线下门店需求预测模型设计

香港中文大学(深圳)---全棉时代---深圳市大数据研究院

日期：02/21/23

目录

图目录

表目录

一、总体模型思路

1. 建模目标

批注 [R1]: 是不是可以采用一些比较数学的方式去描述问题，达到目标。

批注 [2R1]: 不可以，没必要

需求预测的准确率直接影响到公司的固定资本、库存周转率、运营成本等重要的财务指标，如何提升需求预测的准确率一直是很多企业供应链管理的重中之重。为了提高全棉时代线下门店补货过程的效率，以及降低补货过程中，由于对需求的错误结论带来的库存成本增加以及需求流失等问题。需要在对全棉时代业务逻辑进行准确把握的基础上，充分的收集和分析各种影响用户购买需求的直接和间接的因素，以构建门店的需求预测模型。

2. 整体规划

本次建模涉及全国线下所有的门店内裤和婴儿连体服三级品类下的全部SKU，涉及的区域广，产品数量多，以及因素组成复杂等问题。针对此，设计如图 1 所示的需求预测建模思路。在该思路中，主要包含数据处理、特征工程、数据拆分和模型建立及评价四个主要步骤。其中，数据处理表示对来自公司内部和外部收集的相关数据进行合并和清洗；特征工程则使用可视化、统计分析方法检验各种可能影响需求的因素的效果；数据拆分则基于有监督分类和无监督聚类手段对数据集进行拆分便于建模，增加模型的可解释性；而模型建立和评价则借助机器学习模型预测产品需求，并使用平均预测误差（MAPE）和平均预测准确率（1-MAPE）评价模型的泛化效果以确定最终的模型。

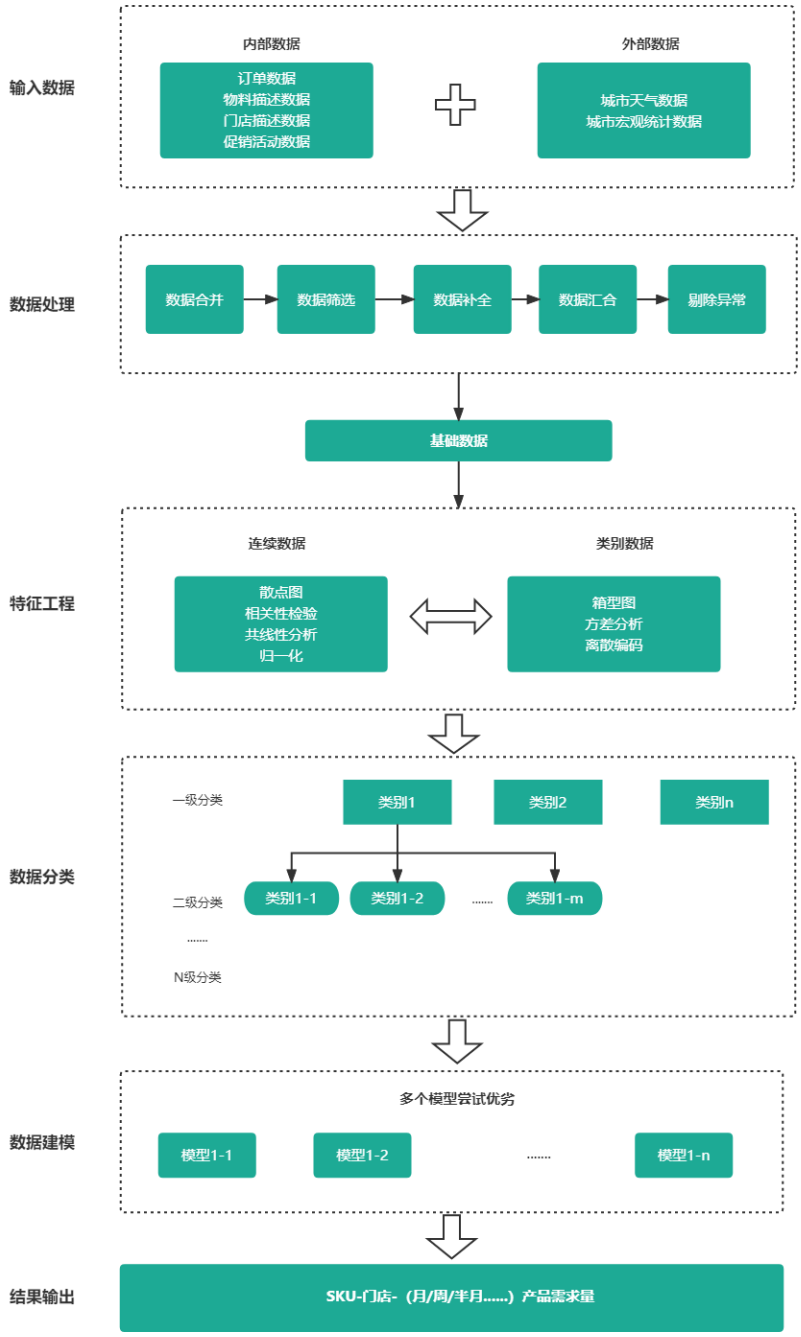


图 1 需求预测建模整体技术路线

初步选定的预测算法及其基本描述如下表所示：

表 1 备选机器学习算法及描述

模型名称	模型简介	优缺点
Lasso	在拟合广义线性模型的同时进行变量筛选和复杂度调整。	可以有效的避免过拟合。
XGBoost	XGBoost 是经过优化的分布式梯度提升库，旨在高效、灵活且可移植，是大规模并行 boosting tree 的工具，它是目前最快最好的开源 boosting tree 工具包，比常见的工具包快 10 倍以上。	XGBoost 对损失函数进行了二阶泰勒展开，以增加精度。预排序过程的空间复杂度过高，不仅需要存储特征值，还需要存储特征对应样本的梯度统计值的索引。
Decision Tree	决策树是一种树形结构，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。	决策树算法易理解，机理解释起来简单，可以用于小数据集。对连续性的字段比较难预测，容易出现过拟合。
Random Forest	随机森林本质上属于机器学习的一大分支——集成学习，是将许多棵决策树整合成森林并用来预测最终结果的方法。	能够处理很高维度的数据，并且不用做特征选择。随机森林在解决回归问题时，没有像它在分类中表现的那么好。
MLP	多层感知机是一种前向结构的人工神经网络，包含输入层、输出层及多个隐藏层。	在非线性数据上表现非常好。容易过拟合，计算复杂度和网络复杂度成正比，可解释性不强。
Gradient Tree Boosting(GBDT)	梯度提升树，是属于集成算法中 boosting 类的一种算法。这个算法是现有机器学习算法中相对较实用的	可以灵活处理各种类型的数据，包括连续值和离散值。由于弱学习器之间存在依赖

	算法。	关系，难以并行训练数据。

3. 结果初览

此次建模，采用分层分类建模预测的思想，分别在城市、店铺等级和省份三个维度下对内裤和婴儿连体服三级品类进行了分类建模，其中，**内裤则从店铺等级的角度出发，**将同一等级下的门店销售数据汇总建模；**婴儿连体服则基于省份维度**。相关预测结果汇总如表 2 所示，从表中可以看出，在单个 sku-单门店-周的颗粒度下，**内裤品类的预测平均准确率为 72%，婴儿连体服品类的预测平均准确率为 70%。**从上述结果可以看出，由于内裤和婴儿连体服历史数据本身的波动情况比较小，机器学习模型有更大的优势来刻画其需求趋势。因此，**在未来需求预测过程中，需要结合数据的波动情况来选择合适的模型。**

批注 [R3]: 待验证

批注 [4R3]: 按交付来

表 2 需求预测建模结果汇总

三级品类	划分簇	14 天移动平均准确率	机器学习建模准确率	提升情况
内裤	店铺等级	13%	(Decision Tree)72%	454%
婴儿连体服	省份	27%	(Decision Tree)70%	159%

批注 [R5]: 填写读取数据范围，预测范围，待验证

批注 [6R5]: 按交付来，这是初级版本，不是模型开发文档，主要是选特征

二、内裤和婴儿连体服三级品类模型设计

1. 内裤品类

1.1 数据预处理

1.1.1 数据合并

首先，合并公司内部数据。按照数据之间的主键，将订单主表、订单明细表、物料表和店铺表进行拼接，以获得该品类下各个 SKU 的历史数据，拼接过程如图 2 历史数据拼接过程所示。具体筛选步骤如下所示。

（1）选出内裤品类的所有订单

14160050 条

（2）删除 skc_code 以 LUN、LDS 开头的 skc 的订单

```
Sku = Sku[Sku["skc_code"].str.startswith("LUN") == False]
Sku = Sku[Sku["skc_code"].str.startswith("LDS") == False]
'''剔除skc_code中含有“LUN”、“LDS”的skc'''
```

13690935 条

（3）删除 shop_id 以 4 和 6 为开头的订单

```
Order_selectp=Order_selectp[Order_selectp["shop_id"].str.startswith("4")==False]
Order_selectp=Order_selectp[Order_selectp["shop_id"].str.startswith("6")==False]
'''删除shop_id为4or6开头的'''
Order_selectp = Order_selectp[Order_selectp.division_code == 10000]
'''筛选出线下门店的订单数据'''
```

13008612 条

（4）删除订单明细表中 sku_cnt 小于 0 的部分

```
ORDER_DETAIL.drop(ORDER_DETAIL[ORDER_DETAIL.sku_cnt <= 0].index, inplace=True)
'''删除销量小于0部分的订单'''
```

11688936 条

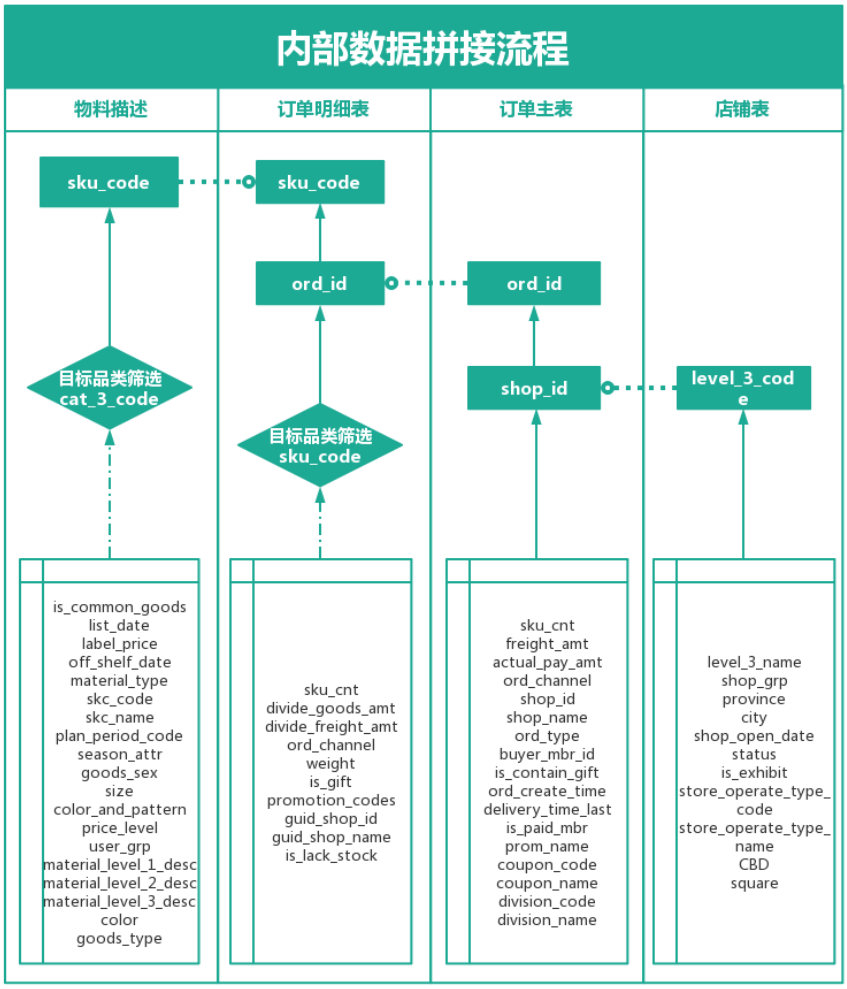


图 2 历史数据拼接过程

然后，合并其他内外部数据，包括 2020-2022 公司一二三级促销行事表、品类陈列数据、国家法定节假日、中国春夏秋冬季节划分、2020-2023 中国城市月平均气温等数据、最低折扣计划数据。按照订单创建时间（ord_create_time）和门店所在城市（city）作为主键将相关字段合并。

1.1.2 数据预处理

（1）数据情况

批注 [R7]: 为什么没有用门店的库存（含在途）信息？

批注 [8R7]: 你们的模型你们可以用，非常建议，将销售还原

批注 [R9]: 合并其他内部数据多少条

批注 [10R9]: 不影响品类拓展

按照 division_code=10000，is_exhbit=N 标准筛选出在营业的线下门店（非展销）的所有订单数据：

```
Order_selectp = Order_selectp[Order_selectp.division_code == 10000]
'''筛选出线下门店的订单数据'''
Shop = SHOP[SHOP.is_exhibit != 'Y']
'''只保留非展销门店'''
```

最终获得全国 74 个城市的 366 个门店的总计 5691183 条历史订单数据，其中涉及 4240 个 SKU（1192 个 SKC），将筛选出的数据剔除值唯一的列和空缺值在 80%以上的字段后，具体字段剔除说明如表 3 所示。

表 3 内裤品类剔除字段说明

序号	字段	描述	剔除原因
1	Is_contain_gift	是否包含赠品	值唯一
2	Division_code	事业部编码	值唯一
3	Division_name	事业部名称	值唯一
4	status	店铺状态	值唯一
5	Is_exhibit	是否展销门店	值唯一
6	Receiver_province_code	收货人省份编码	空缺值大于 80%
7	Receiver_city_code	收货人城市编码	空缺值大于 80%
8	Receiver_area_code	收货人所属县编码	空缺值大于 80%
9	Receiver_province_name	收货人省份名称	空缺值大于 80%
10	Receiver_city_name	收货人城市名称	空缺值大于 80%
11	Receiver_area_name	收货人所属县_名称	空缺值大于 80%

（2）按“SKU-门店-周”groupby-resample

- 销量值异常

首先共有 195 条销量<=0 的数据，将其剔除，此时还有 5691183 条数据。

通过对销量值的分布密度情况，如图 3 左所示，从该图中可以发现，历史销量值大多较小，比较密集的集中在 0-30 之间，而大于 30 的部分占比较低。

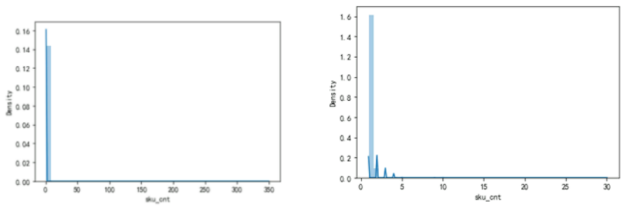


图 3 内裤品类销量密度函数图

- 将各个产品按照“SKU-门店”的维度进行分类汇总，然后再按照 ord_create_time 以每周一（’W-MON’）的维度进行下采样，以获得各个 sku 在每一个门店，每一周的历史销量数据。其中汇总时候，各个字段的汇总方式如图 4 所示。

```
'''将原始的订单数据按照sku-shop-week的格式重新分类汇总，即每个sku-门店-周下的所有订单要合并为一条数据'''
def Week_resample(self,df):
    groupby_columns = ['sku_code', 'shop_id']
    '''合并时候作为关键主键的字段'''
    resample_columns = ['ord_create_time']
    '''时间字段，用来重新归总'''
    sum_columns = ['sku_cnt']
    '''合并时候求和的字段'''
    min_columns = ['prom_level_new']
    '''合并时取最小值的字段，促销等级的值越小，促销等级越大'''
    data1 = df[groupby_columns + resample_columns + sum_columns]
    data1.loc[:, 'ord_create_time'] = pd.to_datetime(data1.loc[:, 'ord_create_time'])
    d1 = data1.groupby(groupby_columns).resample(self.frequency, label='left', closed='left',
                                                 on="ord_create_time").agg({x: "sum" for x in sum_columns})
    d1 = pd.DataFrame(d1).reset_index()
    '''合并第一个数据集'''
    data2 = df[groupby_columns + resample_columns + min_columns]
    data2.loc[:, 'ord_create_time'] = pd.to_datetime(data2.loc[:, 'ord_create_time'])
    d2 = data2.groupby(groupby_columns).resample(self.frequency, label='left', closed='left',
                                                 on="ord_create_time").agg({x: "min" for x in min_columns})
    d2 = pd.DataFrame(d2).reset_index()
    '''合并第二个数据集'''
    group_week_reault = pd.concat([d1, d2[min_columns]], axis=1)
    '''将按照不同方式合并的数据拼接在一起'''
    return group_week_reault
```

图 4 内裤品类字段汇总及其方式

汇总后，数据总共包含 8782572 条订单数据。

批注 [R11]: 这里为什么比步骤（1）5691183 筛选完的有效总记录还要多？

批注 [12R11]: 下采样，没有销售的周要补全

（3）数据补全

数据中，存在缺失值的字段说明如表 4 所示：

表 4 内裤品类包含缺失值字段说明

序号	字段	描述	缺失原因	是否补全
1	price_level	价格档位	包含 unk	经与业务沟通补全
2	color	颜色	包含 unk	无需补全
3	shop_name	店铺名称	包含 unk	不使用，无需补全
4	ord_type_new	订单类型（用以剔除展销、团购）	包含 unk	不使用，无需补全
5	buyer_mbr_id	买家会员 id	包含 unk	不使用，无需补全
6	delivery_time_last	最后出库时间	包含 unk	不使用，无需补全
7	receiver_country_code	收货人国家编码	包含 unk	不使用，无需补全
8	is_paid_mbr	是否付费会员	包含 unk	不使用，无需补全
9	prom_level_new	当前周包含的最大促销等级	包含空值	‘IV’补全
10	avg_temperature	平均温度	包含空值	均值补全
11	temperature_diverge	温度偏移指数	包含控制	

- price_level 的补全

价格档位显示为 unk 的如下所示：

	sku_name	sku_cnt
0	21秋女童平角裤|130/65|蓝底星星+太空流星|2条装	1
1	21秋男童平角裤|120/60|太空恐龙+全棉条纹|2条装	1813
2	21秋男童平角裤|110/60|太空恐龙+全棉条纹|2条装	1944

经与业务确认补全为高价格档产品。补全后价格等级与吊牌价对应关系如下图所示：

```
if len(df[df['price_level'] == 'unk']) > 0:
    '''价格水平补全'''
    low = df[df['price_level'] == '低']['label_price'].max()
    upper = df[df['price_level'] == '高']['label_price'].min()
    df.loc[(df['price_level'] == 'unk') & (df['label_price'] <= low), 'price_level'] = '低'
    df.loc[(df['price_level'] == 'unk') & (df['label_price'] > low) & (
            df['label_price'] <= upper), 'price_level'] = '中'
    df.loc[(df['price_level'] == 'unk') & (df['label_price'] > upper), 'price_level'] = '高'
```

补全后的结果如下图所示：

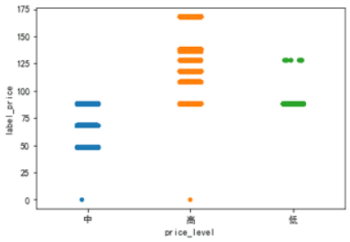


图 5 内裤品类补全后价格水平与吊牌价对应关系

- prom_level_new 的补全

缺失值表示在订单创建日期及对应的渠道下不包含任何的促销活动，因此用”IV”标签替代空值，即四级促销活动：

```
df['prom_level_new'] = df['prom_level_new'].fillna('IV')
'''促销等级补全'''
```

- avg_temperature 的补全

使用门店所在城市对应的年平均温度来补全：

```
if sum(df['avg_temperature'].isnull()) > 0:
    '''温度补全'''
    for c in list(df.city.unique()):
        df.loc[(df['avg_temperature'].isnull().values == True) & (df['city'] == c), 'avg_temperature'] = \
        df[df.city == c]['avg_temperature'].mean()
```

- waist 的补全

使用中腰来补全

```
if self.Cat=='内裤':
    df['waist'] = df['waist'].fillna('中腰')
    '''内裤腰型补全'''
```

- partition 的补全

使用D区陈列来补全

```
df['exhibited_area'] = df['exhibited_area'].fillna('D')
'''#陈列补全'''
```

1.2 特征工程

1.2.1 指标分类

根据数据字段的组成，将相关的指标进行分类，如表 5 所示，便于后期的分析和处理。

表 5 内裤品类指标分类表

连续型	离散型	其他
label_price	price_level	ord_create_time
divide_good_amt	user_grp	shop_open_date
temperature	user_grp_coarse_grain	status
sold_days	物料类别（materia1_level_desc materia3_level2_desc materia3_level1_desc）	list_date
store_square	goods_sex	off_shelf_date
is_mbr	size	
Discount	goods_type	
	is_workday	
	color	
	is_holiday	
	season	
	prom_level	
	shop_grp	
	province	
	city	
	CBD	
	partition	
	waist	

批注 [R13]: 缺少中英文对照

1.2.2 指标分析

（1）连续型指标分析

- 吊牌价销量影响分析

整体上看销量和吊牌价之间的关系如下图左上所示，从该图（图 6 内裤品类产品吊牌价和销量之间的关系）中可以看出，由于吊牌价实际定位并不呈现完美的连续型情况，集中在 50-300 之间，整体上看销量和吊牌价之间存在负相关关系，进一步选择单个门店来展开进一步挖掘。选择其中 shop_id=3289 的数据，分析其与销量的关系，结果如图 6 左上会发现，标价越高，销量越低。然后，借助相关性检验的方法，分析门店吊牌价和销量之间的关系，整体结果显示，吊牌价和销量之间的皮尔逊相关系数为-0.06，与销量之间的相关性较弱，因此作为待定指标。进一步依赖折扣和均摊价来作为价格相关的指标来分析与销量的关系。

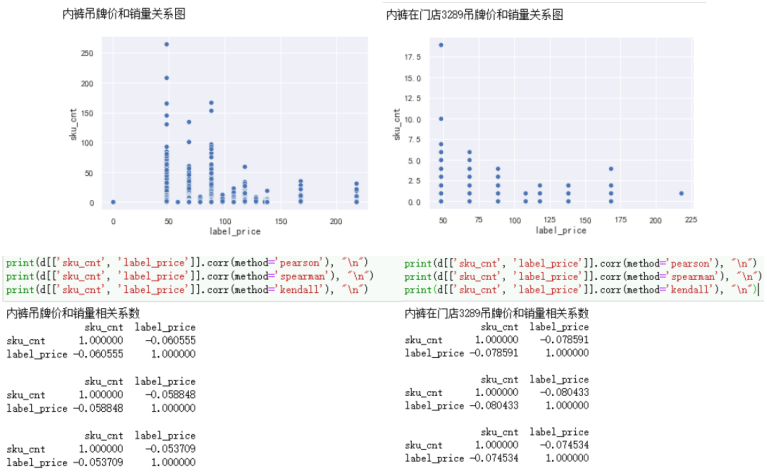


图 6 内裤品类产品吊牌价和销量之间的关系

- 均摊价销量影响分析

从散点图可以看出，在 20-40 之间，销量比较高，进一步根据相关性计算可以看出，销量和均摊价的相关性为-0.01。加之，未来的均摊价无法观测，无法实现批量预测，因此不作为建模的指标。

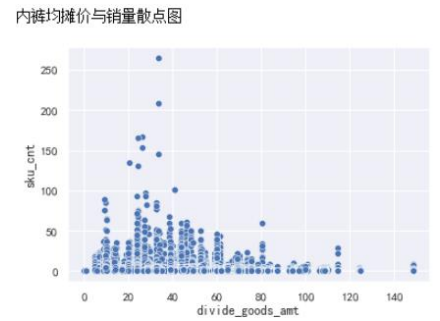
批注 [R14]: 本节前半部分对于连续型指标的分析图结果显示这样处理数据的话其没有具体的相关性，与结论有些不符。

批注 [R15]: 此处吊牌价与销量的相关性是否是针对同种 SKU?
同 SKU 的话吊牌价是否为常量，如果对于分属不同批次的同一 SKU 而言 label price 是变量的话，结合实际的折扣策略，单件产品实际价格计算公式 actual price per sku = sum(label price * discount) / sum(count)

批注 [16R15]: 你再想想怎么算

批注 [R17]: 按字面意思如果是单件实际均价的话此处得出无相关性的结论与后续折扣-销量有相关性的结论就比较矛盾，因为单件实际均价肯定是因为存在折扣（在吊牌价一样时），如果折扣仅仅是按等级划分的话，折扣在价格方面的体现就是单件实际均价，也就是均摊价，这与接下来的折扣销量分析是矛盾的。

批注 [18R17]: 均摊价和折扣的关系再去问一下业务




```
print(dl['sku_cnt', 'divide_goods_amt']].corr(method='pearson'), "\n")
print(dl['sku_cnt', 'divide_goods_amt']].corr(method='spearman'), "\n")
print(dl['sku_cnt', 'divide_goods_amt']].corr(method='kendall'), "\n")
```

内裤均摊价与销量相关系数

	sku_cnt	divide_goods_amt
sku_cnt	1.000000	-0.048212
divide_goods_amt	-0.048212	1.000000

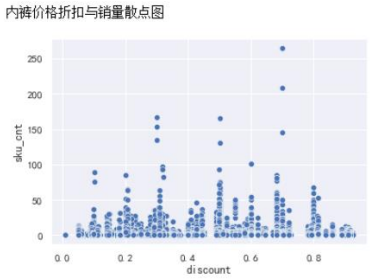
	sku_cnt	divide_goods_amt
sku_cnt	1.000000	-0.011456
divide_goods_amt	-0.011456	1.000000

	sku_cnt	divide_goods_amt
sku_cnt	1.000000	-0.009484
divide_goods_amt	-0.009484	1.000000

图 7 内裤品类均摊价与销量关系

- 折扣销量影响分析

折扣使用计划最低折扣作为特征，其中 2021.11 及之前，折扣并没有对应促销等级，之后折扣开始于促销等级对应，即每一个促销等级下有不同的折扣。从折扣与销量之间的散点图可以看出，折扣集中在 4-6 折之间。进一步计算折扣与销量之间的相关系数，结果显示皮尔孙相关系数为 0.12，因此将折扣作为建模指标之一。折扣反映了吊牌价与均摊价之间的关系，因而选择折扣作为特征后不再考虑吊牌价影响。



批注 [R19]: 如果吊牌价一定，那么商品实际价格的外在表现形式确实主要体现在折扣上。但个人认为消费端的折扣策略可能是一个完善的规则系统，将之简单抽象为“prom_level”、“discount”或者上一步的均摊价格“divide_good_amt”均不足以表征综合的价格因素这一关键特征，这也是我认为此处得出弱相关性的原因之一

批注 [20R19]: 你可以放任何你喜欢的特征，关键是这些特征可以获取，或者你建议业务去获取

```
print(d[['sku_cnt', 'discount']].corr(method='pearson'), "\n")
print(d[['sku_cnt', 'discount']].corr(method='spearman'), "\n")
print(d[['sku_cnt', 'discount']].corr(method='kendall'), "\n")
```

内裤价格折扣与销量相关系数

	sku_cnt	discount
sku_cnt	1.000000	0.022529
discount	0.022529	1.000000

	sku_cnt	discount
sku_cnt	1.000000	0.125554
discount	0.125554	1.000000

	sku_cnt	discount
sku_cnt	1.000000	0.111401
discount	0.111401	1.000000

图 8 内裤品类折扣与销量关系

➢ 温度指标分析

温度偏移指数：

$$T_{diverge} = \frac{|T_{temaperature} - mean|}{mean}$$

其中 mean 指所在城市当前所处季节的平均温度。

计算温度偏移指数与销量相关系数如下图所示，发现存在一定相关性，但并不明显。

批注 [R21]: 看这个结果数据貌似应该得出无相关性的结论

批注 [22R21]: 确实，但你可以测试模型，剔除天气精度的改变情况来反推要不要放这个特征

```
print(d[['sku_cnt', 'temperature_diverge']].corr(method='pearson'), "\n")
print(d[['sku_cnt', 'temperature_diverge']].corr(method='spearman'), "\n")
print(d[['sku_cnt', 'temperature_diverge']].corr(method='kendall'), "\n")
```

内裤温度偏移指数与销量相关系数

	sku_cnt	temperature_diverge
sku_cnt	1.000000	-0.007368
temperature_diverge	-0.007368	1.000000

	sku_cnt	temperature_diverge
sku_cnt	1.000000	-0.053482
temperature_diverge	-0.053482	1.000000

	sku_cnt	temperature_diverge
sku_cnt	1.000000	-0.042687
temperature_diverge	-0.042687	1.000000

进一步选择中国 5 个大区的标志城市分析温度偏移指数和销量之间的关系，去除掉促销等级为一级、二级的数据后作散点图如下图所示，从左到右上到下依次为武汉，深圳，上海，北京和贵阳。从图中可以发现，各个地区的温度变化偏移情况是不同的，整体来看深圳的天气变化较为平稳，也较为集中，北京的天气变化较为明显。但各个城市的偏移指数与销量关系图都可以看出，除北京以外，其他几个城市偏移指数越接近 0（即温度越接近于该城市该季节的平均温度），销量整体越高。说明温度对用户购买行为会产生影响，因此将温度偏移指数作为待建模的指标。

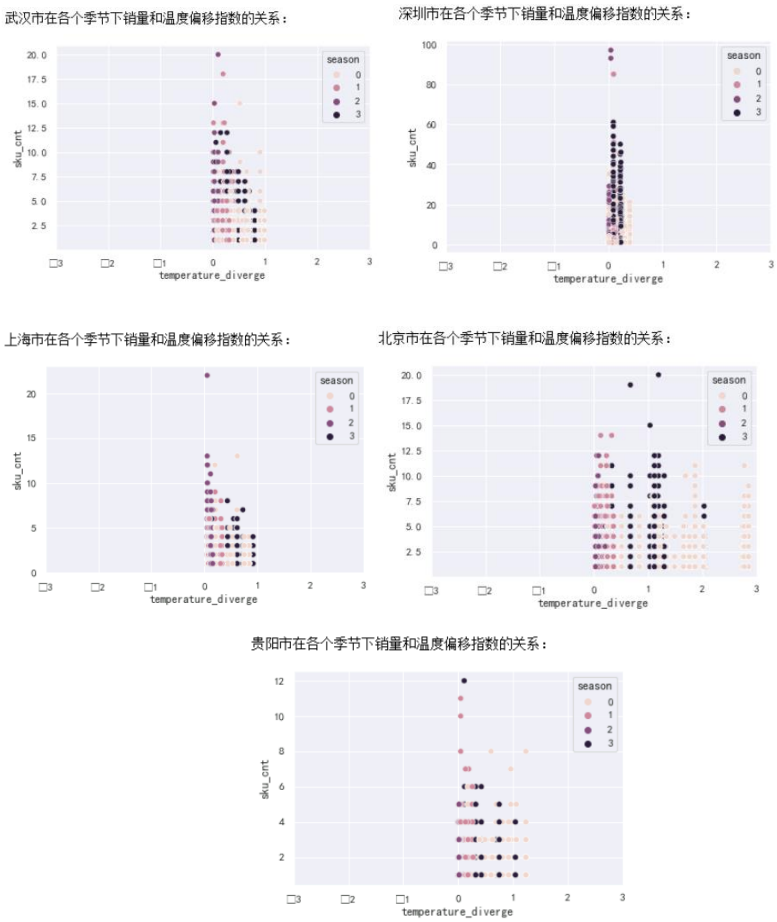
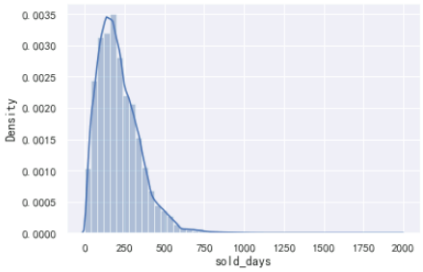


图 9 内裤品类温度偏移指数和销量之间的关系

- 销售天数

销售天数的密度分布情况如下，从图中可以看出，产品销售日期集中在 250 天之间，大于 500 天的产品占比较小。进一步计算二者之间的相关系数，结果显示相关性较弱，因此不作为建模指标。

所有订单的销售日期距离上架日期长度的分布：




```
print(d[['sku_cnt', 'sold_days']].corr(method='pearson'), "\n")
print(d[['sku_cnt', 'sold_days']].corr(method='spearman'), "\n")
print(d[['sku_cnt', 'sold_days']].corr(method='kendall'), "\n")
```

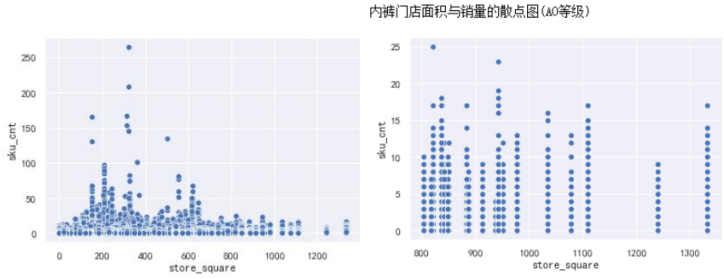
内裤已上市天数与销量相关系数

```
             sku_cnt  sold_days
sku_cnt     1.000000  -0.056291
sold_days  -0.056291   1.000000

             sku_cnt  sold_days
sku_cnt     1.000000  -0.124003
sold_days  -0.124003   1.000000

             sku_cnt  sold_days
sku_cnt     1.000000  -0.099365
sold_days  -0.099365   1.000000
```

图 10 内裤品类销售天数分析图

- 门店面积与销售等级

由散点图可以看出，销售多集中发生在 200-400 平米的门店之间，虽然并不能看出完美的线性关系，通过计算门店面积与销量相关性，整体计算发现相关系数并不高。



内裤门店面积与销量相关系数

	sku_cnt	store_square
sku_cnt	1.000000	0.009055
store_square	0.009055	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.009726
store_square	0.009726	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.007782
store_square	0.007782	1.000000

内裤门店面积与销量相关系数(A0等级)

	sku_cnt	store_square
sku_cnt	1.000000	0.068553
store_square	0.068553	1.000000

	sku_cnt	store_square
sku_cnt	1.00000	0.05074
store_square	0.05074	1.00000

	sku_cnt	store_square
sku_cnt	1.000000	0.041491
store_square	0.041491	1.000000

图 11 各销售等级下店铺面积与销量散点图

下钻到各个促销等级计算相关性，如下所示，发现具有一定相关性，因此将销售等级和门店面积均纳入建模特征。

批注 [R23]: 为什么加入促销等级就有一定相关性

批注 [24R23]: 是在同一促销等级下比较。控制变量法

内裤门店面积与销量相关系数(I级促销)

	sku_cnt	store_square
sku_cnt	1.000000	0.026379
store_square	0.026379	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.028964
store_square	0.028964	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.022728
store_square	0.022728	1.000000

内裤门店面积与销量相关系数(II级促销)

	sku_cnt	store_square
sku_cnt	1.00000	-0.00158
store_square	-0.00158	1.00000

	sku_cnt	store_square
sku_cnt	1.000000	0.005564
store_square	0.005564	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.004449
store_square	0.004449	1.000000

内裤门店面积与销量相关系数(III级促销)

	sku_cnt	store_square
sku_cnt	1.000000	-0.032326
store_square	-0.032326	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	-0.011123
store_square	-0.011123	1.000000

	sku_cnt	store_square
sku_cnt	1.00000	-0.00879
store_square	-0.00879	1.00000

内裤门店面积与销量相关系数(IV级促销)

	sku_cnt	store_square
sku_cnt	1.000000	-0.001872
store_square	-0.001872	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.006832
store_square	0.006832	1.000000

	sku_cnt	store_square
sku_cnt	1.000000	0.005488
store_square	0.005488	1.000000

- 工作日天数

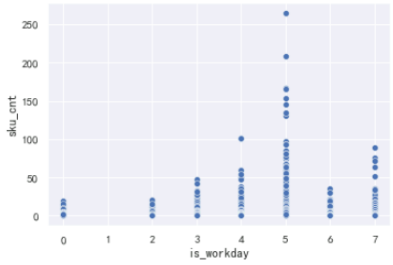
在工作日和节假日与销量的关系图中，工作日天数表示每个销售周包含的

工作天，与节假日呈线性关系。从散点图中可以看出，工作日天数较多，购买基础棉柔巾的销量比较高，这可能是因为人们比较多的选择在工作日出行。通过计算他们与销量之间的相关系数可以看出，与销量的相关性比较高，因此作为建模的指标之一。

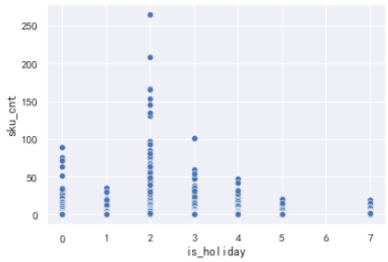
批注 [R25]: 业务判断是否可以合理

批注 [26R25]: 不合理你就不用就好啦，我选什么特征我只考虑我的精度

内裤工作日天数与销量的散点图



内裤节假日天数与销量的散点图




```
print(d[['sku_cnt', 'is_workday']].corr(method='pearson'), '\n')
print(d[['sku_cnt', 'is_workday']].corr(method='spearman'), '\n')
print(d[['sku_cnt', 'is_workday']].corr(method='kendall'), '\n')
```

内裤工作日与销量相关系数

	sku_cnt	is_workday
sku_cnt	1.000000	-0.028882
is_workday	-0.028882	1.000000

	sku_cnt	is_workday
sku_cnt	1.000000	-0.054083
is_workday	-0.054083	1.000000

	sku_cnt	is_workday
sku_cnt	1.000000	-0.050555
is_workday	-0.050555	1.000000

```
print(d[['sku_cnt', 'is_holiday']].corr(method='pearson'), '\n')
print(d[['sku_cnt', 'is_holiday']].corr(method='spearman'), '\n')
print(d[['sku_cnt', 'is_holiday']].corr(method='kendall'), '\n')
```

内裤节假日与销量相关系数

	sku_cnt	is_holiday
sku_cnt	1.000000	0.028882
is_holiday	0.028882	1.000000

	sku_cnt	is_holiday
sku_cnt	1.000000	0.054083
is_holiday	0.054083	1.000000

	sku_cnt	is_holiday
sku_cnt	1.000000	0.050555
is_holiday	0.050555	1.000000

（2）离散型指标分析

批注 [R27]: 对于离散指标的方差检验得出的 P 值 F 值貌似都很好，但统计原理应用前提我觉得是有问题的，一般来讲方差分析无论是单因素还是多因素方差分析都需要固定其他因素来做试验，另外比如样本数据是否服从正态分布、是否独立、是否具备等差性都是其应用前提。

批注 [28R27]: 这是应用，不是写论文，我甚至可以完全不做任何分析，直接拿我想要的特征来做。现实中你觉得有那么多正态分布嘛？

- 门店等级

从门店面积的分布来看，多数门店的面积在 200-600 平米之间，进一步分析门店面积和门店等级之间的关系，可以看出门店的等级在每一个类里面呈明显的降序关系（如图 12 所示），而不同类中，同一等级之间是一样的，因此将

店铺等级划分为 A 类=[AA]，B 类=[A1,B1,C1]，C 类=[A2,B2,C2,OL]，D 类=[A3,B3,C3] ，E 类=[A4,B4,C4,C5]。然后进一步分析在各门店等级下，门店面积和销量之间的关系。为了避免门店面积量级过大带来的影响，需要先对门店面积指标进行归一化处理，处理方式如下：

$$square_{new} = \frac{square_{old} - \mu}{\sigma}$$

其中，$square_{old}$是数据值，μ是门店面积的平均值，σ是门店面积的标准差。

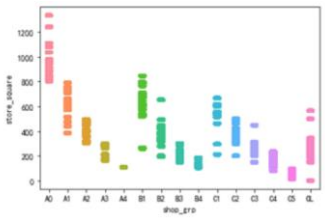


图 12 门店面积与等级关系分析图

对门店等级重新分类后，门店面积与门店等级对应关系如下图所示。

批注 [R29]: 得出了什么结论

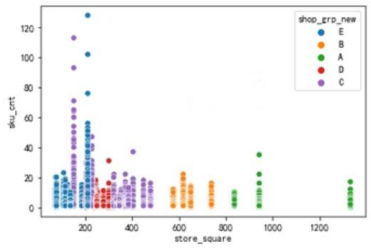
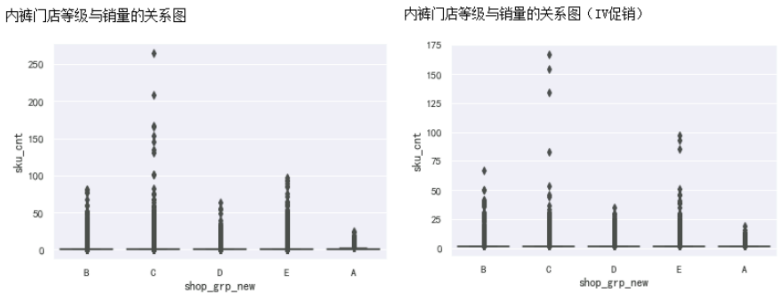


图 13 重新划分后的门店等级与面积对应关系

接着分析门店等级与销量间的关系，从下面的箱型图中可以看出，不同等级的门，销量情况不同，通过使用方差检验，结果显示，P 值很小，F 值很大，因此将其作为建模的特征之一。



内裤门店等级与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
shop_grp_new	4.0	8.983979e+02	224.599475	351.259538	5.908135e-303
Residual	8782567.0	5.615677e+06	0.639412	NaN	NaN

内裤门店等级与销量的方差检验（IV促销）

	df	sum_sq	mean_sq	F	PR(>F)
shop_grp_new	4.0	253.319112	63.329778	77.409259	9.302109e-66
Residual	803180.0	657094.662267	0.818116	NaN	NaN

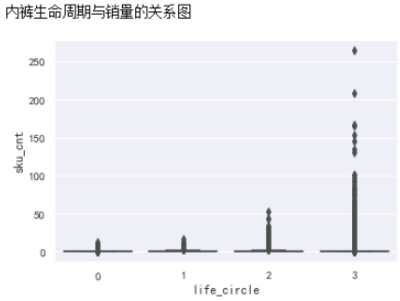
图 14 内裤品类门店等级与销量关系图

- 生命周期分析

  根据企业现有划分逻辑：

180天	弱季节性	0-14天	15-28天	29-135天	剩余天数

分别赋值 0，1，2，3，做散点图分析生命周期与销量间关系。方差检验结果表明各生命周期间均存在显著差异，因此将生命周期纳入建模特征。



内裤生命周期与销量的方差检验

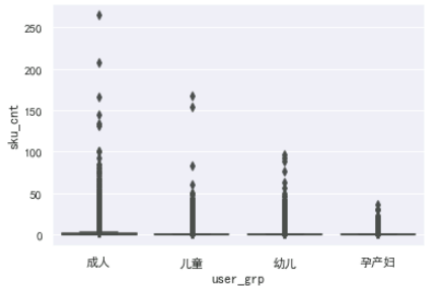
	df	sum_sq	mean_sq	F	PR(>F)
life_circle	1.0	4.199165e+04	41991.653013	66156.448867	0.0
Residual	8782570.0	5.574583e+06	0.634733	NaN	NaN

图 15 内裤品类生命周期与销量箱型图

➢ 用户群

从图中可以看出，不同的用户群销量情况不同，进一步进行方差检验，ANOVA 分析结果中，F 值较大，p 值接近于 0，说明适用人群对销量的影响非常显著。

内裤用户群与销量的关系图



内裤用户群与销量的方差检验

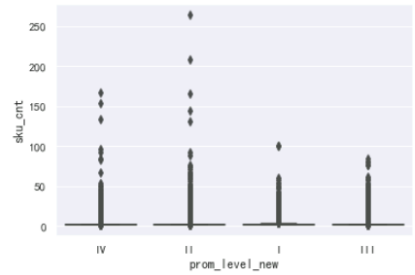
	df	sum_sq	mean_sq	F	PR(>F)
user_grp	3.0	2.101225e+04	7004.082639	10993.323782	0.0
Residual	8782568.0	5.595563e+06	0.637121	NaN	NaN

图 16 内裤品类适用人群分析

➢ 促销等级分析

通过促销等级与销量箱型图，可以看出，发现各销售等级销量间存在一定区别。对促销等级进行方差检验。方差检验结果表明各促销等级间均存在显著差异（reject 这一列为 True 的话则说明两个处理间存在差异），因此将促销等级纳入建模特征。

内裤促销等级与销量的关系图



内裤促销等级与销量的方差检验

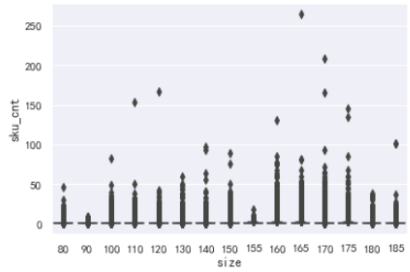
	df	sum_sq	mean_sq	F	PR(>F)
prom_level_new	3.0	1.965194e+04	6550.648287	5541.046172	0.0
Residual	2169607.0	2.564919e+06	1.182204	NaN	NaN

图 17 促销等级与销量箱型图

➢ 尺码

由下图可以看出，160-175 尺码销量较高，为热门尺码。进行 ANOVA 分析结果显示，ANOVA 分析结果中，F 值较大，p 值接近于 0，说明尺码对销量的影响非常显著。

内裤尺码与销量的关系图



内裤尺码与销量的方差检验

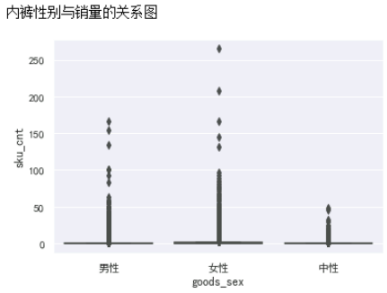
	df	sum_sq	mean_sq	F	PR(>F)
size	1.0	5.677652e+03	5677.652268	8887.059629	0.0
Residual	8782570.0	5.610897e+06	0.638867	NaN	NaN

图 18 内裤品类与尺码关系分析结果

➢ 性别

由下图可以看出，女性产品销量整体较高，中性产品销量较低。进行 ANOVA 分析，ANOVA 分析结果中，F 值较大，p 值接近于 0，说明性别对销

量的影响非常显著，因此作为建模的指标之一。



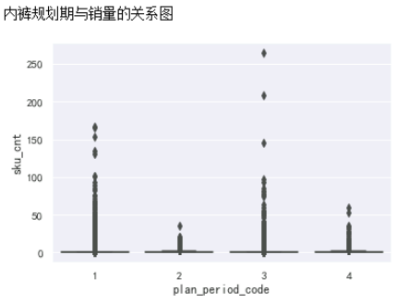
内裤性别与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
goods_sex	2.0	6.397335e+03	3198.667473	5007.42043	0.0
Residual	8782569.0	5.610178e+06	0.638785	NaN	NaN

图 19 内裤品类性别与销量箱型图

- 规划期编码

规划期编码（上市时间季度），根据箱型图可以看出，编码为 2 和 4 的销量较少，编码为 3 的销量最多。进行 ANOVA 分析，ANOVA 分析结果中，F 值较大，p 值接近于 0，说明规划期编码对销量的影响非常显著，因此作为建模的指标之一。



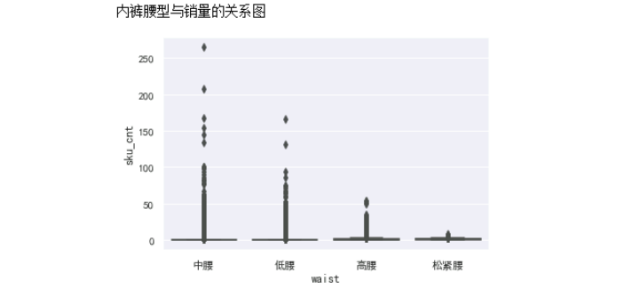
内裤规划期与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
plan_period_code	1.0	1.352937e+03	1352.937275	2116.081321	0.0
Residual	8782570.0	5.615222e+06	0.639360	NaN	NaN

图 20 内裤品类规划期编码与销量关系图

➢ 腰型

绘制箱型图可以看出各类别间有较为明显差别。对腰型进行方差检验，结果如下。通过方差检验，即各类别有显著差异，将腰型纳入建模特征。



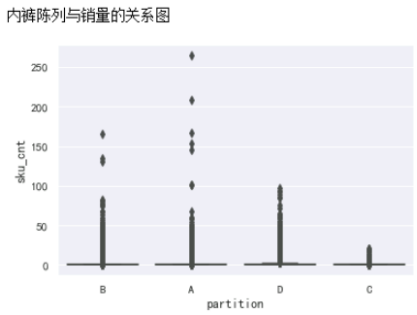
内裤腰型与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
waist	3.0	7.038574e+03	2346.191258	3673.313273	0.0
Residual	8782568.0	5.609536e+06	0.638713	NaN	NaN

图 21 内裤品类腰型与销量关系图

➢ 陈列信息

绘制箱型图可以看出各类别间有较为明显差别。对陈列分区进行方差检验，结果如下，过方差检验，即各类别有显著差异，将陈列分区纳入建模特征。



内裤陈列与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
partition	3.0	2.959153e+03	986.384490	1543.210158	0.0
Residual	8782568.0	5.613616e+06	0.639177	NaN	NaN

图 22 内裤品类陈列分区与销量箱型图

- 颜色和图案（color_and_pattern/color）

批注 [R30]: 站在消费者角度讲，个人认为颜色和图案指标（特别是颜色指标）是服装类 SKU 很关键的指标之一

批注 [31R30]: 未来颜色拓展怎么办？你觉得颜色永远只有红黄蓝？未来新出现一种颜色，模型并没有这个颜色的参数，怎么办？你拓展你想怎么做都行，我这里我就这搞了

```
data['color_and_pattern'].unique()
```

```
array(['浪花海豚+海底小鱼', '欢乐小鱼+水上小鸭', '海底珊瑚+海草水母', '海底游鱼+手绘条纹', '黑白条纹',
       '彩棉棕+彩棉条纹', '浅蓝波纹', '净粉色', '粉花灰', '净黄色', '全棉字母', '方块格纹', '深蓝花灰',
       '灰色波点', '红蓝几何', '浅蓝花灰', '烟粉色', '白底绣球', '白花灰', '粉底卷叶', '紫底蒲公英',
       '绿底满叶', '肤色', '蓝底绣球', '香芋紫', '黑色', '墨紫色', '靛蓝色', '卡其大格', '红蓝格子',
       '花灰色', '花灰蓝格', '藏青底千鸟格', '藏青底白格', '浅橙花灰', '浅花灰', '粉底满叶', '绿底卷叶',
       '芽黄花灰', '蓝条花灰', '薄荷绿', '藕粉色', '黄底小白花', '黄底黄碎花', '灰色', '手绘气球+气球云朵',
       '夏日沙滩+夏日冰淇淋', '手绘西瓜+爱心西瓜', '气球小象+游乐气球', '爱心礼物+爱心乐符', '粉底天鹅+花丛天
       '呦呦里里+星空奇遇', '小长颈鹿+丛林狮子', '白云棉呦呦+云朵鸽子', '白底柠檬+字母柠檬', '全棉小猫+几何波
       '白色+紫波点', '浅蓝+浅黄碎花', '绿底爱心+手绘爱心', '三角几何+撞色横条', '倒影帆船+海浪纸船',
       '绿底线条+花灰波点条', '浅绿碎花', '爱心条纹', '浅水绿棉朵+藕粉棉朵', '浅灰色+粉花灰', '浅蓝花灰+枝繁叶茂
       '白花灰+粉花灰', '粉色+浅蓝花灰', '肤色+浅花灰', '肤色+粉色', '肤色+粉色枝蔓', '蓝色棉朵+紫色棉枝',
       '粉色波点', '肤色波点', '墨绿色', '深蓝条花灰', '深蓝色', '红蓝条纹', '蓝竹节花灰', '黑底几何',
       '幸福红', '白底双色花', '鳞叶满底', '绯红色', '星月蓝', '黄绿小叶+白灰色', '星点花卉+蓝底花影',
       '蓝色棉朵+粉花灰', '白花灰+烂漫粉枝叶', '粉肤色+浅蓝花灰', '白灰色+粉花灰', '蓝色条纹', '墨蓝条花灰',
       '散点字母', '蓝底棕叶', '规则松树', '粉底郁金香', '白底郁金香', '线条红果', '白底满叶', '蓝底雏菊',
       '岩石灰', '柔光蓝', '幸运红', '紫花灰', '桃粉色', '钴蓝色', '红条花灰', '白麻条花灰', '深花灰',
```

```
data['color'].unique()
```

```
array(['多件组合色', '一件混合色', '浅蓝', '粉红', '浅黄', '深蓝', '中灰', '浅灰', '浅紫', '浅绿',
       '杏/肤', '黑色', '深紫', '浅橙', '深绿', '中红', '中蓝', '深红', '深灰', '白色', '中绿',
       '中紫', '浅黄组合多', '棕/咖', '原棉色', '中橙', '浅蓝色', '深蓝色', '暗灰色', '淡蓝绿',
       '中蓝色', '素紫红', '素绿蓝', '艳蓝色', '浅绿蓝', '浅橙色', '浅紫红', '浅红色', '浅黄绿',
       '浅紫色', 'unk', '浅黄色', '淡紫色', '淡蓝色', '浅蓝绿', '暗蓝色', '淡灰色', '深艳蓝',
       '素绿色', '深艳红', '淡绿蓝', '艳红色', '素灰色', '艳黄绿', '中黄', '亮黄色', '尽黑',
       '无荧光尽白', '亮橙色', '艳黄色', '中红色', '淡红色', '中紫红', '淡黄绿'], dtype=object)
```

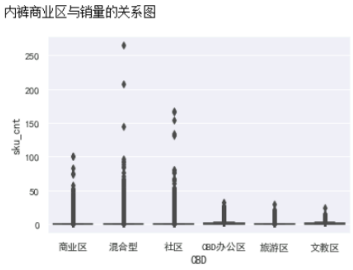
由于该特征下的分类过多，以及后面还会继续拓展新的款色图案，因此不作为特征。

批注 [R32]: 陈列信息没用上

批注 [33R32]: 往上看

- CBD

绘制箱型图可以看出各类别间有较为明显差别。对 CBD 进行方差检验，结果如下，过方差检验，即各类别有显著差异，将陈列分区纳入建模特征。



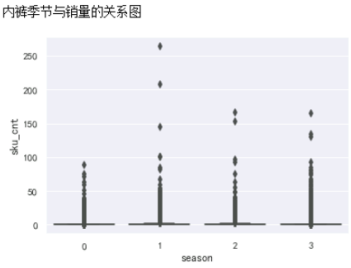
内裤商业区与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
CBD	5.0	4.856266e+03	971.253171	1520.050384	0.0
Residual	8782566.0	5.611719e+06	0.638961	NaN	NaN

图 23 内裤商业区与销量关系分析

➢ 季节

通过分析季节与销量之间的关系可以看到不同季节的销量情况不一样，进一步使用方差检验，结果显示季节存在显著的影响，因此将其作为建模的特征之一。



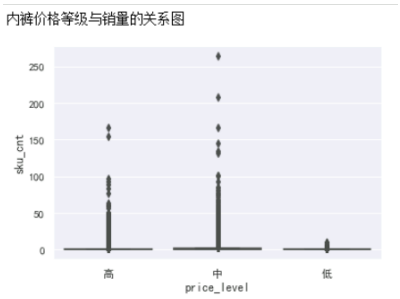
内裤季节与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
season	1.0	2.825269e+03	2825.268731	4420.061812	0.0
Residual	8782570.0	5.613750e+06	0.639192	NaN	NaN

图 24 内裤季节与销量关系分析

➢ 价格等级

通过分析价格等级与销量之间的关系可以看到不同季节的销量情况不一样，进一步使用方差检验，结果显示价格等级存在显著的影响，因此将其作为建模的特征之一。



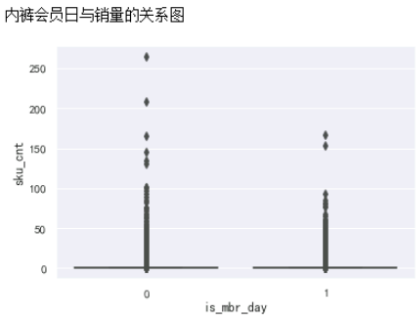
内裤价格等级与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
price_level	2.0	1.960355e+04	9801.774990	15380.59777	0.0
Residual	8782569.0	5.596971e+06	0.637282	NaN	NaN

图 25 内裤价格等级和销量之间的关系图

➢ 会员日

由下图可以看出，虽然会员日和非会员日下的销量关系并不如期望的那般有明显的的落差，但因为非会员日的占比更高，因此使用方差检验进一步分析二者之间的关系，结果显示，F 值很大，p 值很小，因此将其作为建模的特征之一。



内裤会员日与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
is_mbr_day	1.0	6.391036e+02	639.103560	999.472207	2.406369e-219
Residual	8782570.0	5.615936e+06	0.639441	NaN	NaN

图 26 内裤会员日与销量之间的关系图

1.2.3 建模特征选择及编码

根据上述特征工程分析结果，内裤品类选择的建模特征如下：

表 6 内裤品类所选特征字段说明表

序号	字段	字段说明
1	discount	折扣
2	Plan_period_code	规划期编码
3	Goods_sex	性别
4	size	尺码
5	Price_level	价格等级
6	User_grp	适用人群
7	Waist	腰型

8	Is_workday	是否工作日
9	Is_holiday	是否假期
10	season	季节
11	Store_square	店铺面积
12	CBD	商圈
13	Temperature_diverge	温度偏移指数（处理方式如上文）
14	Life_circle_new	生命周期（按企业现有规则划分）
15	Shop_grp_new	店铺等级（处理方式如上文）
16	Prom_level_new	促销等级（处理方式如上文）
17	Partition	陈列分区信息
18	Is_mbr_day	是否会员日

1.3 模型设计

1.3.1 模型设计

- 模型思路

通过将各个 SKU 按照门店划分，统计其历史数据量，如下图所示，从图中可以发现各 SKU 的历史数据量较少，按单个 SKU 预测会导致历史数据量较少，模型的精度效果降低，因此，采用分层聚类的方法。

批注 [R34]: 哪里采用了这个方法

批注 [35R34]: 按城市，省份，可以说是分类，不是聚类

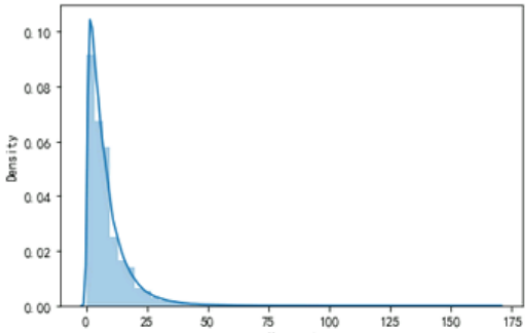


图 27 内裤品类 SKU 在各门店下历史数据量

在时间点t（天/周/半月等）下，隶属于类型$c(c \in C)$的门店$s(s \in S)$下产品（SKU）$i(i \in I^{cs})$的特征向量为$X_{i,t}^{cs}$，包括物料描述、温度、季节、折扣、促销等。其中I^{cs}表示类型c的门店s下包含的所有内裤类产品（SKU）。$d_{i,t}^{cs}$则为该产品在时间点t下的真实需求量。

当$t \le 10$表示该产品为全新品，没有历史数据，需要借助已有产品的历史数据作为协变量来预测：

$$\hat{d}_{i,t}^{cs} = \sum_{j \in I^c, j \neq i} w_{ij} \sum_{t,x,s} X_{j,t,x}^{cs} B_x(t)$$

批注 [R36]: 这是什么模型，这里面没写出具体模型

批注 [37R36]:

其中，$\hat{d}_{i,t}^{cs}$为该产品在时间点t下的预测需求量；$X_{j,t,x}^{cs}$表示产品$j \in I^c$于时间t时在门店s下的第x个特征；$B_x(t)$表示对特征x在时间t对销售的影响进行建模的系数函数；w_{ij}表示产品j对产品i的影响权重；

当$t > 10$表明该产品已经累积了一定的历史数据，可以结合自身时序需求数据和部分协助数据来建模：

$$\hat{d}_{i,t}^{cs} = \alpha * F\left(X_{i,t}^{cs}, X_{i,t-1}^{cs}, \dots, X_{i,0}^{cs}\right) + (1-\alpha) * \sum_{j \in I^{cs}, j \neq i} w_{ij} \sum_{t,x} X_{j,t,x}^{cs} B_x(t)$$

批注 [R38]: 这里也不清晰

其中，$F(\cdot)$表示对产品i自身历史数据建模的函数；$\alpha \in (0,1]$表示自身历史数据对未来需求贡献的权重。

- 模型的预测效果评价指标

平均绝对百分比误差计算公式：

$$\text{MAPE} = \frac{1}{\text{I} * \text{T}} \sum_{\text{i}=1}^{I} \sum_{\text{t}=1}^{\text{T}} \frac{|\hat{d}_{i,t}^{cs} - d_{i,t}^{cs}|}{d_{i,t}^{cs}}$$

其中， MAPE 值越小说明误差越小；

需求预测精度：

$$ACU = 1 - MAPE$$

按全棉 14 天滚动预测方法计算出来的预测误差为$MAPE_P$，港中深团队预测方法的误差为$MAPE_G$，则本项目误差改善百分比为：

$$ACP = \frac{MAPE_P - MAPE_G}{MAPE_P} * 100\%$$

针对补货需求预测的精度为，ACP 的值。

批注 [R39]: 业务的验证方法有变更。

1.3.2 模型搭建

基于特征工程的结果，将选中的特征放入模型，并使用不同的建模维度尝试模型的预测效果，选择最好的结果。

批注 [R40]: 这个写特征工程，用到的方差方法和归一法等等需要把方法写出来。

（1）分类建模的维度——c 的选择

批注 [R41]: 运用了什么参数去选的分类维度

表 7 内裤不同建模维度方式结果

分类维度	算法	MAPE 值	ACP
City	DT	0.30	0.56
Province	DT	0.28	0.58
Shop_grp_new	DT	0.24	0.61

批注 [R42]: 模型选择，搭建思路和方案

根据上述的结果，将历史数据**门店按照店铺等级**来划分，并按各门店等级的线下门店数据来展开需求预测，基本思路如下图所示。

批注 [R43]: 只是在 DT 的算法里是最好的，那其他的算法怎么证明。

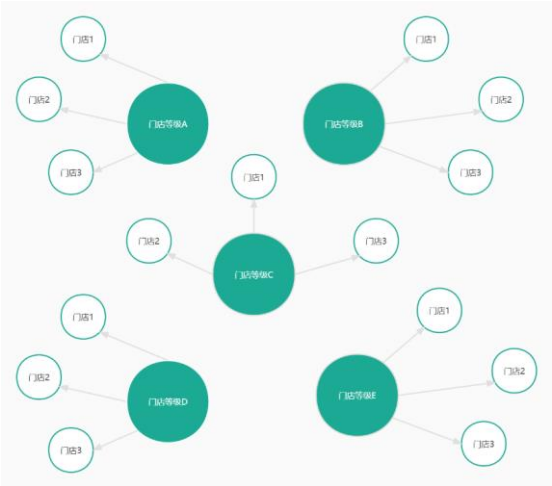


图 28 内裤品类预测模型设计

（2）建模算法选择

使用不同的算法尝试模型的预测效果，选择最好的结果对应的算法作为目标算法。

表 8 内裤不同建模维度方式结果

分类维度	算法	MAPE 值	ACP
Shop_grp_new	DT	0.24	0.61
	XGB	0.28	0.58
	RF	0.64	0.33
	GBR	0.81	0.16

根据上述的结果，将使用**决策树模型**来建模。

批注 [R44]: 需要结合分类维度一起判断

（3）历史数据筛选

使用不同的历史数据范围尝试模型的预测效果，选择最好的结果对应的算法作为目标算法。

表 9 内裤不同建模维度方式结果

分类维度	算法	历史数据量	MAPE 值	ACP
Shop_grp_new	DT	2021.01 至今	0.30	0.55
		2021.06 至今	0.13	0.68

批注 [R45]: 判断数据是否疫情封控数据影响

		2022.01 至今	0.13	0.99
		2022.06 至今	0.21	0.99

根据上述的结果，将使用 2022 之后的历史数据来建模。

批注 [R46]: 跟目前代码运行数据时间不相符

（4）异常值剔除

使用不同的历史数据范围尝试模型的预测效果，选择最好的结果对应的算法作为目标算法。

表 10 内裤不同建模维度方式结果

分类维度	算法	周销量最大阈值	MAPE 值	ACP
Shop_grp_new	DT	500	0.30	0.51
		300	0.30	0.55
		100	0.31	0.54
		50	0.30	0.54

根据上述的结果，将使用**周销量小于 500** 的部分来建模。

根据上述步骤，最终确认的方法为——使用 2022.01 至今的历史订单数据，然后将历史周销量大于等于 500 的部分剔除，然后使用决策树算法训练模型；

2. 婴儿连体服品类

2.1 数据预处理

2.1.1 数据合并

按照基础棉柔巾品类相同的操作进行内部数据的拼接筛选，同时合并外部数据。涉及婴儿连体服品类从 2020 年至 2022 年 8 月 30 日的订单，删除步骤为：1）删除 order_type_new 里包含展销（**pos_exhibition_sale**）和团购（**pos_group_purchase**）以及不需要的事业部（division_name=里物、大客户、总部、稳健、津梁、津梁官网），甚于 2134010 条数据；2）删除 **skc_code** 以 **LUN、LDS** 开头的订单，剩余 2134010 条订单，以及删除 **shop_id** 以 **4 和 6** 为开头的订单之后剩的订单，剩余 2131895 条订单；3）进一步剔除销量 **sku_cnt**

小于 0 的订单，剩余 2056855 条订单。然后将列值唯一的列和缺失值在 80%以上的列删除。经过上诉步骤，最终获得 2056855 条订单数据，数据字段有 57 个。

2.1.2 数据预处理

（1）数据情况

首先，剔除 shop_id 以 4 和 6 为开头的订单，删除 skc_code 以 LUN、LDS 开头的 skc 的订单。然后，按照 devision_code=10000，is_exhbit=N 两个个标准筛选出线下门店（非展销）的所有与婴儿连体服的订单数据，最终获得全国 74 个城市的 365 个门店的总计 993699 条历史订单数据，其中涉及 722 个 SKC 和 2280 个 SKU，将筛选出的数据进一步剔除值唯一的列：

表 11 婴儿连体服删除数据字段

字段	描述	字段	描述
is_exhibit	是否展销门店	cat_1_code	一级大类编码
goods_type	例如:成人服饰	cat_1_name	一级大类名称
season_attr	季节属性	cat_2_code	二级大类编码
is_common_goods	是否常规商品	cat_2_name	二级大类名称
is_contain_gift	是否包含赠品	cat_3_code	三级大类编码
division_code	事业部编码	cat_3_name	三级大类名称
division_name	事业部名称		

空缺值在 80%以上的字段：

表 12 婴儿连体服空缺率大于 80%的字段

保留的字段为：

字段	描述	字段	描述
promotion_codes	促销代码	receiver_province_code	收货人省份编码
is_lack_stock	是否缺货	receiver_city_code	收货人城市编码
receiver_city_name	收货人城市	receiver_area_code	收货人所属县
receiver_area_name	收货人所属县	receiver_province_name	收货人省份

表 13 婴儿连体服保留数据字段

字段	描述	字段	描述
sku_code	物料编码	skc_code	款色编码
sku_name	物料名称	skc_name	款色名称
list_year	上市年份	spu_code	款编码
list_date	上市日期	price_level	价格档位
label_price	零售价/吊牌价	coupon_code	优惠券编码
off_shelf_date	下架日期	coupon_name	优惠券名称
sold_days	订单创建日期与上架日期的差	material_level_1_desc	物料 1 级描述,针织
life_circle	下架日期与上架日期的差	material_level_2_desc	物料 2 级描述,针织单面
error_off	下架日期与订单创建日期的差	material_level_3_desc	物料 2 级描述,针织平纹
is_workday	订单创建日期是否为工作日	sku_cnt	有效商品数量，取消商品的时候会变
is_holiday	订单创建日期是否为节假日	divide_goods_amt	均摊价包含积分抵扣,不含优惠,不含运费
season	订单创建日期所在的季节	divide_freight_amt	均摊运费
prom_level	各渠道下订单创建日期下包含的促销等级	id	订单明细 ID,订单中心维护
store_square	门店面积	ord_id	订单 ID，主表的 ID 字段
CBD	门店商圈	weight	重量
temperature	该城市在订单创建日期下的平均温度	is_gift	是否赠品

discount	价格折扣，到手价除以吊牌价	freight_amt	运费
ord_create_time	下单时间	actual_pay_amt	实付金额
delivery_timelast	最后出库时间	ord_code	订单编号
shop_name	店铺名称	ord_channel	订单渠道
ord_type	订单类型 #ENUM#{\"omni\":\"全渠道订单\",\"pos\":\"POS 订单\",\"reissue\":\"补发订单\"},补发的可能要过滤	shop_id	店铺编码对应主数据门店编码
ord_type_new	剔除展销与团购的字段	plan_period_code	规划期编码,格式 1，2，3，4
goods_sex	商品性别	Sku_total_cnt	订单商品总数量
color_and_pattern	花色	is_paid_mbr	是否付费会员
size	尺码:170,175	buyer_mbr_id	买家会员 ID
color	白色，红色，多件组合色	receiver_country_code	收货人国家
shop_open_date	店铺开业日期	receiver_country_name	收货人国家
city	线下门店城市	shop_grp	店铺等级,即店群
status	店铺状态	province	线下门店省
partition	该品类在该门店的陈列区域（A/B/C）		

（2）按“SKU-门店-周”groupby-resample

- 将各个产品按照“SKU-门店”的维度进行分类汇总，然后再按照

ord_create_time 以每周一（'W-MON'）的维度进行下采样，以获得各个 sku 在每一个门店，每一周的历史销量数据。其中汇总时候，各个字段的汇总方式如图 4 所示，汇总后最终的订单数据为 1440199 条。

(3)数据补全

表 14 婴儿连体服字段缺失率

数据中，存在缺失的列以及对应的缺失率为：

字段	缺失率	补全方案	字段	缺失率	补全方案
material_level_1_desc	0.003	不补全	shop_name	0.065	不补全
material_level_2_desc	0.003	不补全	ord_type_new	0.001	不补全
material_level_3_desc	0.003	不补全	prom_level	0.451	按没有任何促销活动补全，即'IV'
buyer_mbr_id	0.005	不补全	temperature	0.001	使用门店所在城市对应的年平均温度
is_paid_mbr	0.001	不补全	delivery_time_last	0.001	不补全
coupon_code	0.689	不补全	receiver_country_code	0.001	不补全
coupon_name	0.689	不补全			

- price_level 的补全

经与业务确认补全为高价格档产品。补全后价格等级与吊牌价对应关系如下图所示：

```
if len(df[df['price_level'] == 'unk']) > 0:
    '''价格水平补全'''
    low = df[df['price_level'] == '低']['label_price'].max()
    upper = df[df['price_level'] == '高']['label_price'].min()
    df.loc[(df['price_level'] == 'unk') & (df['label_price'] <= low), 'price_level'] = '低'
    df.loc[(df['price_level'] == 'unk') & (df['label_price'] > low) & (
            df['label_price'] <= upper), 'price_level'] = '中'
    df.loc[(df['price_level'] == 'unk') & (df['label_price'] > upper), 'price_level'] = '高'
```

- prom_level_new 的补全

缺失值表示在订单创建日期及对应的渠道下不包含任何的促销活动，因此用”IV”标签替代空值，即四级促销活动：

```
df['prom_level_new'] = df['prom_level_new'].fillna('IV')
'''促销等级补全'''
```

- avg_temperature 的补全

使用门店所在城市对应的年平均温度来补全：

```
if sum(df['avg_temperature'].isnull()) > 0:
    '''温度补全'''
    for c in list(df.city.unique()):
        df.loc[(df['avg_temperature'].isnull().values == True) & (df['city'] == c), 'avg_temperature'] = \
        df[df.city == c]['avg_temperature'].mean()
```

- partition 的补全

使用 D 区陈列来补全

```
df['exhibited_area'] = df['exhibited_area'].fillna('D')
'''#陈列补全'''
```

2.2 特征工程

2.2.1 指标分类

根据数据字段的组成，将相关的指标进行分类，便于后期的分析和处理，划分结果如表 15 所示。

表 15 婴儿连体服指标及其分类

连续型	描述	离散型	描述	其他	描述
label_price	吊牌价	price_level	价格水平	ore_create_time	订单所处周

devide_good_amt	均摊价	materia3_level_desc	物料描述	shop_open_date	开业日期
temperature	温度	is_workday	周工作日天数	status	门店状态
sold_days	已上市天数	is_holiday	周节假日天数	list_date	上架日期
life_circle	生命周期	season	季节	off_shelf_date	下架日期
store_square	门店面积	prom_level_new	促销等级	ord_channel	订单渠道
Sku_cnt	周销量	shop_grp_new	店群	ord_type	订单类型
CBD	门 店商圈	province	省	ord_type_new	订单类型
partition	陈列	city	市	Is_mbr_day	是否包含会员日

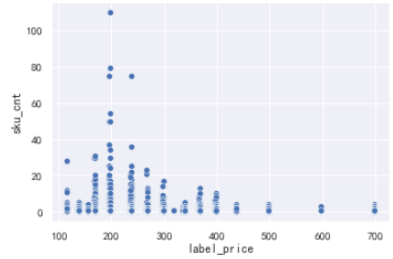
2.2.2 指标分析

（1）连续型指标分析

- 产品价格分析

整体上看销量和吊牌价之间的关系下图所示，从该图中可以看出，价格水平为中的产品价格比低价格水平的产品要高，高价格水平的产品，低吊牌价的销量好一些。然后，借助 Pearson 相关系数进行相关性检验，分析产品吊牌价和销量之间的关系，发现**吊牌价与销量之间的相关系数为 0.02，相关性较弱，甚至没有，因此将其放弃纳入模型。**

婴儿连体服吊牌价和销量关系图




```
print(dl['sku_cnt', 'label_price']].corr(method='pearson'), '\n')
print(d[['sku_cnt', 'label_price']].corr(method='spearman'), '\n')
print(d[['sku_cnt', 'label_price']].corr(method='kendall'), '\n')
```

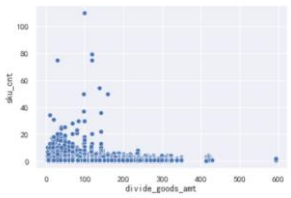
婴儿连体服吊牌价和销量相关系数

```
             sku_cnt  label_price
sku_cnt      1.000000     0.027175
label_price  0.027175     1.000000

             sku_cnt  label_price
sku_cnt      1.000000     0.033865
label_price  0.033865     1.000000

             sku_cnt  label_price
sku_cnt      1.000000     0.028888
label_price  0.028888     1.000000
```

图 29 婴儿连体服吊牌价分析

到手价与销量之间的关系分析，从下图左上所示，可以看出对于高价格水平的产品，如果成交价较吊牌价相应的降低，可以带来销量的增加。进一步计算各个 sku 的销量和到手价之间的相关系数，结果如图下所示，从图中可以看出，相关系数的为-0.04，即到手价越低，销量越高，但是未来是无法准确地观测到到手价指标的，将其强行纳入模型会导致模型无法实现预测，**因此放弃将到手价作为建模的指标之一。**

婴儿连体服均摊价与销量散点图




```
print(d[['sku_cnt', 'divide_goods_amt']].corr(method='pearson'), '\n')
print(d[['sku_cnt', 'divide_goods_amt']].corr(method='spearman'), '\n')
print(d[['sku_cnt', 'divide_goods_amt']].corr(method='kendall'), '\n')
```

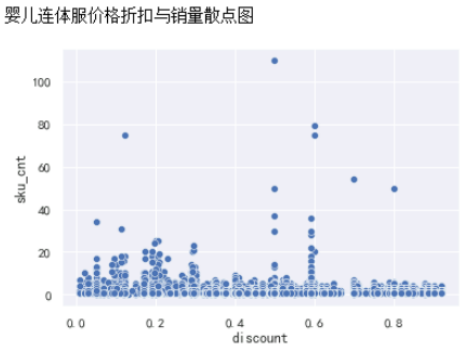
内裤均摊价与销量相关系数

```
                sku_cnt  divide_goods_amt
sku_cnt        1.000000         -0.048212
divide_goods_amt -0.048212       1.000000

                sku_cnt  divide_goods_amt
sku_cnt        1.000000         -0.011456
divide_goods_amt -0.011456       1.000000

                sku_cnt  divide_goods_amt
sku_cnt        1.000000         -0.009484
divide_goods_amt -0.009484       1.000000
```

图 30 基础棉柔巾到手价与销量之间的关系

在价格折扣对销量的影响分析中，价格折扣比较分散，并没有明显的线性关系，进一步借助相关性检验的方法，计算各个 sku 的价格折扣和销量的相关系数，发现杰哥折扣和销量之间相关系数为 0.09，**因此选择折扣作为建模指标**

之一。




```
print(d[['sku_cnt', 'discount']].corr(method='pearson'), "\n")
print(d[['sku_cnt', 'discount']].corr(method='spearman'), "\n")
print(d[['sku_cnt', 'discount']].corr(method='kendall'), "\n")
```

婴儿连体服价格折扣与销量相关系数

	sku_cnt	discount
sku_cnt	1.000000	0.091815
discount	0.091815	1.000000

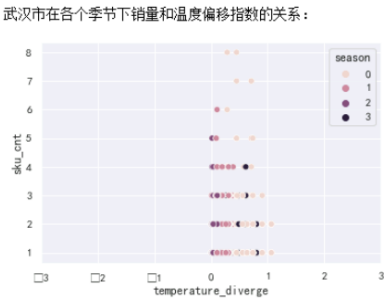
	sku_cnt	discount
sku_cnt	1.000000	0.113985
discount	0.113985	1.000000

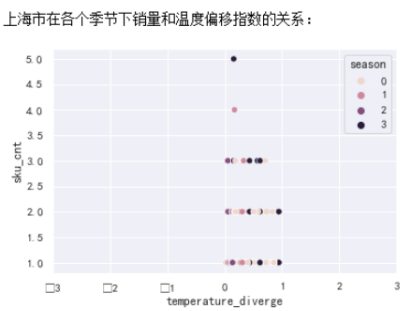
	sku_cnt	discount
sku_cnt	1.000000	0.099158
discount	0.099158	1.000000

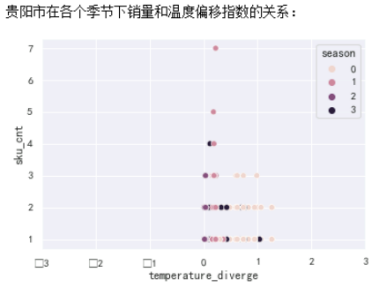
图 31 婴儿连体服价格折扣分析

➢ 温度指标分析

选择中国5个大区的标志城市分析温度偏移指数和销量之间的关系，如图所示。从下面五个图中可以发现，对于各个城市来说，该季节下的温度偏移程度越小，销量就越集中，而随着温度偏移程度的增加，购买行为的发生也变小，可见温度对用户线下购买行为的影响。**因此，将温度偏移指数作为建模的指标之一。**







婴儿连体服温度偏移指数与销量相关系数

	sku_cnt	temperature_diverge
sku_cnt	1.000000	0.000989
temperature_diverge	0.000989	1.000000

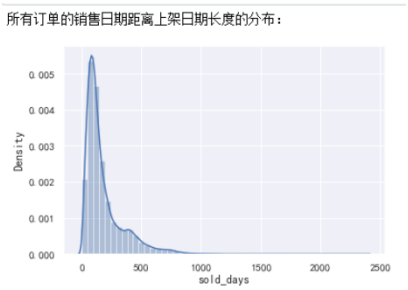
	sku_cnt	temperature_diverge
sku_cnt	1.000000	-0.024541
temperature_diverge	-0.024541	1.000000

	sku_cnt	temperature_diverge
sku_cnt	1.000000	-0.019767
temperature_diverge	-0.019767	1.000000

图 32 婴儿连体服温度偏移指数和销量之间的关系

➢ 已上市天数分析

从已上市天数的原始分布来看，婴儿连体服的销售周期在 0-1000 天之间，然后进一步分析与销量之间关系，从下图可以看出，虽然已上市天数与销量之间的相关关系为-0.07，但是产品的已上市天数呈现随时间递增的趋势，在未来清货的时候产品的销量是会增长的，在通过对历史数据标准化之后，新增的数据的已上市天数已经与历史数据不在同一个分布下面了，因此会导致数据标准化的偏差，不做为特征放入模型。



婴儿连体服已上市天数与销量相关系数

	sku_cnt	sold_days
sku_cnt	1.000000	-0.079401
sold_days	-0.079401	1.000000

	sku_cnt	sold_days
sku_cnt	1.000000	-0.159825
sold_days	-0.159825	1.000000

	sku_cnt	sold_days
sku_cnt	1.000000	-0.129807
sold_days	-0.129807	1.000000

图 33 婴儿连体服生命周期分析

➢ 门店面积分析

在门店面积中，门店面积和销量之间呈一定的一元二次函数的关系，即门店面积在 400-800 平之间，销量情况反而更好，因此进一步选择同一等级下的门店观察其与销量之间的关系，可见呈现并不明显的线性关系，因此将门店面积作为建模的特征之一。

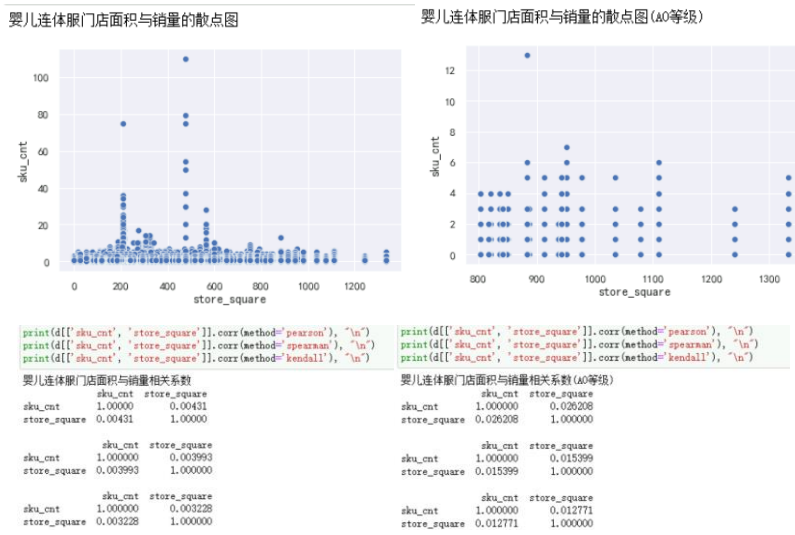
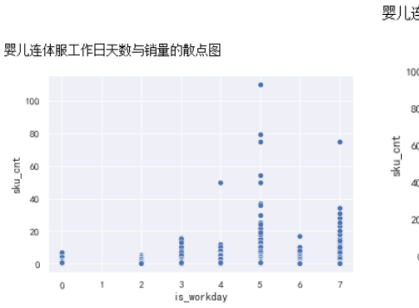
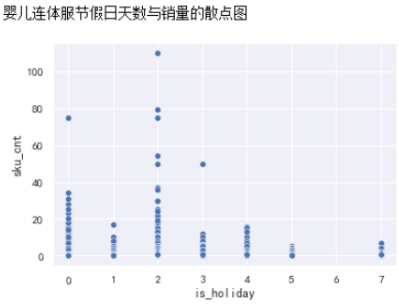


图 34 婴儿连体服门店面积与店铺等级分析

➢ 工作日和节假日天数

在工作日和节假日与销量的关系图中，工作日天数表示每个销售周包含的工作天，与节假日呈线性关系。从散点图中可以看出，工作日天数较多，购买婴儿连体服的销量比较高，这可能是因为人们比较多的选择在工作日出行。通过计算他们与销量之间的相关系数可以看出，与销量的相关性比较高，因此作为建模的指标之一。





婴儿连体服工作日与销量相关系数

	sku_cnt	is_workday
sku_cnt	1.000000	-0.022776
is_workday	-0.022776	1.000000

	sku_cnt	is_workday
sku_cnt	1.000000	-0.031194
is_workday	-0.031194	1.000000

	sku_cnt	is_workday
sku_cnt	1.000000	-0.029496
is_workday	-0.029496	1.000000

婴儿连体服节假日与销量相关系数

	sku_cnt	is_holiday
sku_cnt	1.000000	0.022776
is_holiday	0.022776	1.000000

	sku_cnt	is_holiday
sku_cnt	1.000000	0.031194
is_holiday	0.031194	1.000000

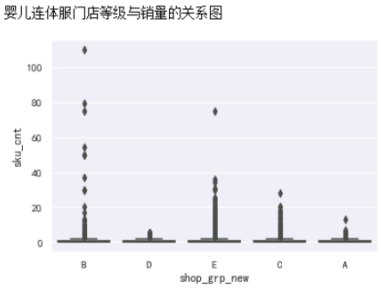
	sku_cnt	is_holiday
sku_cnt	1.000000	0.029496
is_holiday	0.029496	1.000000

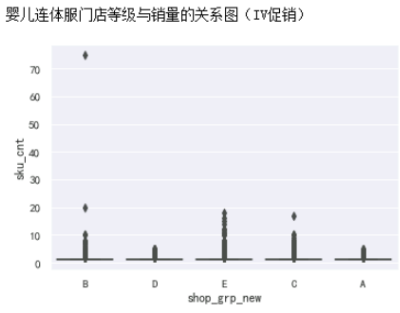
图 35 婴儿连体服节假日工作日指标分析

（2）离散型指标分析

➢ 门店等级

从下图可以看出，不同的门店等级下销量情况有一定的差别，然后通过使用方差检验，结果显示门店等级存在显著的影响，因此作为建模的特征之一。





婴儿连体服门店等级与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
shop_grp_new	4.0	604.349117	151.087279	441.92201	0.0
Residual	1440194.0	492383.244406	0.341887	NaN	NaN

婴儿连体服门店等级与销量的方差检验（IV促销）

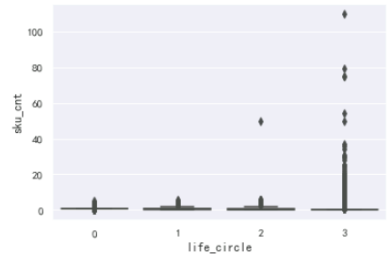
	df	sum_sq	mean_sq	F	PR(>F)
shop_grp_new	4.0	96.821569	24.205392	71.719969	8.204783e-61
Residual	177634.0	59951.234072	0.337499	NaN	NaN

图 36 婴儿连体服门店等级指标分析

➢ 生命周期分析

从下图可以看出，婴儿连体服品类在第一和第二个生命周期的销量并不高，这是因为该产品刚刚进入市场，很多门店甚至还没有开始配货，所以销量低迷。但整体上呈现一个越往后走，销量越高的趋势。方差检验的结果同样显示，在不同的生命周期下，销量存在显著的区别，因此将其当作建模的特征之一。

婴儿连体服生命周期与销量的关系图



婴儿连体服生命周期与销量的方差检验

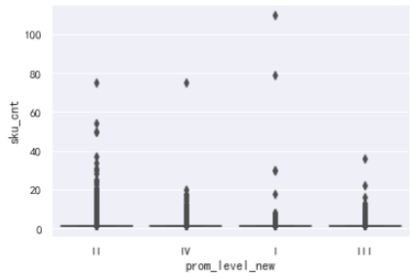
	df	sum_sq	mean_sq	F	PR(>F)
life_circle	1.0	10117.102884	10117.102884	30175.008631	0.0
Residual	1440197.0	482870.490639	0.335281	NaN	NaN

图 37 婴儿连体服生命周期分析

➢ 促销分析

由于不同的促销等级持续的时间周期长度不同，现对各个促销等级进行下抽样，之后再重新画图。其中，四个促销等级中，三级促销的数据量最少，将其他三个等级的历史数据也采样为同三级促销一样的比例。从图 38（左上）可以看出，不同的促销等级带来的销量增长情况不同，而且在无促销时候的销量情况反而更好，进一步对历史数据按照促销等级作为指标进行下采样后再展开分析。从图 38（右上）可以看出，一级促销的销量更大，各个级别的销量依次递减。并且方差分析结果显示，销售等级之间的销量存在显著的差别，因此，**将促销等级作为特征纳入模型。**

婴儿连体服促销等级与销量的关系图



婴儿连体服促销等级与销量的方差检验

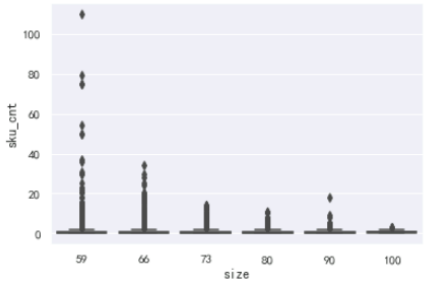
	df	sum_sq	mean_sq	F	PR(>F)
prom_level_new	3.0	155.40800	51.802667	164.983696	6.902061e-107
Residual	400161.0	125645.18457	0.313987	NaN	NaN

图 38 婴儿连体服促销等级分析

➢ 尺码

对于婴儿连体服品类，通过尺码和销量的散点图可以看出，不同的尺码销量情况并不一致，但 73-100 的尺码销量情况基本集中在 0-20 之间，进一步使用方差检验，发现 F 只小于 300，因此放弃其作为建模的特征之一。

婴儿连体服尺码与销量的关系图



婴儿连体服尺码与销量的方差检验

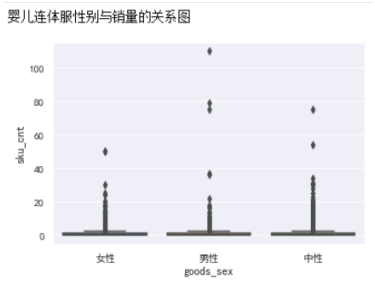
	df	sum_sq	mean_sq	F	PR(>F)
size	1.0	73.625733	73.625733	215.119812	1.057916e-48
Residual	1440197.0	492913.967790	0.342255	NaN	NaN

图 39 婴儿连体服尺码指标分析

➢ 性别

对于婴儿连体服品类，通过性别和销量的散点图可以看出，除了个别异常值之外，三个性别的销量情况基本集中，进一步以来方差检验的结果，显示三

个类别并没有显著的区别，因此放弃其作为建模的指标之一。



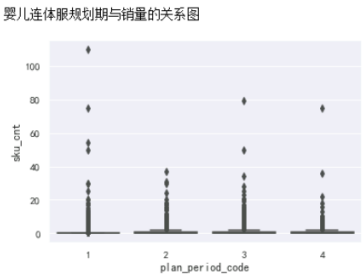
婴儿连体服性别与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
goods_sex	2.0	96.753913	48.376956	141.354421	4.135990e-62
Residual	1440196.0	492890.839610	0.342239	NaN	NaN

图 40 婴儿连体服性别指标分析

➢ 规划期

由于连体服是强季节性产品，不同的季节有对应的款式，所以需要考虑一下产品季节和对应的春夏秋冬之间的关系，并且方差检验结果显示，不同的规划期，存在显著的区别，因此作为建模的特征之一。



婴儿连体服规划期与销量的方差检验

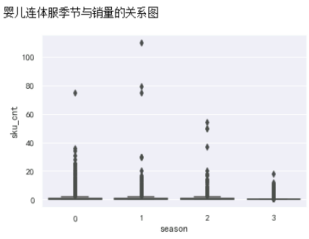
	df	sum_sq	mean_sq	F	PR(>F)
plan_period_code	1.0	2003.263729	2003.263729	5876.143572	0.0
Residual	1440197.0	490984.329793	0.340915	NaN	NaN

图 41 婴儿连体服规划期分析

➢ 季节

由于连体服是强季节性产品，不同的季节有对应的款式，所以需要考虑一下产品季节和对应的春夏秋冬之间的关系，并且方差检验结果显示，不同的季节，

存在显著的区别，因此作为建模的特征之一。



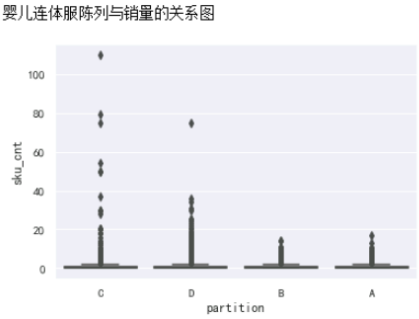
婴儿连体服季节与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
season	1.0	1111.011032	1111.011032	3253.000471	0.0
Residual	1440197.0	491876.582491	0.341534	NaN	NaN

图 42 婴儿连体服季节分析

- 陈列

产品的陈列影响顾客的注意力，在比较明显的区域，顾客更容易选择，方差检验的结果也显示，区域对于销量存在明显的影响，因此将其作为特征放入模型。



婴儿连体服陈列与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
partition	3.0	414.693513	138.231171	404.163204	1.844798e-262
Residual	1440195.0	492572.900010	0.342018	NaN	NaN

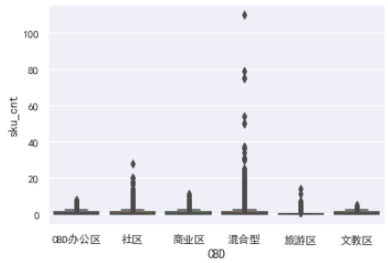
图 43 婴儿连体服陈列区域分析

- CBD

CBD 的是因为，在社区以及混合型区域，家庭分布更密集，对连体服的需求更大，箱型图的统计结果也坐实了这一观点，方差检验进一步说明了不同的

区域销量情况不同，因此作为建模的特征引入模型中。

婴儿连体服商业区与销量的关系图



婴儿连体服商业区与销量的方差检验

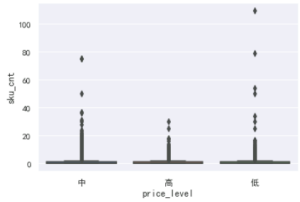
	df	sum_sq	mean_sq	F	PR(>F)
CBD	5.0	684.225258	136.845052	400.328939	0.0
Residual	1440193.0	492303.368265	0.341832	NaN	NaN

图 44 婴儿连体服商业区分析

➢ 价格等级

不同的价格水平，会影响顾客选择，因为连体服是单件的产品，所以价格水平上会比棉柔巾更为明显，方差检验结果也显示，不同的价格水平之间的销量存在明显的区别，因此将其作为建模的候选特征之一。

婴儿连体服价格等级与销量的关系图



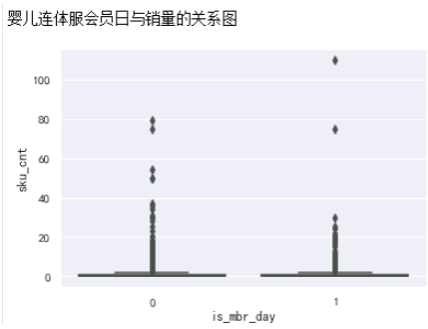
婴儿连体服价格等级与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
price_level	2.0	554.958097	277.479049	811.53073	0.0
Residual	1440196.0	492432.635426	0.341921	NaN	NaN

图 45 婴儿连体服价格等级分析

➢ 会员日

由于对于连体服来说，销量集中在 0-5 件的区间之内，很少能观测到销量的暴涨，即便是在会员日，因此，从方差检验的结果可以看出，是否会员日对于连体服品类的影响并不大，但为了尽量将大销量的情况捕捉，还是将其作为模型的特征之一。



婴儿连体服会员日与销量的方差检验

	df	sum_sq	mean_sq	F	PR(>F)
is_mbr_day	1.0	21.540345	21.540345	62.929972	2.143399e-15
Residual	1440197.0	492966.053178	0.342291	NaN	NaN

图 46 婴儿连体服是否会员日分析

2.2.3 建模特征选择

根据上述特征工程分析结果，基础棉柔巾品类选择的建模特征如表 16：

表 16 婴儿连体服特征选择结果

字段	描述	字段	描述
divide_goods_amt	到手价	price_level	价格水平
discount	价格折扣	prom_level_new	按销量情况重新归类后的促销等级
season	当天所处季节	is_workday	工作日天数
life_circle	订单日所处的生命周期 0，1，2	is_holiday	节假日天数
store_square	门店面积	CBD	所处商圈
plan_period_code	规划期	Is_mbr_day	是否包含会员日
temperature_diverge	温度偏移指数	partiton	陈列区域

在选择好各个指标之后，同基础棉柔巾品类一样，将上述的连续型指标进行归一化和标准化处理，然后将离散型指标进行 0-1 编码。其中，将门店面积

等其他指标进行标准化缩放，缩放方式如下：

$$x_{new} = \frac{x_{std}^{old} - \mu}{s}$$

其中，μ指指标值的均值，s为指标值的标准差。

2.3 模型设计

2.3.1 模型设计

➢ 模型思路

通过将各个 SKU 按照门店划分，统计其历史数据量，如下图所示，从图中可以发现各SKU的历史数据量较少，按单个SKU预测会导致历史数据量较少，模型的精度效果降低，因此，采用分层聚类的方法。

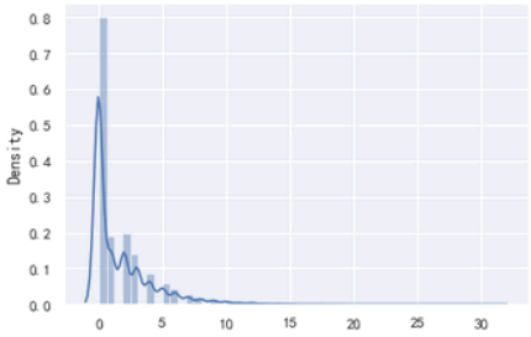


图 47 婴儿连体服门店数据量统计

在时间点t（天/周/半月等）下，隶属于类型$c(c \in C)$的门店$s(s \in S)$下产品（SKU）$i(i \in I^{cs})$的特征向量为$X_{i,t}^{cs}$，包括物料描述、温度、季节、折扣、促销等。其中I^{cs}表示类型c的门店s下包含的所有婴儿连体服类产品（SKU）。$d_{i,t}^{cs}$则为该产品在时间点t下的真实需求量。

当$t \leq 10$表示该产品为全新品，没有历史数据，需要借助已有产品的历史数据作为协变量来预测：

$$\hat{d}_{i,t}^{cs} = \sum_{j \in I^c, j \neq i} w_{ij} \sum_{t,x,s} X_{j,t,x}^{cs} B_x(t)$$

其中，$\hat{d}_{i,t}^{cs}$为该产品在时间点t下的预测需求量；$X_{j,t,x}^{cs}$表示产品$j \in I^c$于时间t时在

门店s下的第x个特征；$B_x(t)$表示对特征x在时间t对销售的影响进行建模的系数函数；w_{ij}表示产品j对产品i的影响权重；

当$t > 10$表明该产品已经累积了一定的历史数据，可以结合自身时序需求数据和部分协助数据来建模：

$$\hat{d}_{i,t}^{cs} = \alpha * F\left(X_{i,t}^{cs}, X_{i,t-1}^{cs}, \dots, X_{i,0}^{cs}\right) + (1-\alpha) * \sum_{j \in I^{cs}, j \neq i} w_{ij} \sum_{t,x} X_{j,t,x}^{cs} B_x(t)$$

其中，$F(\cdot)$表示对产品i自身历史数据建模的函数；$\alpha \in (0,1]$表示自身历史数据对未来需求贡献的权重。

- 模型的预测效果评价指标

平均绝对百分比误差计算公式：

$$\mathrm{MAPE} = \frac{1}{\mathrm{I} * \mathrm{T}} \sum_{\mathrm{i}=1}^{I} \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{|\hat{d}_{i,t}^{cs} - d_{i,t}^{cs}|}{d_{i,t}^{cs}}$$

其中， MAPE 值越小说明误差越小；

需求预测精度：

$$\mathrm{ACU} = 1 - \mathrm{MAPE}$$

按全棉 14 天滚动预测方法计算出来的预测误差为$\mathrm{MAPE_P}$，港中深团队预测方法的误差为$\mathrm{MAPE_G}$，则本项目误差改善百分比为：

$$\mathrm{ACP} = \frac{\mathrm{MAPE_P} - \mathrm{MAPE_G}}{\mathrm{MAPE_P}} * 100\%$$

针对补货需求预测的精度为，ACP 的值。

2.3.2 模型搭建

基于特征工程的结果，将选中的特征放入模型，并使用不同的建模维度尝试模型的预测效果，选择最好的结果。

（1）分类建模的维度——c 的选择

表 17 婴儿连体服不同建模维度方式结果

分类维度	算法	MAPE 值	ACP
City	DT	0.32	0.80
Province	**DT**	**0.29**	**0.83**
Shop_grp_new	DT	0.30	0.82

根据上述的结果，将历史数据**门店按照省份**来划分，并按各门店等级的线下门店数据来展开需求预测，基本思路如下图所示。

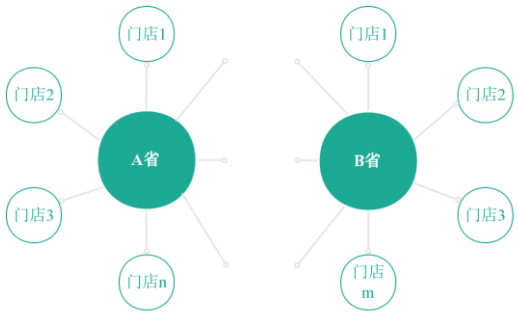


图 48 婴儿连体服需求预测模型设计

（5）建模算法选择

使用不同的算法尝试模型的预测效果，选择最好的结果对应的算法作为目标算法。

表 18 婴儿连体服不同建模维度方式结果

分类维度	算法	MAPE 值	ACP
Province	DT	**0.29**	**0.53**
	XGB	0.81	0.30
	RF	0.58	0.50
	GBR	0.89	0.0

根据上述的结果，将使用决策树 DT **模型**来建模。

（6）历史数据筛选

使用不同的历史数据范围尝试模型的预测效果，选择最好的结果对应的算法作为目标算法。

表 19 婴儿连体服不同建模维度方式结果

分类维度	算法	历史数据量	MAPE 值	ACP
Shop_grp_new	DT	2021.01 至今	0.32	0.81
		2021.06 至今	0.33	0.78
		2022.01 至今	**0.07**	**0.99**
		2022.06 至今	0.07	0.99

根据上述的结果，将使用 **2022 年之后的历史数据**来建模。

（7）异常值剔除

使用不同的历史数据范围尝试模型的预测效果，选择最好的结果对应的算法作为目标算法。

表 20 婴儿连体服不同建模维度方式结果

分类维度	算法	周销量最大阈值	MAPE 值	ACP
Shop_grp_new	DT	500	**0.07**	**0.99**
		300	**0.07**	**0.99**
		100	**0.07**	**0.99**
		50	**0.07**	**0.99**

根据上述的结果，发现，对于连体服品类，高销量占比并不高，因此不影响模型的精度，因此历史数据销量异常只剔除大于 900 的部分（特别异常的团购）。

根据上述步骤，最终确认的方法为——**使用 2022.01 至今的历史订单数据，然后将历史周销量大于等于 900 的部分剔除，然后使用决策树算法训练模型**；

三、预测结果总结

1. 模型总结

本次建模针对内裤和婴儿连体服三级品类的线下门店的需求来建模，模型考虑的特征包括物料描述、促销、折扣等内部特征，以及天气、季节、节假日等外部特征来建模。通过对原始数据的筛选、清洗和分类汇总等预处理，分别获得了两个品类下的历史需求数据。在此基础上，两个品类分别基于不同的维度来施行分类建模，预测颗粒度细至 SKU-单门店-周的维度。其中，内裤基于店铺等级的维度来划分建模，共涉及 4240 个 SKU 在 366 个门店的预测模型设计；而婴儿连体服品类基于省份的维度来划分建模，共涉及 2280 个 SKU 在 365 个门店的预测模型设计。

模型样本外预测结果显示，模型的预测效果均在公司目前按历史 14 天滚动

平均预测的方法有了提升，对于两个品类，由于销量的数据波动集中在 0-5 之间，移动平均法可以很好的刻画模型的需求，因此基于机器学习建模的方法获取的精度提升并不明显。

总之，对于 14 天移动平均的方法比较适用于需求趋势较为平缓的数据，而对于受外界因素影响较明显的产品，刻画其需求模式需要结合特征工程和机器学习。而判断数据波动可以结合产品的变异系数来决定，对于变异系数大于 1 的产品，使用分层机器学习的方法可以获得更好的预测精度，而对于变异系数在 1 以下的产品，可以采纳 14 天滚动平均的方法。

2. 模型比较

对内裤、婴儿连体服三个三级品类分别建模，预测每个品类下单个门店单个 sku 的销量。

首先我们对三个品类的产品分别分析影响销量的因素，选取合适的作为建模特征，最终选定的特征如下表所示。

➢ 三个品类特征选择结果

序号	内裤	婴儿连体服
1	sku_code	sku_code
2	shop_id	shop_id
3	ord_create_time	ord_create_time
4	sku_cnt	sku_cnt
5	shop_grp_new	province
6	discount	discount
7	plan_period_code	life_circle
8	goods_sex	store_square
9	size	plan_period_code
10	prom_level_new	temperature_diverge
11	user_grp	partition
12	waist	price_level

13	partition	prom_level_new
14	is_workday	is_workday
15	is_holiday	is_holiday
16	season	CBD
17	store_square	season
18	CBD	is_mbr_day
19	temperature_diverge	
20	life_circle	
21	prom_level	
22	is_mbr_day	

➢ 结果比较

我们首先通过公司现有算法进行了需求预测，并计算了需求预测误差。之后对各个品类分别尝试多种机器学习算法进行预测，最终选择表现最好的作为模型输出。针对这两个品类，我们最终选定的算法相比公司现有算法在预测准确性上都有了很明显的提升。

对于精确到单门店单 sku 单周的需求预测，我们的方法**对于内裤品类可以保证 75%的单门店 sku 预测误差在 0.3 以内，预测准确率在 70%以上；对于婴儿连体服品类可以保证 50%以上的单门店 sku 预测误差在 0.28 以内，预测准确率大于 72%**。

品类	14 天滚动平均	机器学习算法	机器学习精度	提升情况
内裤	13%	Lasso		
		XGBoost		
		Decision Tree		
		Random Forest		
		MLP		
		GBDT		
婴儿连体服	27%	Lasso		
		XGBoost		

		Decision Tree		
		Random Forest		
		MLP		
		Gradient Tree Boosting(GBDT)		

3. 业务建议

- 改进建议

影响线下购物的因素很多，比如在疫情当下，门店可能因为疫情原因选择闭店，或者周围的居民区可能临时封控，导致线下门店的需求转移到线上去，而这些因素目前都无法观测，可能会导致目前的变量特征没办法完全捕捉需求。随着线上购物的兴起，线上给顾客提供了更透明的价格，更多的产品选择，用户在促销、天气极端等环境下可能会从线下转移至线上购物，将线下和线下剥离，忽略全渠道的模式，也是线下需求预测精度提升的较大阻碍之一。

现在业界对需求预测的精度衡量标准并不相同，而且预测的颗粒度不同，预测的颗粒度越细，模型的预测效果越发难以保证，具体到 SKU-天-门店属于最细颗粒度的需求预测，影响需求的因素更加复杂多变，不可估计。

因此，在后续的需求预测中，可以引入更多的内外部特征，把更多的精力留给数据分析挖掘的工作，深入考虑各个特征对于需求的影响，而不是完全依赖业务经验来建模预测。

- 提升预测精准度的建议

增加对线下门店店庆等活动的记录；

门店主推产品，与销售员业绩有挂钩的产品，记录要具体到时间范围、SKU、门店等；

对促销活动，除记录时间之外，记录活动机制；

增加对城市发展水平、消费指数、新生儿数量等宏观数据的收集；

保留门店产品陈列位置及相关的轮换数据记录；

增加对用户/会员画像信息的收集与刻画，其中消费者基础特征包括年龄、性别、受教育程度、区域、顾客分类，消费者历史购买特征包括历史线上线下

购买情况统计、购物偏好、兴趣等。

给门店的团购行为主动的打标签，标记好团购数据，方便后期模型训练时候，主动将不可预估的因素剔除掉；

附录

1. 成本节省估算

当预测销量低于实际销量时将差值记为因缺货造成的需求损失成本，当预测销量高于实际销量时将差值记为因库存过多造成的库存成本（注：上述成本剔除了管理、运费等，只以吊牌价估算），具体计算逻辑如下：

当真实销量高于预测销量时，部分顾客需求无法得到满足，因此产生需求损失，计算公式为$(\textbf{真实销量} - \textbf{预测销量}) \times \textbf{吊牌价}$；

当真实销量低于预测销量时，会导致部分库存积压，因此产生库存成本，计算公式为$(\textbf{预测销量} - \textbf{真实销量}) \times \textbf{吊牌价}$。

取内裤品类的两个 sku 为例估算建模获得的成本节约（单位为元）。从下表的结果可以看出，算法产生的结果，均可以在目前 14 天滚动的方法上，节省一定的成本，其中，2000180004-080 在 3 个月的周期内，节省了 5632 元的成本。而该成本取决于需求预测的准确度，准确度越高可以保证成本越小。

Sku	方法	库存成本（元）	需求损失（元）	总成本（元）	节省（元）
2000180004-080（吊牌价 88 元，2022/5/1-2022/8/1）	全棉	40*88	31*88	6248	5632
	算法	2*88	5*88	616	
200017007-100（吊牌价 88 元，2022/5/1-2022/8/14）	全棉	14*88	13*88	2376	2200
	算法	1*88	1*88	176	

2. 业界需求预测参考结果梳理

- 宝洁依据过去积累的大数据提高精度，使需求预测的“准确率”从过去的 3 成左右增加到了 8 成左右。在推行 Business Sphere 后，宝洁在全球的库存减少了约 25%，节约资金数千万美元，已经逐渐显现出了成效。

  链接：https://mp.weixin.qq.com/s/5918aj0EkobeNZV1wpBAEw

- 在需求预测方面，京东采集线上的销售数据和浏览数据，对评论进行分析，收集购物车放置痕迹，与行业和媒体合作分析和理解用户的行为，并通过第三方数据公司进行行业趋势分析，同时，监控突发事件、政府政策和社会新闻，关注品类实际发生数据，都为需求分析提供了大数据支持。为了保证预测的准确，在京东内部，数据都是透明的，营销策略也是透明的，所有的协调和沟通机制也是透明的，因此，京东的预测准确率才能达到 95%以上，更可以准确预测单品类、单仓、单供应商、甚至是 SKU，这样才可以有效的降低物流成本同时又提升用户体验。

  链接：https://mp.weixin.qq.com/s/na38NRq5xHTxZ1OjKYbdaQ

- 屈臣氏供应链部门的预测团队每次促销前需要花费大量精力对未来促销商品进行到仓层面的预测，预测的准确率（sell thru）只有 40%；而商品到店的预测只能通过阈值（期望求出）从仓开始切分，预测的准确率极低；促销档期结束后，没有全面地对预测偏差原因洞察分析，无法快速定位问题和提升管理效率。和九章数据合作后，通过整合多渠道销售和市场数据，考虑促销力度、广告投放力度、节假日等多种因素，目前屈臣氏已经实现上千余种 SKU 周颗粒度的滚动销量预测，为其多渠道、多级库存网络下的库存优化提供参考。在商品/门店/周颗粒度下，需求预测准确率从 40%提升至 85%，平均日缺货率下降 1.73% ，平均月周转天数下降 4.5%，店+仓库库存成本下降 30%。九章数据致力于从需求端切入，为零售企业打造敏捷、弹性、可持续的供应链。

  链接：https://mp.weixin.qq.com/s/AkFh-U1IQsD9V2Q777dmYw

- 美国有个叫 ToolsGroup 的公司做过调研，说消费品的预测准确度在 85%上下，零售业跟消费品差不多，而工业品就低得多，只有 70%不到。根据 Gartner 的调查，消费品公司的预测准确度在 50%到 60%之间。这一数字明

显低于 ToolsGroup 的统计。不过 Gartner 的统计方法相当严苛，是基于 SKU 和库位层面，统计方法是绝对差的百分比（MAPE）。

链接：https://mp.weixin.qq.com/s/GkHXW6GPsWDcUnsKWuIqUA

3. 论文需求预测参考结果梳理

- 京东新产品生命周期预测结果

论文作者使用生命周期拟合的预测方式，考虑了促销因素，使用京东 2019-2020 年推出的 330 款笔记本电脑产品（sku）的销售数据预测新上市产品销量。在京东原有方法的基础上提升了 33.64%（减少了 33.64%的预测误差），已被京东采用。

Table 5 New Product Forecasting Performance Comparison (Normalized Forecasts)

Model		Overall RMSE		Relative reduction in RMSE over JD.com's forecasts (across SKUs)	
		Prelaunch	4-Week Update	Prelaunch	4-Week Update
FB	time-invariant	0.0613	**0.0502**	**4.52%**	**33.64%**
	time-varying	**0.0610**	0.0504	4.10%	31.05%
FPCA	time-invariant	0.0700	0.0638	2.31%	10.02%
	time-varying	0.0647	0.0643	2.72%	10.50%
poly4		0.0768	0.1274	-2.03%	-19.81%
TiGo-ETS		0.0694	0.0625	2.57%	27.18%
Sood et al. (2009)		-	0.0626	-	29.86%
JD.com		0.0744	0.0649	-	-

链接：https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4014586

- 阿里巴巴需求估计结果

论文主要预测淘宝推荐页产品顾客的购买概率，从而给顾客推荐更合适的产品集合来提高顾客购买概率从而提高顾客访问页面的收益。阿里巴巴原有方法每次客户访问产生 4.04 元收益，改进后每次产生 5.17 元收益，使客户访问的收益提高了 24%。

下表展现了论文使用的需求预测方法的效果，准确率衡量是否成功预测到顾客对推荐页产品的购买行为，其中表现最好的模型准确率为 77.5%。

Table 4. Model Prediction Performance: Summary Statistics of Prediction Performance on Purchases

	MNL	SF-ML	MNL	AF-ML
ClassificationAccuracy	36.31%	74.55%	36.31%	77.50%
Difference (all *p*-values)	38.24% (< 0.0001)		41.19% (< 0.0001)	
AverageRank	2.51	1.51	2.51	1.43
Difference (all *p*-values)	1.00 (< 0.0001)		1.08 (< 0.0001)	
Observations	82,957	68,395	82,957	86,238

Notes. Standard errors are robust and clustered at the customer level. Reported is the average prediction power of customers' purchasing behaviors during our experiment.

链接：https://doi.org/10.1287/opre.2021.2158