队伍编号	202305203054
题号	B

基于 LSTM 模型的产品需求分析和预测问题研究

摘要

近年来，全球经济环境的不确定性和复杂性不断加剧，在当前外部环境日趋复杂，企业供应链管理面临着前所未有的挑战。作为企业供应链的第一道防线，需求预测显得愈加重要。在这种情况下，如何准确预测市场需求成为了供应链管理的重要问题。本文拟开发一个产品订单的数据分析于需求预测模型。首先对训练数据中各个影响因素对于需求量的影响进行多维度的深入分析，在分析的基础上，对比多个模型，最终建立基于 LSTM 的预测模型，分别按天、周、月的时间粒度对未来三个月的月需求量进行预测，并分析不同时间粒度对预测精度的影响。

针对问题一，我们以训练数据中给出的不同特征为切入点，将特征分为离散型变量与连续型变量分析其对产品需求量的影响。我们首先采用**相关性分析方法**和**分位数回归模型**对于产品价格对需求量的影响进行量化分析，给出协方差矩阵和热力图。对于离散型数据，我们采用可视化方法辅助分析。对于节假日与促销日，我们挑选了部分代表性日期及其前后几天的销售量进行对比分析；对于销售方式、时间段、季节等影响因素，我们采用**独立检验**及分类汇总方法对数据进行分析，得出普遍性规律。

*针对问题二，*采用了 **LSTM、随机森林、决策树**以及 **XGBoost** 模型对未来三个月的月需求量进行预测。首先基于线性插值法和需求量的年周期性填补缺失值；基于第一问的需求分析，选择特征，建立预测模型，并通过多次实验进行参数调优。实验表明，LSTM 模型优于其它模型，在需求量预测中精度最高，故选用 LSTM 为最终的预测模型。最后对不同的时间粒度进行分析，发现以日为周期的时间粒度由于前期数据预处理的插值填补，对于空缺数据较多的产品存在误差累积效应，故精度低于以周、月为粒度的预测。

关键词：需求分析，需求预测，LSTM，相关性分析，分位数回归

目录

一、 问题重述

1.1 问题背景

随着经济全球化、通信技术革新换代，许多企业面临着用户饱和、主营业务收入下滑等压力，在这个背景下，国内各大企业都意识到可持续的供应链管理将成为企业的竞争优势资源，希望通过采购精细化管理、合理控制库存、控制成本等方面构建低成本、高效益供应链管理机制，为企业决策提供强有力的支撑，强化企业的核心竞争力。于此同时，全球经济环境的不确定性和复杂性不断加剧，让企业供应链管理面临着越来越多的挑战。在这种情况下，如何准确预测市场需求成为了供应链管理的重要问题。由于市场需求受多种因素的影响，如消费者行为变化、市场竞争状况、自然环境等，对于企业来说，进行准确的需求预测非常具有挑战性。为此，需要通过更加先进的算法和技术手段来提高需求预测的准确率。

准确的需求预测对于企业管理层制定销售和运营计划、目标以及资金预算等方面具有重要参考价值。通过对市场需求的准确预测，企业可以更加精细地调整生产和供应链管理策略，从而提高运营效率和降低成本。此外，需求预测还能够帮助企业进行采购计划和生产计划的制定，减少业务波动的影响，进一步提高供应链的稳定性和可靠性；如果企业没有进行准确的需求预测，或者预测结果不准确，会导致很多内部关于销售、采购、财务预算等决策都只能依靠经验来做出。这样的情况下，企业很难对市场趋势进行准确的把握，可能会存在产品过剩或过少的库存问题，进而增加了企业的库存成本和资金风险。

因此，提高需求预测的准确性成为建立高效、稳定供应链的关键之一。通过使用先进的算法和技术手段，企业可以更好地应对不确定的外部环境，实现供应链管理的优化和协调。

1.2 问题重述

根据附件中提供的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（order_train1.csv），需要对数据进行充分、深入的分析，并对产品的未来需求量进行预测。本文拟解决以下问题：

针对问题一：对附件中数据的需求量进行多方面、多角度的分析，包括以下几个方面：

（1） 产品的不同价格对需求量的影响；

（2） 产品所在区域对需求量的影响，不同区域的产品需求量有何特性；

（3） 不同销售方式（线上和线下）的产品需求量的特性；

（4） 不同品类之间的产品需求量有何不同点和共同点；

（5） 不同时间段（例如月头、月中、月末等）产品需求量有何特性；

（6） 节假日对产品需求量的影响；

（7） 促销（如 618、双十一等）对产品需求量的影响；

（8） 季节因素对产品需求量的影响。

针对问题二：基于对数据的综合分析，建立数学模型，对附件给出的预测数据（predict_sku1.csv）中给出的产品，分别按天、周、月的时间粒度进行预测，给出未来 3 月（即 2019 年 1 月、2 月、3 月）的月需求量，并分析不同预测粒度对预测精度的影响。

二、 模型假设

（1）假设附件中提供的所有数据都是真实可信的；

（2）假设除价格、产品品类、地区、促销、节假日等因素（问题一中分析的因素）外其它因素，如社会事件因素等对产品需求的影响作用不显著；

（3）假设未来一段时间不存在突发重大事件对需求市场产生重大影响；

（4）假设未来一段时间内市场总体需求量没有发生突变，不存在通货膨胀和通货紧缩。

三、 相关理论研究分析

3.1 相关研究概述

需求分析旨在准确理解用户、项目和产品的功能、性能、可靠性等具体要求，并将用户非形式的需求转化为完整的需求定义。它的应用广泛，可用于新产品的设计、现有产品的改进等方面。在当今互联网时代，随着技术和市场的变化，人们对产品的需求也不断变化，因此需求分析的重要性日益凸显。除了传统的需求收集和分析方法外，现代的数据分析、人工智能等技术也逐渐应用于需求分析，以提高效率和准确性。

对于大数据时代的到来，麦肯锡最早做出了预测：“数据，已经渗透到当今每一个行业和业务领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来[1]”。当今时代，大数据将不断改变企业的经营管理方式，基于大数据挖掘作为决策依据的模式已经为不少的企业带来经济效益和管理提升。

3.1.1 国内相关研究

对于需求预测领域来说，大数据分析技术的应用已成为供应链效能提升的有效手段，也是企业参与竞争的强大动力。陈爱菊在采购计划制定中运用时间序列的预测模型和经济库存模型结合的方式降低了采购库存成本[2]。资武成提出基于企业级的大数据生态子系统的构建策略[3]，通过存量大数据分析搭建新型客户关系管理体系，推动数据挖掘向客户行为延伸，在客户行为分析的基础上准确预测市场变化规律并合理应对。

3.1.2 国外相关研究

国外许多学者也提出了切实可行的理论和方法来构建大数据供应链。在数据预测方面，Hua 经研究发现了时间序列的特征，并选择最适合的 ARIMA 模型开展序列建模[4]；俄国数学家 A.A.Markovr 提出的马尔科夫算法，基于历史数据分析历史状态分布，预测将来的发展趋势[5]。Xu 等人在供应链的各关键节点上运用需求预测模型，并根据预测结果向供应商下订单，有效缓解了超龄库存形成[6]。Erik Hofmann 经研究发现大数据的高速率对供应链的牛鞭效应有较好的控制效果[7]。通过大数据预测，能够有效提升供应链管理效能，赋能企业管理能力的提升。

3.2 数据挖掘常用技术

数据挖掘技术的应用领域快速拓展，其实质是知识的转化，对数据中蕴含的规律进行分析及把握并预测未来的发展动态，针对未来发展趋势，采取相应的措施[8]。其常用技术主要有以下几种：

(1) 分类：是通过对一些已知类别标号的训练数据进行分析，找到一种可以描述和区分数据类别或概念的模型，然后用这个模型来预测未知类别标号的数据所属的类别。分类模型

的形式有许多，例如决策树，神经网络，朴素贝叶斯分类器，支持向量机和 KNN 分类器等[9]。

(2) 回归：对具有连续取值的函数进行建模。回归分析是一种统计方法，常用于数值预测。回归分析法是在需求量和影响其变化的自变量间找到其存在的线性关系，通过训练出拟合函数，建立模型实现数据预测[8]。

(3) 离群点分析：离群点可以通过统计测试进行检测，即假设数据集服从某一个概率分布，看某个对象是否在该分布范围之内。也可以使用距离测量，将那些与任何聚类都相距较远的对象视为离群点，对于局部区域内的离群点也可采取密度相关的检测手段来进行[9]。

3.3 产品订单需求预测算法原理

3.3.1 LSTM 算法原理

LSTM (长短时记忆网络) 是一种用于处理和预测时间序列中间隔和延迟相对较长的重要事件的递归神经网络模型。LSTM 模型的主要思想是引入门机制，通过控制信息的输入、遗忘和输出，从而有效地解决长期依赖性问题。

LSTM 中的每个单元包含一个状态向量 c_t，它负责保存上一步的记忆，并通过门机制控制当前输入数据的影响。具体来说，LSTM 中有三个门：输入门、遗忘门和输出门。它们通过一些权重矩阵和偏置项来控制状态 c_t 的更新。

- 输入门：将输入数据与前一时刻的状态向量 h_{t-1} 经过一个 sigmoid 函数计算，得到更新状态的系数 $\Gamma_i(t)$。然后将输入数据通过一个 tanh 函数映射到一个新的向量 $\tilde{C}_t$。最后，将这两个值相乘并加到c_{t-1}上，上，得到下一时刻的状态向量

$$C_t = f_t \cdot C_{t-1} + \Gamma_j(\boldsymbol{t}) \cdot C_t \tag{1}$$

其中 f_t 是遗忘门。

- 遗忘门：计算 $f_t = \sigma(W_f \cdot [h_{t-1}, \chi_t] + b_f)$，其中 σ 是 sigmoid 函数，W_f，b_f 是网络学习到的参数。f_t 负责决定在当前时间步长中保留上一时刻状态向量c_{t-1}的哪些部分。
- 输出门：计算 $\phi_t = \sigma(W_0 \cdot [h_{t-1}, x_t] + b_0)$，其中$W_0$，$b_0$ 是网络学习到的参数。然后，通过一个 tanh 函数将当前状态向量c_t映射到一个新的向量$\tilde{h}_t$。最后，将$\tilde{h}_t$ 乘以o_t，得到输出向量$h_t = O_t \cdot \tanh(C_t)$。

3.3.2 随机森林算法原理

随机森林是以决策树为估计器的 Bagging 算法，将多个决策树结合在一起，每次数据集是随机有放回的选出，同时随机选出部分特征作为输入。

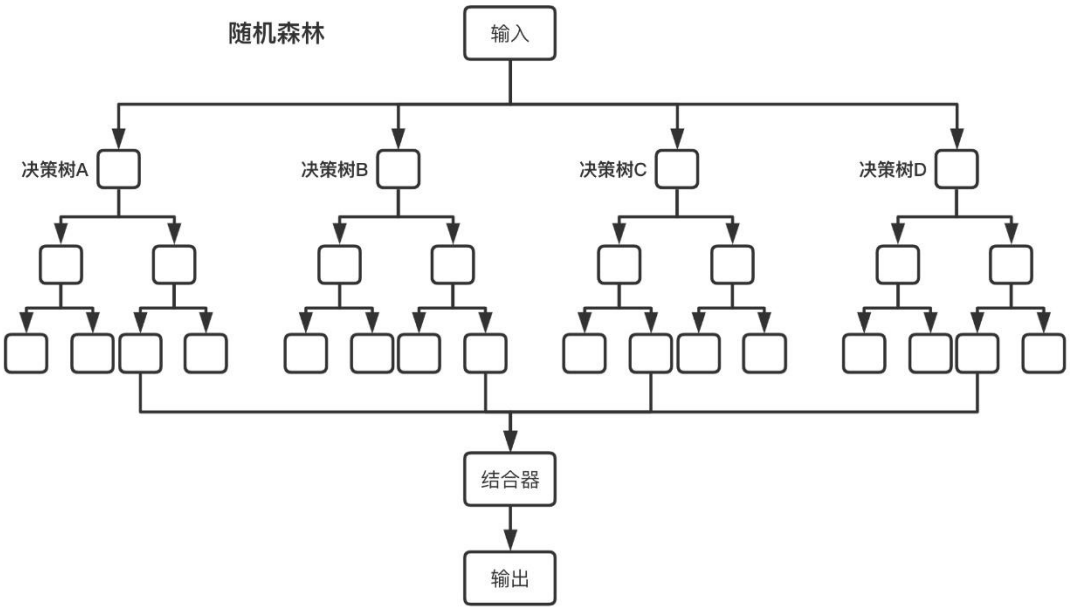

图 1：随机森林算法示意图

其中决策树是数据挖掘技术中的一种重要的分类与回归方法，它是一种以树结构(包括二叉树和多叉树)形式来表达的预测分析模型。决策树以信息熵为评估标准，对不同的特征进行分类，从而得到预测的结果。其中，随机森林和决策树的重要性特征由基尼系数（不纯度）来衡量，其表达式为：

$$G(p)=\sum_{k=1}^{K} p_k(1-p_k)=1-\sum_{k=1}^{K} p_k^2 \tag{2}$$

其中 p 为事件的概率，k 为事件发生种类，我们根据归一化后的基尼系数来筛选特征，值越大代表该特征越重要。

3.3.3 **决策树算法原理**

决策树回归模型是一种基于树结构的回归模型。它将数据集划分为多个子集，每个子集对应一个节点，最终形成一棵以每个节点为分裂点的决策树。对新数据进行预测时，将其从根节点开始沿着树的分支走到叶子节点，叶子节点所对应的值即为预测结果。

决策树回归模型的关键是如何划分数据集和构造回归树，一般采用贪心算法进行构建。具体地，首先选取一个划分特征和阈值，然后将数据集划分为左右两个子集，每个子集对应一个节点，并计算其平方误差。接着对两个子集分别递归上述过程，直到满足停止条件为止。

回归树的节点划分标准可以采用平方误差最小化准则，即：

$$\min_{j,s}\ I\min_{\chi_i}\sum_{\chi_i\in\mathcal{R}(j,s)}(y_i-\mathcal{C}_1)^2+\min_{\mathcal{C}_2}\sum_{\chi_i\in\mathcal{R}_2(j,s)}(y_i-\mathcal{C}_2)^2] \tag{3}$$

其中，j表示划分特征，表示划分阈值，$\mathcal{C}_1$，$\mathcal{C}_2$表示左右子集的预测值，$\mathcal{R}_1$，$\mathcal{R}_2$表示左右子集。

当样本点集合D中的每个样本都被预测为c_D时，平方误差为：

$$\sum_{x_i\in D}(y_i-c_D)^2 \tag{4}$$

对于给定的特征j和阈值s，将数据集划分为$D1$和 D2 两个子集，分别对应节点 t1 和 t2，则可以定义平方误差损失函数为：

$$L(\hat{t}) = \sum_{x_i \in D_t} (y_i - \hat{c}_i)^2 \tag{5}$$

其中$\hat{c}_i$表示节点 t 对应的预测值，节点 t 的最优化问题即为：

$$\min_{j,s} [\min_{c_1} \sum_{x_i \in D_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in D_2(j,s)} (y_i - c_2)^2] \tag{6}$$

最终得到的决策树模型可表示为：

$$f(x) = \sum_{m=1}^{M} c_m \cdot I(x \in R_m) \tag{7}$$

其中，M 表示叶子节点数，R_m表示第 m 个叶子节点对应的样本区域，$I(x \in R_m)$表示样本 x 的特征取值是否落在区域R_m内。

3.3.4 XGBoost 算法原理

XGBoost 的全称是 eXtreme Gradient Boosting，它是经过优化的分布式梯度提升库，旨在高效、灵活且可移植。XGBoost 算法是一种以 CART 决策树模型为基础的集成学习方法，但与 CART 模型不同，XGBoost 以损失函数的二阶泰勒展开式作为其替代函数，求解其最小化来确定回归树的最佳切分点和叶节点输出数值。

此外，XGBoost 通过在损失函数中引入子树数量和子树叶节点数值等，充分考虑到了正则化问题，能够有效避免过拟合。在效率上，XGBoost 通过利用独特的近似回归树分叉点估计和子节点并行化等方式，加上二阶收敛的特性，建模效率较一般的 GBDT 有了大幅的提升。

其目标函数如下：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i)^2 \tag{8}$$

其中，t 为迭代次数，n 为样本总量。上式的第二部分是将棵树的复杂度进行求和，添加到目标函数中作为正则化项，用于防止模型过度拟合，具体定义如下：

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2 \tag{9}$$

T 为叶子数量，$\omega^2{}_j$为叶子结点权重向量的L_2范数，λ，γ分别为两项的调节系数。

联立损失函数，可得最终目标函数为：

$$Obj^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, {\hat{y}_i}^{t-1}) + L' f_t(x_i) + \frac{1}{2} L'' {f^2}_t(x_i)] + \Omega(f_t) + const \tag{10}$$

3.4 预测模型误差评价指标

对需求量的预测，由于受到各种因素的影响，很难做到全面考虑，必定会产生一定的误差，只能尽量调整模型降低误差。误差越小，精度越高。为了对预测模型进行客观的评估，常用的误差评价指标有以下几种：

平均绝对误差（mean absolute error，MAE）：

对预测值和真实值之间绝对误差平均值的评价，MAE 越小，模型精度越高。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\,y_i-\hat{y}_i\,| \tag{11}$$

平均绝对百分比误差（mean absolute percentage error，MAPE）：

预测值和真实值的差与真实值比值的绝对误差平均值，相比平均误差，MAPE 的离差值被绝对化，不存在误差被正负抵消的情况，因此能更好的体现预测结果的误差情况。同样的，MAPE 越小，模型精度越高。

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}|\frac{y_i-\hat{y_i}}{y_i}|\times 100\% \tag{12}$$

均方根误差（root mean square error，RMSE）：

加强了对大误差在指标中的影响作用，从而使得该指标的灵敏度更高。RMSE 越小，模型精度越高。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2} \tag{13}$$

均方误差（mean square error，MSE）：

通过求误差的平方和再平均来计算误差，以衡量模型的真实值和预测值之间的偏差。MSE 越小，模型精度越高。

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{14}$$

其中，y_i为真实值，$\hat{y}_i$为预测值，n是样本数目

四、 需求分析

根据所给的附录所给的数据，本文将通过多个双变量分析进行需求分析，探究不同变量对需求的影响。根据变量的类型不同（连续型或类别型），本文将从定量数据影响分析和定类数据影响分析两大部分展开。

4.1 定量数据的影响

4.1.1 产品价格于需求量

本文首先对数据中的价格和需求量进行相关性分析，根据公式

$$\mathrm{Correlation}=\frac{\mathrm{Covariance}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{15}$$

计算得到产品价格和需求量的协方差矩阵如下：

$$\begin{matrix} 1 & -0.12078224 \\ -0.12078224 & 1 \end{matrix} \tag{16}$$

根据协方差矩阵画出热力图：

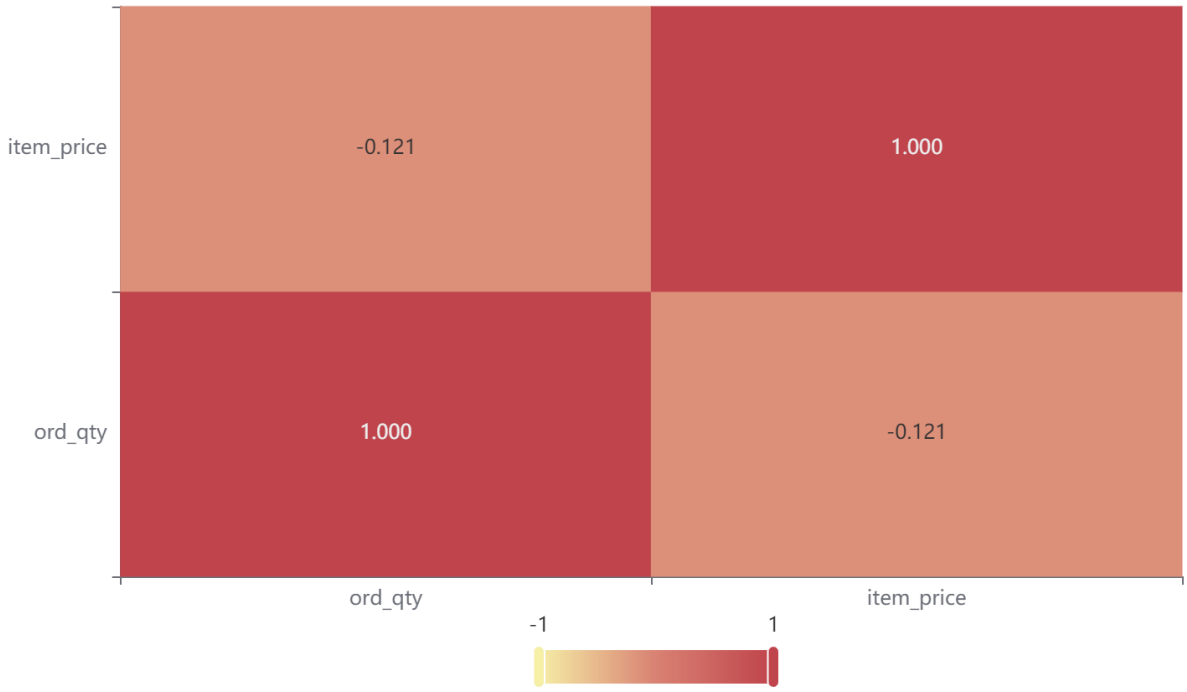

图 2：价格和需求量相关系数热力图

协方差用来衡量两个变量的总体误差，取值区间为[-1,1], 如果两个变量的变化趋势一致，协方差为值，说明两个变量正相关。如果两个变量的变化趋势相反，协方差为负值，说明两个变量负相关。如果两个变量相互独立，协方差为 0，说明两个变量不相关。热力图通过颜色的深浅可以直观地表示相关性系数的大小。

由热力图和相关系数矩阵可知，产品价格和需求呈负相关性（-0.12078224），也就是说，产品价格越高，需求量越低，但两者相关性较弱。考虑到不同品类之间的价格差异较大，且各品类产品本身的需求弹性也存在差异，对整体价格和需求量进行分析时无法得到较为显著的结果。因此，本文将通过 item_code（产品编码）对产品进行划分，分析在特定编码下产品不同价格对需求量的影响，并从特殊到一般，总结出产品价格对需求量影响的一般规律。由于产品众多，本文认为对需求量大的产品更具有代表性。通过统计各品类产品的总需求量，本文选取需求量前 3 的产品，分别为 21271，20973 和 21619，进行进一步的分析。

表 1：各产品需求总量统计

item_code	订单总需求量
21271	8455
20973	8239
21619	6450
……	……
21217	1
21747	1

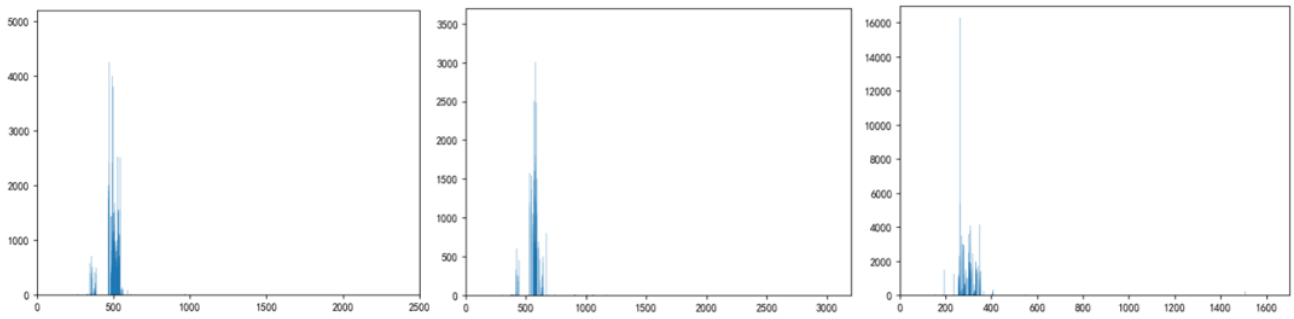

图 3：产品 21271，20973 和 21619（从左至右）不同价格与需求

由图可得，产品需求量都集中在[200,700]的价格区间，且随着价格的上升，需求量都呈现先上升后下降的趋势，基本体现正态分布特性。当产品价格处于某个特定阈值左右时，需求量集中且达到最大；当价格小于阈值时，价格上升需求量也上升；当价格大于阈值时，需求量下降且下降幅度大，总体符合负相关性。每个产品的阈值不同，体现了不同产品的不同需求弹性。为进一步量化探究价格对需求量的影响变化趋势，本文将产品价格作为自变量，订单需求量作为因变量进行分位数回归。

分位数回归既能研究在不同分位点处自变量对于因变量的影响变化趋势，也能研究在不同分位点处的哪些自变量是主要影响因素。原理是将数据按因变量进行拆分成多个分位数点，研究不同分位点情况下时的回归影响关系情况。步骤如下：

step1 通过运算，得到不同分位数的回归估计系数以及拟合效果(R^2)。

step2 通过作图，得到每个回归系数在不同分位数下的回归系数及其置信区间。

表 2：分位数回归结果表

	分位数 0.10	分位数 0.20	分位数 0.30	分位数 0.40	分位数 0.50	分位数 0.60	分位数 0.70	分位数 0.80	分位数 0.90
const	13.815 (0.084*)	57.863 (0.000***)	131.214 (0.000***)	209.473 (0.000***)	338.715 (0.000***)	446.183 (0.000***)	616.860 (0.000***)	944.640 (0.000***)	1322.991 (0.000***)
item_price	-0.011 (0.498)	-0.057 (0.000***)	-0.134 (0.000***)	-0.214 (0.000***)	-0.347 (0.000***)	-0.457 (0.000***)	-0.630 (0.000***)	-0.967 (0.000***)	-1.315 (0.000***)
R^2	0.001	0.005	0.010	0.011	0.014	0.013	0.013	0.016	0.017
因变量：ord_qty									

注：***、**、*分别代表 1%、5%、10%的显著性水平

对应某一分位数，若回归系数的 P 值小于 0.05，就说明该自变量是主要影响因素。上表展示了分位数回归的参数结果，包括分位数点、变量、样本量、拟合度 R^2 等。可从两方面来进行分析：

● 在不同分位数处产品价格（自变量）对需求量（因变量）的回归系数呈现的变化趋势

对于产品价格（item_price），回归系数始终为负，说明价格与需求量呈负相关；回归系数的绝对值随着分位数的增大而不断增大，这说明随着产品价格的不断提高，对产品需求量的负向影响逐渐增大。

● 在不同分位数处产品价格（自变量）的显著性

从 0.2 分位数开始，产品价格的系数都是显著的（p 值小于 0.05），说明产品价格对需求量

有显著影响。

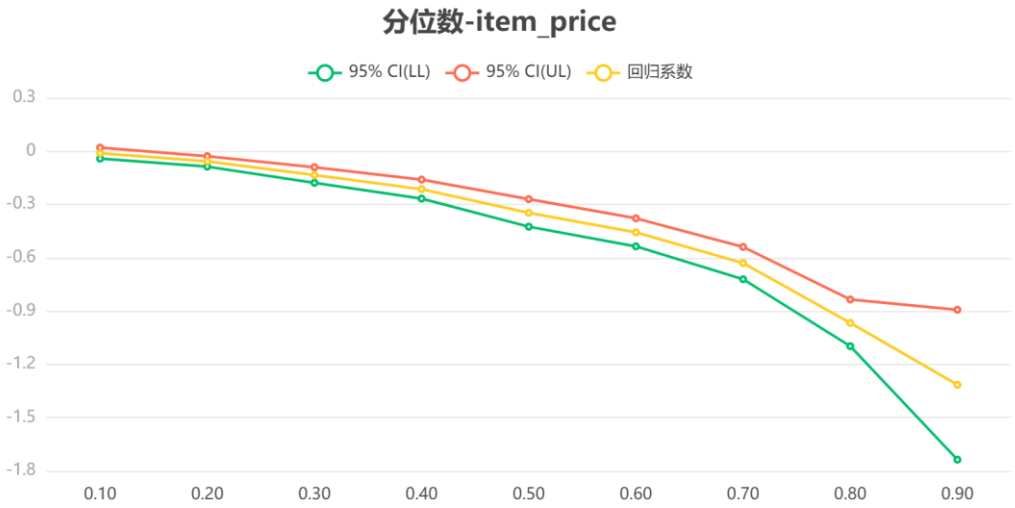

图 4：item_price 回归系数及置信区间

上图展示了分位数回归的参数结果。对于产品价格，回归系数始终为负值，分位点的回归系数绝对值整体上逐渐增大，并且在 0.9 分位点处对需求量的影响是最高的，这说明随着价格的不断提高，对需求量的负向影响逐渐增大。

产品价格对需求量有着明显的影响。一般而言，随着产品价格的上升，需求量会下降；反之，随着产品价格的下降，需求量会上升。这是由于消费者在购买产品时会考虑到产品的价格与自己的购买能力，价格越高，购买者的数量就会减少；价格越低，购买者的数量就会增加。此外，在一定范围内，价格的变化也会对消费者的购买决策产生较大的影响，价格的适度波动可能会改变消费者的心理预期和购买意愿。

4.2 定类数据的影响

4.2.1 产品所在区域与需求量

首先对各个销售区域需求量占比进行分析。本文统计了不同区域（sales_region_code）下需求量的总额，下图显示了不同区域订单需求额度占比：

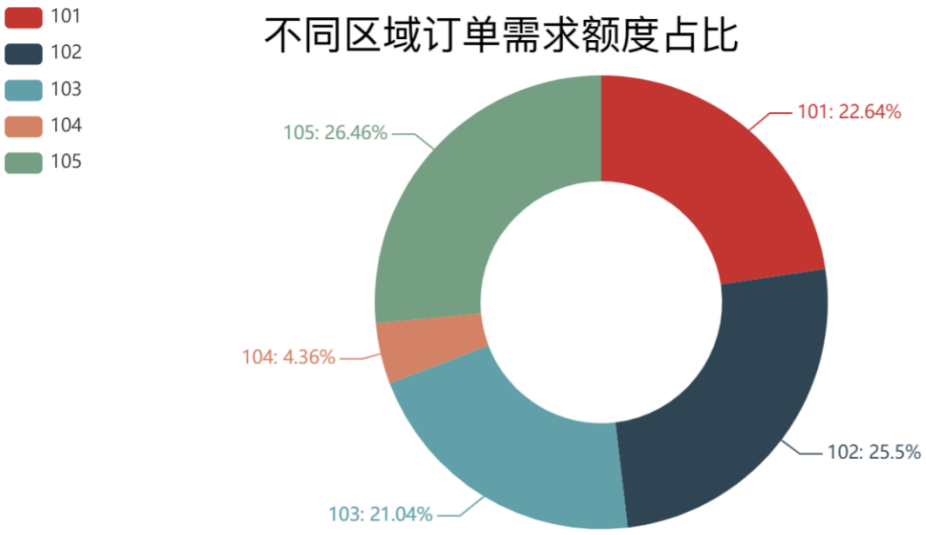

图 5：不同区域需求额度占比

由上图可知，105 区域需求量总额占比最大，为 26.46%，104 区域需求量占比最小，仅占 4.63%，而 101，102 和 103 区域的需求量占比相近，均与区域 105 相差不大，不同区域间需求占比极差较大。为进一步分析不同区域间的需求量差异大小，我们对不同地区对不同大类产品的需求进行了统计分析。

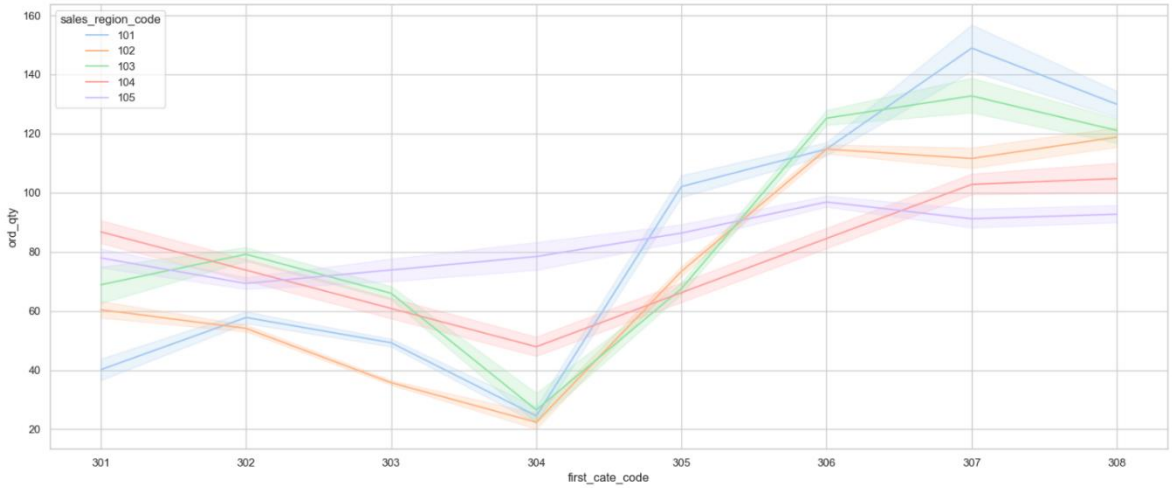

图 6：不同地区对不同大类产品的需求

由上图可得，不同销售地区对不同大类产品的需求量有明显差异，但对不同大类产品的需求占比趋势都大致相同。大部分地区对大类 304 产品的需求量都较低，而对大类 307 产品需求量都较高。

综上，产品销售区域对需求量产生较大的影响。按照产品的市场特征、消费人群、地理位置等因素，划分不同的销售区域可以更好地进行市场策略定位和实施。这样有利于企业更好地了解各个销售区域的客户需求、购买能力以及市场竞争状况，并基于这些信息制定更加有效的营销策略和销售计划，以提高销售收益和市场份额。此外，针对不同销售区域的目标客户特征，企业还可以进行不同的产品价格、促销活动和品牌推广等方面的差异化策略。

4.2.2 产品品类与需求量

本文根据产品大类将产品划分为不同品类，拟探究不同产品品类的需求量有何共同点和不同点。

首先研究不同品类产品间线上、线下需求量的异同点。分别计算各个品类下线上和线下的需求量品均值如下：

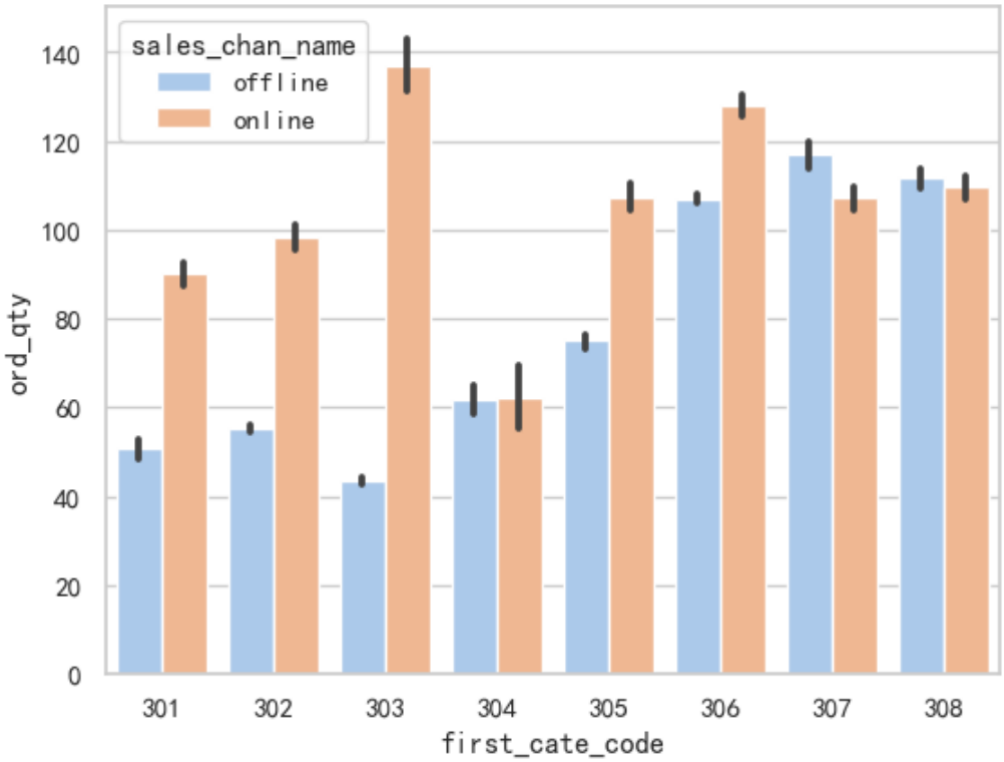

图 7：不同大类产品线上、线下需求量平均值

从上图中可以看出，不大部分大类产品平均线上需求量大于线下，其中大类 301，302，303 线上需求量远大于线下需求量，而大类 304 线上线下需求量基本持平，只有大类 307、308 两大类产品的平均线下需求量大于线上。

再研究不同品类产品的需求量面对促销时有何异同点。

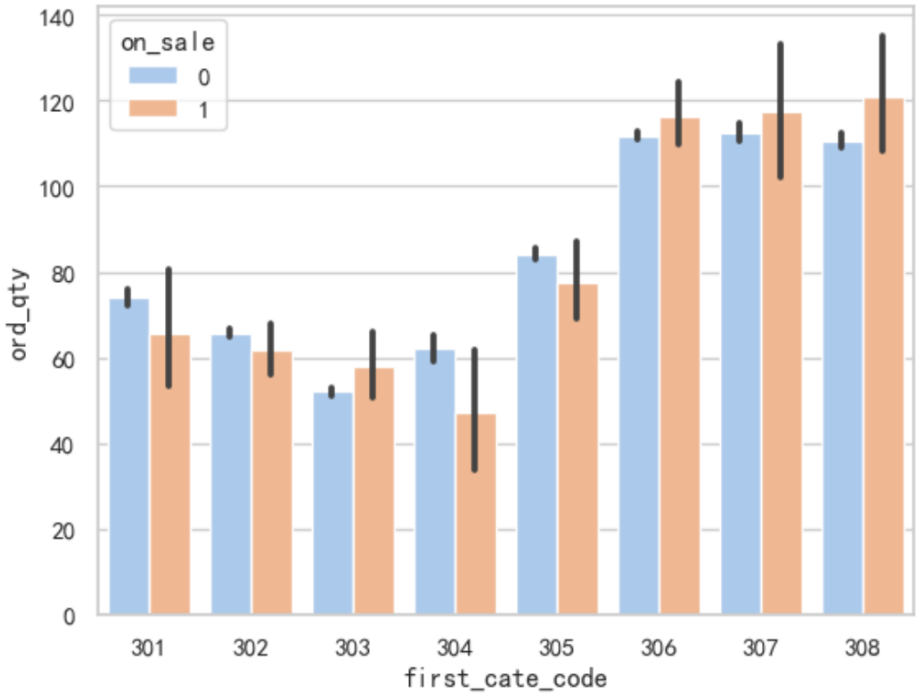

图 8：不同大类产品促销日和非促销日需求量平均值

由图可得，大部分产品在促销日的平均需求并无明显变化，即促销日和非促销日的平均需求差异不大。大类产品 304 在非促销日时的平均需求相比于促销日明显提高，而大类产品

306、307、308 在促销日时的平均需求较非促销日有略微提升，但提升不大。

4.2.3 不同销售方式与需求量

为了分析不同销售方式对产品需求量的影响，本文对线上、线下需求量作核密度图如下：

图 9：线上、线下需求量核密度图

从核密度图中可以看出，线下销售方式下的产品需求量分布相对于线上销售方式更加集中，呈现出一个明显的峰态；而线上销售方式下的产品需求量分布比较平滑，没有出现明显的峰态。同时，线下销售方式下的产品需求量整体偏高，而线上销售方式下的产品需求量整体偏低。

4.2.4 时间段与需求量

为了分析不同时间段对产品需求量的影响以及不同时间段下产品需求量的特性，本文将每月划分为月初、月中和月末三个时间段，并对附件中训练数据的时间数据根据此三个时间段打上标签（period）。其中，每月 1-10 号为月中，11-20 号为月中，21-30 号为月末。每月的月初、月中和月末组成一个时间周期。

表 3：时间段划分

日期（每月）	时间段	period
1-10 号	月中	begin
11-20 号	月初	mid
21-30 号	月末	end

本文分别计算每个大类产品月初、月中、月末的需求量平均值如下：

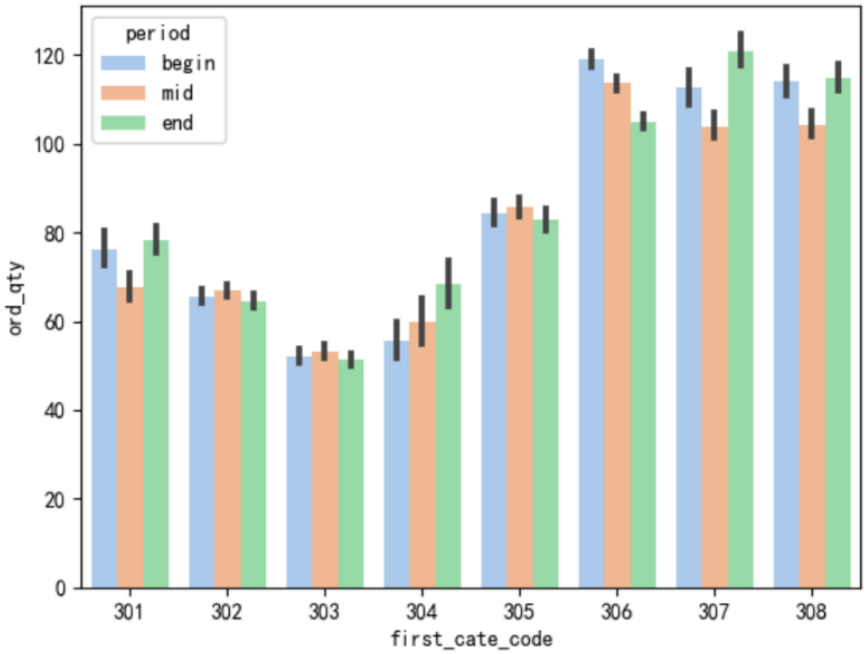

图 10：每个产品大类不同时间段下的需求平均值

从上图中可以看出，每个时间段的需求在不同大类产品下有不同的表现形式。有些大类产品的需求量有明显的时间周期性，如大类 301、307 和 308 在月初和月末的平均需求量大于月中，即在一个时间周期内需求量先上升后下降再上升，大类 306 的平均需求在一个时间周期内呈下降趋势，而大类 304 的平均需求在一个时间周期内呈上升趋势；有些大类产品则没有明显的时间周期性，如大类 302，303 和 305，在一个时间周期内平均需求平稳，无明显波动。需求的时间周期性受不同产品自身特性影响，在不同大类产品下表现不同。

为探究不同时间段内需求量的趋势变化，本文随机选取了几个月并对每月不同时间段的需求进行趋势分析。

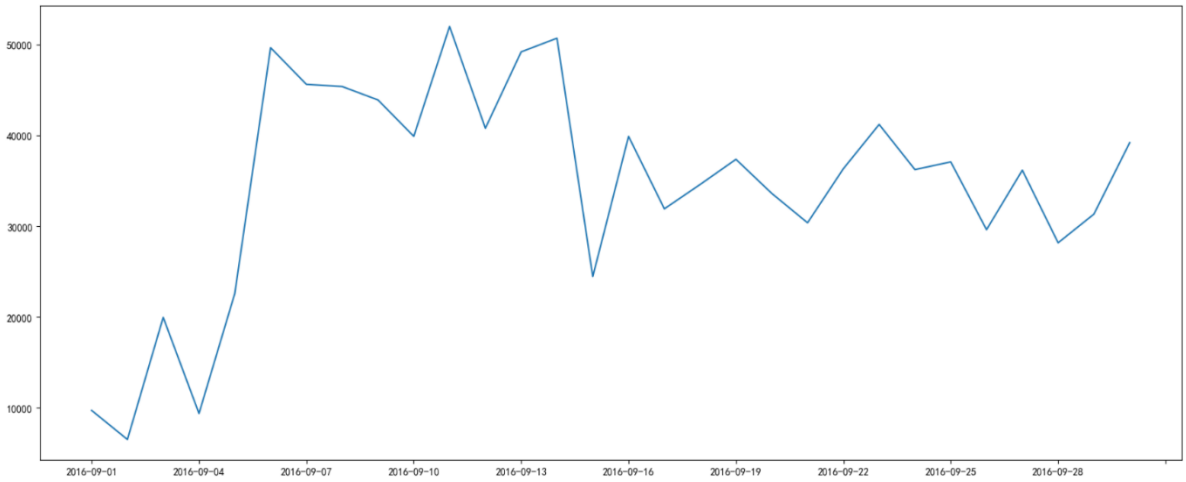

图 11：2016 年 9 月需求量变化曲线

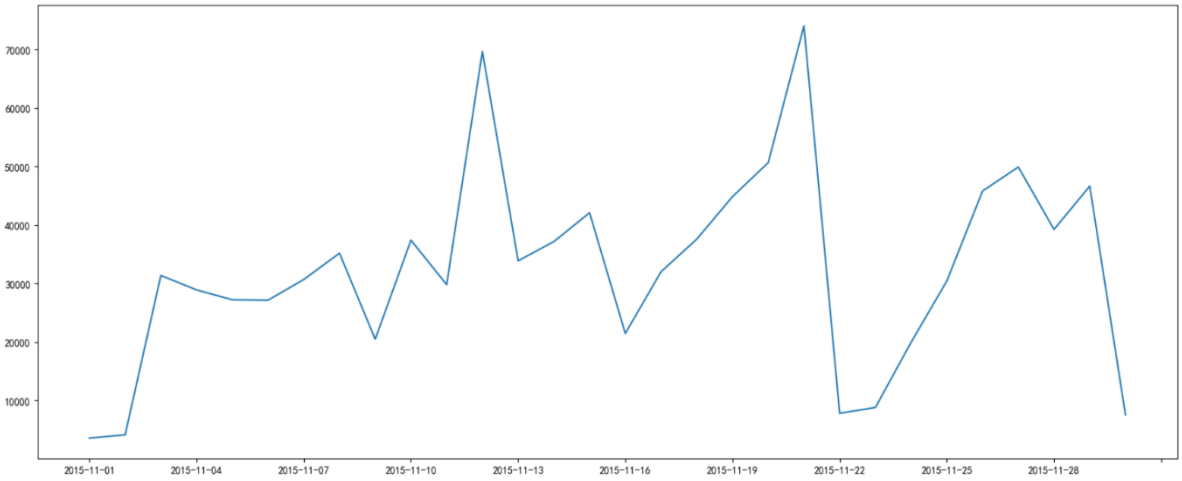

图 12：2015 年 11 月需求量变化曲线

一般而言，在月初因为薪资入账等原因，消费者购买力比较强，需求量相对较高；在月中，消费者购买动力逐渐降低，需求量有所下降；到了月末，消费者可能已经花光了部分资金或者考虑到下一个月的开销，需求量再次有所下降。此外，不同时间段内的需求量具有一定的特性，例如在一些特殊节假日或者促销活动期间，消费者的购买欲望会更加强烈，需求量上升；而在传统淡季，需求量则可能出现下降趋势。

4.2.5 节假日与需求量

中国法定节假日包括元旦、春节、清明节、劳动节、端午节、中秋节和国庆节。根据国务院对于节假日公休安排的通知，本文将节假日定义为公休安排中的假期，并对附件中的训练数据打上标签（is_holiday: 其中 1 为节假日，0 为普通日期）。附件中的训练数据无春节期间的数据，故附件数据中包含的节假日如下：

表 4：2015 年 9 月 1 日至 2018 年 12 月 20 日节假日统计

年份	节假日	日期（年/月/日）
2015	中秋节	2015/9/27
	国庆节	2015/10/1 - 2015/10/7
2016	元旦	2016/1/1 - 2016/1/3
	清明节	2016/4/2 - 2016/4/4
	劳动节	2016/4/30 - 2016/5/2
	端午节	2016/6/9 - 2016/6/11
	中秋节	2016/9/15 - 2016/9/17
	国庆节	2016/10/1 - 2016/10/7
2017	元旦	2016/12/31 - 2017/1/2
	清明节	2017/4/2 - 2017/4/4
	劳动节	2017/4/29 - 2017/5/1
	端午节	2017/5/28 - 2017/5/30
	中秋节	2017/10/4
	国庆节	2017/10/1 - 2017/10/8
2018	元旦	2017/12/30 - 2018/1/1
	清明节	2018/4/5 - 2018/4/7

	劳动节	2018/4/29 - 2018/5/1
	端午节	2018/6/16 - 2018/6/18
	中秋节	2018/9/22 - 2018/9/24
	国庆节	2018/10/1 - 2018/10/7

本文将节假日与非节假日各大类产品订单平均需求量进行统计分析，结果如下图所示：

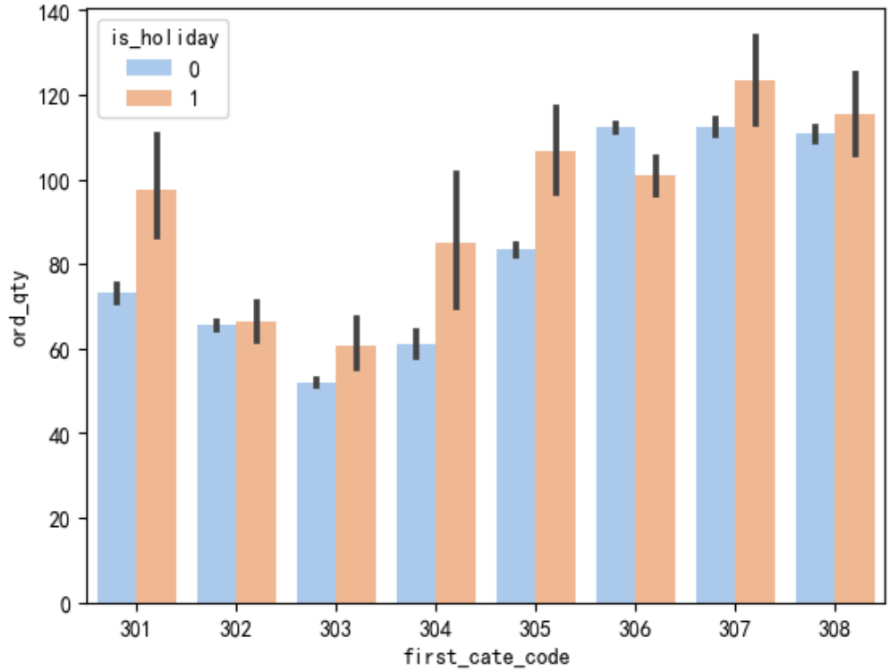

图 13：节假日与非节假日各大类产品平均订单需求量

绝大部分大类产品节假日期间的平均需求量都高于非节假日，因此可以看出，节假日期间消费者的购买力会更加集中和突出。为进一步研究节假日对需求量的影响，我们选取 2018 年的节假日数据并向前、向后截取一定的天数对其需求量进行比对分析。

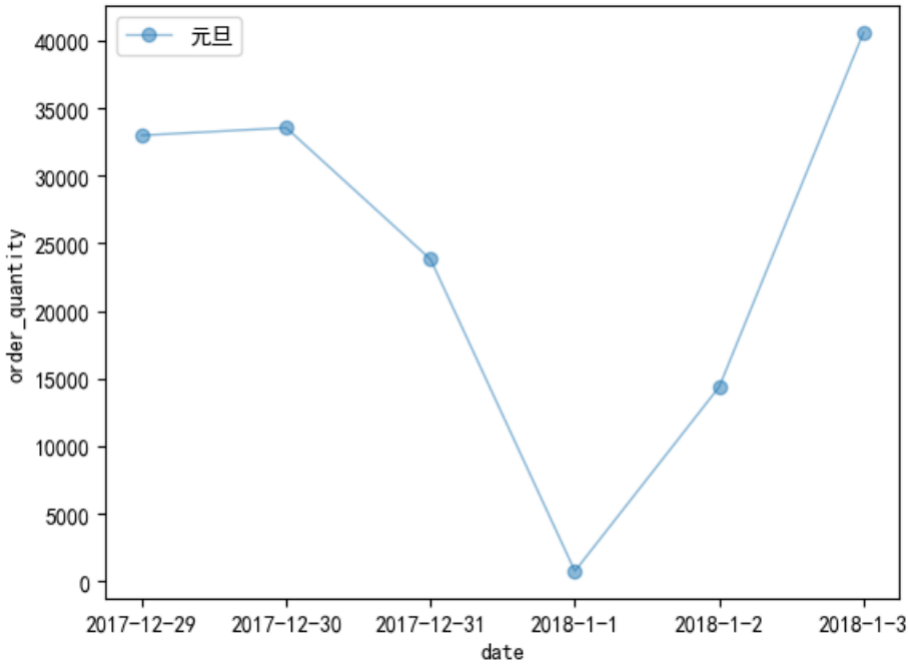

图 14：元旦前后需求量对比

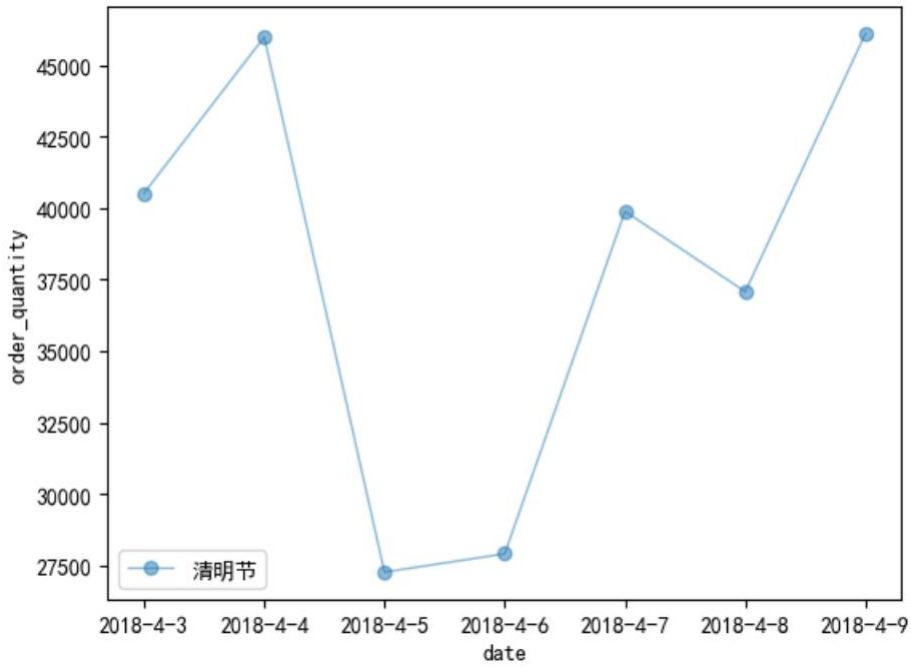

图 15：清明前后需求量对比

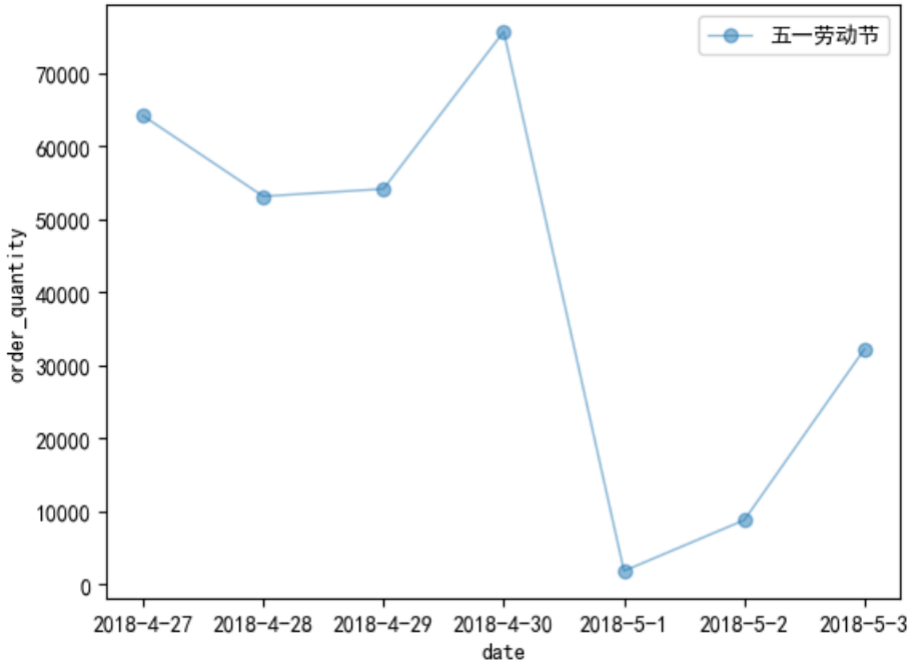

图 16：五一劳动节前后需求量对比

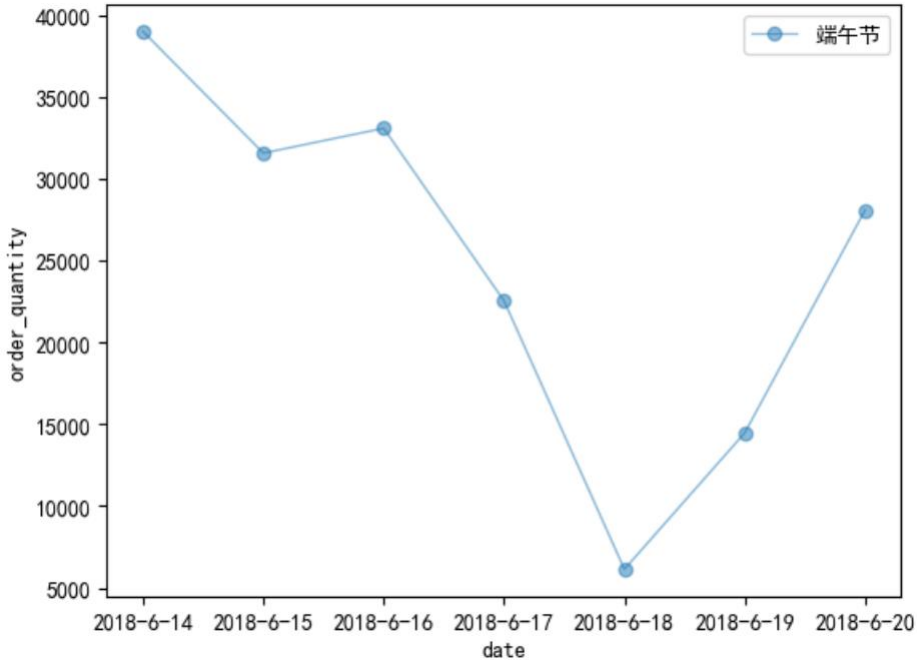

图 17：端午节前后需求量对比

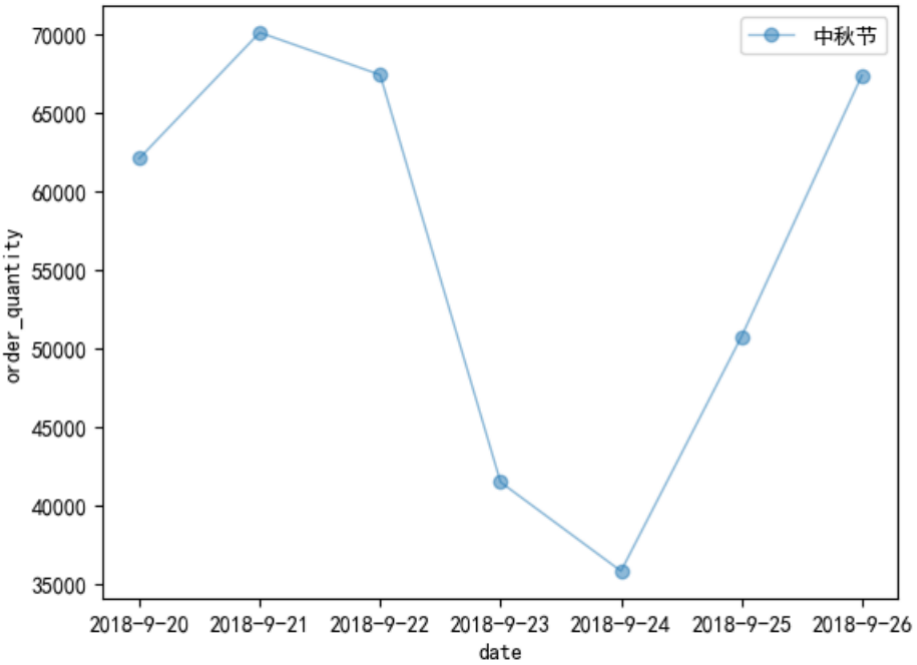

图 18：中秋节前后需求量对比

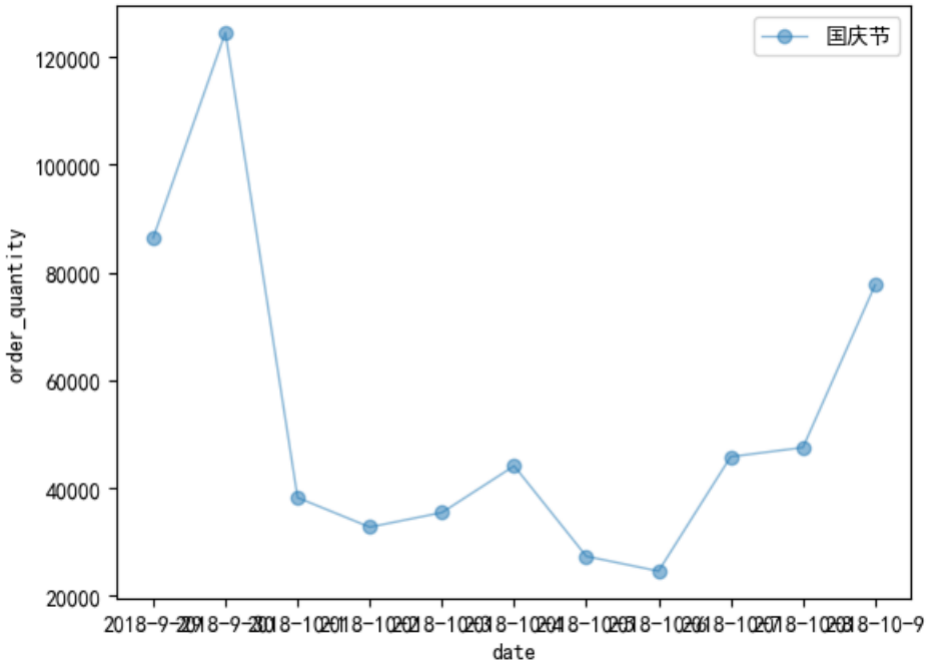

图 19：国庆前后需求量对比

通过对各个节假日前后的需求量对比，我们发现在部分节假日（如清明节），在节假日前需求量会大幅度回落，在节假日期间，需求量逐步提升；而又有部分节假日（如五一），节假日前的需求回落一直持续到节假日结束，而在节假日结束后需求量才逐步攀升。

4.2.6 促销与需求量

本文选取了“38 节”“618”“双 11”“双 12”四个具有代表性的促销日，研究促销对需求量的影响。

进一步，本文用独立样本 MannWhitney 检验分析促销（定类变量）与需求量（定量变量）之间有无明显差异。分析步骤如下：

step1 根据定类变量（促销）对定量字段（需求量）进行分组，分别检验其正态性，查看数据的总体分布是否呈现正态性分布，若检验通过，建议采用独立样本 T 检验。

step2 查看 MannWhitney 检验表，若呈现显著性，可以查看中位数对差异进行分析，反之则表明不呈现差异性。

step3 若独立样本 MannWhitney 检验呈现显著性，也可借助效应量化分析对差异性进行量化分析。

下表展示了订单需求量（ord_qty）的描述性统计和正态性检验的结果，包括均值、标准差等，用于检验数据的正态性。

表 5：变量 ord_qty 正态性检验结果

变量名	样本量	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
ord_qty	597694	91.651	199.843	10.981	373.474	0.413(0.000***)	0.325(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

需求量（ord_qty）样本 N≥5000，采用 K-S 检验，显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，因此数据不满足正态分布，可以进行独立样本 MannWhitney 检验。

表 6：MannWhitney U 检验结果表

变量名	变量值	样本量	中位数	标准差	统计量	P	中位数值差值	Cohen's d 值
ord_qty	0	588428	29	200.045	2707796678	0.264	0	0.013
	1	9266	29	186.553				
	合计	597694	29	199.843				

注：***、**、*分别代表 1%、5%、10%的显著性水平

由上表可知，促销标签 0、1 在订单需求量上的中位数都为 29；检验结果 P 值为 0.264，因此统计结果不显著，促销标签 0、1 在订单需求量上不存在显著差异；其差异幅度 Cohen's d 值为：0.013，差异幅度非常小。因此是否为促销日在订单的需求量上不存在显著差异。

考虑到促销标签仅为促销日当天，本文对促销前后的需求量进行对比分析。

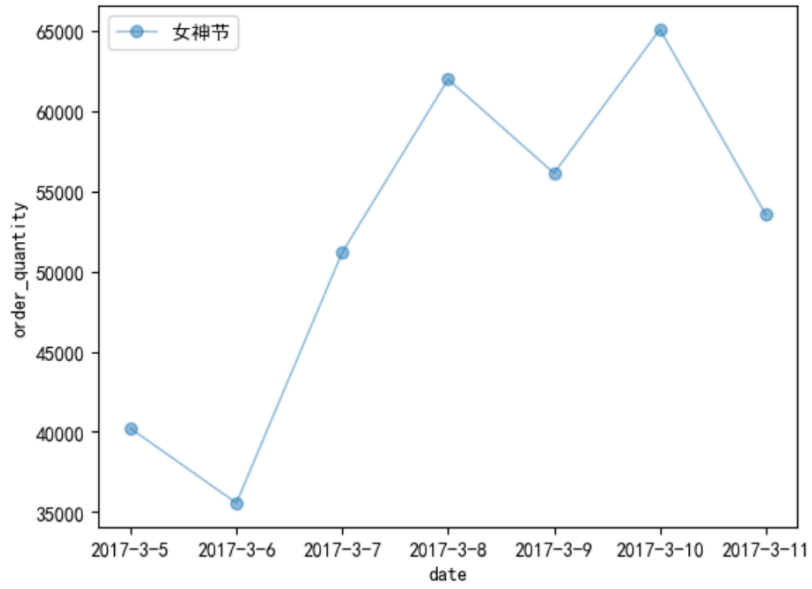

图 20：“3.8 女神节”促销活动前后需求量对比

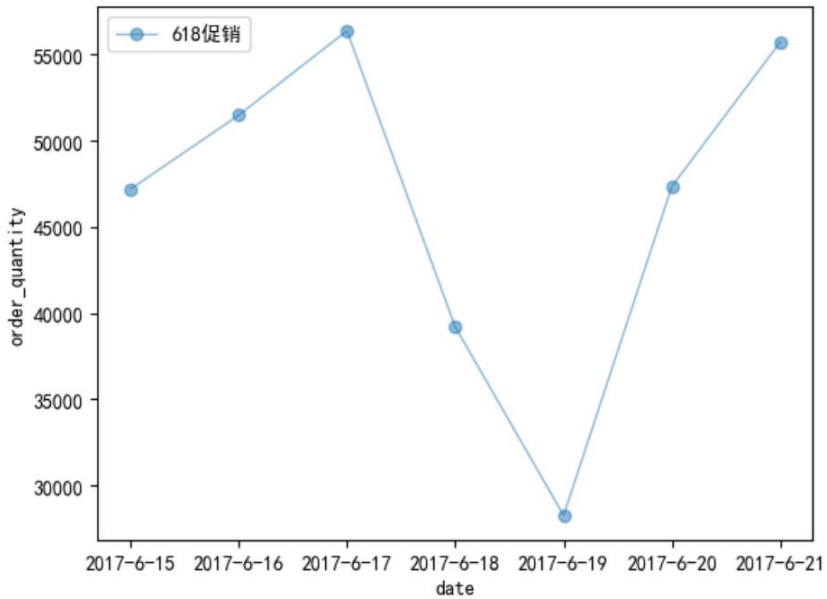

图 21：“6.18”促销活动前后需求量对比

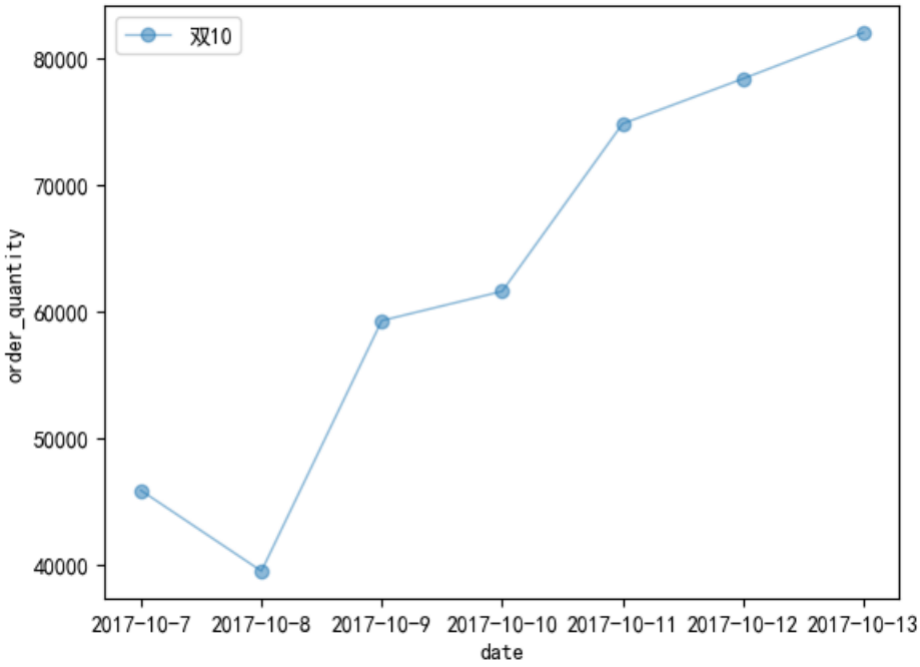

图 22：“双 10”促销活动前后需求量对比

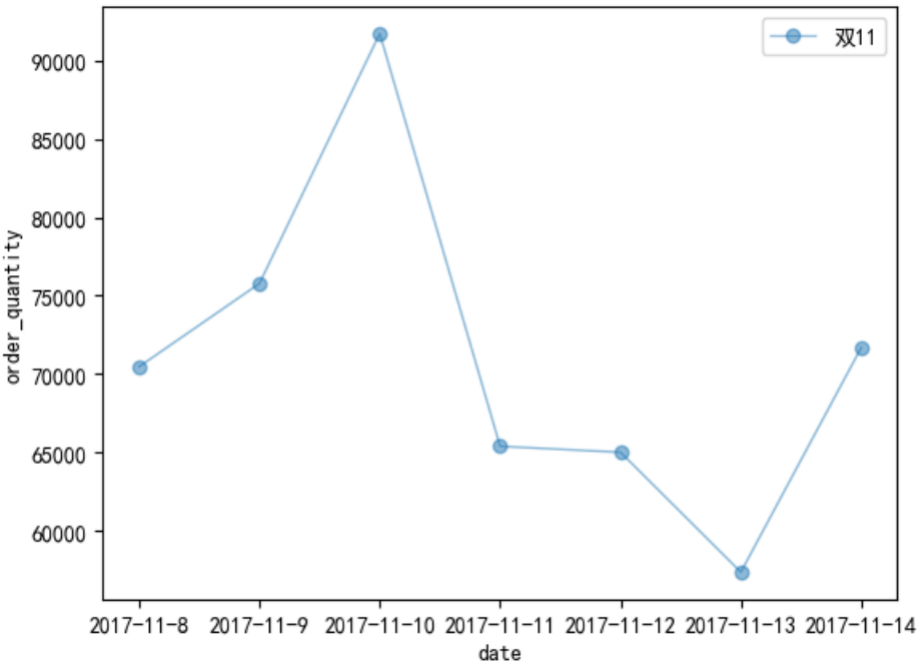

图 23：“双 11”促销活动前后需求量对比

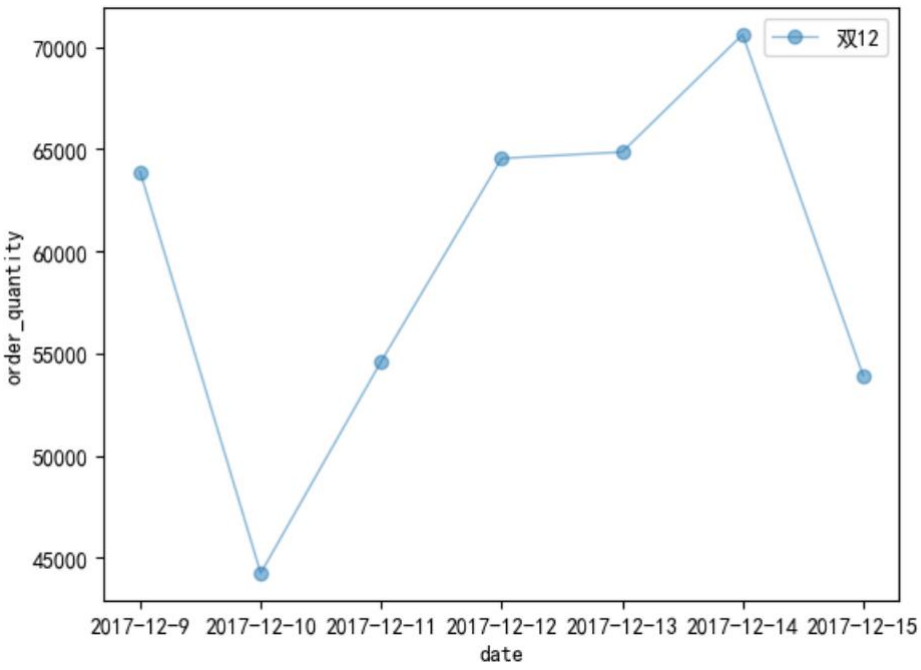

图 24：“双 12”促销活动前后需求量对比

通过上述促销活动前后需求对比分析，我们发现需求量在大部分促销日前有明显的提升，而在促销日时需求量一般小于促销前。猜测目前促销活动会提前进行“预热”，刺激消费者的购买欲望，增加需求量。而到促销日时，促销活动结束，需求回落。

4.2.7 季节与需求量

本文将季节按照二十四节气中的立春、立夏、立秋、立冬来进行划分，并对附件中训练数据打上标签（season：其中 Spring 为春，Summer 为夏，Autumn 为秋，Winter 为冬）。附件数据中的季节划分如下：

表 7：2015 年 9 月 1 日至 2018 年 12 月 20 日季节划分

季节	日期（年/月/日）
Spring	2016/2/4-2016/5/4；2017/2/3-2017/5/4；2018/2/4-2018/5/4
Summer	2016/5/5-2016/8/6；2017/5/5-2017/8/6；2018/5/5-2018/8/6；
Autumn	2015/9/1-2015/11/7；2016/8/7-11/6；2017/8/7-11/6；2018/8/7-11/6
Winter	2015/11/8-2016/2/3；2016/11/7-2017/2/2；2017/11/7-2018/2/3；2018/11/7-2018/12/20

本文分别计算每个大类产品春季、夏季、秋季、冬季的需求量平均值如下：

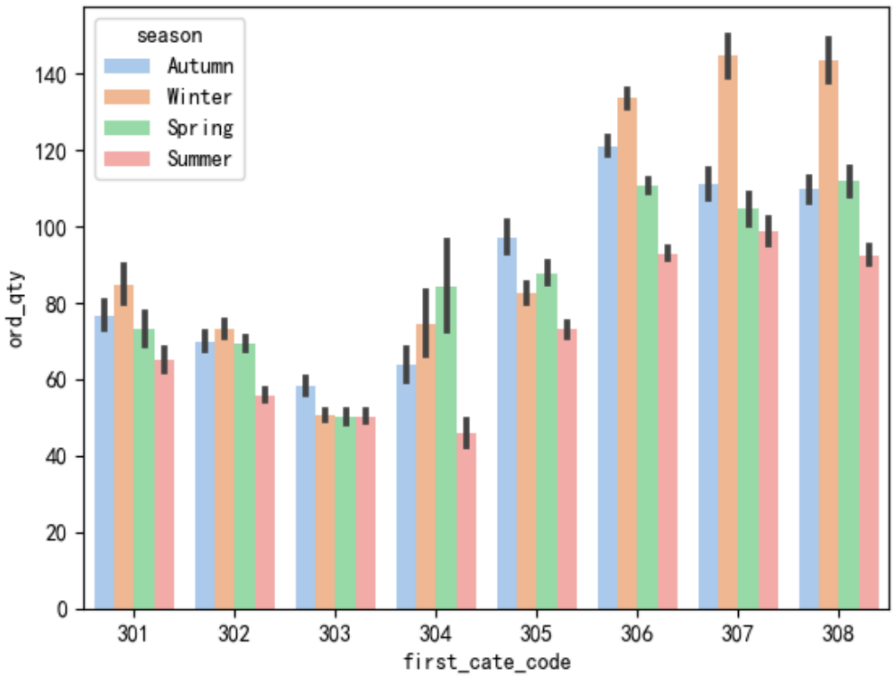

图 25：各大类产品季度需求平均值对比

总体而言，需求量在不同的季节具有明显的波动性，但受不同类型产品本身特性影响，季节性在不同大类产品中有不同的表现。因此，本文将对不同季节的总体需求量进行进一步的分析。由于附件中训练数据在 2016 年和 2017 年是完整的，本文选取 2016 年和 2017 年的需求量进行季节分析，四个季节组成一个年周期。

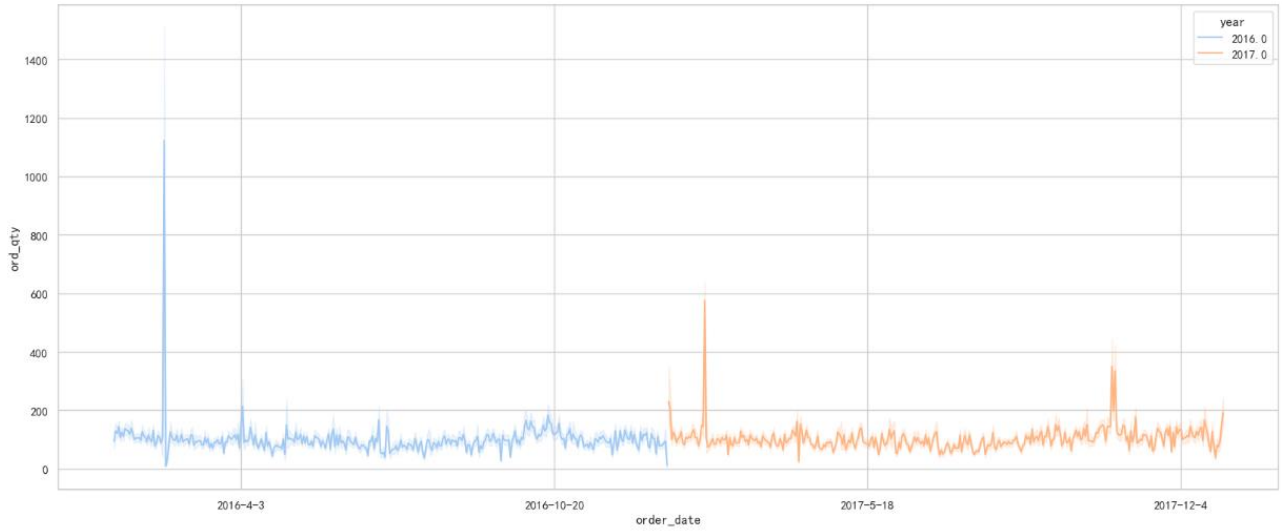

图 26：2016-2017 年需求量曲线图

由图可知，需求量与季节有明显的关系，需求量整体以季节呈年周期变化。以年为周期，需求量有两次明显提升，分别是冬季的大幅度提升和秋季的小幅提升，但冬季和秋季的需求量波动也较大。相比之下，夏季和春季的需求量较为平稳。进一步地，我们对不同季节的需求量作如下描述性统计量分析：

表 8：不同季节的需求量描述性统计量分析

season	均值	中位数	标准差	最大值	最小值	总量
Autumn	98.017	30.000	234.056	16308.000	1.000	14788609.000

Spring	92.091	33.000	178.065	5998.000	1.000	11489532.000
Summer	77.790	22.000	169.849	8035.000	1.000	15352485.000
Winter	105.444	36.000	217.297	9874.000	1.000	13148331.000

其中不同季节的平均需求量如下图：

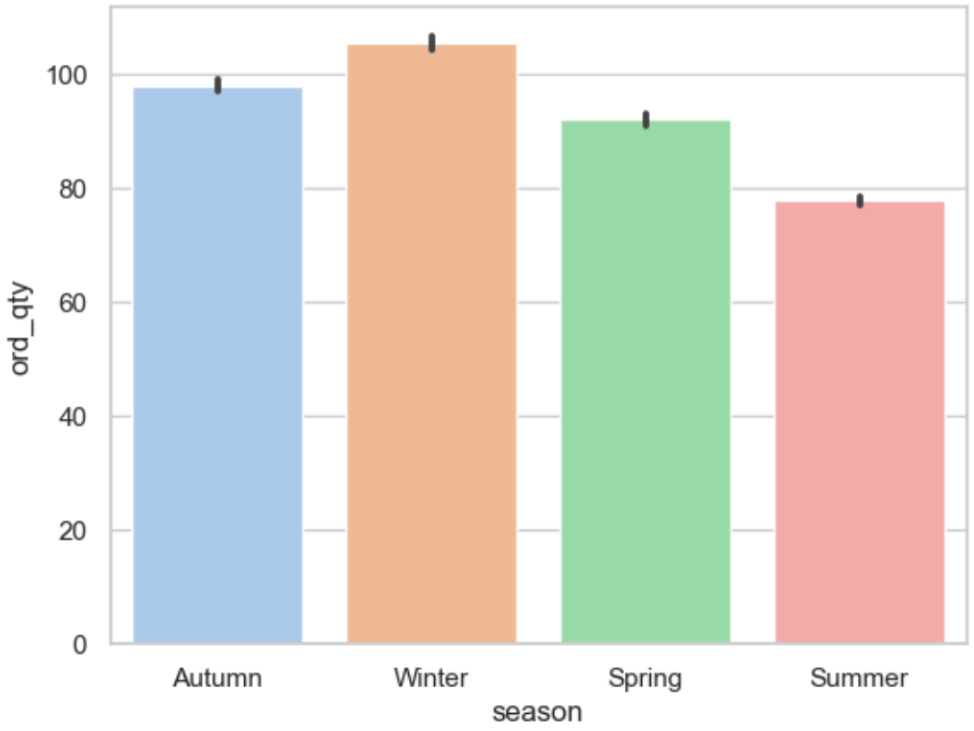

图 27：不同季节平均需求量

由图可得，冬季的平均需求量最大，夏季的平均需求量较少，秋冬的需求量大于春夏的需求量。

不同季节的需求量标准差如下图：

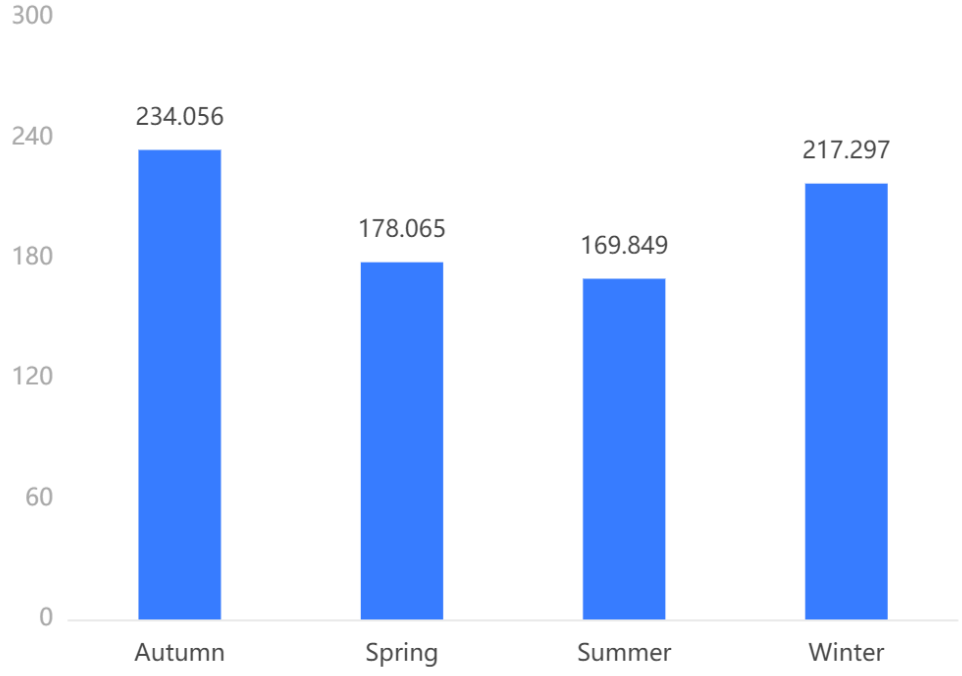

图 28：不同季节需求量标准差

标准差能够反映数据的波动幅度，由图可得，秋季和冬季的标准差大于春季和夏季，因而秋冬时需求量的波动程度大，这与图 26 反映的波动幅度一致。

考虑到可能是大部分的促销都集中在秋冬季，且秋冬季包含了春节这个大长假，极大程度地刺激了冬季的总体需求，也同时造成了需求的波动。

不同季节的需求量总额如下图：

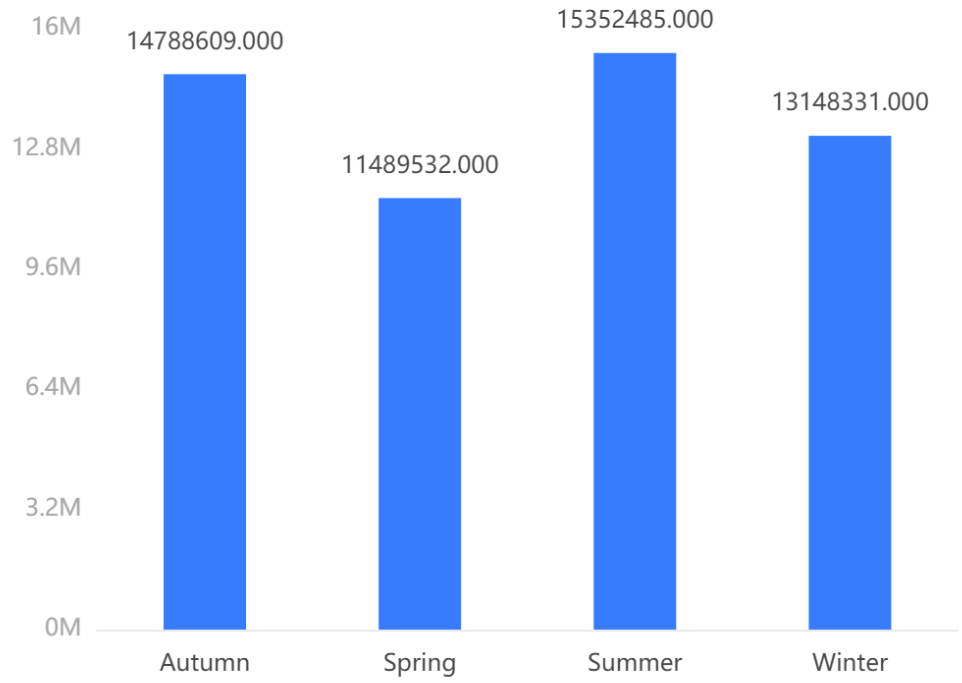

图 29：不同季节需求量总额

虽然秋冬有两次较为明显的需求量提升，但夏季的总体需求量是四季中最大的，考虑到可能是夏季中有多个小长假分布。

综上，不同季节会对产品需求量产生影响。因此，在进行市场预测和计划时，需要考虑到季节性因素的影响，以便更好地了解市场需求。

五、 需求预测

5.1 数据预处理

首先从附件中的训练数据（order_train1.csv）中获得运算需要的数据，再从附件中的预测数据（predict_sku1.csv）中获得需要预测的 sales_region_code 和 item_code，按照这两个指标将数据进行聚类分割。

5.1.1 缺失值处理

附件中的训练数据（order_train1.csv）提供了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，排列成的时间序列数据应当具有连续性。按照产品编码和销售区域编码对数据进行聚类分割后，观察到部分产品数据量无法构成连续的时间序列，且有的产品数据量过少，影响后续预测的准确性。因此，需要对缺失值进行填补以保证时间序列的连续性，进而保证序列的时间特性和周期性的完整。由于不同年份同一日期时的订单需求量之间存在一定的相似性，且间隔相差不大的相邻时间点之间的订单需求量也存在连续趋势性，因此本文对各产品时间序列上的数据缺失值分两种情况进行填补：

（1） 时间序列上存在少量缺失值，用线性差值法：

$$y_i = \frac{y_{i+n} + y_{i-n}}{2n}, n = 1,2,3 \quad (17)$$

y_i 为缺失日期的订单需求量数据，y_{i+n} 和 y_{i-n} 分别是缺失值相邻的前 n 天和后 n 天的订单需求量数据，本文规定连续缺失 10 天以内的产品需求量都可用此公式修正。

（2） 时间序列上存在大量缺失值，用年周期平均法：

由于订单需求量数据具有年周期性，不同年份相同日期的需求量之间具有相似性，因此可以寻找缺失值日期对应的相邻几年同日期的数据，对其求平均进行填补。本文规定连续缺失大于 10 天的数据用年周期平均法。

5.1.2 重复数据处理

经过描述性分析我们发现同一日相同地点同一货品会有多条数据，但均为有效数据。因此，本文将该类数据的需求量进行求和合并，将该类数据的价格用日平均值进行替代。以产品编号为 20021 在 2016-7-17 的数据为例：

表 9：产品编号为 20021 在 2016-7-17 的数据

order_date	sales_region_code	item_code	first_cate_code	second_cate_code	sales_chan_name	item_price	ord_qty
2016-7-17	101	20021	305	412	offline	1013	106
2016-7-17	101	20021	305	412	offline	1008	154

进行求和合并后：

表 10：产品编号为 20021 在 2016-7-17 经处理后的数据

order_date	sales_region_code	item_code	first_cate_code	second_cate_code	sales_chan_name	item_price	ord_qty
2016-7-17	101	20021	305	412	offline	1010.5	260

5.1.3 归一化处理

在使用模型进行预测时，通常要对输入的特征数据进行归一化处理，这样做的好处是：

（1）避免各特征与目标值的量纲不同对预测性能造成影响；

（2）加快梯度下降的速度；

（3）使数据更方便模型处理。

5.2 模型的建立、求解与误差分析

本文以 2019 年 1 月 1 日前后划分比例大致为 7：3 的训练集与测试集，分别建立 LSTM，随机森林，决策树，XGBoost 模型，可视化预测值与真实值的拟合图以及误差曲线图，从而清晰直观地看出模型的拟合效果，再将 MSE，RMSE，MAE，MAPE 作为评价指标分析各模型的预测精度，最后选用效果较好的模型分别按天、周、月的时间粒度测来预测未来三个月的月需求量，并分析不同时间粒度对预测精度的影响。

5.2.1 长短期记忆递归神经网络（LSTM）

本文选用多变量 LSTM 模型进行需求预测。LSTM 模型的优势在于它能够解决传统循环神经网络模型（例如 RNN 模型）存在的长期依赖和短期记忆问题。具体来说，LSTM 模型通过引入门机制来控制信息的输入、遗忘和输出，从而能够更有效地保存和处理长期依赖性信息，使得在处理时间序列数据时能够取得更好的效果。

首先将所有处理过后的指标数据投入到多变量 LSTM 模型当中，通过不断调整其中参数，以获得最优参数的 LSTM 模型。处理后的指标数据包括 sales_region_code（销售区域编码）、item_code（产品编码）、first_cate_code （产品大类编码）、second_cate_code （产品细类编码）、sales_chan_name （销售渠道名称）、item_price （产品价格）、is_holiday（是否是节假日）、season（季节）、period（时间段）、on_sale（是否是促销日），其中 is_holiday、season、period、on_sale

为本文在附件训练数据上新增的标签。

（1）模型搭建

本文借助于 python 自带的 TensorFlow 库实现 LSTM 模型的构建。

step1 定义各类相关指标，设置 learning rate, num_layer, size_layer, forget_bias 等系数。

step2 定义神经网络的基本框架。借助于 LSTMCell 函数构建单层神经网络，并按照 size_layer 的大小构建全神经网络。

step3 利用 placeholder, dropoutwrapper, dynamic_rnn 等 TensorFlow 函数完成输入、输出、过拟合跳出、忘记门、更新门、输出门等的设置。

为减少模型结果的特殊性以影响模型选择，对于同一参数，我们通过 5 次模拟结果取平均相对误差以选择最优参数。

表 11：LSTM 模型调参

参数名称	调参范围	调参结果
模拟次数 simulation_size	[4,10]	5
LSTM 单元个数 num_layers	[1,3]	1
一层网络中的神经元个数 size_layer	[128,512]	128
时间步长 timestamp	[3,10]	5
迭代次数 epoch	[50,500]	150
跳出参数 dropout_rate	[0.4,0.9]	0.8
学习率 learning_rate	[0.001,0.5]	0.01

通过对比各参数下 LSTM 模型的平均相对误差，我们不难发现在忘记偏置为 1.0，LSTM 单元数为 1 时多变量 LSTM 模型取得最优预测结果。

表 12：各参数下的平均相对误差

平均相对误差		
forget bias=1.0 LSTM 单元数=1	forget bias=0.7 LSTM 单元数=2	forget bias=1.0 LSTM 单元数=14
0.1522	0.2119	0.2888
forget bias=0.4 LSTM 单元数=7	forget bias=0.4 LSTM 单元数=14	
0.2264	0.3072	

（2）数据输入

对于训练集，对其迭代每一天，每次获取 1 时间步长单位的训练数据加入到训练集当中，以获取足够数据进行预测。用前 t-时间步长天预测未来 1 天的方式预测训练集的每一天，以便获得更加优质的模型。

对于测试集，从 2019 年 1 月 1 日以 1 步长单位进行迭代，每次同样获取 1 时间步长数据加入到测试集当中，测试时可直接通过移动数据窗口获取强 t-时间步长天的数据进行预测。

（3）模型迭代，Sanity Check 和结果可视化

对于每一个固定参数，我们都会进行 5 次模拟以保证其结果的准确性与稳定性。在模型迭代结束后，我们将会对这 5 次结果进行 Sanity Check。对于预测曲线中的其中一个数值小于实际结果中的最小值，同时其中一个数值大于实际结果最大值的 2 倍，那么我们就认为其没有通过 Sanity Check，并剔除该曲线结果。

在获得最终结果曲线后，我们将计算所有曲线的平均准确度或者平均相对误差，并将通过 Sanity Check 的曲线与真实值曲线绘制在同一个图中以便更加清晰直观的观察预测结果。

（4）模型评估

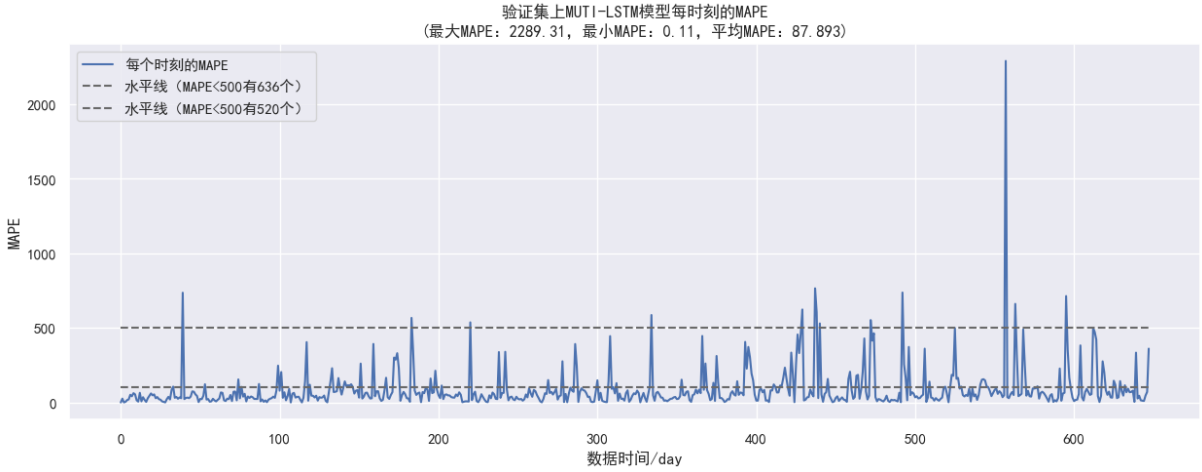

图 30：验证集上 MUTI—LSTM 每时刻的 MAPE

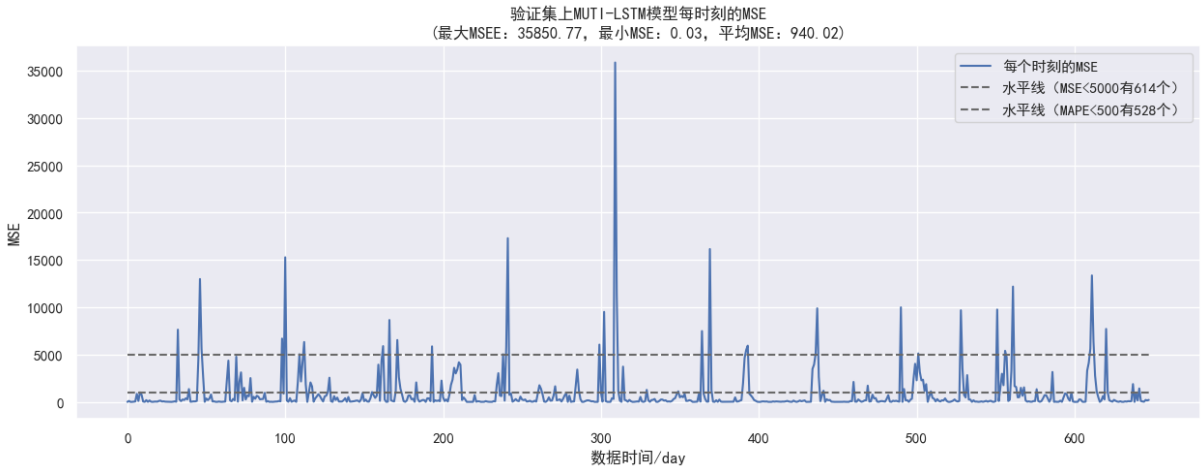

图 31：验证集上 MUTI—LSTM 每时刻的 MSE

与真实值对比，最大的 MAPE 是 228931%，最小 MAPE 是 11%，平均 MAPE 为 8789.3%；误差大部分集中在水平线附近，少部分时刻有较大的误差；最大的 MSE 是 3585077%，最小 MSE 是 3%，平均 MAPE 为 94002%。

5.2.2 随机森林回归（RFR）

随机森林回归通过对多个决策树进行训练和组合来提高预测准确度。对于每个决策树，它对数据的解释能力是受限的。但是随着决策树的数量增加，随机森林回归模型的拟合能力也随之提高，从而达到更好的预测效果。因此，决策树的数量的选择是模型构建的关键。

（1） 模型搭建

一般来说，决策树数量太小容易欠拟合，太大容易过拟合，所以需选择一个适中的值；再调整决策树的最大深度，控制模型的复杂度，默认为 None，即在建立子树时不会限制子树的深度。

本文确定 RFR 的模型参数如下：

表 13：RFR 模型参数

参数名	参数值
训练用时	0.114s
数据切分	0.7
数据洗牌	是
交叉验证	是
节点分裂评价准则	mse

划分时考虑的最大特征比例	None
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10
叶子节点的最大数量	50
节点划分不纯度的阈值	0
决策树数量	100
有放回采样	true
袋外数据测试	false

（2） 特征重要性

RFR 模型输出的特征重要性排序如下：

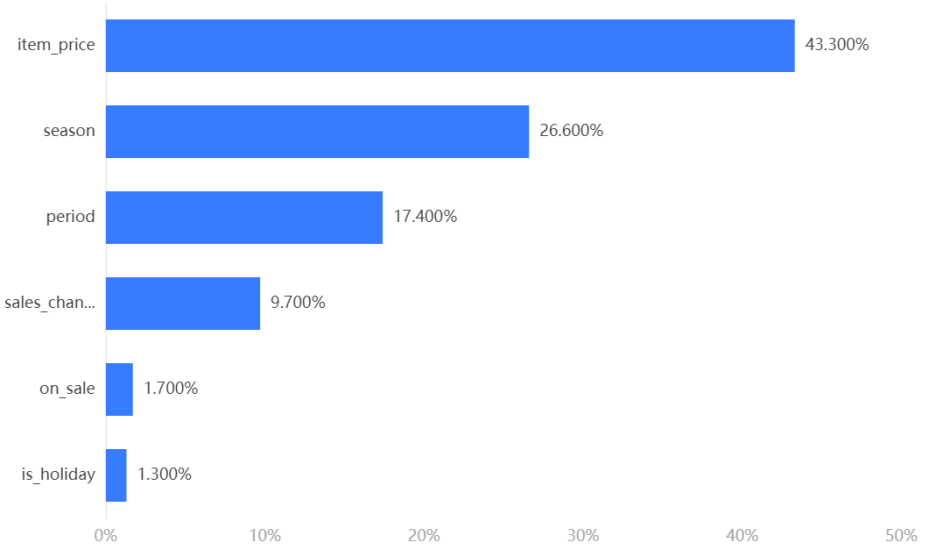

图 32：RFR 特征重要性排序

其中，产品价格的重要程度最高，节假日的重要程度最低。

（3） 模型评估

表 14：RFR 模型评估

	MSE	RMSE	MAE	MAPE	R^2
训练集	39372.5	6248.9	2887.2	85683	0.00734
测试集	45526.1	5187.7	2986.9	41586.5	0.0699

随机森林回归模型的各项误差均较大，因此预测效果不理想。

5.2.3 决策树

决策树回归模型的构建关键在于定义节点分裂准则。回归树的构建过程是递归地将数据集分成不同的区域，对每个区域内的数据进行回归分析。因此，决策树回归模型的分裂准则至关重要。

（1） 模型搭建

表 15：决策树模型参数

参数名	参数值
训练用时	0.098s

数据切分	0.7
数据洗牌	是
交叉验证	是
节点分裂评价准则	friedman_mse
特征划分点选择标准	best
划分时考虑的最大特征比例	None
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10
叶子节点的最大数量	50
节点划分不纯度的阈值	0

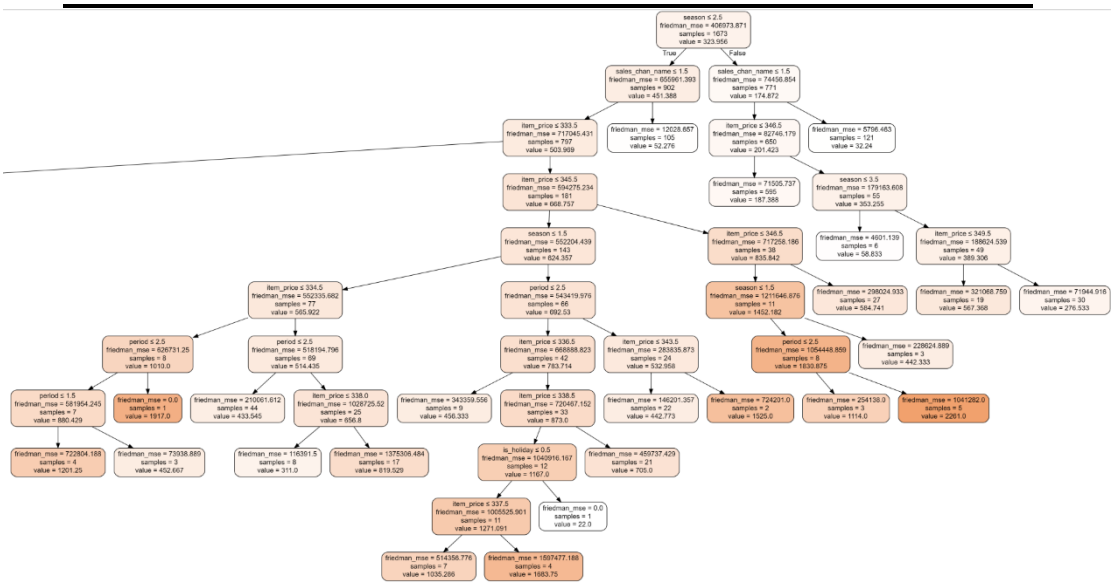

图 33：决策树结构（部分）

（2） 特征重要性

决策树模型输出的特征重要性排序如下：

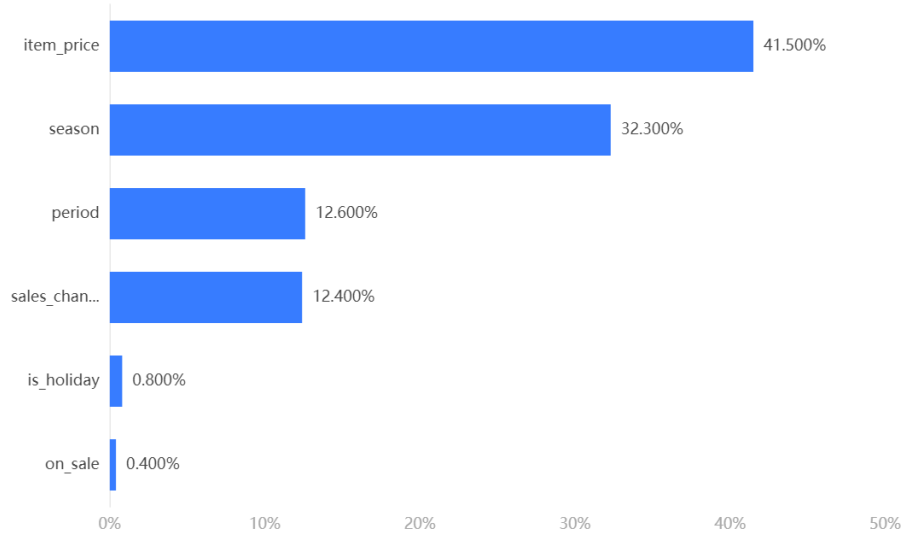

图 34：决策树模型输出的特征重要性

其中产品价格的重要性程度依旧最大。

（3） 模型评估

表 16：RFR 模型评估

	MSE	RMSE	MAE	MAPE	R^2
训练集	38854.4	5889.8	3200.4	9144.3	0.056
测试集	36612	6551.7	3419.3	40990.9	0.058

决策树回归模型的各项误差均较大，因此预测效果不理想。

5.2.4 XGBoost

（1）模型搭建

首先选择较高的学习率(learning_rate)，如 0.1，可以减少迭代用时。先确定决策树数目，再调整决策树的最大深度、样本的采样率、以及叶子节点中样本的最小权重，最后再调整学习率，确定最佳参数。

表 17：XGBoost 参数设置

参数名	参数值
训练用时	1.778s
数据切分	0.7
数据洗牌	否
交叉验证	否
基学习器	gbtree
基学习器数量	100
学习率	0.1
L1 正则项	0
L2 正则项	1
样本征采样率	0.54
树特征采样率	0.15
节点特征采样率	0.4
叶子节点中样本的最小权重	0.3
树的最大深度	10

（2）特征重要性

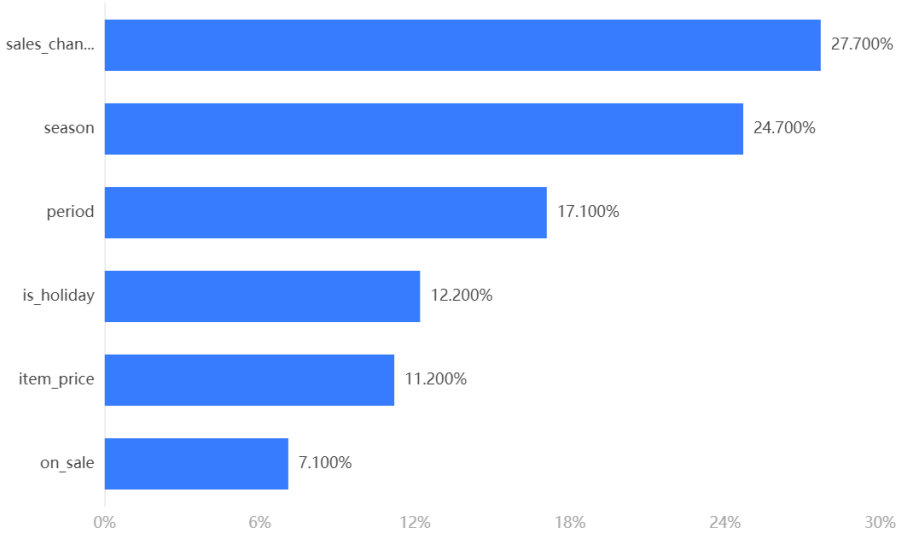

图 35：XGBoost 模型输出的特征重要性

（3）模型评估

表 18：XGBoost 模型评估

	MSE	RMSE	MAE	MAPE	R^2
训练集	21375.9	5808.7	21478	64975	0.035
测试集	31830.7	54904.	37897	49431	0.0019

XGBoost 回归模型的各项误差均较大，因此预测效果不理想。

5.3 不同时间粒度对预测精度的影响分析

通过对各个模型的误差和拟合效果进行综合分析，我们选择 LSTM 为最终预测模型。

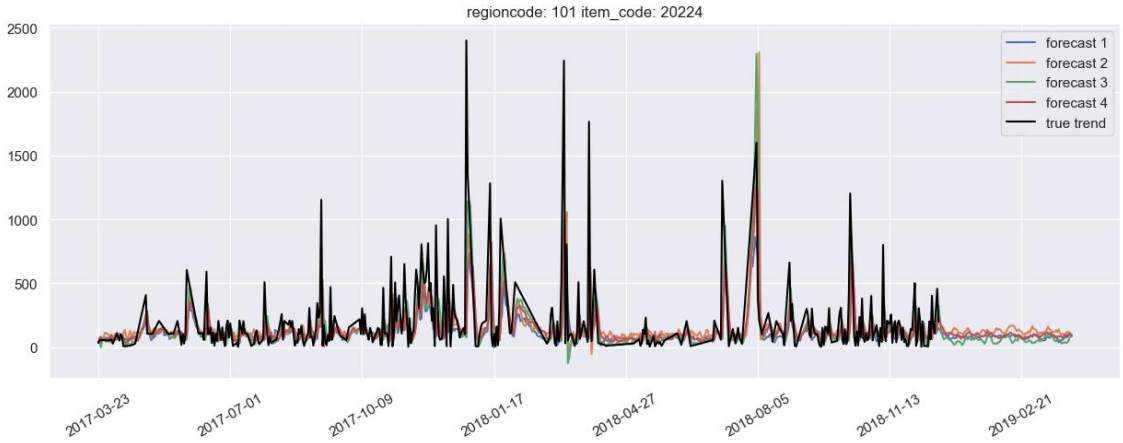

图 36：20224 产品以日为单位粒度进行预测

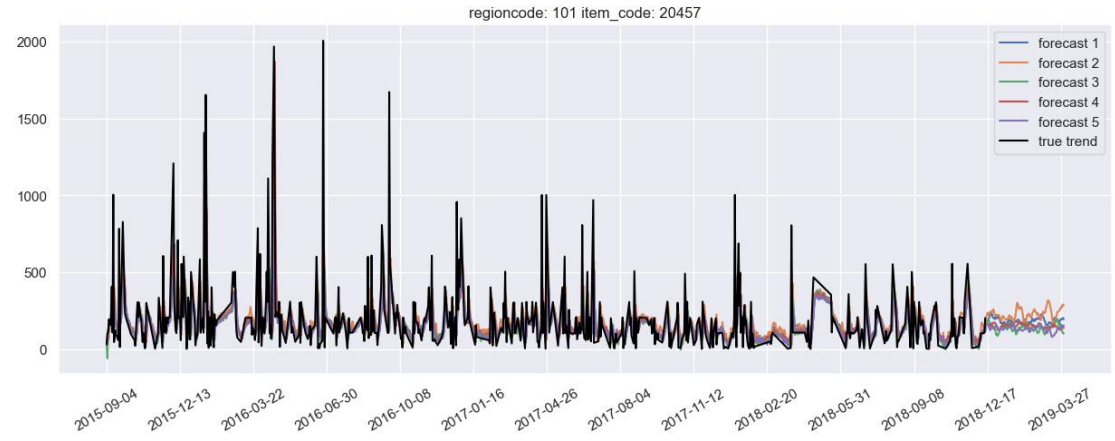

图 37：20457 产品以日为粒度进行预测

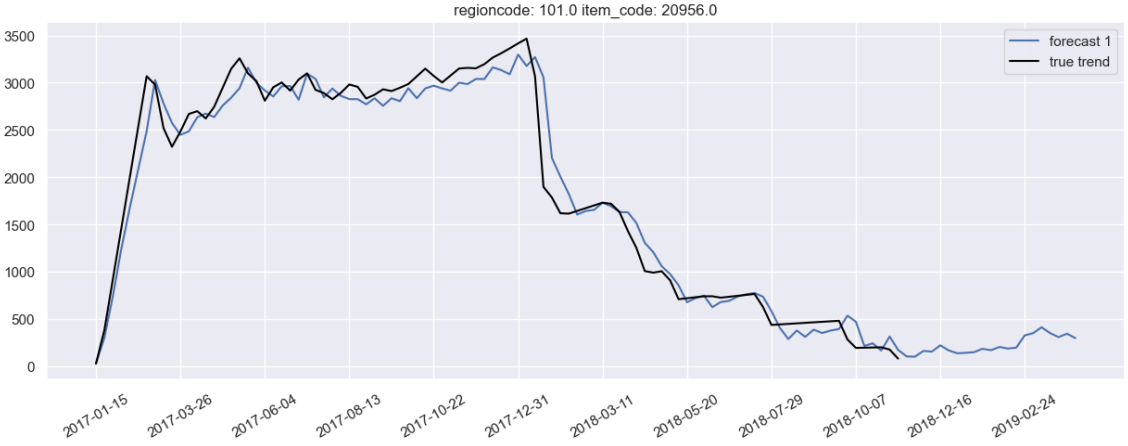

图 38：20956 产品以周为粒度进行预测

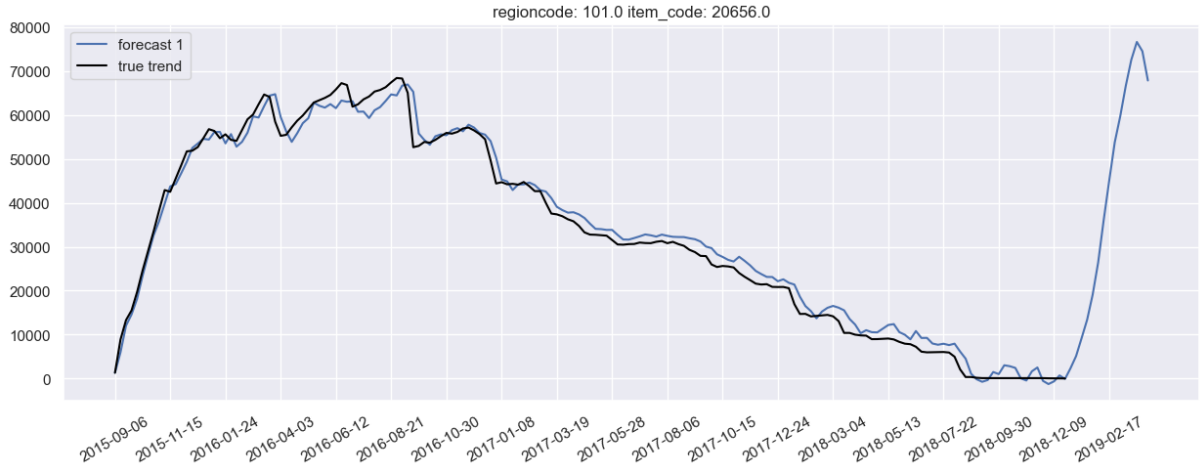

图 39：20656 产品以周为粒度进行预测

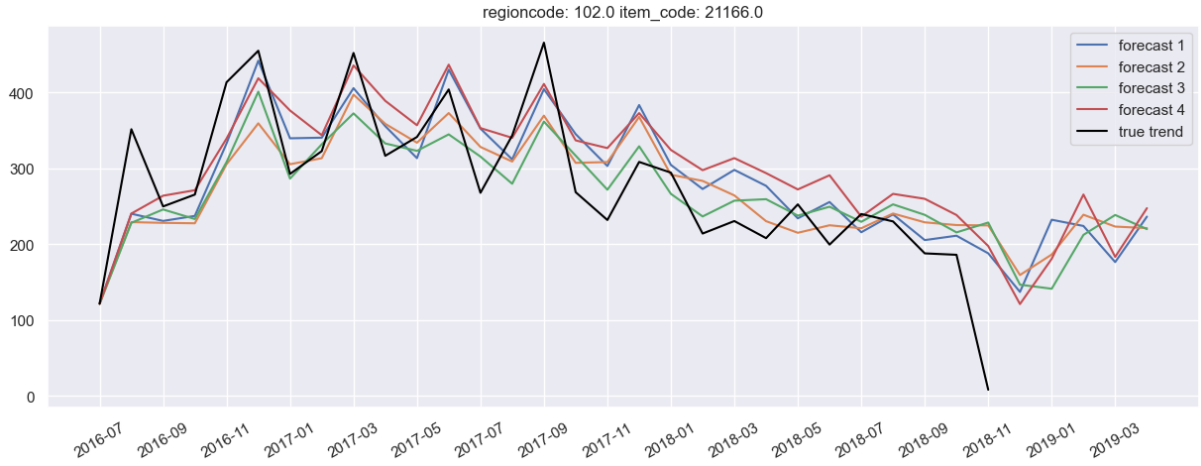

图 40：21166 产品以月为粒度进行预测

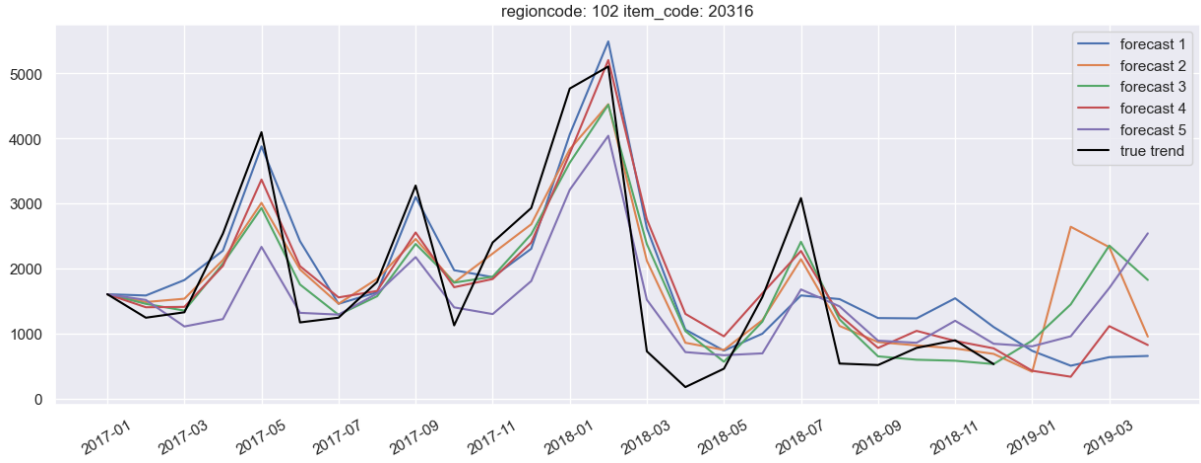

图 41：20316 产品以月为粒度进行预测

由于某些产品编码下的产品存在大量数据缺失，因此在以日为颗粒度进行预测时，由于数据的缺乏，插值所填补的数据所带来的误差会在以日为时间粒度进行预测时累积，因此以日为时间粒度进行预测相比以周、月为时间粒度存在一定程度上的精度损失。

六、 模型的优缺点与改进

6.1 模型的优点

（1）本文将时序特征转换为更具代表性的假日促销因子与时间段因子，特征工程处理更为巧妙与符合产品需求量变化。

（2）本文采用了时序预测模型与回归模型进行需求量预测，经模型调参优化后，最优模型能更加匹配需求量预测值。

（3）本文从天、周、月三个时间粒度进行预测，能从微观至宏观掌握市场动态，预测产品需求量变化。

6.2 模型的缺点

(1) 本文使用的模型在训练集上有较好的表现，但现实场景更加复杂，不能保证模型具有良好的延展性。

(2) 对于部分数据量较少的产品，本文模型训练度较差，无法预测较为精准的数值。

七、 参考文献

[1] 页川. 大数据时代背景下挖掘教育数据的价值——教育部科学技术研究重点项目成果《教育数据挖掘:方法与应用》出版[J]. 中国远程教育, 2013(04):96.

[2] 王红春, 刘帅. 大数据在供应链管理中的应用研究综述[J]. 物流科技, 2017(8).

[3] 陈爱菊．制造型企业备件采购与库存控制系统研究．武汉理工大学学报， 2010,32(14):179-182.

[4] 牟敬锋, 赵星, 樊静洁,等. 基于 ARIMA 模型的深圳市空气质量指数时间序列预测研究[J]. 环境卫生学杂志, 2017, 000(002):P.102-107

[5] Kimelffeld B,Re C.Transducing Markov sequences.[J]JOURNAL OF THE ACM,2014,61(5):359-373.

[6] Xu K, Dong Y, Evers P T. Towards better coordination of the supply chain [J]. Transportation Research, 2001, Part E 37:35-54.

[7] 闫敏.基于大数据的电商物流价值链分析[J]. 商业时代, 2015, 000(024): 49-50.

[8] 李林汉, 韩祝华. 多元线性回归模型对于荒漠区植物生物量的分析[J]. 中小企业管理与科技, 2015, 000(020):98-98,99.

[9] 彭艳兵, 冯利容. 基于网格概率的离群点检测算法[J]. 计算机系统应用, 2016(4):215-220,共 6 页.