

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 -](#_Toc531541041)

[1.1背景 - 3 -](#_Toc531541042)

[1.2任务分析 - 6 -](#_Toc531541043)

[1.3研究现状 - 7 -](#_Toc531541044)

[1.4技术路线图 - 8 -](#_Toc531541045)

[二、团队介绍 - 9 -](#_Toc531541046)

[三、数据探索性分析 - 9 -](#_Toc531541047)

[2.1 数据分析与可视化 - 9 -](#_Toc531541048)

[2.2 特诊（征）工程 - 9 -](#_Toc531541049)

[四、个人房产信用评分模型 - 9 -](#_Toc531541050)

[3.1 基于传统的信用评分卡建模 - 9 -](#_Toc531541051)

[3.2 基于机器学习或者深度学习预测 - 13 -](#_Toc531541052)

[3.3 加权求和 - 13 -](#_Toc531541053)

[五、模型评估 - 14 -](#_Toc531541054)

[5．1赛题评分模型 - 14 -](#_Toc531541055)

[5.2模型 预测结果 - 14 -](#_Toc531541056)

[5.3总结与展望 - 14 -](#_Toc531541057)

[六、参考文献 - 14 -](#_Toc531541058)

# 一、个人房产信用评分介绍

## 1.1背景

**1.1.1大赛背景**

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。

**1.1.2国内外信用研究现状**

随着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。

**A、国外信用评价体系情况**

国外对信用评价体系的研究主要在评价方法、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，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。

**B、国内信用评价体系探索**

我国的信用评价体系发展较晚，采用了第三方评级机构和公共征信体系并行发展的模式。

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。

**C、当前信用评价体系存在的问题**

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。

**1.1.3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，以做房产市场交易评估之用。

## 1.2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，以此为参与房产交易的个人进行信用评分，以做业内参考之用。

## 1.3研究现状

依据赛题提供2017年1月-2018年5月购房者的房产交易数据与房屋状态数据。用赛题中以下20项相关数据指标为依据：

年龄 性别 电话归属地 省份 国籍

房屋属性 房屋用途 建筑面积 网签时间 解封次数

备案时间 贷款方式 付款方式 查封次数 历史卖房次数

抵押开始时间 房屋编号 房屋成交总价 抵押结束时间 历史买房次数

通过相关性分析对上述指标进行相关度进行排列，剔除关联度特别小的部分，进一步优化评价指标体系。另外考虑到指标的科学性、实用性，我们又采用文献阅读、专家交谈、房产行业从业人员咨询等方法来确定初始的指标集，在阅读相关文献和参考其他领域的信用指标体系后，初步建立指标集合。由于赛题提供的数据指标体系有限，在原有的数据指标体系中，我们参考银行的个人信用评级，又加入文化程度、职业、职务、个人年收入、家庭年收入等几个关键性指标。通过对房产从业人员的咨询，结合其从业经验，我们最终元数据中选出16个对其业务发展更为重要的相关指标：年龄、性别、省份、国籍、房屋属性、房屋用途、建筑面积、网签时间、备案时间、贷款方式、付款方式、历史卖房次数、历史买房次数、房屋成交总价、查封次数和解封次数。再加上文化程度、职业、职务、个人年收入和家庭年收入等四个关键性指标，共构成二十个指标体系。接下来采用层次分析法和德尔菲法计算各指标权重，构建评估模型采用评分表和模糊隶属度分析法计算定性指标和定量指标分值。最后通过抽取部分数据进行实证，并与商业银行评估分值作对比。具体的数据分析部分见下面的数据探索性分析。所构建评分指标体系如下：

表一：个人房产信用原始指标库

|  |  |  |  |
| --- | --- | --- | --- |
| 三维信用体系 | 一级指标 | 二级指标 | 指标分值说明 |
| 个人  基本  信息 | 身份特质(40%) | 年龄5 | 按照0-5表示：  0:20岁以下（1分）、1:20-30岁（3分）、2:30-40岁（5分）、3:40-50岁（4分）、  4:50-60岁（3分）、  5:60岁以上（2分） |
| 性别2 | 0:男（2分）、1:女（1分） |
| 文化程度6 | 1:初中以下（1分）、  2:中专或高中（2分）、  3:专科（3分）、  4:本科（4分）、5:硕士（5分）、6:博士（6分） |
| 个人收入及家庭收入10 | 数据暂无 |
| 职务9 | 数据暂无 |
| 职业8 | 数据暂无 |
| 其他相关信息(10%) | 国籍1 |  |
| 省份和  电话归属地6 | 两项指标具有很强关性:1：省内（6分） 2：省外（2分） |
| 房产  基本  情况 | 房产信息(20%) | 房屋属性8 | 0：存量房（6分）、1：商品房（8分）、 |
| 房屋面积5 |  |
| 房屋用途2 | 0；住宅（1分）、1：非住宅（2分） |
| 建筑面积5 | 0:小于60平方米1分、1:60-90平方米2分、2:90-120平方米3分、3:120-140平方米4分、4:大于140平方米5分 |
| 历史  交易  情况 | 房产履约状况(20%) | 房屋成交价格2 |  |
| 付款方式5 | 1:一次性付款（5分）  2:分期付款（4分）、  3:抵押贷款（3分）、  4：其他情况（2分） |
| 贷款方式3 |  |
| 查封次数5 |  |
| 解封次数5 |  |
| 抵押开始 |  |
| 抵押结束 |  |
| 信用行为活跃度(10%) | 历史买房次数6 |  |
| 历史卖房次数4 |  |

## 1.4技术路线图

# 二、团队介绍

队 长: 闫 强

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浙江省“新世纪151人才工程”第三层次，浙江省人民政府。

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、

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。

# 三、数据探索性分析

## 2.1 数据分析与可视化

数据可视化分析：是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。

## 2.2 特诊（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的作用的过程。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：

# 四、个人房产信用评分模型

本部分模型的建立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 基于传统的信用评分卡建模

信用评分技术是一种应用统计模型，其作用是对贷款申请人（信用卡申请人）做风险评估分值的方法。信用评分卡模型是一种成熟的预测方法，尤其在信用风险评估以及金融风险控制领域更是得到了比较广泛的使用。信用评分卡可以根据客户提供的资料、客户的历史数据、第三方平台（芝麻分、京东、微信等）的数据，对客户的信用进行评估。信用评分卡的建立是以对大量数据的统计分析结果为基础，具有较高的准确性和可靠性。为此对本赛题的相关数据做初步筛选后，通过信用评分卡建模对数据进行预处理。

3.1.1数据说明

数据来源于全球数据资源开发者大赛举办方提供算法赛赛题六中完整的房产交易数据与名下房屋查封情况数据，其中数据有44多万条，包含20个变量，具体如下字段（表2）：

表2;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 |

3.1.2数据获取及查看

执行如下源代码;

得到如下结果：

结论：

3.1.3数据预处理

a.缺失值处理

数据缺失情况在现实问题中非常普遍，这会导致一些不能处理缺失值的分析方法无法应用，因此，在信用风险评级模型开发的第一步就需要进行缺失值处理。缺失值处理的方法，包括如下几种：

（1） 直接删除含有缺失值的样本。

（2） 根据样本之间的相似性填补缺失值。

（3） 根据变量之间的相关关系填补缺失值。

变量MonthlyIncome缺失率比较大，所以我们根据变量之间的相关关系填补缺失值，采用随机森林法进行填补。而NumberOfDependts缺失较少，直接删除。

源代码如下：

b.异常值处理

缺失值处理后，需要处理异常值。异常值一般是指偏离数据较大的值。在统计学中，常把低于Q1-1.5IQR的值和高于Q3+1.5IQR的值作为异常值。通过绘制箱型图能很明显的看到异常值，例如

c.单变量探索性分析

在建立模型之前，一般会对现有数据进行探索性数据分析（Exploratory Data Analysis）。EDA是指对已有的数据(特别是调查或观察得来的原始数据)在尽量少的先验假定下进行探索。常用的探索性数据分析方法有：直方图、散点图和箱线图等。例如，我们对特征量年龄和月收入进行分析，如下。

3.1.4变量选择

（下面内容继续这样写，太过于详细繁杂，个人觉得应该是关键步骤需要提一下；整体缩短到3-5步为益）

3.1.5建模开发与评估

WOE/IV

WOE转换WOE(Weight of Evidence)叫做证据权重，IV(Information Value)叫做信息价值，是一组评估变量的预测能力的指标。也就是说，当我们想要拿出证据证明“年龄”这个变量对于违约概率是否有影响的时候，可以使用这个指标评估年龄到底对违约概率的影响有多大。

下面表格展示的就是年龄、性别及婚姻状况三个变量相关的好坏样本数据以及计算出的对应的WOE及IV值。WOE的计算公式是：ln[(违约/总违约)/(正常/总正常)]。比如对于年龄18~25的组别，WOE=In[(131/总违约样本数)/(1016/总正常样本数)]。根据WOE值，可以进一步计算出IV值。

3.1.6信用评分卡创建

根据参考文件所描述的，记录如下：

3.1.7自动评分系统创建

3.1.8总结及展望

相关理论的嵌套：

KS值模型能够帮我们挑选到最多的好客户，同时不要放进来那么多坏客户。KS值的取值范围是[0，1]。通常来说，值越大，表明正负样本区分的程度越好。一般，KS值>0.2就可认为模型有比较好的预测准确性。

2.ROC/AUC

ROC(receiver operating characteristic curve)，中文叫“接受者操作特性曲线”。

这里需要补充的基础知识是关于经典的混淆矩阵（confusion matrix）的内容。这个分析的经典之处在于，它超出了我们以往理解“正确率”的限制，使得我们有更多的维度去评价一个模型的预测能力。（哇喔我们已经不知不觉的涉及到了机器学习的内容）

以下的表格就是混淆矩阵的内容。怎么来理解呢？在使用一个模型去判断一个样本是“正样本”还是“负样本”的时候，模型会输出“正”或者“负”两种预测，叫做预测值；而实际上这个样本会有正样本或者负样本两种情况，称为实际值（类似我们利用模型预测“好人”及“坏人”）。预测值和实际值各有两种情况，这两种情况的交叉组合就形成了以下的混淆矩阵。

5.PSI

PSI(Population Stability Index)叫做群体稳定性指标，用于衡量两组样本的评分是否有显著差异。PSI = sum(（实际占比-预期占比）\*ln(实际占比/预期占比)

举个栗子，假设在训练一个评分模型时，我们将样本评分按从小到大排序分成10组，那么每组会有不同的样本数量占比P1；评分模型制作出来之后，我们试用这个模型去预测新的一组数据样本，按上面的方法同样按评分分成10组，每组也会有一定的样本数量占比P2。PSI可以帮助我们量化P1和P2，即预期占比与实际占比的差距。

## 3.2 基于机器学习或者深度学习预测

根据CFCS(查封次数)这个字段对用进行分类：

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

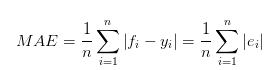
比如我们模型预测一个用户为1的概率为0.8，预测为0的概率为0.2，那么该用户的评分可以概率px100=0.2x100=20

## 3.3 加权求和

将上述两部分评分加权求和

# 五、模型评估

## 5.1赛题评分模型

本赛题用实际个人房产信用得分（风控模型评分）先行评价，指标维度如下：

其中，fi是预测值，yi是真实值，ei=|fi-yi|即是绝对误差。

## 5.2模型 预测结果

In[2]：data['label'] = clf.labels\_ *# 对原数据表进行类别标记*

data['label'].value\_counts()

In[2]：data['label'].value\_counts().plot(kind='barh')

plt.showd()

In[2]：ata['label']=data['label'].apply(lambda x:(10-x)\*9)

data[['USER\_ID','label']].to\_csv('submission.csv',index=None)

## 5.3总结与展望

a.总结

由于初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做了相关的初步探索。而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，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。

b.展望

# 六、参考文献

[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