**** 

**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 3](#_Toc3354)

[表目录 4](#_Toc9737)

[一、总体模型思路 6](#_Toc32072)

[1. 建模目标 6](#_Toc6682)

[2. 整体规划 6](#_Toc482)

[3. 结果初览 9](#_Toc24065)

[二、内裤和婴儿连体服三级品类模型设计 9](#_Toc25776)

[1. 内裤品类 9](#_Toc16175)

[1.1数据预处理 9](#_Toc25855)

[1.2 特征工程 16](#_Toc4309)

[1.3 模型设计 33](#_Toc15854)

[2. 婴儿连体服品类 37](#_Toc19333)

[2.1数据预处理 37](#_Toc12297)

[2.2 特征工程 42](#_Toc14035)

[2.3 模型设计 55](#_Toc12521)

[三、 预测结果总结 58](#_Toc21755)

[1. 模型总结 58](#_Toc14517)

[2. 模型比较 59](#_Toc18660)

[3. 业务建议 61](#_Toc9394)

[附录 63](#_Toc28908)

[1. 成本节省估算 63](#_Toc8001)

[2. 业界需求预测参考结果梳理 64](#_Toc1379)

[3. 论文需求预测参考结果梳理 65](#_Toc16649)

### 图目录

[图 1 需求预测建模整体技术路线 7](#_Toc27221)

[图 2 历史数据拼接过程 11](#_Toc31930)

[图 3 内裤品类销量密度函数图 13](#_Toc2215)

[图 4 内裤品类字段汇总及其方式 13](#_Toc7633)

[图 5 内裤品类补全后价格水平与吊牌价对应关系 15](#_Toc22724)

[图 6 内裤品类产品吊牌价和销量之间的关系 17](#_Toc8666)

[图 7 内裤品类均摊价与销量关系 18](#_Toc7551)

[图 8 内裤品类折扣与销量关系 19](#_Toc3490)

[图 9 内裤品类温度偏移指数和销量之间的关系 20](#_Toc9075)

[图 10 内裤品类销售天数分析图 21](#_Toc6015)

[图 11 各销售等级下店铺面积与销量散点图 22](#_Toc29638)

[图 12 门店面积与等级关系分析图 24](#_Toc10092)

[图 13 重新划分后的门店等级与面积对应关系 24](#_Toc30954)

[图 14 内裤品类门店等级与销量关系图 25](#_Toc21238)

[图 15 内裤品类生命周期与销量箱型图 26](#_Toc2426)

[图 16 内裤品类适用人群分析 26](#_Toc24387)

[图 17 促销等级与销量箱型图 27](#_Toc6756)

[图 18 内裤品类与尺码关系分析结果 27](#_Toc9300)

[图 19 内裤品类性别与销量箱型图 28](#_Toc23629)

[图 20 内裤品类规划期编码与销量关系图 28](#_Toc13703)

[图 21 内裤品类腰型与销量关系图 29](#_Toc18481)

[图 22 内裤品类陈列分区与销量箱型图 29](#_Toc29536)

[图 23 内裤商业区与销量关系分析 30](#_Toc24394)

[图 24 内裤季节与销量关系分析 31](#_Toc27272)

[图 25 内裤价格等级和销量之间的关系图 31](#_Toc8756)

[图 26 内裤会员日与销量之间的关系图 32](#_Toc9238)

[图 27 内裤品类SKU在各门店下历史数据量 34](#_Toc846)

[图 28 内裤品类预测模型设计 36](#_Toc19214)

[图 29 婴儿连体服吊牌价分析 44](#_Toc22627)

[图 30 基础棉柔巾到手价与销量之间的关系 44](#_Toc19239)

[图 31 婴儿连体服价格折扣分析 45](#_Toc3978)

[图 32 婴儿连体服温度偏移指数和销量之间的关系 46](#_Toc19666)

[图 33 婴儿连体服生命周期分析 46](#_Toc11082)

[图 34 婴儿连体服门店面积与店铺等级分析 47](#_Toc12919)

[图 35 婴儿连体服节假日工作日指标分析 48](#_Toc18946)

[图 36 婴儿连体服门店等级指标分析 49](#_Toc31490)

[图 37 婴儿连体服生命周期分析 49](#_Toc15496)

[图 38 婴儿连体服促销等级分析 50](#_Toc11678)

[图 39 婴儿连体服尺码指标分析 50](#_Toc28689)

[图 40 婴儿连体服性别指标分析 51](#_Toc12992)

[图 41 婴儿连体服规划期分析 51](#_Toc12113)

[图 42 婴儿连体服季节分析 52](#_Toc15976)

[图 43 婴儿连体服陈列区域分析 52](#_Toc32483)

[图 44 婴儿连体服商业区分析 53](#_Toc12196)

[图 45 婴儿连体服价格等级分析 53](#_Toc6378)

[图 46 婴儿连体服是否会员日分析 54](#_Toc2389)

[图 47 婴儿连体服门店数据量统计 55](#_Toc19449)

[图 48 婴儿连体服需求预测模型设计 57](#_Toc26677)

### 表目录

[表 1 备选机器学习算法及描述 8](#_Toc16434)

[表 2 需求预测建模结果汇总 9](#_Toc30554)

[表 3 内裤品类剔除字段说明 12](#_Toc8759)

[表 4 内裤品类包含缺失值字段说明 14](#_Toc28357)

[表 5 内裤品类指标分类表 16](#_Toc24908)

[表 6 内裤品类所选特征字段说明表 32](#_Toc23148)

[表 7 内裤不同建模维度方式结果 35](#_Toc27255)

[表 8 内裤不同建模维度方式结果 36](#_Toc17574)

[表 9 内裤不同建模维度方式结果 36](#_Toc30959)

[表 10 内裤不同建模维度方式结果 37](#_Toc4479)

[表 11 婴儿连体服删除数据字段 38](#_Toc22831)

[表 12 婴儿连体服空缺率大于80%的字段 38](#_Toc31334)

[表 13 婴儿连体服保留数据字段 39](#_Toc25161)

[表 14 婴儿连体服字段缺失率 41](#_Toc29222)

[表 15 婴儿连体服指标及其分类 42](#_Toc6525)

[表 16 婴儿连体服特征选择结果 54](#_Toc1304)

[表 17 婴儿连体服不同建模维度方式结果 56](#_Toc13322)

[表 18 婴儿连体服不同建模维度方式结果 57](#_Toc2454)

[表 19 婴儿连体服不同建模维度方式结果 57](#_Toc27460)

[表 20 婴儿连体服不同建模维度方式结果 58](#_Toc13490)

## 一、总体模型思路

### 1. 建模目标

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。

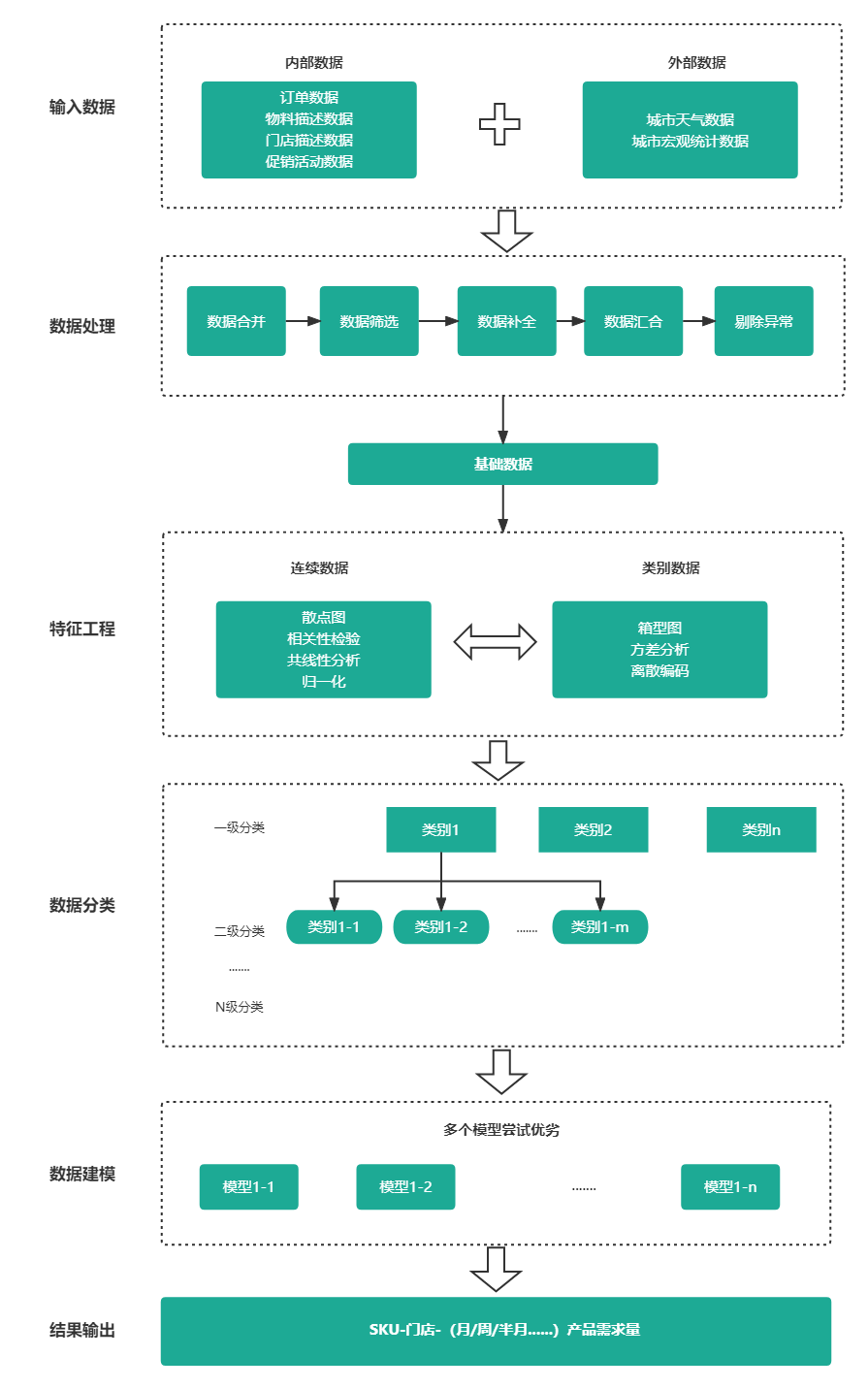


图 1 需求预测建模整体技术路线

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算法。 | 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关系，难以并行训练数据。 |

### 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。**

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% |

## 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

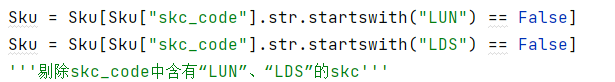
##### 1.1.1 数据合并

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

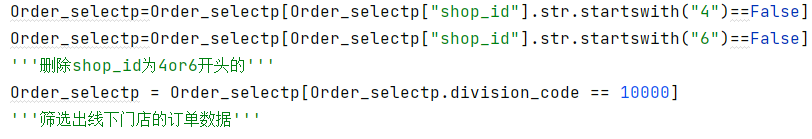
14160050条

（2）删除skc\_code以LUN、LDS开头的skc的订单



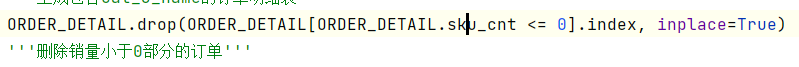
13690935条

（3）删除shop\_id以4和6为开头的订单



13008612条

（4）删除订单明细表中sku\_cnt小于0的部分



11688936条

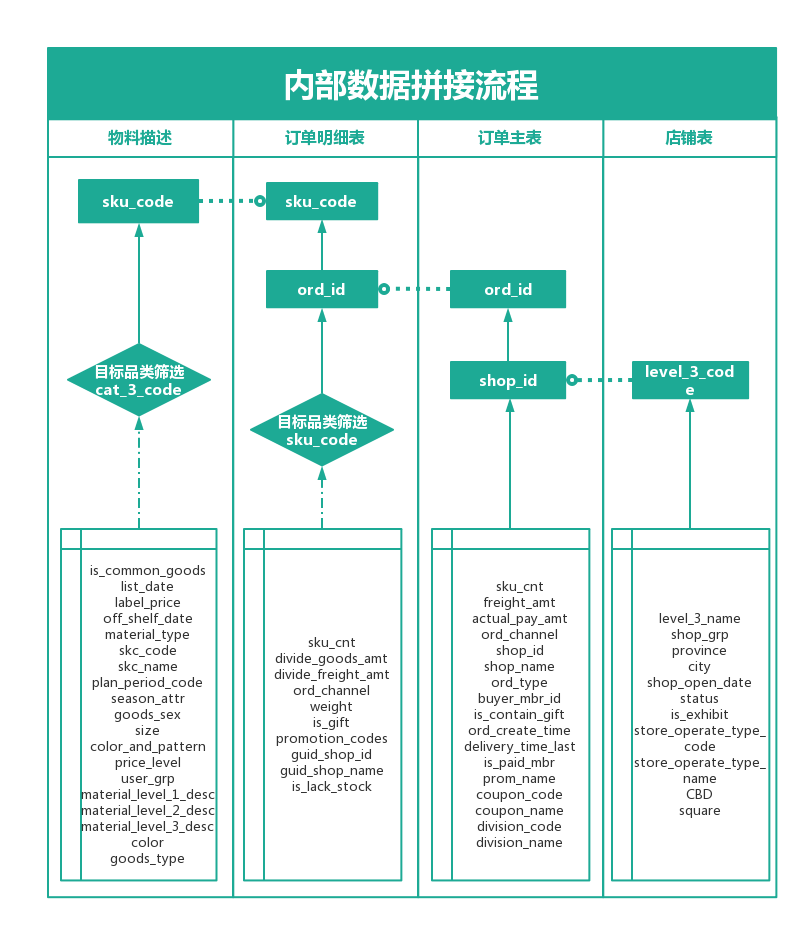


图 2 历史数据拼接过程

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

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：





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。

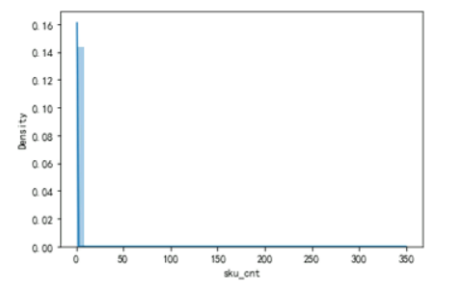
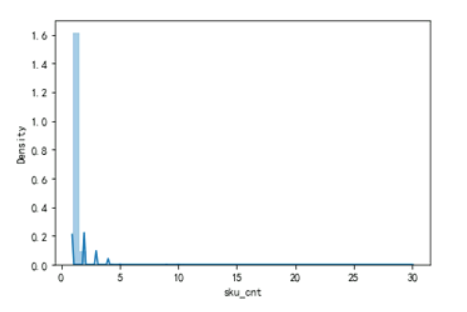
 

图 3 内裤品类销量密度函数图

* 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。



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。

（3）数据补全

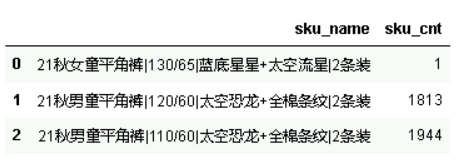
数据中，存在缺失值的字段说明如表 4所示：

表 4 内裤品类包含缺失值字段说明

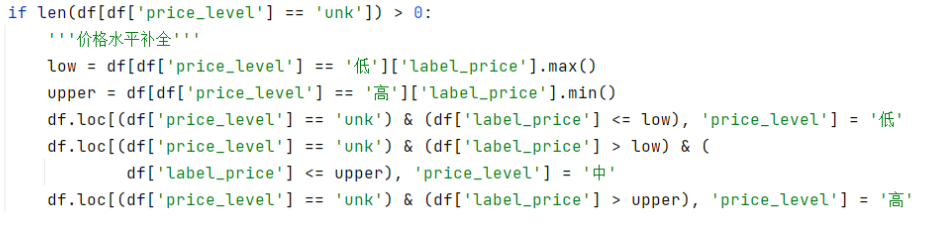
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 字段 | 描述 | 缺失原因 | 是否补全 |
| 1 | price\_level | 价格档位 | 包含unk | 经与业务沟通补全 |
| 2 | color | 颜色 | 包含unk | 无需补全 |
| 3 | shop\_name | 店铺名称 | 包含unk | 不使用，无需补全 |
| 4 | ord\_type\_new | 订单类型（用以剔除展销、团购） | 包含unk | 不使用，无需补全 |
| 5 | buyer\_mbr\_id | 买家会员id | 包含unk | 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：



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：



补全后的结果如下图所示：

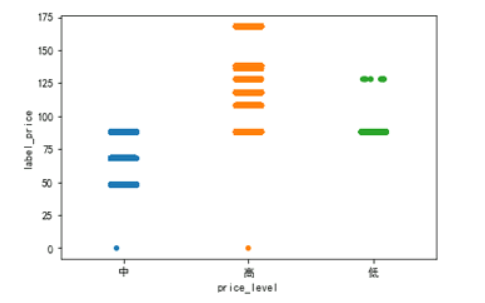


图 5 内裤品类补全后价格水平与吊牌价对应关系

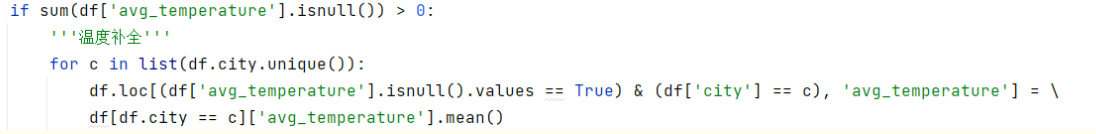
* prom\_level\_new的补全

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：



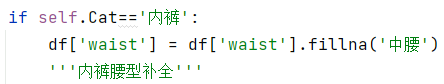
* avg\_temperature的补全

使用门店所在城市对应的年平均温度来补全：



* waist的补全

使用中腰来补全



* partition的补全

使用D区陈列来补全

#### 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

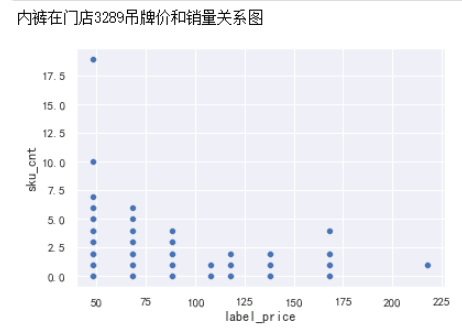
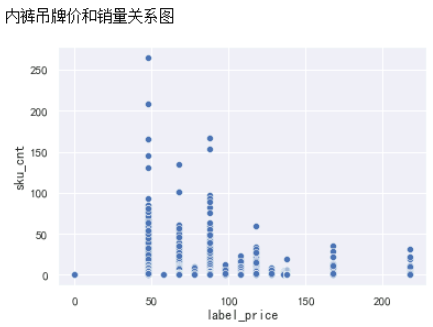
|  |  |  |
| --- | --- | --- |
| 连续型 | 离散型 | 其他 |
| label\_price | price\_level | ord\_create\_time |
| divide\_good\_amt | user\_grp | shop\_open\_date |
| temperature | user\_grp\_coarse\_grain | status |
| sold\_days | 物料类别（materia1\_level\_desc  materia3\_level2\_desc materia3\_level1\_desc） | list\_date |
| 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 |  |

##### 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。



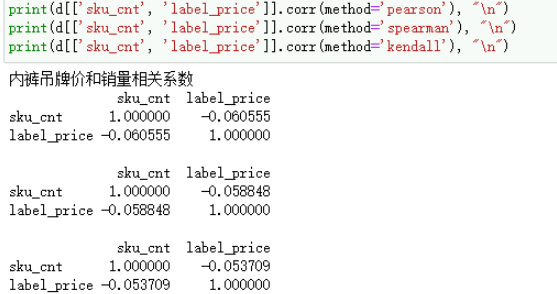


图 6 内裤品类产品吊牌价和销量之间的关系

* 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。

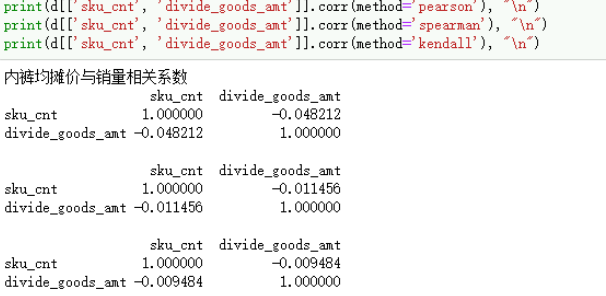
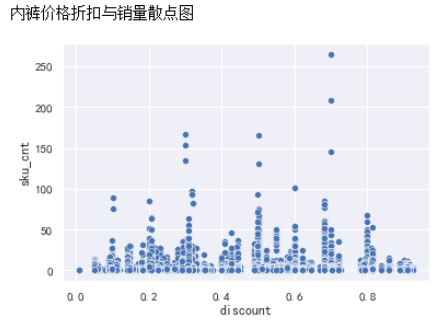


图 7 内裤品类均摊价与销量关系

* 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。



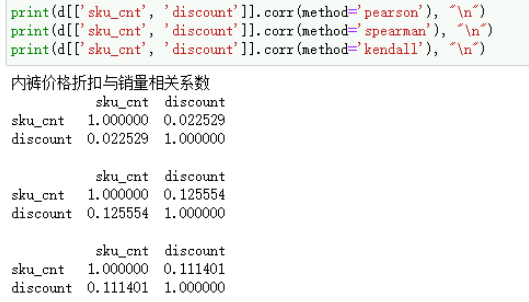


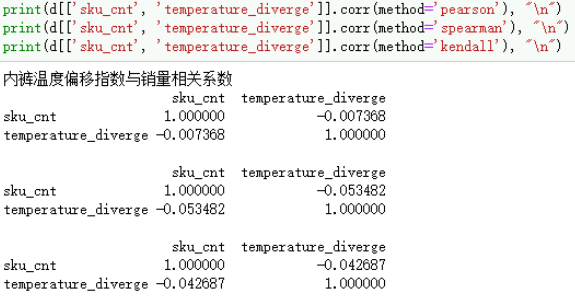
图 8 内裤品类折扣与销量关系

* 温度指标分析

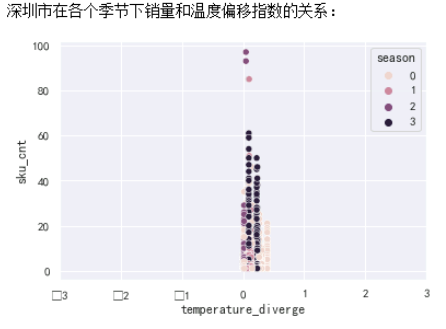
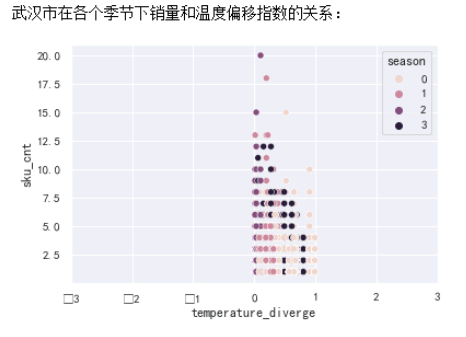
温度偏移指数：

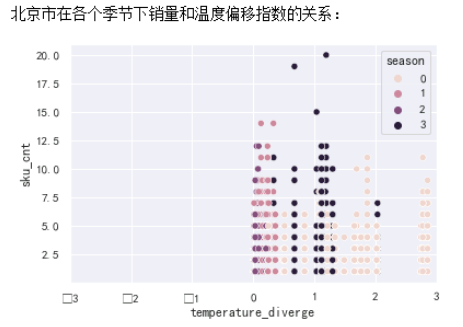
其中mean指所在城市当前所处季节的平均温度。

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。



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。





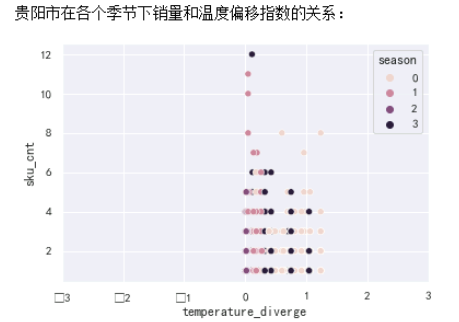
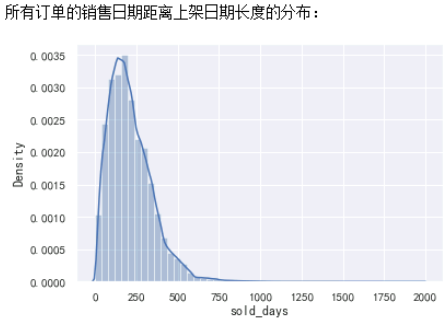


图 9 内裤品类温度偏移指数和销量之间的关系

* 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。



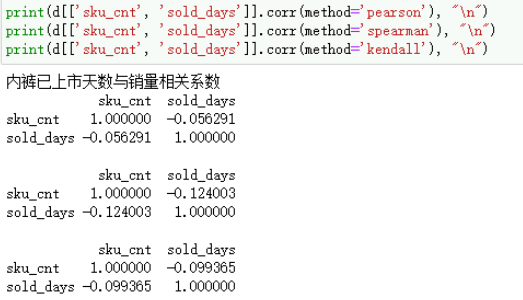
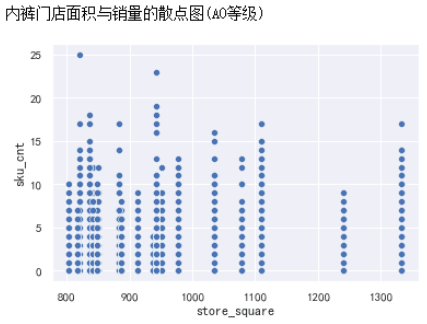
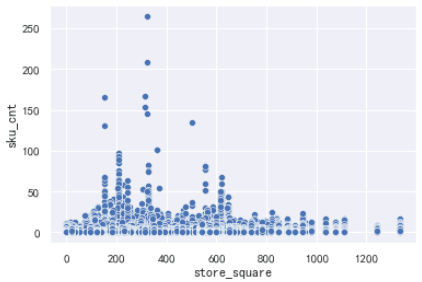


图 10 内裤品类销售天数分析图

* 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。



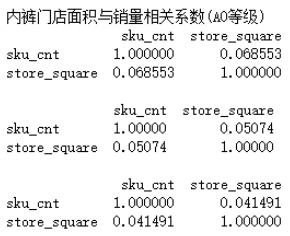
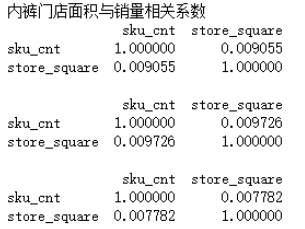
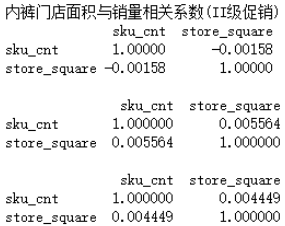
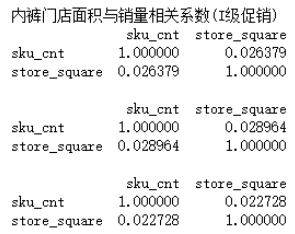
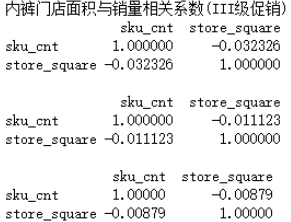


图 11各销售等级下店铺面积与销量散点图

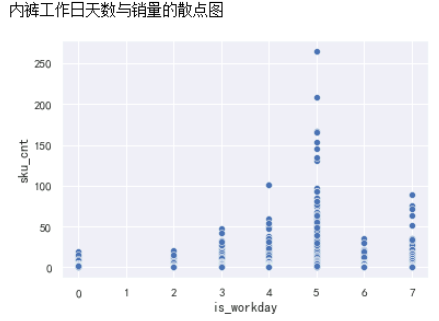
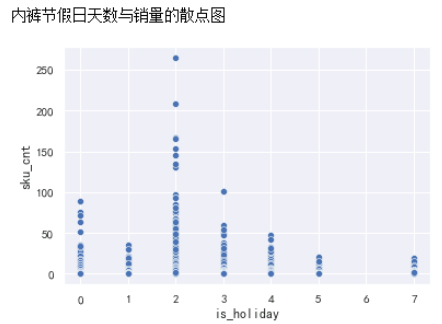
下钻到各个促销等级计算相关性，如下所示，发现具有一定相关性，因此将销售等级和门店面积均纳入建模特征。

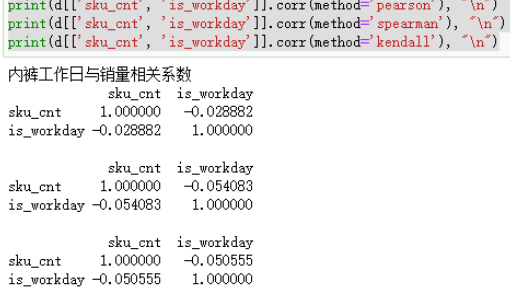


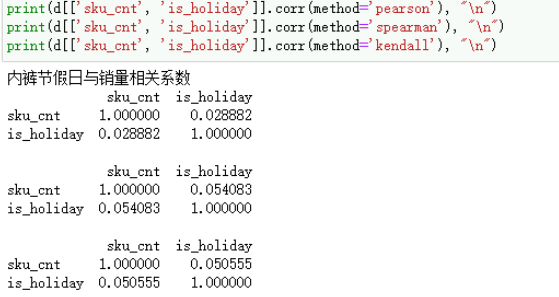


* 工作日天数

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。





（2）离散型指标分析

* 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：

其中，是数据值，是门店面积的平均值，是门店面积的标准差。

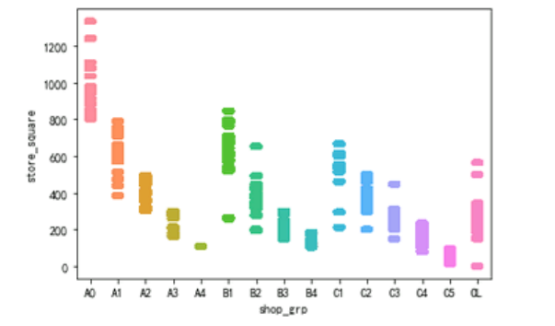


图 12门店面积与等级关系分析图

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。

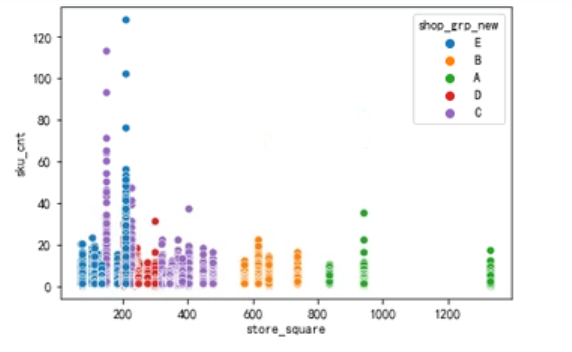
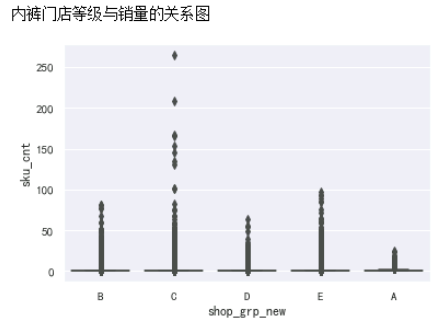
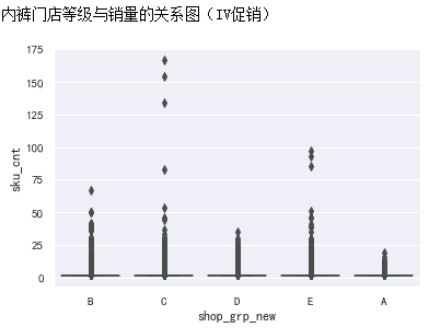


图 13 重新划分后的门店等级与面积对应关系

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。

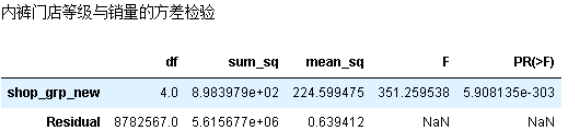




图 14 内裤品类门店等级与销量关系图

* 生命周期分析

根据企业现有划分逻辑：



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。

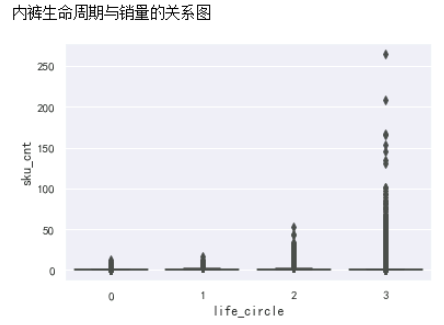
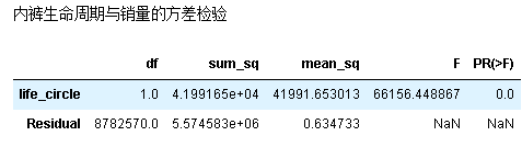
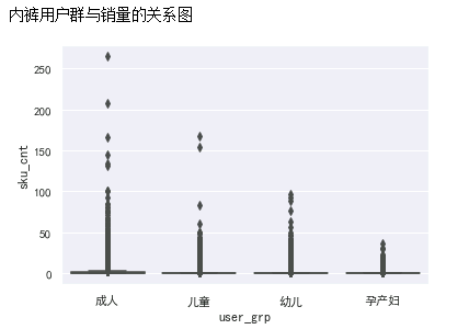
 

图 15 内裤品类生命周期与销量箱型图

* 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。



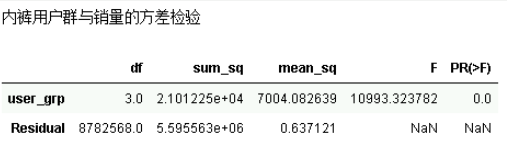


图 16 内裤品类适用人群分析

* 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。

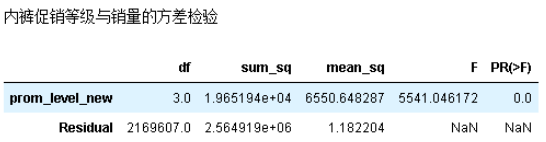
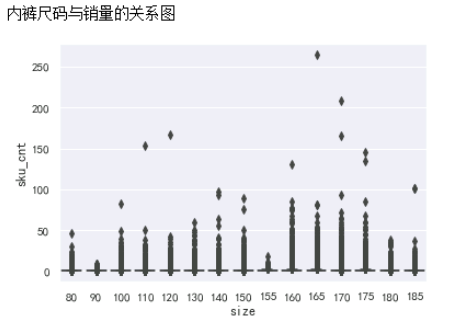


图 17 促销等级与销量箱型图

* 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。



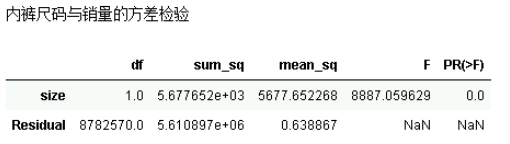


图 18 内裤品类与尺码关系分析结果

* 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。



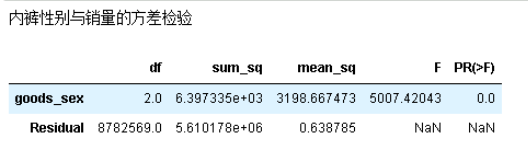


图 19 内裤品类性别与销量箱型图

* 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。



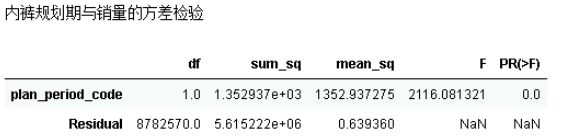
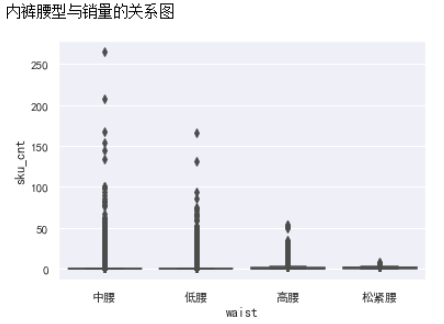


图 20 内裤品类规划期编码与销量关系图

* 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。



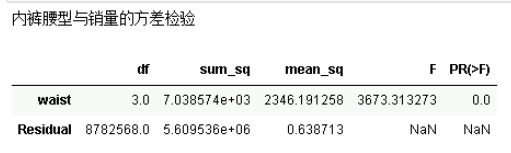
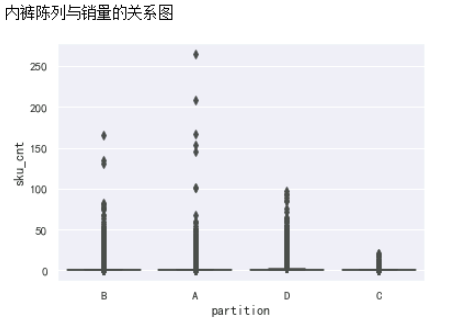


图 21 内裤品类腰型与销量关系图

* 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。



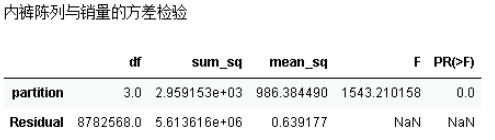
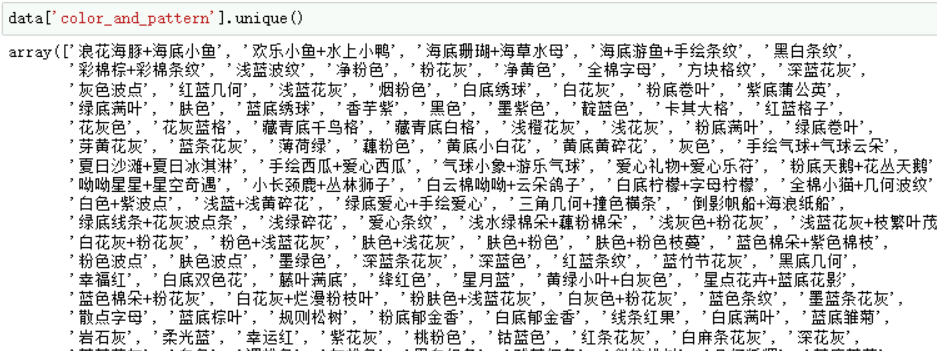
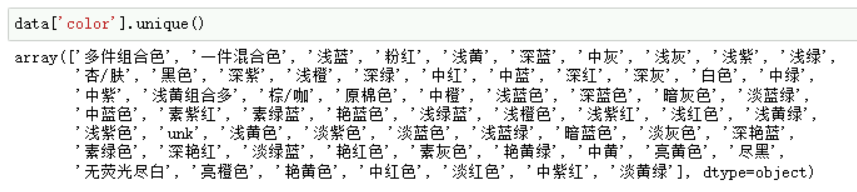


图 22 内裤品类陈列分区与销量箱型图

* 颜色和图案（color\_and\_pattern/color）

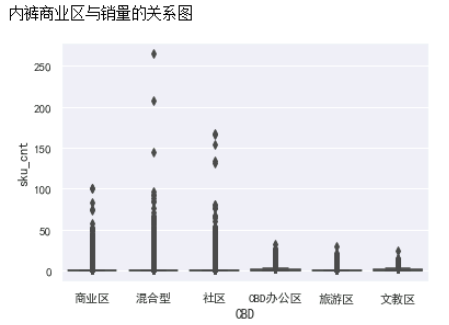




由于该特征下的分类过多，以及后面还会继续拓展新的款色图案，因此不作为特征。

* 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。



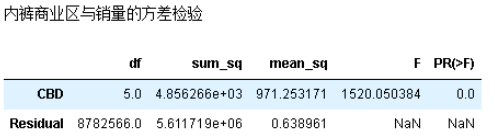
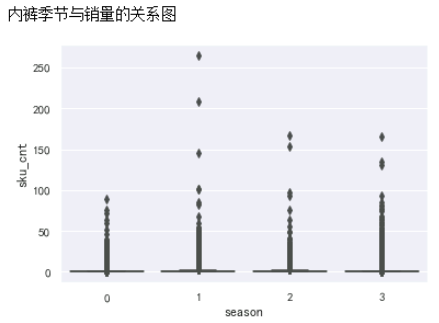


图 23 内裤商业区与销量关系分析

* 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。



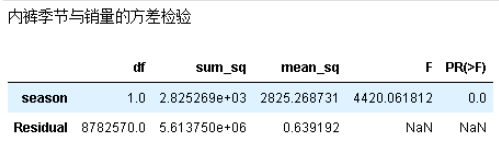


图 24 内裤季节与销量关系分析

* 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。



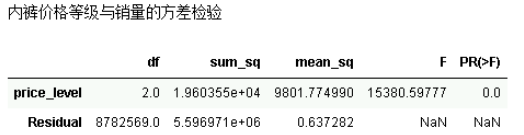
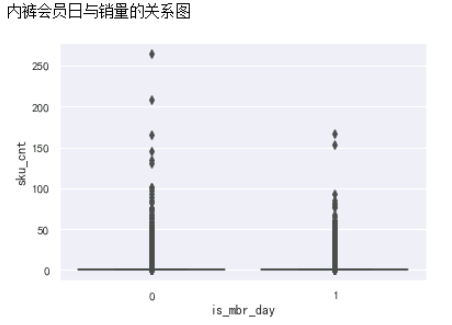


图 25 内裤价格等级和销量之间的关系图

* 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。



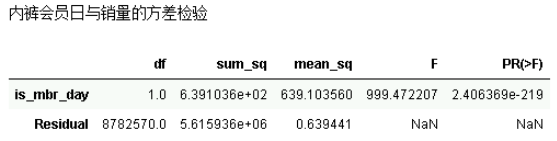


图 26 内裤会员日与销量之间的关系图

##### 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。

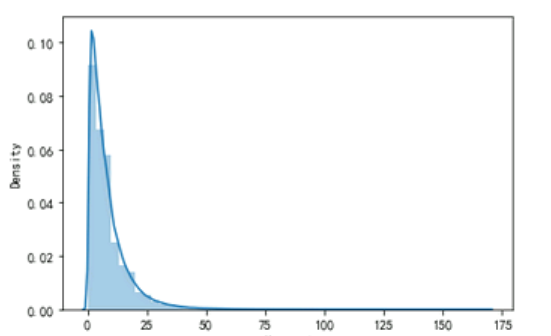


图 27 内裤品类SKU在各门店下历史数据量

在时间点（天/周/半月等）下，隶属于类型的门店下产品（SKU）的特征向量为，包括物料描述、温度、季节、折扣、促销等。其中表示类型的门店下包含的所有内裤类产品（SKU）。则为该产品在时间点下的真实需求量。

当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：

其中，为该产品在时间点下的预测需求量；表示产品于时间时在门店下的第个特征；表示对特征在时间对销售的影响进行建模的系数函数；表示产品对产品的影响权重；

当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：

其中，表示对产品自身历史数据建模的函数；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：

其中， MAPE值越小说明误差越小；

需求预测精度：

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

针对补货需求预测的精度为，ACP的值。

**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。

（1）分类建模的维度——c的选择

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** |

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。

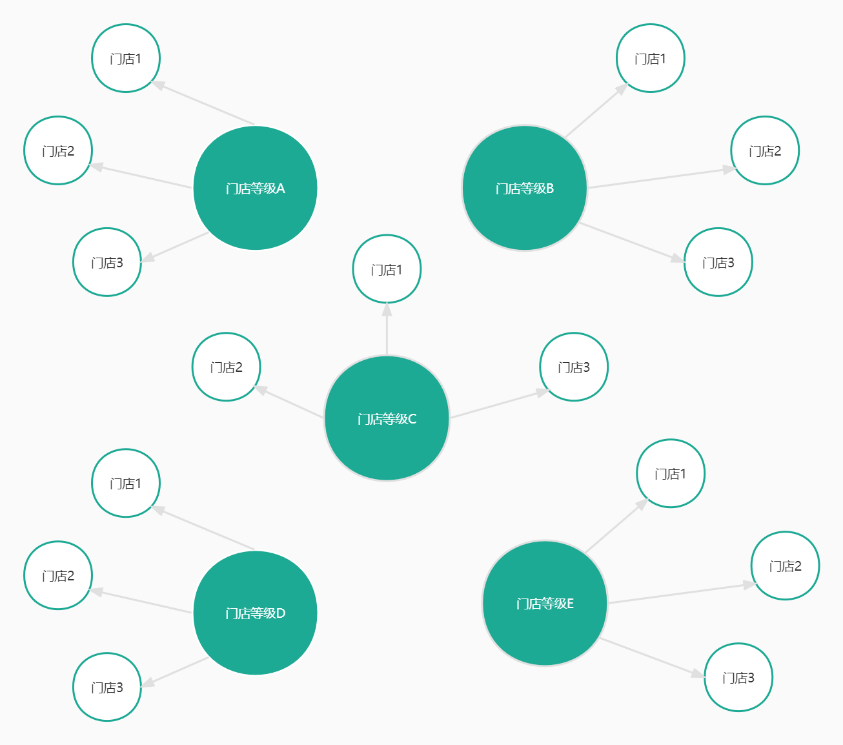


图 28 内裤品类预测模型设计

1. 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。

1. 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 |
| **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。

1. 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\_purchas**e）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

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

保留的字段为：

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

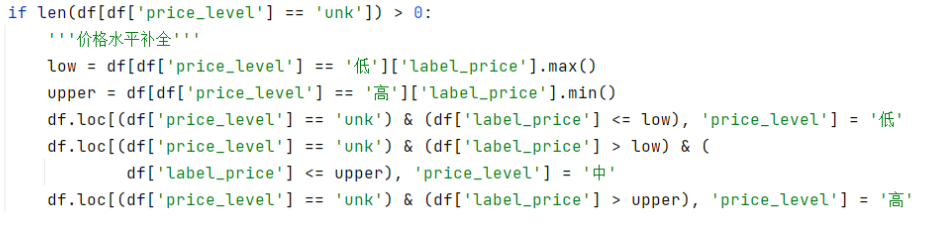
表 14 婴儿连体服字段缺失率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |  |  |  |

数据中，存在缺失的列以及对应的缺失率为：

* 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：



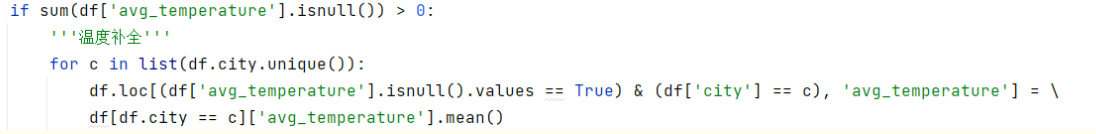
* prom\_level\_new的补全

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：



* avg\_temperature的补全

使用门店所在城市对应的年平均温度来补全：



* partition的补全

使用D区陈列来补全



#### 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。**

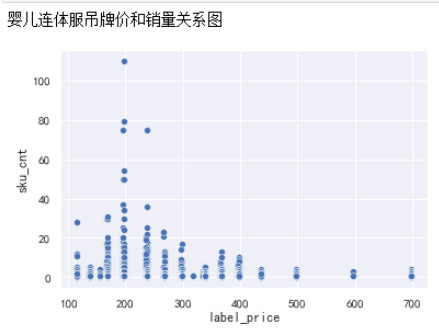


图 29 婴儿连体服吊牌价分析

**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。**

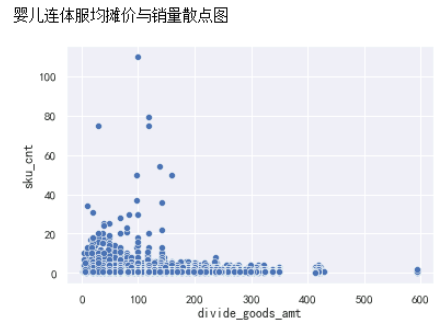
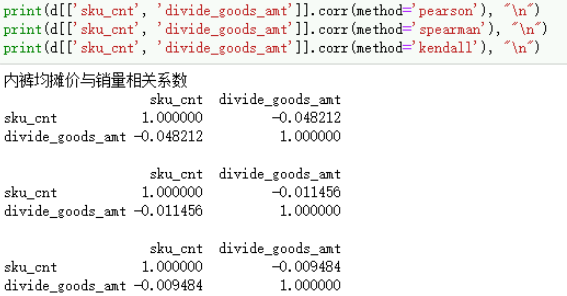
 

图 30 基础棉柔巾到手价与销量之间的关系

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。**

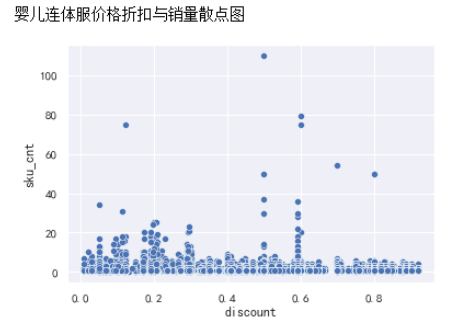




图 31 婴儿连体服价格折扣分析

* 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。**

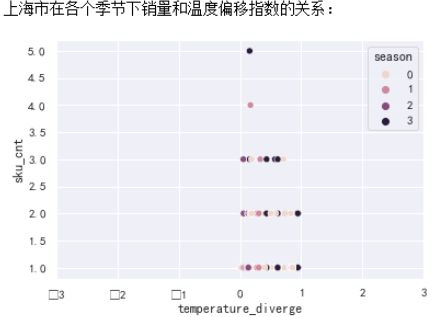
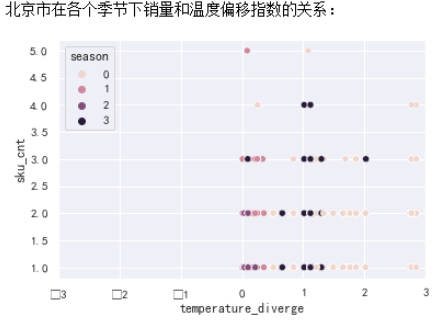
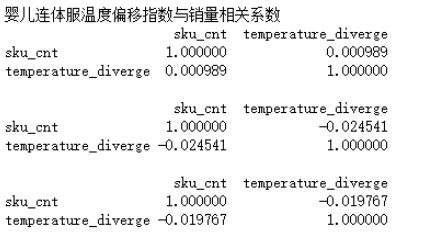
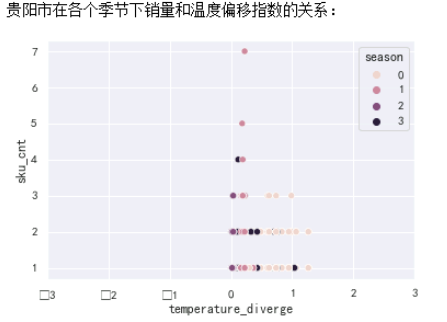
    

图 32 婴儿连体服温度偏移指数和销量之间的关系

* 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。

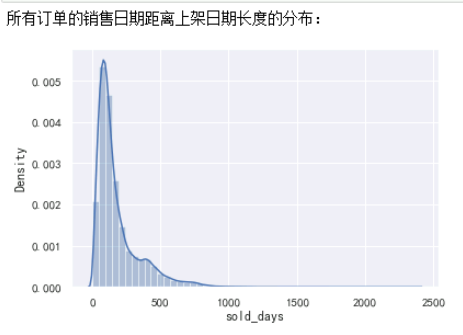
 

图 33 婴儿连体服生命周期分析

* 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。

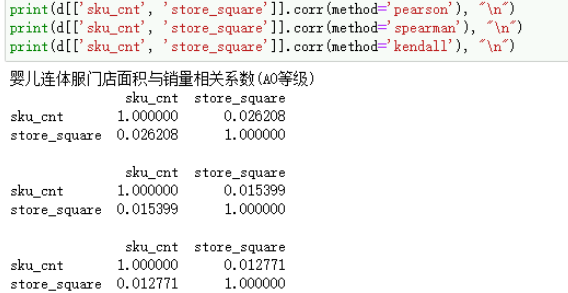
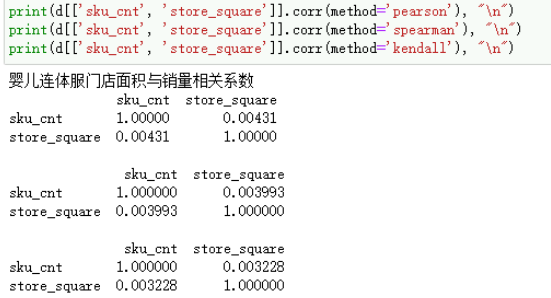
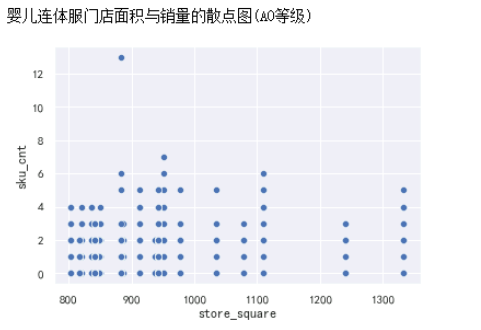
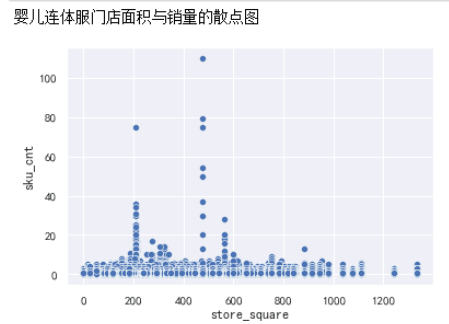


图 34 婴儿连体服门店面积与店铺等级分析

* 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。

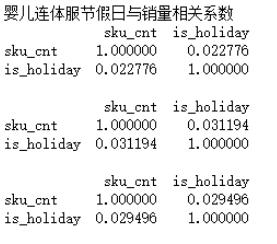
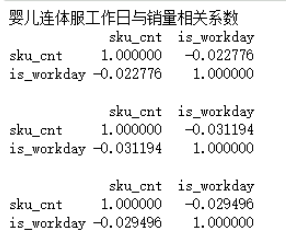
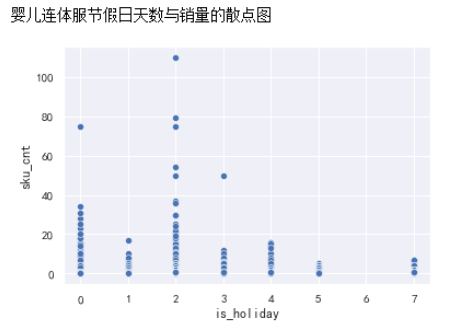
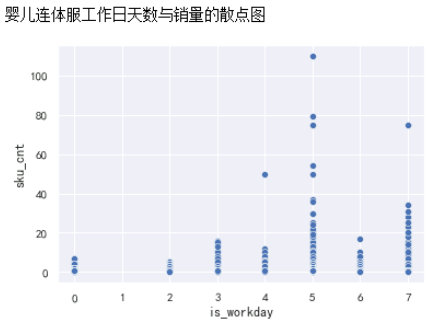


图 35 婴儿连体服节假日工作日指标分析

（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。

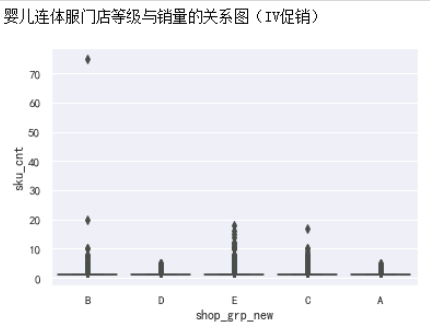
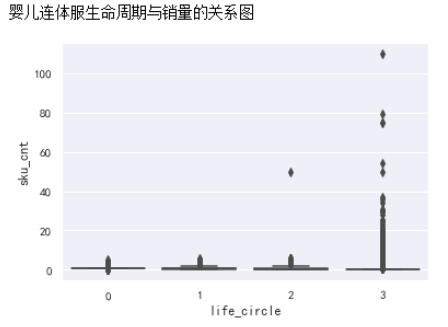
  

图 36 婴儿连体服门店等级指标分析

* 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。



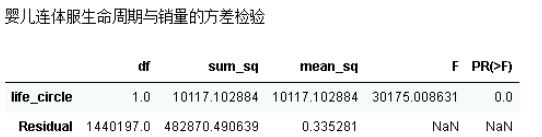
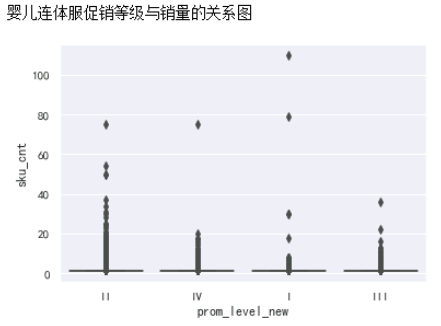


图 37 婴儿连体服生命周期分析

* 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。**



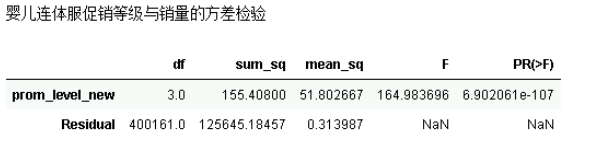
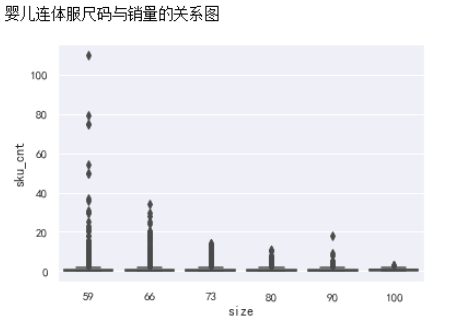


图 38 婴儿连体服促销等级分析

* 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。



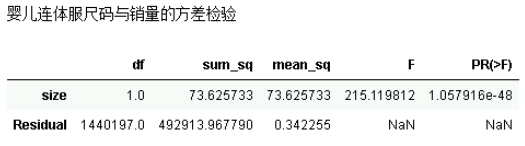
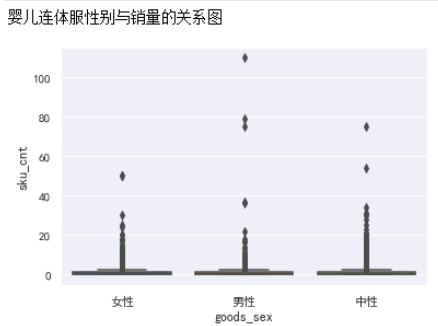


图 39 婴儿连体服尺码指标分析

* 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。



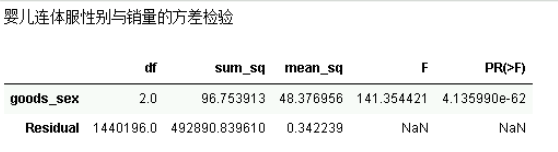
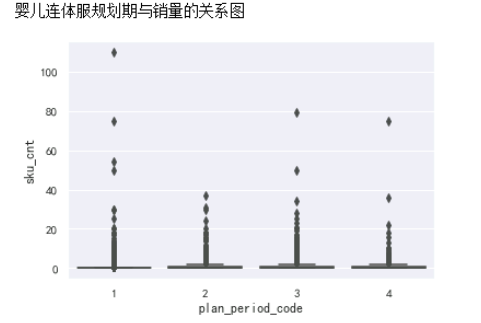


图 40 婴儿连体服性别指标分析

* 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。



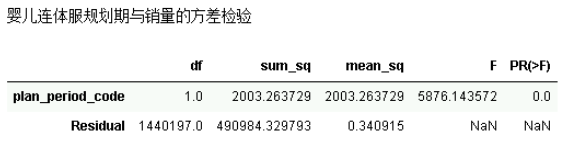
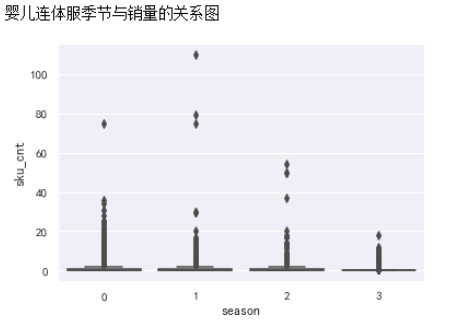


图 41 婴儿连体服规划期分析

* 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。



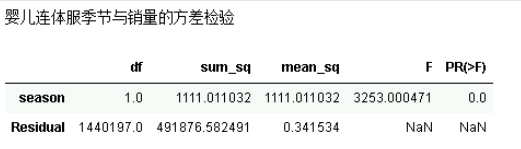
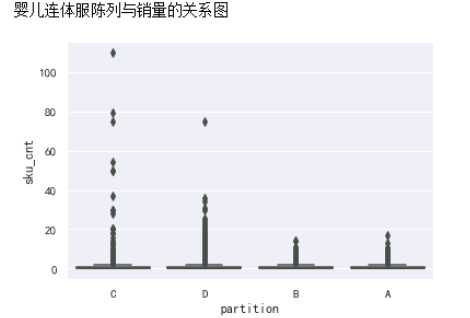


图 42 婴儿连体服季节分析

* 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。



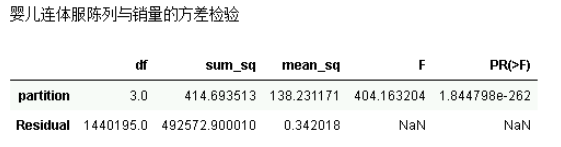
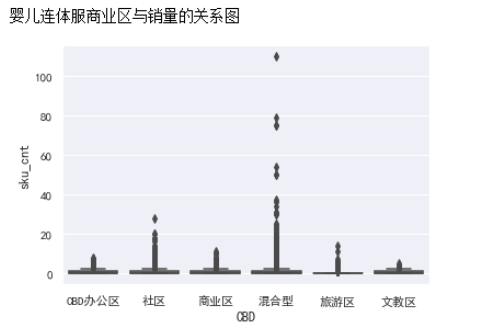


图 43 婴儿连体服陈列区域分析

* 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。



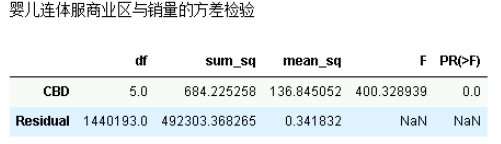


图 44 婴儿连体服商业区分析

* 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。

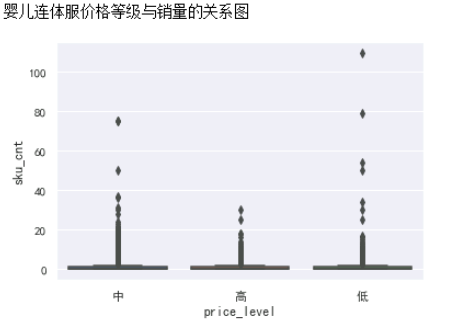
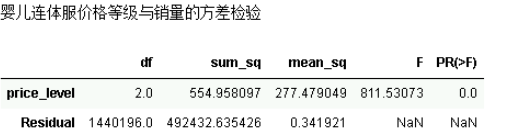
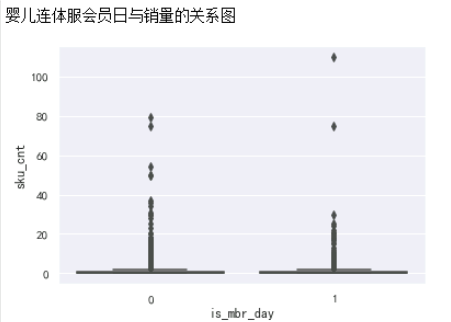
 

图 45 婴儿连体服价格等级分析

* 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。



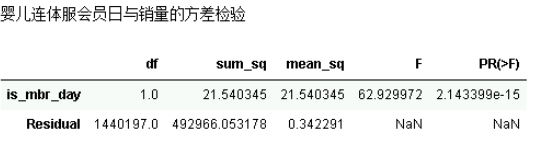


图 46 婴儿连体服是否会员日分析

##### 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：

其中，指指标值的均值，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（天/周/半月等）下，隶属于类型的门店下产品（SKU）的特征向量为，包括物料描述、温度、季节、折扣、促销等。其中表示类型的门店下包含的所有婴儿连体服类产品（SKU）。则为该产品在时间点下的真实需求量。

当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：

其中，为该产品在时间点下的预测需求量；表示产品于时间时在门店下的第个特征；表示对特征在时间对销售的影响进行建模的系数函数；表示产品对产品的影响权重；

当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：

其中，表示对产品自身历史数据建模的函数；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：

其中， MAPE值越小说明误差越小；

需求预测精度：

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

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。

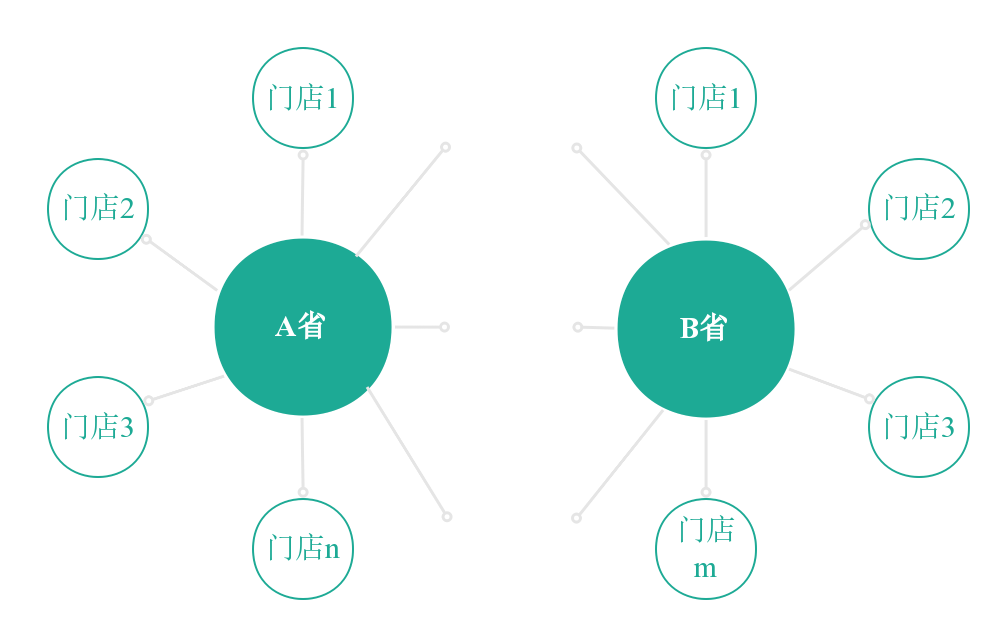


图 48 婴儿连体服需求预测模型设计

1. 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。

1. 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。

1. 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**；

## 预测结果总结

### 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。

### 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) |  |  |

### 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

### 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；

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。

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 |

### 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>

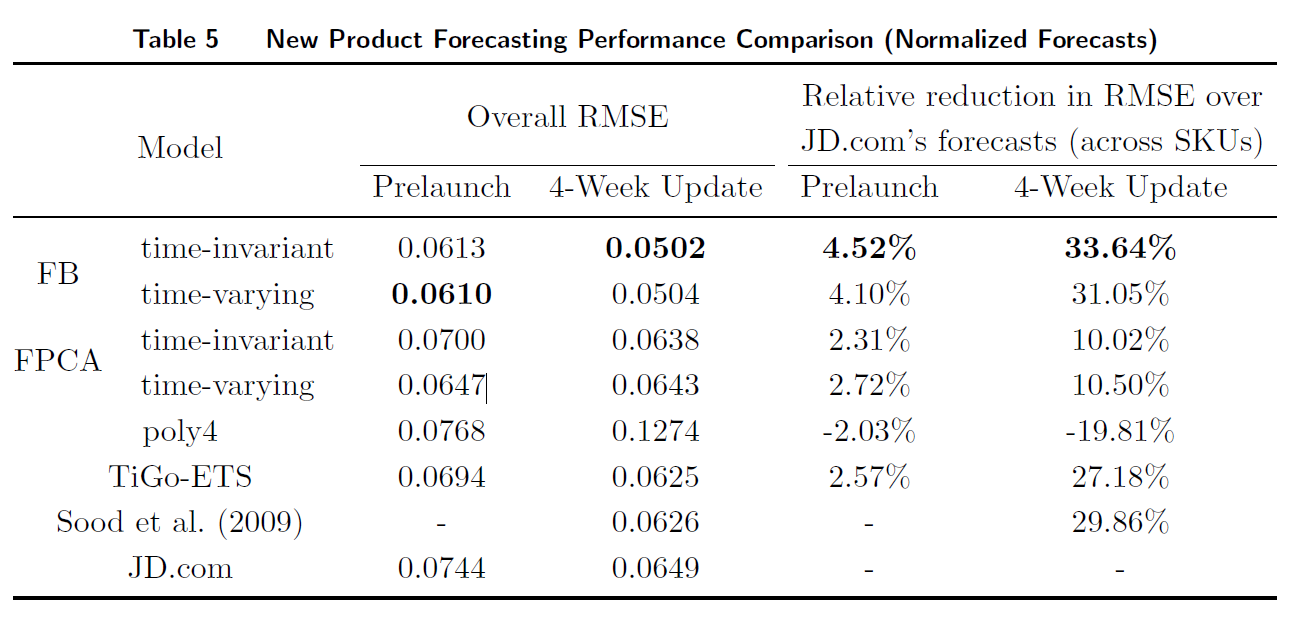
* 美国有个叫ToolsGroup的公司做过调研，说消费品的预测准确度在85%上下，零售业跟消费品差不多，而工业品就低得多，只有70%不到。根据Gartner的调查，消费品公司的预测准确度在50%到60%之间。这一数字明显低于ToolsGroup的统计。不过Gartner的统计方法相当严苛，是基于SKU和库位层面，统计方法是绝对差的百分比（MAPE）。

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

### 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。

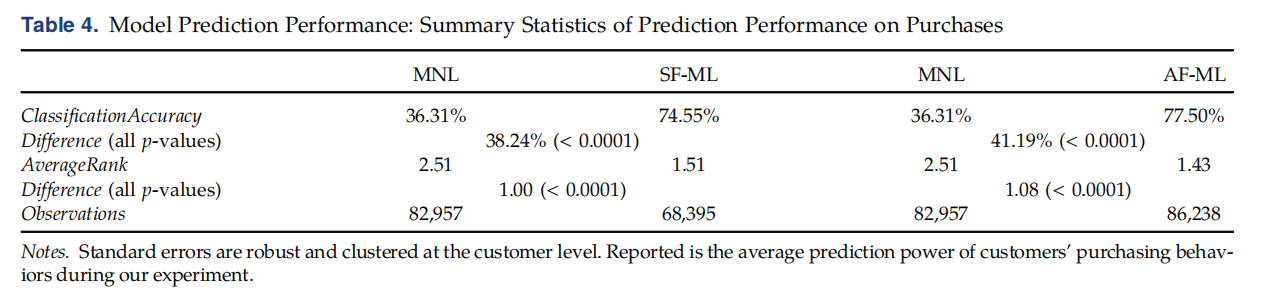


链接：<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4014586>

* 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%。



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