**华中科技大学计算机科学与技术学院**

**《机器学习》课堂三结课报告**



专 业： 计算机科学与技术

班 级：

学 号：

姓 名：

成 绩：

指导教师： 何琨

**完成日期： 2021年 12月 06日**

目录

[购房贷款违约预测 2](#_Toc89534780)

[一、 实验题目：购房贷款违约预测 2](#_Toc89534781)

[二、实验要求 2](#_Toc89534782)

[三、算法设计 2](#_Toc89534783)

[四、实验环境与平台 9](#_Toc89534784)

[五、程序实现 9](#_Toc89534785)

[六、实验结果 9](#_Toc89534786)

[七、结果分析 10](#_Toc89534787)

# 购房贷款违约预测

## 实验题目：购房贷款违约预测

## 二、实验要求

**2.1题目背景**

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，这些人口经常被不可靠的放贷人利用。住房信贷通过提供积极和安全的借贷经验，努力扩大无银行账户人群的金融包容性。为了确保服务不足的人群拥有良好的贷款经验，利用各种替代数据来预测其客户的还款能力。

**2.2数据集**

附件给定的tran.csv文件，包含若干相关信息，以及是否违约的label字段。

**2.3任务描述**

使用机器学习相关知识完成购房贷款违约预测，给定特征字段，输出是否会发生逾期的预测。

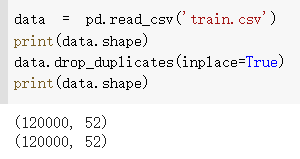
**2.4评测标准**

在线educoder平台评测，输入结果给出F1Score。

## 三、算法设计

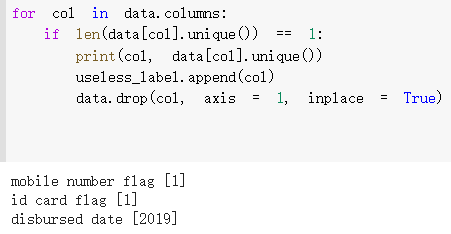
**3.1数据处理**

1. 非空数据



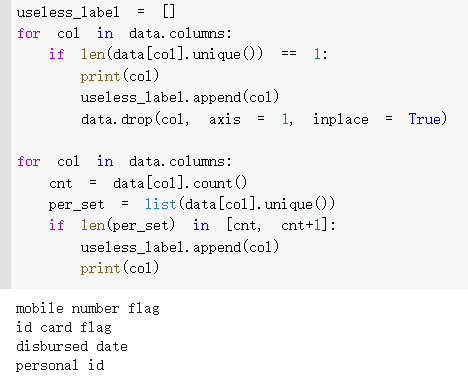
发现不含空行， 忽略该处理步骤

2．忽略一致性数据



当单一值的数量和值的数量的总和相等时，就可以初步证明该数据完全一致，不能当做参考，处理得出的结果为”mobile number flag”、”id card flag”、”disbursed data”三种数据

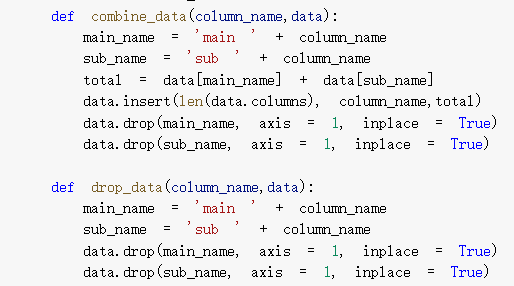
3.忽略标识性数据



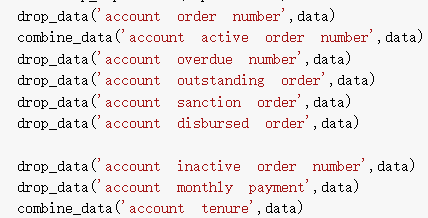
一些唯一性数据，比如个人id，所有的样例都具有不同的值，不做为参考。

4. 重复数据

分析给定的Readme表格，含有许多数据同时存在子账户XXX、主账户XXX，总账户XXX，为明显的数据重复，选择保留主账户数据，删除其他的数据，部分没有主账户的项将两项合并为新的主账户项。



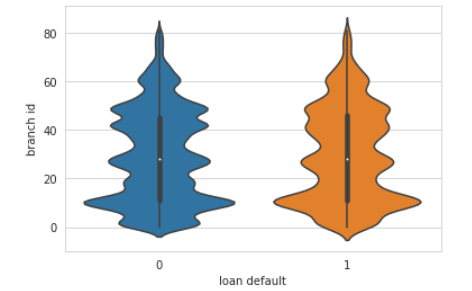
定义了两种函数辅助处理。



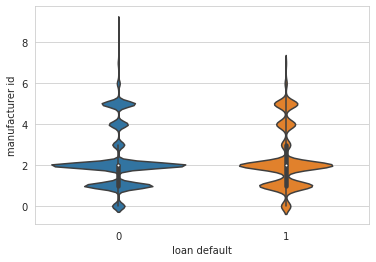
调用对应的处理函数

5.相关性较差的数据

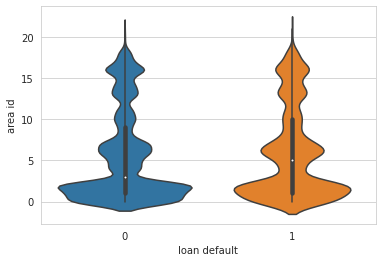
从逻辑上判断，大部分贷款情况与id、年龄等数据并不相关，为验证想法的合理性，依次输出怀疑数据的小提琴图，比较二者的形状验证猜想。



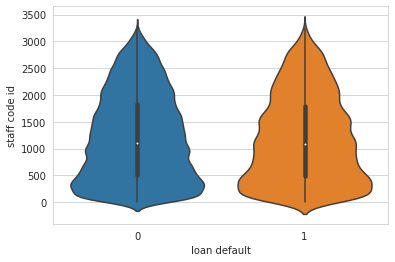
贷款发起银行的id，除了部分波动值几乎一致。



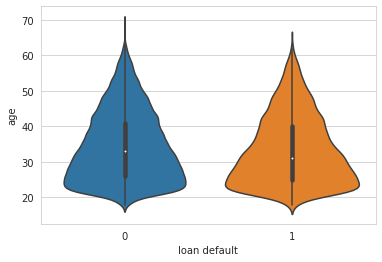
开发商id，形状基本一致，单位有所不同。



区域id，形状基本一致



职员id，形状基本一致（事实上因果关系也不是很大）



年龄，除边界值外其他基本一致。

6.非法数据

首先是部分inf值，对与一些flag值，直接填入0



其他值直接取中位数作为缺省值。



在观察数据时，发现贷款分期数据有的过于离谱，产生了大于一万的分期，这和正常100以内的分期现实显然是矛盾的，直接将相关数据抛弃（实际上也尝试修正值，但训练效果不如直接丢） 

7.test数据处理

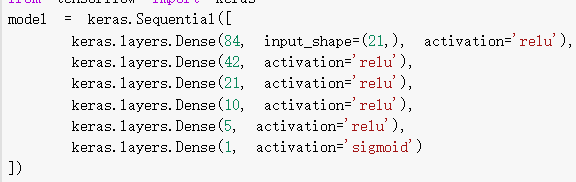
将上述数据处理过程封装为函数，一些会对修改原数据的地方设置flag，避免对test数据集中的相关数据进行修改，影响预测。

**3.2模型算法**



将train.csv的数据划分训练集和测试集。

选用多层神经网络，结合之前筛选出的数据维数设置基本神经网络。



其他层使用relu激活函数，输出层激活函数设置为sigmoid。

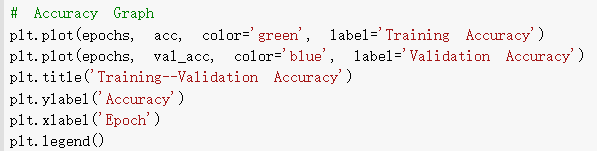


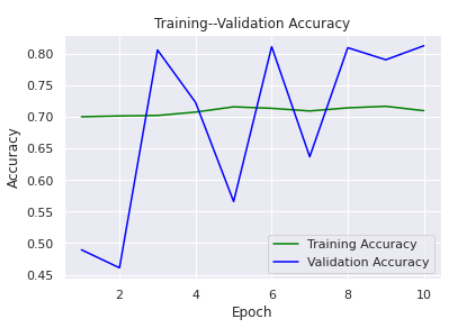
迭代次数设置为10，过多会产生过拟合。

**3.3模型评估**

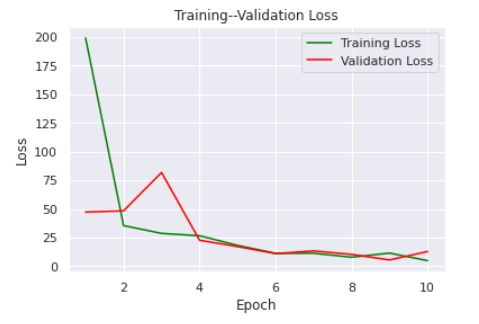
使用相关绘图API

绘制随迭代次变化的训练集、验证集准确率图像



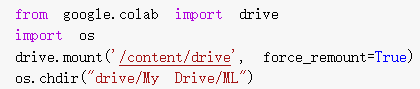


绘制随迭代次数变化的loss损失函数相关图像

****

从图中可以看出，准确率和loss函数总体趋势正常

## 四、实验环境与平台



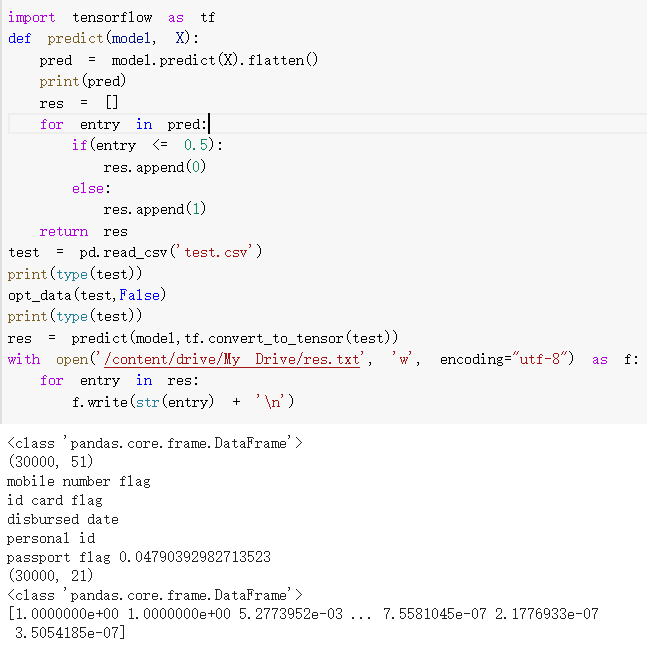
GoogleDirve，使用.ipynb在线编译，将相关的train.csv、test.csv上传到云端文件中；使用相关shell命令转移到对应文件夹中，对云端文件进行操作。

## 五、程序实现

见附加的.ipynb文件

## 六、实验结果

使用训练出的模型预测训练集，将阈值设置为0.5，并将结果输出到 google硬盘的对应文本文件中。



最后label参见附件中的文本文件。

## 七、结果分析

Loss函数波动较大，模型选取有待进一步优化。

时间不足，有空应尝试随机数森林的使用。