产品订单需求预测分析

摘要

商品需求预测和库存控制是企业营中两个重要的组成部分，库存控制又依赖于准确的需求预测,因为准确的需求预测能够让企业实现更好的库存计划，制定科学的推广方案以及有效地提高客户满意度和服务品质。同时，有效的库存控制策略可以减少仓库的占用，降低总的库存费用，加速企业的资金流动。因此，商品需求预测具有重要的现实意义。

本文从需求预测方面出发，首先通过数据清洗与数据预处理将数据集进行处理，使数据集更具有参考性，本文利用了数学统计的常用方法，如直方图、箱线图、饼图等，对需求特性进行分析。本文着重考虑了时间特性（月的时间段、月份、季节、是否节假日、是否促销日）以及产品特性在不同区域、不同销售方式（线上与线下）等多个特征值对需求的影响，并基于此，利用 one-hot 编码建立特征工程。

进行预测模型搭建环节，首先采用留出法对数据集划分为训练集与测试集，为了本文选用了在预测领域主流的 XGBOOST、随机森林回归、线性回归、决策树回归、LGBM 回归、神经网络（MLP）六种模型进行预测，通过观察数据拟合程度，最终选择了 XGBOOST、随机森林两种模型，并对 XGBOOST 随机森林、决策树分别使用网格搜索和交叉验证进行参数调优得到最优参数，再次进行预测回归两种模型分别进行日、周、月三种时间粒度的需求预测，通过最终准确度进行比较，随机森林以月为颗粒度进行预测，训练集 Mape 值为 0.310582，测试集 Mape 值为 1.492073；以周为颗粒度进行预测，训练集 Mape 值为 0.245801，测试集 Mape 值为 1.451065；以日为颗粒度进行预测，训练集 Mape 值为 0.326993，测试集 Mape 值为 1.206496。xgboost 以日为颗粒度进行预测，训练集 Mape 值为 0.017077，测试集 Mape 值为 3.020385；以周为颗粒度进行预测，训练集 Mape 值为 0.000439，测试集 Mape 值为 2.412107；以月为颗粒度进行预测，训练集 Mape 值为 0.000028，测试集 Mape 值为 3.060715。相比之下，使用随机森林进行预测的误差更小，拟合度更高。

因此最终选择了随机森林回归作为预测模型最稳本文的预测模型，并通过准确度评估验证了模型的有效性及预测的准确性，所提融合模型能有效地应用于实际的企业商品需求预测中，该过程具有科学性。

最后，我们从商品销售领域出发，分析目前我国供给侧改革与扩大内需的政策的大背景下，需求预测的对各行各业积极响应供给侧改革政策的重大意义，分析了商品预测对促进产业智能化科学化，提升产业效能的相关积极意义；以及对为消费者提供更好优质的消费体验等重要现实意义，进一步说明商品需求预测对行业发展有着深远的影响，并对此提出相关发展展望。

abstract

Commodity demand forecast and inventory control are two important components of enterprise operation, and inventory control depends on accurate demand forecast, because accurate demand forecast can enable enterprises to achieve better inventory plan, develop scientific promotion plan and effectively improve customer satisfaction and service quality. At the same time, the effective inventory control strategy can reduce the occupation of the warehouse, reduce the total inventory cost, and accelerate the capital flow of enterprises. Therefore, the commodity demand prediction is of important practical significance.

Starting from the aspect of demand prediction, this paper first processes the data set through data cleaning and data pre-processing to make the data set more reference. This paper uses the common methods of mathematical statistics, such as histogram, box chart, pie chart, etc., to analyze the demand characteristics. This paper focuses on the impact of time characteristics (time period, month, season, whether holidays, whether promotion day) and product characteristics in different regions, different sales methods (online and offline) on demand, and based on this, the one-hot code is used to establish feature engineering.

To build up the prediction model, First, the data set was used to divide it into the training set and the test set, In order to select six models, including XGBOOST, random forest regression, linear regression, decision tree regression, LGBM regression and neural network (MLP), in the prediction field, By looking at the extent of the data fit, Two models, XGBOOST and random forest, were finally selected, And parameter tuning for XGBOOST random forest, decision tree using grid search and cross validation to obtain the optimal parameters, respectively, Make the demand prediction of the three time granularity of day, week and month respectively, For comparison by the final accuracy, Random forest predictions using the month as the granularity, The Mape value in the training set was 0.310582, The Mape value of the test set was 1.492073; With weeks as granularity, The Mape value in the training set was 0.245801, The Mape value of the test set was 1.451065; Predictions by using the day as the granularity, The Mape value in the training set was 0.326993, The test set Mape value was 1.206496. xgboost For the prediction, the Mape value is 0.017077, and the test Mape value is 3.020385; in the training set, the Mape value is 0.000439 and the Mape value is 2.412107; the Mape value is 0.000028 and the Mape value is 3.060715. In contrast, prediction using random forest yielded less error and higher fit.

Finally, from the field of commodity sales, we analyze the significance of the demand prediction on the promotion of the industry intelligence and the industrial efficiency; and provide consumers with better quality consumption experience, further illustrate the significance that the commodity demand forecast has a profound influence on the development of the industry, and put forward relevant development prospects.

目录

产品订单需求预测分析 1
摘要 1
abstract 2
1. 绪论 4
1.1 问题背景 4
1.2 研究意义 4
2. 数据预处理 4
2.1 数据清洗 5
2.1.1 数据异常值与空值处理 5
2.2 数据预处理 5
2.2.1 数据异常值及处理 5
2.2.1 产品价格与产品需求数据平滑验证 7
3. 需求特性分析与特征工程集选择 8
3.1 经销商需求量特性分析 8
3.1.1 时间特性 8
3.1.2 节假日特性 10
3.1.3 区域特性 12
3.1.4 产品特性 13
3.1.5 销售方式特性 15
3.2 基于模型特征工程集的筛选与构造 17
3.2.1 数据预处理 17
3.2.2 特征构建 18
3.2.3 特征构建常用方法 18
3.3.2 决策树 25
3.3.3 XGBOOST 29
3.3.4 多元线性回归预测 30
3.3.5 LightGBM 回归预测模型 30
3.3.6 多层神经网络（MLP） 31
3.4 模型调参 GridSearchCV（网格搜索） 31
3.4.1 网格搜索（GridSearchCV） 31
3.4.2 Grid Search 网格搜索 32
3.5 模型训练 32
4.总结与展望 40
4.1 总结 40
4.2 展望 41
参考文献： 42

1. 绪论

1.1 问题背景

近年来企业外部环境越来越不确定，复杂多变的外部环境，让企业供应链面临较多难题。需求预测作为企业供应链的第一道防线，重要程度不言而喻，然而需求预测受多种因素的影响，导致预测准确率普遍较低，因此需要更加优秀的算法来解决这个问题。需求预测是基于历史数据和未来的预判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及运营计划、目标，资金预算做决策参考；其次，需求预测有助于采购计划和安排生产计划的制定，减少受业务波动的影响。如果没有需求预测或者预测不准，公司内部很多关于销售、采购、财务预算等决策都只能根据经验而来了，会导致对市场预测不足，产生库存和资金的积压或不足等问题，增加企业库存成本。

1.2 研究意义

供应链是一条由供应商、制造商、零售商及终端顾客等组成的复杂网链。因而，供应链上各方的决策充满了诸多不确定性，比如供应不确定性、生产不确定性、需求不确定性及运输不确定性等。而这些不确定因素不仅会影响企业供应链的运作绩效，如生产延迟、成本增加、存货数量提高等，还会导致顾客的满意度下降等负面效果。并且，需求作为牵动整个企业供应系统的源动力，对上下游节点的商品需求进行预测是供应链上的各方成员进行订货策略制定、库存管理及配送规划等经营决策的基础，有利于企业合理安排生产发展，减少仓储成本。企业根据需求量的变化快速的响应市场的变化，能够更加清晰的了解消费者的行为和需求，帮助企业与客户建立更加紧密的联系。[2]因此，如何降低商品需求的不确定性，对于供应链的运作及管理策略的合理制定具有重要的现实意义。[4]

2.数据预处理

由于在出货数据进行录入时可能存在缺失、异常、重复等脏数据，往往在构建模型之前，都需要进行相关预处理工作：对异常值、缺失无效值进行清洗以及数据转换等，通过数据清洗以及数据转换等工作，可以使原始数据对预测结果的表达更加清晰。经过处理的数据集对

于提升模型的预测效果有很大的作用，本节我们首先对整体数据进行初步探索，分析丢失数据和异常数据的原因及特征，再进行一系列的数据预处理和特征过程步骤。

首先对本数据集中相关数据项说明：

项目表中名称	代表含义
order_date	订单日期
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
ord_qty	订单需求量

2.1 数据清洗

由于在出货数据进行录入时可能存在缺失、异常、重复等脏数据，往往在构建模型之前，都需要进行相关预处理工作：对异常值、缺失无效值进行清洗以及数据转换等，通过数据清洗以及数据转换等工作，可以使原始数据对预测结果的表达更加清晰。经过处理的数据集对于提升模型的预测效果有很大的作用，本节我们首先对整体数据进行初步探索，分析丢失数据和异常数据，再进行一系列的数据预处理和特征过程步骤。

2.1.1 数据空值与重复值处理

通过对数据集中各项值的数据筛查，未发现数据缺失值；对检测到的重复项进行删除处理，保留唯一项。为方便后续进行数据分析，对数据集中的异常值进行预处理。若出现 2 及 2 行以上属性值都相同的数据，则称为有重复。

2.2 数据预处理

2.2.1 数据异常值及处理

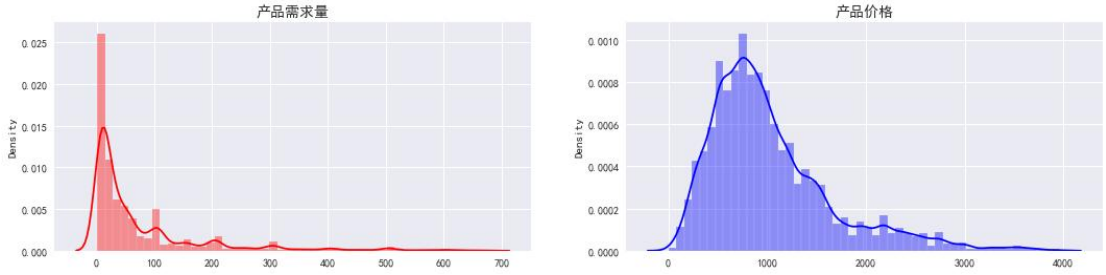
观察清洗后的数据，猜想其满足正态分布，其中，正态分布公式为若随机变量 X 服从一个数学期望为 μ、方差为 σ 2 的正态分布，记为 N(μ， σ 2)。其概率密度函数为正态分布的期望值 μ 决定了其位置，其标准差 σ 决定了分布的幅度。

μ 是正态分布的位置参数，描述正态分布的集中趋势位置。概率规律为取与 μ 邻近的值的概率大，而取离 μ 越远的值的概率越小。正态分布以 X= μ 为对称轴，左右完全对称。正态分布的期望、均数、中位数、众数相同，均等于 μ 。

σ 描述正态分布资料数据分布的离散程度，σ 越大，数据分布越分散，σ 越小，数据分布越集中。也称为是正态分布的形状参数，σ 越大，曲线越扁平，反之，σ 越小，曲线越瘦高。

正态分布	均值（μ）	标准差（σ）
产品需求量	1027.246	624.337
产品价格	73.402	107.996

将产品价格与产品需求作为频数，绘制直方图，并基于上述数据绘制正态分布函数曲线，存在个别数据的值差异过大，出现离群点的情况，则需要对数据集进行平滑处理。

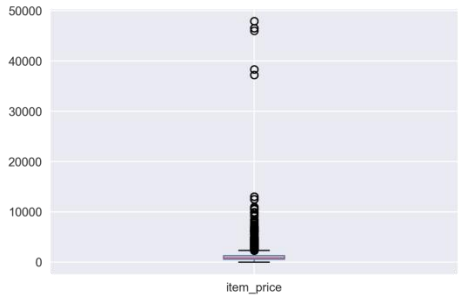


产品需求量与产品价格直方图

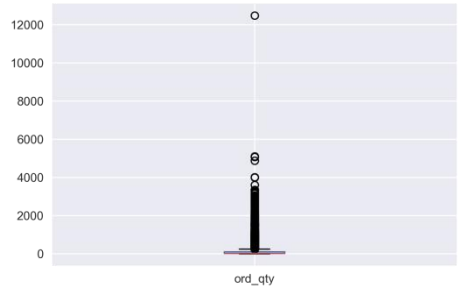
通过观察数据可以发现，某些数据出现了与其他值差异过大的情况，并猜测该情况可能是由于记录有误导致该值其他正常值差异过大，如果将异常值一同计入统计，可能会导致数据难以分析，影响后续数据预测准确程度等问题，因此将对异常值进行检测及处理。

异常值检测常用方法有简单统计分析、3σ 原则、箱线图、聚类算法等，其中箱线图为一种常用统计图，是通过最小值、下四分位数、中位数、上四分位数和最大值进行绘制箱线图，并以此展现样本数据大致的分布情况。

通过代码分别绘制产品价格与产品需求量的箱线图：

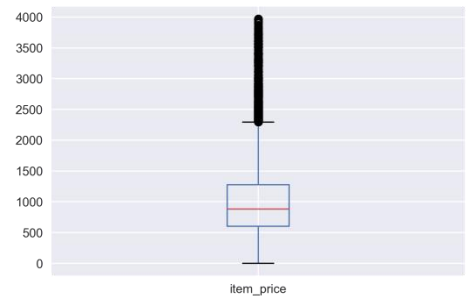


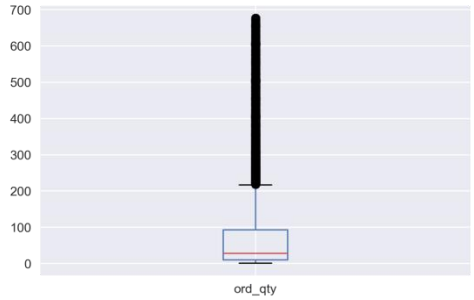
产品价格箱线图



产品需求量箱线图

通过观察产品价格箱线图与产品需求箱线图可知，产品价格和各类产品的需求量分布较分散，样本集中存在许多异常值，有部分产品价格较高异常值过多会影响数据的分析结果，因此需要对样本中的异常值使用正态分布的 3σ 原则进行处理，去除正态分布中 3σ 以外的数据，并再次绘制产品价格箱线图与产品需求箱线图。

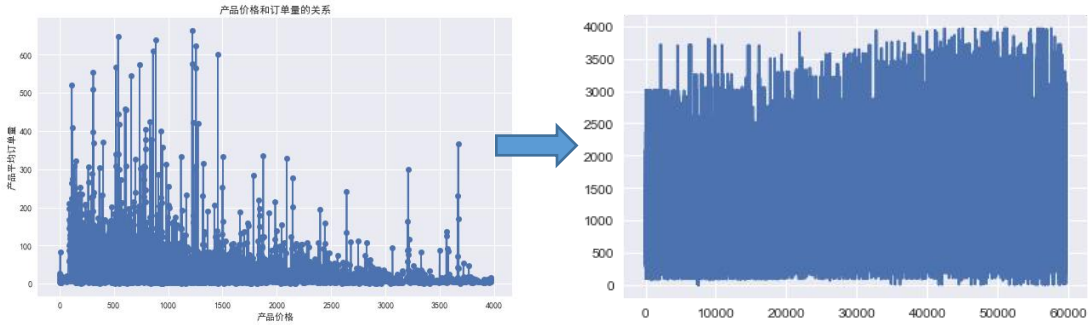




通过观察新的产品价格箱线图与产品需求箱线图可知，图中的散点异常值明显减少，说明数据异常值已被去除。产品大部分价格在 800 元左右，产品的需求量在 100 以内分布较平均。

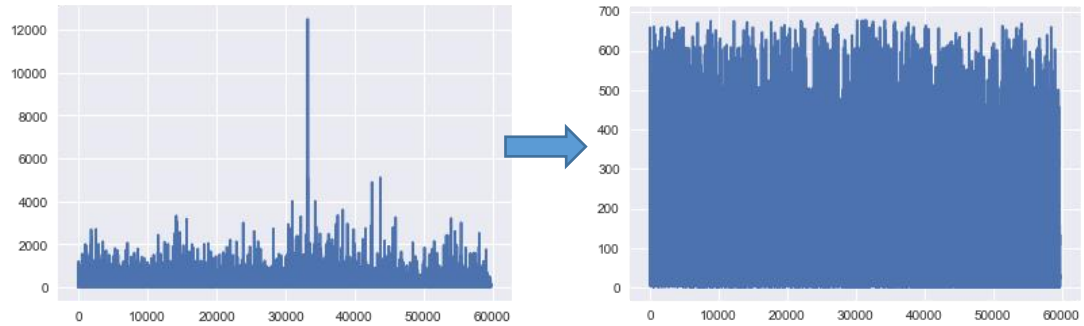
2.2.1 产品价格与产品需求数据平滑验证

通过分别绘制产品需求量的直方图处理前与处理后的图，可以发现进行数据异常值处理前产品订单量较分散，规律不明显，经过平滑处理后的数据更为集中，一些高异常的值向下收缩，最终使得数据分布更为集中。



产品价格分布直方图处理前后对比

（左：处理前、右：处理后，纵轴为订单量，横轴为产品价格）



产品需求量分布直方图处理前后对比

（左：处理前、右：处理后，纵轴为订单量，横轴为需求量）

3.需求特性分析与特征工程集选择

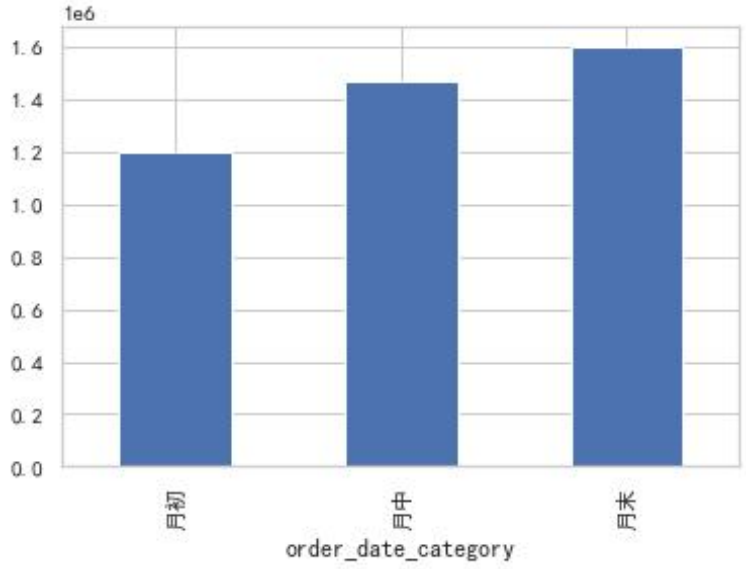
3.1 经销商需求量特性分析

3.1.1 时间特性

时间特性是指在具体的某一特殊时间区间内订单量的变化，通过分信息一个月中的各个时间段（月初、月中、月末）、每个月、以及每个季度三个时间维度，以及额外分析特殊时间段（例如节假日，促销日等）对经销商的订单需求量的影响，以此总结经销商的需求量所内涵的时间特性。

1.月内周期性

首先，通过统计每个月的不同时间段（月初、月中、月末）的需求量总和。

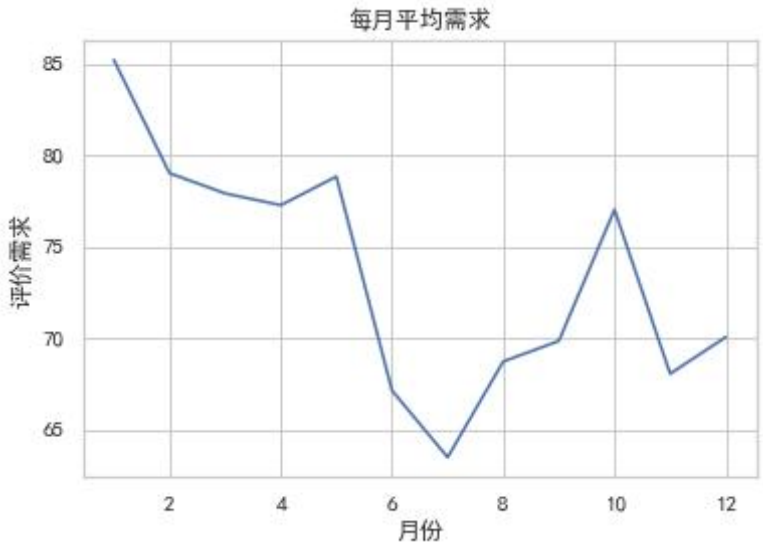


月不同时间段需求量总和直方图

通过直方图可以看出，经销商更加倾向于在月中和月末订购商品。经销商在月头需完成采购相关的准备工作，如提前制定每月的采购计划、寻找合适供应商、指定月销售额指标等前期工作，因此在每个月的月中与月末经销商需求量较大，月初需求量较少。

2.月周期性

通过统计每个月的平均需求量，并绘制折线图，以此分析每个月份与需求量的关系。



12 个月平均需求折线图

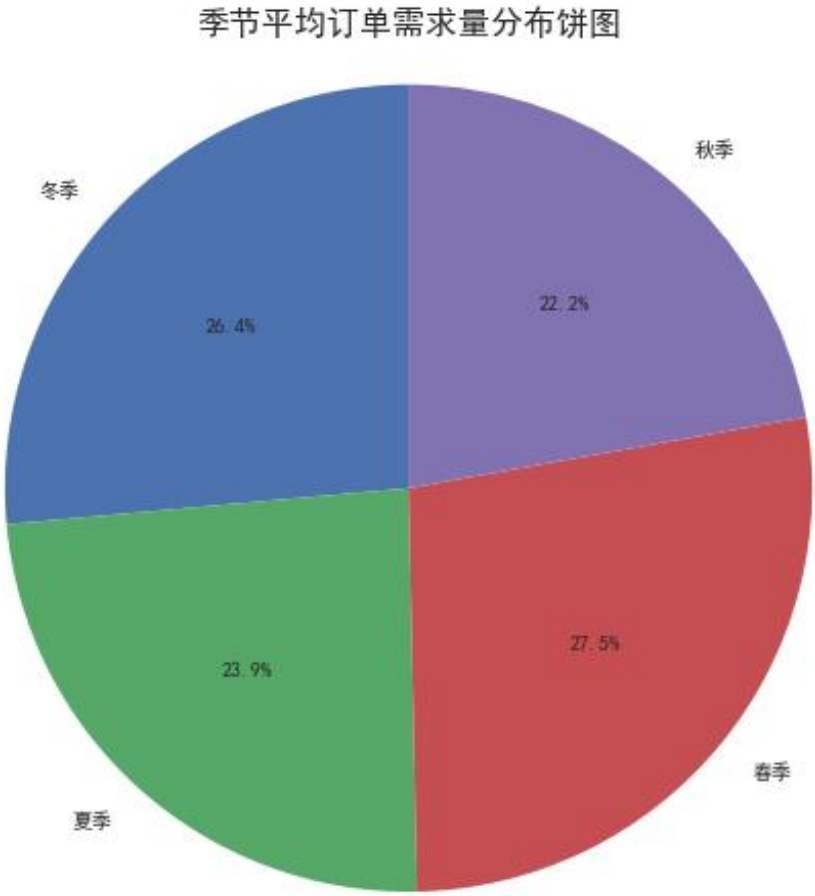
由折线图可知，全年需求量最高峰为一月，其次在五月、十月出现小高峰，一月到七月供应商订单需求量总体呈下降趋势，七月后上升。一月份、五月份、十月份为传统大宗商品采购旺季，经销商进货量较多。一月为春节假期时期，市场消费能力较强，供应商需求量大；春节消费能力势头过后，3、4 月为消费高峰过后的缓冲期，因此持续到五月份，在五一长假、端午假期等作用下，消费需求迎来第二个高峰，紧接着再次进入缓冲期，十月份过后，受中秋节以及十一黄金周影响，消费需求再一次出现高峰，因此经销商在一月、五月、十月需求量较大。

3.季节周期性

按我国气象部门的气象划分法，将四季按如下月份划分：

季节	月份
春季	3 月、4 月、5 月
夏季	6 月、7 月、8 月
秋季	9 月、10 月、11 月
冬季	12 月、1 月、2 月

通过划分好的月份统计每个季节的平均需求量，并画出四个季节的平均订单量的饼图。



由图可知，春季占比较多，占 27.5%；秋季占比较少，占 22.2%，符合月周期性中的规律。

3.1.2 节假日特性

源于国务院对于节假日公休安排的通知：

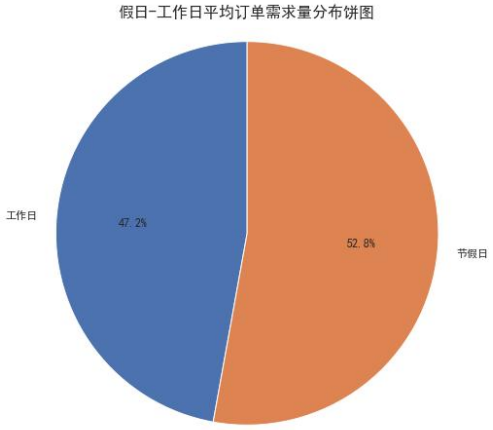
年份	放假时间
2015 年	1、元旦：1 月 1 日至 3 日放假调休，共 3 天。1 月 4 日（星期日）上班。 2、春节：2 月 18 日至 24 日放假调休，共 7 天。2 月 15 日（星期日）、2 月 28 日（星期六）上班。 3、清明节：4 月 4 日放假，4 月 6 日（星期一）补休。 2015 4、劳动节：5 月 1 日放假，与周末连休。 5、端午节：6 月 20 日放假，6 月 22 日（星期一）补休。 6、胜利日：9 月 3 日~9 月 5 日，9 月 4 日调休、9 月 6 日（星期日）上班

	7、中秋节: 9 月 27 日放假。 8、国庆节: 10 月 1 日至 7 日放假调休，共 7 天。10 月 10 日（星期六）上班
2016 年	1、元旦: 1 月 1 日放假，与周末连休。 2、春节: 2 月 7 日至 13 日放假调休，共 7 天。2 月 6 日（星期六）、2 月 14 日（星期 8)上班。 3、清明节:4 月 4 日放假，与周末连休。 4、劳动节:5 月 1 日放假，5 月 2 日（星期一）补休。 5、端午节:6 月 9 日至 11 日放假调休，共 3 天。6 月 12 日（星期日）上班。 6、中秋节:9 月 15 日至 17 日放假调休，共 3 天。9 月 18 日（星期日)上班。 7、国庆节:10 月 1 日至 7 日放假调休，共 7 天。10 月 8 日（星期六)、10 月 9 日（星期日）
2017 年	1、元旦:1 月 1 日放假，1 月 2 日（星期一）补休，与 2016 年 12 月 31 日（星期六)连休，共 3 天。 2、春节:1 月 27 日至 2 月 2 日放假调休，共 7 天。1 月 22 日（星期日)、2 月 4 日（星期六）上班。 3、清明节:4 月 2 日至 4 日放假调休，共 3 天。4 月 1 日（星期六）上班。 4、劳动节:5 月 1 日放假，与周末连休，共 3 天 5、端午节:5 月 28 日至 30 日放假调休，共 3 天。5 月 27 日（星期六)上班。 6、国庆节、中秋节: 10 月 1 日至 8 日放假调休，共 8 天，9 月 30 日（星期六）上班。
2018 年	1、元旦:1 月 1 日放假，与周末连休。 2、春节:2 月 15 日至 21 日放假调休，共 7 天。2 月 11 日(星期日)、2 月 24 日(星期六)上班。 3、清明节:4 月 5 日至 7 日放假调休，共 3 天。4 月 8 日(星期日)上班。 4、劳动节:4 月 29 日至 5 月 1 日放假调休，共 3 天。4 月 28 日(星期六)上班 5、端午节:6 月 18 日放假，与周末连休。 6、中秋节:9 月 24 日放假，与周末连休。 7、国庆节:10 月 1 日至 7 日放假调休，共

	7 天。9 月 29 日(星期六)、9 月 30 日(星期日)上班。

将以上节假日日期以及 2015 年至 2018 年的周末设定为节假日，并基于此进行数据分析。

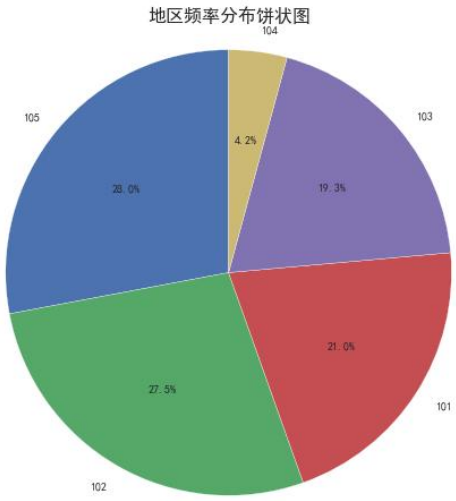
根据数据集，统计出节假日与非节假日需求量均值，并绘制出分布饼图。



由图可知，节假日的订单需求量占 52.8%，工作日的订单需求量占 47.2%，节假日对订单需求量的提升有一定的影响，但影响几率相对较小，说明节假日不是影响供应商需求的决定性因素。

3.1.3 区域特性

本数据集中包含五个地区编码（101-105），基于数据集统计各地区的需求量，并绘制各地区产品需求量的饼图，以此分析各大地区的需求量占比情况。



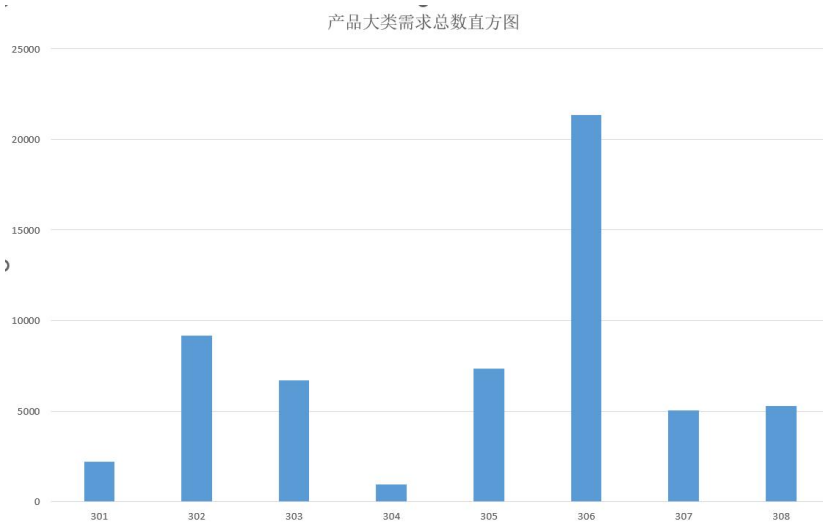
由图可知，102（绿色）地区的产品需求量最大，占 27.5%；105（蓝色）、103（紫色）与 101（红色）三个地区的产品需求量较大，104（黄褐色）产品需求量占比最少。可以说明 102 地区的供应商需求量较大，101、103、105 地区经销商需求量较均匀，104 地区经销商

需求量较少，因此可以更多拓展对 102 地区的业务，如开拓新路线，设置更多下级仓库。104 地区的市场需求量潜力大，可以在 104 地区加强宣传力度，吸引更多经销商，开设更多销售渠道。

3.1.4 产品特性

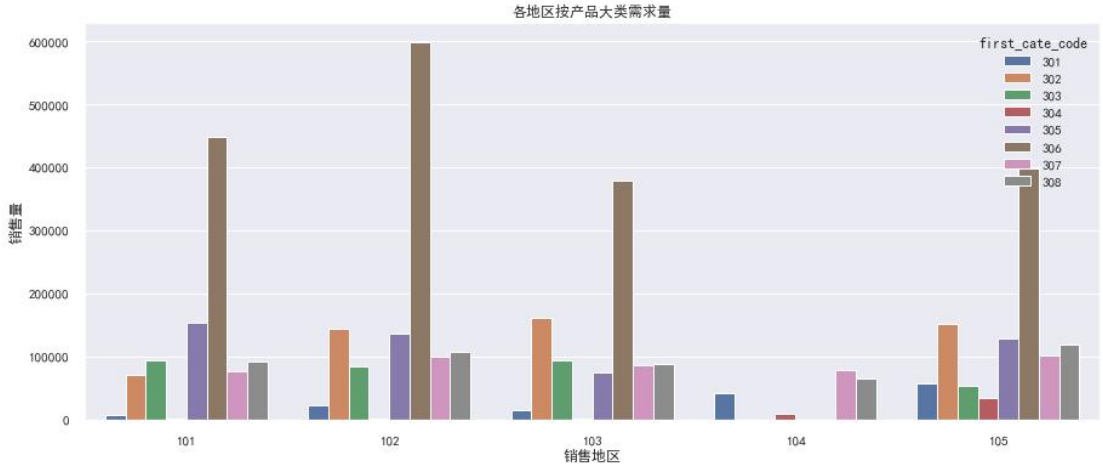
1.大类产品特性

大类产品为大宗商品的基本分类，不同大类产品的的需求情况不同，因此需要统计不同大类产品需求情况，以此分析不同大类产品的需求特点，可以更准确地把握需求市场。



由图可知，306 大类产品较为受欢迎，301 与 304 大类产品需求较少，其余产品需求量较为平均，可适当增加 306 大类的产品的仓储面积。

2.产品大类与区域性结合分析



各地区产品大类需求量直方图

由图可知，102（绿色）地区的产品需求量较大，占 27.5%；105（蓝色）、103（紫色）与 101（红色）三个地区的产品需求量较为一致，104（黄褐色）产品需求量占比最少。

基于各销售地区的各大类需求量直方图进行分析，初步得出结论：

104 地区的总体产品需求量大大低于其他四个地区，其余四个的总体产品需求量较均匀。

301 类（蓝色）产品占总体需求量较少，且在 101、102 与 103 地区需求量远小于 104 与 105 地区；306（褐色）产品占总体需求量较多。

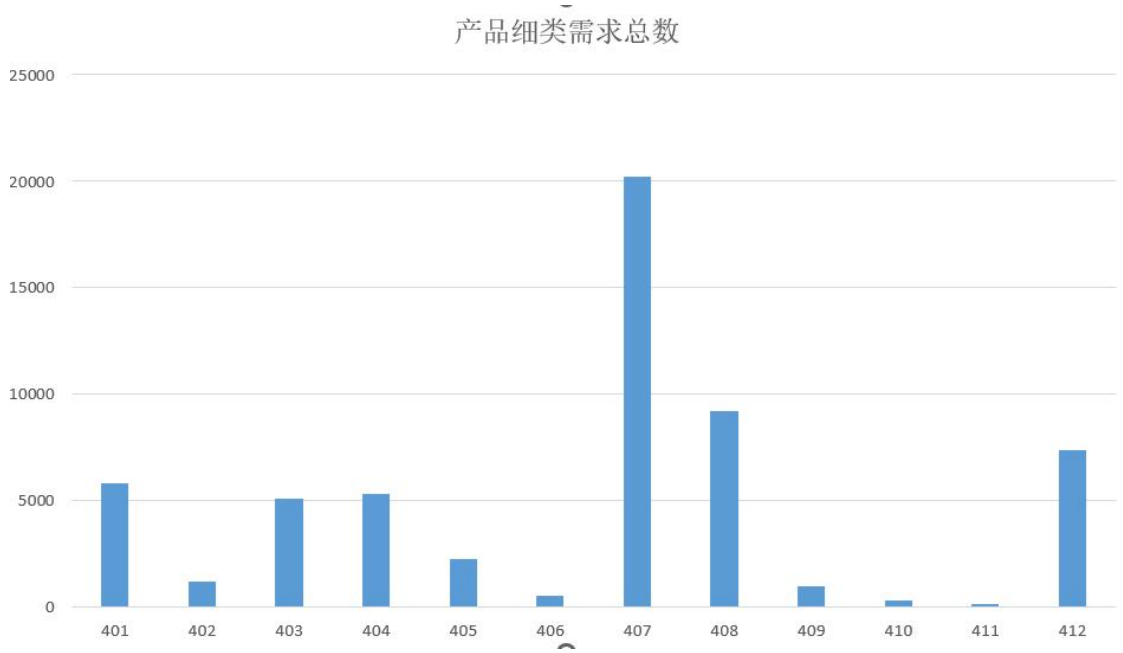
302（橙色）303（绿色）、305（紫色）、307（粉色）四种产品大类在除 104 以外的四个地区占比数值较为平均。

304 大类产品只在 104 与 105 地区有需求，且需求量不大，基于此推测 304 大类产品为 104、105 地区的特色产品。

通过观察数据，将产品大类与所包含地产品细类通过表格形式展示：

产品大类	产品细类
301	405
302	408
303	401、406、410、411
304	409
305	412
306	402、407
307	403
308	404

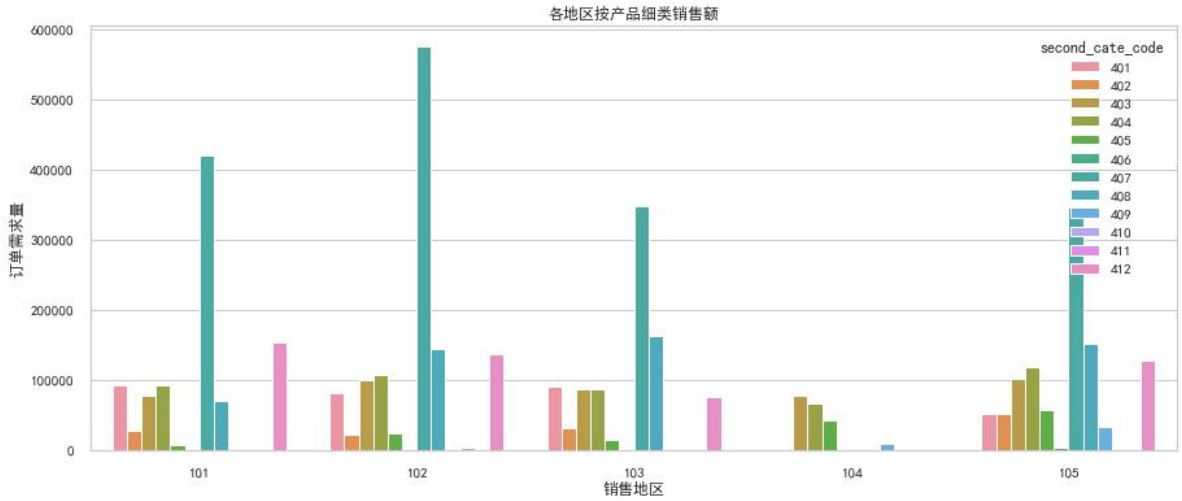
通过统计各细类需求数量，并对此绘制直方图：



产品细类直方图

由图可知，407 细类产品需求量远超于其他细类产品，其中 407 为 306 大类产品下的细类产品，因此在增加 306 大类产品的生产与仓储投入时，应着重增加 407 细类产品的投入。

3.产品细类与区域性结合分析：



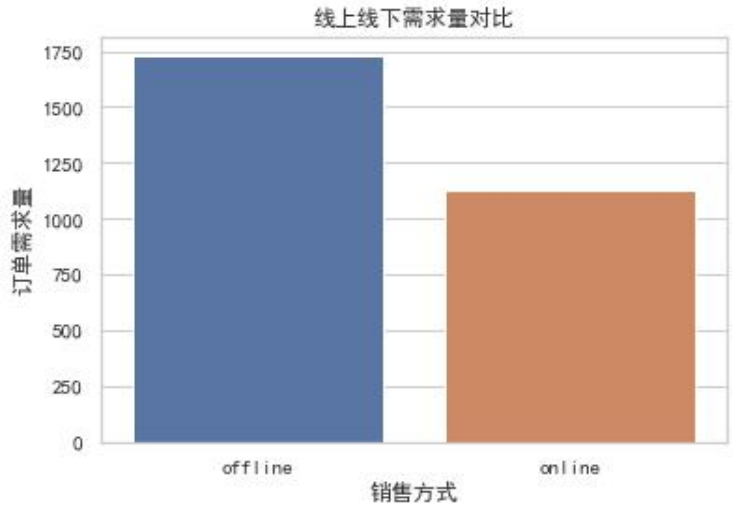
各个区域中不同细类产品需求量直方图

由图可知，407 细类产品在除 104 区域外占比最高，符合上述结论。

3.1.5 销售方式特性

同销售方式的总体分布情况

首先统计线上线下两种方式的需求量进行求和统计，对此绘制直方图。

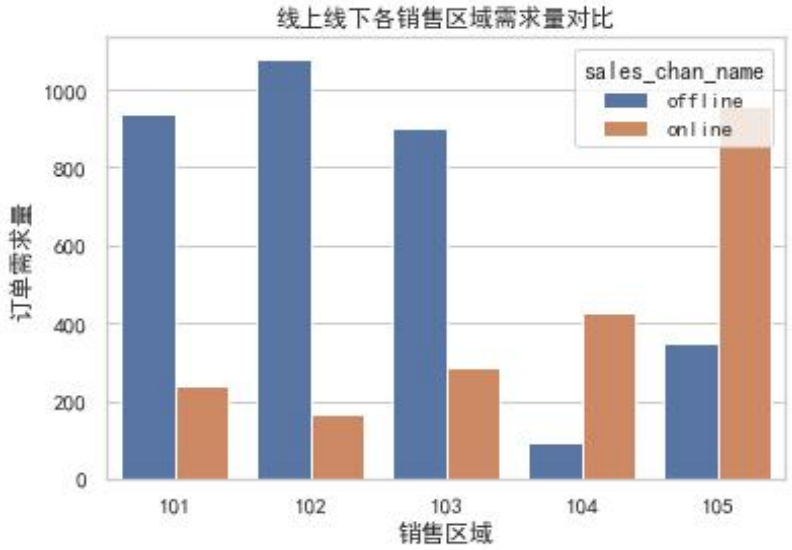


由图可知，从总的需求量分布上，经销商通过线下销售方式的需求量更多，由于经销商在线下可以实际考察产品的治疗与生产流程，企业的规模大小等，因此更倾向于线下订购。

总结，经销商的产品需求受其所在区域以及产品大类影响，且不同区域不同商品影响程度不同，后续构造特征工程时需单独考虑不同组合的影响。

1. 销售方式与区域性结合分析

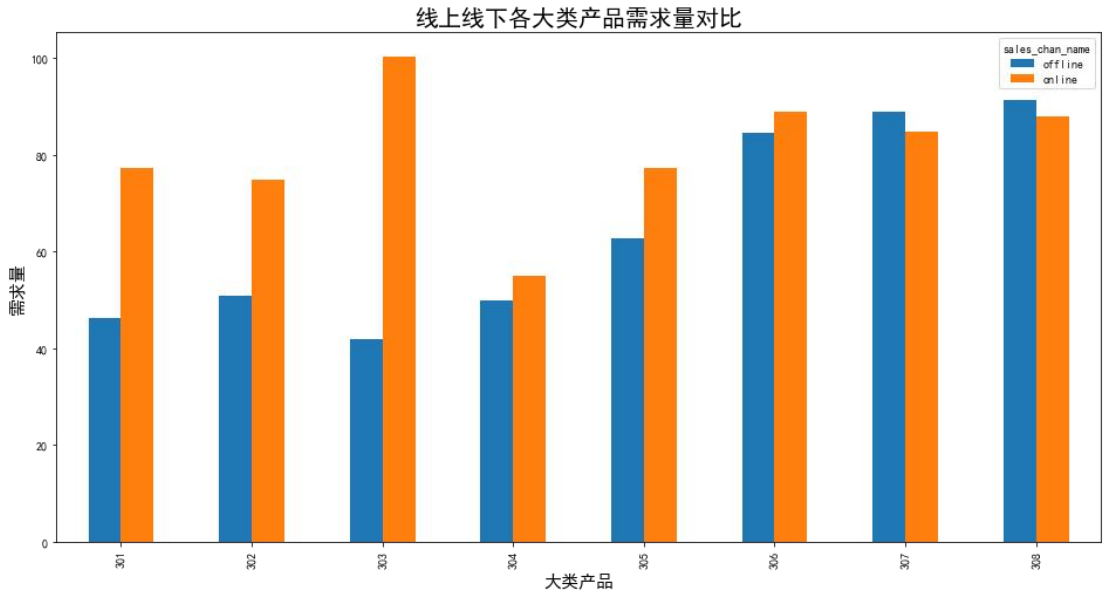
通过绘制各大类的销售方式的需求量的直方图，进一步分析各个大类产品与不同的销售方式的关系。



由图可知，在 101、102 和 103 地区，经销商对线下销售（offline）方式需求量更大，说明在 101、102 与 103 地区线下销售方式更受欢迎；而在 104、105 地区，经销商对线上销售（online）方式需求量更大，说明在 104、105 地区线上销售方式更受欢迎，由此可以总结，销售方式对于需求量有显著影响，且不同地区的影响方式不同，在 101、102 与 103 地区的经销商更加倾向于线下销售（offline）方式，而在 104、105 地区的经销商更倾向与选择线上销售方式。

2. 产品大类不同销售方式结合分析

通过绘制各个大类产品在不同销售方式的直方图，并分析对不同的大类产品与销售方式的影响。



由图可知，不同大类产品受销售方式影响不同，其中：301、302 与 303 大类产品受销售方式影响较大，且经销商更倾向于线上销售；305 与 306 大类产品的需求量受销售方式的影响，但影响较小；其余 304、307 与 308 大类产品在线上与线下中销售方式中，订单需求量相差不大。

3.2 基于模型特征工程集的筛选与构造

前面已对需求特性进行了详尽的分析，分析结果表明需求与时间特性（月的时间段、月份、季节、是否节假日、是否促销日）以及产品特性在不同区域、不同销售方式（线上与线下）关系较为紧密，因此，商品需求模型应考虑商品订购时间、何种商品、何地订购、如何订购作为输入变量。此外，由于数据集还有商品细类，但是为了接下来的负荷预测，我们需要确定具有较大影响力的特征，所以我们将考虑这些因素的影响，先将需求数据按商品区域分组，形成 2618 个负荷子序列，再对各个子序列分别建模。通过回归模型的方法选择对需求数据有解释力的特征。基于上述考虑，可以对需求数据进行如下处理并建立模型：

3.2.1 数据预处理

这里我们采用去除重复值，以及异常值（订单需求量或者产品价格均不能为负数），经过数据预处理后，我们得到的数据集如下：

	A	B	C	D	E	F	G	H	I	J	K
1	订单日期	销售区域编	产品编码	产品大类编	产品细类编	销售渠道名	产品价格	订单需求量	订单日期_I	订单日期_J	订单日期_月
2	2015/9/1	104	22069	307	403	offline	1114	19	1	36	9
3	2015/9/1	104	20028	301	405	offline	1012	12	1	36	9
4	2015/9/2	104	21183	307	403	online	428	109	2	36	9
5	2015/9/2	104	20448	308	404	online	962	3	2	36	9
6	2015/9/2	104	21565	307	403	offline	1400	3	2	36	9
7	2015/9/2	104	20606	308	404	offline	1614	14	2	36	9
8	2015/9/2	104	20606	308	404	offline	1615	17	2	36	9
9	2015/9/2	104	20028	301	405	offline	1005	11	2	36	9
10	2015/9/2	102	20323	305	412	offline	99	502	2	36	9
11	2015/9/2	102	21350	305	412	offline	267	107	2	36	9
12	2015/9/2	101	20657	303	410	offline	2996	18	2	36	9
13	2015/9/2	102	20457	305	412	offline	164	308	2	36	9
14	2015/9/3	102	21052	303	401	offline	1507	14	3	36	9
15	2015/9/3	102	22046	305	412	offline	1204	88	3	36	9
16	2015/9/3	102	20518	305	412	offline	1069	15	3	36	9
17	2015/9/3	102	20020	305	412	offline	1918	18	3	36	9
18	2015/9/3	102	20459	305	412	offline	1666	6	3	36	9
19	2015/9/3	102	20797	305	412	offline	2619	31	3	36	9
20	2015/9/3	102	21745	303	401	offline	1713	41	3	36	9
21	2015/9/3	102	20844	303	401	offline	2449	17	3	36	9
22	2015/9/3	102	20354	303	401	offline	1420	64	3	36	9
23	2015/9/3	102	20717	303	401	offline	2351	20	3	36	9
24	2015/9/3	102	21499	303	401	offline	2006	17	3	36	9
25	2015/9/3	102	22035	305	412	offline	2008	26	3	36	9
26	2015/9/3	102	21347	303	401	offline	1131	168	3	36	9
27	2015/9/3	102	20008	303	401	offline	667	44	3	36	9
28	2015/9/3	102	21644	303	401	offline	319	9	3	36	9
29	2015/9/3	102	21271	306	407	offline	459	303	3	36	9
30	2015/9/3	102	21003	306	407	offline	594	155	3	36	9

共有 597694 条数据

3.2.2 特征构建

特征构建：机器学习模型一般只能处理向量化的数据，因此在建模过程中，需要将收集到的原始数据进行转化，构建出机器学习模型可以利用的数据形式（即向量化的数据），这个过程就是特征构建的过程。特征构建从收集到的机器学习模型的原始数据中提取出特征，将原始数据空间映射到新的特征空间，使得在新的特征空间中，模型能够更好地学习数据中的规律。

很多经典的机器学习模型，比如 logistic 回归、线性模型、FM 等需要进行精细的特征构建才能达到很好的效果。在如今的深度学习时代，由于数据量大，数据只需要进行简单的处理就可以灌入到深度学习模型中，最终可以获得比较好的效果。虽说如此，但是在很多时候特征构建是必须的，特征构建相当于通过人类的思考和理解，期望抓住问题的本质，辅助机器学习模型获得更好的效果。

3.2.3 特征构建常用方法

（1）离散特征构建

离散特征是非常常见的一类特征，推荐系统中的用户属性数据、物品属性数据中就包含大量的类别特征，如性别、学历、视频的类型、标签、导演、国别等等。对于离散特征，一般可以采用如下 4 种方式对特征进行编码（即特征构建）。

one-hot 编码

one-hot 编码通常用于离散特征（也叫类别特征），如果某个类别特征有 k 类，我们将这 k 类固定一个序关系，我们可以将每个值映射为一个 k 维向量，其中这个值所在的分量为 1，其他分量为 0。比如性别进行编码的话，男可以编码为(1，0)，女可以编码为(0，1)。该方法当类别的数量很多时，特征空间会变得非常大。

散列编码

对于有些取值特别多的类别特征，使用 one-hot 编码得到的特征矩阵非常稀疏，如果再进行特征交叉（5.1.4 会讲到），会使得特征维度爆炸式增长。特征散列的目标就是是把原始的高维特征向量压缩成较低维特征向量，且尽量不损失原始特征的表达能力，其优势在于实现简单，所需额外计算量小。降低特征维度，也能加速算法训练与预测，降低内存消耗，但代价是通过哈希转换后学习到的模型变得很难检验(因为一般哈希函数是不可逆的)，我们很难对训练出的模型参数做出合理解释。特征散列的另一个问题是可能把多个原始特征哈希到相同的位置上，出现哈希冲突现象，但经验表明这种冲突对算法的精度影响很小，通过选择合适的 hash 函数也可以减少冲突概率。

计数编码

就是将所有样本中该类别出现的次数或者频次作为该特征的编码，这类方法对异常值比较敏感(拿电影的标签来说，很多电影包含“剧情”这个标签，计数编码会让剧情的编码值非常大)，也容易产生冲突(两个不同类别的编码一样，特别是对于出现很稀少的标签，编码值一样的概率非常大)。

这里使用 one-hot 编码（热编码），新增月日周以及季度列：

月时间段	**0 月初**	**1 月中**	**2 月末**	
季节	**0 否**	**1 是**		
是否节假日	**1 春季**	**2 夏季**	**3 秋季**	**4 冬季**

通过上述特性分析， 以日为粒度的订单需求量与节假日有关，同时与月中月末有关； 以月周为粒度的订单需求量与季节有关。

下图为特征构建完成后的表格，该表格将作为训练以日为粒度的模型：

	订单日期	销售区域编码	产品编码	产品大类编码	产品细类编码	销售渠道名称	产品价格	订单需求量	日	周	月	月时间段	是否节假日
0	2015-09-01	104	22069	307	403	1	1114.0	19	1	36	9	0	0
1	2015-09-01	104	20028	301	405	1	1012.0	12	1	36	9	0	0
2	2015-09-02	104	21183	307	403	0	428.0	109	2	36	9	0	0
3	2015-09-02	104	20448	308	404	0	962.0	3	2	36	9	0	0
4	2015-09-02	104	21565	307	403	1	1400.0	3	2	36	9	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
597689	2018-12-20	102	20994	302	408	1	843.0	59	20	51	12	1	0
597690	2018-12-20	102	21875	302	408	1	762.0	502	20	51	12	1	0
597691	2018-12-20	102	20215	302	408	1	2013.0	106	20	51	12	1	0
597692	2018-12-20	102	20195	302	408	1	2120.0	187	20	51	12	1	0
597693	2018-12-20	102	20321	302	408	1	1244.0	205	20	51	12	1	0

下图为特征构建完成后的表格，该表格将作为训练以月和周为粒度的模型：

	订单日期	日	周	月	销售区域编码	产品编码	产品大类编码	产品细类编码	订单需求量	季节
0	2015-09-01	1	36	9	104	22069	307	403	19	3
1	2015-09-01	1	36	9	104	20028	301	405	12	3
2	2015-09-02	2	36	9	104	21183	307	403	109	3
3	2015-09-02	2	36	9	104	20448	308	404	3	3
4	2015-09-02	2	36	9	104	21565	307	403	3	3
...	...	...	...	...	...	...	...	...	...	...
597689	2018-12-20	20	51	12	102	20994	302	408	59	1
597690	2018-12-20	20	51	12	102	21875	302	408	502	1
597691	2018-12-20	20	51	12	102	20215	302	408	106	1
597692	2018-12-20	20	51	12	102	20195	302	408	187	1
597693	2018-12-20	20	51	12	102	20321	302	408	205	1

（2）模型构建

训练集划分：

留出法:

将数据集 D 划分为 2 个互斥子集，其中一个作为训练集 S，另一个作为测试集 T，即有：

$$D = S \cup T, \quad S \cap T = \varnothing$$

用训练集 S 训练模型，再用测试集 T 评估误差，作为泛化误差估计。

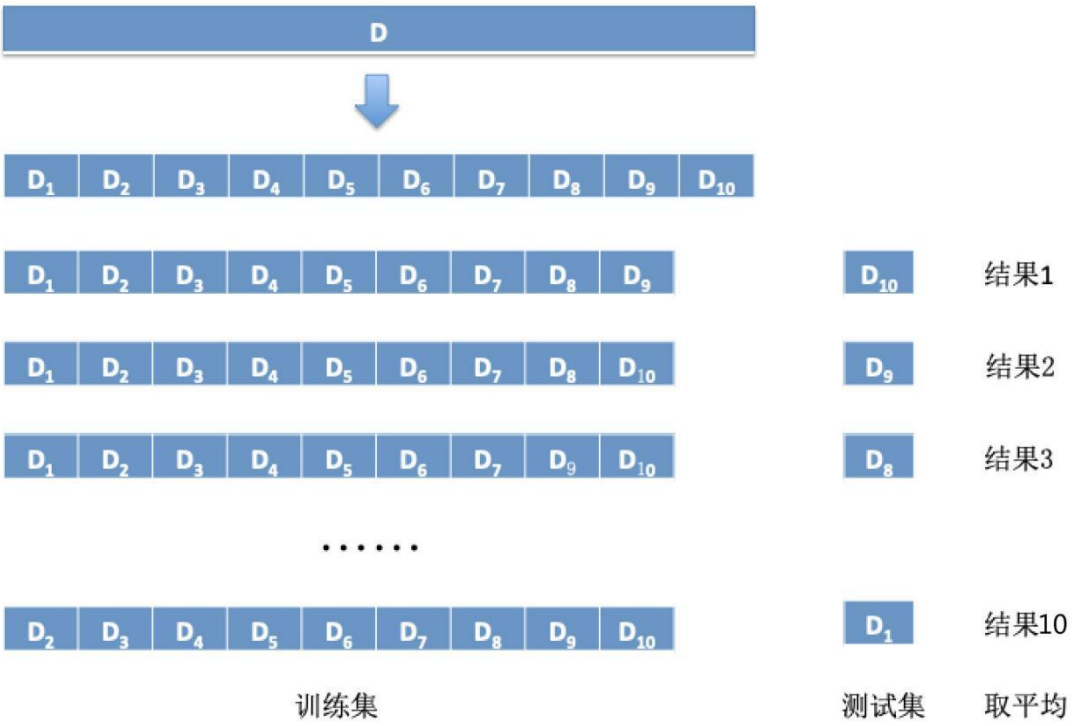
特点：单次使用留出法得到的估计结果往往不够稳定可靠，故如果要使用留出法，一般采用若干次随机划分，重复进行实验评估后，取平均值作为最终评估结果。

交叉验证法：

将数据集 D 划分为 k 个大小相似的互斥子集，即：

D = D1 ∪ D2 ∪ D3 ∪ ..... ∪ Dk

有：Di ∩ Dj=∅ ，每个子集 Dr 都尽可能保持数据分布的一致性，即：从 D 中通过分层采样得到，每次使用 k-1 个子集的并集作为训练集 S，余下的一个作为测试集 T，最终返回的是 k 个测试结果的均值。因其稳定性与保真性取决于 k 值，故又称为 k 折交叉验证（K-fold cross validation），其中 k 最常用取值为 10，又称 10 折交叉验证。



特点：当数据集 D 中数据量较大时，训练 m 个模型的开销过大。这里我们采取第一种方法来划分训练集。

（3）回归模型常用评估指标

按天、周、月的时间粒度对产品订单量和产品价格进行预测，由于受现实各种因素的影响，一些问题没有得到全面的考虑，同时预测集中出现了训练集没有的产品，产生误差不可避免，只能尽量调整模型参数降低误差。预测精度评判标准是：误差越小，精度就越高，反之精度就越低。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：

1.均方误差（mean-square error, MSE）

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

均方误差(MSE)是反映估计量与被估计量之间差异程度的一种度量。均方误差(MSE)是最常用的回归损失函数，计算方法是求预测值与真实值之间差值的平方和，再求平均。对比平均绝对误差，均方误差对异常值更敏感，因为均分误差对误差进行了放大（乘方）。该指标是线性回归的最小二乘法损失函数，将其作为模型的预测性能评估指标，比较简单直观。

通过求误差的平方和再平均来计算误差，可以衡量模型的真实值和预测值之间的偏差。取值范围[0, + ∞)，指标越小，模型精度越高。公式如下，其中是y_i真实值，是$\hat{y}_i$预测值，n 是样本数。

2.平均绝对误差（Mean Absolute Error，MAE)

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

平均绝对误差 MAE 与原始数据单位相同， 它仅能比较误差是相同单位的模型。量级近似与 RMSE，但是误差值相对小一些。MAE 对预测值和真实值的残差直接进行计算，且量纲与计算值相同，可以比较直观地体现模型预测结果的误差情况。是对预测值和真实值之间绝对误差平均值的评价，取值范围[0, + ∞)，指标越小，模型精度越高。公式如下，其中是y_i真实值，是$\hat{y}_i$预测值，n 是样本数。

3.均方根误差（Root Mean Squared Error，RMSE）

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

RMSE 是一个衡量回归模型误差率的常用公式。RMSE 针对异常值会比较敏感，若出现预测值和真实值偏差较大，则 RMSE 值就会很大，可以用来衡量模型的稳定性。均方根误差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，也常作为评价指标之一。取值范围[0, + ∞)，指标越小，模型精度越高。公式如下，其中是y_i真实值，是$\hat{y}_i$预测值，n 是样本数。

4.决定系数R^2（Coefficient of Determination）

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_1 - \bar{y})^2}$$

R^2决定系数是一种用于评估回归模型拟合优度的指标，它反应了拟合模型模型对样本数据的拟合程度。表示模型能够解释数据方差的比例，通常用于比较不同模型的表现。

假设有 n 个样本，真实值分别为 y_1，y_2，……，y_n ，预测值分别为$\hat{y}_1$，$\hat{y}_2$，……，$\hat{y}_n$ 。首先，我们可以定义总方差（Total Sum of Squares，TSS）为真实值 y 的方差，即：$TSS = \Sigma(y_i - \bar{y})^2 / n$，$(i=1,2,\cdots,n)$ 。其中，$\bar{y}$为所有真实值的平均数。我们希望得到模型的解释方差，即预测值能够解释的数据方差。因此，可以定义残差平方和（Residual Sum of Squares，RSS）为：$RSS = \Sigma(y_i - \hat{y}_i)^2 / n$，$(i=1,2,\cdots,n)$。其中，$y_i - \hat{y}_i$为第 i 个样本的残差，表示预测值与真实值之间的差。最后计算R^2决定系数为解释方差占总方差的比例，即：$R^2 = 1 - (RSS / TSS)$。R^2决定系数的取值范围为[0,1]，当R^2为 1 时，表示模型完美预测了数据；当R^2为 0 时，表示模型无法解释数据方差。在实际应用中，R^2决定系数通常用于比较不同模型的表现，取值越接近 1，表示模型解释的数据方差越小，表现越好。

在原始参数下各模型拟合效果如下图：

模型名称	Mae	Mse	Rmse	R2
XGBOOST	37.154	1.453	1083.3863825808626	0.6238415007785094
随机森林回归	0.482	0.740	840.6454971986391	0.7251488666920374
线性回归	1.275	1.284	1857.5698532078165	0.012491683034555612
决策树回归	0.000	3.866	929.0867520250849	0.6467246595558439

LGBM 回归	1.001	1.049	1473.2504501618341	0.3264812749158805

由上图对比可得出：决策树回归、随机森林回归和 XGBOOST 拟合效果相对较好。

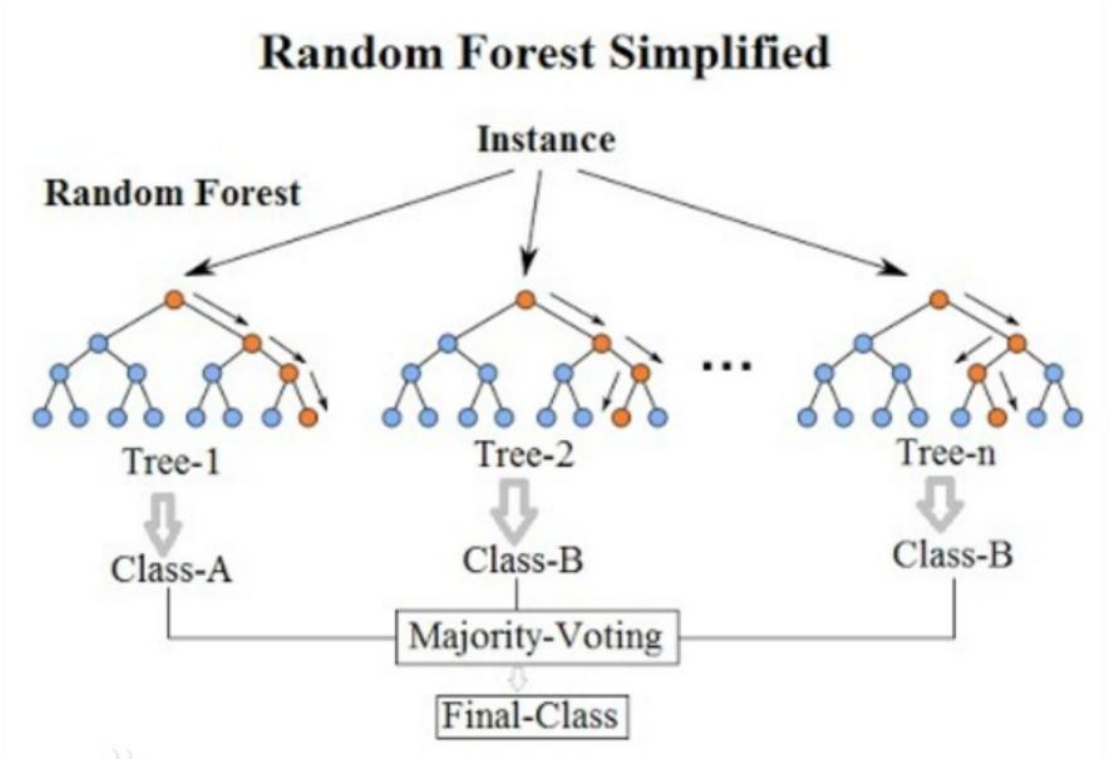
下面着重介绍随机森林、决策树回归以及 XGBOOST 回归模型：

3.3 机器模型

3.3.1 随机森林

随机森林（Random Forest, RF）是以 CART 作为基学习器的集成学习算法，利用多棵树对样本进行训练并预测的一种分类器，并且其输出的类别是由个别树输出的类别的众数而定，可以基于影响需求数据中的不同特征值进行分类，并利用该模型对未来需求量进行预测。

定义 1 RF 是一个由一组决策树作为基学习{ $h(X,\Theta_k);k=1,\cdots,K$ }组成的集成学习器，其基分类器是由 CART 学习算法构建的未剪枝的分类、回归树，其中{ $\Theta_k,k=1,2,\cdots,K$ }为独立同分布的随机向量，K 表示 RF 中基学习器的个数，在进行对训练数据的学习时，每个基分类器训练出一个结果，并根据相应的集成策略对各基分类器的结果进行集成。在回归问题中每棵决策树的结果进行加权平均作为最终模型的输出，在分类问题中，进行投票法输出结果。



随机森林流程示意图

构造随机森林的过程：

1.假如有 N 个样本，则有放回的随机选择 N 个样本(每次随机选择一个样本，然后返回继续选择)。这选择好了的 N 个样本用来训练一个决策树，作为决策树根节点处的样本。

2.当每个样本有 M 个属性时，在决策树的每个节点需要分裂时，随机从这 M 个属性中选取出 m 个属性，满足条件 m << M。然后从这 m 个属性中采用某种策略（比如说信息增益）来选择 1 个属性作为该节点的分裂属性。

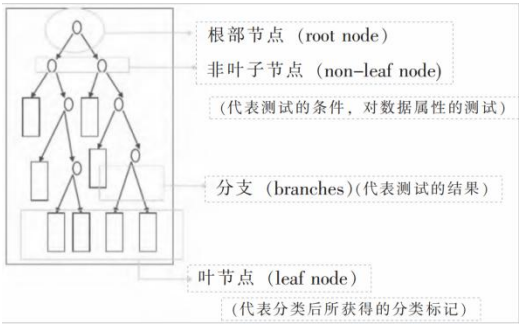
3.决策树形成过程中每个节点都要按照步骤 2 来分裂（已知，若下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到了叶子节点，无须继续分裂了）。一直到不能够再分裂为止。注意整个决策树形成过程中没有进行剪枝。

4.按照步骤 1~3 建立大量的决策树，这样就构成了随机森林。

3.3.2 决策树

决策树是一个预测模型；是通过分析寻找对象属性与对象值之间的一种映射关系，对数据进行处理。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表

示的对象的值，可以又分叉路径代表所有特征值，而基于根节点到叶节点的路径，反应不同的特征值的形成的组合，并以此训练，使得决策树对更多种组合进行分析，使预测需求更加精确。



决策树属于有监督学习，是一个典型的树模型，它体现了特征与标签的一种函数映射关系。决策树中非叶子节点表示样本在某个特征上的划分，根据样本集在该特征上的不同取值将其划分成若干个子树，直至叶子节点或满足某个条件就不再进行样本集分裂，每个叶节点对应着一个分类。构造一棵决策树的核心问题是如何选择一个合适的特征对当前样本集做拆分。决策树的典型的决策过程是从根节点开始，根据节点特征进行判断往哪棵子树进行决策分解，每次决策都采用新的特征(信息)，直到叶子节点或者不需要再进行决策。[5]

决策树学习过程可以分为 3 步，不同的决策树算法，区别在于每个步骤上进行了不同的考虑和优化，这 3 个步骤是：特征选择、决策树生成、决策树剪枝。

特征选择：这一步主要根据特征选择算法来选取合适的特征来对当前样本集进行划分。在训练样本集中，每个样本的特征可能有很多个，不同特征对分类标签的作用有大有小。因而特征选择的作用就是筛选出跟分类标签最相关性的特征。在特征选择中通常使用的准则是：信息增益、信息增益率、基尼指数。

信息增益：信息熵表示对随机变量 X 不确定性的度量。就是衡量一组数据是否不纯的指标。熵越高，信息越多；熵越低，信息越少。信息熵如公式（1）所示：

$$E(s) = \sum_{i=1}^{n} -p_i \, log_2 \, p_i \qquad (1)$$

其中 s 表示随机变量，c 表示随机变量的取值数量，pi 表示每个取值时的概率。当使用决策树中的节点将训练实例划分为更小的子集时，熵会发生变化。某个特征的信息增益是熵变化程度的量度。简单来说，就是信息增益是决定何时停止或不停止分裂的标准之一，如果熵的减少太小，则需要进一步拆分。在这种情况下，我们可以选择熵减少的较多的结果进行拆分，因为一直拆分会有过度拟合的风险。

基尼（Gin）i 指数:如果所有元素都被正确地划分为不同的类（理想情况），则该划分被认为是纯的。Gini 指数用于衡量随机选择的样本被某个节点错误分类的可能性。

Gini 指数的程度始终介于 0 和 1 之间，其中 0 表示所有元素都属于某个类别（或划分为纯的），1 表示元素随机分布在各个类中。Gini 杂质为 0.5 表示元素均匀分布到某些类别中。基尼杂质度量的数学符号如公式（3）所示：

$$\mathrm{Gini} = 1 - \sum_{i=1}^{n} p_i^2 \qquad (3)$$

其中 pi 是特定元素属于特定类别的概率。

通过从根节点开始计算信息增益，寻找信息增益最大的特征值，也就是说其是最好的分裂特征，所以在根节点使用该特征作为特征。

决策树生成：选好特征后，就从当前树的根节点出发，根据该特征的不同取值建立子节点；对每个子节点使用相同的方式生成新的子节点，直到信息增益很小（满足一定的阈值）或者当前样本集不需要再划分为止。
决策树剪枝：即为剪枝指砍掉考虑低重要性特征分支的过程。决策树剪枝的主要防止过拟合，通过提前去掉部分分支来降低过拟合的风险。剪枝的阈值设置是一个关键。[6]

决策树算法介绍：
（常见的决策树算法有 ID3，C4.5 和 CART，它们的主要区别是特征选择的算法不同。）
(1) CLS 算法：是最原始的决策树分类算法，基本流程是，从一棵空数出发，不断的从决策表选取属性加入数的生长过程中，直到决策树可以满足分类要求为止。CLS 算法存在的主要问题是在新增属性选取时有很大的随机性。
(2) ID3 算法:ID3 算法摒弃了对 CLS 算法的属性选择的随机性，利用信息熵的下降速度作为属性选择的度量。ID3 是一种基于信息熵的决策树分类学习算法，以信息增益和信息熵，作为对象分类的衡量标准。ID3 算法结构简单、学习能力强、分类速度快适合大规模数据分类。但同时由于信息增益的不稳定性，容易倾向于众数属性导致过度拟合，算法抗干扰能力差。
ID3 核心思想：ID3 算法根据样本子集属性取值的信息增益值的大小来选择决策属性(即决策树的非叶子结点)，并根据该属性的不同取值生成决策树的分支，再

对子集进行递归调用该方法，当所有子集的数据都只包含于同一个类别时结束。最后，根据生成的决策树模型，对新的、未知类别的数据对象进行分类。

ID3 算法优点：方法简单、计算量小、理论清晰、学习能力较强、比较适用于处理规模较大的学习问题。

ID3 算法缺点：倾向于选择那些属性取值比较多的属性，在实际的应用中往往取值比较多的属性对分类没有太大价值、不能对连续属性进行处理、对噪声数据比较敏感、需计算每一个属性的信息增益值、计算代价较高。

(2) C4.5 算法:C4.5 是基于 ID3 算法的改进，主要包括：使用信息增益率替换了信息增益下降度作为属性选择的标准；在决策树构造的同时进行剪枝操作；避免了树的过度拟合情况；可以对不完整属性和连续型数据进行处理；使用 k 交叉验证降低了计算复杂度；针对数据构成形式，提升了算法的普适性。

(3)CART（ClassificationandRegressionTree） :分类回归树算法(Classification and Regression Trees,简称 CART 算法)是一种基于二分递归分割技术的算法。该算法是将当前的样本集，分为两个样本子集，这样做就使得每一个非叶子节点最多只有两个分支。因此，使用 CART 算法所建立的决策树是一棵二叉树，树的结构简单，与其它决策树算法相比，由该算法生成的决策树模型分类规则较少。

CART 分类算法的基本思想是：对训练样本集进行递归划分自变量空间，并依次建立决策树模型，然后采用验证数据的方法进行树枝修剪，从而得到一颗符合要求的决策树分类模型。

CART 分类算法和 C4.5 算法一样既可以处理离散型数据，也可以处理连续型数据。CART 分类算法是根据基尼(gini)系数来选择测试属性，gini 系数的值越小，划分效果越好。

设样本集合为 T，则 T 的 gini 系数值可由下式计算：

$$\text{Gini（T）} = 1 - \sum_{i=1}^{n} p_i^2$$

其中，pj 是指类别 j 在样本集 T 中出现的概率。若我们将 T 划分为 T1、T2 两个子集，则此次划分的 gini 系数的值可由下式计算：

$$gini_{split}(T) = \frac{s_1}{s} gini_{split}(T_1) + \frac{s_2}{s} gini_{split}(T_2)$$

其中，s 为样本集 T 中总样本的个数，s1 为属于子集 T1 的样本个数，s2 为属于子集 T2 的样本个数。

CART 算法优点：除了具有一般决策树的高准确性、高效性、模式简单等特点外，还具有一些自身的特点。如，CART 算法对目标变量和预测变量在概率分布上没有要求，这样就避免了因目标变量与预测变量概率分布的不同造成的结果；CART 算法能够处理空缺值，这样就避免了因空缺值造成的偏差；CART 算法能够处理孤立的叶子结点，这样可以避免因为数据集中与其它数据集具有不同的属性的数据对进一步分支产生影响；CART 算法使用的是二元分支，能够充分地运用数据集中的全部数据，进而发现全部树的结构；比其它模型更容易理解，从模型中得到的规则能获得非常直观的解释。

决策树模型中 ID3 和 CART 算法

(1) ID3 算法流程

1)根节点的构造，遍历所有特征，找到那个使分类信息增益最大的特征，将其设置为根节点，并且将这个特征删除。

2)由于根节点已经将数据分叉，递归的方式寻找每个分枝的最优特征 3.

3)ID3 采用信息增益来选取最优分裂特征。

(2) CART 算法流程

CART 算法输入是训练集 D，基尼系数的阙值，样本个数闯值，输出是决策树 T 对训练样本集 D，计算所有的基尼杂质分数。

比较基尼杂质分数，选择杂质分数最低的特征对当前样本集进行分离。

3)重复上面步骤，直至决策树到叶子节点，或者达到剪枝的闽值，或者当前样本集已经不需再分 (无杂质)。

3.3.3 XGBOOST

XGBoost 的基本原理：

XGBoost 是基于预排序的决策树算法。基本思路就是不断生成新的树，每棵树都是基于上一颗树和目标值的差值来进行学习，从而降低模型的偏差。

其具体步骤是首先将所有特征按照特征的数值进行排序，在遍历分割点的时候有 O（#data）代价找到一个特征的最好分割点，将其分为左右节点。可以将不同特征值作为节点，快速准确地对供应商需求量进行预测。该模型通过自动调用多

线程并行计算，构建并结合多个弱学习器以实现高精度预测。将 XGBoost 与商品需求智能化评估需求相结合，具体的方法原理描述如下:

给定数据集 D＝{(xi，yi)} ，xi 和 yi 分别为输入数据和输出目标值，则模型可表示为:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)，f_k \in F$$

式中: $\hat{y}_i$为输出预测值; K 为回归树数量; $f_k(x_i)$为第 k 棵回归树预测样本 i 的值。

XGBoost 目标函数为:

$$f_{\text{Obj}} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

式中: n 为样本数量; l(•) 为误差函数，衡量预测值与目标值 yi 之间的误差; Ω(fk) 为正则化项，用于控制模型的复杂度，避免模型过拟合，其表达式为:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\kappa \parallel w \parallel^2 = \gamma T + \frac{1}{2}\kappa \sum_{j=1}^{T} w_j^2$$

式中: γ 为叶子结点的惩罚系数; κ 为正则项系数; T 和 w 分别为第 k 棵树的叶子数目和权重; wj 为第 j 个叶子的权重系数。

XGBoost 利用前向分步算法对目标函数训练，进而通过数据驱动方式建立输入数据与动态频率响应间的良好映射关系。[13]

其余机器学习算法介绍：

3.3.4 多元线性回归预测

线性回归预测法就是寻找变量之间的因果关系，并将这种关系用数学模型表示出来，通过历史资料计算这两种变量的相关程度，从而预测未来情况的一种方法。一元线性回归分析预测法，是根据自变量 x 和因变量 Y 的相关关系，建立 x 与 Y 的线性回归方程进行预测的方法。由于需求受多种因素的影响，而并不是仅仅受一个因素的影响。所以应用多元线性回归分析预测法，将多个特征值作为变量，以此构造多元线性回归模型，对未来需求量进行预测。[9]

3.3.5 LightGBM 回归预测模型

ghtGBM 模型是一个梯度提升决策树(GBDT)的实现，其本质原理就是利用基分类器(决策树)训练集成，得到最优的模型，与 XGBoost 类似，并在 XGBoost 基础上增加了直方图算法（解决了预排序算法效率低的问题）、带深度限制的 Leaf-wise 算法（解决了容易过拟合的问题）、单边梯度采样算法（计算信息增益会保留梯度大的样本并随机采样一些梯度小的样本，使得能保持原来数据的分布)、互斥特征捆绑算法（减少高纬度系数特征数据的损耗性）。

3.3.6 多层神经网络（MLP）

MLP(Multi-Layer Perceptron)，即多层感知器，是一种趋向结构的人工神经网络，映射一组输入向量到一组输出向量。MLP 可以被看做是一个有向图，由多个节点层组成，每一层全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元(或称处理单元)。

3.4 模型调参 GridSearchCV（网格搜索）

3.4.1 网格搜索（GridSearchCV）

GridSearchCV 的名字其实可以拆分为两部分，GridSearch 和 CV，即网格搜索和交叉验证。这两个名字都非常好理解。网格搜索，搜索的是参数，即在指定的参数范围内，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，这其实是一个训练和比较的过程。

GridSearchCV 可以保证在指定的参数范围内找到精度最高的参数，但是这也是网格搜索的缺陷所在，他要求遍历所有可能参数的组合，在面对大数据集和多参数的情况下，非常耗时。

3.4.2 Grid Search 网格搜索

Grid Search：一种调参手段；穷举搜索：在所有候选的参数选择中，通过循环遍历，尝试每一种可能性，表现最好的参数就是最终的结果。其原理就像是在数组里找到最大值。这种方法的主要缺点是比较耗时。

所以网格搜索适用于三四个（或者更少）的超参数（当超参数的数量增长时，网格搜索的计算复杂度会呈现指数增长，这时候则使用随机搜索），用户列出一个较小的超参数值域，这些超参数至于的笛卡尔积（排列组合）为一组组超参数。网格搜索算法使用每组超参数训练模型并挑选验证集误差最小的超参数组合。

3.5 模型训练

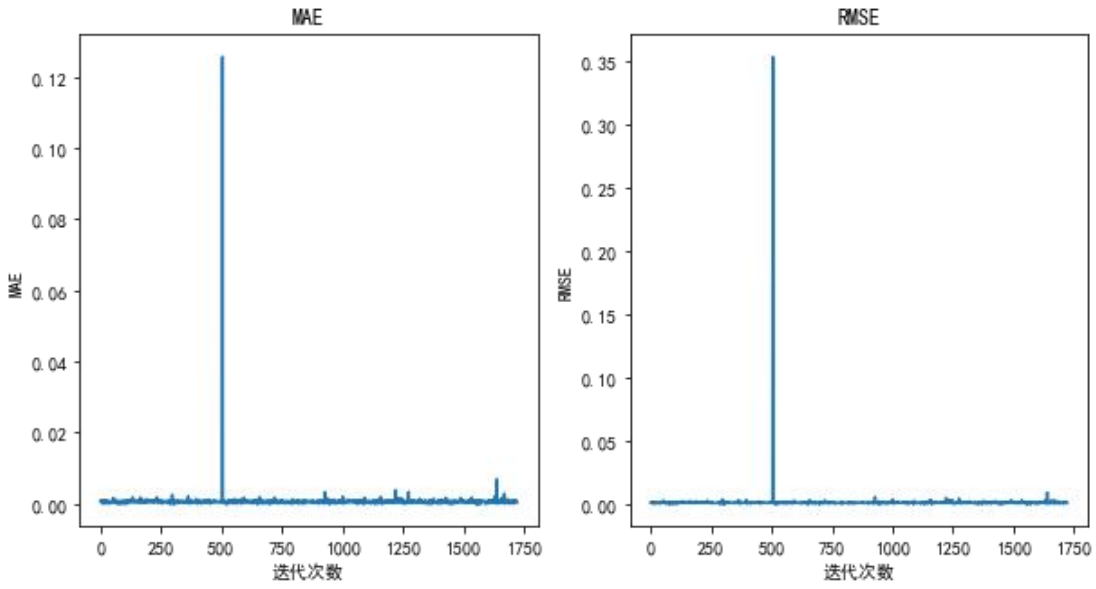
下面对 XGBOOST、 随机森林和决策树分别使用网格搜索和交叉验证进行参数调优

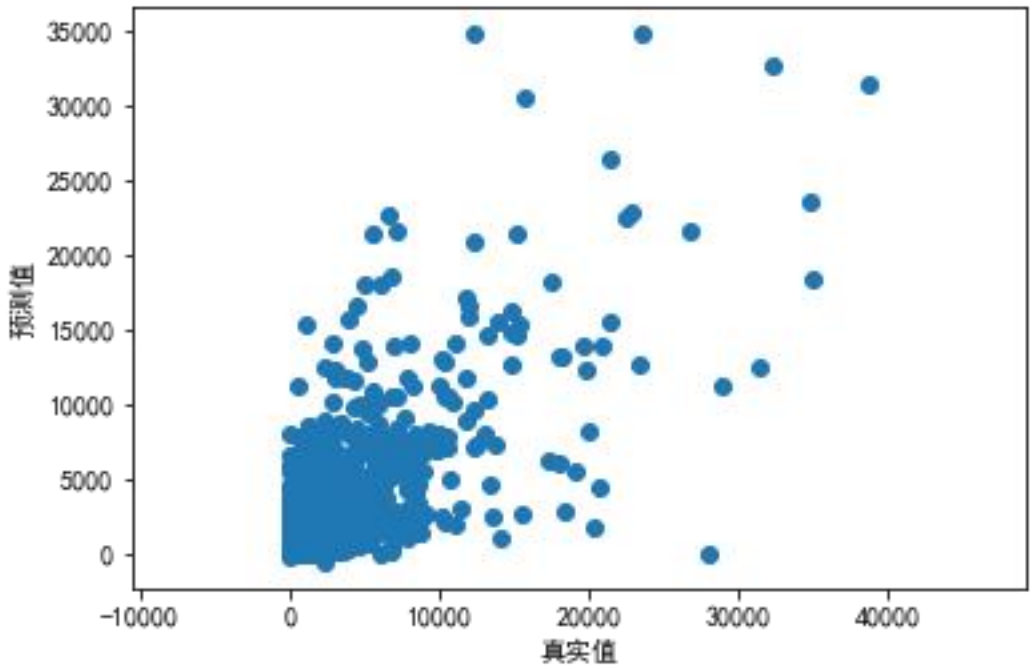
得到最优参数：

1.以月为时间粒度对上述三个模型分别用最优参数训练

下面对 XGBOOST 训练得到测试数据如下

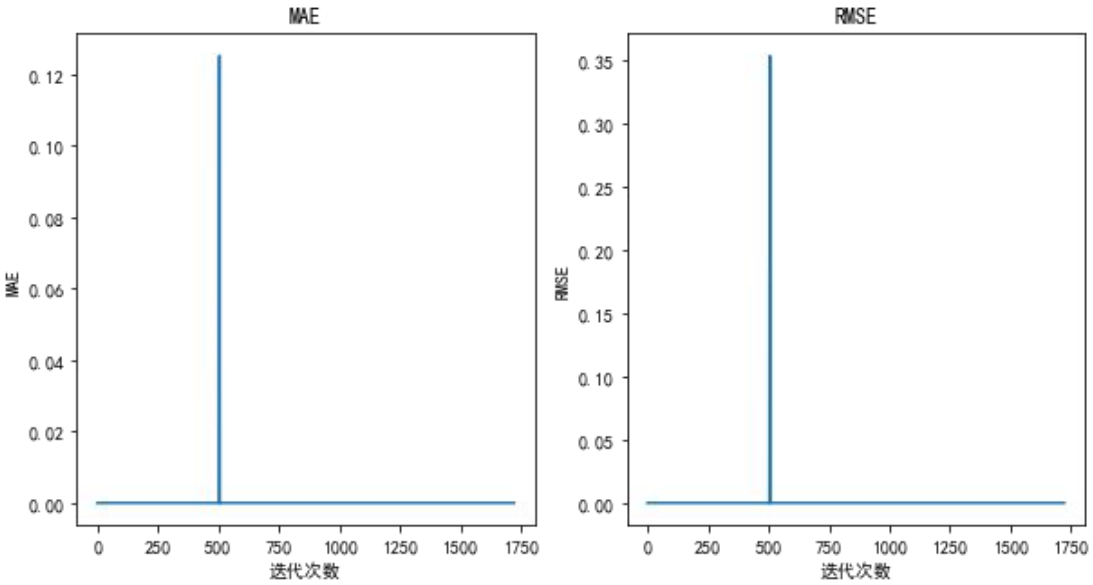
	组合	Mae	Mse	rmse	mape_val	r2	2019年1月预测需求量	2019年2月预测需求量	2019年3月预测需求量
0	10120002	0.000874	8.932974e-07	0.000945	0.002147	0.836705	19.467485	40.724682	19.000746
1	10120003	0.000565	5.971880e-07	0.000773	0.000209	0.956956	519.440613	246.365967	935.518555
2	10120006	0.000621	6.316665e-07	0.000795	0.000657	0.829795	145.258896	120.167206	156.157471
3	10120014	0.000642	1.044886e-06	0.001022	0.002921	0.832781	16.000307	2.005131	2.005131
4	10120016	0.000594	6.198534e-07	0.000787	0.000064	0.925147	1513.518677	147.859146	553.000366

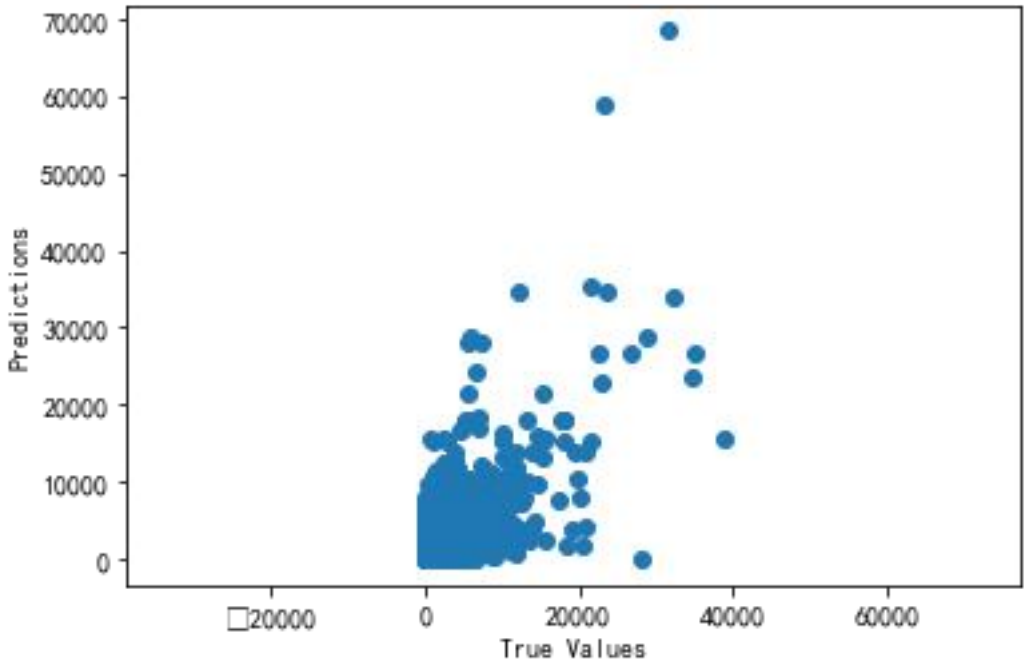




下面是对决策树回归模型训练得到的数据

	组合	Mae	Mse	Rmse	Mape_val	R2	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
0	10120002	0.0	0.0	0.0	0.0	0.779051	19.0	36.0	19.0
1	10120003	0.0	0.0	0.0	0.0	0.923881	661.0	85.0	661.0
2	10120006	0.0	0.0	0.0	0.0	0.560356	138.0	119.0	299.0
3	10120014	0.0	0.0	0.0	0.0	0.549115	16.0	2.0	2.0
4	10120016	0.0	0.0	0.0	0.0	0.873920	553.0	193.0	1603.0

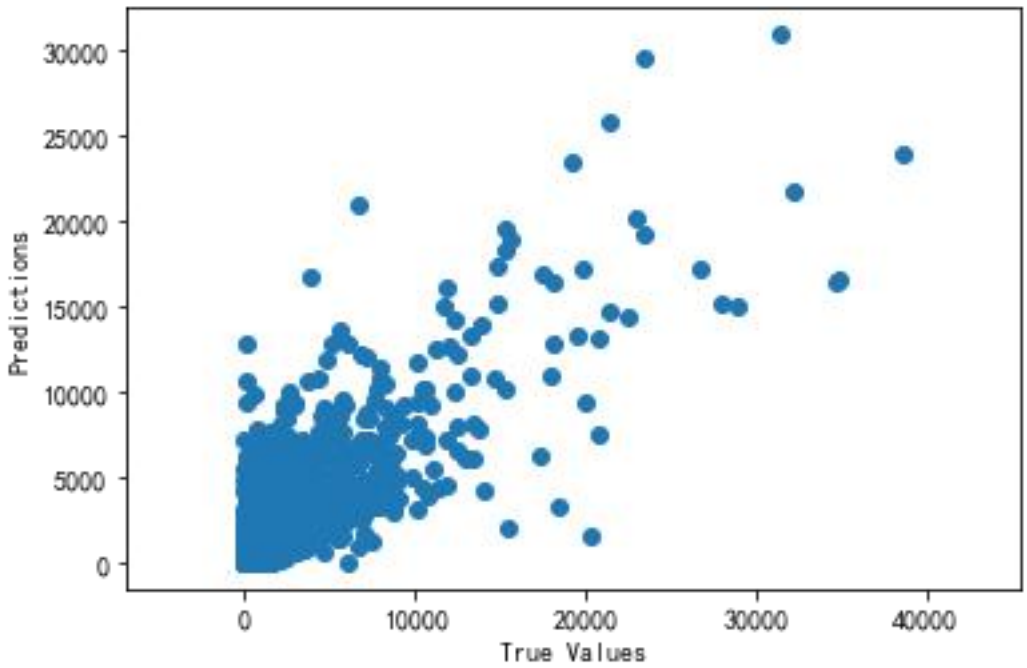




下面是对随机森林回归模型训练得到的数据

```
model_result3.head()
```

	组合	Mae	Mse	Rmse	Mape_val	R2	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
0	10120002	8.679000	1.153371e+02	10.739511	18.354134	0.805274	32.055	39.540	30.252
1	10120003	144.501389	3.916686e+04	197.906188	61.137395	0.741423	533.611	320.042	558.476
2	10120006	43.162611	3.522144e+03	59.347653	38.333875	0.808441	130.018	127.359	124.378
3	10120014	1181.037385	2.299933e+06	1516.552841	606.849863	0.821624	193.971	202.105	202.105
4	10120016	373.493300	3.093344e+05	556.178352	29.430104	0.862310	977.059	371.454	1001.156



想决定系数 R2 越接近 1，必须满足 MSE 越小，也就是真实值与预测值相差不大，

也就是模型拟合程度高，同时 var 方差越大，也就是我们的样本离散程度大，对应的我们实际采样过程中，就是要求样本是随机性，以及全面性，覆盖度广。因为数据量比较大，这里算出的评估指标均采用平均值。

模型名称	Mae	Mse	Rmse	Mape_val
XGBOOST	0.000767	0.000078	0.001117	0.002786
随机森林回归	295.602087	594098.856208	387.193014	186.199066
决策树回归	0.000077	0.000077	0.000218	0.000865

可以看出决策树在 MAE RMSE 平稳性，以及散点图分布更加集中在 Y=X 直线上，由此分析决策树回归拟合效果比 XGBoost 好。根据散点图可以看出随机森林比决策树的拟合效果好。

综上所述我们采用随机森林+GridSearchCV 自动寻参作为本题的模型。首先构造组合这一列(销售区域编码+产品编码)，方便后序对每个组合进行遍历。

2. 按月为时间粒度训练模型

构建年月这一列，方便后续进行分组，训练集如下：

	订单日期	日	周	月	销售区域编码	产品编码	产品大类编码	产品细类编码	订单需求量	组合	年月
0	2015-09-01	1	36	9	104	22069	307	403	19	10422069	2015-09
1	2015-09-01	1	36	9	104	20028	301	405	12	10420028	2015-09
2	2015-09-02	2	36	9	104	21183	307	403	109	10421183	2015-09
3	2015-09-02	2	36	9	104	20448	308	404	3	10420448	2015-09
4	2015-09-02	2	36	9	104	21565	307	403	3	10421565	2015-09
...	...	...	...	...	...	...	...	...	...	...	...
597689	2018-12-20	20	51	12	102	20994	302	408	59	10220994	2018-12
597690	2018-12-20	20	51	12	102	21875	302	408	502	10221875	2018-12
597691	2018-12-20	20	51	12	102	20215	302	408	106	10220215	2018-12
597692	2018-12-20	20	51	12	102	20195	302	408	187	10220195	2018-12
597693	2018-12-20	20	51	12	102	20321	302	408	205	10220321	2018-12

使用随机森林回归模型，对每组进行遍历，对于每组数据，提取订单需求量数据，

对数据进行偏移，使用前 3 个值预测后一个值，循环迭代三次，得到未来三个时间周期的预测值。

```
model_result3.head()
```

	组合	Mae	Mse	Rmse	Mape_val	R2	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
0	10120002	8.679000	1.153371e+02	10.739511	18.354134	0.805274	32.055	39.540	30.252
1	10120003	144.501389	3.916686e+04	197.906188	61.137395	0.741423	533.611	320.042	558.476
2	10120006	43.162611	3.522144e+03	59.347653	38.333875	0.808441	130.018	127.359	124.378
3	10120014	1181.037385	2.299933e+06	1516.552841	606.849863	0.821624	193.971	202.105	202.105
4	10120016	373.493300	3.093344e+05	556.178352	29.430104	0.862310	977.059	371.454	1001.156

经过随机森林训练后得到的预测数据

3. 按周为时间粒度训练模型

首先构建年和周这一列，方便后续进行分组

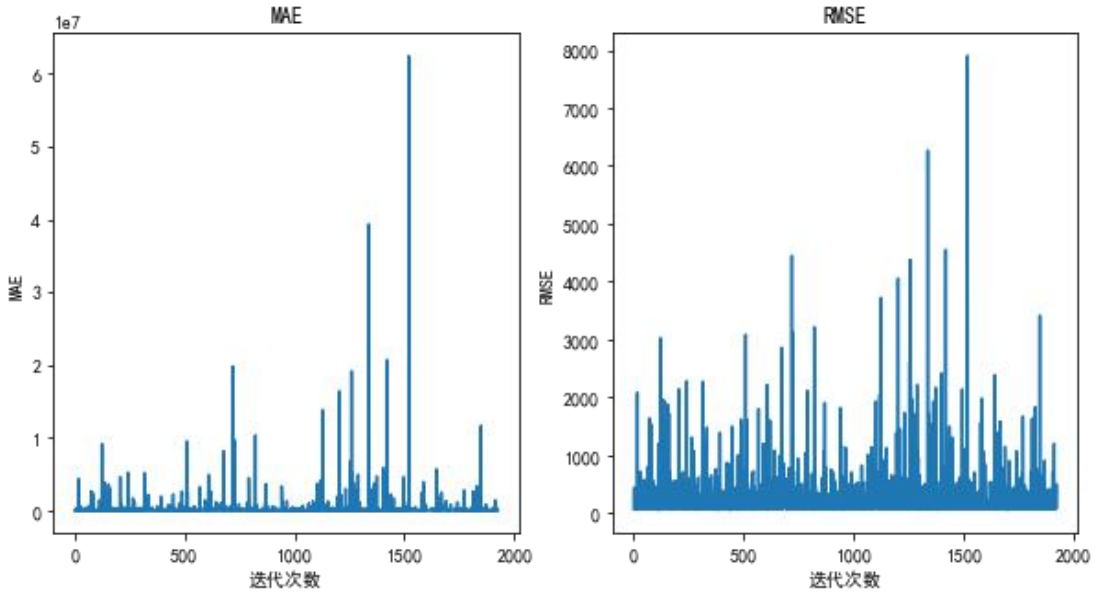
	订单日期	日	周	月	销售区域编码	产品编码	产品大类编码	产品细类编码	订单需求量	组合	年月	年周
0	2015-09-01	1	36	9	104	22069	307	403	19	10422069	2015-09	2015-36
1	2015-09-06	6	36	9	102	22046	305	412	208	10222046	2015-09	2015-36
2	2015-09-06	6	36	9	102	21379	305	412	84	10221379	2015-09	2015-36
3	2015-09-06	6	36	9	105	20007	305	412	1007	10520007	2015-09	2015-36
4	2015-09-06	6	36	9	105	21744	302	408	53	10521744	2015-09	2015-36
...	...	...	...	...	...	...	...	...	...	...	...	...
597689	2018-02-27	27	9	2	102	21253	306	407	305	10221253	2018-02	2018-9
597690	2018-02-27	27	9	2	103	20244	308	404	10	10320244	2018-02	2018-9
597691	2018-02-27	27	9	2	103	20704	307	403	56	10320704	2018-02	2018-9
597692	2018-02-27	27	9	2	103	21940	303	401	14	10321940	2018-02	2018-9
597693	2018-03-04	4	9	3	101	21394	306	407	142	10121394	2018-03	2018-9

使用随机森林回归模型，对每组进行遍历，对于每组数据，提取订单需求量数据，对数据进行偏移，使用前 3 个值预测后一个值，循环迭代 12 次，将本次得到的数据放回训练数据，以前一周预测第二周，再以第一周第二周预测第三周，通过迭代得到 13 周的数据。

经过训练后得到的预测数据，将得到的数据集前四周的数据加上第五周前四天的数据得到第一个月的需求总量，将第四周的数据后三天、第五周、第六周、第七周、第八周前四天加起来得到第二个月的需求总量。将第八周后三天、第九周、第十周、第十一周、第十二周的订单需求量加起来得到第三个月的需求总量。

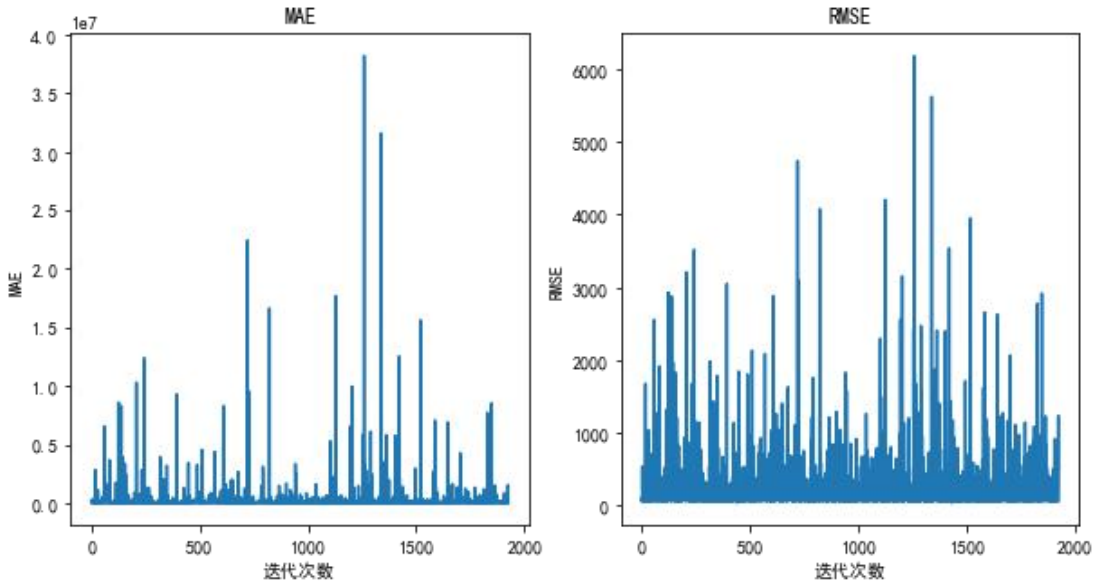
下表（图）为按周为时间粒度的预测数据（XGBoost）：

	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
0	101	20002	109.106613	90.784309	93.142298
1	101	20003	1006.281119	976.636773	1074.685425
2	101	20006	285.581105	187.346033	278.414173
3	101	20014	5538.070243	5138.464155	4461.210473
4	101	20016	676.786045	1199.805828	1380.848300
...	...	...	...	...	...
1889	105	22059	54.193289	38.449593	37.353132
1890	105	22066	6455.089379	5057.263951	5919.624041
1891	105	22072	2002.267862	1375.545421	1535.453250
1892	105	22083	1843.554143	1927.737015	2171.412517
1893	105	22084	231.652575	266.359993	316.061291



下表（图）为按周为时间粒度的预测数据（随机森林）：

	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
0	101	20002	89.028143	82.710857	90.003000
1	101	20003	555.356143	588.222143	969.925714
2	101	20006	302.791429	286.217286	309.832286
3	101	20014	5907.297143	6006.314286	5160.226571
4	101	20016	1092.263143	1171.867571	1149.268286



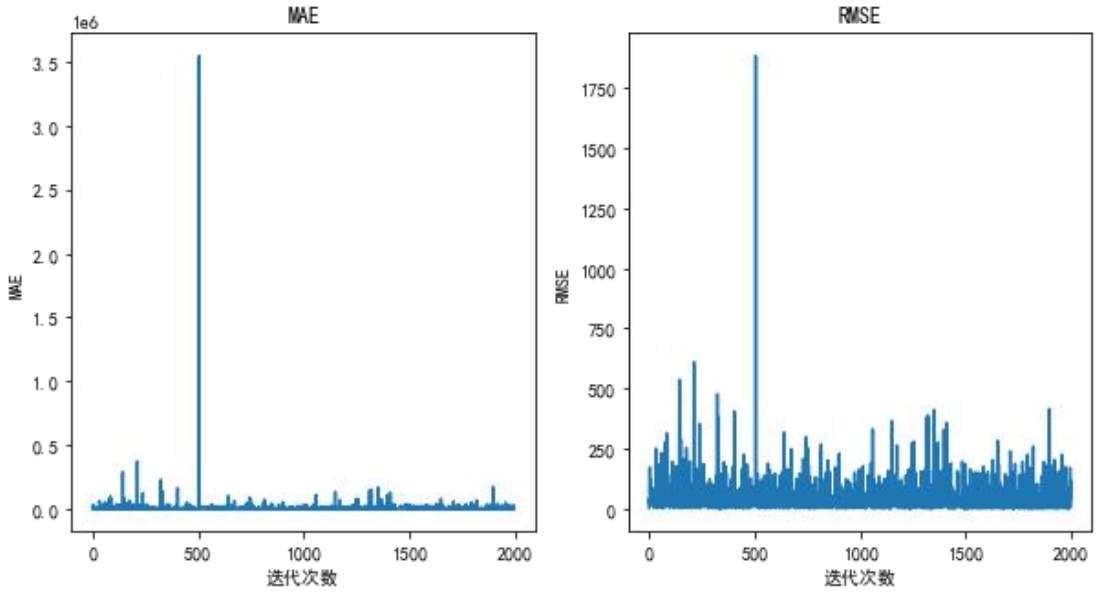
4. 按天为时间粒度训练模型

首先构建年天这一列，方便后续进行分组，使用随机森林回归模型，对每组进行遍历，对于每组数据，提取订单需求量数据，对数据进行偏移，使用前 3 个值预测后一个值，循环迭代 90 次，将本次得到的数据放回训练数据，以前一天预测第二天，再以第一天第二天预测第三天，通过迭代 90 次得到 90 天的数据。经过训练后得到的预测数据。

	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
0	101	20002	409.19900	362.077000	405.261000
1	101	20003	4137.86500	3615.790000	4152.234000
2	101	20006	1442.98000	1242.243000	1391.882000
3	101	20014	9421.52970	12454.556333	14606.213333
4	101	20016	4372.59121	4782.327288	5106.450564
...	...	...	...	...	...
1967	105	22059	255.34400	233.352000	258.626000
1968	105	22066	14185.85600	12864.789000	14892.826000
1969	105	22072	1576.06800	1291.828000	1431.017000
1970	105	22083	3632.40100	3203.406000	3518.912000
1971	105	22084	267.83900	249.766000	274.553000

取数据前 31 天作为第一个月的需求总量，取数据 31 天-59 天作为第二个月的需求总量，取数据 59-90 天作为第三个月的需求总量，得到按天为时间粒度的预测数据（XGBoost）。

	标识	训练集Mape	测试集Mape	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
0	10120002	0.000129	0.253494	349.830118	321.535332	444.200970
1	10120003	0.000413	4.765406	3989.474098	2781.312044	4423.372459
2	10120006	0.000035	5.870019	1974.464196	1258.383326	1824.705070
3	10120014	0.010881	2.670774	12406.372803	11011.595015	20773.196991
4	10120016	0.040092	1.077220	4637.293388	3725.857956	6210.673431
...	...	...	...	...	...	...
1967	10522059	0.000113	0.102159	356.583781	342.735027	478.496788
1968	10522066	0.024811	3.849449	16235.634228	15213.016216	20062.666468
1969	10522072	0.000022	5.489137	2622.147446	2603.978308	3463.232567
1970	10522083	0.003644	0.763426	2394.587674	2336.379478	3076.469095
1971	10522084	0.000119	0.563401	902.261551	871.788376	1182.857028



通过上图可知，训练集和预测集的平均绝对百分比误差数值大部分都在【0,1】区间内，数值较小，说明数据拟合效果较好，预测精度比较高。

4.总结与建议

4.1 总结

随着信息及云存储技术的发展，企业产生了大量的生产数据。数据的价值逐渐体现。尤其是以机器学习和深度学习改变了这一状况。随着人们看到大数据带来的增值效益，迫切需要建立大数据部门，利用大数据技术提高现有企业的核心竞争力，辅助企业决策。在大数据技术中，特征工程是机器学习的核心内容，如何选择好的特征，对模型的构建至关重要。所以对特征工程的研究和集成学习算法的研究，适应了现如今的社会发展需求，因而具有现实意义。本文对集成学习模型的商品需求预测的研究及创新点如下：

首先，通过对初始数据集进行数据清洗，其次通过箱线图观察其存在异常值，通过 3σ 原则对数据进行预处理，去除异常值，其次利用直方图、饼图、散点图等常用数据统计分析方法对数据集的对影响需求特性的特征量进行画图分析，找寻其与需求量的影响关系。

第二，进行特征工程搭建环节，首先对前期分析的影响因素进行特征构建，使用 one-hot 编码对特征来进行编码，该过程不仅对数据集已有的特征量进行编码，还要新增前期分析得出的对需求有影响但未再原有数据集的特征量，并对新特征量进行编码。

第三，进行预测模型搭建环节，首先采用留出法对数据集划分为训练集与测试集，我们将预备选择的模型大致分为了两类：时间序列与非时间序列模型，时间序列模型采用了主流的 ARIMA 模型与 Lstm 模型，但是预测结果发现测试集中存在训练集没有的新产品，预测结果不太理想。非时间序列模型使用了主流的非时间序列参数模型可以对其他特征值进行参数设置，考虑时间因素的同时，使预测结果更加精确，因而选用主流的 XGBOOST、随机森林回归、线性回归、决策树回归、LGBM 回归、神经网络（MLP）六种模型进行预测。最后通过 Mae、Mse、Rmse、R2 四种评估标准发现，决策树回归、随机森林回归、XGBOOST、拟合效果相对较好，并对 XGBOOST、随机森林、决策树分别使用网格搜索和交叉验证进行参数调优得到最优参数，再次进行预测，得到随机森林以月为颗粒度进行预测，训练集 Mape 值为 0.310582，测试集 Mape 值为 1.492073；以周为颗粒度进行预测，训练集 Mape 值为 0.245801，测试集 Mape 值为 1.451065；以日为颗粒度进行预测，训练集 Mape 值为 0.326993，测试集 Mape 值为 1.206496。xgboost 以日为颗粒度进行预测，训练集 Mape 值为 0.017077，测试集 Mape 值为 3.020385；以周为颗粒度进行预测，训练集 Mape 值为 0.000439，测试集 Mape 值为 2.412107；以月为颗粒度进行预测，训练集 Mape 值为 0.000028，测试集

Mape 值为 3.060715。发现随机森林拟合程度更好，预测精度更优。基于此，选择随机森林+GridSearchCV 作为本题预测模型，分别通过日周月三种时间粒度进行产品需求预测。

尽管产品的需求往往受到季节和环境的影响，很难实现零库存成本。但随着大数据时代的到来，可以利用历史数据描绘商品需求模式，实现对商品的准确预测，可以降低仓储成本，改变传统的体验式企业需求预测，为整个供应链过程提供辅助决策。

本文重点研究了特征选择、集成学习模型融合和验证相关方法在商品中的应用，本文分析了考虑了外部环境的不稳定性的影响。因此，这对降低库存成本、优化供应链决策、提高核心竞争力、降低能源消耗具有重要意义。本文可以为供应商提供理论和实践支持。

5.展望

高质量发展是全面建设社会主义现代化国家的首要任务。高质量发展就是能够更好满足人民日益增长的美好生活需要的发展，就是创新成为第一动力、协调成为内生特点、绿色成为普遍形态、开放成为必由之路、共享成为根本目的的发展。[1]

高质量发展需要深化供给侧结构性改革，只有持续深化供给侧结构性改革，不断实现经济发展的质量变革、效率变革、动力变革，才能推动我国经济走上更高质量、更有效率、更加公平、更可持续、更为安全的发展;在供给侧改革的基础上，扩大内需，实现生产、分配、流通、消费循环畅通。利用大数据技术对企业的历史数据进行分析，通过机器学习与深度学习，对商品带的需求量进行预测，既有利于科技助力企业进行供给侧改革，又有利于将生产过程科学化智能化，提升生产效率。

因此应推动各企业进行数字化改革，发掘数据的经济价值，将深度学习的商品预测推广至更多行业，合理安排生产活动，推进产业智能化、现代化发展，让科技发展成果惠及更多的人，对于企业，基于深度学习的商品预测方式有利于减轻企业仓储成本，避免过多的货物囤积，减少资源浪费；对于消费者，有利于购买到生产日期更为新鲜的产品，提升购买体验。因此国家应更多支持传统行业与新兴科技行业的融合发展，促进产业现代化，推动高质量发展。

参考文献：

[1]刘良军.更好地统筹供给侧结构性改革和扩大内需[J].三晋基层治理,2023(01):20-23.
[2]丁敬安. 特征选择下的集成学习模型商品需求预测及分仓研究[D].安徽工业大学,2018.
[3]蔡尉尉. 基于多模态数据B电商平台商品需求预测与库存控制研究[D].中南林业科技大学,2022.DOI:10.27662/d.cnki.gznlc.2022.000725.
[4]邱萍萍. 供应链情境下基于销售记录的商品需求预测[D].华南理工大学,2020.DOI:10.27151/d.cnki.ghnlu.2020.001961.
[5]黄世反. 基于决策树的档案文本自动分类算法研究[D].云南大学,2015.
[6]陈旭. 一种改进的决策树分类算法[D].华中师范大学,2016.
[7]赵柯. 面向离散属性的决策树分类方法研究[D].大连海事大学,2017.
[8]王栎桥. 基于智能算法的神经网络训练研究[D].贵州大学,2021.DOI:10.27047/d.cnki.ggudu.2021.002544.
[9]钱莹,方秀男.多元线性回归模型及实例应用[J].中国科技信息,2022,No.669(04):73-74.
[10] 袁志聪.人工智能——随机森林技术分析[J].科技创新与应用,2020,No.298(06):151-152.
[11] 汪靖翔.决策树算法的原理研究和实际应用[J].电脑编程技巧与维护,2022,No.446(08):54-56+72.DOI:10.16184/j.cnki.comprg.2022.08.043.
[12] 舒仕文.LightGBM模型及其应用[J].信息记录材料,2022,23(07):219-222.DOI:10.16009/j.cnki.cn13-1295/tq.2022.07.003.
[13] 于琳琳,王泽,郝元钊,晏昕童,张丽华,严格,文云峰.基于XGBoost的电力系统动态频率响应曲线预测方法[J].电力建设,2023,44(04):74-81.