**《机器学习》 课 程 论 文**

**(2020-2021 学年第2 学期)**

**信用评分卡模型**

**学生姓名：朱飞龙 王志荣 李采妍 熊尹芝 林铭璐**

**提交日期：2021 年 7 月 01 日 学生签名：**

|  |  |  |  |
| --- | --- | --- | --- |
| **学 号** |  | **座位编号** |  |
| **学 院** | **经济与金融学院** | **专业班级** | **18级金融学（汇丰精英班）** |
| **课程名称** | **机器学习** | **任课教师** | **贺建风** |
| **教师评语：** | | | |
| **本论文成绩评定：** **分** | | | |

**目录**

[**一、研究背景和国内已知研究成果 1**](#_Toc31895)

[1.1研究背景和意义 1](#_Toc30161)

[1.2国内研究现状 3](#_Toc5716)

[1.3常用评估并预测店铺信用的计数 3](#_Toc3174)

[1.3.1 店铺信用概念 3](#_Toc28366)

[1.3.2 信用评价方法 4](#_Toc22387)

[1.3.3目前常用店铺信用预测模型及问题 5](#_Toc5341)

[1.4论文创新点 6](#_Toc16836)

[（1） 店铺信息与店铺内商品信息结合 6](#_Toc27152)

[（2） 多种特征衍生方式 6](#_Toc2263)

[（3）将评论情感得分作为特征衍生的一部分 6](#_Toc14190)

[**二、数据获取 7**](#_Toc21779)

[2.1爬虫概述 7](#_Toc24918)

[2.2数据来源:爬虫 7](#_Toc10404)

[**三、数据清洗 9**](#_Toc31590)

[3.1缺失值处理 9](#_Toc587)

[3.1.1看店宝中电器店铺数据 9](#_Toc17548)

[3.1.2看店宝中食品店铺数据 10](#_Toc15271)

[3.1.3天猫中电器店铺数据 11](#_Toc5296)

[3.1.4天猫中食品店铺数据 13](#_Toc9915)

[3.2数据格式清洗 14](#_Toc20630)

[3.2.1看店宝中电器店铺数据 14](#_Toc19652)

[3.2.2看店宝中食品店铺数据 14](#_Toc24040)

[3.2.3天猫中电器品店铺数据 15](#_Toc156)

[3.2.4天猫中食品店铺数据 16](#_Toc8337)

[3.3异常值处理 16](#_Toc18739)

[3.3.1看店宝中电器店铺数据 16](#_Toc5056)

[3.3.2看店宝中食品店铺数据 18](#_Toc3308)

[3.3.3天猫中电器店铺数据 20](#_Toc25691)

[3.3.4天猫中食品店铺数据 20](#_Toc27881)

[**四、特征工程 21**](#_Toc24483)

[4.1卖家服务状况 21](#_Toc30586)

[4.1.1卖家服务专业性（内容质量） 21](#_Toc24661)

[4.1.2卖家服务及时性 22](#_Toc21331)

[4.1.3卖家服务一致性 22](#_Toc30104)

[4.1.4售后服务 22](#_Toc20797)

[4.1.5物流服务 23](#_Toc6759)

[4.1.6顾客满意度 23](#_Toc74)

[4.2历史交易状况 23](#_Toc18792)

[4.2.1销售历史情况 24](#_Toc233)

[4.2.2销售稳定程度 24](#_Toc22597)

[4.2.3评价真实程度 24](#_Toc30020)

[4.3待交易商品特征 25](#_Toc7939)

[4.3.1待交易商品价格 25](#_Toc8644)

[4.3.2待交易商品种类/品类 26](#_Toc21233)

[4.3.3待交易商品人气 26](#_Toc9927)

[4.3.4待交易商品质量 27](#_Toc31264)

[4.3.5待交易商品基础承诺状况 27](#_Toc17590)

[**五、特征衍生 28**](#_Toc17282)

[5.1对X1特征的衍生 28](#_Toc26148)

[5.2对X2特征的衍生 30](#_Toc12580)

[5.3对X3特征的衍生 30](#_Toc18336)

[5.4根据店铺评论NLP进行特征衍生 32](#_Toc11555)

[**六、建模 36**](#_Toc4961)

[6.1数据降维 36](#_Toc10053)

[6.2 聚类结果 38](#_Toc14063)

[6.3信用风险评估结果 40](#_Toc2142)

[6.3.1主成分分析 40](#_Toc11127)

[6.3.2 信用风险评估 41](#_Toc29167)

[**七、未来建议和展望 42**](#_Toc28017)

[（1） 数据获取困难、数据缺乏权威且难以被量化 42](#_Toc13211)

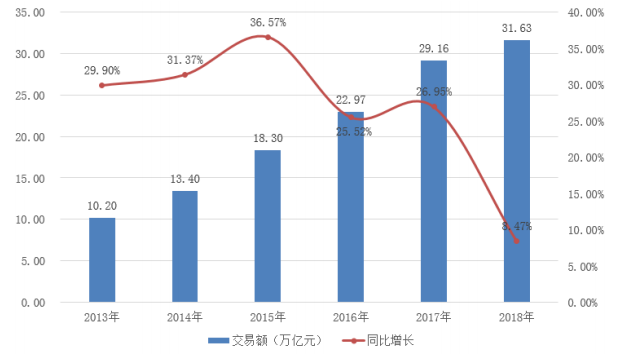
[（2）电商店铺相关数据储存于于中心节点上，数据可信性度略有折扣 42](#_Toc1726)

[（3）尚未建立成熟可靠的大数据电商信贷风控系统 43](#_Toc28006)

## 一、研究背景和国内已知研究成果

### 1.1研究背景和意义

1998年，中国首个互联网在线交易成功完成；1999年，中国建立了8848等B2C网站，标志着我国电子商务交易开始进入实用阶段；2001年，随着企业电子商务成为中国电子商务的新主体，中国的电子商务开始了20年的快速发展时期。2016年，电子商务交易额在我国国民生产总值中占接近于40%。根据中国工业研究院统计的数据，在 2017年，我国的电子商务交易额达到29.16万亿元，同比增长了11.7％；其中商品、服务类电商交易额占总交易额的74.86%，同比增长了24.0％；合约类电子商务交易额占总交易额的25.14%，同比降低了 28.7％。2018年我国电子商务交易额为31.63万亿元，同比增长了14.5%；其中合约类电子商务交易额占总交易额的3.23%，同比降低了51.3%。下图显示了我国电子商务市场变化的趋势图。



随着大数据和机器智能的快速发展，新一轮技术革命开创了电子商务的新 局面，全球化的发展对电子商务提出了新的要求。经济和社会结构的转型为电子商务开辟了新的空间。中国的电子商务规模将继续扩大，其结构将不断优化，活力将不断增加，进入一个新的发展阶段。整体而言，中国的电子商务将呈现出服务化、多元化、国际化和标准化的趋势。为了加快建立公正、公开、诚信的电子商务的市场秩序，我国通过创新监管的方法来形成共识，形成政策合力。为了保障电子商务有效健康的发展，企业建立了“反垄断联盟”等自发性组织，发展改革委员会等32个部门还成立了相互监督相互协调的跨部门工作组，不断加强内部管理，有效抑制了不良因素的滋生。

电子商务标准化进程的不断加快，增大了对市场秩序规范的难度，也使电子商务相关信贷的可获得性难度加大。信用作为市场经济的根蒂和生命线，影响着电子商务这种新兴的贸易活动，信用的好坏是电子商务未来发展的关键。电子商务和信用服务属于新兴起的社会领域，市场前景一片光明，从这两者关系来看，电子商务发展不能脱离信用体系的支持，信用体系的价值也能在电子商务这个社会领域充分体现，二者息息相关，不可分离。

多方研究表示，是否存在着对电商店铺信用的真实且可靠的评估，影响着社会普罗大众在进行互联网购物时的的合法权益和信贷行业针对电商贷款需求的放贷标准。

根据大数据分析报告显示，2018 年，当全网民都关注的“双十一”正式结束时，天猫“双十一”全球嘉年华营业额达到了 2135 亿元，再创新高。在“双十一”期间，电商通过发放商品优惠券这种低成本的诱惑，同时配合各种其它 营销方式引起了消费者大量购物的热潮。然而，在“双十一”狂欢节即电子商务推广促销之后，通常是消费者投诉纠纷的高峰期。从近几年的“双十一”消费记录来看，网络运营商欺骗消费者的现象时有发生，致使出现价格虚高、刷单、炒店铺信用和信息泄漏等各种问题。这些行为导致网上购物投诉创历史新高，同时也反映出了电子商务市场的竞争秩序并不和谐，这种情况已经引起了各界的广泛关注。

因为电商店铺的信用问题而导致的风险同样出现在针对电商的贷款领域。淘宝天猫等电商店铺创业者多属于小微企业，针对小微电商企业贷款频率快、贷款周期短和贷款需求量大等特征,部分电商企业及银行为其提供创新服务和新型信贷业务。然而小微企业由于自身实力低下、入门槛较低、融资成本高、融资渠道单一、融资效率低等因素的制约,小微电商企业发展虽然迅速但是后继无力。同时银行还面临着小微企业及其严重且不可控的信用风险。从目前小微电商网络贷款的途径来看,除银行界推出的网络联保信贷,其背后依赖的是具有雄厚资本和信贷经验的金融机构外,包含京东、阿里集团和苏宁在内的各类小微电商网络信贷模式,其目前提供的贷款均为纯信用贷款,无需提供任何抵押或担保。 因此一旦小微企业违约,以上各方并没有任何可减少损失的方式。 但这也意味着,微电商网络信贷的过程中,存在着大量的信用违约的情况。 如果是小微电商企业群体在贷款过程中,大量存在违约情况或者违约行为,京东、阿里集团和苏宁在内的各类小微电商网络信贷模式就有极大的可能停止向小微企业贷款,那么小微企业的贷款途径就会减少,甚至更为单一。

综上，建立切实可行、有效可靠的电商店铺信用评价体系对于消费者的权益保障和促进针对电商店铺的小小微企业贷款有正面作用，创新性地提出信用评估模型成为刚需。

### 1.2国内研究现状

根据《网络零售市场信用机制优化研究》我国网络零售市场信用等级计算采用的是一种累加的信用模型，鉴于我国网络零售市场好评率普遍较高的情况，这种信用评价机制主要反映了一个卖家经营的资历，卖家经营的时间越久，其信用等级越高，这样一来，新卖家的信用等级大多不如资历老的卖家，这对些后进入市场的卖家而言是极为不利的。这样的信用评价机制既不利于公平竞争，也不能反映卖家真实的信用状况。有不少网购过的买家表示，自己曾经在一些经营资历老、信用等级高的店铺买到过假货。

根据《浅论电子商务交易中虚假信用评价的法律规制》一文所述，近年来虚假信用评价问题越来越严重。对于消费者来说，这些虚假信用评价加大了他们购物决策的风险， 害了消费者的知情权和选择权，违背了民商事交易中的诚实守 信原则和公平原则；对于其他遵循交易规则的商家来说，虚假 用评价行为，会导致一些直接处于竞争关系的合法经营商家损失了商业机会，而实际造成的损失却又难以估算；对于平台方来说，这些虚假信用评价的行为不仅是对平台交易规则的无视和践踏，更玷污了平台辛苦经营的良好信用环境，长此以往，必然会使得平台长期积累的信誉毁于一旦。

根据《基于交易三阶段的电子商务卖家信用评价模型研究》一文，在电子商务单笔交易中，随着交易流程的推进，卖家履约范围会发生变化，信用水平也会相应波动。同时，消费者对于卖家的履约行为的感知也存在动态性。现有电子商务信用评价模型的信用值计量方法仅设置3个总体评价等级，无法区分消费者在不同阶段不同感知下的信用评价结果。

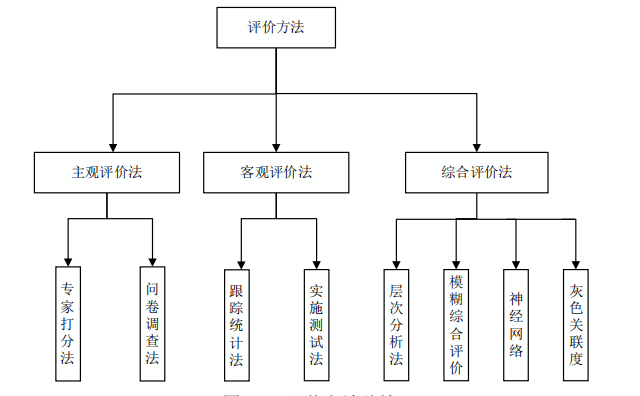
### 1.3常用评估并预测店铺信用的计数

#### 1.3.1 店铺信用概念

电子商务信用是指电子商务大环境下的信用，即每个交易实体在电子商务环境中遵守市场规章制度水平不确定程度。很多因素都能影响电子商务信用，例如虚拟市场的规章制度、买卖双方的品德与文化、买卖双方的信用能力等，因此构建电子商务信用评价是非常必要的。

#### 1.3.2 信用评价方法

目前关于电子商务店铺信用评价模型的研究还比较匮乏，本文通过归纳总结电子商务评价的方法，提出了一个适用于电商店铺信用评价模型的方法。电子商务评价方法可分为主观评价法、客观评价法和综合评价法。



**（1）主观评价法**

主观评价法是基于评价者主观意愿的一种评价方法，该方法操作简单、方便实施，适用于不需要定性的调查研究。但是该方法受评价者主观意愿的影响较 大，没有统一的标准去衡量，通常不能获得客观、高准确率的结果，因此常不 单独使用。

**（2）客观评价法**

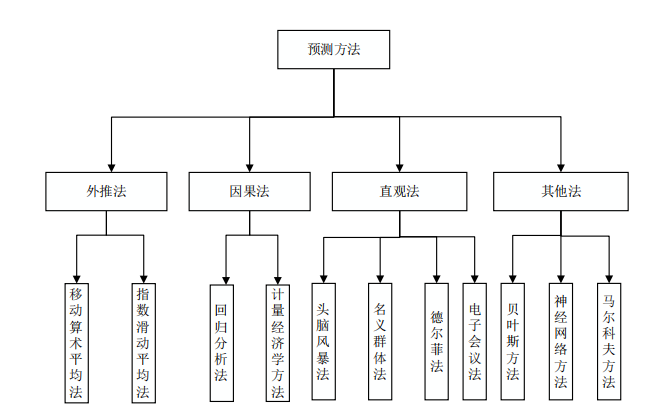
客观评价法是基于评价对象的实际情况的一种评价方法，该方法可以获得较客观的结果。电子商务信用评价中常用的客观评价方法包括跟踪统计法和实时测试法。

**（3）综合评价法**

综合评价方法是结合主观评价法和客观评价法的一种综合评价方法。常用的综合评价方法有层次分析法、神经网络、模糊综合评价法等。该方法综合考虑了主观和客观双方面的影响，通常可以获得客观、高准确率的结果。

#### 1.3.3目前常用店铺信用预测模型及问题

预测是通过某些因素和迹象以预测未来的发展和变化。在各个领域都要用大数据来预测，其方法也有很多，主要预测方法分为四大类。



**（1）外推法**

外推法是指使用过去的数据来预测未来状态的方法。它的最大优点是只要获得有关以往的真实信息，就可以简单且容易的预测未来。因为它并未从因果角度分析过去与未来的关系，所以长时间的预测可靠性并不高，外推法广泛用于短期和近期预测。时间序列法是根据时间排列过去统计的数据，观察其发展趋势的一种方法，属于外推法，其最重要的特点是它数据不规则。其常用的方法是移动算术平均法和指数滑动平均法。此方法仅仅采用了历史数据进行简单的推测性计算，无法应对复杂的如周期性变化等预测需求

**（2）因果法**

因果法包括两种经典的研究定量方法：回归分析法和计量经济学法，两种方法都是为了研究变量存在的因果关系，然后预测未来的状况。然而它们只能预警电子商务信用风险存在的一些周期性风险，无法预警一些时变的风险，缺陷十分明显。

**（3）直观法**

直观法是指靠自身的经验和综合能力去预测未来的一种方法。它是一种常见的简单方法，例如头脑风暴法、德尔菲法等。 此方法主观性较强，不适用与需要准确预测的环境

**（4）其他预测方法**

预测方法的新进展促进了新学科的出现，即预测学，现代预测学在现有计量经济预测模型的深入研究以及新的计量经济预测模型的不断探索方面取得巨大的进步。关于线性预测模型，主要有贝叶斯方法；对于非线性问题，主要有神经网络方法、马尔科夫方法，非线性电子商务信用风险预警方法存在一些不足，如神经网络由于参数优化问题，收敛速度慢、风险预警精度低，因此电子商务信用风险预警有待进一步研究

### 1.4论文创新点

本文创新点主要体现在数据的获取和特征衍生方面，模型上则采用PCA方法进行数据降维，并使用K-means方法对降维后的数据进行聚类。

1. **店铺信息与店铺内商品信息结合**

以往的针对电商店铺的信贷分析多是采集针对店铺本身整体性的数据，本文在分析店铺经营状况时不仅获取了描述比率、物流比率等常见的公开店铺相关数据，还获取了店铺内每个商品的详细状况，包括其近30天销量、最大折扣、评论中标签等数据，不仅可以辅助鉴别店铺的最近经营能力，而且可以根据商品相关数据来鉴别有刷单行为存在的可能性的大小，有助于识别虚假繁荣，获取接近数据脱水后的真实情况。

1. **多种特征衍生方式**

在特征衍生中，我们将数据本身之间差距很小的属性进行分箱，将数量级较大的数据取对数，对某些数据取平方防止倒U效应的出现，并且找出一些影响较大的数据进行处理，包括平方立方等，用于扩大某些特征的影响力。

**（3）将评论情感得分作为特征衍生的一部分**

在此次针对店铺信息的评估中，我们添加了店铺情感得分，将字符型的数据集变为一个可以通过一个分布在0到1之间的数字表达的情感倾向。我们小组组员爬取了300家店铺每家店铺最新的前2000条评论，共计60万条评论，并通过公开数据获取词典库，进行文本清洗和jieba分词，然后生成生成文档-词矩阵，下一步则利用SnowNLP包来计算每家店铺的情感得分，最后DBSCAN进行聚类分析，将以上店铺分成8个簇，构建出基于词频的一个综合评级。

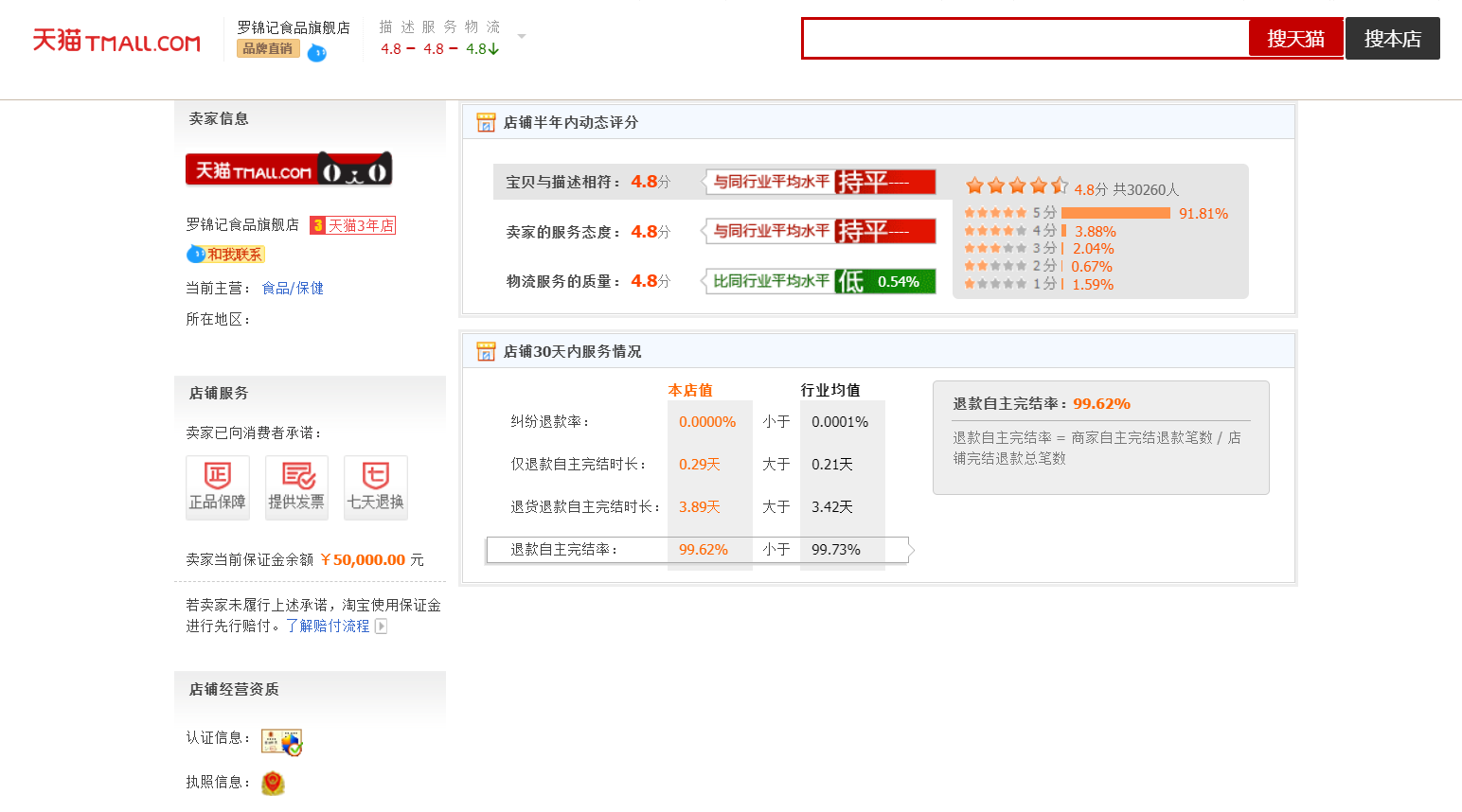
## 二、数据获取

### 2.1爬虫概述

网络爬虫是一种非常实用的技术，它可以通过提前定义好的规则自动从网页上采集文本信息，从网络上下载网页，并反映在搜索引擎中，因此它是搜索引擎的重要组成部分。最常用的网络爬虫有以下几种：常规 web 爬虫、聚焦web 爬虫、增量式 web 爬虫、深层 web 爬虫。

### 2.2数据来源:爬虫

我们的数据主要来自于天猫网页平台和第三方平台看店宝，利用python的selenium模拟器从天猫爬取数据。首先加载出天猫的登录页面，然后用手机扫码登录天猫，获取到cookies信息，然后用这个cookies信息访问天猫店铺页面。



从上图中可以获取到描述，物流，服务，纠纷退款率等一系列的信息，所以我们从天猫获取信息的主要来源就是这个页面。因为天猫的反爬机制比较强，所以我们一次致爬取9家店铺的信息，并且每家店铺之间间隔4秒。但是在这个过程中偶尔还是会被识别出来，要求做滑块验证，从而使程序退出正常运行状态。原因可能是我们并没有对webdriver.Chrome()的浏览器内部设置做改变，导致平台是可以检测出浏览器是selenium在驱动而不是人为的。

看店宝平台有一个店铺分析的板块，里面可以查到对天猫平台店铺的一些统计信息，同时还有部分评论信息。



如上图所示，我们可以输入要查询的店铺名称，然后就可以看到与该店铺相关的一些信息。

如果注册成为账号会员可以查询所有的店铺，否则只能试用，且一个账号每天只能查询10次，总共能查询7天。我们需要近300家店铺，所以我们总共注册了10个左右的账号，然后利用selenium模拟人的行为去做查询，将最终得到的数据用python整合起来。

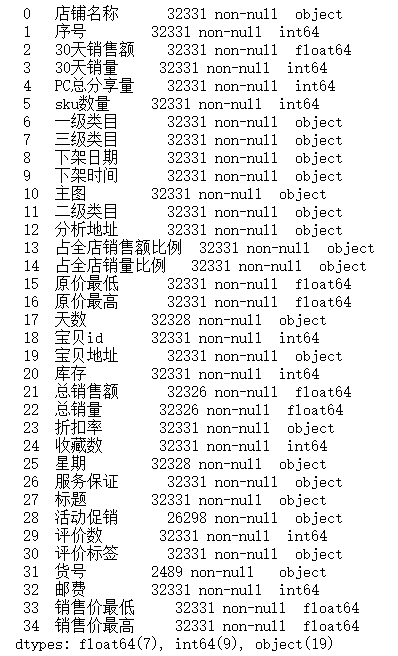
## 三、数据清洗

原始数据一共有四个表格，分别是来自看店宝的电器和食品电铺数据、来自天猫的电器和食品店铺数据。

### 3.1缺失值处理

#### 3.1.1看店宝中电器店铺数据

**（1）查看数据结构**



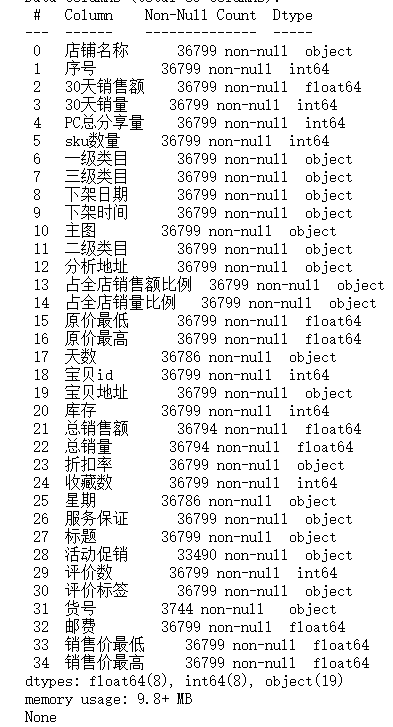
从结果上看，一共有32331个样本，其中有缺失值的数据类型：天数，总销售额，总销量，星期，活动促销，货号。

**（2）处理缺失值**

|  |  |
| --- | --- |
| 属性 | 处理方式 |
| 天数 | 用00:00:00填充。天数指的是商品存在的时间，基于这个含义如果商品的存在时间缺失则可以粗略认为此值为0。 |
| 总销售额 | 用0填充。由于这个数据无法从其他数据中观察出来，商品的销售额是无法根据其他商品销售额来粗略估计的，所以缺失的数据用0填充。 |
| 总销量 | 用0填充。由于这个数据无法从其他数据中观察出来，商品的销售量是无法根据其他商品销售量来粗略估计的，所以缺失的数据用0填充。 |
| 星期 | 用一填充。统一认为没有星期的数据都是星期一。 |
| 活动促销 | 用无填充。活动促销没有数据可认为其没有活动促销，所以用无填充。 |
| 货号 | 用无填充。缺失的货号可认为该商品没有货号，所以用无填充。 |

#### 3.1.2看店宝中食品店铺数据

**（1）查看数据结构**



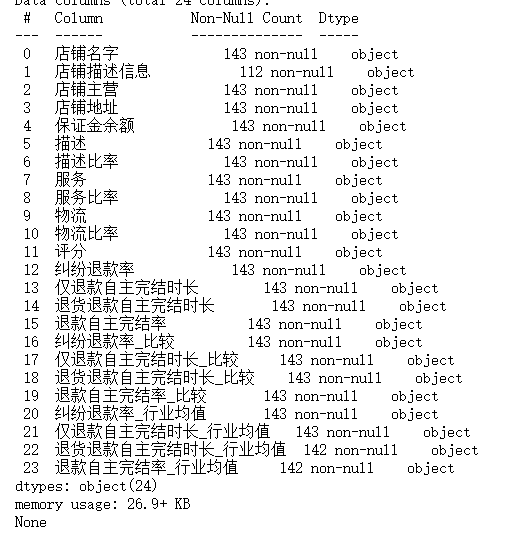
从结果上看，一共有36486个样本，其中有缺失值的数据类型：天数，总销售额，总销量，星期，活动促销，货号。

**（2）处理缺失值**

与看店宝中电器数据一致。

#### 3.1.3天猫中电器店铺数据

（1）**查看数据结构**



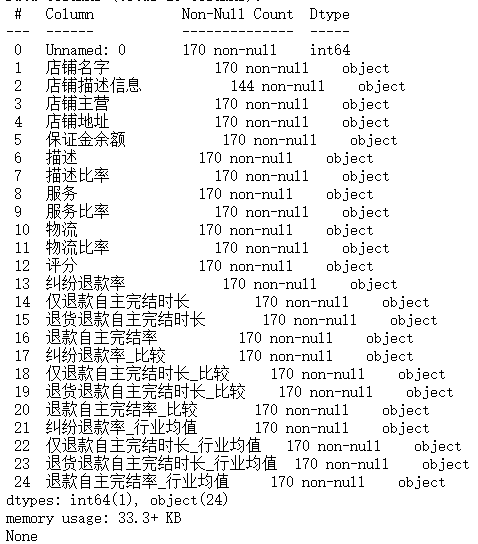
从结果上看，一共有143个样本，其中有缺失值的数据类型：店铺描述信息，退货退款自主完结时长\_行业均值，退款自主完结率\_行业均值。

1. **处理缺失值**

|  |  |
| --- | --- |
| 属性 | 处理方式 |
| 店铺描述信息 | 由于无法找到该数据的值，并且不适合拿其他店铺的数据衍生的来，所以只能用无填充。 |
| 退货退款自主完结时长\_行业均值 | 由表格数据看出只有一个店铺是没有这个值的，根据仅退款自主完结时长\_行业均值这个属性来看，该店铺和其上面的店铺值相等，所以可以粗略的用上一个店铺的退货退款自主完结时长\_行业均值来填充该店铺的缺失值。 |
| 退款自主完结率\_行业 | 与退货退款自主完结时长\_行业均值属性操作相同。 |

#### 3.1.4天猫中食品店铺数据

**（1）查看数据结构**



从结果上看，一共有170个样本，其中有缺失值的数据类型：店铺描述信息.

**（2）处理缺失值**

店铺描述信息：该数据是表示的是年份的信息，是一个非常重要的特征，所以不能简单的用无填充，应该用均值填充更加准确。

### 3.2数据格式清洗

#### 3.2.1看店宝中电器店铺数据

**（1）目标**

将所有可以变成数值特征的文本特征变成数值型特征。

**（2）需要转换的数据类型：**

|  |  |
| --- | --- |
| 属性 | 处理方式 |
| 占全店销售额比例 | 占全店销售额比例是字符串类型要转换为浮点数类型，它的原始数据格式是数字加%，要将其转化为没有百分号的浮点数。其中要用到正则表达式匹配%前的数字，将其取出后再除100，使数值大小保持一致。 |
| 占全店销量比例 | 与占全店销售额比例的处理方式一致。 |
| 天数 | 于天数有两种格式的数据（xx天xx时xx分xx秒和0:0:0），处理逻辑较为复杂，所以写了一个transDayCount函数进行处理，该函数运用正则表达式进行匹配，对于xx天xx时xx分xx秒格式的数据，由于该格式的数据可能不是固定的，可能会缺失单位（如没有分），所以可以用多个条件语句分别提取出天、时、分、秒的数字。对于0:0:0格式的数据可以用一个条件语句一个正则表达式则匹配提取的出时分秒的数据。最后函数的返回值以天为单位，返回一个浮点型的数据。 |
| 折扣率 | 该数据的格式也是数字加%，所以处理逻辑和占全店销售额比例一致。 |
| 星期 | 将汉字一、二、三等转换为整数1，2，3等。由于要用到多次判断所以写了一个transDayNum函数进行处理。 |
| 活动促销 | 将活动促销的数据转换为列表。由于活动促销有多个，并且用|隔开，所以在处理的时候直接调用字符串的函数split即可。 |

#### 3.2.2看店宝中食品店铺数据

**（1）目标**

将所有可以变成数值特征的文本特征变成数值型特征。

**（2）需要转换的数据类型**

占全店销售额比例,占全店销量比例，天数，折扣率，星期，服务保证。处理方式与电器数据一致。

#### 3.2.3天猫中电器品店铺数据

**（1）目标**

将所有可以变成数值特征的文本特征变成数值型特征。

**（2）需要转换的数据类型**

除了店铺描述信息之外的所有属性都需要。

|  |  |
| --- | --- |
| 属性 | 处理方式 |
| 店铺地址 | 将店铺地址规范化，只留下具体地址，把前面的无关字眼删除。 |
| 保证金余额 | 转换为浮点型。把￥和，都去除，剩下的数字转化为浮点型。 |
| 描述 | 转换成浮点型。把分字去除后剩下的数字变为浮点型。 |
| 服务 | 转换成浮点型。与描述处理方法一致。 |
| 物流 | 转换成浮点型。与描述处理方法一致 |
| 描述比率 | 由于数据中包含----的数值，其意思为与平均值持平，即应该为0%，所以先将这些值转化成0.00%的字符串，再将这些字符串转化为值相等的浮点型。 |
| 服务比率 | 与描述比率处理方法一致。 |
| 物流比率 | 与描述比率处理方法一致。 |
| 评分 | 转换成浮点型。与描述处理方法一致。 |
| 纠纷退款率 | 转换成浮点型。先把%号去除，然后将剩下的数字转化成浮点数后再除100，得到与原来等值的数。 |
| 仅退款自主完结时长 | 换成浮点型。与描述处理方法一致。 |
| 退货退款自主完结时长 | 换成浮点型。与描述处理方法一致 |
| 退款自主完结率 | 转换成浮点型。与纠纷退款率处理方法一致。 |
| 纠纷退款率\_行业均值 | 转换成浮点型。与纠纷退款率处理方法一致。 |
| 仅退款自主完结时长\_行业均值 | 换成浮点型。与描述处理方法一致。 |
| 退货退款自主完结时长\_行业均值 | 换成浮点型。与描述处理方法一致。 |
| 退款自主完结率\_行业均值 | 转换成浮点型。与纠纷退款率处理方法一致。 |
| 纠纷退款率\_比较 | 将有小于，大于的值的属性改成数值型，小于为0，大于为1。 |
| 仅退款自主完结时长\_比较 | 与纠纷退款率\_比较方法一致。 |
| 退货退款自主完结时长\_比较 | 与纠纷退款率\_比较方法一致。 |
| 退款自主完结率\_比较 | 与纠纷退款率\_比较方法一致。 |

#### 3.2.4天猫中食品店铺数据

**（1）目标**

将所有可以变成数值特征的文本特征变成数值型特征。

**（2）需要转换的数据类型**

除了店铺描述信息之外的所有属性都需要。处理方式与天猫中电器店铺数据一致。

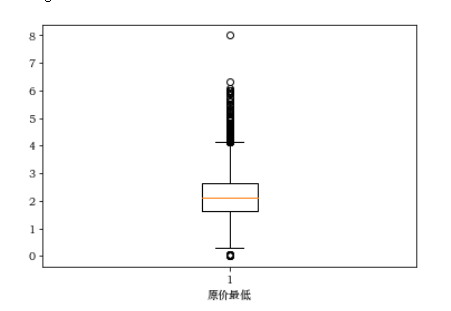
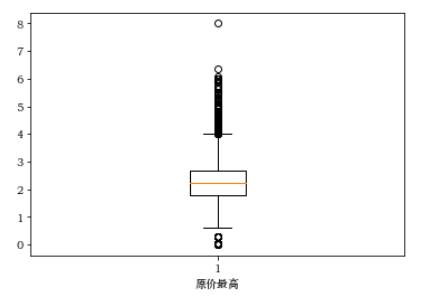
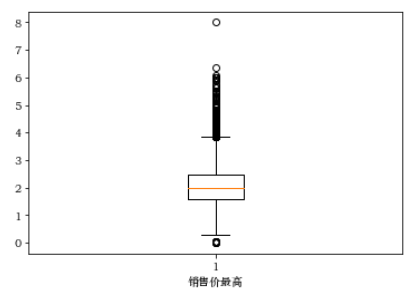
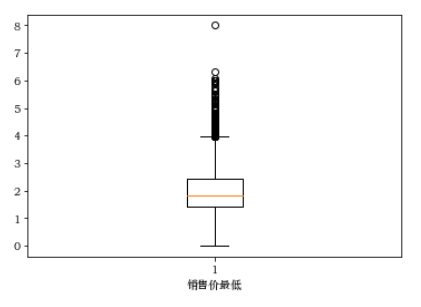
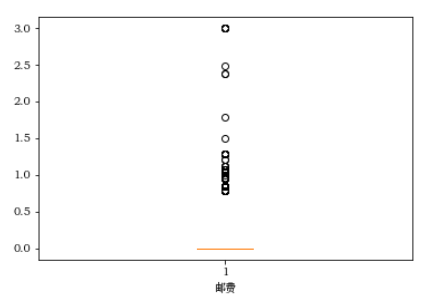
### 3.3异常值处理

#### 3.3.1看店宝中电器店铺数据

**（1）用describe()函数观察数据的分布情况**

得出有异常值的属性有：原价最低，原价最高，邮费，销售价最低，销售价最高。

**（2）用箱线图观察这些属性**

****

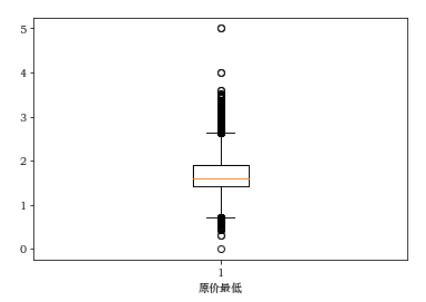
由于要去除这些异常值的话要删除整个样本，所以最后决定保留这些值。

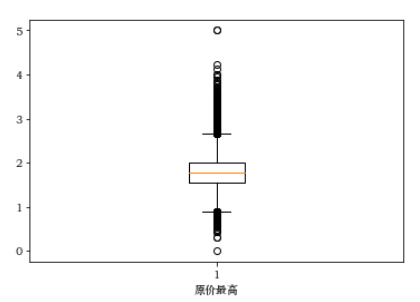
#### 3.3.2看店宝中食品店铺数据

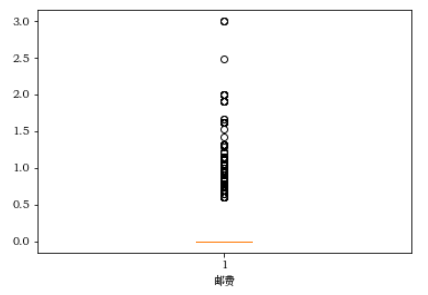
**（1）用describe()函数观察数据的分布情况**

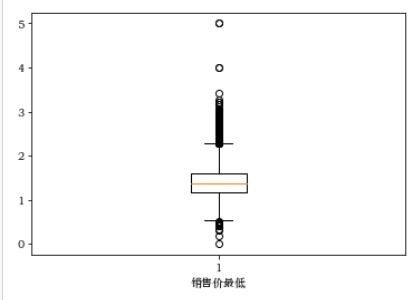
得出有异常值的属性有：原价最低，原价最高，邮费，销售价最低，销售价最高。

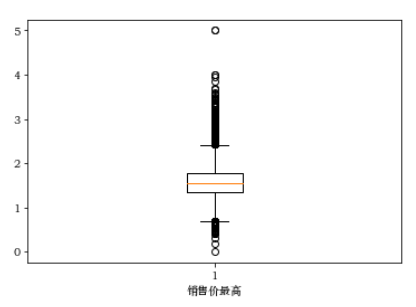
**（2）用箱线图观察这些属性**











由于要去除这些异常值的话要删除整个样本，所以最后决定保留这些值。

#### 3.3.3天猫中电器店铺数据

用describe()函数观察数据的分布情况，没有发现有异常值。

#### 3.3.4天猫中食品店铺数据

用describe()函数观察数据的分布情况，没有发现有异常值。

## 四、特征工程

我们将特征分为三大类：卖家服务状况，历史交易状况，待交易商品特征。分别在各类中做特征工程。

在数据处理的过程中我们的命名规则为：X+类别序号+类别中的分类序号+\_+属性名称（例：描述卖家服务状况中卖家服务专业性的描述得分命名为X11\_描述得分）

在特征工程开始前，我们在天猫数据的表格中加入了描述比率正负，服务比率正负，物流比率正负三个特征变量，这三个特征变量由各自的评分与各自类别的行业平均值范围（比率中为0的评分）的比较得出，大于该类别平均值的值为1，小于该类别平均值的值为0，等于该类别平均值的值为2。

表格最终格式为，将看店宝数据和天猫数据按照电器和食品两个类别根据店铺名称合并为两张张表格。

由于食品数据和店铺数据处理的方式一致，所以下面只分析电器数据，食品数据不多赘述。

### 4.1卖家服务状况

#### 4.1.1卖家服务专业性（内容质量）

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X11\_描述得分 | 天猫数据中的描述属性中的值 |
| X11\_描述得分\_比较正负 | 天猫数据中的描述比率正负中的值 |
| X11\_描述得分\_比较比率 | 将天猫数据中描述比率加上正负号，正负号的判断来自描述比率正负中的值。 |
| X11\_开店时长 | 将天猫数据中店铺描述信息中的值的数字提取出来，代表开店的时长，如果该值为无，得到的结果为空。 |
| X11\_在售商品数 | 将看店宝数据根据店铺名称聚合，然后用count聚合函数来计算序号属性中的值的总数。 |
| X11\_保证金金额 | 天猫数据中的保证金余额属性中的值 |
| X11\_总评分 | 将天猫数据中描述、服务、物流三个属性的值相加 |
| X11\_描述评分占总评分 | 将天猫数据中描述属性的值除X11\_总评分 |

#### 4.1.2卖家服务及时性

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X12\_仅退款自主完结时长 | 天猫数据中的仅退款自主完结时长属性中的值。 |
| X12\_仅退款自主完结时长\_比较 | 天猫数据中的仅退款自主完结时长\_比较属性中的值。 |
| X12\_退货退款自主完结时长 | 天猫数据中的仅退货退款自主完结时长属性中的值。 |
| X12\_退货退款自主完结时长\_比较 | 天猫数据中的退货退款自主完结时长\_比较属性中的值。 |

#### 4.1.3卖家服务一致性

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X13\_服务得分 | 天猫数据中的服务属性中的值。 |
| X13\_服务得分\_比较正负 | 天猫数据中的服务比率正负属性中的值。 |
| X13\_服务得分\_比较比率 | 将天猫数据中服务比率加上正负号，正负号的判断来自服务比率正负中的值。 |
| X13\_服务评分占总评分 | 将天猫数据中服务属性的值除X11\_总评分 |

#### 4.1.4售后服务

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X14\_纠纷退款率 | 天猫数据中的纠纷退款率属性中的值。 |
| X14\_纠纷退款率\_比较 | 天猫数据中的纠纷退款率\_比较属性中的值。 |
| X14\_退款自主完结率 | 天猫数据中退款自主完结率属性中的值。 |
| X14\_退款自主完结率\_比较 | 天猫数据中退款自主完结率\_比较属性中的值 |

#### 4.1.5物流服务

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X15\_物流得分 | 天猫数据中的物流属性中的值。 |
| 'X15\_物流得分\_比较正负 | 天猫数据中的物流比率正负属性中的值。 |
| X15\_物流得分\_比较比率 | 将天猫数据中物流比率加上正负号，正负号的判断来自物流比率正负中的值。 |
| X15\_物流评分占总评分 | 将天猫数据中物流属性的值除X11\_总评分 |
| X15\_发货好标签评论数 | 统计看店宝数据中评价标签属性中的文字出现发货快、发货好、发货赞、发货好评的商品的个数。 |
| X15\_物流好标签评论数 | 统计看店宝数据中评价标签属性中的文字出现物流快，物流好，物流给力的商品的个数。 |
| X15\_包装好标签评论数 | 统计看店宝数据中评价标签属性中的文字出现包装好，包装实，包装精的商品的个数。 |

#### 4.1.6顾客满意度

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X16\_服务高标签评论数 | 统计看店宝数据中评价标签属性中的文字出现服务好、服务棒、服务一流、服务热情、服务到位的商品的个数。 |

### 4.2历史交易状况

在清洗完的看店宝数据中加入评价标签-new得分属性，该属性运用了SnowNLP库中的sentiments方法对看店宝数据中的评价标签数据的每个标签进行情感分析，将其加总得到对应商品的评价标签的总情感值，该值越高情绪越积极，其取值范围在0-1之间。

#### 4.2.1销售历史情况

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X21\_30天总销售额 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算30天销售额属性中的值的总和。 |
| X21\_30天总销量 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算30天销量额属性中的值的总和。 |
| X21\_30天销售均价 | 用X21\_30天总销售额除X21\_30天总销量 |

#### 4.2.2销售稳定程度

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X22\_最畅销产品占全店销售额比例 | 将看店宝数据根据店铺名称聚合，然后用max聚合函数来选出占全店销售额比例属性中的值的最大值。 |
| X22\_总销量 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算总销量属性中的值的总和。 |
| X22\_近30天销量占总销量比例 | 用X21\_30天总销量除X22\_总销量 |

#### 4.2.3评价真实程度

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X23\_平均收藏数 | 将看店宝数据根据店铺名称聚合，然后用mean聚合函数来计算出收藏数属性的平均值。 |
| X23\_平均评价数 | 将看店宝数据根据店铺名称聚合，然后用mean聚合函数来计算出评价数属性的平均值。 |
| X23\_平均总销量 | 将看店宝数据根据店铺名称聚合，然后用mean聚合函数来计算出总销量属性的平均值。 |
| X23\_总销量 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算总销量属性中的值的总和。 |
| X23\_总收藏数 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算收藏数属性中的值的总和。 |
| X23\_总评价数 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算评价数属性中的值的总和。 |
| X23\_每家店评价标签情感平均值 | 将看店宝数据根据店铺名称聚合，然后用mean聚合函数来计算出评价标签-new得分属性的平均值。 |

### 4.3待交易商品特征

#### 4.3.1待交易商品价格

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X31\_最高商品原价 | 将看店宝数据根据店铺名称聚合，然后用max聚合函数来选出原价最高属性中的值的最大值。 |
| X31\_最低商品原价 | 将看店宝数据根据店铺名称聚合，然后用min聚合函数来选出原价最低属性中的值的最小值。 |
| X31\_店铺内商品原价最大差价 | 用X31\_最高商品原价-X31\_最低商品原价 |
| X31\_最高商品销售价 | 将看店宝数据根据店铺名称聚合，然后用max聚合函数来选出销售价最高属性中的值的最大值。 |
| X31\_最低商品销售价 | 将看店宝数据根据店铺名称聚合，然后用min聚合函数来选出销售价最低属性中的值的最小值。 |
| X31\_店铺内商品销售价最大差价 | 用X31\_最高商品销售价-X31\_最低商品销售价。 |
| X31\_最大折扣力度 | 用1-将看店宝数据根据店铺名称聚合，然后用min聚合函数来选出折扣率属性中的值的最小值。 |
| X31\_最小折扣力度 | 用1-将看店宝数据根据店铺名称聚合，然后用max聚合函数来选出折扣率属性中的值的最大值。 |
| X31\_平均折扣力度 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算折扣率属性中的值的总和，将总和除该店铺样本总个数得到该值 |

#### 4.3.2待交易商品种类/品类

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X32\_一级类目数量 | 将看店宝数据根据店铺名称聚合，然后计算出各个店铺一级类目中不同类别的数量 |
| X32\_二级类目数量 | 将看店宝数据根据店铺名称聚合，然后计算出各个店铺二级类目中不同类别的数量 |
| X32\_三级类目数量 | 将看店宝数据根据店铺名称聚合，然后计算出各个店铺三级类目中不同类别的数量 |

#### 4.3.3待交易商品人气

在看店宝数据中加入新属性

|  |  |
| --- | --- |
| 是否存在评价标签 | 看店宝数据中该商品评价标签属性是否有值，1为有0为没有 |
| 评价标签个数 | 将看店宝数据中的评价标签属性用split函数将字符串按；划分开，分开后统计该列表长度 |
| 活动促销 | 将看店宝数据中的活动促销属性用split函数将字符串按|划分开，分开后统计该列表长度 |

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X33\_店铺商品总PC分享量 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算PC总分享量属性中的值的总和 |
| X33\_店铺商品平均PC分享量 | X33\_店铺商品总PC分享量除该店铺样本总个数 |
| X33\_店铺含有评论标签商品总数 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算评价标签个数属性中的值的总和 |
| X33\_店铺商品含有评论标签商品概率 | 将看店宝数据根据店铺名称聚合，用sum聚合函数计算出是否存在评价标签属性的总和，用该总和除该店铺样本的个数得到该值 |
| X33\_店铺商品含有评论标签平均个数 | 将看店宝数据根据店铺名称聚合，用sum聚合函数计算出评价标签个数属性的总和，用该总和除该店铺样本的个数得到该值 |
| X33\_店铺商品平均活动促销数 | 将看店宝数据根据店铺名称聚合，用sum聚合函数计算出活动促销属性的总和，用该总和除该店铺样本的个数得到该值 |

#### 4.3.4待交易商品质量

在看店宝数据中加入新属性

|  |  |
| --- | --- |
| 货号 | 看店宝数据中该商品货号属性是否有值，1为有0为没有 |

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X34\_店铺存在货号商品总数 | 将看店宝数据根据店铺名称聚合，然后用sum聚合函数来计算货号属性中的值的总和。 |
| X34\_店铺商品存在货号概率 | 将看店宝数据根据店铺名称聚合，用sum聚合函数计算出货号属性的总和，用该总和除该店铺样本的个数得到该值。 |

#### 4.3.5待交易商品基础承诺状况

在看店宝数据中加入新属性

|  |  |
| --- | --- |
| 服务保证 | 将看店宝数据中的评价标签属性去掉[]符号，并用split函数将字符串按，划分开。 |
| 服务数量 | 服务保证的列表长度 |

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X35\_店铺商品平均服务数量 | 将看店宝数据根据店铺名称聚合，用sum聚合函数计算出服务数量属性的总和，用该总和除该店铺样本的个数得到该值。 |

## 五、特征衍生

特征衍生整体上采用了两个方法，分别是取对数方法和分箱方法。

**（1）取对数方法**

取对数分为取2或10为底的对数，看数据的范围，若数量级小于等于4则以2为底，否则以10为底。要求数据范围为非负，数据范围为全体实数，所以需要数据正向化后再进行取对数。

**（2）分箱方法**

以5,7,9,11的箱数进行等宽分箱。

### 5.1对X1特征的衍生

大部分特征衍生如表格所示，此处解释部分特殊特征衍生原因：第一个，开店时长添加平方项，进行此特征衍生是用于判断是否可能出现倒U型效应；第二个是纠纷退款率的相关特征衍生，因为这个特征大部分的店铺都为0，因此一旦有一个店铺不为0，说明这个店铺可能相对于其他的店铺来说更不好，我们应该加强这个特征的影响效果，先计算得出一个纠纷退款率因子，再衍生出纠纷退款率因子的平方项，三次方项，加强纠纷退款率因子的影响。

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X11\_描述得分\_比较正负 | 修改：-1为负，0为持平，1为正。将0变-1，2变0，1不变。 |
| X13\_服务得分\_比较正负 | 修改：-1为负，0为持平，1为正。将0变-1，2变0，1不变。 |
| X15\_物流得分\_比较正负 | 修改：-1为负，0为持平，1为正。将0变-1，2变0，1不变。 |
| X11\_描述得分\_分箱 | 将dianqi\_X1数据中的X11\_描述得分属性按7个箱数等宽分箱。 |
| X13\_服务得分\_分箱 | 将dianqi\_X1数据中的X11\_服务得分属性按7个箱数等宽分箱。 |
| X15\_物流得分\_分箱 | 将dianqi\_X1数据中的X11\_物流得分属性按7个箱数等宽分箱。 |
| X11\_描述得分\_比较比率\_分箱 | 将dianqi\_X1数据中的X11\_描述得分\_比较比率属性按9个箱数等宽分箱。 |
| X13\_服务得分\_比较比率\_分箱 | 将dianqi\_X1数据中的X13\_服务得分\_比较比率属性按9个箱数等宽分箱。 |
| X15\_物流得分\_比较比率\_分箱 | 将dianqi\_X1数据中的X15\_物流得分\_比较比率属性按9个箱数等宽分箱。 |
| X11\_在售商品数对数\_以2为底 | 将dianqi\_X1数据中的X11\_在售商品数属性以2为底取对数。 |
| X11\_开店时长\_平方 | 将dianqi\_X1数据中的X11\_开店时长属性平方 |
| X14\_是否存在纠纷退款失败 | 判断ianqi\_X1数据中的X14\_纠纷退款率属性是否为0，如果为0则该值为0，如果不为0则该值为1 |
| X14\_纠纷退款率因子 | dianqi\_X1数据中X14\_纠纷退款率\*10n次方，n取值为所有店铺纠纷退款率最小(非0)的小数点位数，例如：五个店铺纠纷退款率为(0,0,0.1,0.01,0.001)，则n为1000 |
| X14\_纠纷退款率因子\_平方 | X14\_纠纷退款率因子的平方 |
| X14\_纠纷退款率因子\_三次方 | X14\_纠纷退款率因子的立方 |
| X14\_退款自主完结率\_分箱 | 将dianqi\_X1数据中的X14\_退款自主完结率属性按11个箱数等宽分箱。 |
| X15\_发货好标签评论数\_以2为底 | 将dianqi\_X1数据中的X15\_发货好标签评论数属性以2为底取对数。 |
| X15\_物流好标签评论数\_以2为底 | 将dianqi\_X1数据中的X15\_物流好标签评论数属性以2为底取对数。 |
| X15\_包装好标签评论数\_以2为底 | 将dianqi\_X1数据中的X15\_包装好标签评论数属性以2为底取对数。 |
| X16\_服务高标签评论数\_以2为底 | 将dianqi\_X1数据中的X16\_服务高标签评论数属性以2为底取对数。 |

### 5.2对X2特征的衍生

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X21\_描述比率正负 | 修改：-1为负，0为持平，1为正。将0变-1，2变0，1不变。 |
| X23\_30天总销售额对数\_以10为底 | 将dianqi\_X2数据中的X23\_30天总销售额对数属性以10为底取对数。 |
| X23\_30天总销量对数\_以10为底 | 同上 |
| X23\_30天销售均价对数\_以2为底 | 同上 |
| X23\_平均收藏数对数\_以10为底 | 同上 |
| X23\_平均评价数对数\_以10为底 | 同上 |
| X23\_平均总销量对数\_以10为底 | 同上 |
| X23\_总收藏数对数\_以10为底 | 同上 |
| X23\_总评价数对数\_以10为底 | 同上 |
| X23\_总销量对数\_以10为底 | 同上 |
| X23\_评价数占销量率 | dianqi\_X2数据中X23\_总评价数除X23\_总销量 |
| X23\_收藏数占销量率 | dianqi\_X2数据中X23\_总收藏数除X23\_总销量 |
| X23\_收藏数占评价数率 | dianqi\_X2数据中X23\_总收藏数除X23\_总评价数 |
| X23\_评价数占店收藏数率 | dianqi\_X2数据中X23\_总评价数除X23\_总收藏数 |
| X23\_每家店评价标签情感平均值\_分箱 | 将dianqi\_X2数据中的X23\_每家店评价标签情感平均值属性按11个箱数等宽分箱。 |

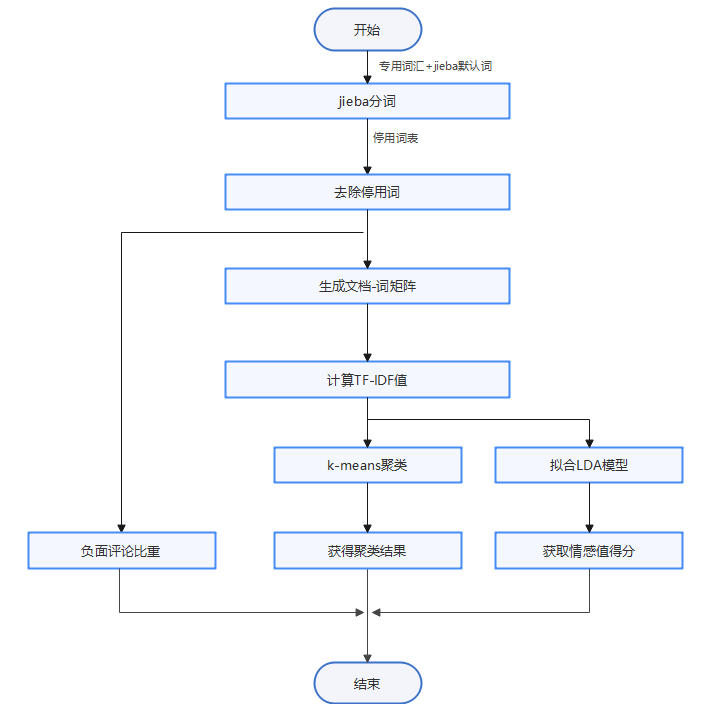
### 5.3对X3特征的衍生

大部分特征衍生如表格所示，此处特别解释存在货号概率这个指标，这个指标表示的是在一个店铺购买的商品存在货号的概率，存在货号的商品比不存在货号的商品，其正品概率更大，因此该指标可以衡量一个店铺的正品概率。放大该特征的影响：衍生出平方及三次项，因为其原本范围在0-1，应该\*100再平方和三次方。

|  |  |
| --- | --- |
| 特征 | 处理过程 |
| X31\_最高商品原价对数\_以10为底 | 将dianqi\_X3数据中的X31\_最高商品原价对数属性以10为底取对数。 |
| X31\_最低商品原价对数\_以2为底 | 同上 |
| X31\_店铺内商品原价最大差价对数\_以10为底 | 同上 |
| X31\_最高商品销售价对数\_以10为底 | 同上 |
| X31\_最低商品销售价对数\_以2为底 | 同上 |
| X31\_店铺内商品销售价最大差价对数\_以10为底 | 同上 |
| X31\_最大折扣力度\_分箱 | 将dianqi\_X1数据中的X31\_最大折扣力度属性9个箱数等宽分箱。 |
| X31\_最小折扣力度\_分箱 | 同上 |
| X31\_平均折扣力度\_分箱 | 同上 |
| X32\_一级+二级类目数量 | 将dianqi\_X1数据中X32\_一级类目数量+X32\_二级类目数量 |
| X32\_二级+三级类目数量 | 同上 |
| X32\_一级+二级+三级类目数量 | 同上 |
| X33\_店铺商品总PC分享量对数\_以10为底 | 对数操作同上 |
| X33\_店铺商品平均PC分享量对数\_以2为底 | 对数操作同上 |
| X34\_店铺商品存在货号概率\_衍生 | 将dianqi\_X3数据中X34\_店铺商品存在货号概率属性的数值\*100 |
| X34\_店铺商品存在货号概率\_平方 | X34\_店铺商品存在货号概率\_衍生数值的平方 |
| X34\_店铺商品存在货号概率\_三次方 | X34\_店铺商品存在货号概率\_衍生数值的立方 |
| X34\_店铺商品存在货号概率\_分箱 | 等宽分箱，分箱数为9。分享操作同上。 |
| X34\_店铺商品存在货号概率\_平方\_分箱 | 等宽分箱，分箱数为9。分享操作同上。 |
| X34\_店铺商品存在货号概率\_三次方\_分箱 | 等宽分箱，分箱数为9。分享操作同上。 |

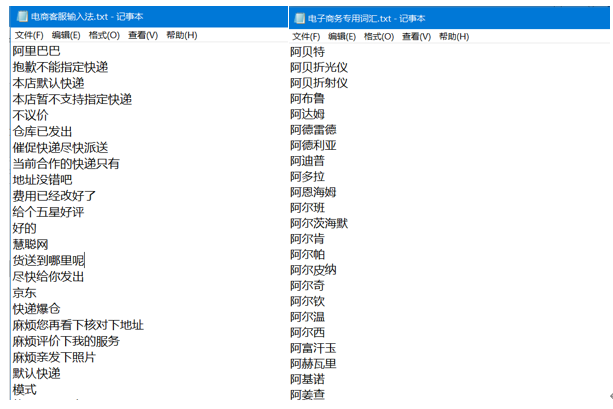
### 5.4根据店铺评论NLP进行特征衍生

我们的评论数据来源是看店宝网站和天猫网站。每家店铺选取截止爬取前最新的2000条评论，约300家店铺，共计约60万条评论。NLP分析的流程如下图所示。

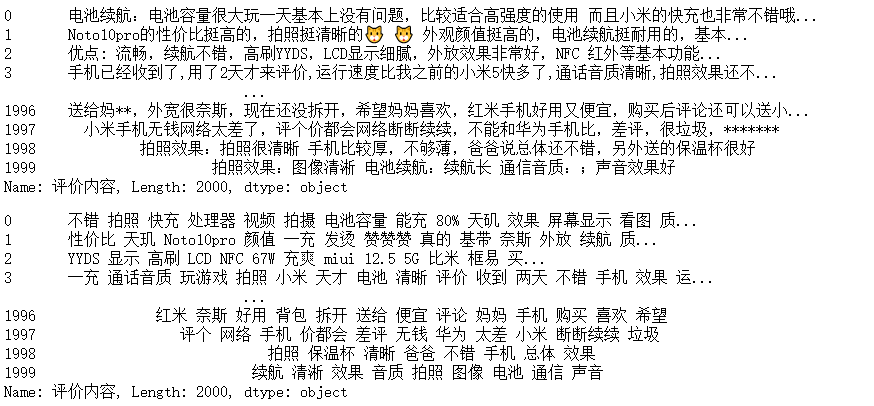


#### 5.4.1生成词典库并分词

首先我们从搜狗词库找到了”电商客服输入法“和”电子商务”专用词库”两个词库，然后采用”深蓝词库转换”工具转成.txt文件，将其导入程序中，再加上python的jieba库默认的词库来组成我们的词典库。



然后依赖构建的词典库对获取的评论数据进行分词，之后再根据停用词表对分词后的一些特殊字符，表情，无意义的字符等进行剔除。得到所有有意义的词汇信息。如下图所示，



对比分词前后，我们可以明显地看到已经成功剔除了很多无关的信息，最终词汇也都基本很好的被分出来了，包括“YYDS”,”赞赞赞”这些明显带有互联网色彩的词也都被很好的分出来了。

以“小米官方旗舰店”为例，制作词云如下：



可以看到出现频率最高的是手机，电池，速度，续航等信息，这反映出了小米的主要业务信息，同时也可以看到不错，性价比，流畅等词出现的频率也很高，这反映出了顾客对于这家店铺的主流评价。

#### 5.4.2生成文档-词矩阵

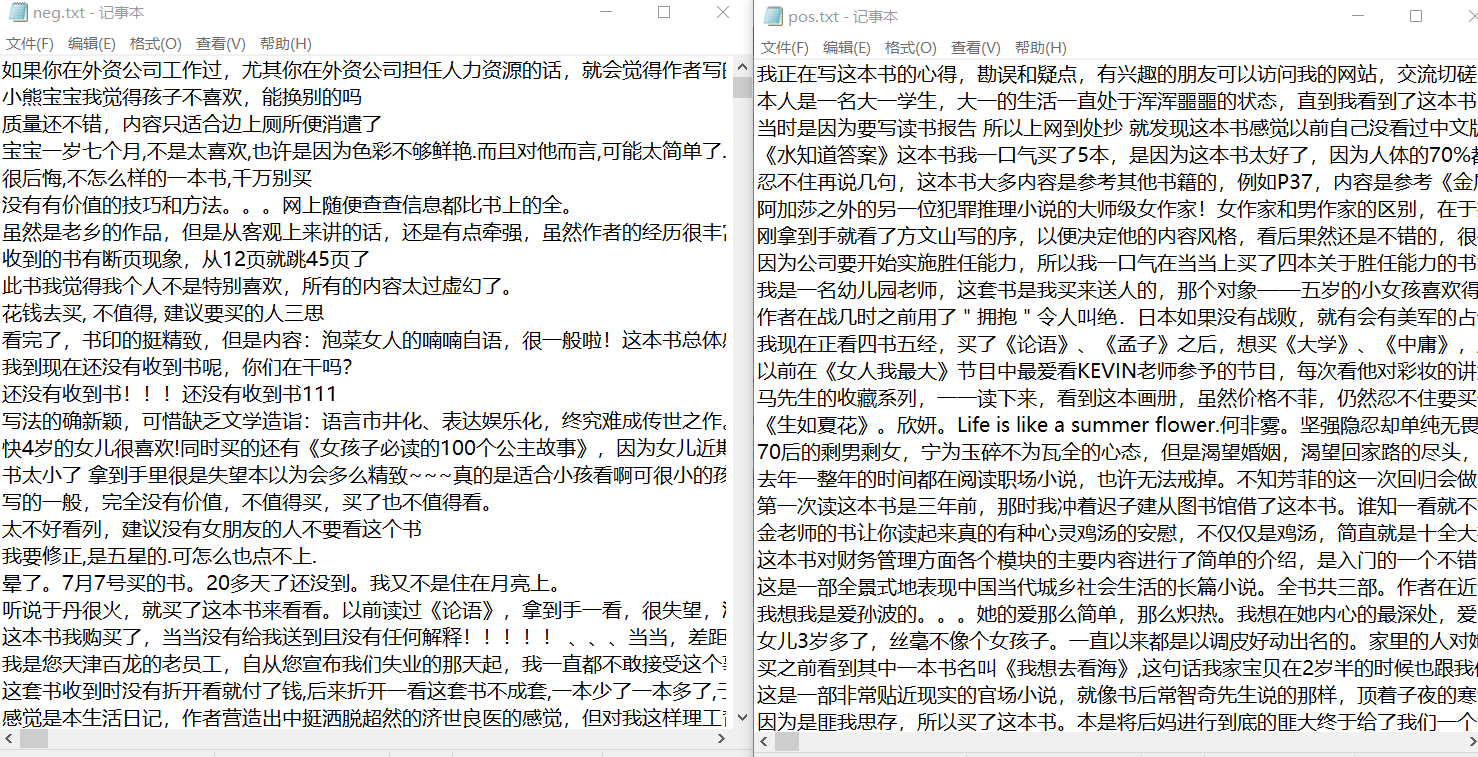
在成功分词之后，我们使用python的sklearn包生成了文档-词矩阵，方便我们根据词向量进行后续的建模。为了使我们的矩阵不至于太过稀疏，在生成的过程中只有在2个以上用户中出现的词我们才会保留。然后基于词频矩阵X计算TF-IDF值，得到每个词在文档中的重要性。之后再根据LDA模型抽取文档的主题。我们提取了文档的前10个主题，每个主题选择前10个关键词。

#### 5.4.3计算情感得分

之后我们希望获取每家店铺的情感值得分，这里我们采用的python的snownlp库，SnowNLP可以方便的处理中文文本内容，它自带了一些训练好的字典。下图中的neg.txtx和pos.txt就是消极和积极的词库文件。



同时，也可以通过训练自己提供的语料来制定与自己分析内容更为贴近的模型。我们分析分档中的内容发现里面的信息设计的方面很广，是属于百科知识类型的，其中也包括商品评论的数据。因为我们自身的数据量就有限，所以这里我们直接采用默认的词典库来进行构建情感值得分。



snownlp根据这两个库中的信息对用户新输入的字符串进行分析，给出不同的情感值。这里情感值用0-1之间的数字来表示，0表示绝对消极的态度，1表示绝对积极的态度，从0-1态度逐渐由消极转变为积极。我们对每个主题的关键字情感值进行求和，然后对10个主题的情感值求平均得到该文档的最终情感值得分。

因为评论数据有些可能是玩家为了获取店铺奖励或者是平台默认超过规定期限不评价就会是好评，所以可借鉴的价值不高。相比之下我们更关心一些负面的评论信息，所以我们从分词之后就获取每个用户评论的情感值，然后根据其情感值得分判断评论的态度，判定的界限为0.5，小于0.5的为消极评论，大于0.5的表示积极评论。然后我们计算出每家店铺消极评论所占的比重，比重越大代表该店铺的信用值越低。

#### 5.4.4聚类分析

以上过程都是根据文档的情感值来进行分析，却忽略了词频本身具有的价值，所以我们在获取个词在文档中的比重之后采用DBSCAN进行聚类分析，将以上店铺分成8个簇，构建出基于词频的一个综合评级。因为词向量不好对模型的结果进行解释，所以我们不再对聚类的结果进行分析，而是直接采用哑变量的形式消除各个类别之间的差异。最后将我们从文本中提取出的特征代入到主特征工程模块中进行最后的模型构建。

## 六、建模

### 6.1数据降维

经过特征工程以及特征衍生，我们一共得到了3个一级指标，一级指标下有14个二级指标，特征维数为133。而这个数据集不能直接进入聚类模型，必须先通过数据降维，然后使用降维后的数据进行建模。降维就是一种对高维度特征数据预处理方法。降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

#### 6.1.1 PCA

PCA(Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第1,2个轴正交的平面中方差最大的。依次类推，可以得到n个这样的坐标轴。通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面k个坐标轴中，后面的坐标轴所含的方差几乎为0。于是，我们可以忽略余下的坐标轴，只保留前面k个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为0的特征维度，实现对数据特征的降维处理。得到这些包含最大差异性的主成分方向的方法：通过计算数据矩阵的协方差矩阵，然后得到协方差矩阵的特征值特征向量，选择特征值最大(即方差最大)的k个特征所对应的特征向量组成的矩阵。这样就可以将数据矩阵转换到新的空间当中，实现数据特征的降维。

#### 6.1.2 PCA结果分析

我们使用Python3.7中的Sklearn包中的PCA模块，对本项目的数据进行数据降维。首先先数据进行标准化，将标准化后的数据集进行PCA降维，得到18个主成分， 提取这18个主成分对应的特征值和方差百分比作表如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 提取载荷平方和 | | | |
| 主成分 | 总计 | 方差百分比 | 累计 % |
| **0** | 23.07584 | 0.172147 | 0.172147 |
| **1** | 18.23361 | 0.136024 | 0.308171 |
| **2** | 9.351281 | 0.069761 | 0.377932 |
| **3** | 8.03837 | 0.059967 | 0.437899 |
| **4** | 6.785594 | 0.050621 | 0.48852 |
| **5** | 6.325377 | 0.047188 | 0.535707 |
| **6** | 5.495874 | 0.041 | 0.576707 |
| **7** | 4.409337 | 0.032894 | 0.609601 |
| **8** | 3.73956 | 0.027897 | 0.637498 |
| **9** | 3.312926 | 0.024715 | 0.662213 |
| **10** | 2.866463 | 0.021384 | 0.683597 |
| **11** | 2.796546 | 0.020862 | 0.704459 |
| **12** | 2.672845 | 0.01994 | 0.724399 |
| **13** | 2.506304 | 0.018697 | 0.743096 |
| **14** | 2.176263 | 0.016235 | 0.759331 |
| **15** | 2.054365 | 0.015326 | 0.774656 |
| **16** | 1.823274 | 0.013602 | 0.788258 |
| **17** | 1.762792 | 0.013151 | 0.801409 |

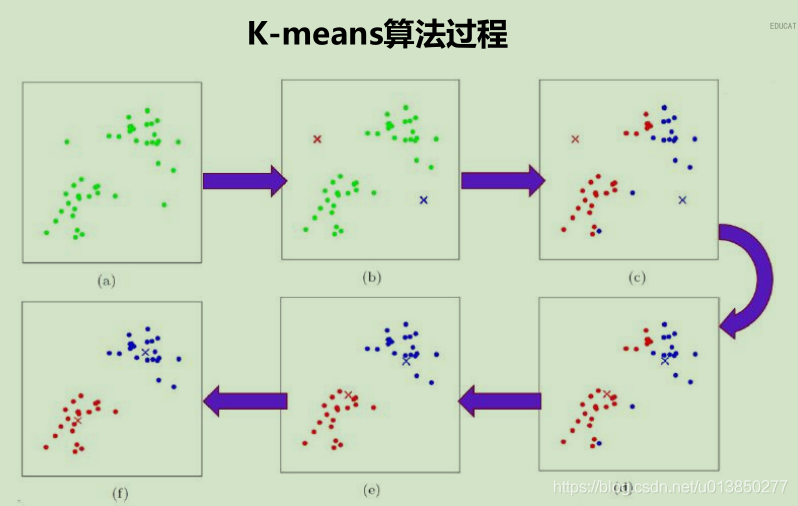
从上表可以看出主成分1的方差百分比达到了17.22%，主成分2的方差百分比达到了13.60%，说明主成分1解释了原来变量的17.22%的内容，主成分2解释了原来变量的13.60%的内容。从第三个主成分开始方差百分比就仅仅只有6.97%，再后面的主成分就更小了。我们在后续建模中选择使用方差贡献率累计达到80%的18个主成分。

### 6.2 聚类结果

#### 6.2.1 K-means

K-means是机器学习中一个比较常用的算法，属于无监督学习算法，其常被用于数据的聚类，只需为它指定簇的数量即可自动将数据聚合到多类中，相同簇中的数据相似度较高，不同簇中数据相似度较低。

其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695)。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，[误差](https://baike.baidu.com/item/%E8%AF%AF%E5%B7%AE/738024" \t "https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95/_blank)[平方和](https://baike.baidu.com/item/%E5%B9%B3%E6%96%B9%E5%92%8C/783894)局部最小。



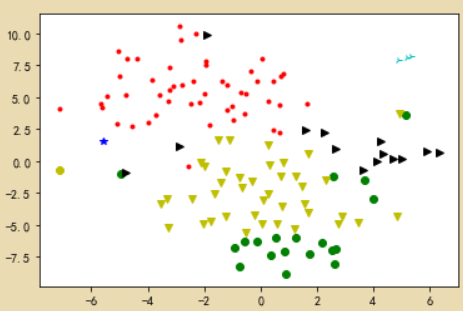
优点：容易理解，聚类效果不错，虽然是局部最优， 但往往局部最优就够了；处理大数据集的时候，该算法可以保证较好的伸缩性；当簇近似高斯分布的时候，效果非常不错；算法复杂度低。

缺点：K 值需要人为设定，不同 K 值得到的结果不一样；对初始的簇中心敏感，不同选取方式会得到不同结果；对异常值敏感；样本只能归为一类，不适合多分类任务；不适合太离散的分类、样本类别不平衡的分类、非凸形状的分类。

#### 6.2.2 K-means结果分析

我们将提取的18个主成分数据来进行Kmeans聚类，选择聚类簇数为3-15，通过观察发现簇数为7的聚类效果最好，因此我们后续的建模过程中选择聚类簇数为7。

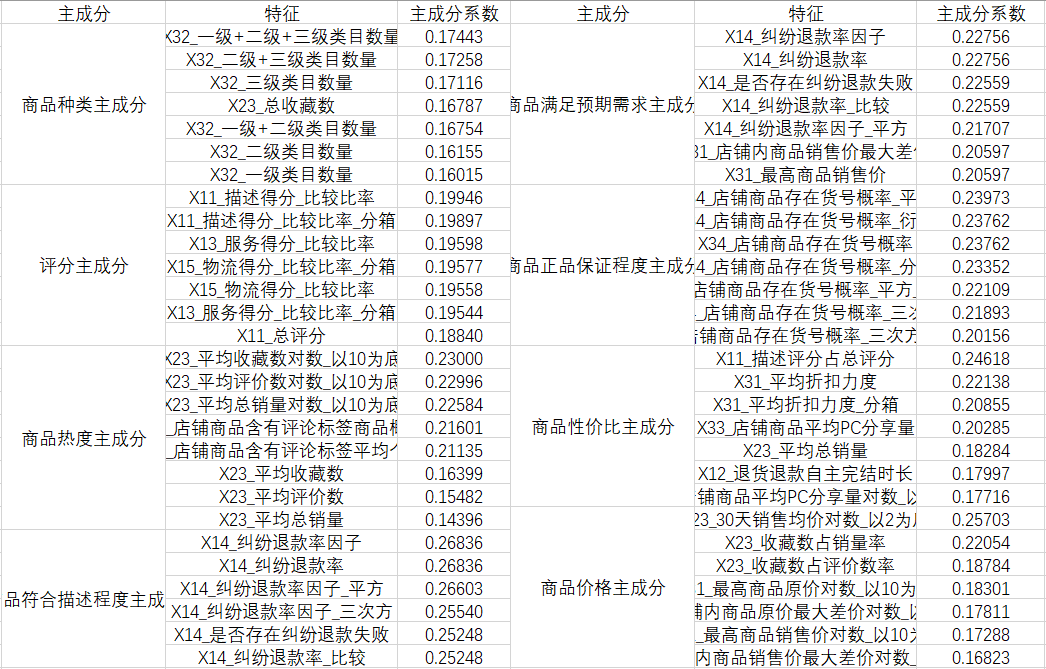
下图是使用sklearn中的TSNE模块对聚类结果进行可视化，可以发现该聚类效果较好。



### 6.3信用风险评估结果

#### 6.3.1主成分分析

前面我们对特征数据进行PCA后得到18个主成分数据，然而我们发现这18个主成分并不是都对原有特征数据产生相同的影响，因此我们只针对前8个主成分（方差贡献率>3%）进行分析，我们将这8个主成分进行以下分析：第一个主成分其最大的主成分系数对应的特征都与商品种类相关，因此我们把第一个主成分归类为商品种类主成分；第二个主成分其最大的主成分系数对应的特征都与店铺评分相关，因此我们把第二个主成分归类为评分主成分，其余六个主成分类似。



#### 6.3.2 信用风险评估

因为我们最后的聚类簇数为7，因此后续我们将信用风险评估的评级为7级，分别为AAA、AA、A、BB、B、C、D。我们按每个簇分组，取每个主成分的平均，然后按对应主成分的方差贡献率加权计算每个簇的信用得分，按照信用得分高低来进行信用评级，最终得到以下的信用风险评估结果。

|  |  |  |  |
| --- | --- | --- | --- |
| 信用评级 | 信用情况 | 公司数量 | 聚类簇数 |
| AAA | 信用极好 | 1 | 6 |
| AA | 信用优良 | 1 | 2 |
| A | 信用较好 | 14 | 4 |
| BB | 信用一般 | 50 | 0 |
| B | 信用欠佳 | 40 | 3 |
| C | 信用较差 | 3 | 5 |
| D | 信用极差 | 19 | 1 |

## 七、未来建议和展望

本文通过爬取网络公开数据，完成数据清洗，多种方式进行特征衍生，得到可分析的数据，后利用PCA方法将113条衍生属性降维为17个属性，并通过Kmeans聚类方法依据降维后的属性数据将分别将电器和食品行业的数据进行无监督的聚类分级，最终得出针对不同店铺的一个评级结果。

以上成果表明研究所设计的小微企业信用评级模型能为小微企业提供一种客观的评价，能成为当前我国信用评级体系的有益补充。在大数据时代，可以采集的数据越来越多，因此也可以从更多的维度评价一个企业，能够从更细微的角度衡量一个企业，如果后续有更多的私有数据开放，则可以将更多的私有信息融入到模型，实现不同的数据互补。

在大数据时代，获取公开数据进行主体评级，用以估计主体在经济活动状况中的信用程度和经济能力，并将评级结果作为重要指标，甚至让其拥有与传统店铺财务指标一样的评估效力，已经成为互联网经济下针对电子商务店铺信贷评估的一个重要手段，其有助于促进我国电商甚至小微企业未来的征信模式的发展和革新。

但是在利用大数据进行信用评估时，仍存在着一些不足，针对这些问题，本文在此一一提出建议。

1. **数据获取困难、数据缺乏权威且难以被量化**

面对这些问题，政府需规范企业数据开放模式，完善小微企业信息采集通道，明确小微企业信用信息数据采集规范。党的十九大报告指出“要建设数字中国、智慧社会，发展数字经济、共享经济”。显然丰富的数据有助于更精准的评估企业信用。建议国家出台相关政策，在数字中国建设背景下，规范数据采集方式，同时实现电商、社交、交通等企业数据共享，实现工商、税务、法院、知识产权等部门的相关数据共享。这样有助于研究者专注于技术开发，提高评级模型的精度。

**（2）电商店铺相关数据储存于于中心节点上，数据可信性度略有折扣**

面对数据的造假问题，在后续研究中，可以探索将各种数据都保存在区块链系统上，利用区块链的去中心化、防伪溯源、信任机制等技术特征保障已有数据的不可篡改性，进一步提升企业信用评级的可信度。

**（3）尚未建立成熟可靠的大数据电商信贷风控系统**

针对大数据评估系统的建立，未来应当在无监督学习模型的基础上，通过加强数据处理系统，将理论模型与信贷实践结合起来，建立一个完善的电商信贷风控体系。通过分析多方数据域实际情况，建立相应的评估指标来检测其区分度、稳定性和可解释性能力。除此之外还设置风险预警机制，处理模型无法解决的风险，同时也要提前设置好风险止损机制，防止损失进一步扩大。由于完全依靠数据的风控模型可能在预测能力和逻辑处理方面还具备一定局限性，因此也要注重大数据技术能力的提升，加大技术研发投入并且引进专业信贷风控人才。

综上，尽管通过将机器学习与大数据结合起来，可以建立初步的电商店铺信用评估模型，并且这个模型具有一定的科学性和可靠性。但是其是否真正具备足够的实践价值，还需同电商产业整体数据存储和获取大政策大环境的配合，并在实践研究中不断更新迭代自身模型。唯有通过此方法进步自身，机器学习和大数据才能在电商信贷评估领域发挥自己的真正价值。