南华大学

计 算 机 学 院

课 程 报 告

（ 2023 ~ 2024 学年度 第 二 学期 ）

|  |  |
| --- | --- |
| 课程名称 | 深度学习 |
| 题 目 | Home Credit Default Risk |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 刘靖 | 学号 | 20210441019 |

|  |  |  |  |
| --- | --- | --- | --- |
| 专业 | 软件工程 | 班级 | 软工2101班 |

目录

[1、 课设选题来源和内容 3](#_Toc169354971)

[a) 选题来源 3](#_Toc169354972)

[b) 选题内容 3](#_Toc169354973)

[2、 数据集描述 4](#_Toc169354974)

[2.1 application\_train/test 4](#_Toc169354975)

[2.2 bureau 6](#_Toc169354976)

[2.3 bureau\_balance 7](#_Toc169354977)

[2.4 previous\_application 7](#_Toc169354978)

[2.5 pos\_cash\_balance 8](#_Toc169354979)

[2.6 credit\_card\_balance 9](#_Toc169354980)

[2.7 instalments\_payments 9](#_Toc169354981)

[3、 评测方案 9](#_Toc169354982)

[使用逻辑回归方法训练得出解。 10](#_Toc169354983)

[4、 课设设计说明（算法与模型）   11](#_Toc169354984)

[第一步、在设计之前，读入数据并分析 11](#_Toc169354985)

[第二步、检查Target的分布 12](#_Toc169354986)

[第三步、检查缺失值 13](#_Toc169354987)

[第四步、查看各种数据类型的列数 14](#_Toc169354988)

[第五步、调整训练集和测试集 15](#_Toc169354989)

[第六步、EDA（检查异常值） 16](#_Toc169354990)

[第六步、查看相关性 17](#_Toc169354991)

[第七步、构建模型 19](#_Toc169354992)

[5、 结果与分析 20](#_Toc169354993)

1. 课设选题来源和内容
2. **选题来源**

该比赛是Home Credit公司拿出他们公司的实际问题设置题目，传统的金融机构需要客户有良好且连续的信用卡使用记录才给贷款，Home Credit想用科技的方法解决这个问题，找到那些有实际还款能力但是被传统金融机构拒绝的客户。

本赛题是要预测客户是否会违约，违约的客户会给公司带来实际的经济损失，因为我们要尽量的避免给这部分客户放贷。违约一般分两部分，一个是无还款能力、一个是无还款意愿。当然也有既无还款能力又无还款意愿的人，在这里我们认为无还款能力导致的欠款都是由于无还款能力。

比赛链接如下：<https://www.kaggle.com/competitions/home-credit-default-risk/overview>

1. **选题内容**

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

译文：

由于信用记录不足或不存在，许多人难以获得贷款。而且，不幸的是，这些人经常被不可信的贷方利用。

捷信致力于通过提供积极和安全的借贷体验，为无银行账户的人群扩大金融包容性。为了确保这些服务不足的人群获得积极的贷款体验，捷信利用各种替代数据（包括电信和交易信息）来预测客户的还款能力。

虽然捷信目前正在使用各种统计和机器学习方法来做出这些预测，但他们正在向 Kagglers 提出挑战，以帮助他们释放数据的全部潜力。这样做将确保有能力还款的客户不会被拒绝，并且贷款的本金、到期日和还款日历将使他们的客户能够取得成功。

1. 数据集描述

该比赛一共有十个数据表

## 2.1 application\_train/test

训练集30.75万（SK\_ID\_CURR是主键，其中违约人数是24825，占比8.07%），测试集4.87万，每个训练集&测试集都行都对应一个贷款申请，训练测试的比例是6.31:1。

训练集&测试集的特征及其含义如下：

|  |  |  |
| --- | --- | --- |
| 特征名 | 特征含义 | 特征类型 |
| name\_contract\_type | 合同类型是现金还是循环贷 | 枚举值 |
| code\_gender | 客户性别 | 枚举值 |
| flag\_own\_car | 客户是否有汽车 | 布尔 |
| flag\_own\_realty | 客户是否有房产或公寓 | 布尔 |
| cnt\_children | 客户的孩子数量 | 数值 |
| amt\_income\_total | 客户的总收入 | 数值 |
| amt\_credit | 客户的贷款信贷额度 | 数值 |
| amt\_annuity | 每年要还的贷款额度 | 数值 |
| amt\_goods\_price | 对于消费贷来说，这个字段是要买的商品价格，从数据看每个贷款都对应某个商品，难道业务全部是消费贷吗 | 数值 |
| name\_type\_suite | 办理贷款的时候是谁跟着一起来的：孩子、家人、配偶、自己、… | 枚举值 |
| name\_income\_type | 收入来源：企业家、打工人、政府职员、养老金 | 枚举值 |
| name\_education\_type | 最高学历 | 枚举值 |
| name\_family\_status | 家庭状态：已婚、民事婚姻、单身、离婚、丧偶 | 枚举值 |
| name\_house\_type | 住房状态：租房、和父母住等 | 枚举值 |
| region\_population\_relative | 数字越大代表客户住的地方人口越密集 | 数值 |
| days\_birth | 出生到现在的天数 | 数值 |
| days\_employed | 从事当前工作多少天了，365243是异常数据，要剔除 | 数值 |
| days\_registration | 多少天之前注册的 | 数值 |
| days\_id\_publish | 多少天之前修改了身份文件 | 数值 |
| own\_car\_age | 开的汽车的车龄 | 数值 |
| flag\_mobil | 是否提供手机号，训练集中只有1个没提供 | 布尔 |
| flag\_emp\_phone | 是否提供工作电话，25万人提供了 | 布尔 |
| flag\_work\_phone | 是否提供家庭电话，6万人提供 | 布尔 |
| flag\_cont\_mobile | 电话能不能打通，574个人打不通 | 布尔 |
| flag\_phone | 是否提供家庭电话，8万人提供 | 布尔 |
| flag\_email | 是否提供email，1.7万提供 | 布尔 |
| occupation\_type | 职业类型 | 枚举值 |
| cnt\_fam\_members | 家庭成员数量 | 数值 |
| REGION\_RATING\_CLIENT/\_W\_CITY | 居住区域的等级 | 顺序枚举值 |
| WEEKDAY\_APPR\_PROCESS\_START | 周几申请的贷款，一般工作日多 | 枚举值 |
| hours\_appr\_process\_start | 几点申请的贷款 | 24小时 |
| REG\_REGION\_NOT\_LIVE\_REGION | 区域地址、城市地址的一系列验证，1是不一致、0是一致 | 布尔 |
| orgnization\_type | 工作企业类型：相当于行业：广告、农业、工业、军队等 | 枚举值 |
| ext\_source\_1/2/3 | 外部数据打分 | 数值 |
| APARTMENTS\_AVG/FLOORSMAX\_MODE，一共47个特征 | 是一系列描述客户家房子的数据，有均值、众数、中位数 | 数值 |
| obs\_30\_cnt\_social\_circle、共4个特征 | 客户的社会关系中有多少30天/60天逾期 | 数值 |
| days\_last\_phone\_change | 申请贷款前的多少天换了手机 | 数值 |
| flag\_document\_2-21，共20个特征 | 是否提供文件XX | 布尔 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR，共计6个特征 | 申请贷款前1小时/天/周/月/季度/年查询了多少次征信 | 数值 |

## 2.2 bureau

征信机构合计有171万数据（SK\_ID\_BUREAU是主键），对应30.58万贷款ID（其中26.35万是训练集，4.23万是测试集，比值是6.23:1，和之前的训练测试之比大致相当）。

征信机构特征如下：

|  |  |  |
| --- | --- | --- |
| 特征名 | 特征含义 | 特征类型 |
| credit\_status | Active：630607、Bad debt：21、Closed：1079273、Sold：6527 | 枚举值 |
| CREDIT\_CURRENCY | 信用货币类型，currency1-4，共四个特征 | 枚举值 |
| days\_credit | 客户在申请日前多少天申请的征信机构信用 | 枚举值 |
| credit\_day\_overdue | 申请贷款时客户的逾期天数 | 数值 |
| days\_credit\_enddate | 客户的在征信机构还有多少天的信用时间 | 数值 |
| days\_enddate\_fact | 客户在征信机构关闭了多久的信用 | 数值 |
| amt\_credit\_max\_overdue | 客户到目前为止的最大逾期额度 | 数值 |
| cnt\_credit\_prolong | 在征信机构有几次延期 | 数值 |
| amt\_credit\_sum | 客户当前在信用机构的信用额度 | 数值 |
| amt\_credit\_sum\_debt | 客户在信用机构的当前债务 | 数值 |
| amt\_credit\_sum\_limit | 信用卡的信用限额 | 数值 |
| amt\_credit\_sum\_overdue | 客户在信用机构的违约之和 | 数值 |
| credit\_type | 信用机构的信用类型，车贷、房贷、信用卡、经营贷等 | 枚举值 |
| days\_credit\_update | 信用机构的最近一次信息更新是多少天前 | 数值 |
| amt\_annuity | 信用年金 | 数值 |

## 2.3 bureau\_balance

征信机构的账单信息表共有2729万条数据，对应81万SK\_ID\_BUREAU，占全部征信记录171万的47%（主键为SK\_ID\_BUREAU和MONTHS\_BALANCE）。该表的唯一可用字段是STATUS，C代表closed、X是未知、0是无逾期、1是逾期在1-30天、2是逾期31-60天、3是逾期61-90天、4是逾期91-120天、5是逾期120天以上。

## 2.4 previous\_application

该表是客户在申请这次贷款之前的申请记录，共计167万条（主键是SK\_ID\_PREV），对应33万SK\_ID\_CURR，可见当前训练和测试集中只有2万没有之前的贷款申请记录。

|  |  |  |
| --- | --- | --- |
| 特征名 | 特征含义 | 特征类型 |
| name\_contract\_type | 贷款合同类型，相比1.1多了消费贷和XNA | 枚举值 |
| amt\_annuity | 同1.1 | 数值 |
| amt\_application | 之前的贷款申请了多少钱数值 | 数值 |
| amt\_credit | 这个和申请不一样，这个额度是实际批下来的金额 | 数值 |
| amt\_down\_payment | 之前贷款的首付款 | 数值 |
| amt\_goods\_price | 同1.1 | 数值 |
| weekday\_appr\_process\_start | 同1.1 | 日期 |
| hour\_appr\_process\_start | 同1.1 | 日期 |
| flag\_last\_appl\_per\_contract | 有时一个合同会被错误的提交多次申请，这个字段用来标志是不是一个贷款的最后一次申请 | 布尔 |
| nflag\_last\_appl\_in\_day | 是不是当天的最后一次申请 | 布尔 |
| nflag\_micro\_cash | 是不是小微金融贷 | 布尔 |
| rate\_down\_payment | 归一化的贷款首付比例 | 数值 |
| rate\_interest\_primary | 主要贷款利息的归一化值 | 数值 |
| rate\_interest\_privileged | 优惠贷款利息的归一化值 | 数值 |
| name\_cash\_loan\_purpose | 贷款用途 | 枚举值 |
| name\_contract\_type | 贷款合同类型，approved、cancel等 | 枚举值 |
| days\_decision | 相对于当前贷款，上一次申请的决定是什么时候做的 | 数值 |
| name\_payment\_type | 客户选择上一次贷款的付款方式，现金、电子支付、XNA | 枚举值 |
| code\_reject\_reason | 被拒原因，XAP、HC、LIMIT | 枚举值 |
| name\_type\_suite | 办贷款时是和谁一起来的 | 枚举值 |
| name\_client\_type | 客户是新客还是老客 | 二分类枚举值 |
| name\_goods\_category | 贷款是为了买什么类型的东西 | 枚举值 |
| name\_portfolio | pos、cash、cards、xna | 枚举值 |
| NAME\_PRODUCT\_TYPE | x-sell、walk-in、xna | 枚举值 |
| channel\_type | 获客渠道，country wide、contact center、stone、AP+（cash loan） | 枚举值 |
| sellerplace\_area | 销售区域的面积 | 数值 |
| name\_seller\_industry | 卖家（应该是客户要买商品的卖家）的行业，消费电子、衣服、工业等 | 枚举值 |
| cnt\_payment | 之前贷款分为多少期还款 | 数值 |
| name\_yield\_group | 贷款利息，small、medium、high | 枚举值 |
| product\_combination | 产品组合、 PRODUCT\_COMBINATION、Cash X-Sell: low、Cash、POS household with interest | 枚举值 |
| days\_first\_drawing | 相对于当前申请日期，上一次贷款的首次发放时间 | 数值 |
| days\_first\_due | 相对于当前申请日期，上一次贷款的首次逾期时间 | 数值 |
| days\_last\_due | 最近一次逾期 | 数值 |
| days\_last\_due\_1st\_version | 和days\_last\_due的区别？没搞懂 | 数值 |
| nflag\_insured\_on\_approval | 之前的申请有没有要求保险 | 数值 |

## 2.5 pos\_cash\_balance

共计1000万条数据（主键是SK\_ID\_PREV和月份），对应93万SK\_ID\_PREV，163万前贷款申请中57%有pos机流水记录。

|  |  |  |
| --- | --- | --- |
| 特征名 | 特征含义 | 特征类型 |
| months\_balance | 距今的月份 | 枚举值 |
| cnt\_installment | 贷款期数 | 数值 |
| cnt\_installment\_future | 贷款剩余期数 | 数值 |
| name\_contract\_status | 合同状态，Active、Signed、Complete | 枚举值 |
| sk\_dpd | 当月逾期了多少天 | 数值 |
| sk\_dpd\_def | 忽略金额比较低的贷款，逾期了多少天 | 数值 |

## 2.6 credit\_card\_balance

这个表是信用卡账单表，合计384万条数据（主键是MONTHS\_BALANCE和SK\_ID\_PREV）。

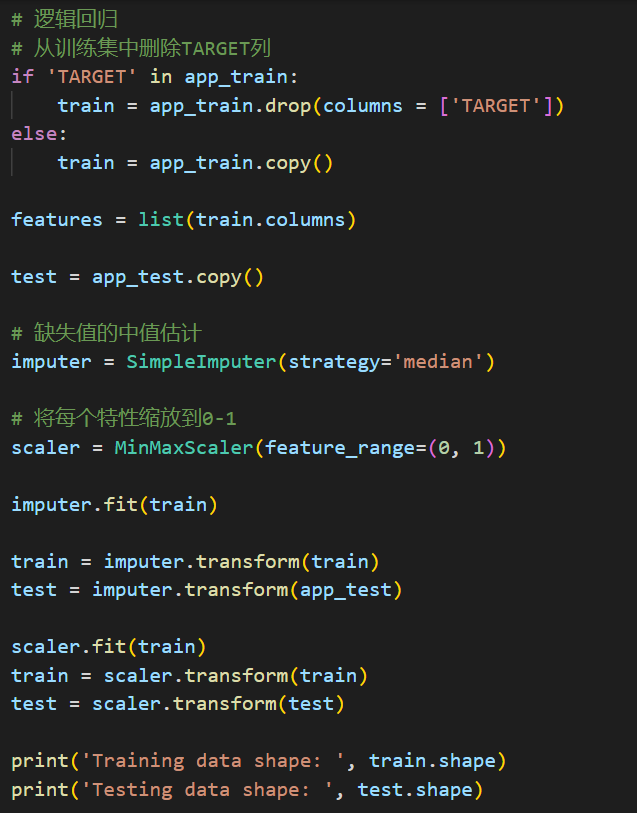
## 2.7 instalments\_payments

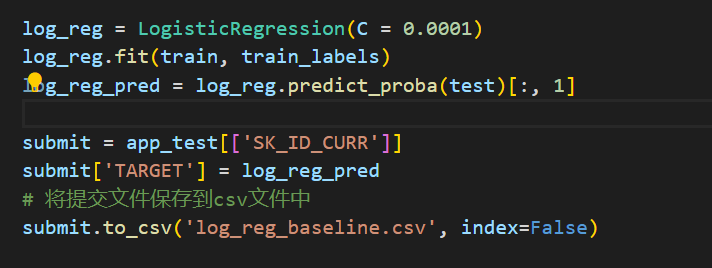
分期付款表共计1360万条数据（主键是SK\_ID\_PREV、分期的第几期、分期版本），对应99万SK\_ID\_PREV，占全部163万前贷款申请的60%。

|  |  |  |
| --- | --- | --- |
| 特征名 | 特征含义 | 特征类型 |
| name\_instalment\_version | 分期的版本数（0是信用卡），版本变化贷款还款日历发生了变化 | 枚举值 |
| name\_instalment\_number | 现在是第多少期分期 | 数值 |
| days\_instalment | 相对于当前申请日期，之前的贷款应该在什么时间支付 | 数值 |
| days\_entry\_payment | 相对于当前申请日期，之前贷款的实际支付时间 | 数值 |
| amt\_instalment | 贷款分期的约定付款金额 | 数值 |
| amt\_payment | 分期的实际付款金额 | 数值 |

1. 评测方案

## 使用逻辑回归方法训练得出解。



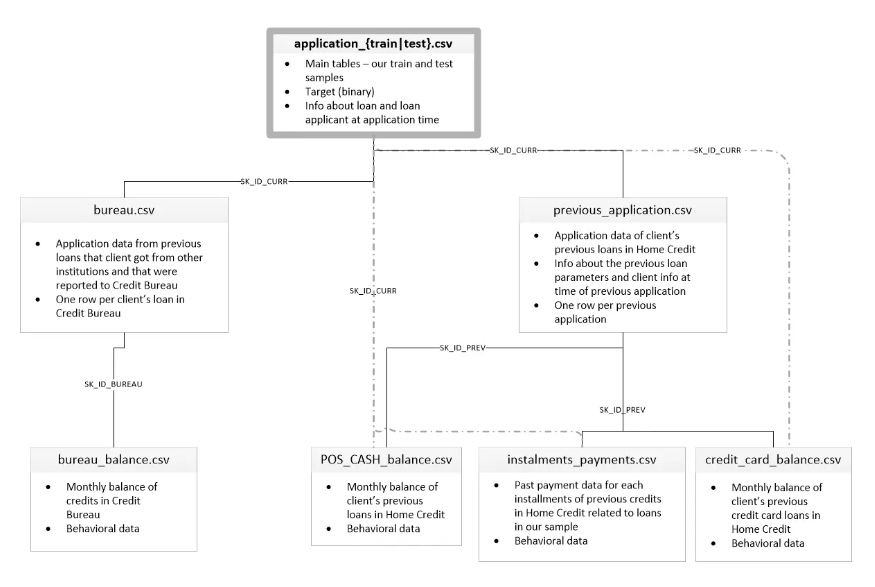


1. 课设设计说明（算法与模型）

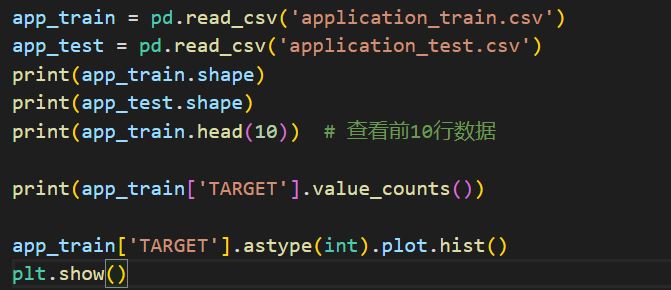
以下我将分为多步来实现该课设。

## 第一步、在设计之前，读入数据并分析

Kaggle上直接给出了各个数据之间的关系图如下：



然后我们读取数据：



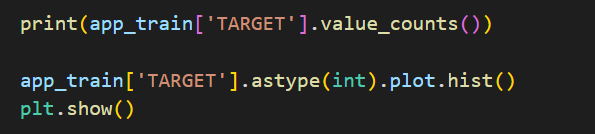
得出结果：



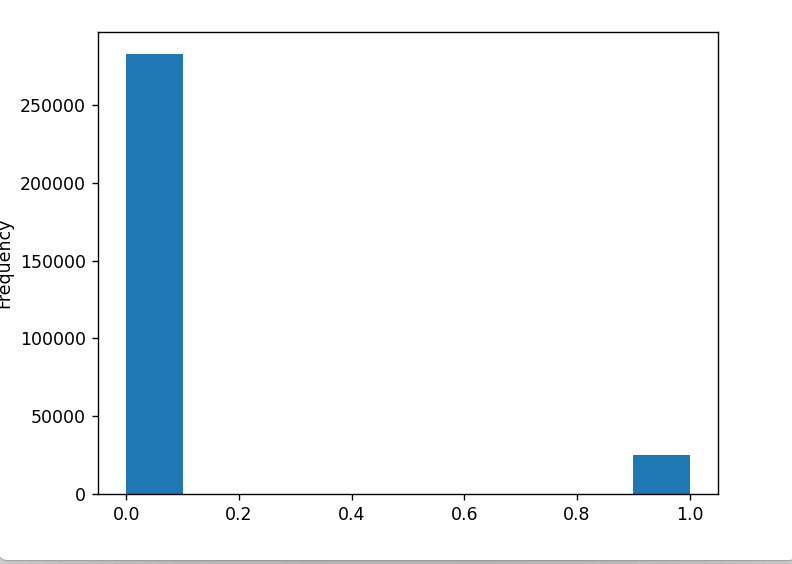
由此可知，可以看出训练集有307511个观察值(每个单独的贷款)和122个特征(变量)。

## 第二步、检查Target的分布

TARGET是我们要预测的：0表示贷款按时偿还，1表示贷款未按时偿还。我们先检查每一类贷款的数量。



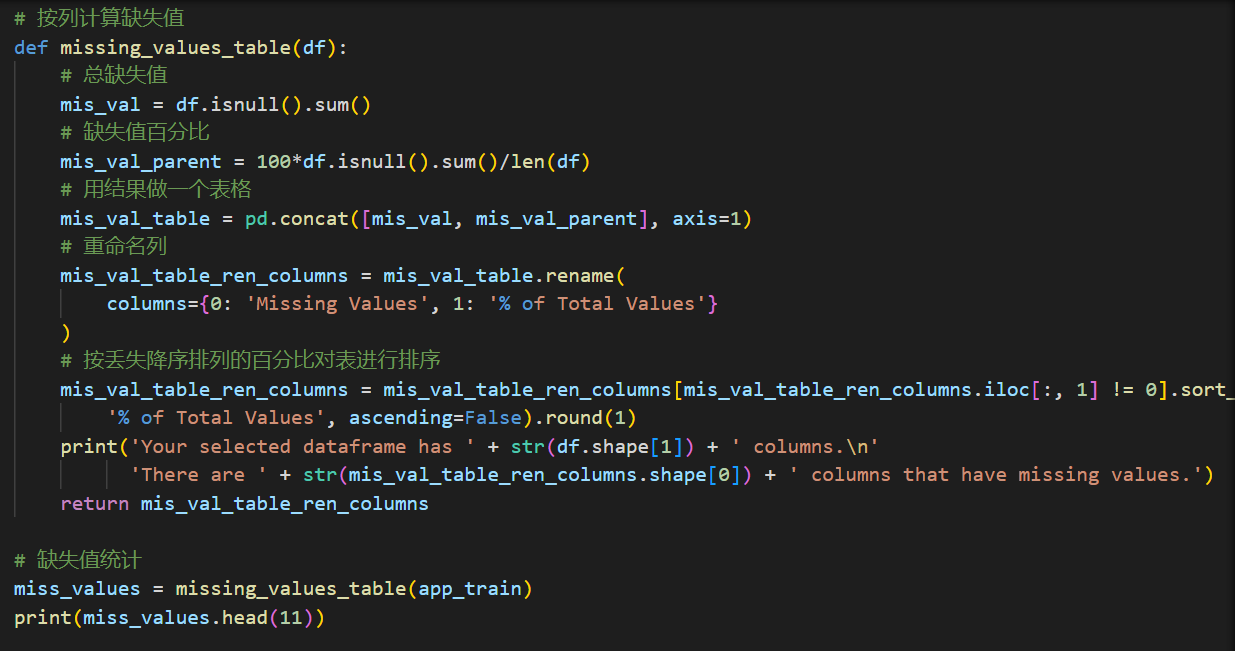
结果图：



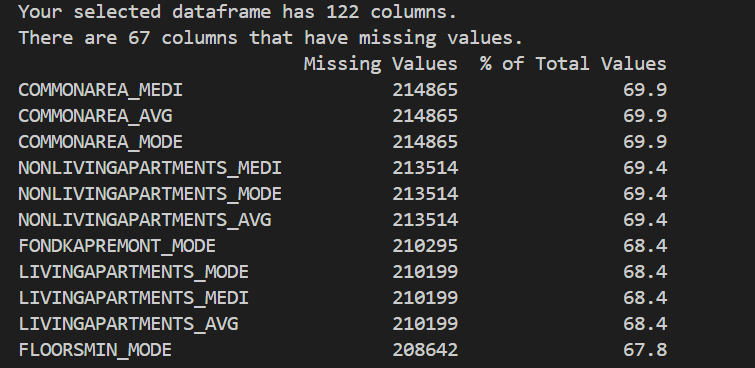
由此可知：这是一个类别不平衡的问题，我们可以很明显的看出按时偿还贷款的远远比未按时偿还贷款的多。

## 第三步、检查缺失值

接下来，我们可以查看每一列中缺失值的数量和百分比。



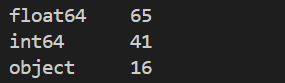
结果如下：



在构建模型之前，我们必须填补这些缺失的值。不然就只能选择一些特殊的模型，比如XGBoost之类的模型，这些模型可以处理缺失的值。

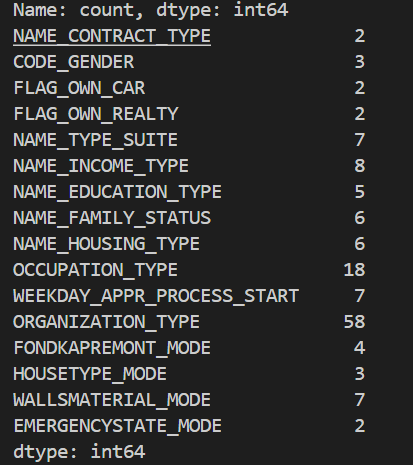
## 第四步、查看各种数据类型的列数



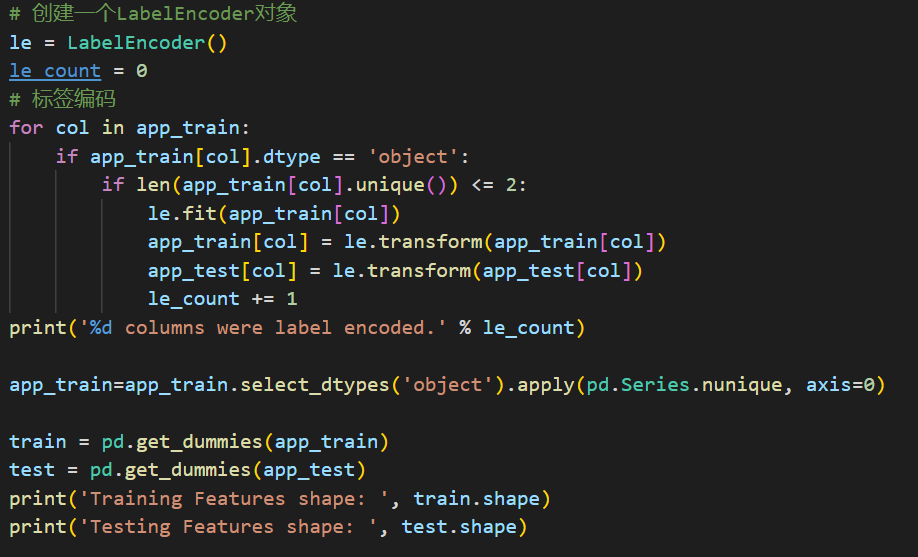


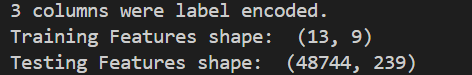
由于机器学习不太能处理object类型的数据，所以要将其转化以下。



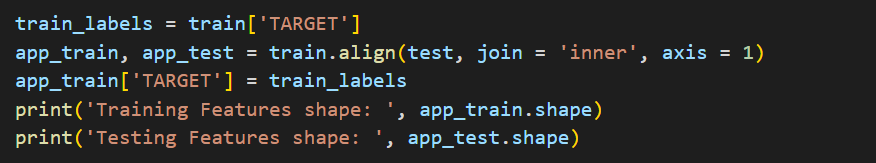


## 第五步、调整训练集和测试集





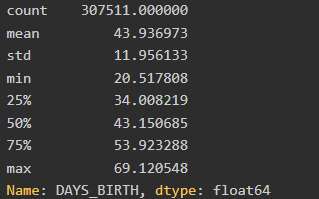
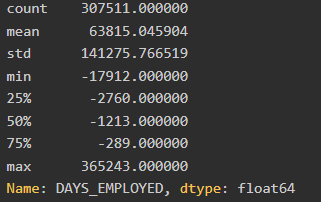
在训练集和测试集中需要有相同的特征(列)。独热编码在训练集中创建了更多的列，因为训练集中有TARGET列，而测试集中没有。





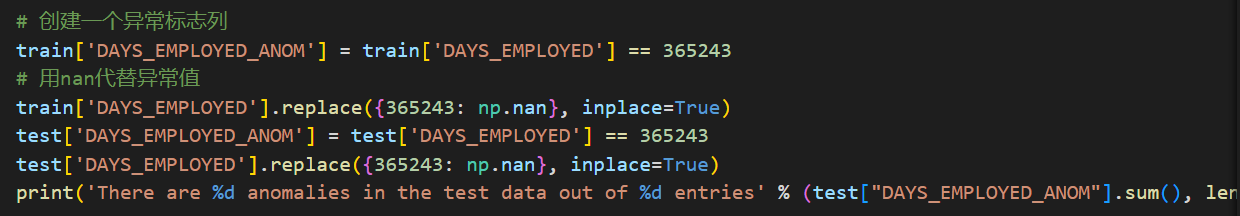
## 第六步、EDA（检查异常值）



第一个看上去是没有异常值的，但是第二个有，最大值是365243天，差不多1000年了，明显有问题。

处理异常取决于具体情况，没有固定的规则。最安全的方法之一就是将异常设置为一个缺失的值，然后在使用算法之前填充它们。在这种情况下，由于所有的异常值都是相同的，所以我们希望用相同的值来填充它们，以防所有这些贷款都有一个共同点。异常值似乎有一定的重要性。作为一种解决方案，我们将用非数字(np.nan)填充异常值，然后创建一个新的布尔列，指示该值是否异常。

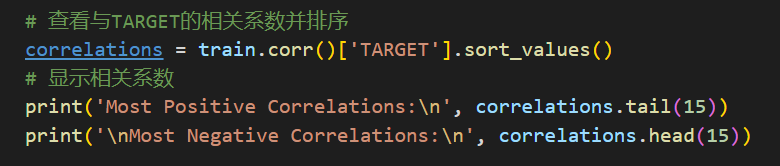


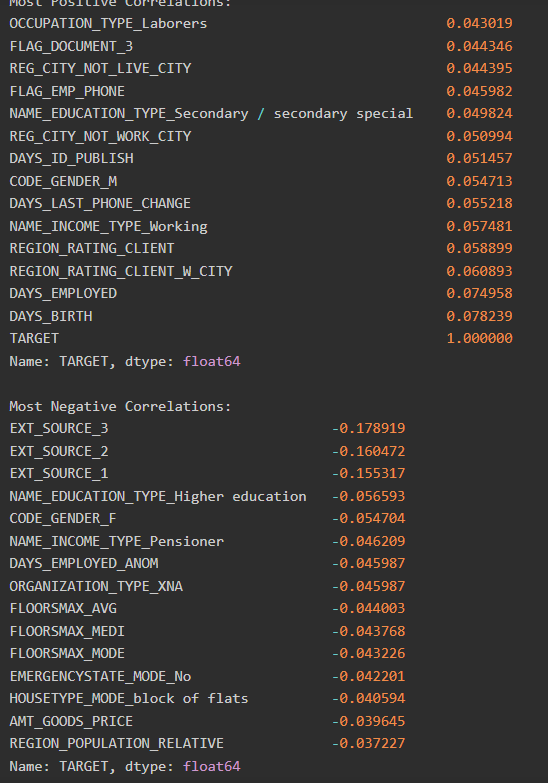


## 第六步、查看相关性

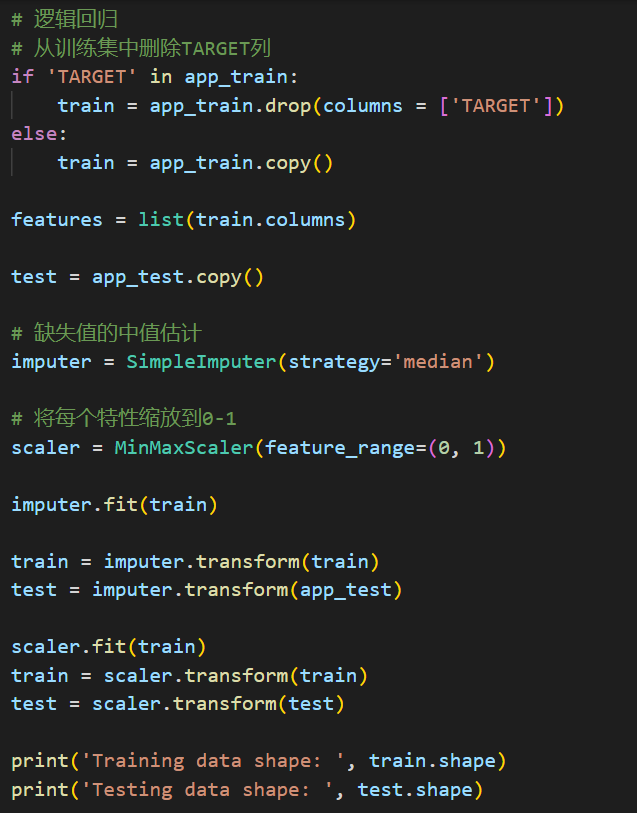
现在我们已经处理了分类变量和异常值，让我们继续EDA。一种尝试和理解数据的方法是寻找特征和目标之间的关联。我们可以使用.corr方法计算每个变量与目标之间的Pearson相关系数。  
相关系数并不是表示特征“相关性”的最佳方法，但它确实让我们了解了数据中可能存在的关系。对相关系数绝对值的一般解释是：

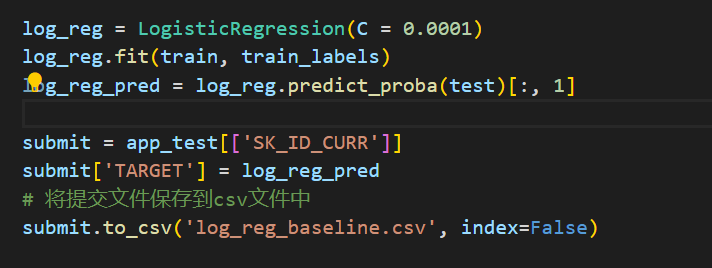
* .00-.19 “very weak”
* .20-.39 “weak”
* .40-.59 “moderate”
* .60-.79 “strong”
* .80-1.0 “very strong”





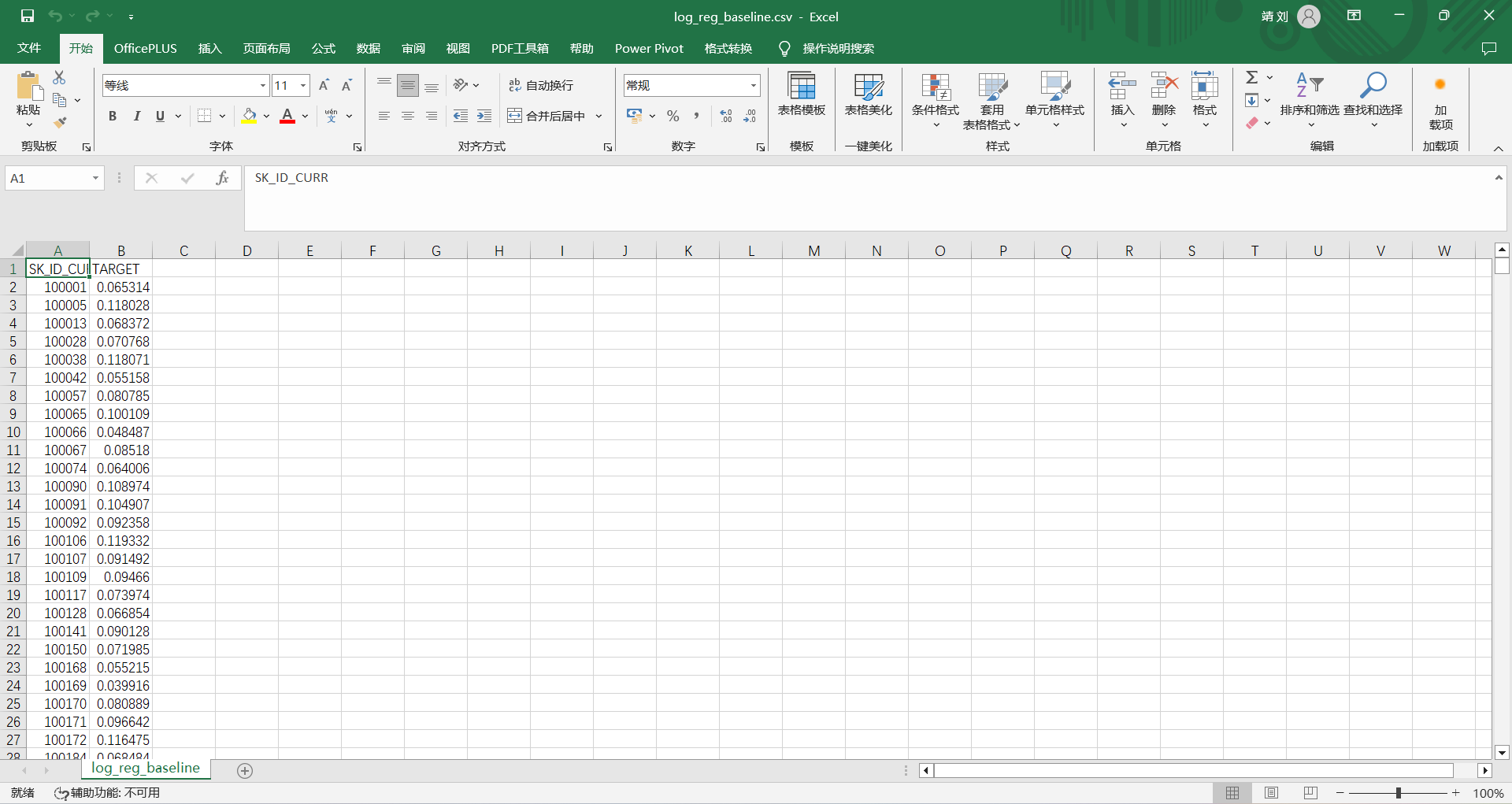
## 第七步、构建模型





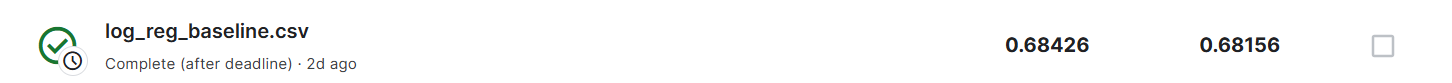
得出结果：





1. 结果与分析

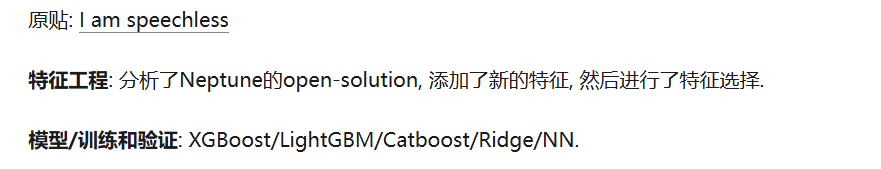
在得出结果后将其提交到kaggle上，得分如下：



在完成该比赛之前，我首先了解数据（这里其实还查了不少资料，看了以下别人对于数据的理解，然后整理成了表格）。然后，我们执行了一个简单的EDA，来识别可能有助于建模的关系、趋势或异常。在此过程中，我们执行了必要的预处理步骤，比如对分类变量进行编码、输入缺失的值以及将特性缩放到一个范围。然后，我使用了逻辑回归方法来构建一个模型，并成功进行了预测。

我其实有去看这个比赛的前几名是怎么做的，但奈何自己比较菜，看了也没法让自己得分也这么高。

比如第一名：



看了以下原帖甚至没咋看懂，还是以后多看多练，提升一下自己的知识量吧。