消费者人群画像—信用智能评分

比赛真的太难了

胡敏  
 西南大学  
 中国-重庆

张楚瑜  
武汉大学  
中国-武汉

周洋洋  
 华中科技大学  
中国-武汉

匡祯辉

福州大学  
中国-福州

田冬雪

东北林业大学

中国-哈尔滨

团队简介

胡敏，西南大学研二，研究方向：数据挖掘，国家电投光伏预测比赛 TOP 1，融360第三届金融算法挑战赛 TOP 2，马上金融用户违约风险预测 TOP 5。

匡祯辉，福州大学研二，研究方向：数据挖掘，海上风场数据缺失智能修复 TOP 6。

田冬雪，东北林业大学研三，研究方向：遥感影像数据挖掘、分布式数据处理服务，国家电投“安全帽检测” TOP 5。

周洋洋，华中科技大学研一，研究方向：自然语言处理、数据挖掘。

张楚瑜，武汉大学大三，研究方向：计算机视觉、数据挖掘。

摘要

本篇文章，主要阐述“消费者人群画像-信用智能评分”比赛的思路。

数据挖掘的第一步是数据探索和数据预处理，我们也不例外。在数据预处理部分，我们首先处理一些错误数据，包括年龄、用户话费敏感度和缴费用户最近一次缴费金额（元），之后处理了一些离群数据，本次比赛中，离群数据有很多，包括，各类应用的使用次数、当月通话交往圈人数和用户当月账户余额（元）等，我们主要处理了特征重要性比较高的数据，主要是消费类特征和应用使用情况特征。探索数据的同时，我们也留意的训练集和测试集的数据分布，发现两者分布基本一致，只有个别离群点有些差异。

数据预处理完后，我们查了一些资料，理解字段的意思和业务意义，我们了解到，信用分主要由五部分组成，身份特征、消费能力、信用历史、人脉关系、行为偏好。我们主要从消费能力、行为偏好两个方面构造特征。

对于回归问题，我们常用MAE损失或者MSE损失，这里我们采用Huber损失。模型方面，我们尝试过线性回归、神经网络和树模型，最后，我们主要采用三种树模型，Lightgbm、XGBoost、CatBoost，五折交叉后融合三个模型的结果，对融合结果，进行分段取整处理，从而得到最后结果。

我们的模型容易解释，速度也比较快，可以部署工业应用。

关键词

数据预处理、业务意义、树模型

1、数据探索与数据预处理

**1.1赛题回顾**

赛题背景：

随着社会信用体系建设的深入推进, 社会信用标准建设飞速发展，相关的标准相继发布，包括信用服务标准、信用数据釆集和服务标准、信用修复标准、城市信用标准、行业信用标准等在内的多层次标准体系亟待出台，社会信用标准体系有望快速推进。社会各行业信用服务机构深度参与广告、政务、涉金融、共享单车、旅游、重大投资项目、教育、环保以及社会信用体系建设，社会信用体系建设是个系统工程，通讯运营商作为社会企业中不可缺少的部分 同样需要打造企业信用评分体系，助推整个社会的信用体系升级。同时国家也鼓励推进 第三方信用服务机构与政府数据交换，以增强政府公共信用信息中心的核心竞争力。

传统的信用评分主要以客户消费能力等少数的维度来衡量，难以全面、客观、及时的反映客户的信用。中国移动作为通信运营商拥有海量、广泛、高质量、高时效的数据，如何基于丰富的大数据对客户进行智能评分是中国移动和新大陆科技集团目前攻关的难题。运营商信用智能评分体系的建立不仅能完善社会信用体系，同时也中国移动内部提供了丰富的应用价值，包括全球通客户服务品质的提升、客户欠费额度的信用控制、根据信用等级享受各类业务优惠等，希望通过本次建模比赛，征集优秀的模型体系，准确评估用户信用分值。

赛题任务：

中国移动福建公司提供2018年x月份的样本数据（脱敏），包括客户的各类通信支出、欠费情况、出行情况、消费场所、社交、个人兴趣等丰富的多维度数据，参赛者通过分析建模，运用机器学习和深度学习算法，准确评估用户消费信用分值。

**1.2数据探索性分析和数据处理**

1.2.1 信用分分布探索

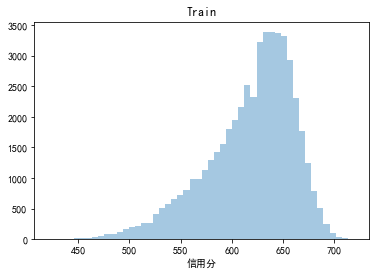


图 1信用分分布

由图1可以知道，信用分分布在[400, 750]之间，均值为618，中位数627，数据整体左偏。

1.2.2错误数据填充

在数据探索中，我们主要处理三类错误数据，分别是，用户年龄，话费敏感度，缴费用户最近一次缴费金额（元）。

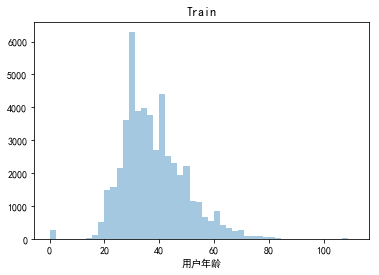


图 2 训练集中用户年龄分布

由图2可知，用户年龄出现等于零的错误数据，而且还有超过一百的情况异常数据，对于错误数据，我们采用了众数填充，对于异常数据，我们采用截断处理，将大于100的数据，截断。

在官方给出的用户话费敏感度说明表中指出，用户话费只有五个等级，而在数据中却出现六个等级，我们判断等级为零的为错误数据，我们使用众数填充。

缴费用户最近一次缴费金额（元）与用户最近一次缴费距今时长（月）有一些冲突。当用户最近一次缴费距今时长（月）为1时，代表这个月缴费了，那么缴费用户最近一次缴费金额（元）应该大于零，但是出现了等于零的情况，这很难解释，因此我们判断为错误数据，对于这个错误数据，我们是将它标记为nan，交给模型去处理。

1.2.3 离群数据处理

图3与图4给出，当月网购类应用使用次数、当月金融理财类应用使用总次数分布情况。

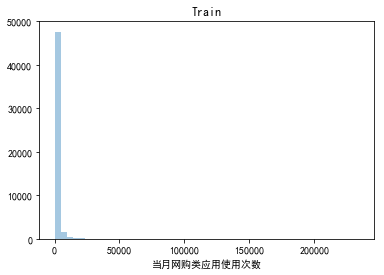


图 3 训练集中当月网购类应用使用次数分布



图 4 训练集中当月金融理财类应用使用次数分布

由统计分析可以知道，数据中存在离群点，同样情况的还有，当月飞机类应用使用次数、当月火车类应用使用次数、当月旅游资讯类应用使用次数、当月通话交往圈人数（如图5）等属性。



图 5训练集中当月通话交往圈人数分布

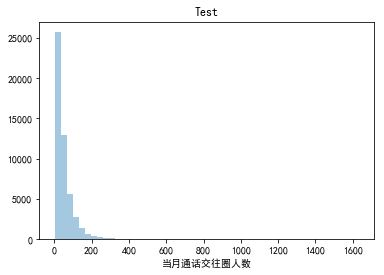


图 6测试集中当月通话交往圈人数分布

对于离群点，我们采用截断处理：



与分别是下限和上限，下限与上限通过数据的分位点确定。由上述几幅图，可知，数据呈右偏分布，为了将数据转化为正态分布，我们对截断后数据取对数，将右偏数据转化为近似正态分布。训练集的当月通话交往圈人数平滑后的分布如图7。

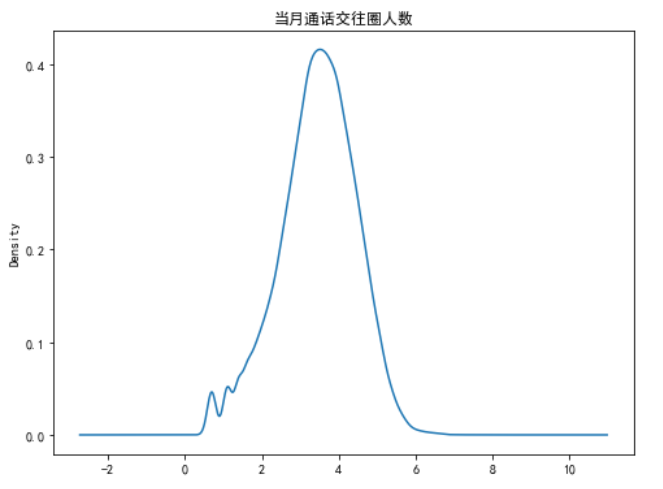


图 7 训练集当月通话交往圈人数平滑后分布

1.2.4相关性分析

下表为一些相关性分析。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 特征 | 用户网龄（月） | 用户近6个月平均消费值（元） | 当月通话交往圈人数 | 用户账单当月总费用（元） |
| 相关系数 | 0.547 | 0.494 | 0.477 | 0.442 |
| 特征 | 缴费用户最近一次缴费金额（元） | 当月是否景点游览 | 当月是否体育场馆消费 | 用户最近一次缴费距今时长（月） |
| 相关系数 | 0.334 | 0.271 | 0.233 | 0.215 |

表 1 部分属性相关性分析表

由表知，用户网龄（月）、话费属性、社交属性的相关性比较高，下面具体探索一下用户网龄和用户近6个月平均消费值（元）。

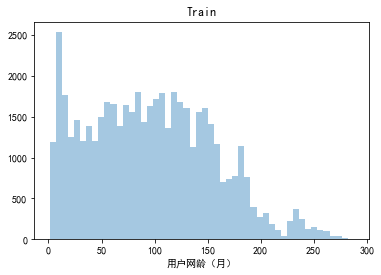


图 8 训练集中用户网龄（月）分布

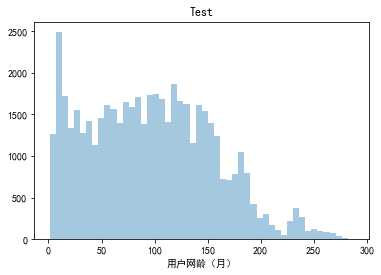


图 9测试集中用户网龄（月）分布

图8和图9，给出了训练集和测试集的用户网龄（月）分布，由图可以知道用户网龄分布主要分布在200以内，而且训练集的数据分布和测试集的数据分布基本相同。

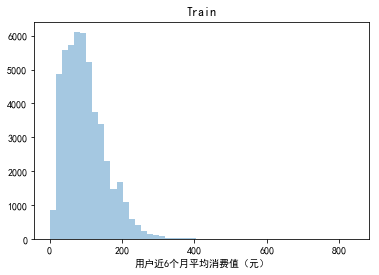


图 10训练集中用户近六个月平均消费值

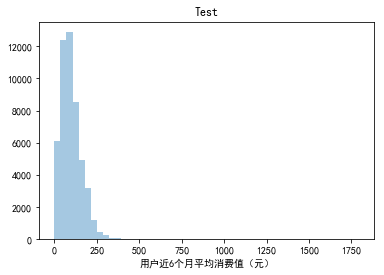


图 11 测试集中用户近六个月平均消费值（月）

由图10和图11知，训练集用户近六个月平均消费值（月）的最大值大于训练集用户近六个月平均消费值（月），这一点导致数据的分布有所差异，除去较大的值，其它数据分布基本相同。另外，从其它属性的比较中，可以知道训练集和测试集的数据分布基本相同。

1. 特征工程

从互联网上，我们找到了用户信用分评分构成部分，如下图：

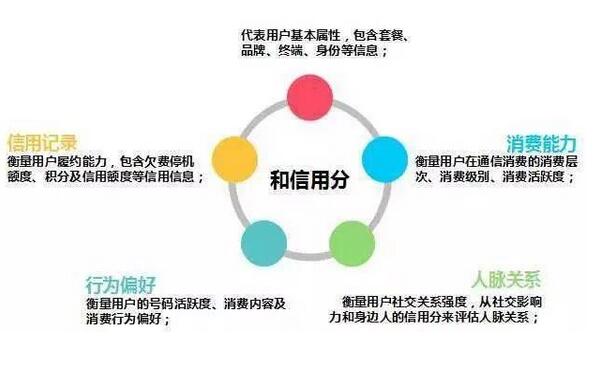


图 12 信用分组成部分

基于上述数据探索性分析和信用分评分构成部分，我们将从三个方面来构建我们的特征，分别是消费能力，行为偏好，基本属性。

**2.1消费能力**

2.1.1 消费等级划分

基于当月缴费金额，我们划分用户的消费等级，这样做，可以防止模型过拟合。此外，这样做也符合实际情况。

2.1.2 缴费类型特征

分析缴费金额可以知道，很多缴费金额是包含小数，根据现实情况，可以知道这是通过互联网缴费的，相反，小数位不为零的，则是营业厅充值的。这一特征，体现了用户的消费习惯。

2.1.3 话费类特征

话费类特征代指一类特征，它由原始话费特征生成，对模型的提升很大。

* 当月账单是否超过平均消费：用户账单当月总费用（元）- 用户近6个月平均消费值（元）
* 最近一次交费是否超过平均消费额：缴费用户最近一次缴费金额 - 用户近6个月平均消费值（元）
* 当月话费稳定性特征：用户账单当月总费用（元）/（用户近6个月平均消费值（元） + 1）
* 用户账单当月总费用（元）/用户当月账户余额（元）：用户账单当月总费用（元）/（用户当月账户余额（元） + 1）
* 用户近5个月平均消费值（元）：（用户近6个月平均消费值（元）\* 6 - 用户账单当月总费用（元）） / 5
* 用户6-5月平均消费差值（元）：用户近6个月平均消费值（元） - 用户近5个月平均消费值（元）
* 最近一次缴费后花费了多少：缴费用户最近一次缴费金额（元）-用户当月账户余额（元）
* 通话人均花费：用户账单当月总费用（元）/ (当月通话交往圈人数+1)

上述特征，有些相似。在不同的模型中使用都能达到很好的效果，特征的差异性也提高了模型的差异性，对后续的融合比较有益。

上述特征的构造过程可以分为两类，一类是原始特征的差值，一类是原始特征的比值（为了防止除零，分母加了一），这两种构造方法也是特征构造的基本方法，这些特征之间有些重复，也有些不同。例如用户前5个月平均消费值与用户近6个月平均消费值相似度很高，很多文献和实验[5]表明特征选择时去除相关性高的特征是有利的，但是在本问题中我们发现，虽然特征间相关性很高，但是这种做法可以提高准确度，我们认为这有可能与单个树模型在抽样特征的时候，重复特征会有更大的机率被抽到有关。

**2.2行为属性**

2.2.1 各类应用使用总和

将各类应用的使用次数加起来，得到各类应用使用总和。这一个特征反应了用户应用使用习惯。但是这个特征有一个缺点，就是应用使用次数的区间相差很大，会出现一类应用使用次数过多，导致其它应用的贡献为零。

2.2.2网购类应用使用次数比例

网购类应用使用次数代表了一个人的消费能力，这里用网购类应用使用次数与各类应用总和的比值，进一步表征用户的消费能力。这一特征比原始特征更具有表达能力，因为比值代表了两类的信息的综合。

2.2.3 交通类应用总和

原始特征中，包含当月飞机类应用使用次数、当月火车类应用使用次数，这两个特征表达的都是交通类信息，我们这里将两个特征求和，得到交通类应用总和。

2.2.3 布尔类型特征组合

原始特征中，包含很多布尔类型特征，例如当月是否逛过福州仓山万达、当月是否到过福州山姆会员店、当月是否看电影、当月是否景点游览、当月是否体育场馆消费、是否经常逛商场的人等，将这些特征组合起来，可以构造新的特征。由于组合比较多，我们只尝试了几种：

* 是否\_电影\_体育馆：当月是否看电影\*当月是否体育场馆消费
* 是否\_电影\_旅游：当月是否看电影\*当月是否景点游览
* 是否\_旅游\_体育馆：当月是否景点游览\*当月是否体育场馆消费
* 是否去过高档商场：当月是否到过福州山姆会员店+当月是否逛过福州仓山万达
* 是否商场\_体育馆\_电影\_旅游：是否去过高档商场\*当月是否体育场馆消费\*当月是否看电影\*当月是否景点游览

上述组合只是众多组合特征中几种，此外，我们这里没有涉及到其他布尔类型属性，例如是否大学生客户、是否黑名单客户等，主要是因为这样组合不具有可解释性，同类特征的组合，可以很好的解释。

**2.3基本属性**

这里基本属性的构造主要围绕用户网龄（月）和用户年龄。在数据预处理的时候，我们已经对用户年龄缺省值进行了处理，对离群点也截断处理了。下面利用这两个属性来构造新的特征。

2.3.1用户网龄（年）

原始特征中，用户网龄是按照月给出的，这里将用户网龄按照年来表达，使得老客户在评分是更具有优势。

2.3.2用户上网年龄

前一步得到了用户网龄（年），这里取用户年龄和用户网龄（年）的差值，构造用户上网年龄，这一特征与年龄特征比较相似。

2.3.3网龄十年

构造一个布尔类型特征，网龄大于十年为一，小于十年为零。这一特征表达了对老用户的友好。

1. 模型选择与调试

**3.1 Huber、MAE与MSE**

此外，对于回归问题，我们常用两种损失函数，MAE和MSE，公式如下。



其中，为模型预测值，为实际值。

对所有的观测数据，如果我们只给一个预测结果来最小化MSE，那么该预测值应该是所有目标值的均值。但是如果我们试图最小化MAE，那么这个预测就是所有目标值的中位数。中位数对于离群点比平均值更鲁棒，这使得MAE比MSE更加鲁棒。

使用MAE损失（特别是对于神经网络）的一个大问题是它的梯度始终是相同的，这意味着即使对于小的损失值，其梯度也是大的。这对不利于模型的学习。为了解决这个问题，我们可以使用随着接近最小值而减小的动态学习率。MSE在这种情况下的表现很好，即使采用固定的学习率也会收敛。MSE损失的梯度在损失值较高时会比较大，随着损失接近0时而下降，从而使其在训练结束时更加精确。图13和图14很好的展示了，两者训练时的区别。

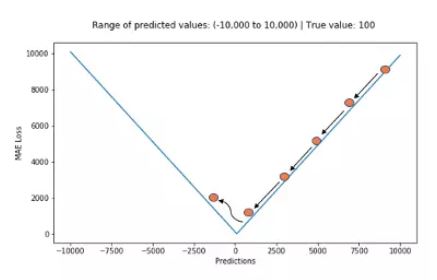


图 13 MAE训练过程图示

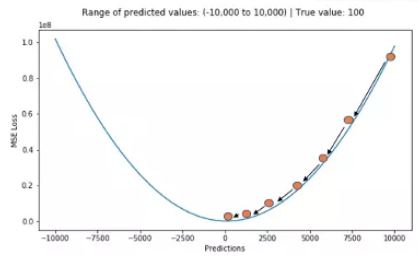


图 14 MSE训练过程图示

Huber 损失综合MSE和MAE优缺点，对两者折中处理。公式如下：



Huber损失对数据离群点的敏感度低于平方误差损失。超参数δ可调，当 δ ~ 0时， Huber损失接近MAE，当 δ ~ ∞（很大的数）时，Huber Loss接近MSE。

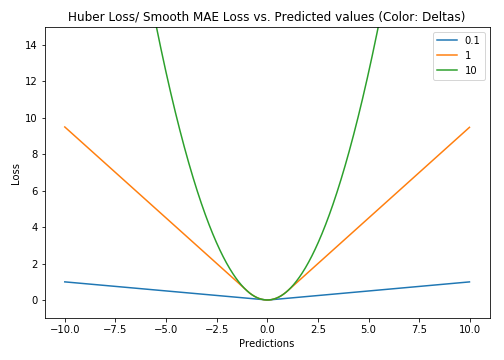


图 15 Huber损失随δ的变化

我们通过交叉验证来选择δ。

**3.2预测模型整体结构**

这场比赛我们采用树模型回归，主要使用Lightgbm、XGBoost、CatBoost，三个模型使用不同的特征进行训练，五折交叉得到三个结果，再对三个结果线性组合，得到组合结果。

组合结果是带小数的，我们尝试了四舍五入、向上取整和向下取整三种方式，最后，我们确定，对不同范围的分数做不同的处理。例如，对分数小于500的向上取整，对分数大于600的向下取整，对中间段的四舍五入取整。

****

图 16 模型整体结构图

**四、****总结与展望**

本次比赛，我们从多个角度构建特征，刻画消费者画像，从而给消费者评分。整体上，我们的模型具有两点业务价值：

* 模型的构造均为业务特征、模型的可解释性强！
* 模型的运行时间短，泛化能力较好，易于工业部署。

在比赛的前期，我们尝试过用神经网络来解决这个问题，但是效果不是很好，可能是数据的分布导致模型不收敛，后续就没有做过多尝试。此外，也尝试过将回归问题转化为分类问题，即将连续的信用分转换成离散的信用分，为了保证可靠性，我们尝试过转换为30分类问题，但是训练速度太慢，修改起来十分耗时，便放弃了。

总的来说，这次比赛收获很大。一步步的迭代，损失一步步降低，这也许就是数据挖掘的乐趣吧。

致谢

非常感谢福建省数字福建建设领导小组办公室、福建省工业和信息化厅、福州市人民政府、中国电子信息产业发展研究院、数字中国研究院和中国互联网投资基金和DF平台举办这样一个比赛，在比赛中，我们收获了友谊和知识，锻炼了能力，开拓了视野。会议名称：ACM伍德斯托克会议

会议简称：WOODSTOCK'18

会议地点：美国德克萨斯州埃尔帕索

ISBN：978-1-4503-0000-0 / 18/06

年份：2018

时间：6月

版权年份：2018年

版权声明：维护权利

DOI：10.1145 / 1234567890

RRH：F。Surname等。

价格：$ 15.00