

**个人房产信用评估**

**----方案说明文档**

[一、个人房产信用评分介绍 - 3 -](#_Toc532934609)

[1.1 背景 - 3 -](#_Toc532934610)

[1.2 国内外信用研究现状 - 3 -](#_Toc532934611)

[1.3 任务分析 - 6 -](#_Toc532934612)

[1.4 准备事项 - 6 -](#_Toc532934613)

[1.5 技术路线图 - 8 -](#_Toc532934614)

[二、数据探索性分析 - 9 -](#_Toc532934615)

[2.1 数据分析与可视化 - 9 -](#_Toc532934616)

[2.2 特诊（征）工程 - 17 -](#_Toc532934617)

[三、个人房产信用评分模型 - 18 -](#_Toc532934618)

[3.1 基于传统的信用评分卡建模 - 19 -](#_Toc532934619)

[3.2 基于机器学习或者深度学习预测 - 19 -](#_Toc532934620)

[3.3 加权求和 - 21 -](#_Toc532934621)

[四、模型评估 - 22 -](#_Toc532934622)

[4.1赛题评分模型 - 22 -](#_Toc532934623)

[4.2模型预测结果 - 23 -](#_Toc532934624)

[4.3总结与展望 - 24 -](#_Toc532934625)

[五、团队介绍 - 24 -](#_Toc532934626)

[六、参考文献及资料 - 25 -](#_Toc532934627)

## 一、个人房产信用评分介绍

## 1.1 背景

随着国家政策调控趋严，各地政府加大对房产市场的调控，房地产行业贷款额度趋紧。作为一个资金密集型的产业，投资规模大、周期长，融资渠道相对单一，对银行贷款依赖过高是其显著特点，因此资金筹措是其发展的重要瓶颈，也是风险来源的渠道。在上述背景下，房地产行业信用体系建设对行业发展来说愈发重要，为此，本方案在防范其风险上综合考虑房地产全生命周期过程中所有行为后，深度分析企业、个人的房产信用。以大赛举办方提供的2017年1月-2018年5月购房者的房产交易数据与房屋状态海量数据为依据，针对房产交易双方，利用真实成交数据，从房地产交易维度对交易参与人进行评价，计算个人房产信用评分。以期助力市场监督与检查，建设领域内信用，赋能行业发展。

## 1.2 国内外信用研究现状

互联网的快速发展及云计算、大数据、移动互联网、物联网等技术日渐成熟，政府机关、金融机构、互联网公司、公共服务行业等都积累了丰富的数据资源，这些数据资源只有在大数据环境下才能得到充分利用，才能挖掘出这些数据与信用的关联关系，促进对征信数据进行更加充分的利用，进而更好的为市场企业服务。

信用评价体系主要包含评价方法、评价指标、评价程序、信用等级等内容，广义上还包括信用评价应用。目前，发达资本主义国家都建立了比较完善的信用评价体系，我国在这方面也进行了积极探索。总体来说国外学者在个人信用体系研究领域多偏重商业贷款、个人纳税信用体系方面，而国内很多学者关于信用体系的研究大都聚焦在商业信贷、银行信贷、社会信用体系等方面。

**1.2.1 国外信用评价体系概况**

国外对信用评价体系的研究主要围绕评价方法、评价机构的评价程序、评价等级符号三方面围绕开展。国际上早期评价方法包括财务分析法与要素分析法，现代主要采用模型分析法。

国际上一些知名的信用评价机构采用的评价程序不完全相同，但其评价的周期都较长，比如穆迪在收集了足够的评价信息后，组建评级委员会，从评价委员会的首次讨论开始到评价结果发布，大约需要60-90天时间。国际上对企业信用等级一般采用等级符号来区分，三大信用评估机构所采用的符号不完全相同，但划分标准和等级数量相差不大，如标准普尔和惠誉是20个等级，最高级都是AAA，最低级是D，穆迪是19个等级，最高级是Aaa，最低级是C。FICO模型对个人信用等级用信用分表示，取值范围为300-850分，680分以上通常会被认为是信用极好。

过去几年里，发达资本主义国家已经出现了一些新兴企业利用大数据来重新定义信用评估方式。比如美国的Neo Finance公司利用申请人的工作历史和其在LinkedIn（一家面向商业客户的社交网络）上的人际关系来决定是否向他提供贷款。另一家名为Zest Fiance的公司是前谷歌首席信息官创办的，它利用顾客各方面的信用数据来做评价，包含成千上万的参数，比如借款方搬家的次数、他们在网页上使用大写字母的能力等，这些参数可能与借款方的信用没有直接的因果关系，但是通过大数据手段可以发现一定的关联关系，因此也会对借款方的信用状况产生影响。这些企业通过对庞大的数据量进行深入挖掘，以此找到一条新的利润增长点。

**1.2.2 国内信用评价体系探索**

我国的信用评价体系发展较晚，主要采用了第三方评级机构和公共征信体系并行发展的模式。目前，国内信用评级业在自主发展的同时，也与国际信用评价机构积极开展合作，比如惠誉持有联合资信49%的股权、穆迪与大公国际曾签署为期3年的技术合作协议等。国内信用评价机构充分借鉴了国外在评价方法、评价指标、评价程序、信用等级等方面的经验，同时也在评价指标、评价模型等方面进行了积极探索。

目前比较典型的是2006年中国人民银行印发的《信用评级管理指导意见》，将企业评级要素分为企业素质、经营能力、获利能力、偿债能力、履约情况和发展前景六个方面；在中国信用行业标准技术委员会2009年编制的《企业信用评价标准》将企业信用评价过程分为外部环境、经营状况、管理状况、财务指标、公共信用记录和招投标信用记录五大部分。个人信用评价方面，目前国内主要使用的评价模型有：上海资信评估模型、深圳鹏远公司建立的“鹏元”模型和蚂蚁金服的芝麻信用评分模型。其中芝麻信用评分模型与美国的FICO模型类似，通过身份特质、履约能力、信用历史、人脉关系、行为偏好五个维度对个体信用进行分析。芝麻信用在评分的基础上扩展了信用贷、信用住、信用租车、出国免签等多种应用。关于房产领域个人信用体系问题的探讨还少有人涉及，而大数据个人房产信用体系领域研究方面更鲜有人涉及。

**1.2.3 当前信用评价体系存在的问题**

当前信用评价体系存在的问题主要有三方面：一是评价结果易受到外部因素的影响。主权评级容易受到政治因素的干预，比如2008年金融危机爆发以后，国际三大评级机构频繁下调欧洲国家的主权信用评级，导致欧债危机加剧，间接帮助美国度过危机。对社会主体的评价结果容易受到评价委员会人员构成的影响，评价委员会的业务能力、经验、主观好恶、品德都是重要的影响因素。二是评价指标不全面。传统信用评价指标比较注重金融消费领域的指标，许多信用评价机构并未将公共管理领域和互联网数据纳入信用评价体系，即便纳入也只是少量指标，比如德国将地铁逃票作为个人征信的指标。三是运用范围较窄。传统信用评价主要运用于金融领域，比如银行贷款、发行证券、商业合作等，也有些信用评价产品会被用在招聘求职等方面。近年来在我国兴起的芝麻信用也是主要运用于提供小额贷款、免押金等领域，芝麻信用的出国免签是一大亮点。但是由于信用数据来源和评价机制的局限性，传统信用评价产品还无法广泛运用到社会管理、公共服务等各个领域。

**1.2.4 行业需求**

随着《社会信用体系建设规划纲要（2014-2020年）》和《关于做好个人征信业务准备工作的通知》》的颁布，我国个人信用行业迎来了发展的黄金时期。信用将在社会经济生活中承担越来越重要的作用，拥有信用者可以获得机会、形成社会关系或达成信用交易。信用正成为社会资源配置的新依据，改变了传统经济仅仅依靠实物资本进行资源配置的方式，促进缺乏实物资本但是拥有信用资本的主体发展。为此，各方都在积极地为社会信用体系的构建而努力，而个人房产信用体系作为构成社会信用体系的重要分支，对于社会信用体系的建立与完善，有着非常重要的作用，其中个人房产信用体系的建立完善，不仅对稳定房产市场、规范房产市场行为、促进房产市场稳定有序发展具有重要意义，而且对整合社会系统信用体系的构建具有重要影响。

目前我国房地产信用体系正处在初级建设阶段，暂时没有形成比较完备的个人房产信用评估体系，相关的研究也只停留在定量和定性分析层面，主观性偏大，实施成本高。而面对房产交易市场纷繁复杂的情况，急需一套科学的且具有实际商业应用价值的个人房产信用评估体系，以作房产市场交易评估之用。

## 1.3 任务分析

通过对国内外信用体系的初步研究，结合大赛的背景和提供的相关数据信息；我们得出本次方案的两个核心点是：数据指标的选取和算法模型的构建。围绕这两点，方案立足于房产交易的全过程，建立一套“用数据说话、用数据评价”的个人房产信用评价程序模型，进一步对现实商业贷款进行拟合分析，以此对参与房产交易的个人进行信用评分，以作业内参考之用。

## 1.4 准备事项

（1）数据准备

数据来源于全球数据资源开发者大赛举办方针对行业算法赛赛题六提供的2017年1月-2018年5月购房者的房产交易与房屋状态数据，共计44.1万条，包含20个变量，具体如下字段（表 1）：

表1 个人房产信用评价体系指标表

|  |  |  |
| --- | --- | --- |
| USER\_ID | 购房人ID | 抽样&脱敏 |
| Age | 年龄 | 枚举值，按照0-5表示：0:20岁以下、1:20-30岁、2:30-40岁、3:40-50岁、4:50-60岁、5:60岁以上 |
| Sex | 性别 | 枚举值，按照0-1表示：0:男、1:女 |
| TEL\_ID | 电话归属地 | 枚举值，按照0-1表示：0:浙江省、1:其他 |
| Province | 省份 | 请下载省份字典进行参照 |
| NATIONALITY | 国籍 | 请下载国籍字典进行参照 |
| HOU\_ID | 房屋编号 | 整数型 |
| Property\_ID | 房屋属性 | 枚举值，按照0-1表示： 0：存量房、1：商品房、 |
| Property\_Usage\_Type | 房屋用途 | 枚举值，按照0-1表示： 0；住宅、1：非住宅 |
| Property\_Architectural\_Area | 建筑面积 | 枚举值，按照0-4表示： 0:小于60平方米、1:60-90平方米、2:90-120平方米、3:120-140平方米、4:大于140平方米 |
| Property\_Sign\_Date | 网签时间 | 以YYYY/MM表示 |
| Property\_Record\_Date | 备案时间 | 以YYYY/MM表示 |
| Property\_Price | 房屋成交总价 | 枚举值，按照0-5表示： 0:100万以下、1:100-200万、2:200-300万、3:300-400万、4:400-500万、5:500万以上 |
| Property\_Loan\_Way | 贷款方式 | 枚举值，按照1-5表示： 1:组合贷款、2公积金贷款、3:商业贷款、4、银行贷款5、其他 |
| Property\_Payment | 付款方式 | 枚举值，按照0-3表示： 1:一次性付款、2:分期付款、3:抵押贷款、4：其他方式、5：其他情况 |
| CFCS | 查封次数 | 整数型 |
| JFCS | 解封次数 | 整数型 |
| Purchase\_houcs | 历史买房次数 | 整数型 |
| Sale\_houcs | 历史卖房次数 | 整数型 |
| Mortgage\_starttime | 抵押开始时间 | 以YYYY/MM表示 |
| Mortgage\_endtime | 抵押结束时间 | 以YYYY/MM表示 |

（2）软件工具准备

Python 环境

Python2.7.12

支持numpy, scipy, pandas, tensorflow1.4, tflearn, keras

操作系统：Ubuntu16

内存：128GB

GPU：Nvidia Quadro M4000 8G

Java环境

JDK1.8

操作系统：Ubuntu16

内存：128GB

GPU：Nvidia Quadro M4000 8G

## 1.5 技术路线图

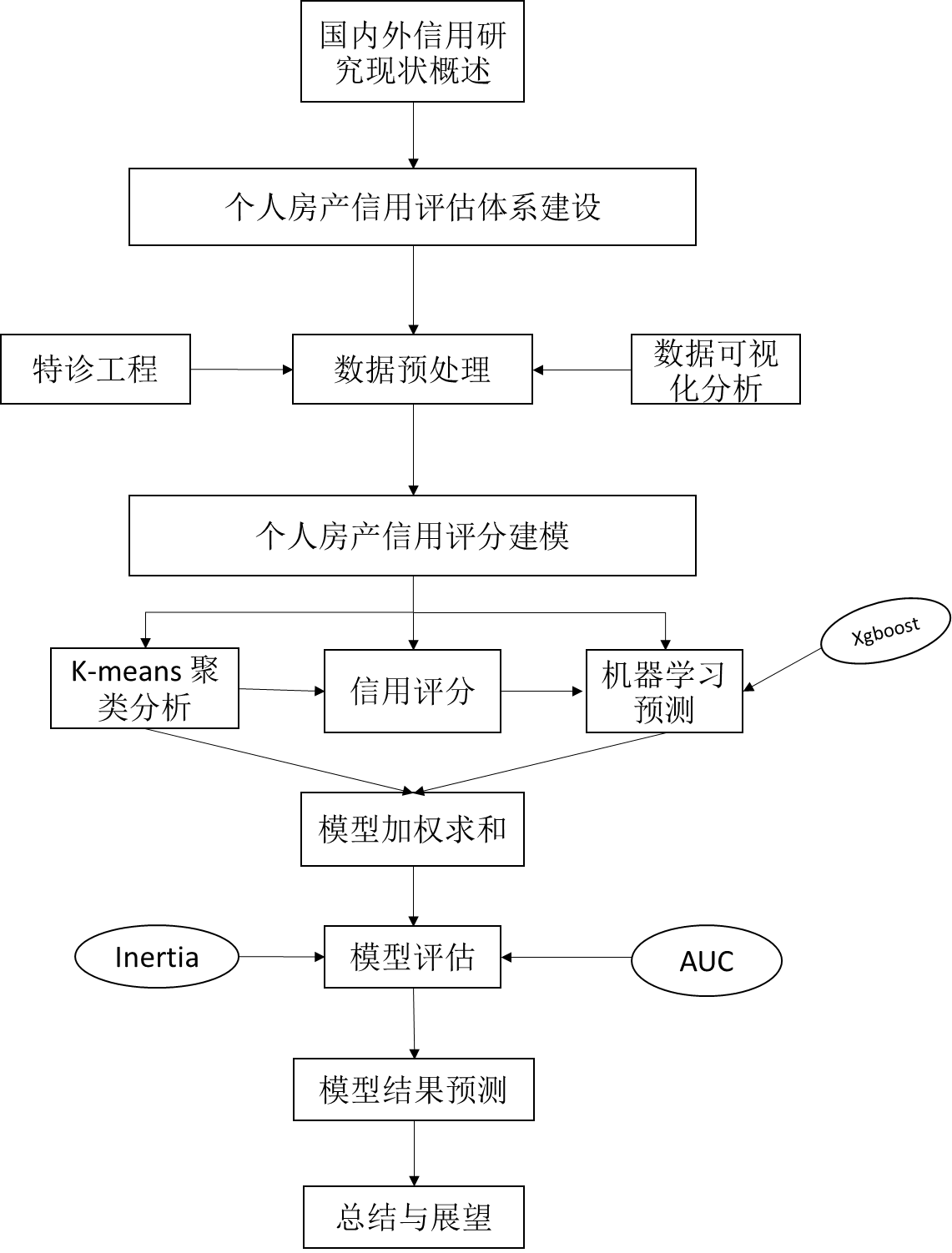


图1 技术路线图

## 二、数据探索性分析

## 2.1 数据分析与可视化

通过对数据进行探索性分析，将分析的数据结果通过图形画的方式展现出来，让用户更加直观的感受到数据的分布和一些其他信息。由于具体过程如下：

（1）年龄分析

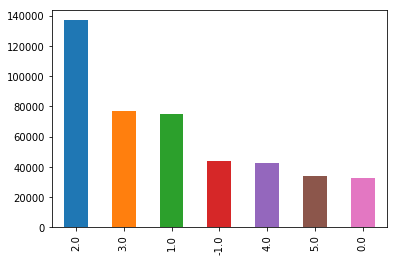


图2 年龄状况

从图2可以看出级别集中在2.0和3.0，对应的实际年龄分别为30-40岁和40-50岁之间，此外发现年龄存在-1，属于异常值，需要处理掉。

（2）性别分析



图3 性别状况

性别是按照0-1表示：0代表男性、1代表女性。从图3可以看出年龄多条为空值，在这里看到男女比例基本一样，我们可以将年龄缺失值随机用0,1填充或者众数填充，或者不选去性别这个特征。

（3）电话归属地



图4 电话归属地状况

电话归属地按照0-1表示：0代表浙江省、1代表其他省份，从图4可以看出两个类别之间的数量差异较大。

（4）省份分析

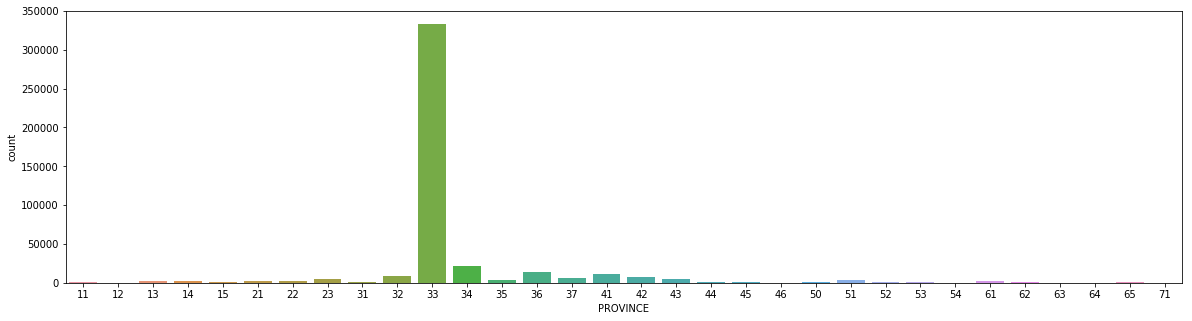


图5 省份分布状况

如图5所示，数字33代表浙江省份，其他数字为其他省份，可以看出购房用户主要分布在浙江省。

（5）国籍统计

表2 国籍状况

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 9 | 16 |
| 224307 | 47 | 15 | 216668 |

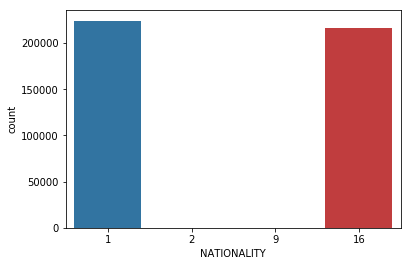


图6 国籍状况

如表2和图6所示，1代表中国，2代表美国，9代表法国，16代表特殊情况。可以看出除了大部分用户为中国籍之外，接近一半的用户国籍为特殊情况，法国和美国国籍较少。

（6）房屋属性

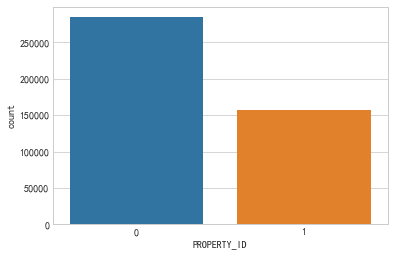


图7 房屋属性状况

如图7所示，房屋属性中0代表存量房，1代表商品房，其中存量房的数量明显多于商品房，这是一个区别性比较大的特征。

（7）房屋用途

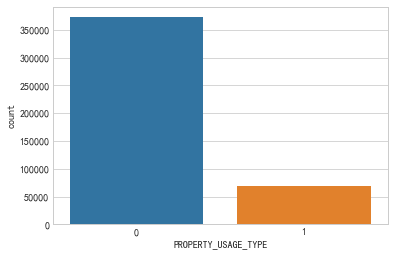


图8 房屋用途状况

如图8所示，0代表住宅，1代表非住宅，其中住宅房的数量明显多于非住宅的数量，这也是一个区别性比较大的特征。

（8）房屋建筑面积

表3 房屋建筑面积状况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 |
| 105665 | 136928 | 71606 | 67812 | 59026 |

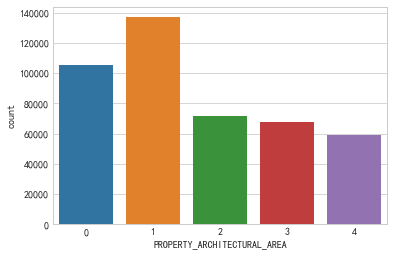


图9 房屋建筑面积状况

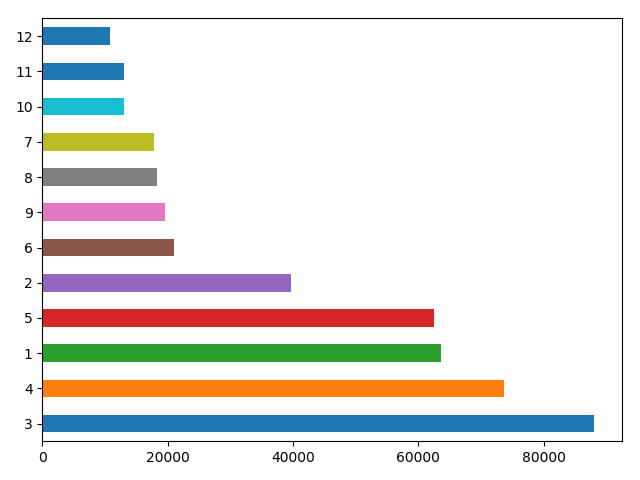
如表3和图9所示，0代表小于60平方米，1代表60-90平方米，2代表90-120平方米，3代表120-140平方米，4代表大于140平方米。可以看出房屋建筑面积主要集中在60-90平米。

（9）时间特征

抵押开始时间和抵押结束时间虽然有相距时间，但是超过99%的数据这两个特征为空，只有69个数据不为空，我们将这两个时间特征舍弃。

备案时间和网签时间两个时间相距为0，备案时间和网签时间同时进行的，且都是是2018年的数据，所以我们将这两个时间特征的月份（1-12月）作为离散特征处理，下面是新增特征“PROPERTY\_DATE”：

df[**'PROPERTY\_DATE'**]=df[**'PROPERTY\_SIGN\_DATE'**].apply(**lambda** x:x.split(**'/'**)[-1]).astype(**'int32'**)



（10）房屋成交总价

表4房屋成交总价状况

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 |
| 151489 | 133680 | 76311 | 36327 | 18350 | 24880 |

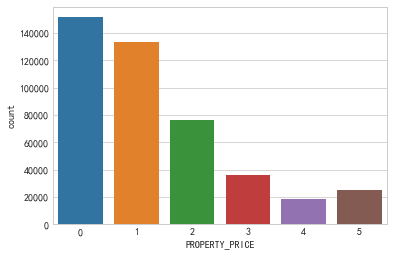


图10 房屋成交总价状况

房屋成交总价是一个枚举值，按照0-5表示：0代表100万以下、1代表100-200万、2代表200-300万、3代表300-400万、4代表400-500万、5代表500万以上。由表4和图10可以看到房屋成交总价基本上以中低价为主，高价房较少。

（11）贷款方式

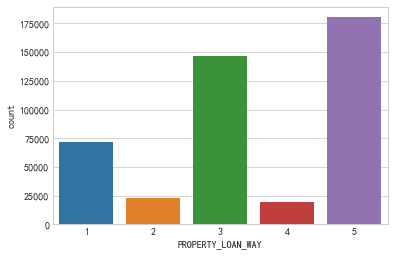


图11 贷款方式状况

如图11所示，贷款方式为枚举值，按照1-5表示：1代表组合贷款、2代表公积金贷款、3代表商业贷款、4代表银行贷款、5代表其他。其中贷款方式“其他”和“商业贷款”数量比较多。

（12）付款方式

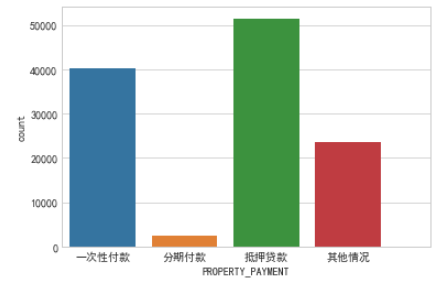


图12 付款方式状况

从图12可以看出，抵押贷款最多，其次是一次性付款，一次性付款的基本上可以认定为是信用比较良好的用户。

（13）查封次数

我们对查封次数做了一个统计，由图13所示，基本上有99%的用户不存在查封记录，存在查封记录的用户有3000人左右，可以认为这些用户是信用比较差的。

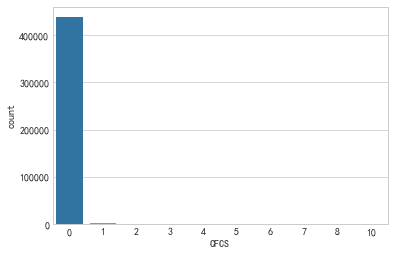


图13 查封次数情况

（14）解封次数

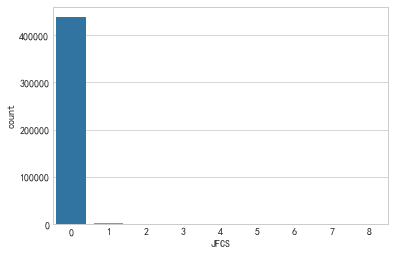


图14 解封次数情况

解封次数和查封次数基本上相对应，如果一个用户不存在查封记录，那么也不存在解封记录；如果一个用户存在查封记录，查封次数和解封次数基本上相等的。

（15）历史买房次数

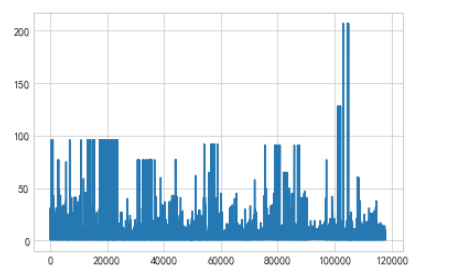


图15 历史买房次数

（16）历史卖房次数

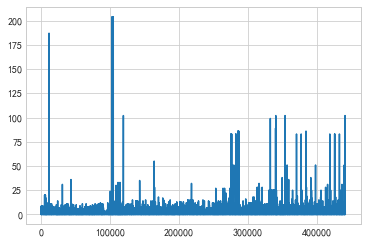


图16 历史卖房次数

## 2.2 特征工程

特征工程是使用专业背景知识和技巧处理数据，使得特征能在机器学习算法上发挥更好作用的过程。过程包含了特征提取、特征构建、特征选择等模块。其目的是筛选出更好的特征，获取更好的训练数据。因为好的特征具有更强的灵活性，可以用简单的模型做训练，更可以得到优秀的结果。为此，为了剔除数据指标的冗杂性、残缺性等方面所带来的误差，用特征工程对数据进行初步的预处理。具体过程如下：

（1）缺失值处理

*# 去除用户USER\_ID 为空的数据*df.dropna(subset=[**'USER\_ID'**], inplace=**True**, axis=0)  
*# 使用众数填充AGE的缺失值*df[**'AGE'**] = df[**'AGE'**].fillna(df[**'AGE'**].median())  
*# 使用0 填充Sale\_houcs/JFCS和CFCF的缺失值*df[**'Purchase\_houcs'**] = df[**'Purchase\_houcs'**].fillna(0).astype(**'int32'**)  
df[**'Sale\_houcs'**] = df[**'Sale\_houcs'**].fillna(0).astype(**'int32'**)

（2）One-Hot编码

在很多机器学习任务中，特征并不总是连续值，而有可能是分类值。但是，即使转化为数字表示后，上述数据也不能直接用在我们的分类器中。这个整数特征的表示并不能在分类器中直接使用，因为这样的连续输入，机器会认为类别之间是有序的，但实际却是无序的。(例如：浏览器的类别数据则是任意排序的)。使用one-hot编码，将离散特征的取值扩展到欧式空间，离散特征的某个取值就对应欧式空间的某个点。

将类别型特征使用one-hot编码，作用是让特征之间的距离计算更加合理。比如，有一个离散型特征，代表工作类型，该离散型特征，共有三个取值，不使用one-hot编码，其表示分别是x\_1 = (1), x\_2 = (2), x\_3 = (3)。两个工作之间的距离是，(x\_1, x\_2) = 1, d(x\_2, x\_3) = 1, d(x\_1, x\_3) = 2。那么x\_1和x\_3工作之间就越不相似吗？显然这样的表示，计算出来的特征的距离是不合理。那如果使用one-hot编码，则得到x\_1 = (1, 0, 0), x\_2 = (0, 1, 0), x\_3 = (0, 0, 1)，那么两个工作之间的距离就都是sqrt(2).即每两个工作之间的距离是一样的，显得更合理。

*# 处理类别标签 one-hot*cate\_cols = [**'TEL\_ID'**, **'PROVINCE'**, **'NATIONALITY'**, **'PROPERTY\_ID'**, **'PROPERTY\_USAGE\_TYPE'**, **'PROPERTY\_LOAN\_WAY'**,  
 **'PROPERTY\_PAYMENT'**]  
df = pd.get\_dummies(df, columns=cate\_cols)

（3）类型标签确定

我们使用CFCS作为类别标签，如果该用户存在查封记录，那么标签为1，如果不存在查封记录标签记为0。

df[**'CFCS'**] = df[**'CFCS'**].fillna(0).astype(**'int32'**) *# 类别标签1*df[**'CFCS\_Label'**] = df[**'CFCS'**].apply(**lambda** x: 1 **if** x != 0 **else** 0) *# 类别标签1*

需要注意的是，需要去除JFCS这一列，因为通过我们上面的数据分析发现，JFCS和CFCS存在强关联的关系。不去除该列的话，会导致模型过拟合，缺乏泛化的能力。

## 三、个人房产信用评分模型

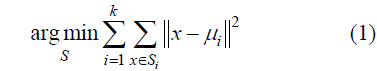
本部分模型建立在大数据处理的基础上，利用已有的数据指标进行模型构建评分。

## 3.1 基于K-means聚类分析的个人房产信用评估模型

**3.1.1 K-means原理**

K-means 聚类分析是一种非监督学习模型, 可以将数据对象分成若干子类。其基本思路是将n 个观测数据划分为k 个互斥簇, 划分结果可以作为风险分级的依据。K-means 方法将数据中的每个观察值视为具有空间位置的对象, 它找到一个分区, 其中每个簇中的对象尽可能地彼此靠近, 并且尽可能远离其他簇中的对象。分区中的每个簇由其成员对象的质心定义。每个簇的质心是该簇中所有对象的距离之和最小化的点。通过运用K-means 方法, 可以将数据对象分成若干簇（n=10）, 通过计算簇质心与某指定参考点(如原点)的距离, 可以确定该簇的信用等级，这里选取聚类中心到原地的距离。

具体地说, 已知观测集{x1, x2 ,…,xn} , 其中每个观测都是一个d 维实向量, K-means 聚类要将这n 个观测划分到k 个集合中(k≤n), 使组内平方和最小。即找到使满足公式(1)的聚类Si。

（1）

其中, μi 是Si 中所有点的均值。

K-means 聚类分析信用评估流程如图17所示:

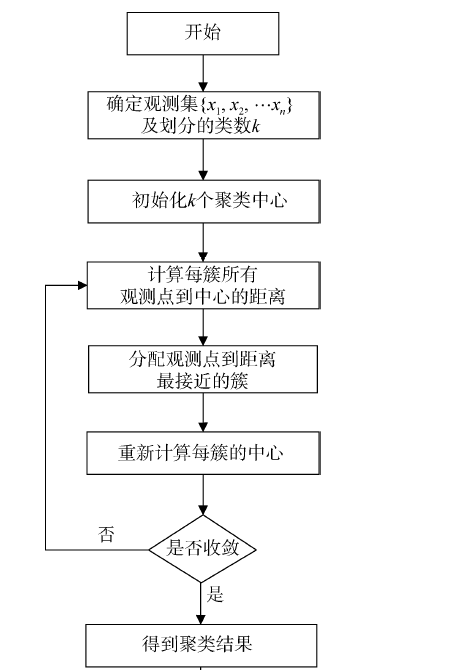


图17 K-means 聚类分析信用评估流程

**3.1.2 信用评分**

传统的个人信用评估模型对客户预测的结果要么是“违约”,要么是“不违约”，这种方式的个人信用评估模型在个人信用评估模型中占据主导地位.现实生活中, 未出现违约情况的人群中也存在违约可能, 而违约人群也有一定概率选择不违约, 鉴于此现象, 很多银行会对情况相似且有相似违规率的客户进行信用等级划分, 针对不同等级的客户实施不同的贷款政策。因此文中从实际情况出发, 提出个人信用度, 再将信用度转变成信用等级, 进而以信用等级作为新的个人信用评估标准, 不再仅仅输出“违约”、“不违约”作为评估标准。

首先先对个人信用度进行定义, 假设信贷个人信息相似情况下信贷客户的信贷违约概率是相同的,个人在贷款后采取不违约行为的概率就是个人信用度。文中的信用度是从相似群体的不违约概率延伸到个人不违约的概率。为了解决样本数据中没有信用度这一问题。通过k-means聚类方法来确定样本信用度, 假设聚类后的聚类个数为k, 其中第i(i≤ k)个聚类中的样本总数记为Ai, 未违约的个数记为Bi, 则第i类样本中的信用度就可以用公式Pi=Bi/Ai计算得到。通过信用度的计算公式可以得到Pi∈ [0, 1]，最终个人信用评分为Pi\*100。

## 3.2 基于机器学习或者深度学习预测

**3.2.1 SMOTE 处理类别不均衡**

大部分分类任务中，各类别下的数据个数基本上不可能完全相等，但是一点点差异是不会产生任何影响与问题的。

在现实中有很多类别不均衡问题，它是常见的，并且也是合理的，符合人们期望的。如，在欺诈交易识别中，属于欺诈交易的应该是很少部分，即绝大部分交易是正常的，只有极少部分的交易属于欺诈交易。这就是一个正常的类别不均衡问题。又如，在客户流失的数据集中，绝大部分的客户是会继续享受其服务的（非流失对象），只有极少数部分的客户不会再继续享受其服务（流失对象）。一般而已，如果类别不平衡比例超过4:1，那么其分类器会大大地因为数据不平衡性而无法满足分类要求的。因此在构建分类模型之前，需要对分类不均衡性问题进行处理。

  在前面，我们使用准确度这个指标来评价分类质量，可以看出，在类别不均衡时，准确度这个评价指标并不能work。因为分类器将所有的样本都分类到大类下面时，该指标值仍然会很高。即，该分类器偏向了大类这个类别的数据。

根据CFCS(查封次数)这个字段对其进行分类：

如果存在查封，该用户标签记为1，那么该用户是信用评分比较差的，这也符合常规；如果不存查封记录或者查封次数为0，该用户标签记为0，那么该用户信用理应比较高。比如我们模型预测一个用户为1的概率为0.8，预测为0的概率为0.2，那么该用户的评分可以概率px100=0.2x100=20。下面表格是标签统计结果：

|  |  |
| --- | --- |
| 0 | 1 |
| 437843 | **3192** |

为了处理类别不均衡问题，我们采用SMOTE算法。SMOTE全称是Synthetic Minority Oversampling Technique即合成少数类过采样技术，它是基于随机过采样算法的一种改进方案，由于随机过采样采取简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，即使得模型学习到的信息过于特别(Specific)而不够泛化(General)，SMOTE算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中，具体如图17所示，算法流程如下：

1. 对于少数类中每一个样本X，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其k近邻。
2. 根据样本不平衡比例设置一个采样比例以确定采样倍率N，对于每一个少数类样本X，从其k近邻中随机选择若干个样本，假设选择的近邻为。
3. 对于每一个随机选出的近邻，分别与原样本按照如下的公式构建新的样本。

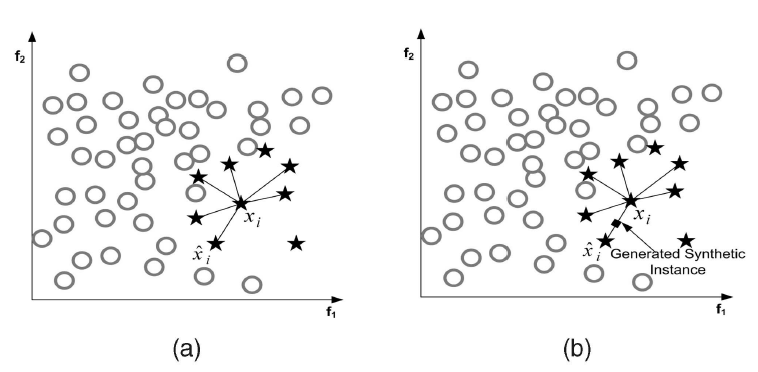


图17 模型的特征重要性

**3.2.2 构建XGBoost模型**

XGBoost算法属于集成学习中的boosting分支，其算法框架遵循1999年Friedman提出的boosting框架，该分支还有GBDT(Gradient Boosting Decision Tree)，boosting集成是后一个模型是对前一个模型产生误差信息进行矫正。gradient boost更具体，新模型的引入是为了减少上个模型的残差(residual)，我们可以在残差减少的梯度(Gradient)方向上建立一个新的模型。

该算法思想就是不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。当我们训练完成得到k棵树，我们要预测一个样本的分数，其实就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数，最后只需要将每棵树对应的分数加起来就是该样本的预测值。

模型我们采用XGBoost，是否存在查封记录作为Y标签，其他变量作为X，最终训练的模型的特征重要性如图18所示。

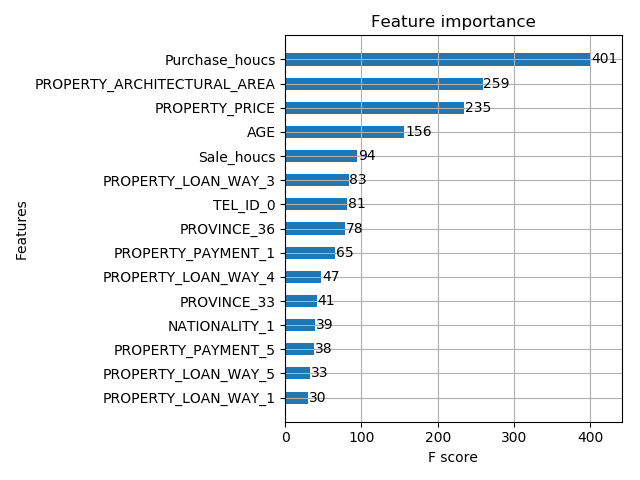


图18 模型的特征重要性

可以看出历史购房次数和购房面积、房屋价格对模型贡献率比较大。这也同时符合现实实际情况，如果一个用户多次购房，并且选择房屋面积不大，房屋价格较低进行倒卖，那么这个用户可能存在倒卖或者其他违规操作，也有可能以前存在查封次数或者违规操作，多次申请购房。

## 3.3 加权求和

因为我们K-means算法是将用户划分为10个等级，每个等级的评分范围是[10,100], 跨度比较大。同时与分类模型相比，考虑到K-means算法的可解释性差，为了不使加权求和的结果偏离分类评分，我们最终将K-means算法的评分结果权重定为0.3，分类评分结果定为0.7。

将上述两部分评分加权求和，具体是将聚类结果赋予0.3的权重，分类评分结果赋予0.7的权重，然后加权求和：

（2）

## 四、模型评估

## 4.1 评分模型评估

(1) Inertia

Inertia是指每个点到其簇的质心的距离之和，如果该值越小代表聚类效果越好,是对Kmeans算法的评估指标，计算公式如下：

https://ask.qcloudimg.com/http-save/yehe-1051732/x7ndrjzlvz.jpeg?imageView2/2/w/1620 （3）

在此案列中Inertia的值为4018905，以下是每个聚类中信用好（没有查封记录）的个数：

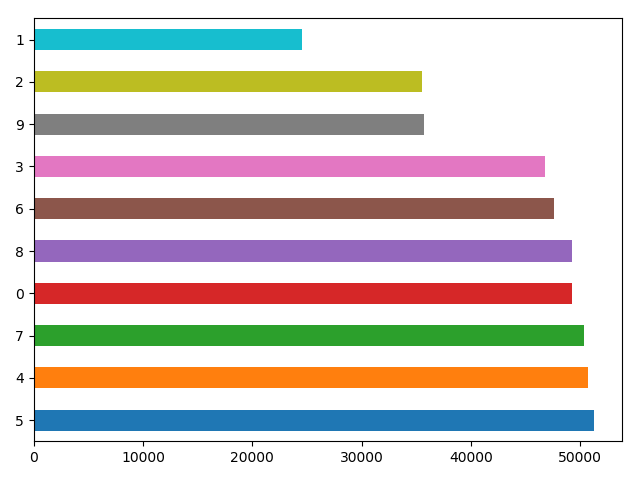


图19 各聚类效果的个数

由图19的统计结果可以得出结论就是，在现实生活中，犯错的人总是少数，守规矩的人总数多数。当其中如果一个聚类数量越多，那么就越代表该聚类用户信用越好。

最终我们将每个聚类的信用度乘以100，计算出用户的得分。

(2) AUC

AUC是一个模型评价指标，只能用于二分类模型的评价，对于二分类模型，还有很多其他评价指标，比如logloss，accuracy，precision。很多机器学习的模型对分类问题的预测结果都是概率，如果要计算accuracy，需要先把概率转化成类别，这就需要手动设置一个阈值，如果对一个样本的预测概率高于这个预测，就把这个样本放进一个类别里面，低于这个阈值，放进另一个类别里面。所以这个阈值很大程度上影响了accuracy的计算。使用AUC或者logloss可以避免把预测概率转换成类别。AUC量化了ROC曲线的分类能力，越大分类效果越好，输出概率越合理。

其中我们的Xgboost分类效果如下：

|  |  |
| --- | --- |
| Precision | AUC |
| 92.37% | 0.978 |

我们模型的准确率达到92%，AUC值高达0.97，可见构建的模型是一个泛化能力强的健壮模型。

## 4.2 模型预测结果

最终我们将聚类评分和分类模型评分加权求和，得到如下结果：

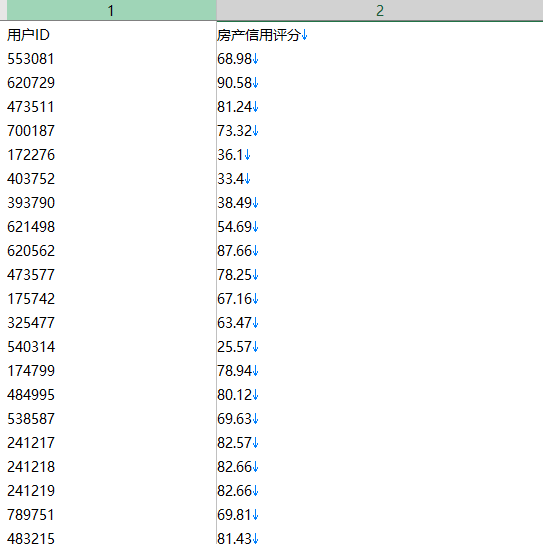


图20 房产信用评分情况

## 4.3总结与展望

由于比赛提供的数据以及相关指标的不完整性，我们针对此类状况仅对已有的指标做了相关的初步探索。大数据真正所潜在的价值在于其模糊性、混杂性、相关性可以通过合理的算法程序进行深度分析，找出其内在的关联，侧重相关关系的分析。就像沃尔玛通过对零售数据分析，找到飓风和蛋挞之间的联系，从而大大提高其销量。个人房产信用评价体系虽然和其差别比较大，但其对数据的挖掘和分析整理的原理是相同的。未来是大数据的时代，各行各业的发展都离不开对大数据的深入挖掘和分析探讨，基于本次大赛提供的数据，结合数据可视化分析，特诊工程，K-means 聚类分析模型、模型评估等完成了数据的处理、特征变量选择、模型开发与评估、信用评分卡和自动评分系统的创建。

该信用评分系统基于机器学习算法，通过Python语言进行创建，可通过数据的更替及算法的提升再次实现模型自建，同时用机器学习预测，使信用评分系统越来越强大。由于时间有限，团队所做成果不足之处很多，敬请各位专家评委指正。

## 五、团队介绍

队 长: 闫 强

中国人民大学在读研究生，信息学院经济信息管理系。研究方向为自然语言处理、金融知识图谱、量化分析等，负责并完成实验室企业知识图谱构建重大项目。在读期间参加多次数据科学比赛，利用机器学习或深度学习解决数据挖掘案例，曾获得神策杯top2%、研究生数学建模三等奖等， 本科期间多次获得学校奖学金。

指导导师：顾 杰

管理学博士、教授、美国北卡罗来纳大学博士后。现任浙江工商大学土地研究所副所长、浙江工商大学硕士生导师、中国高校房地产学者联谊会理事。2009当选浙江省人民政府“新世纪151人才工程”第三层次。

近期主要成果：

[1] 2013-主持国家社科基金“大城市保障性住房居住与就业、公共服务设施空间失配问题与治理途径研究”。

[2] 2008- 2010年主持完成教育部人文社科基金“城市空间增长中的土地、住房价格空间结构演变”。

[3] 2008- 2010年主持完成浙江省自然科学基金“城市空间增长中的土地增值机制与收益测算研究”。

[4] 2011年-2012年主持完成浙江省课题“浙江人口居住状况、住房需求与住房保障研究”。

[5] 2010-2011年主持完成慈溪市国土局课题“慈溪市乡镇土地利用规划修编”。

[6] 2010-2011年主持完成浙江省课题“浙江省房地产投资规模、效益分析。

成 员：张 辉

浙江工商大学在读研究生，土地与房地产方向。个人本科期间擅长文案写作、

商业策划，曾获创青春创新创业大赛湖北赛区铜奖、首届互联网＋创新创业大赛二三等奖、“京东杯”物流设计大赛湖北赛区一等奖等多项省级奖项。目前研究方向主要房产市场信用体系构建、房价增长相关性研究以及土地城乡规划等相关领域。

成 员：刘乐章

浙江工商大学在读研究生，土地与房地产方向。主要研究领域：住房经济与政策。研究生期间参与导师课题有：浙江省自然科学基金“转型期城镇家庭住房消费分化的基本特征、形成机制及其社会效应”，浙江省建设厅课题“浙江省培育房地产租赁市场研究”，嘉善县住建局课题“嘉善县住房租赁市场发展规划”。

## 六、参考文献及资料

[1]王志友.基于社交软件大数据分析的互联网金融征信研究[J].商情,2018,(43):101,151.

[2]袁庆春.个人信用重建的国际经验[J].中国金融, 2017，(23).

[3]赵靓.中国个人征信体系主要模式研究[J].中国市场, 2017，(14).

[4]宋清.论完善.中国个人信用制度[J].环球市场信息导报, 2017，(38).

[5]周荣技.《基于R语言的信用评分卡建模分析》.知乎

[6]iceForest1.《Kaggle|Give Me Some Credit信用卡评分建模分析（R语言）》.CSDN博客

[7]岱峰.《Give me some credit 机器学习练习》.知乎

[8]YoLean.《基于Python的信用评分卡模型分析》.简书

[9]Edwin Jarvis.《结合Scikit-learn介绍几种常用的特征选择方法》.博客园

[10]朱天星，于立新，田慧勇.商业银行个人信用风险评价模型研究[J].金融理论与实践，2011，380(3):64-67.

[11] Altmna E, Marco G and Varetto F.Corporate distress diagnosis: comparison using linear discriminate analysis and neural networks. Journal of Banking and Finance, 2010, 18(3): 505-529

[12] D. Molodtsov. Soft set theory - First results [J]. Computers & Mathematics with Applications, 1999, 37 (4/5): 19- 31.

[13]张成虎,李育林,吴鸣.基于判别分析的个人信用评分模型研究与实证分析[J].大连理工大学学报(社会科学版),2009,30(01):6-10.

[14]王晓慧,李云飞.判别分析和神经网络法的个人信用等级划分模型[J].内江师范学院学报,2018,33(02):64-68.

[15]戴德宝, 倪立平, 薛铭. 基于k-means和SVM的银行个人信用评估应用[J]. 江苏科技大学学报(自然科学版), 2017,(6):836-842