

工程创新

课程报告

**学 院 经济与金融学院**

**专业班级 金融学汇丰班**

**课程名称 机器学习**

**学生姓名 吴泓栎 王馨怡**

**孙智敏 杨子涵**

日期：2021年7月2日

目录

[一、概述 2](#_Toc76075944)

[二、数据来源 2](#_Toc76075945)

[三、数据预处理 5](#_Toc76075946)

[1.数据整合 5](#_Toc76075947)

[2.数据处理与转换 6](#_Toc76075948)

[四、特征工程构建 7](#_Toc76075949)

[1.特征构建 7](#_Toc76075950)

[2.特征提取 8](#_Toc76075951)

[3.特征选择 9](#_Toc76075952)

[五、模型构建 10](#_Toc76075953)

[1.模型建立 10](#_Toc76075954)

[2.模型解释 11](#_Toc76075955)

**淘宝网电商卖家的信用评估**

# 一、概述

电商平台卖家的信用评估体系一直在持续不断地完善与发展之中，本研究结合因子分析、核主成分分析(KPCA)以及kprototype机器学习算法，以淘宝网服装业电商卖家为样本，建立对淘宝网电商卖家的信用评估方案。

工具：Python、SPSS、Excel

# 二、数据来源

1.**淘宝网**(<https://www.taobao.com/>)

* 在进入淘宝网搜索页后，输入“服装”以及筛选店铺为淘宝店铺，排序依据淘宝平台信用评分从高到低返回结果。结果显示有100页，每页24家店铺。利用Python爬虫技术爬取每页的前三家店铺的部分基础信息，共获取300家店铺信息。获取部分如下：

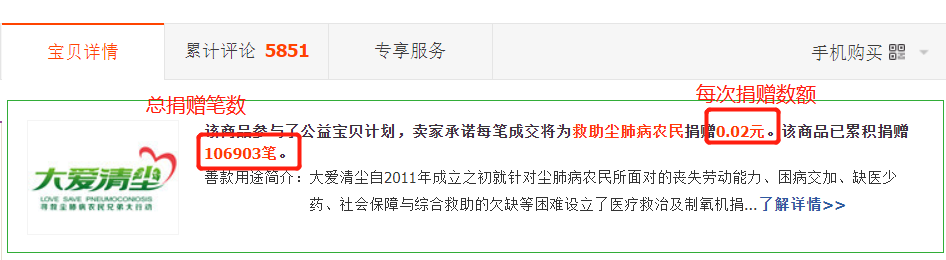


* 对每家店铺，获取进入店铺链接后的部分信息，包括商家基础信息以及商品部分数据，对应每家店铺一张表，爬取部分如下：





* 同时对商品承诺中含有公益宝贝计划的商品，获取对应信息：



**2.店查查**(<https://www.dianchacha.com/>)

* 根据店铺id进行网址拼接，爬取店铺的相关信息：

（1）基础信息



（2）店铺7天销量



（3）综合搜索情况



（4）宝贝列表



（5）月销售预测



**3.淘数据**(<https://www.taosj.com/>)

获取店铺部分销售数据：拼接网址，爬虫实现



# 三、数据预处理

1.数据整合：

（1）.淘宝网店铺基础信息表：

提取好评率的小数，去除网址

（2）.店铺宝贝列表：



计算月销量均值、评价数中位数、收藏数中位数、价格均值作为店铺的宝贝信息

（3）.店铺热度和曝光度表：



宝贝曝光度和热度取均值作为店铺曝光度和热度

（4）.店铺月销预测



转换数据类型

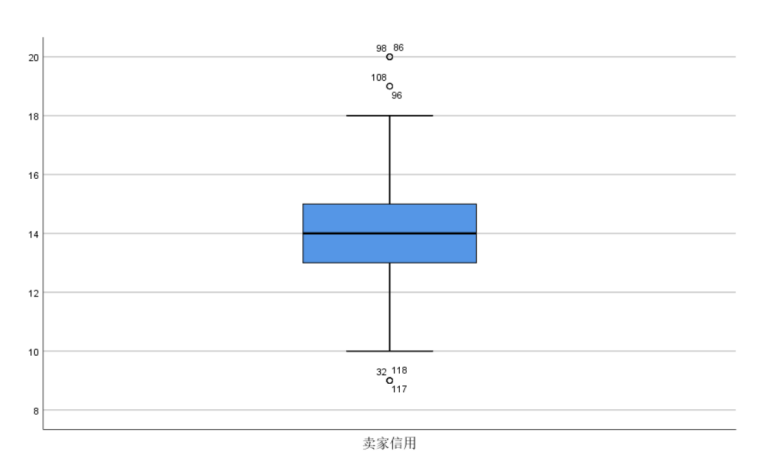
## 2.数据处理与转换

（1）**缺失值**处理

对于**缺失值过多的店铺**选择从数据表中删除；少部分空缺属于爬虫无法爬取的值，手动填充该缺失值；部分无法查询的值，选择通过前后值或中位值填充。

（2）**异常值**检验和处理

采取箱线图逐个对数值型指标进行异常值检验，以卖家信用为例：



存在异常高和异常低的离群点，经过仔细的数据对比，源数据中的真实值就是如此，所以选择保留该数据。

（3）指标转换

将指标开店时间转化成距今的天数，同时删除方差小于设定阈值的数值型指标和频率小于阈值的分类型指标。

整合所有数据表后，得到初始数据表如下：



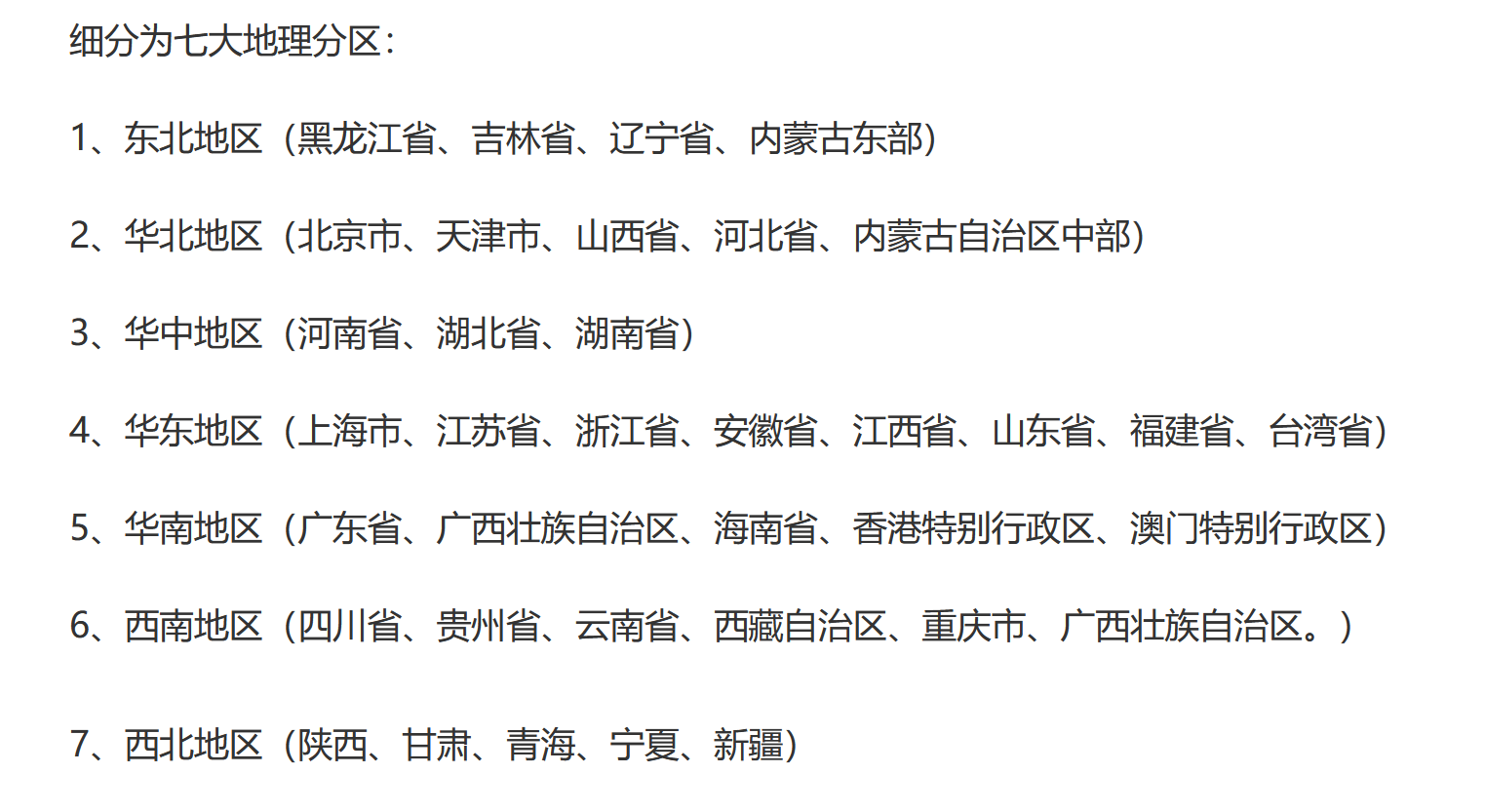
最终保留228条商家数据

# 四、特征工程构建

1.特征构建：

（1）所在地区

地区按照下图对应的地理关系进行编码：



（2）引入哑变量

对于部分分类变量（是否金牌卖家、承诺、男装或女装、支付手段）拆分为多个哑变量，取值0、1。

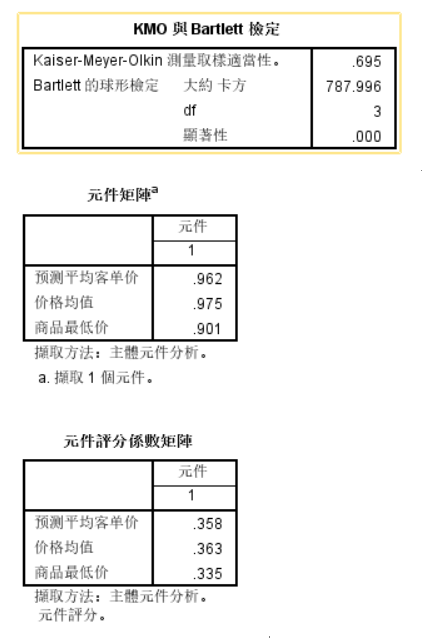
## 2.特征提取

综合后得到43个特征变量，首先利用因子之间的相关性，对各个数值型因子进行聚类划分，将两两之间相关度大于0.5的划分到同一类，根据KMO和Bartlett球体检验对此类别下指标数据是否适合做因子分析进行判断，对于KMO的值大于0.6的且Bartlett球体检验显著性P值小于0.05的指标大类选择保留主因子载荷量最大的那一项。

分类结果：

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 指标 | 类别 | 指标 |
| 第一类 | 交易成功比率 | 第六类 | 预测平均客单价 |
| 第二类 | 销售环比 | 价格均值 |
| 第三类 | 平均月销量 | 商品最低价 |
| 第四类 | 距今天数 | 第七类 | 描述相符 |
| 第五类 | 好评率 | 服务态度 |
| 第八类 | 发货时间 | 物流服务 |
| 第九类 | 曝光度 | 第十一类 | 店铺收藏数 |
| 热度 | 月销量 |
| 第十类 | 保证金 | 预测月销售额 |
| 7天上新数 | 预测成交笔数 |
| 第十二类 | 卖家信用 | 预测平均日销售额 |
| 宝贝数 | 收藏数中位数 |
| 第十三类 | 宝贝公益计划 | 评价数中位数 |
| 第十四类 | 30天销售额 |  |  |

KMO与球型检验结果，以第六类为例：



最终提取的主要数值型因子如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 指标 | 类别 | 指标 |
| 第一类 | 交易成功比率 | 第六类 | 价格均值 |
| 第二类 | 销售环比 | 第七类 | 描述相符 |
| 第三类 | 平均月销量 | 第八类 | 发货时间 |
| 第四类 | 距今天数 | 第九类 | 曝光度 |
| 第五类 | 好评率 | 热度 |
| 第十一类 | 预测成交笔数 | 第十类 | 保证金 |
| 第十二类 | 宝贝数 | 7天上新数 |
| 卖家信用 | 第十三类 | 宝贝公益计划 |
|  |  | 第十四类 | 30天销售额 |

## 3.特征选择

KPCA方法，

**基于核函数的主成分分析**：

（1）方法的基本思想是：对样本进行非线性变换，在变换空间进行主成分分析来实现在原空间的非线性主成分分析

（2）算法步骤：

① 通过核函数计算矩阵，其元素为K_{ij}=k(x_i,x_j)，其中x_i和x_j为原空间的样本，k(\cdot,\cdot)是核函数。

② 计算K的特征值，并从大到小进行排列。找出由特征值对应的特征向量\alpha^l (表示第l个特征向量)，并对进行归一化(||\alpha^l||=1)。

③ 原始样本在第l个非主成分下的坐标为：

Z^l(x)=\sum^n_{i=1}\alpha_i^lk(x_i,x)

这里的x_i是指第i个样本，\alpha^l的维度与样本数相同。如果选择m个非线性主成分(即计算K的前m大个特征值及相应的特征向量)，则样本x在前m个非线性主成分上的坐标就构成了样本在新空间中的表示[Z^1(x),Z^2(x),...,Z^m(x)]^T

**选用原因**：

初步采用PCA进行降维之后，发现数据并没有较好地分离，考虑到KPCA更适合线性不可分的数据，故采用KPCA。同时，PCA分析后，前两个因子的累计方差贡献率已达到99%，故选择降到2维。

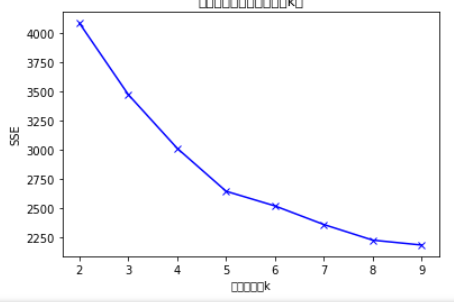
# 五、模型构建

## 1.模型建立

K-means聚类算法

（1）k值的确定：

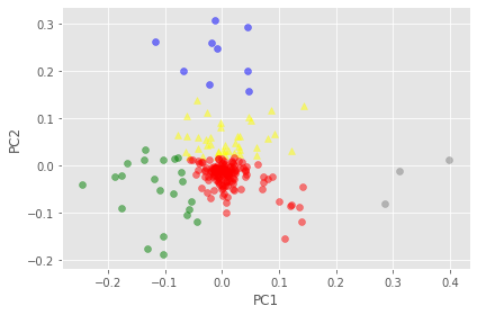
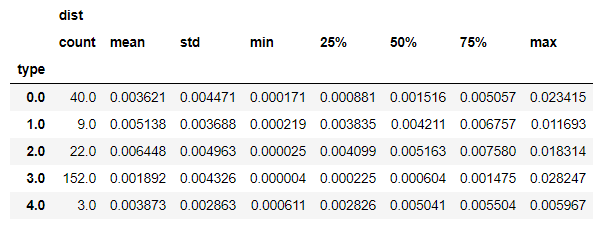
采用肘部法则对k值进行确定，在python中画出不同k值下，整体SSE的变化情况：



可以看到，在k=5时，效果较好，确定k=5

（2）聚类

将降维的数据放入模型中，使k=5，输出聚类结果如下：

可看出，数据较好地相互分离成5个类别。

## 2.模型解释

将各个类别进行筛选，得到对应的描述性统计