401 产品的需求量为例，在需求量 0～10 之间快速增加，在需求量为 10 达到峰值，在需求量 10～100 之间逐步减少。符合此需求量分布的产品有：401，402，409。

③ 以 406 产品的需求量为例，产品的需求分布主要集中在 0～50 之间，在 0～10 之间频率快速上升，在 10～25 之间快速下降，且最高频率在 1000 次一下。符合此需求量分布的产品有：406，410，411。

以月为时间粒度对订单需求量重采样，绘制不同产品细类需求量时序图，如图4-13。

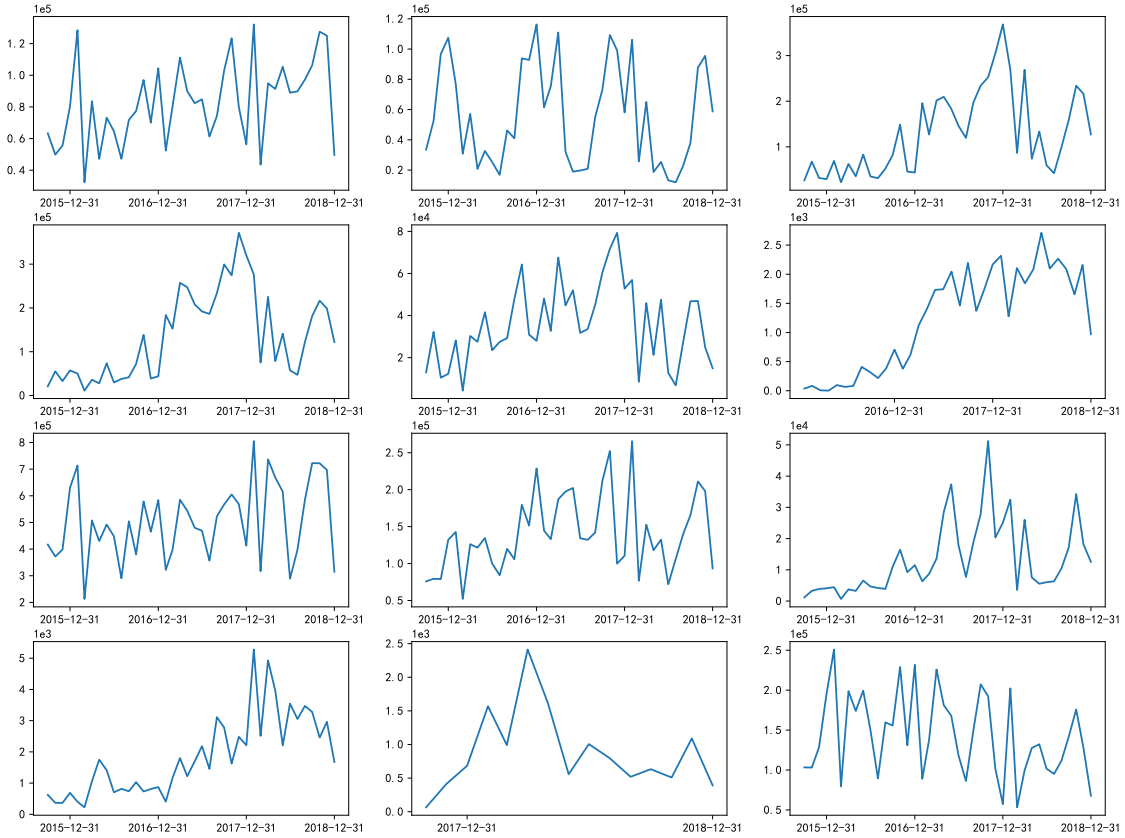

图 **4-13**　各产品细类月需求量时序图 **(**从左到右依次为 **401**～**412)**

观察可知，各产品在月时间粒度下表现出较为明显周期性，比如 402、404 产品，在每年冬季有明显上升趋势，来年春季出现下降。但也存在各月需求量波动较大无明显趋势产品，如 401、412 产品。403、404、406 产品在 2016 年前需求量较低，在 2016 年后，需求量增长较为明显，可能是由于产品经过市场检验后逐渐完善的结果。

4.5 各时段产品需求量特性

按月分组对订单需求量求和、求平均，分别绘制月产品需求总量折线图、月产品需求平均折线图，如图4-14。

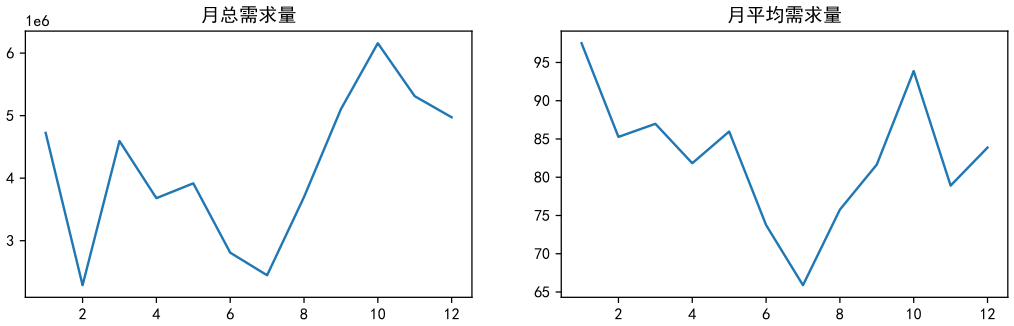

图 **4-14**　月产品需求总量与平均折线图

观察月总需求量图，可以发现在 2 月和 7 月的月总需求量最低，10 月的月总需求量达到峰值。观察月平均需求量图，发现在 2 月时，月平均需求量达到并不低，而其月总需求量较低，这可能与 2 月份的天数较其他月份少造成的。而 7 月的月平均需求量最低，且 7 月的月总需求量也处于较低水平，而 7 月后的产品需求量逐渐增多。

为探究各时间段（月头、月中、月尾）产品需求量特性。按订单日期分组对产品需求量求和，将订单日期分为年、月、日 3 列。添加时段列，根据经验将各月 1~10 日定为月头，10~20 日定为月中，20~30 日定为月末。按时段分组，对产品需求量求和，绘制柱状图，如图4-15。

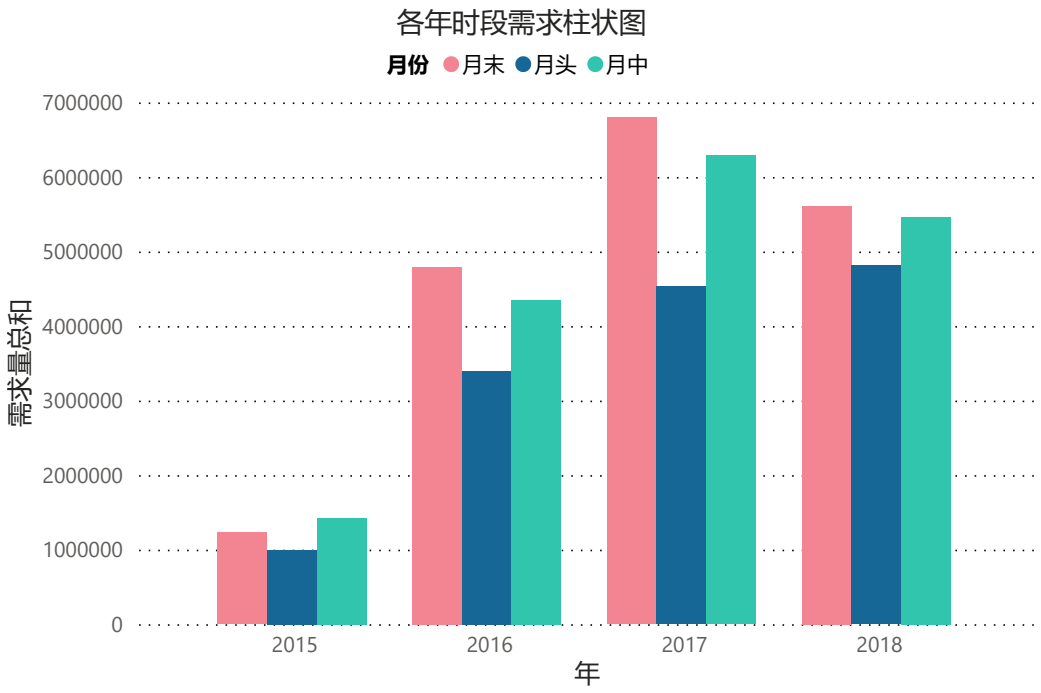

图 **4-15**　各时间段产品需求柱状图

观察可知，每月月头的需求量最少，月末的需求量最大，在 2018 年总体需求量较去年减少。每月的月中是发放工资的时间，在此期间用户可支配资金增加，可以购买需要的产品，而月头由于可支配资金的减少，对于产品的购买欲下降，使月头的需求量减少。

按年和时段分组对产品需求量求和，观察每年各时间段产品需求量特性，绘制柱状图，如图4-16。

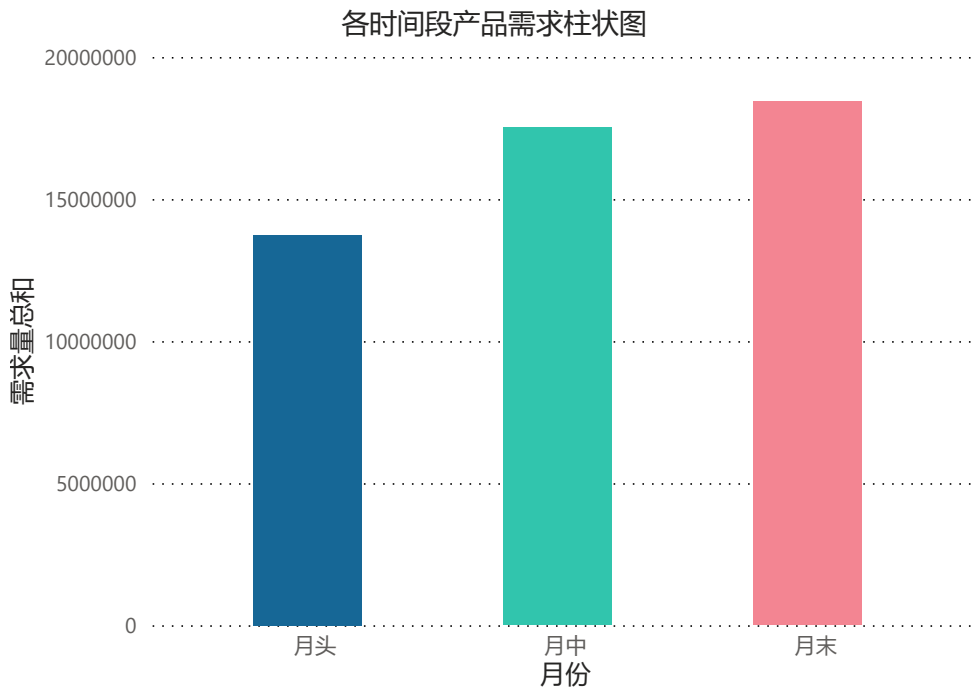

图 **4-16**　各年时段需求柱状图

可以看到，每月的月头产品需求最少，每月月末产品需求量最多。结合图4-15的分析，对产品的供应策略进行优化：根据每月不同时段在调整产品的供应来适应每月需求量的变化，在月中后增加产品供应，在月头减少产品供应。

4.6 节假日产品需求量分析

根据国务院发布的节假日程文件，将日期分为节假日与非节假日。按是否为节假日分组对订单需求量求平均，绘制节假日订单需求柱状图，如图4-17。

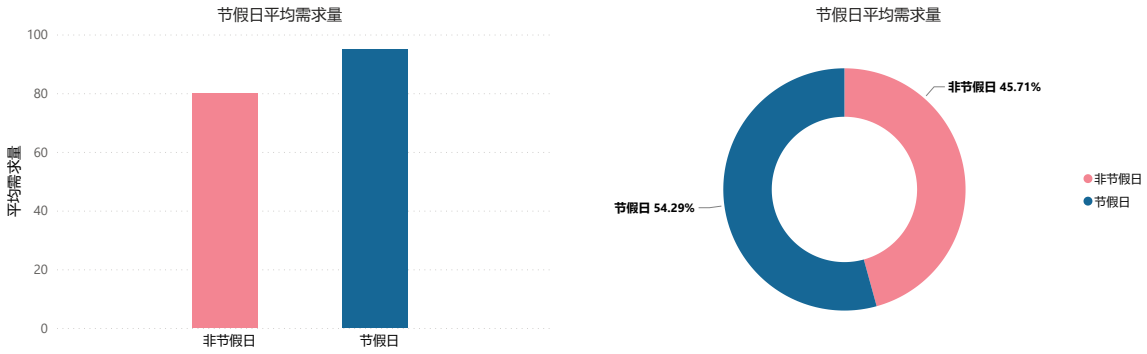

图 **4-17**　节假日产品平均需求量柱状图

观察可知，节假日的产品需求量高于非节假日的产品需求量。这体现了在节假日时，由于很多消费者处于休假状态，有较多时间来挑选与购买产品，对此可以适当增加节假

日的产品的供应，并可以对产品做一些优惠活动，以此来刺激消费者的消费心理，增加销售量提高收益。

为探究不同节日对产品需求量影响，添加节日列。按节日列分组，对订单需求量求平均，绘制不同节日订单需求柱状图，如图4-18。

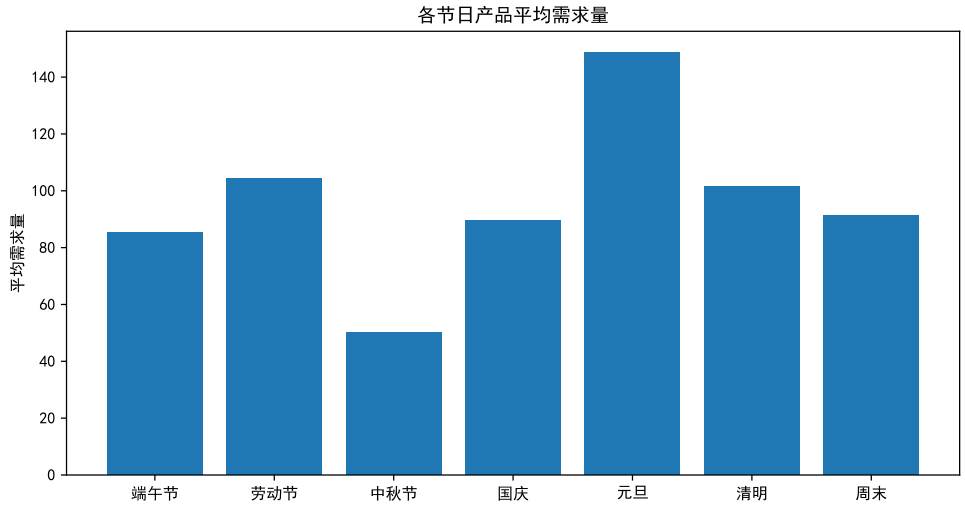

图 **4-18**　各节日产品平均需求量柱状图

可以看到，元旦节平均需求量最大，中秋节平均需求量最小，其他节日的平均需求量较为平均。

4.7 促销与产品需求量分析

根据促销活动 (618，双十一) 历年开始与结束日期，将活动前后 1 星期作为促销区间，添加是否为促销期列。按是否为促销期列分组，对订单需求量求平均，绘制各促销活动订单需求柱状图，如图4-19。

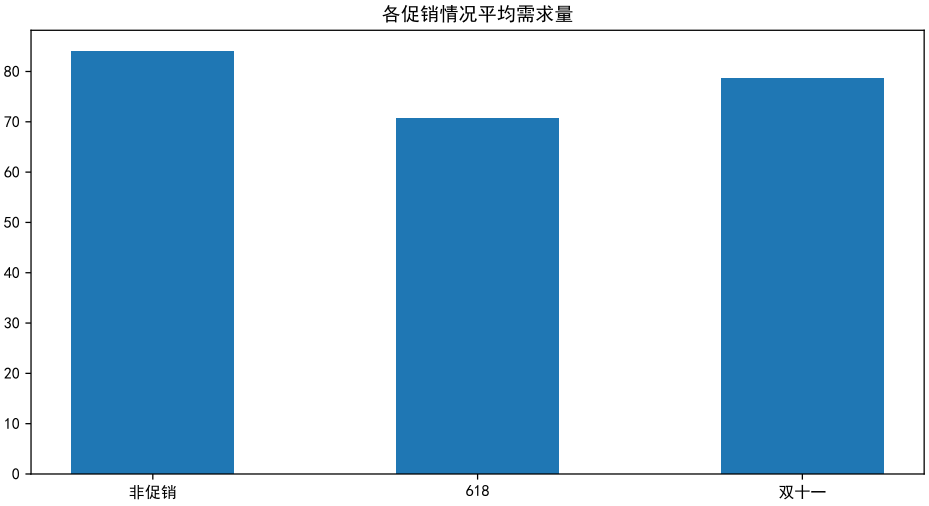

图 **4-19**　各促销情况平均需求量柱状图

观察发现，非促销时期的平均需求量高于 618 以及双 11 促销时的需求量，这说明促销活动对于该产品的销售影响较少，该产品受价格影响的程度较低，价格弹性小。

为观察各促销活动产品需求量均值、中位数等统计指标，绘制各促销情况箱型图。如图4-20。

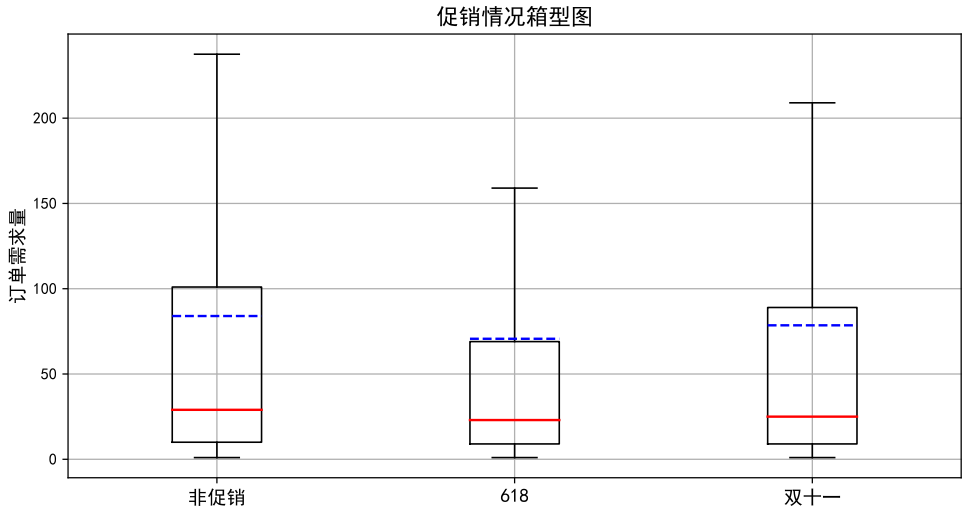

图 **4-20**　促销情况箱型图

可以看到，双十一中位数与 618 基本相同，但 618 平均值低于双十一，说明促销活动订单需求量分布基本相同，但双十一在平均水平上高于 618，618 活动订单量更稳定。说明双十一活动在整体上更具有市场竞争力。

为探究每年促销活动需求量平均水平，绘制各年促销活动产品需求量柱状图。如图4-21。

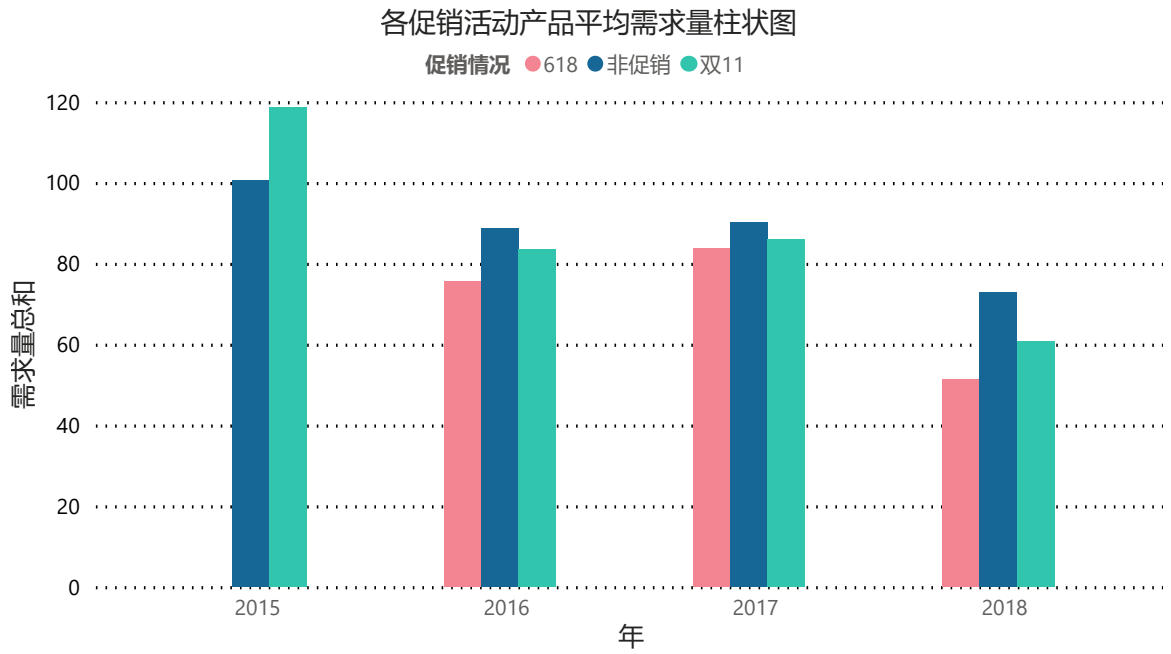

图 **4-21**　各促销活动产品平均需求量柱状图

由图可知，除了 2015 年双 11 促销活动的需求量高于非促销的需求量，2016~2018 年 618 和双 11 促销活动的需求量都低于非促销的需求量。结合图4-19的分析，该产品需求稳定，受价格波动、促销手段的影响小。应考虑平时产品的生产质量，以及相应的销售服务，提升用户体验，增强产品口碑，打造产品的优质形象，以此稳定消费者，巩固市场份额。

2015~2018 年产品平均需求量整体呈下降趋势，说明消费者对于产品的购买量减少，消费热情降低，产品的市场需求占比下降。造成的原因可能是产品的更新迭代，出现了竞争对手，利用新的技术来占领市场，或是出现了同等效果的低价等效产品，以此来打价格战，占领市场。对此应考虑产品的更新换代，用产品的技术升级，来刺激消费者的购买欲望，提升市场占比，增加收益。

4.8 各季节产品需求量分析

为探究季节因素对产品需求量影响，添加季节列。根据经验，将 1~2 月定为冬季，各年 3~5 月定为春季，6~8 月定为夏季，9~11 月定为秋季，12 月定为冬季。按年和季节分组对产品需求量求和，观察每年各季节产品需求量特性，绘制柱状图，如图4-22。

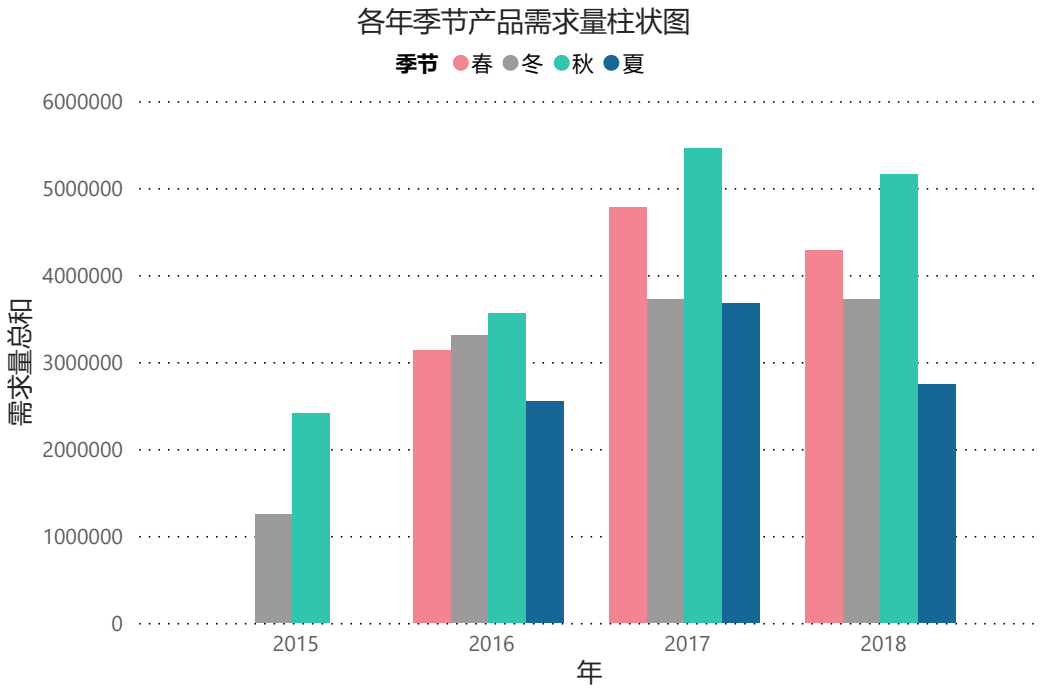

图 **4-22**　各年季节产品需求量柱状图

可以看到，由于 2015 年数据量不足，只考虑 2016~2018 年的产品需求量的情况。观察到每年秋季的产品需求量最大，夏季的产品需求量最小，呈现一种季节波动。根据不同季节的产品需求量来调整产品的供应，在秋天时增加产品供应，在夏天时减少产品供应，以此实现供应的合理分配，减少库存，降低投资成本增加收益。

按季节分组，对产品需求量求和，绘制柱状图，如图4-23。

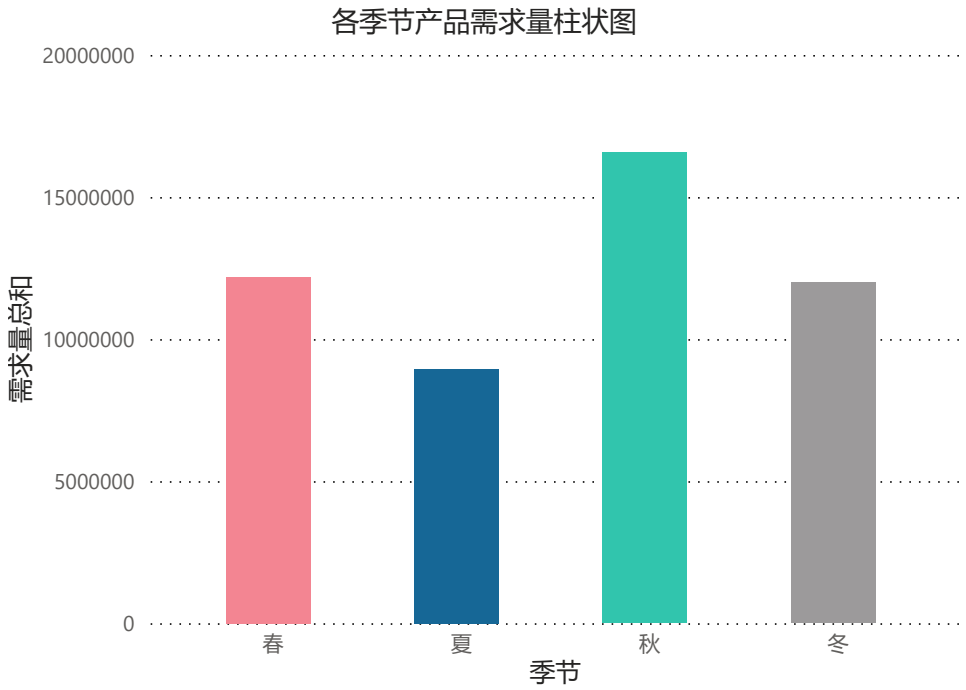

图 **4-23**　各季节产品需求量柱状图

为观察各季节产品需求量均值、中位数等统计指标，绘制各季节产品需求量箱型图。如图4-24。

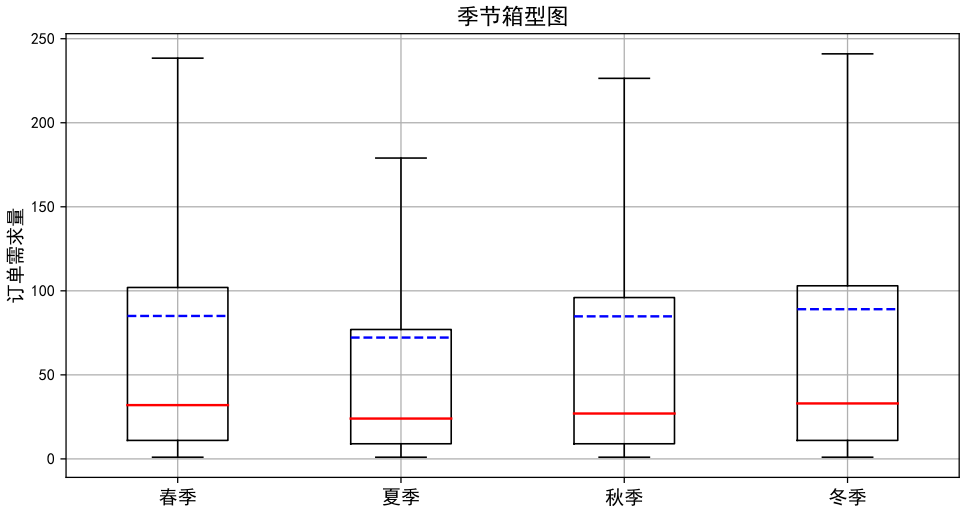

图 **4-24**　各季节产品需求量箱型图

观察发现，各季节产品订单需求量中位数基本一致，春季和冬季略高一点。说明整体上春、冬季产品需求量数值更高一些。春、秋、冬季均值均高于夏季，说明夏季是产品销售淡季。秋季均值最高，说明整体上秋季订单需求量最高，秋季是产品销售旺季。进一步说明产品订单需求量可能与温度存在密切关系。

以季度为时间粒度对订单需求量重采样，绘制各季度产品需求时序图，如图4-25。

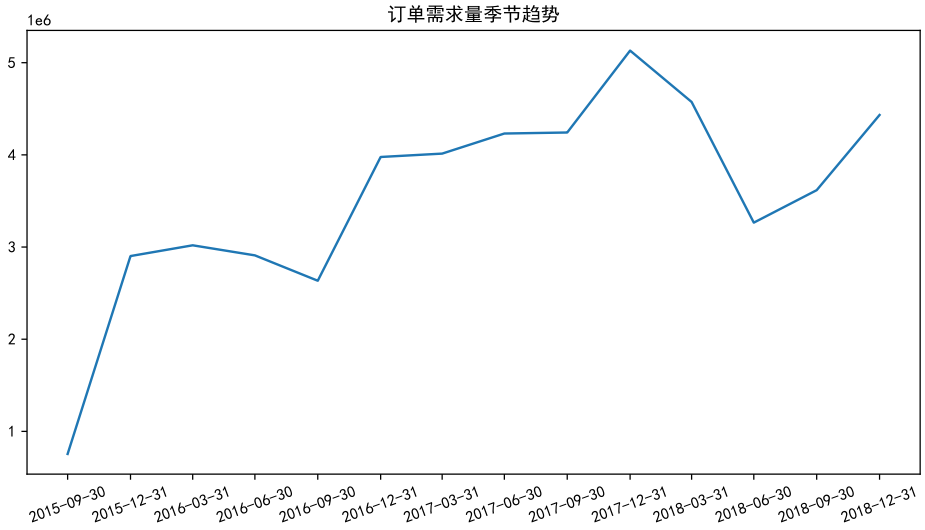

图 **4-25**　订单需求量季节时序图

由图可知，2015 年 9 月 ~2018 年 12 月订单需求量整体呈上升趋势，而在每年的 6 月 ~9 月期间，订单需求量会呈现一种下降或增长减缓的趋势。可以在每年 9 月至次年 6 月期间增加产品的生产，在 6 月至 9 月减少产品的生产，适应市场需求，合理调整产能，来增加收益。

通过观察年份，每年订单需求量在每年 12 月达到当年峰值，说明产品在 12 月需求量最大，可以增加 12 月产品的生产，提升产品的库存，应对 12 月需求的快速增长，实现收益最大化。

五、 需求预测

构建特征工程，使用 MMIFS 方法选取有效特征。针对不同时间粒度做相应重采样，分别训练 CRU、DeepAR、Prophet 模型，使用交叉检验 MSE 评估模型，选取最优模型。分析不同时间粒度对模型精度影响。

5.1 特征工程

新增产品标识列，将销售区域编码、产品编码、产品大类编码，产品细类编码整合到一起。将产品销售数据按订单日期和产品标识分组对订单需求量求和，然后与产品预测数据中产品标识进行匹配，做完全内部连接，共匹配到 2187 件产品。将新表按产品标识分组计数，并按降序排列，部分产品日订单记录如表5-1。

表 **5-1**　部分产品日订单记录条数

产品标识	记录条数
105-21619-306-402	987
102-20973-306-407	961
105-21026-306-407	935
105-21881-306-407	933
102-21271-306-407	913

可以看到最高记录条数仅为 987 条，从 2015 年 9 月 1 日至 2018 年 12 月 20 日应有 1207 条，数据缺失量占完整数据的 18.22%。

将日期拆分为年、月、日，并根据日期添加季度列 (春夏秋冬四季度分别使用独热编码)、周列 (从周一到周日，分别使用 0∼6 代替)。根据国家政府信息网发布 2015 ∼ 2018 年节假日安排文件添加节假日特征。添加 618 与双十一促销活动周期信息。

由于产品销售记录缺失严重，并且数据采集频率不一致，使用固定时间周期滑动窗口提取时间滞后特征。根据上述分析，订单需求量周趋势较为明显，分别提取滞后 1、2、3、5、7 天特征。为进一步探究各变量对订单需求量影响，本文采用 MMIFS[14](Multivariate Mutual Information-based Time Series Feature Selection,MMIFS) 方法，对特征进行选择。

5.2 MMIFS 特征选择

MMIFS 是一种基于多变量互信息的时间序列特征选择方法。具体地，该方法将多变量时间序列中每个变量视为一个节点，并将它们组合成一个无向图，然后，利用互信息度量变量间相关性，并将其用于计算节点间边权重。使用最小生成树算法来选择重要特征。

时间序列互信息是一种衡量两个时间序列间关联程度的方法。对于两个时间序列 X 和 Y，互信息 $I(X:Y)$ 计算公式为：

$$I(X:Y)=\sum^{x\in X}\sum^{y\in Y}p(x,y)log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

其中 $p(x,y)$ 表示 X 和 Y 同时出现的概率，$p(x)$ 和 $p(y)$ 分别表示 X 和 Y 出现的概率。

该算法主要优势在于能够同时考虑多个变量之间的相关性，并且能够在特征选择过程中捕捉自变量与因变量间非线性关系，适合具有任何分布的随机变量。针对产品需求预测采集频率不规律和突变性的特点，具有一定优势。MMIFS 影响因素相关系数如表5-2

表 **5-2　MMIFS** 相关系数

变量	相关系数	变量	相关系数
年份	0.21	是否为促销期	0.44
季度	0.32	促销类型	0.36
月份	0.22	滞后 1 天	0.95
周数	0.37	滞后 2 天	0.93
日期	0.07	滞后 3 天	0.90
是否为节假日	0.56	滞后 5 天	0.89
节假日类型	0.38	滞后 7 天	0.91
产品价格	0.43		

可以看到，滞后特征与因变量间相关性较强，其次为促销期、节假日、周与季度。日期与因变量间相关性较弱，为提高模型预测精度，避免引入噪声，将日期特征剔除。

最后将数据进行归一化处理，为保证日期特征取值相对均匀地分布在一定范围内，更好地与其他特征进行比较和组合，日期特征映射到 $[-0.5, 0.5]$ 区间内。其他特征选择稳健归一化 (RobustScaler)，避免数值大幅度波动导致映射偏移。

5.3 日订单需求量预测

将数据按日为时间粒度进行重采样，将数值规整，选取最后 90 个点作为测试集，以评估模型精度。以 105-21619-306-402 产品为例，预测该产品未来 90 天 (3 个月) 订单需求量。该产品日订单需求时序图如图5-1。

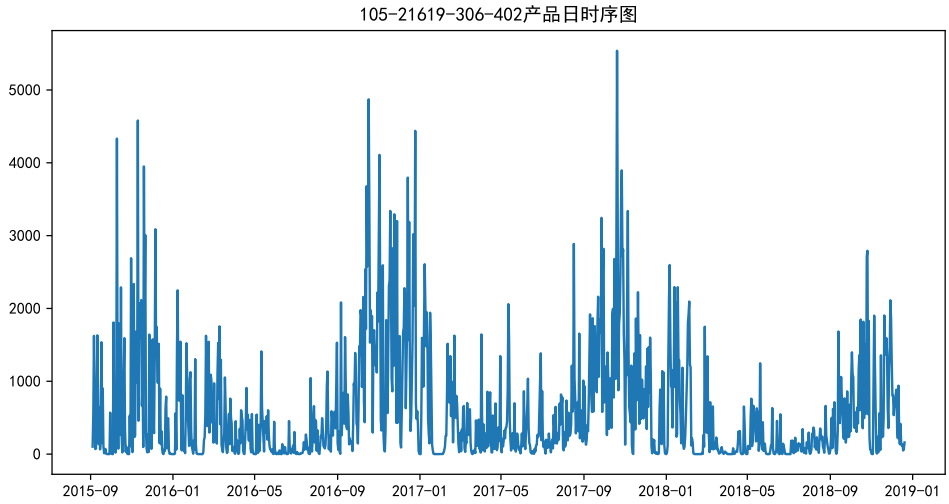

图 **5-1　105-21619-306-402** 产品日时序图

5.3.1 CRU 模型

连续循环单元[5](continuous recurrent units, CRUs)，是一种新型循环神经网络模型，用于建模不规则时间序列数据。传统循环神经网络（RNN）模型，如 LSTM[13] 和 GRU 等，通常采用离散时间步来处理时间序列数据，这在处理不规则时间间隔数据时可能会存在问题。CRU 模型使用连续时间表示，能够有效处理不规则时间间隔时间序列数据。模型原理示意如图5-2

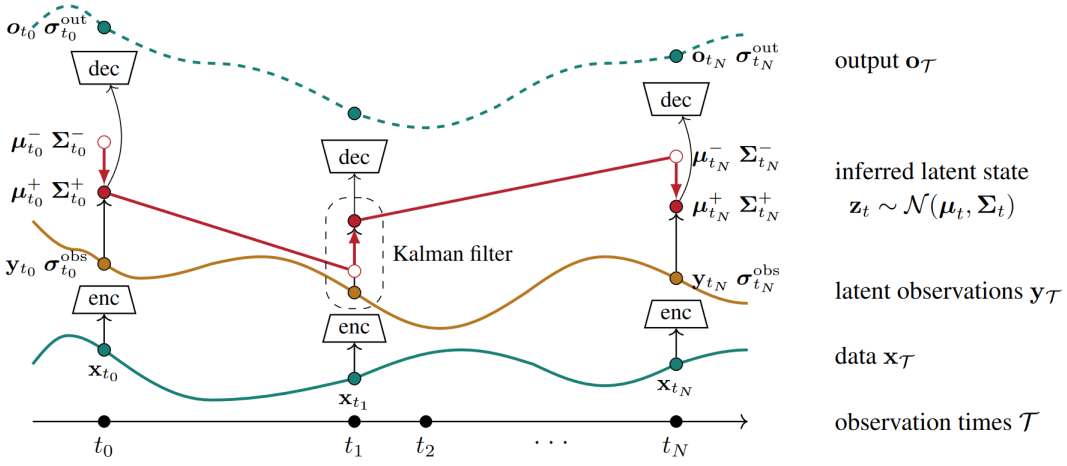

图 **5-2　CRU** 模型原理图

CRU 模型关键是引入了时间增量 (time increments) 这一概念，它表示相邻时间点间时间间隔。在 CRU 模型中，每个时间步的输入不仅包括当前观测值，还包括时间增量，这样网络就能够根据不同时间间隔来调整自己的内部状态。CRU 模型还引入了一个称为时间扭曲层 (time warping layer) 的模块，用于对时间间隔进行建模，从而进一步提高模型性能。

模型重要参数说明：

(1) 回视窗口 (lookback windows)：输入数据特征维度，也代表模型能学习的历史窗口信息。

(2) 隐藏状态维度 (hidden state dimension, hsd)：CRU 模型中隐藏状态维度大小。它决定了 CRU 模型中可以存储和处理信息容量。

(3) 时间增量维度 (time increment dimension, tid)：时间增量维度大小。时间增量维度是一个与时间相关的维度，用于捕捉时间间隔变化。

(4) CRU 层数 (number of CRU layers, ncl)：CRU 模型中 CRU 层数量。增加 CRU 层数可以提高模型表现能力，但也会增加计算复杂度和过拟合风险。

(5) 时间扭曲层数 (number of time warping layers, ntwl)：CRU 模型中时间扭曲层数量。时间扭曲层用于建模不同时间间隔变化模式，增加时间扭曲层数可以提高模型性能。但同时也会提高计算复杂度。

由于以天为时间粒度订单需求量波动和噪点较多，选用 Huber Loss 作为损失函数，当误差较小时，它使用 L2 范数来计算误差；而当误差较大时，它使用 L1 范数，从而保证了对异常值的容忍性。并且 Huber Loss 通过将 L1 范数和 L2 范数相结合，可以让损失函数具有平滑性，从而使得优化更加稳定。

$$\text{HuberLoss}(y,\hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2, & \text{if } |y-\hat{y}| \le \delta \\ \delta\left(|y-\hat{y}| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}$$

保持模型其他外部条件不变，使用 TPE[7](Tree-structured Parzen Estimator, TPE) 算法调整上述参数。TPE 是一种树状贝叶斯优化方法，支持所有类型搜索空间，效率高，无任何依赖组件。以时间滚动交叉检验 MSE 作为评估标准，迭代 1000 次。得到最优参如表5-3。

表 **5-3** **CRU** 模型最优参

hsd	tid	ncl	ntwl
50	3	2	15

固定模型架构，对模型学习率 (learning_rate)，dropout，梯度阈值 (clip_gradient)，优化器权重衰减系数 (weight_decay) 进行调整。首先使用 TPE 算法缩小搜索区间，再使用网格搜索逼近最优参，以时间滚动交叉检验 MSE 作为评估标准。最优参及模型损失如表5-4。

表 **5-4**　**CRU** 模型最优参

learning_rate	dropout	clip_gradient	weight_decay	MSE (10^7)
0.0021	0.01	10	0.9×10^{-8}	0.643

将 Prophet、DeepAR 作为基线模型，CRU 与模型进行对比，各模型损失如表5-5。

表 **5-5**　各模型结果对比

模型	MSE (10^7)	MAE (10^7)
CRU	0.643	0.601
DeepAR	0.779	0.753
Prophet	0.947	0.919

可以看到，CRU 模型相较于其他基线模型具有更高精度，选择 CRU 模型作为日预测模型。模型在测试集上表现及 90 天订单需求量预测效果如图5-3。

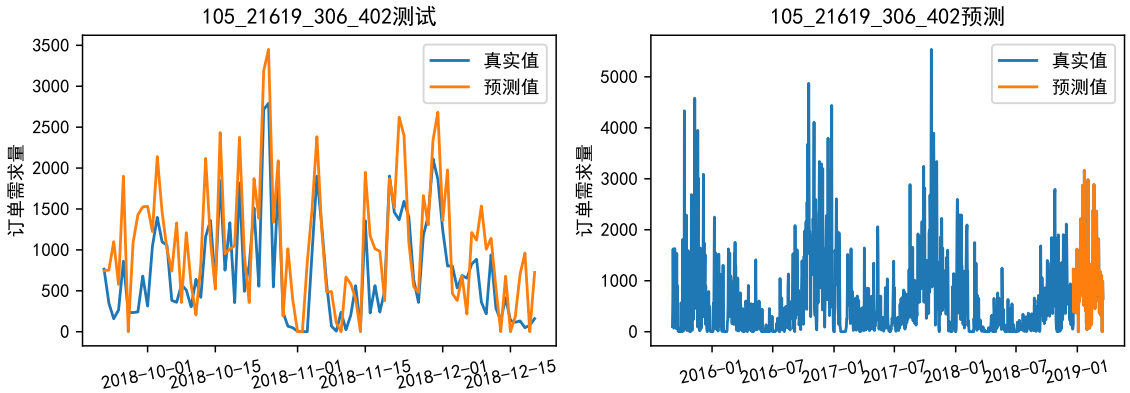

图 **5-3**　**CRU** 效果图

随机选择 3 个产品，使用 CRU 模型预测 90 天订单需求量，测试（上）与预测（下）效果如图5-4。

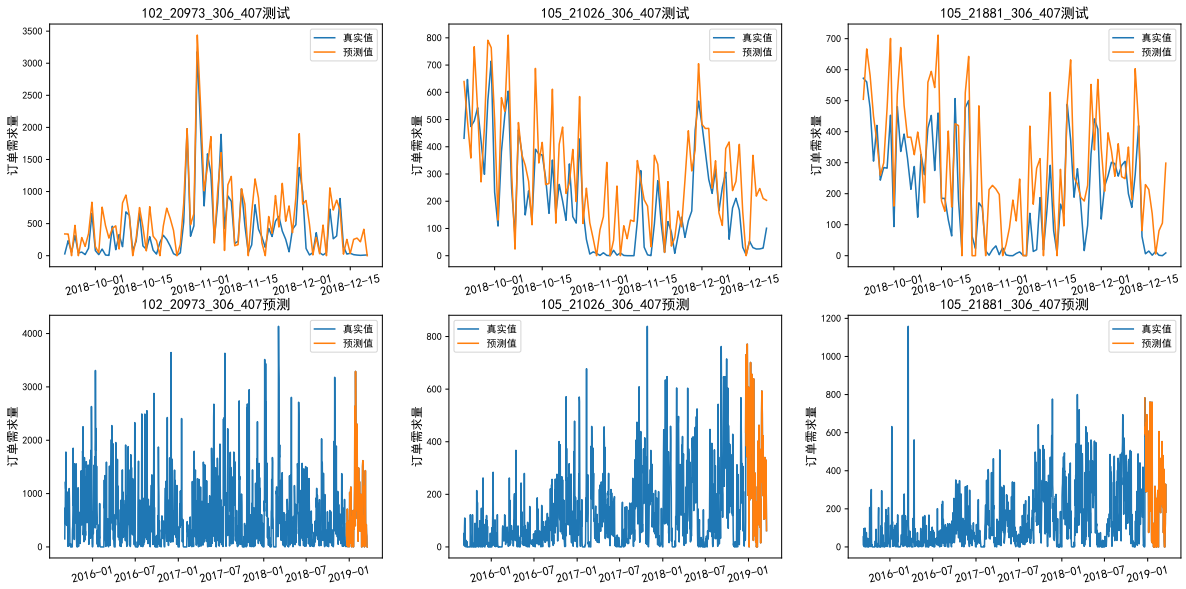

图 **5-4**　部分产品预测效果图

5.4 周订单需求量预测

将数据按周为时间粒度进行重采样，同时将数值规整，选取最后 14 个点作为测试集，以评估模型精度。以 105-21619-306-402 产品为例，预测该产品未来 14 周（3 个月）订单需求量。该产品周订单需求量时序如图5-5。

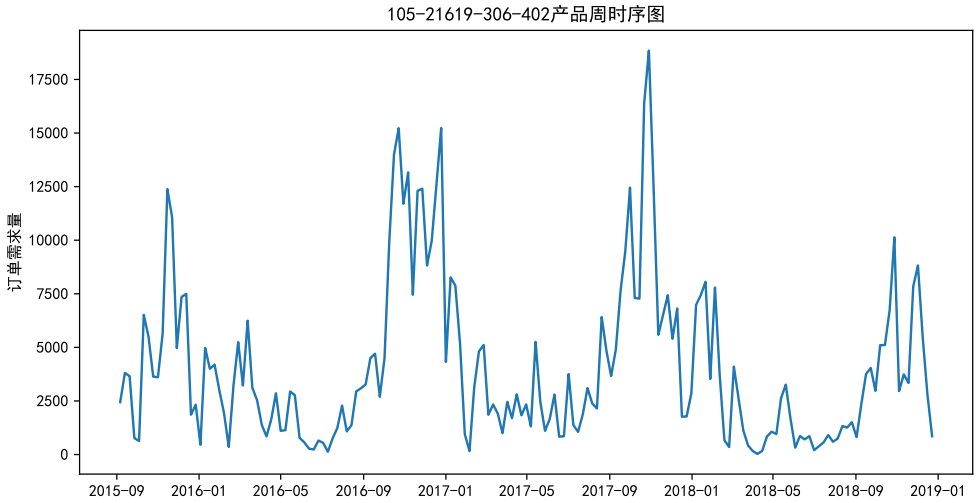

图 **5-5**　**105-21619-306-402** 产品周时序图

5.4.1 DeepAR 模型

DeepAR[8] 模型是一种基于自回归循环网络 (RNN) 的概率预测模型。DeepAR 模型不仅可以处理单变量时间序列，还可以处理包含多个相关变量的多变量时间序列。通过将其他相关变量作为外部协变量，模型可以更准确地捕捉时间序列间依赖关系。该模型

除了能预测单个点结果，还可以生成未来一段时间内概率分布结果，对企业生成计划和决策提供风险度量。模型自回归示意图如图5-6。

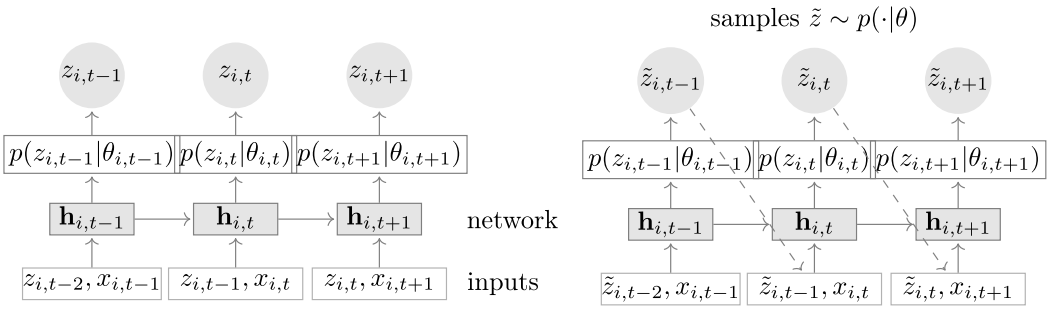

图 **5-6　DeepAR** 自回归原理图

模型重要参数说明：

(1) context_length(回视窗口)：输入模型历史时间步数，用于训练模型。context_length 参数理想状态下取时间序列周期性和波动情况明显是时段。

(2) cell_type(循环单元类型)：循环单元类型，可选 LSTM 或 GRU。

(3) num_layers(循环网络层数)：所选循环单元堆叠层数。

(4) num_cells(循环网络单元数)：模型中循环神经网络单元数量，该参数决定模型复杂度和记忆能力。

(5) embedding(协变量嵌入维度)：协变量嵌入维度，数值越高对协变量信息参考权重越高。

保持模型其他外部条件不变，使用 TPE 算法调整上述参数，以时间滚动交叉检验 MSE 作为评估标准，迭代 1000 次。得到最优参如表5-6。

表 **5-6　DeepAR** 模型最优参

context_length	cell_type	num_layers	num_cells	embedding
28	LSTM	2	32	50

固定模型架构，对模型学习率 (learning_rate)，dropout，梯度阈值 (clip_gradient)，优化器权重衰减系数 (weight_decay) 进行调整。首先使用 TPE 算法缩小搜索区间，再使用网格搜索逼近最优参，以时间滚动交叉检验 MSE 作为评估标准。最优参及模型损失如表5-7。

表 **5-7**　**DeepAR** 模型最优参

learning_rate	dropout	clip_gradient	weight_decay	MSE (10^7)
0.0012	0.05	8	1.2×10^{-8}	0.475

将 Prophet、CRU 作为基线模型，与 DeepAR 模型进行对比，各模型损失如表5-8。

表 **5-8**　各模型结果对比

模型	MSE (10^7)	MAE (10^7)
DeepAR	0.475	0.473
CRU	0.526	0.500
Prophet	0.589	0.554

可以看到，DeepAR 相较于其他基线模型具有更高精度，选择 DeepAR 模型作为周预测模型。模型在测试集上表现及 14 周订单需求量预测效果如图5-7。

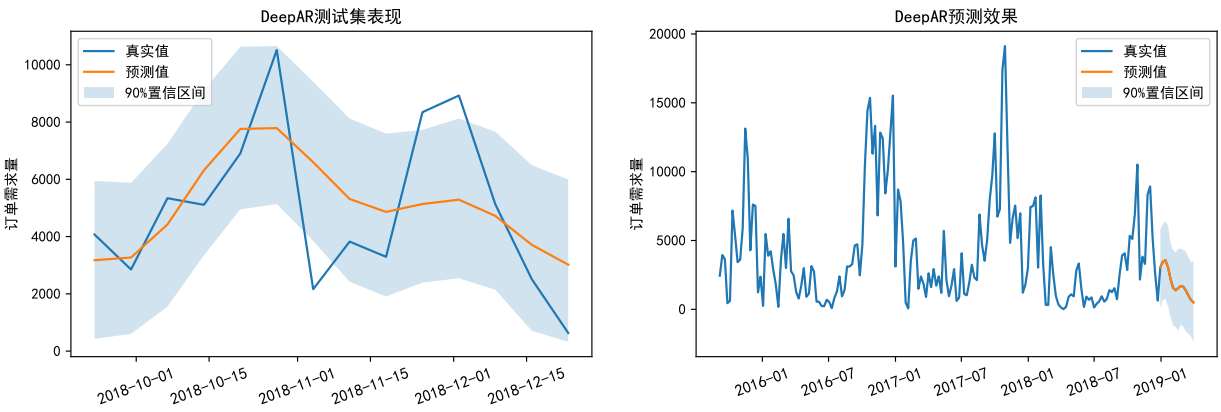

图 **5-7**　**DeepAR** 效果图

随机选择 3 个产品，使用 DeepAR 模型预测 14 周订单需求量，测试（上）与预测（下）效果如图5-8。

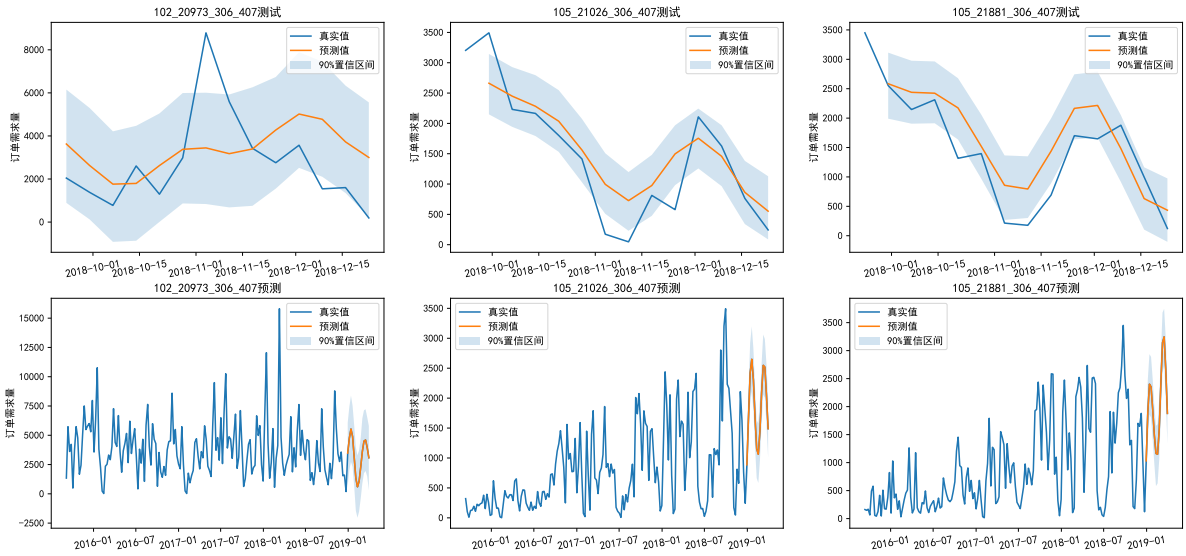

图 **5-8**　部分产品预测效果图

5.5 月订单需求量预测

将数据按月为时间粒度进行重采样，同时将数值规整，选取最后 3 个点作为测试集，以评估模型精度。以 105-21619-306-402 产品为例，预测该产品未来 3 个月订单需求量。该产品月订单需求量时序图如5-9。

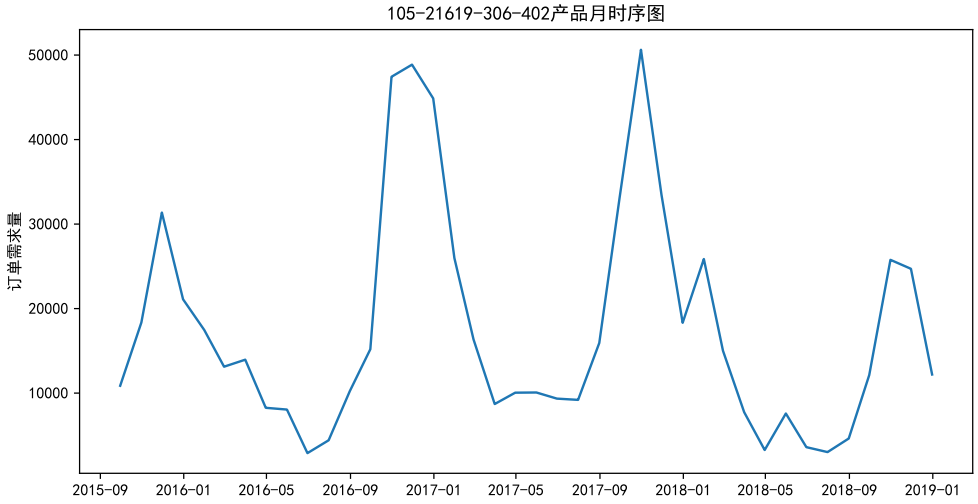

图 **5-9**　**105-21619-306-402** 产品月时序图

5.5.1 Prophet 模型

Prophet[9] 模型是一种时间序列预测模型。它能够对季节性、趋势性和节假日等因素进行建模，从而对未来时间序列进行预测。Prophet 模型优势在于：对异常值和缺失数据具有较好稳健性，能够在数据缺失情况下进行预测。可解释性强，能够给出每个因

素对预测结果影响程度。模型原理为：

$$y(t) = g(t) + s(t) + h(t) + \epsilon \tag{2}$$

其中 $g(t)$ 是趋势函数，$s(t)$ 表示周期性函数，$h(t)$ 表示节假日、假期函数，ϵ 表示误差或者是噪声等。

模型重要参数说明：

(1) n_changepoints(突变点数量，简称 nc)：时间序列中出现突变点数量最大值。

(2) changepoint_range(突变出现范围，简称 cr)：与 n_changepoints 配合使用，changepoint_range 决定了突变点能出现在离当前时间最近的时间点，changepont_range 越大，突变点离当前时间越近。

(3) changepoint_prior_scale(突变点拟合度，简称 cps)：突变点拟合程度，值越大，对历史信息拟合越强，越容易过拟合。

(4) seasonality_mode(季节模型，简称 sm)：季节模型类型，可以选择 additive(加法模型) 或者 multiplicative(乘法模型)。

(5) seasonality_prior_scale(周期性因素占比，简称 sps)：更改周期性影响因素强度，值越大，周期因素在预测中占比越大。

保持模型其他外部条件不变，使用 TPE 算法调整上述参数，以时间滚动交叉检验 MSE 作为评估标准，迭代 1000 次。得到最优参如表5-9。

表 **5-9** **Prophet** 模型最优参

nc	cr	cps	sm	sps	MSE (10^7)
12.00	0.84	0.05	multiplicative	15	0.310

将 DeepAR、CRU 作为基线模型，与 Prophet 模型进行对比，各模型损失如表5-10。

表 **5-10** 各模型结果对比

模型	MSE (10^7)	MAE (10^7)
DeepAR	0.407	0.392
CRU	0.433	0.431
Prophet	0.310	0.308

可以看到，Prophet 相较于其他基线模型具有更高精度，选择 Prophet 模型作为月预测模型。模型在数据集上表现及 3 个月订单需求量预测效果如图5-10。

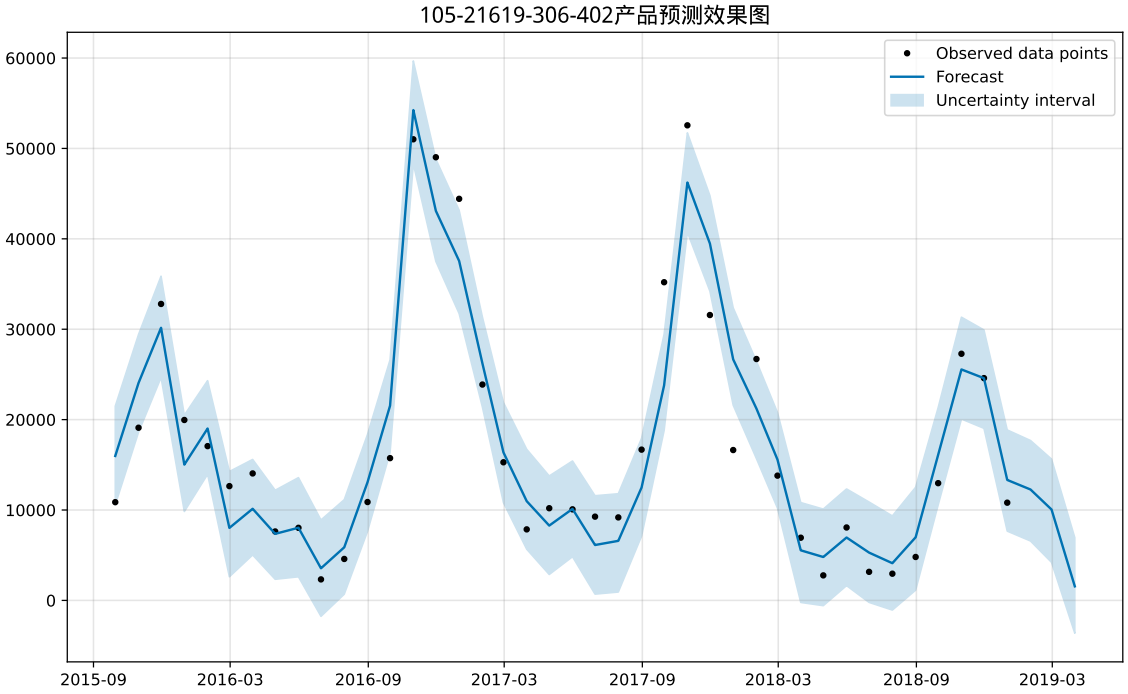

图 **5-10**　模型预测效果图

Prophet 模型输出时间序列趋势分解图可以观察模型预测过程中各趋势所起作用，时序分解如图5-11。

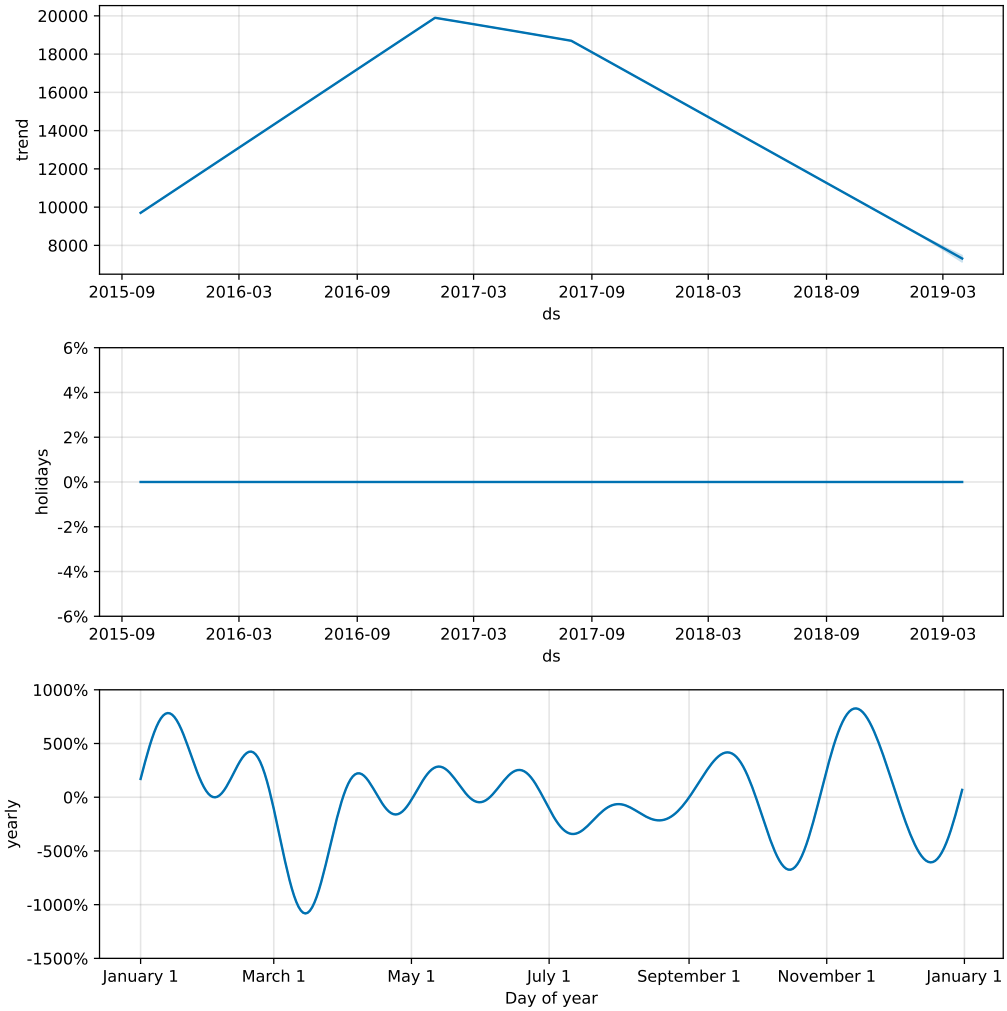

图 **5-11**　模型时序分解图

可以看到，从 2015 年 9 月到 2017 年 1 月产品需求量呈上升趋势，然后到 2017 年 8 月产品需求量开始下降，2017 年 9 月以后下降趋势明显变大。但月份趋势十分有规律，均在冬季有所上升，夏季出现下降。随机选择 3 个产品，使用 Prophet 模型预测 3 个月订单需求量，效果如图5-12。

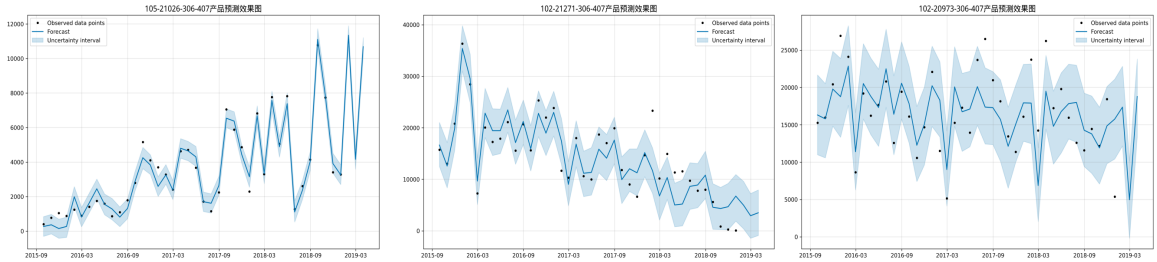

图 **5-12**　部分产品预测效果图

因为预测集中产品存在数据缺失，将样本数小于 3 的产品未来 3 个月订单需求量使用平均值代替，无训练数据产品未来 3 个月订单需求量使用同区域各月份中位数代替。

展示预测集前 5 产品未来 3 个月订单需求量，如表5-11

表 **5-11**　部分产品预测结果展示

销售区域编码	产品编码	1 月需求量	2 月需求量	3 月需求量
101	20002	67	71	75
101	20003	485	489	494
101	20006	213	218	223
101	20011	325	311	208
101	20014	255	324	465

5.6 各时间粒度预测分析

根据上述各时间粒度预测模型效果，可以发现 CRU 模型在天时间粒度上表现较好，这可能是 CRU 模型时间增量模块对不规则时间建模的优势。并且天产品订单需求量波动较大，其他模型很难学习到数据波动特性，往往预测较为保守。而 CRU 模型时间扭曲模块能很好处理这个问题，这也是 CRU 模型能在天时间粒度上保持良好预测性能原因。

观察产品周时序图可以发现，周期性相较于天时间粒度更明显，并且突变和波动都较为平缓。而 DeepAR 模型采用 LSTM 为循环结构单元，对周期信息敏感，能很好的拟合产品周需求量。相对的，CRU 模型中重要模块：时间增量和时间扭曲不再具有优势。Prophet 模型则因为模型复杂度不够，在测试集中表现较差。所以 DeepAR 模型能在周时间粒度上保持良好预测性能。

当产品以月为时间粒度进行重采样时，样本总量不超过 36，趋势与周期性更加明显。CRU 模型和 DeepAR 模型等深度模型在小样本数据上非常容易过拟合，模型过度拟合噪点反而忽略了时序整体特性。而 Prophet 模型采用时序分解模块对年、季、月趋势拟合，以传统时序分析方法为核心，避免了过拟合现象。所以 Prophet 模型能在月时间粒度上保持良好预测性能。

六、 模型评价

模型优点:

- 构建特征工程，并使用 MMIFS 方法选取具有时序特性的特征，客观严谨。
- 针对不同时间粒度特性选择不同建模方法。

- 使用滚动时间交叉验证评估模型。
- 使用 TPE 算法调参，时空复杂度低，模型收敛迅速。

模型改进:

- 样本数量较少，并且时间序列不规整，希望后续获得完整数据进一步提高模型精度。
- 经过上述分析，发现产品订单需求量与季节因素有较强相关性，希望能获得温度、湿度等天气特征，可以进一步提高模型稳健性。

参考文献

[1] 李航, 统计学习方法 [M].2 版. 北京: 清华大学出版社,2019.

[2] 周志华, 机器学习 [M].1 版. 北京: 清华大学出版社,2016.

[3] 侯俊军, 岳有福, 叶家柏. 供需双循环测度与中国经济平稳增长 [J]. 统计研究,2023,40(03):3-17.DOI:10.19343/j.cnki.11-1302/c.2023.03.001.

[4] 马云鹤, 王玉玫, 赵宇帆. 基于时空融合图的共享单车需求预测系统 [J]. 计算机测量与控制,2023,31(02):97-103.DOI:10.16526/j.cnki.11-4762/tp.2023.02.015.

[5] Schirmer M, Eltayeb M, Lessmann S, et al. Modeling irregular time series with continuous recurrent units[C]//International Conference on Machine Learning. PMLR, 2022: 19388-19405.

[6] Chen X, He Z, Sun L. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation[J]. Transportation research part C: emerging technologies, 2019, 98: 73-84.

[7] Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyper-parameter optimization[J]. Advances in neural information processing systems, 2011, 24.

[8] Salinas D, Flunkert V, Gasthaus J, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks[J]. International Journal of Forecasting, 2020, 36(3): 1181-1191.

[9] Taylor S J, Letham B. Forecasting at scale[J]. The American Statistician, 2018, 72(1): 37-45.

[10] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[11] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM’ 08. Eighth IEEE International Conference on.

[12] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation-based anomaly detection.” ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 3.

[13] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[14] Han M, Ren W, Liu X. Joint mutual information-based input variable selection for multivariate time series modeling[J]. Engineering Applications of Artificial Intelligence, 2015, 37: 250-257.