**课程编号：A1901000221**

**《数据科学基础Ⅰ（Matlab）》**

**消费者人群画像-信用智能评分**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **姓名** |  | **学号** | |  |
| **班级** |  | **指导教师** | |  |
| **实验名称** | **消费者人群画像-信用智能评分** | | | |
| **开设学期** | **2021-2022秋季学期** | | | |
| **开设时间** | **第1周——第14周** | | | |
| **报告日期** | **2022年1月5日** | | | |
| **评定成绩** |  | | **评定人** |  |
| **评定日期** |  |

**东北大学计算机学院**

**目录**

[一、 项目名称 3](#_Toc755)

[二、项目背景 3](#_Toc10083)

[三、我的工作 3](#_Toc27388)

[四、数据集说明 4](#_Toc27746)

[4.1数据集名称 4](#_Toc9938)

[4.2数据集来源 4](#_Toc4617)

[4.3数据集介绍 4](#_Toc15743)

[4.4数据集字段说明 5](#_Toc32192)

[4.5数据集概览 6](#_Toc5956)

[五、实验目的 7](#_Toc5458)

[六、实验环境 7](#_Toc28427)

[七、实验步骤 8](#_Toc2850)

[八、实验过程 10](#_Toc3143)

[8.1数据集探索 10](#_Toc21773)

[8.1.1数据集获取和读入 10](#_Toc28872)

[8.1.2数据集基本信息 11](#_Toc27492)

[8.2数据预处理 14](#_Toc16983)

[8.2.1重复值处理 14](#_Toc6302)

[8.2.2缺失值处理 14](#_Toc13846)

[8.2.3异常值处理 15](#_Toc10165)

[8.3数据可视化 17](#_Toc13303)

[8.3.1分类变量分析-用户话费敏感度 17](#_Toc28231)

[8.3.2分类变量分析-用户最近一次缴费距今时长 18](#_Toc23235)

[8.3.3分类变量分析-用户实名制是否通过核实 19](#_Toc22776)

[8.3.4分类变量分析-是否经常逛商场的人 19](#_Toc4493)

[8.3.5连续变量分析-用户账单当月总费用 21](#_Toc5030)

[8.3.6连续变量分析-用户近6个月平均消费值 21](#_Toc31105)

[8.3.7连续变量分析-用户年龄分布分析 22](#_Toc29802)

[8.3.8相关性分析 23](#_Toc27506)

[8.4特征工程 24](#_Toc18303)

[8.4.1特征创造 24](#_Toc18071)

[8.4.2特征筛选 25](#_Toc19333)

[8.4.3特征标准化 25](#_Toc16970)

[8.4.4特征降维 26](#_Toc26812)

[8.5 数据集划分 27](#_Toc17110)

[8.6 决策树回归 27](#_Toc17189)

[8.7 高斯回归 29](#_Toc29073)

[8.8 线性回归 30](#_Toc1714)

[8.9 模型融合-基于回归树集成算法的stacking堆叠法 31](#_Toc6863)

[九、总结 32](#_Toc1514)

# 项目名称

**消费者人群画像-信用智能评分**

# 二、项目背景

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。现在通过使用人工智能技术，它将会从日常生活、学校、工作、休闲度假等多个维度来评定个人的信用度。每个人的社会活动将会被记录，例如如果你与一个人深度接触后再没有了联系，智能系统会认为你们双方兴趣爱好或者价值观不同，但是如果一个人深度接触过很多人之后，这些人后来都远离了他，这时智能系统会认为这个人的信誉很差。同样的道理，无论在学校和工作中任何的社交往来都将成为评定参考标准之一。

**而我们如何根据用户在生活中的各项日常生活中的行为表现，例如消费水平、贷款水平、日常前往商场的次数、以及购物次数等用户日常生活中的特征来衡量用户的信用分，这个问题就是我们需要考虑的问题。我们通过根据用户日常生活的行为表现给该用户的信誉进行一个评级，使用这个评级能够给很多行业带来便利，例如可以银行等金融行业一个很明确的衡量标准，决定是否放贷给该客户。**

# 三、我的工作

**在本项目中我首先使用Matlab对原始数据集进行了数据预处理，包括缺失值，重复值，异常值等特殊数据的处理。然后我使用了Matlab绘图的方法对原始数据进行了可视化处理。接下来我基于有标签的用户信用智能评分的数据，进行了特征工程实现了特征创造和特征降维。最后我使用了回归的算法构建消费者行为的评价模型，对他们的身份特征、消费能力、人脉关系、位置轨迹、应用行为偏好等特征与他们信用智能评分的分数的关联性进行预测。在算法求解过程中，使用了包括决策树回归算法、高斯回归算法和线性回归算法对问题进行求解，并且在决策树回归算法的建模求解过程中使用了交叉验证的方法通过绘制minleaf这个参数的学习曲线，对参数进行了调参。然后我还尝试了stacking堆叠法在上面算法的基础上再使用回归树集成算法进行模型融合。在最后的模型建模过程中我使用了RMSE对模型效果进行评价。**

# 四、数据集说明

## 4.1数据集名称

消费者人群画像-信用智能评分数据集

## 4.2数据集来源

https://www.datafountain.cn/competitions/337/datasets

## 4.3数据集介绍

数据的原始来源是中国移动福建公司提供2018年某月份的样本数据（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。该数据集包含了2个csv文件，train\_dataset.csv文件和test\_dataset.csv文件。其中train\_dataset.csv文件是用来给我们提供训练机器学习算法模型的原始数据文件，包含了某月中各个用户的上网行为信息及其信用评价的分数。数据集大小为5.43MB。整个数据集有50000条记录，用户编码，用户实名制是否通过核实，用户年龄，是否大学生落户，是否黑名单落户，是否4G不健康客户，用户网龄，用户最近一次缴费距今时长，缴费用户最近一次缴费金额，用户近6个月平均消费话费，用户账单当月总费用，用户当月账户余额，缴费用户当前是否欠费缴费，用户话费敏感度等30个特征，其中信用分作为反应变量，其他29个特征作为解释变量，在解释特征中有一个用来标识不同用户上网行为记录的编码。每条记录都由这30个字段组成，代表着一名用户在该月的行为特征及其获得的信用分数。

## 4.4数据集字段说明

数据集的字段说明如下表所示。

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名** | **含义** |
| 1 | 用户编码 | 数值 唯一性 |
| 2 | 用户实名制是否通过核实 | 1为是0为否 |
| 3 | 用户年龄 | 数值 |
| 4 | 是否大学生客户 | 1为是0为否 |
| 5 | 是否黑名单客户 | 1为是0为否 |
| 6 | 是否4G不健康客户 | 1为是0为否 |
| 7 | 用户网龄（月） | 数值 |
| 8 | 用户最近一次缴费距今时长（月） | 数值 |
| 9 | 缴费用户最近一次缴费金额（元） | 数值 |
| 10 | 用户近6个月平均消费话费（元） | 数值 |
| 11 | 用户账单当月总费用（元） | 数值 |
| 12 | 用户当月账户余额（元） | 数值 |
| 13 | 缴费用户当前是否欠费缴费 | 1为是0为否 |
| 14 | 用户话费敏感度 | 用户话费敏感度一级表示敏感等级最大。  根据极值计算法、叶指标权重后得出的结果，根据规则，生成敏感度用户的敏感级别：  先将敏感度用户按中间分值按降序进行排序，前5%的用户对应的敏感级别为一级：  接下来的15%的用户对应的敏感级别为二级；  接下来的15%的用户对应的敏感级别为三级；  接下来的25%的用户对应的敏感级别为四级；  最后40%的用户对应的敏感度级别为五级。 |
| 15 | 当月通话交往圈人数 | 数值 |
| 16 | 是否经常逛商场的人 | 1为是0为否 |
| 17 | 近三个月月均商场出现次数 | 数值 |
| 18 | 当月是否逛过福州仓山万达 | 1为是0为否 |
| 19 | 当月是否到过福州山姆会员店 | 1为是0为否 |
| 20 | 当月是否看电影 | 1为是0为否 |
| 21 | 当月是否景点游览 | 1为是0为否 |
| 22 | 当月是否体育场馆消费 | 1为是0为否 |
| 23 | 当月网购类应用使用次数 | 数值 |
| 24 | 当月物流快递类应用使用次数 | 数值 |
| 25 | 当月金融理财类应用使用总次数 | 数值 |
| 26 | 当月视频播放类应用使用次数 | 数值 |
| 27 | 当月飞机类应用使用次数 | 数值 |
| 28 | 当月火车类应用使用次数 | 数值 |
| 29 | 当月旅游资讯类应用使用次数 | 数值 |
| 30 | 信用分 | 数值 |

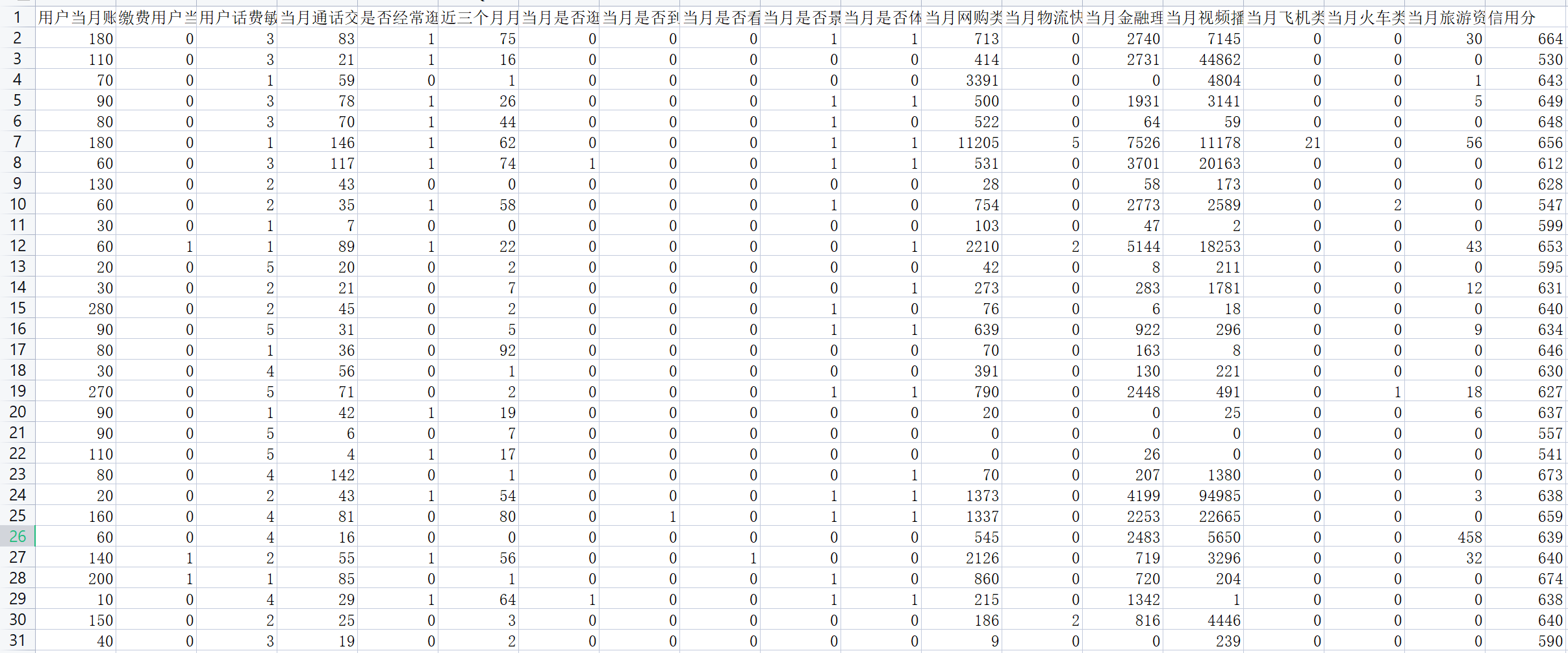
**表1 train\_data数据集各字段含义解释**

## 4.5数据集概览

为了方便查看数据集，先截取了数据集的一部份用来观察数据集的内容。截取train\_dataset.csv文件前50条记录如下图所示。



**图1 train\_dataset数据集概览一**



**图2 train\_dataset数据集概览二**

# 五、实验目的

* 学会如何进行数据分析的项目，拿到一个数据分析的项目后需要进行的数据信息的初步探索，数据预处理，数据可视化，特征工程，算法建模求解，模型评价等步骤。
* 学会针对项目需求进行分析，理清我们要解决的问题，设计合理的实验流程和合理的算法。
* 学会使用Matlab处理原始数据文件，并且能针对性的进行数据预处理。
* 学会使用Matlab处理异常值等特殊数据。
* 学会使用Matlab作图的方法进行数据可视化从而实现数据的探索性分析。
* 学会使用Matlab进行多项式回归等回归算法的设计和建模。
* 学会使用Matlab针对算法进行调参和评价。
* 学会如何使用Matlab执行线性回归、决策树回归、高斯回归、回归树集成等机器学习算法
* 学会模型融合的思想，如何使用stacking堆叠法将一些基本的学习器进行模型融合。

# 六、实验环境

MATLAB 版本：9.10.0.1602886 (R2021a)

MATLAB 许可证编号: 968398

操作系统: Microsoft Windows 10 家庭中文版 Version 10.0 (Build 19043)

Java 版本: Java 1.8.0\_202-b08 with Oracle Corporation Java HotSpot(TM) 64-Bit Server VM mixed mode

# 七、实验步骤

1. 数据的初步探索

* 数据集获取和读入
* 数据集基本信息统计

1. 数据预处理

* 重复值处理
* 缺失值处理
* 异常值处理

1. 数据可视化探索

* 用户话费敏感度分析
* 用户最近一次缴费距今时长分析
* 用户实名制是否通过核实分析
* 是否经常逛商场的人分析
* 用户账单当月总费用分析
* 用户近6个月平均消费值分析
* 用户年龄分析
* 相关性分析

1. 特征工程

* 特征创造
* 特征筛选
* 特征标准化
* 特征降维

1. 决策树回归

* 交叉验证
* minleaf调参
* 特征标准化
* 特征降维

1. 高斯回归

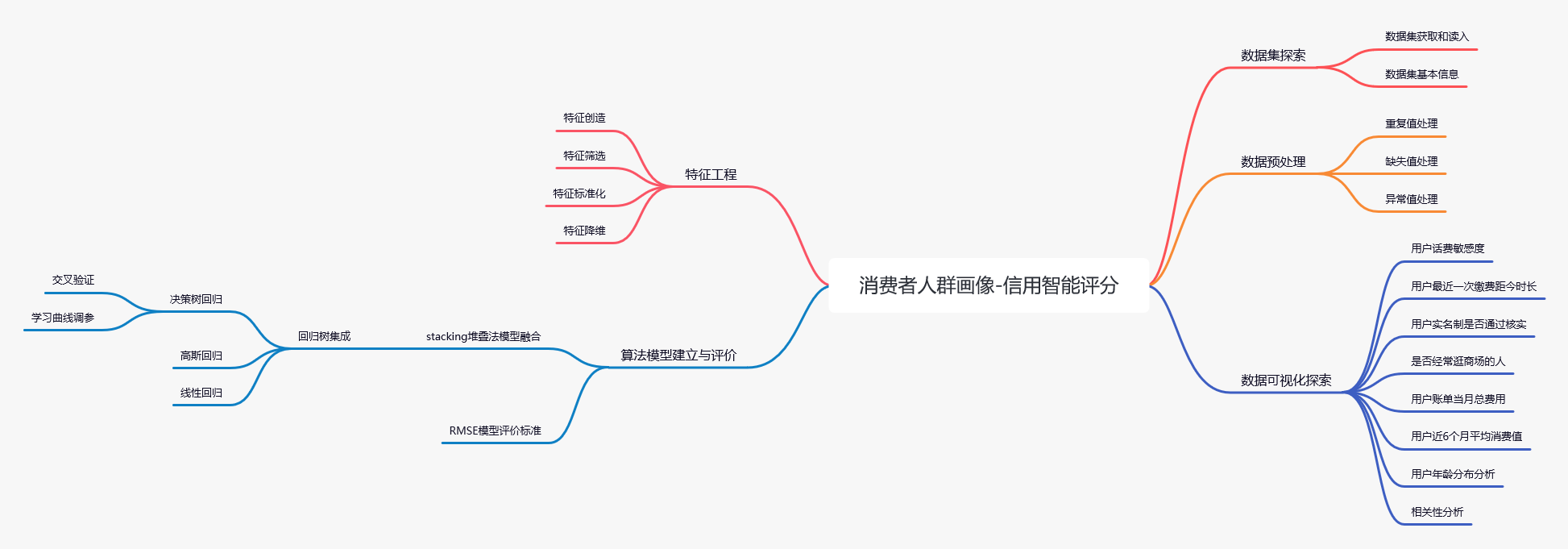
* 建模

1. 线性回归

* 建模

1. Stacking堆叠法模型融合

* 回归树集成算法



**图3 项目思维流程图**

# 八、实验过程

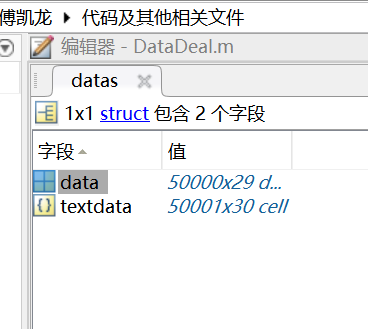
## 8.1数据集探索

### 8.1.1数据集获取和读入

DataFountain是国内领先的数据科学及人工智能协同创新平台，提供AI 竞赛/大数据竞赛、人工智能数据集、开源分享社区、AI实验室等服务。是CCF大赛官方指定竞赛平台。我在DataFountain平台上寻找到用于用户上网异常行为预测的数据集UEBA异常行为分析数据集，数据集来源网址：https://www.datafountain.cn/competitions/337/datasets。由于原始数据集的列名是带有中文的，所以我们无法使用csvread这个函数来读取csv的文件，所以我选择了importdata这个函数来读入原始csv文件中的数据。

1. datas = importdata("train\_dataset.csv");

读入完数据后我们可以看到datas的结构如下图所示。从中我们可以看出来datas是一个结构体，其中data保存了这个结构体的所有数据，而textdata保留这个结构体的特征名字，方便和数据一一对应上。



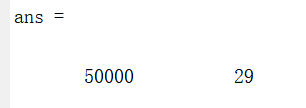
**图4 datas形状**

要想取出这个结构体的数据，我们可以使用getfield这个方法，通过这个方法能够指定结构体对象的名字，和要取出的那一部份数据的名字，从而得到要取出的数据。

1. data = getfield(datas, 'data');
2. name = getfield(datas, 'textdata');

### 8.1.2数据集基本信息

通过size函数查看数据集的大小：

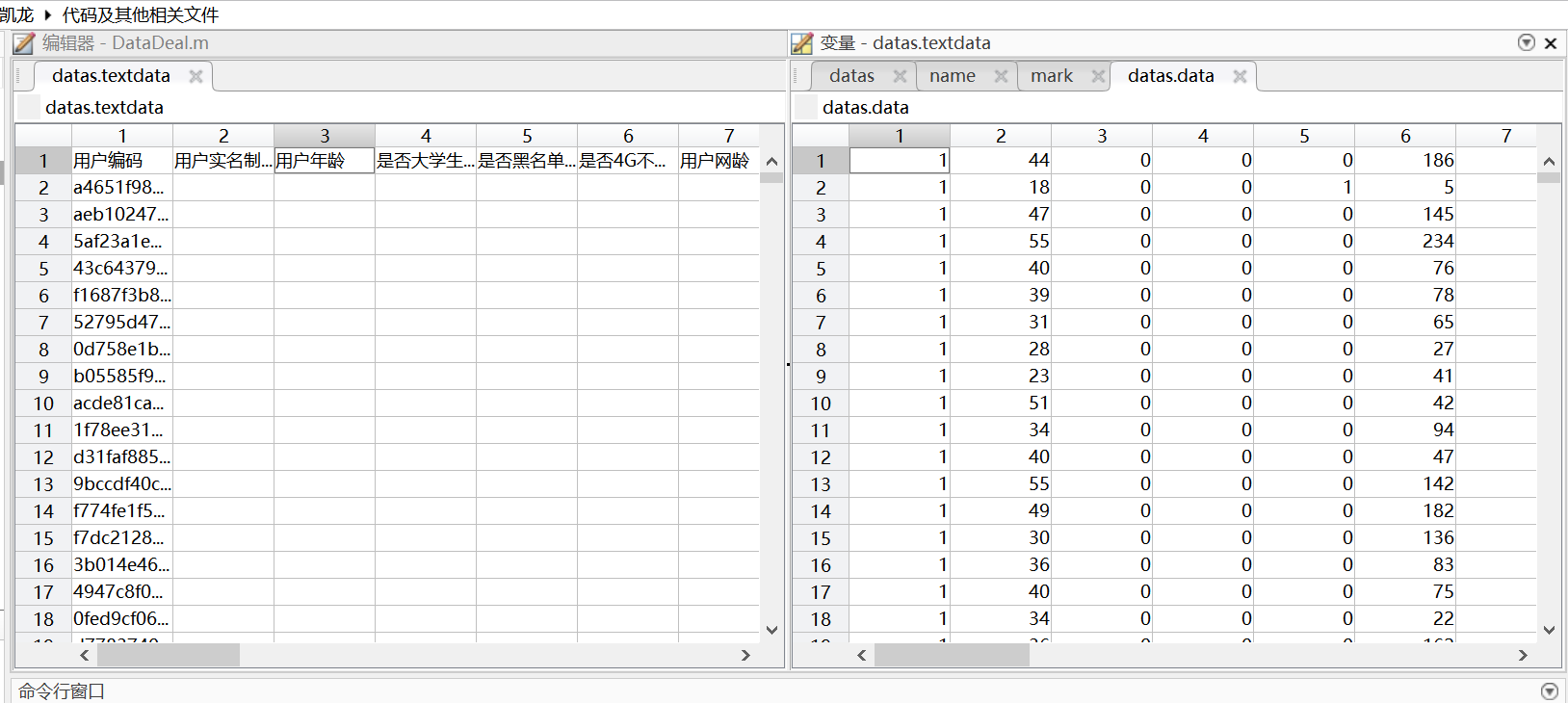


**图5 data形状**

train\_set数据集有50000行，29列。

我们知道偏度Skewness是描述数据分布形态的统计量，其描述的是某总体取值分布的**对称性**，简单来说就是数据的不对称程度：Skewness = 0，分布形态与正态分布偏度相同。Skewness >0，正偏差数值较大，为正偏或右偏。长尾巴拖在右边，数据右端有较多的极端值。Skewness < 0，负偏差数值较大，为负偏或左偏。长尾巴拖在左边，数据左端有较多的极端值。数值的绝对值越大，表明数据分布越不对称，偏斜程度大。峰度Kurtosis是描述某变量所有取值分布形态陡缓程度的统计量，简单来说就是数据分布顶的**尖锐程度：**Kurtosis = 0 与正态分布的陡缓程度相同。Kurtosis > 0 比正态分布的高峰更加陡峭——尖顶峰。Kurtosis < 0 比正态分布的高峰来得平台——平顶峰

鉴于Matlab里面并没有直接提供针对数据集描述性的函数，所以我编写了2个函数来针对数据集进行描述性统计。一个函数是dts，统计了对应列的所有数据的均值，方差，标准差，极差，变异系数，偏度，峰度。另外一个函数是fws，统计了对应列的所有数据的中位数，上下四分位数，四分位极差。



**图6 结构体数据的格式**

由上图6我们可以看出原始数据集中的列名和用来标记不同用户的用户编码字段的数据已经被保存到datas.textdata中，而纯数据保存在datas.data中。由于一方面列过于多，一次性对所有列进行描述性统计不方便观察。另外一方面我们要考虑到程序便于执行，用户不可能打开变量区去看统计的是第几列。所以我设计了input函数，让用户自己输入列名，再输出对应列的统计信息。

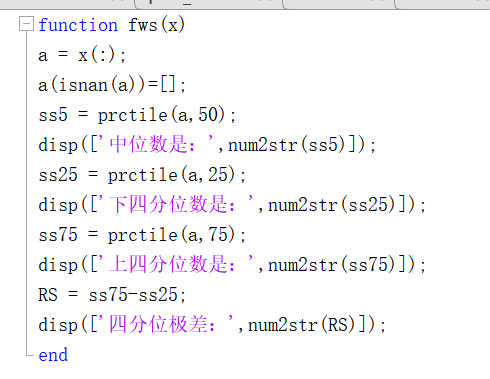
1. str = input('请输入你要进行查看统计信息的数据集列：','s')
2. mark = 0
3. for i = 1:30
4. if strcmp(str,name(1,i))==1
5. mark = i
6. end
7. end;
8. dts(data(:,mark-1))
9. fws(data(:,mark-1))

我设计的dts函数如下所示。



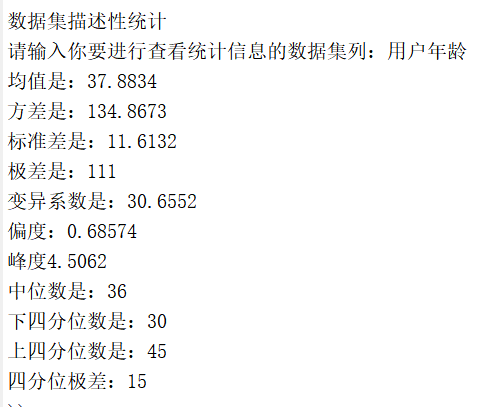
**图7 描述性统计-dts函数**

我设计的fws函数如下所示，其中针对如何求解分位数，我们可以使用prctile这个函数，这个函数。



**图8 描述性统计-fws函数**

以用户年龄这列为例子，我们对它进行描述性统计，统计结果如下所示。从用户年龄的几个分位数我们可以看出，有一半的用户年龄是小于36岁的年轻人，有3/4的用户年龄小于45岁。从偏度和峰度我们可以看出，用户年龄的分布是近似正态分布的，但是比正态分布更陡峭，会有一些年龄特别大的用户的极端值出现，这在后面的数据预处理中要除去极端值。



**图9 用户年龄的描述性统计结果**

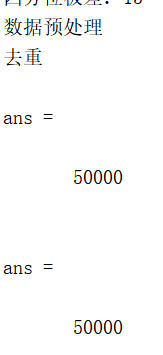
## 8.2数据预处理

### 8.2.1重复值处理

由于我把由于标识用户的唯一标记用户编码字段放在了name里，所以在data里面的只是纯数据，我们要去查看data中是否存在重复值，如果数据集中存在了大量的重复值会影响机器学习算法模型的计算和求解过程。在Matlab中我们可以使用unique这个函数，这个函数能直接得到原始数据去掉完全相同的行后剩下的数据。

1. size(data,1)
2. data=unique(data,'rows');
3. size(data,1)

通过去重前和去重后的data行数我们可以看出来，原始数据是有50000行数据，而新数据也有50000行数据，这说明针对原始数据中不存在重复的数据。



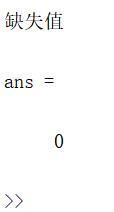
**图10 去重后的数据变化**

### 8.2.2缺失值处理

Matlab里面提供了isnan这个函数来判断数据的缺失情况。对于缺失值，我们可以使用删除法、替换法、填补法。

1. sum(isnan(data))

通过以上命令我们可以得到原始数据集中并不存在缺失值，所以不需要针对缺失值进行单独处理。



**图11 train\_set缺失值**

我们通过观察得到的结果可以知道原始数据集每一列都不存在缺失值，我们不需要进行缺失值的处理，填补或者删除缺失值。

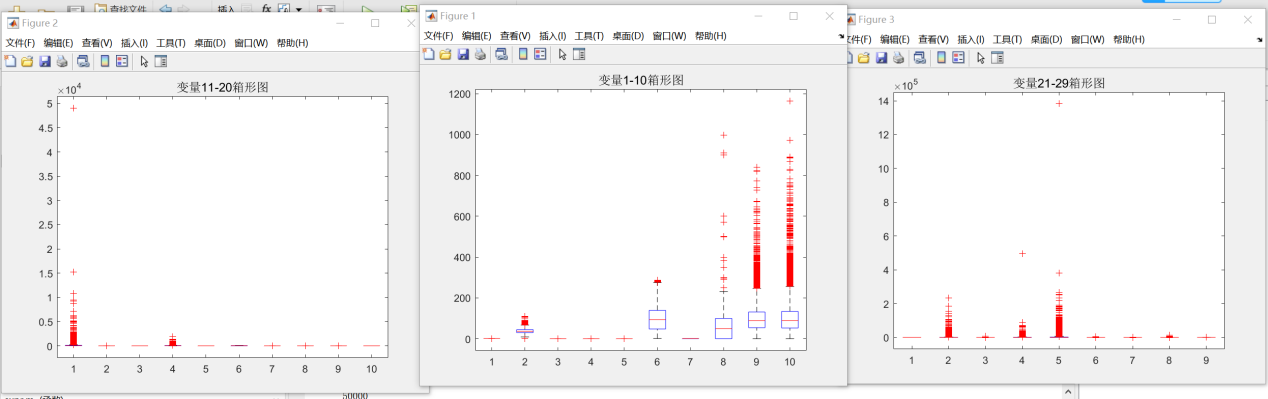
### 8.2.3异常值处理

异常值指的是数据极大或者极小，和正常情况下的数据分布情况偏离了太多，针对这种数据我们可以先使用箱型图对数据分布进行观察。Matlab中提供了boxplot命令让我们绘制箱型图。

由于原始数据有29个不同的特征，绘制在一张图上所以我绘制了3个不同的figure，每个figure上分别绘制10个特征。绘制箱型图的代码如下所示：

1. figure(1);boxplot(data(:,1:10))
2. title('变量1-10箱形图','fontsize',12)
3. figure(2);boxplot(data(:,11:20))
4. title('变量11-20箱形图','fontsize',12)
5. figure(3);boxplot(data(:,21:29))
6. title('变量21-29箱形图','fontsize',12)

绘制出的箱型图图像如下所示。

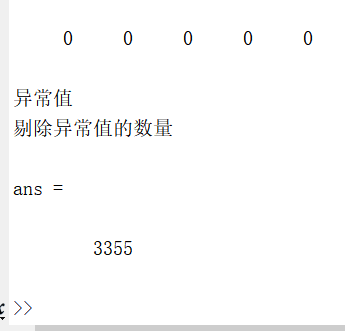


**图12 箱型图结果**

常见的针对异常值进行处理的方法都需要数据满足一定的分布，即正态分布或者类似正态分布。而有些方法需要使用作图工具箱而，所以我在接下来处理异常值的时候使用了肖维勒方法来剔除异常值。其基本思想是规定一个置信水平,确定一个置信限度,凡是超过该限度的误差, 就认为它是异常值,从而予以剔除。

1. [m n] = size(data);
2. Y = [];
3. w = 1 + 0.4\*log(m); % 肖维勒系数（近似计算公式）
4. for i = 1:n
5. x = data(:,i);
6. YiChang = abs(x-mean(x)) > w\*std(x);
7. Y(:,i) = YiChang;
8. end
9. [u v] = find(Y() == 1); % 找出异常值所在的行与列
10. ls = size(u,1);
11. uu = unique(u); % 剔除重复的行数
12. now = size(uu,1);
13. disp("剔除异常值的数量:");
14. size(uu,1)
15. data(uu,:) = [ ]; %令异常值所在行为空,即剔除异常值

我们可以查看统计结果，发现删除了3355个异常值。



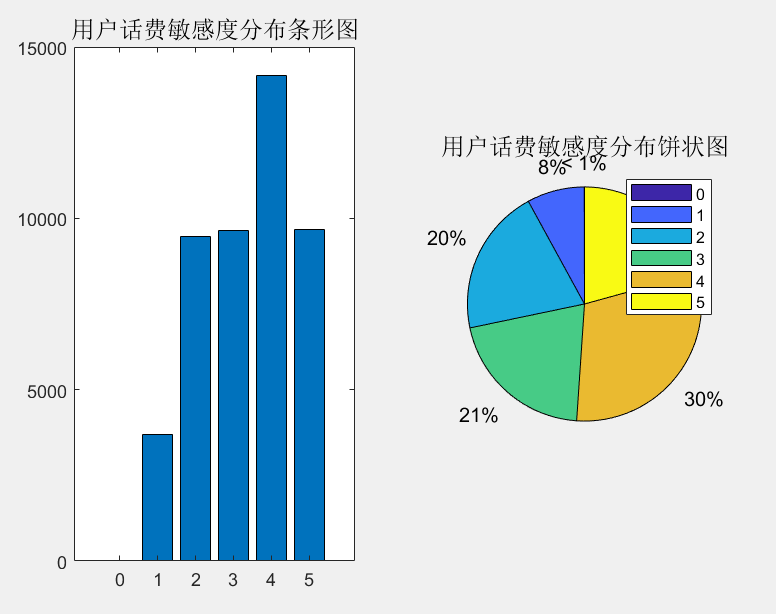
**图13 剔除异常值的数量**

## 8.3数据可视化

### 8.3.1分类变量分析-用户话费敏感度

用户话费敏感度是一个比较重要的特征，下面对它进行探索性分析。我们可以先使用Matlab提供的tabulate函数来统计离散变量的频数分布，这个函数能返回每个函数值和函数值对应的频数值。

1. target=data(:,29);
2. phoneprice=data(:,13);
3. result = tabulate(phoneprice(:));
4. lab = num2str(result(:,1));
5. figure(4)
6. subplot(1,2,1)
7. bar(result(:,1),result(:,2))
8. title('用户话费敏感度分布条形图','fontsize',12)
9. subplot(1,2,2)
10. pie(result(:,2))
11. title('用户话费敏感度分布饼状图','fontsize',12)
12. legend(lab);



**图14 用户话费敏感度饼状图、柱状图**

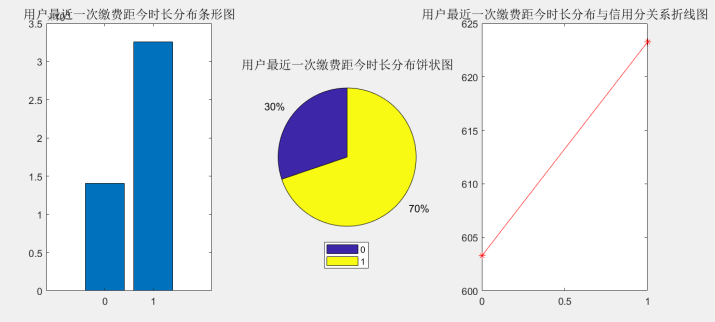
从对于特征名称的解读中可以很明显的发现，用户对于话费的敏感度这个特征可以较为直观地反映出用户当月的收入情况。如果用户对于话费很敏感的话说明该用户该月的收入肯定不高，收入不高会导致该用户该月的信用评分也不高。

从统计结果可以看出来用户话费敏感度较高的人的数量少，敏感度一级的人只占了8%，相对来讲敏感度较低的人数量更多，敏感度四级和五级的人占了50.9%。说明大部分人的月收入还是比较高的，对应的信用分也会较高。

### 8.3.2分类变量分析-用户最近一次缴费距今时长

用户最近一次缴费距今时长越短，说明用户缴费的意愿越高，该用户的收入情况肯定也会越好，其信用分的水平也会越高。我们在这个变量的可视化分析中，为了更进一步看出用户最近一次缴费距今时长这个变量和信用分之间的关系，我统计了用户最近一次缴费距今时长相同的用户的信用分的平均值。

1. disp("用户最近一次缴费距今时长（月）")
2. ms = [0,0];
3. timetonow=data(:,7);
4. for i=1:46645
5. if timetonow(i,1)==0
6. ms(1,1)=ms(1,1)+target(i,1);
7. end
8. if timetonow(i,1)==1
9. ms(1,2)=ms(1,2)+target(i,1);
10. end
11. end
12. result = tabulate(timetonow(:));
13. for i=1:2
14. ms(1,i)=ms(1,i)/result(i,2);
15. end
16. lab = num2str(result(:,1));
17. figure(5)
18. subplot(1,3,1)
19. bar(result(:,1),result(:,2))
20. title('用户最近一次缴费距今时长分布条形图','fontsize',12)
21. subplot(1,3,2)
22. pie(result(:,2))
23. title('用户最近一次缴费距今时长分布饼状图','fontsize',12)
24. legend(lab);
25. subplot(1,3,3)
26. plot(result(:,1),ms,'-\*r');
27. title('用户最近一次缴费距今时长分布与信用分关系折线图','fontsize',12);



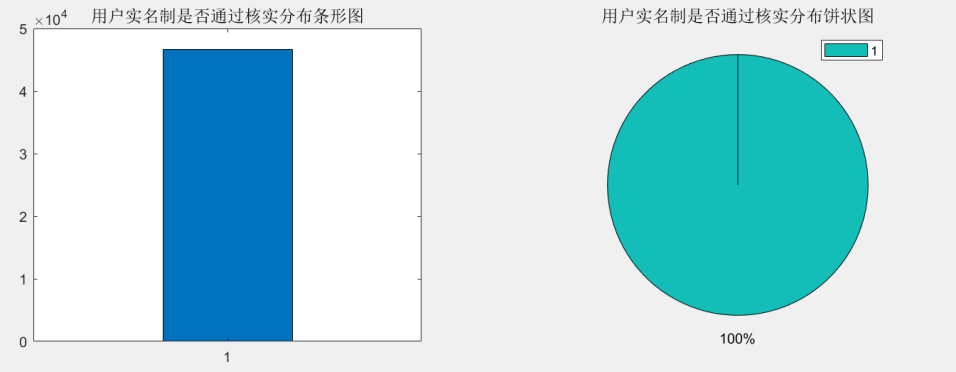
**图15 用户最近一次缴费距今时长可视化分析**

从可视化分析我们作出的图中可以看出来，有30%的用户最近一次缴费距今时长是0个月，有70%的用户最近一次缴费距今时长是1个月。但是实际上最近一次缴费距今时长越短的用户信用分是更低的，这不符合我们之前所做出的的推断，从结果图中我们可以看出来用户最近一次缴费距今时长和信用分是呈现反比关系。

### 8.3.3分类变量分析-用户实名制是否通过核实

用户实名制是否通过核实这个特征能够反映用户的信用安全情况，实名制通过认证的用户，这个账号的信用就将会和这个人绑定，如果用户在这个账号的信用表现不高的情况下，用户就将会进入失信人员名单，对个人生活带来巨大的不便的影响。

1. figure(6)
2. subplot(1,2,1)
3. bar(result(:,1),result(:,2))
4. title('用户实名制是否通过核实分布条形图','fontsize',12)
5. subplot(1,2,2)
6. pie(result(:,2))
7. title('用户实名制是否通过核实分布饼状图','fontsize',12)
8. legend(lab);



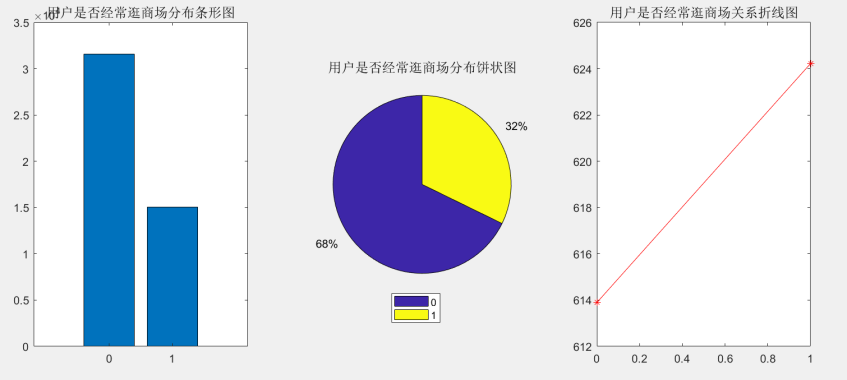
**图16 用户实名制是否通过核实可视化分析**

从可视化分析我们作出的图中可以看出来，所有用户都通过了实名制的核实，这是一个方差为0的变量，不对于区分用户信用分的大小贡献任何信息熵，属于无用值，对于我们机器学习算法建模没有任何帮助，所以我们在特征工程的特征降维中应该要对类似变量进行删除处理。

### 8.3.4分类变量分析-是否经常逛商场的人

用户如果经常逛商场，说明该用户的消费水平比较高，恩格尔系数较低。这种用户的抗风险的能力也较强，不容易发生信用违约的情况，所以这种用户的信用分一般都是较高水平的。

1. figure(7)
2. shop=data(:,15);
3. disp("用户是否经常逛商场")
4. ms = [0,0];
5. timetonow=shop
6. for i=1:46645
7. if timetonow(i,1)==0
8. ms(1,1)=ms(1,1)+target(i,1);
9. end
10. if timetonow(i,1)==1
11. ms(1,2)=ms(1,2)+target(i,1);
12. end
13. end
14. result = tabulate(timetonow(:));
15. for i=1:2
16. ms(1,i)=ms(1,i)/result(i,2);
17. end
18. lab = num2str(result(:,1));
19. subplot(1,3,1)
20. bar(result(:,1),result(:,2))
21. title('用户是否经常逛商场分布条形图','fontsize',12)
22. subplot(1,3,2)
23. pie(result(:,2))
24. title('用户是否经常逛商场分布饼状图','fontsize',12)
25. legend(lab);
26. subplot(1,3,3)
27. plot(result(:,1),ms,'-\*r');
28. title('用户是否经常逛商场关系折线图','fontsize',12)
29. disp("用户实名制是否通过核实")



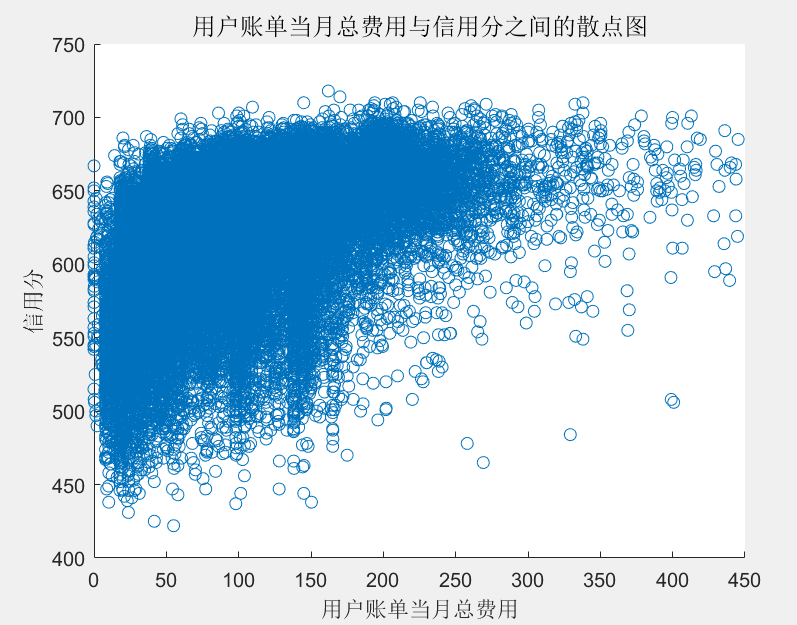
**图17 是否经常逛商场的人可视化分析**

从可视化分析我们作出的图中可以看出来，有32%的用户经常逛商场，有68%的用户不经常逛商场。与此同时经常逛商场的人的信用分越会越高，这符合我们之前针对逛商场的推断。是否经常逛商场与信用分之间是呈现正比的关系。

### 8.3.5连续变量分析-用户账单当月总费用

用户账单当月总费用也是一个重要的特征。不同的人具有不同的消费能力，往往消费能力高的用户的收入水平也较高，它具有比较高的可支配财富，所以他抵御风险的能力也较强。即用户账单当月总费用越高的人收入也应该越高，信用分也应该越高。

1. figure(8)
2. shop=data(:,10);
3. scatter(shop,target)
4. xlabel('用户账单当月总费用')
5. ylabel('信用分')
6. title('用户账单当月总费用与信用分之间的散点图','fontsize',12);



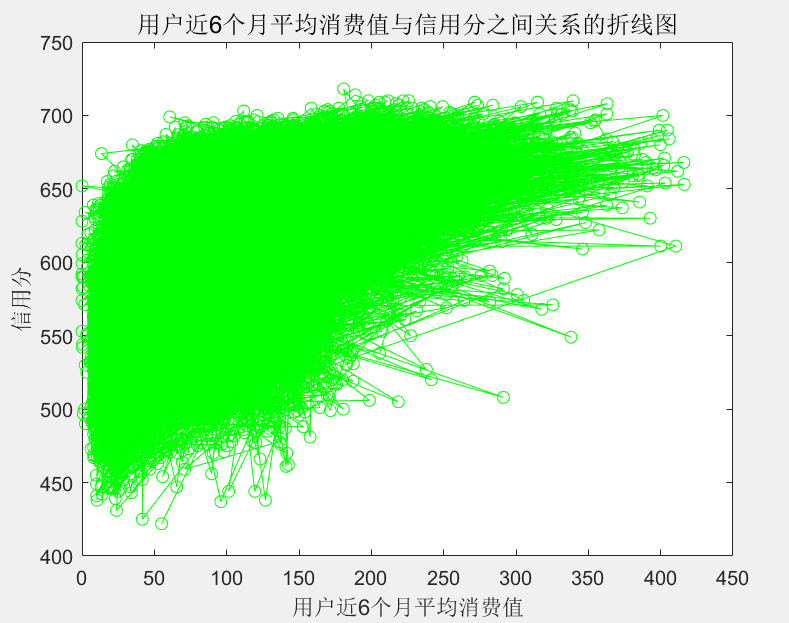
**图18 用户账单当月总费用可视化分析**

根据作出来的用户账单当月总费用分布散点图可以看出绝大部分的用户账单当月总费用金额在小于400元，用户账单当月总费用大于400元的用户少之又少不占大头。从结果可以看出来绝大部分的人收入还不是特别高，对应的信用分数也大概率偏低。图中分布呈现了明显的上三角形的分布情况，这说明用户账单当月总费用一定的情况下，用户的信用分大概率会大于用户账单当月总费用对应的某个确定的值，这也是一种很明显的相关关系。

### 8.3.6连续变量分析-用户近6个月平均消费值

用户近6个月消费值也是一个重要的特征。不同的人具有不同的消费能力，往往消费能力高的用户的收入水平也较高，它具有比较高的可支配财富，所以他抵御风险的能力也较强，相比于用户当月账单总费用，近6个月消费值对应的时间周期更长，反映的情况也更稳定，信用分也应该越高。

1. figure(9)
2. shop6=data(:,9);
3. plot(shop6,target,'-og')
4. xlabel('用户近6个月平均消费值')
5. ylabel('信用分')
6. title('用户近6个月平均消费值与信用分之间关系的折线图','fontsize',12);



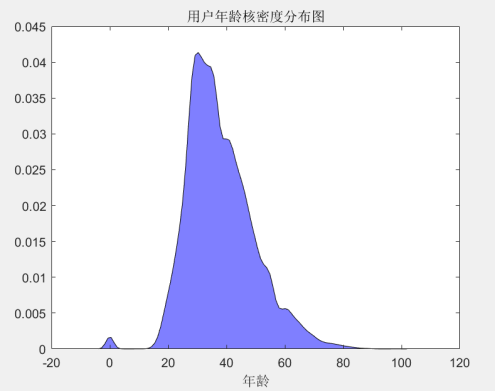
**图19 用户近6个月平均消费值可视化分析**

从结果我们可以看出来用户近6个月平均消费值与信用分之间的关系和用户账单当月总费用与信用分之间的关系类似，都是呈现正比的关系。

### 8.3.7连续变量分析-用户年龄分布分析

用户年龄分布也是一个十分影响结果的变量，对于年轻人来说，因为刚刚步入社会，容易遇到各种各样的生活压力，而这个时候他们自己还没有稳定的工作，所以十分容易发生债务违约的现象，但是对于中年人和老年人来说，他们已经成家立业，步入社会，遇到生活压力时候的抗风险能力也会越强。所以接下来我对用户年龄分布进行了分析。在Matlab中提供了ksdensity函数来让我们绘制核密度图，核密度图显示年龄的概率分布，能直观展现数据分布情况。

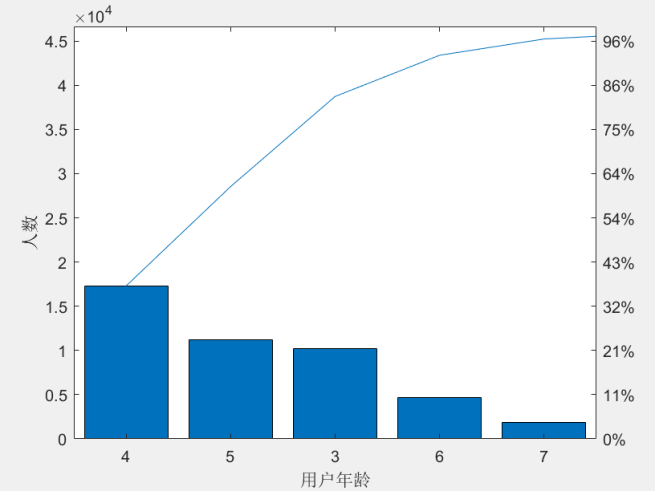
1. figure(11)
2. [f1, x1] = ksdensity(year);
3. t1 = area(x1, f1);t1.FaceColor='b'; t1.FaceAlpha=0.5;
4. title('用户年龄核密度分布图');
5. xlabel("年龄");



**图20 用户年龄可视化分析一**

我们可以把年龄按照10岁为一个区间划分成十个区间，这样能更方便的观察到结果。划分成十个区间后，我们可以作出帕累托图，观察数据的累计分布的情况。在Matlab中我们可以使用pareto函数作出帕累托图。我们在按年龄段统计用户的数量的时候可以直接一次搞定，不需要出现10个if，因为用户的年龄数据除以10后取值得到的结果，就是对应以10为单位的年龄段的下标，直接把下标对应的数加1就行。

1. figure(10)
2. year=data(:,2);
3. result = zeros(1,10);
4. for i=1:46645
5. result(1,floor(year(i,1)/10)+1)=result(1,floor(year(i,1)/10)+1)+1;
6. end
7. pareto(result);
8. xlabel('用户年龄')
9. ylabel('人数');



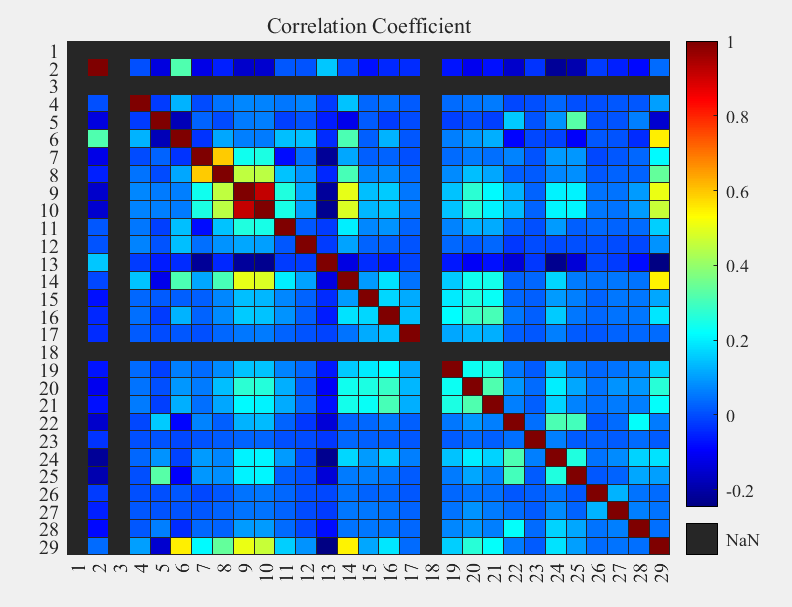
**图21 用户年龄可视化分析二**

从两个针对年龄的可视化分析我们能够看出来，用户中的年龄分布最大的是30-40岁这一个年龄段，其次是40-50岁，这两个年龄段的用户应该都成家立业了，会有比较好的抗风险的能力，对应的信用分应该也是较高的。

### 8.3.8相关性分析

为了直观地得到不同变量之间的关系，我们可以使用相关系数热力图来观察变量之间的相关系数。在Matlab中我们可以使用corr求得相关系数，然后使用colormap绘制出热力图。

1. figure(12)
2. rho = corr(data);
3. string\_name={'1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17','18','19','20','21','22','23','24','25','26','27','28','29'};
4. xvalues = string\_name;
5. yvalues = string\_name;
6. h = heatmap(xvalues,yvalues,rho, 'FontSize',10, 'FontName','Times New Roman');
7. h.Title = 'Correlation Coefficient';
8. colormap(jet)



**图22 数据集中各个特征之间相关系数的热力图**

为了研究自变量和解释变量这些不同变量之间的相关性，用Person相关系数量化计算变量之间的相关性。从上面作出的Person相关系数热力图中可以看出信用分与是否4G不健康用户、缴费用户最近一次缴费金额（元）、用户近6个月平均消费值（元）、用户话费敏感度等数据的关系最大。具有很强的正相关性，所以我们在建模的时候应该着重考虑这些字段的数据。

## 8.4特征工程

### 8.4.1特征创造

在本问题中根据有些关于欠费和账单的特征能进行符合实际生活场景的特征创造，生成新的特征。

对于缴费用户最近一次缴费金额和用户近6个月平均消费值这两个特征。我的理解是对于用户来说，如果用户缴纳的费用比用户账单金额高，说明该用户收入水平比较高足以付清该用户的账单。所以进行特征创造的时候可以使用缴费用户最近一次缴费金额减去用户近6个月平均消费值作为一个新的特征字段用来代表用户的收入水平支付当月消费的账单的能力，这个值是正数代表用户该月收入足以偿还当月的账单，这个值如果是负数代表用户该月的收入不足以偿还当月消费的账单。而且这个只越大代表用户偿还账单的能力越强，可以反映出该用户的信用情况是越好的。

上面的特征创造我们是基于用户的收入和花销间的关系，除此之外我们可以考虑用户的花销本身。如果用户的花销越来越小，说明用户违约的可能性也越来越小。基于次我考虑使用用户账单当月总费用减去用户近6个月平均消费值作为一个新的特征字段用来代表用户花销的变化情况。这个值越小代表用户的花销的变化情况是越来越小的。

进行特征创造的代码如下所示，在进行特征创造时我使用了向量化编程的思想方法，不使用for循环去遍历46635行的原始数据，从而提高程序运行效率：

1. newFea = zeros(46645,1);
2. newFea(:,1)=data(:,8)-data(:,9);
3. data = [data newFea];
4. newFea(:,1)=data(:,10)-data(:,9);
5. data = [data newFea];

### 8.4.2特征筛选

有些字段的数据是完全一样的这对于建立机器学习模型没有任何帮助，反而会加重机器学习算法模型的负担。比如说我们之前讨论过的用户实名制是否通过核实字段，这个字段的数据都是1，并不会对区分不同用户的信用分带来任何帮助，所以我们去除该字段的数据。

1. data(:,1) = [];

### 8.4.3特征标准化

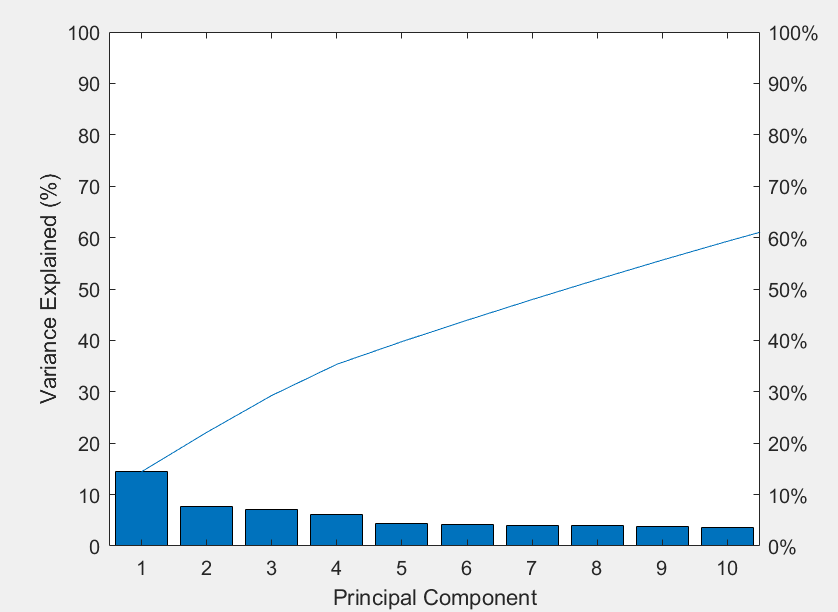
许多机器学习的算法中目标函数的基础都是假设所有的特征都是零均值并且具有同一阶数上的方差。如果某个特征的方差比其他特征大几个数量级，那么它就会在学习算法中占据主导位置，会影响学习器的效率，导致学习器并不能像我们说期望的那样，从其他特征中学习。特征标准化是让不同维度之间的特征在数值上有一定比较性，可以大大提高分类器的准确性。但是对于回归算法来说，本身就需要同时考虑各个不同的特征，如果不同特征之间数量级相差太大，将会影响学习器学习的效率，所以我们应该对原始数据进行特征标准化。在Matlab中提供给我们一种函数zscore，这种函数支持了0标准化。

1. disp('特征标准化')
2. traindata = data;
3. traindata(:,28) = [];
4. newData = zscore(traindata);

### 8.4.4特征降维

由于经过处理后的数据有多达30个特征，有很多特征只是噪声数据，所以我们应该要把它进行pca降维，把降维后的数据作为机器学习算法建模的原始数据。在Matlab中提供函数pca来让我们实现pca降维，我们在进行pca降维后要绘制出降维后的各个主成分的解释贡献率帕累托图来决定我们需要选取那几个主成分进行回归算法的建模。

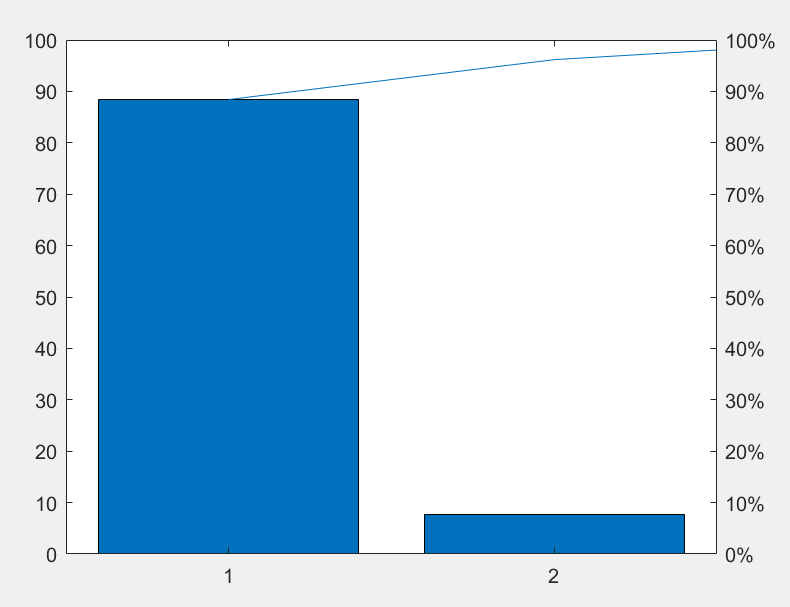
1. disp('特征降维')
2. [COEFF,SCORE,latent,tsquared,explained,mu]=pca(newData);
3. data\_PCA=newData\*COEFF(:,1:10);
4. latent1=100\*latent/sum(latent);%将latent总和统一为100，便于观察贡献率
5. pareto(latent1);%调用matla画图 pareto仅绘制累积分布的前95%，因此y中的部分元素并未显示
6. xlabel('Principal Component');
7. ylabel('Variance Explained (%)');
8. iris\_pac=data\_PCA(:,1:10) ;



**图23 pca降维各主成分的贡献解释图**

但是实际上我们使用标准化后的数据进行机器学习建模的效果并不好，它可能是由于量纲导致的问题，所以我也对未经过标准化后的数据进行了pca降维。

未标准化的数据的主成分贡献解释图如下所示，相比于标准化后的数据，未标准化的数据它有几个成分贡献了绝大部分的作用。这肯定是因为量纲的作用，贡献大部分的数据的量纲较大，它的信息熵也会较大。



**图24 未标准化的数据pca降维各主成分的贡献解释图**

## 8.5 数据集划分

对于机器学习建模来说，通常情况下应该要划分成是三个数据集，训练集用来训练数据，测试集用来调节参数，验证集在未参与模型训练的情况下用来评估泛化能力。所以我先按3:1的大小关系划分出验证集和其他数据集。

1. assess\_x = iris\_pac1(size(iris\_pac1,1)\*0.75:end,:);
2. assess\_y = target(size(target,1)\*0.75:end,:);
3. iris\_pac1 = iris\_pac1(1:size(iris\_pac1,1)\*0.75,:);
4. target = target(1:size(target,1)\*0.75,:);

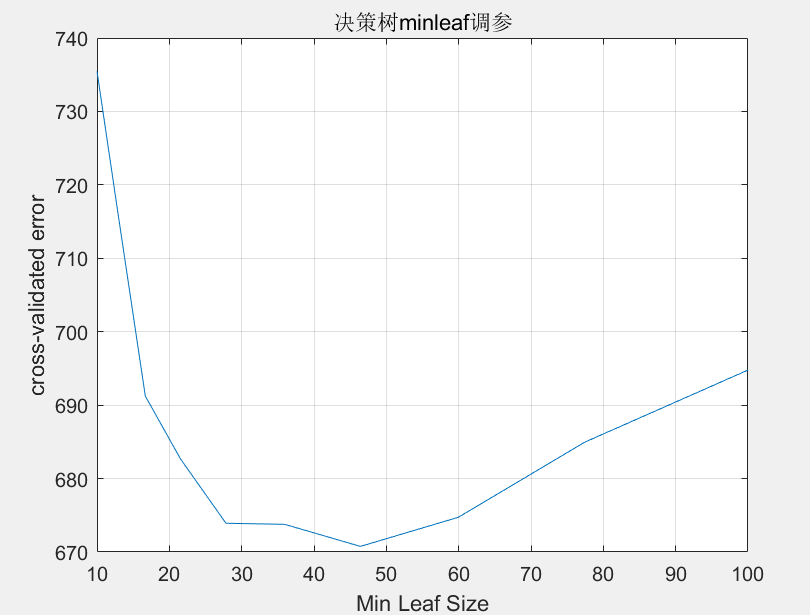
## 8.6 决策树回归

在Matlab中提供了fitrtree函数用来构建决策树的机器学习模型，我设置了参数CrossVal代表交叉验证，通过循环调整minleaf这个参数的取值来调整参数，然后绘制出了每个参数下的决策树模型对应的RMSE的值。



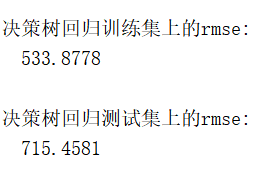
**图25 决策树模型代码**

作出的minleaf学习曲线如下所示。



**图26 minleaf调参**

训练集和验证集上的RMSE值如下所示。

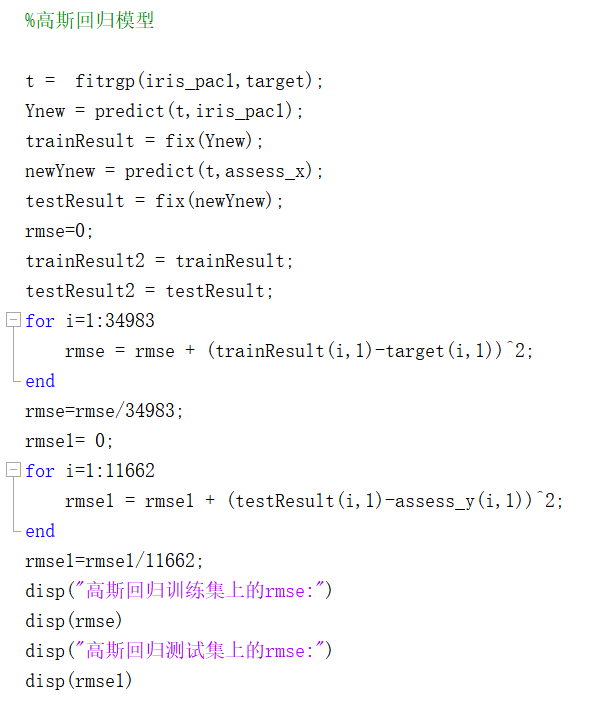


**图27 决策树结果**

在训练集上的RMSE值是533，说明模型在训练集上的表现效果较好。在验证集上的模型的RMSE值是715说明模型的泛化能力还是不错的，但是和在训练集上的表现还是有一定的差距。

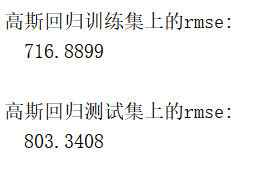
## 8.7 高斯回归

在Matlab中提供了fitrgp函数用来进行高斯回归，我在高斯回归中使用的都是默认参数。



**图28高斯回归代码**

训练集和验证集上的RMSE值如下所示。

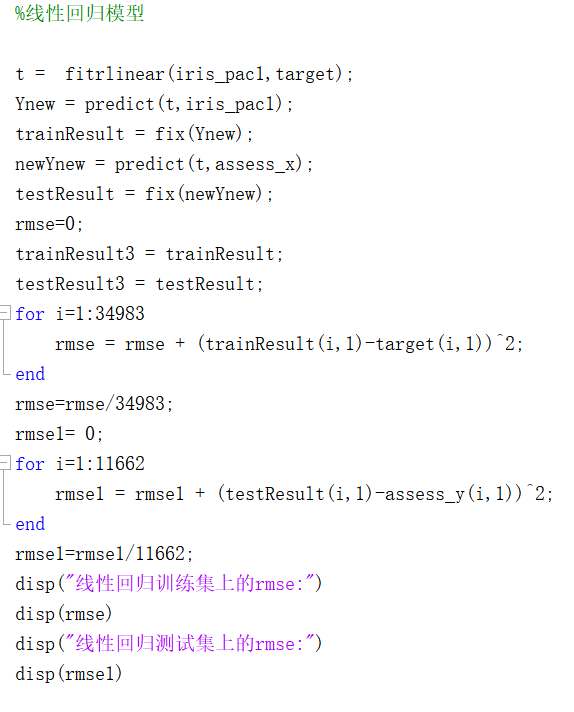


**图29 高斯回归结果**

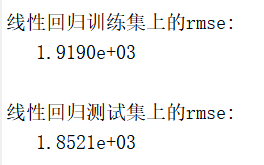
在训练集上的RMSE值是716，说明模型在训练集上的表现效果较好。在验证集上的模型的RMSE值是803说明模型的泛化能力还是不错的，但是和在训练集上的表现还是有一定的差距。但是相比于决策树模型来说，训练集和验证集表现都要差一些，这说明还是决策树模型的效果更好。

## 8.8 线性回归

在Matlab中提供了fitrlinear函数用来进行线性回归，我在线性回归中使用的也都是默认参数。

**图30线性回归代码**

训练集和验证集上的RMSE值如下所示。

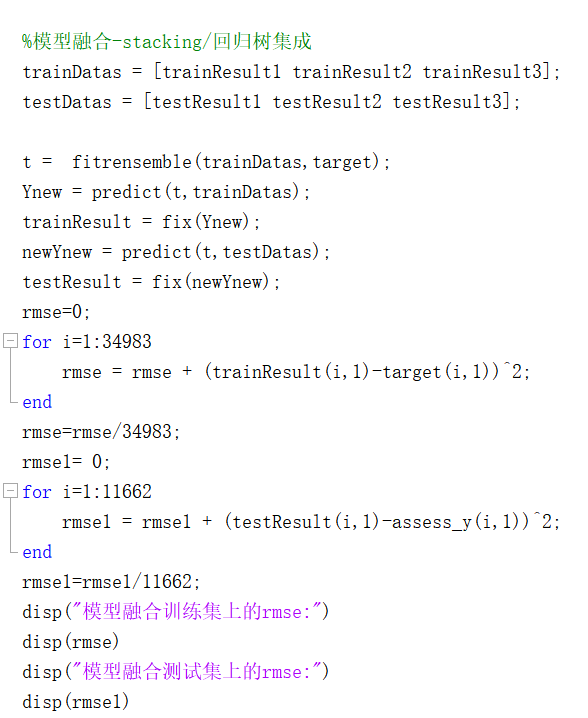


**图31 线性回归结果**

在训练集上的RMSE值是1919，说明模型在训练集上的表现效果较好。在验证集上的模型的RMSE值是1852说明模型的泛化能力还是不错的比训练集上表现还要好。但是相比于决策树和高斯回归模型来说，训练集和验证集表现都要差一些，这说明还是决策树模型的效果更好。

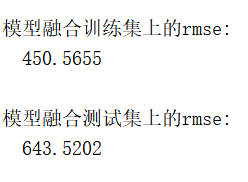
## 8.9 模型融合-基于回归树集成算法的stacking堆叠法

Stacking堆叠法的基本思想是基于几个基本学习器的结果，在它们这些结果的上面一层再构造一层学习器。我stacking堆叠法选择的是回归树集成算法。



**图32 模型融合代码**

训练集和验证集上的RMSE值如下所示。



**图33 模型融合结果**

在训练集上的RMSE值是450，说明模型在训练集上的表现效果较好。在验证集上的模型的RMSE值是643说明模型的泛化能力还是不错的，但是和训练集上还是有一定差距。模型融合算法得到的RMSE值无论是训练集还是验证集上都是最好的，而且相比于其他模型有了不少的提高，说明模型融合法中的stacking堆叠法是一种非常有效地提高模型准确率的方法。

# 九、总结

在本次消费者人群画像-信用值智能评分的项目中，我使用利用线性回归算法、决策树回归算法、高斯回归算法、回归树集成算法等机器学习算法模型，根据用户给定的日常生活中的行为数据进行了针对用户的画像，给定了它的信用分评级，从而使得银行等金融业能够利用信用分数据判断用户的信誉情况决定是否发放贷款给对应用户。

我使用Matlab完成了数据的初步探索，数据的重复值、缺失值、异常值等处理。然后我在数据可视化当中探寻了很多变量和信用分之间的关系，以及变量之间的相关性。接下来我进行了特征工程，进行了特征创造，筛选，标准化，降维。最后我使用了线性回归算法、决策树回归算法、高斯回归算法等机器学习算法进行了单层建模，然后我还使用了stacking模型融合方法进一步优化了模型的预测效果。