数据挖掘课程论文

题　　目：基于决策树的信用风险预测

姓　　名：冯翔

班　　级：经统 1702

学　　号：2120170321

2020 年　6　月　12　日

考察项目	数据符合要求，说明清晰（30%）	软件使用正确（20%）	原理应用正确，结果分析详细正确（40%）	条理清楚（10%）	查重率	总分
得分					10.2%	

一、 引言

信用风险是金融监管机构重点关注的风险，关乎金融系统运行的稳定。在实际业务开展和模型构建中，面临着高维稀疏矩阵以及样本不平衡等各种问题。

因此，如何运用数据挖掘方法提高信用风险的预测的准确率是一项有趣的挑战，同时也是各个金融机构所积极探索的方向。而在实际场景中，更高的预测准确度可以更好地保障金融系统的平稳运行。

二、 实验数据介绍

本次所使用的数据为 2019 年所参加的厦门国际银行“数创金融”数据建模大赛赛题数据。

图 2-1：数据建模大赛

该数据集分为两个部分，一个是 train.csv，该数据集中总共包括 132029 个样本及其对应特征；另一个是 train_target.csv 数据集，该数据集中记录了 132029 个样本所对应的目标变量。前者记录了用户样本的各项特征，后者为各个用户样本所对应的是否违约的标签。

样本数据集中的字段分为三大部分：用户基本信息、借贷相关信息和用户征信相关信息。值得注意的是，由于用户征信相关信息涉及到第三方敏感数据，因此主办方对其做了脱敏处理，同时字段也并未做出进一步说明，而是以 x_1、x_70 等匿名形式给出。而用户基本信息与借贷相关信息的相关字段介绍如下图所示。

字段名称	中文解释	字段名称	中文解释
id	用户唯一标识	edu	学历
target	违约标识，1 违约、0 正常	job	单位类型
certId	证件号	ethnic	民族
gender	性别	highestEdu	最高学历
age	年龄	certValidBegin	证件号起始日
dist	地区	certValidStop	证件号失效日

图 2-2：用户基本信息

字段名称	中文解释	字段名称	中文解释
loanProduct	产品类型	residentAddr	居住地
lmt	预授信金额	linkRela	联系人关系
basicLevel	基础评级	setupHour	申请时段
bankCard	放款卡号	weekday,	申请日（周几）
isNew	是否新增数据		

图 2-3：借贷相关信息

三、 实验用软件工具简介

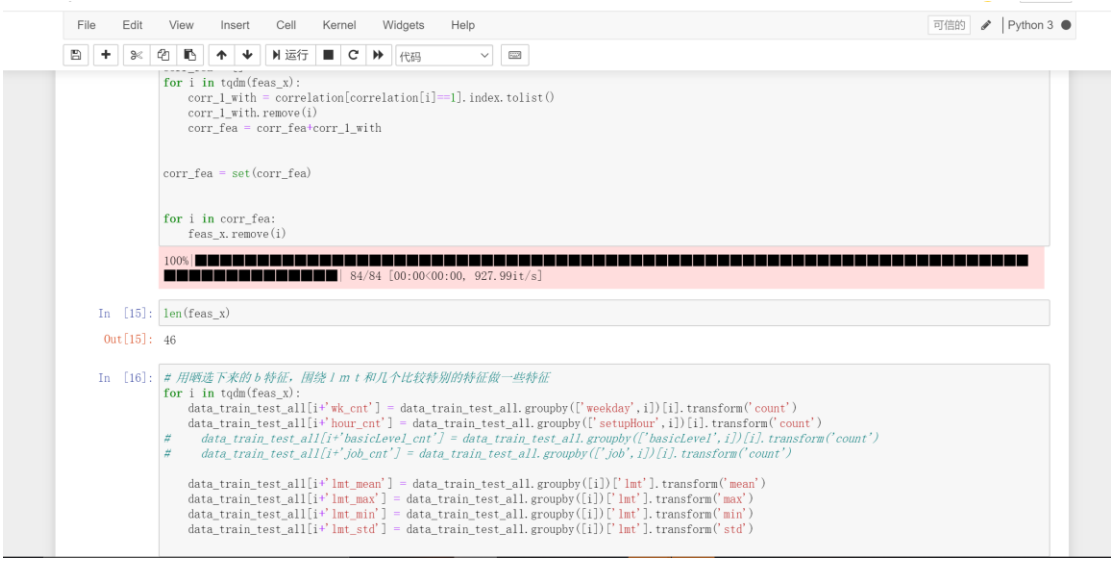

图 3-1：实验所用工具

本次实验所使用工具为 Jupyter notebook，在该环境下使用 Python 语言进行相关操作。Jupyter Notebook 是以网页的形式打开，可以在网页页面中**直接**编写代码和运行代码，代码的运行结果也会直接在代码块下显示。如在编程过程中需要编写说明文档，可在同一个页面中直接编写，便于作及时的说明和解释。因此，十分适合像 Python 这样的解释型脚本语言。

四、 实验所用技术，实验结果及结果分析

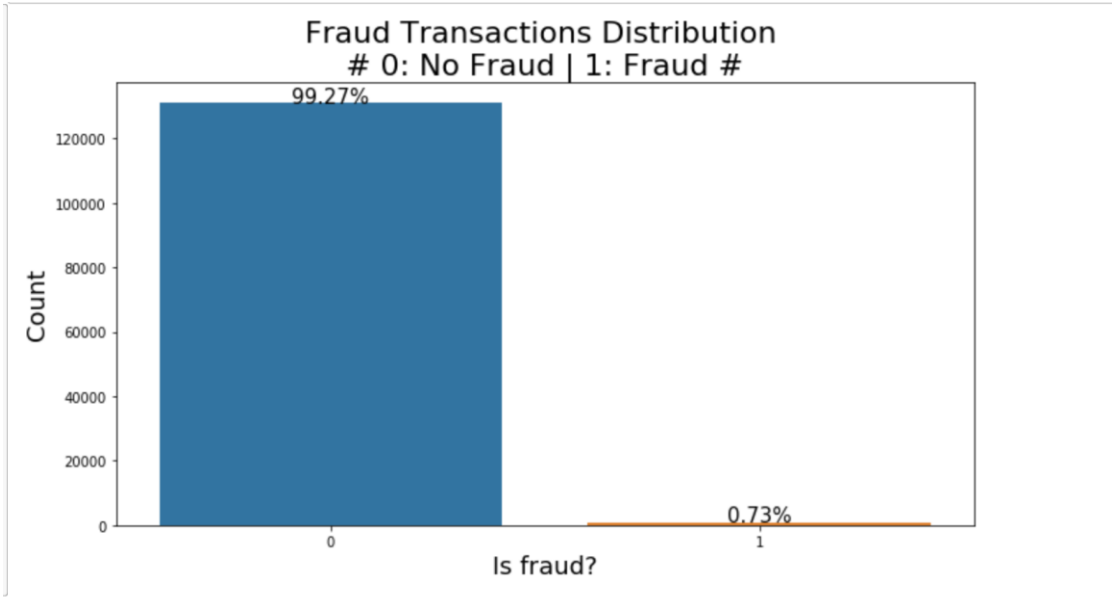

图 4-1：正负样本比例

首先，对训练样本数据集进行探索性分析。如上图 4-1 所示，样本数据集中 99.27%的客户信用良好不存在欺诈行为，数量为 131070，而有 0.73%的客户存在违约行为，数量为 959。显然，样本数据集存在着样本不平衡问题。基于以上事实，选取预测准确率作为度量模型性能的指标意义不大。因此，本次实验使用的是 auc 值作为度量标准。auc 值是指 roc 曲线下方的面积大小，auc 值越趋近于 1 则意味着分类器性能越好，而 auc 值低于 0.5 时意味着分类器性能还不如随机猜测。

通常处理样本不平衡可以通过对训练集重新采样、使用 k 折交叉验证和转化为一分类问题等手段。本文首先尝试了对训练集重新采样的方法，对训练集进行重新采样一方面是减少丰富类的比例，另一方面是增加少数类的样本量。但是效果并不显著，可能是因为正负样本比例过于悬殊。因此，本文主要结合特征工程和五折交叉验证的方法来降低样本不平衡问题的影响。

观察了样本数据集后，发现许多字段都有缺失值的情况。缺失值主要分为两种，一种是 NA，另一种是-999。由于 NA 占比较少，本实验使用 0 来填充。而-999 存在于大多数特征列中，且比例较高，因此本实验将其视为一种值的情形未作处理。

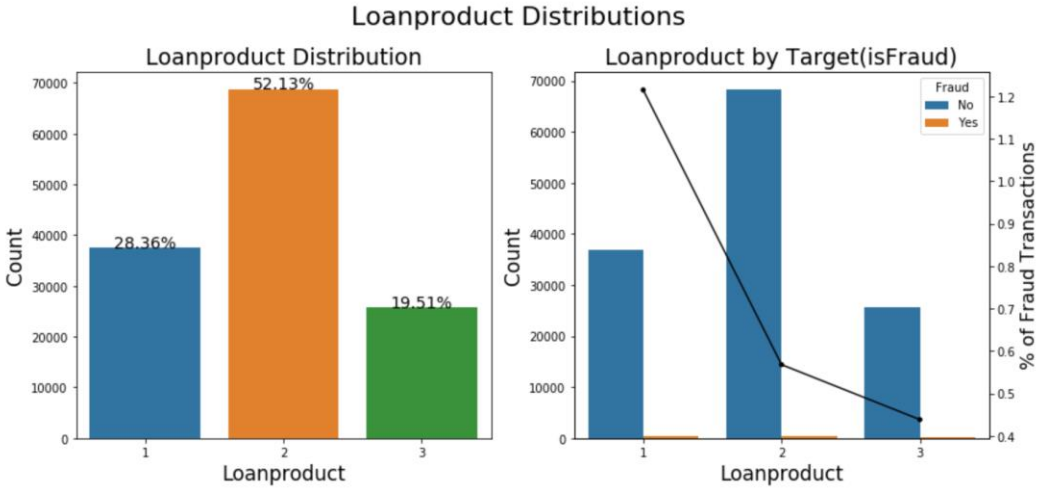

图 4-2：产品类型分布

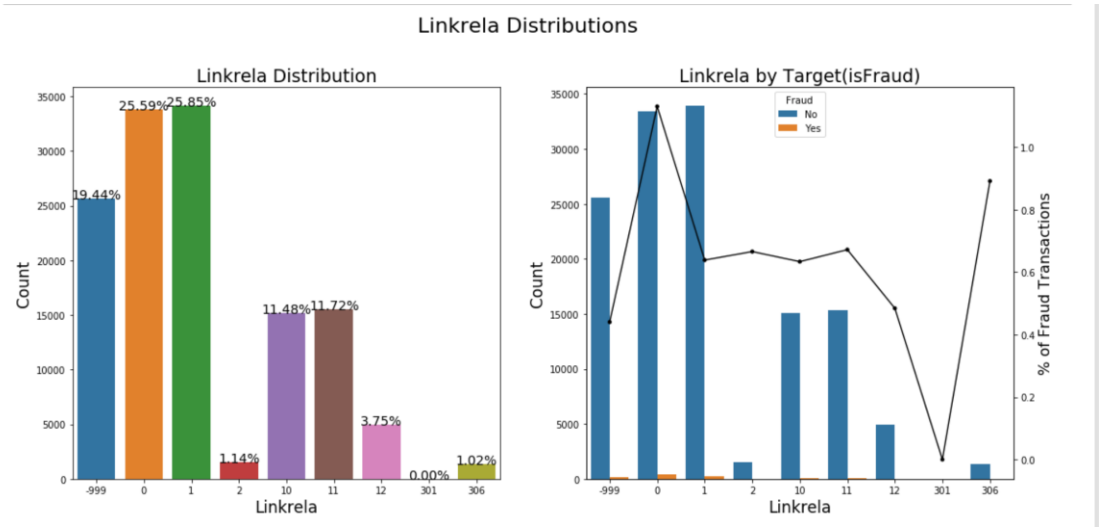

图 4-3：联系人类型分布

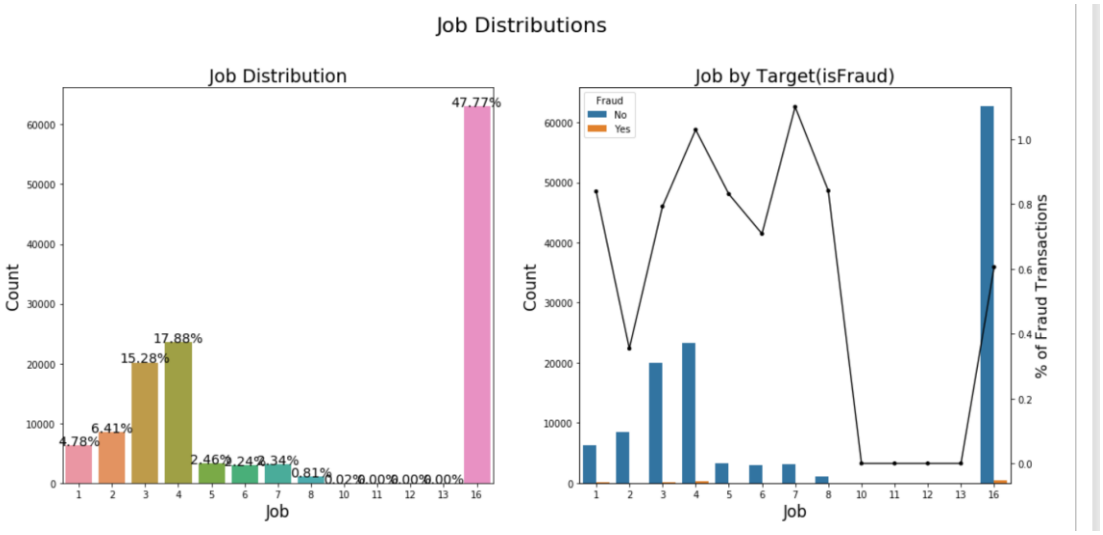

图 4-4：工作类型分布

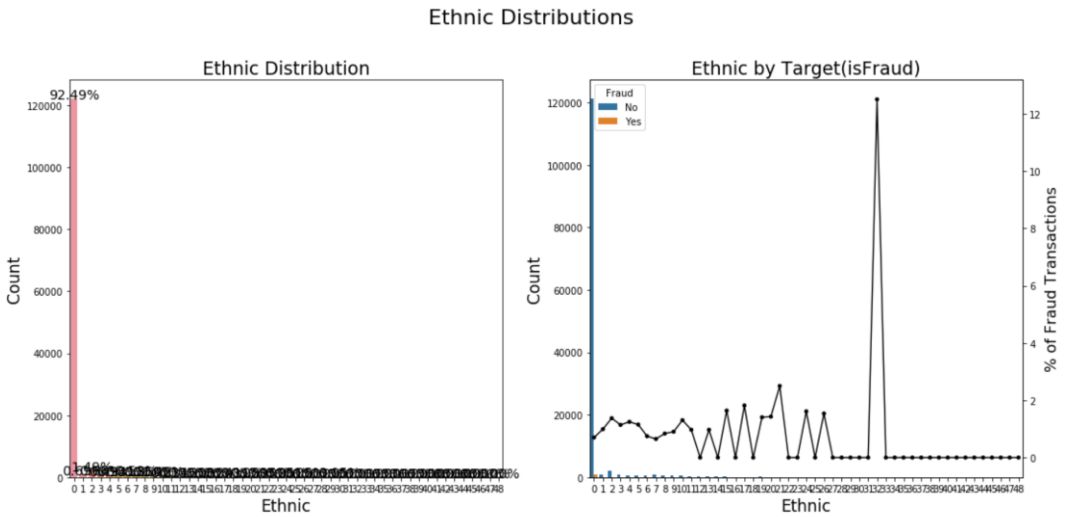

图 4-5 民族分布

对于有明确含义样本特征而非匿名特征，可以简单分为两大类。一类为类别特征，一类为数值特征。

首先，对于类别特征。如图 4-2 至 4-5 所示，本次实验采用较为保守的策略。即对于像产品类型和联系人类型这样的取值数量较少，且各个取值均具有一定比例的特征，本实验采取 one-hot 的方法对其进行编码处理。而对于诸如工作类型和民族这样的取值较多，且各个取值之间占比不均匀的类别特征，尤其是像民族这样的长尾特征，如果对其进行 one-hot 编码，那么将付出巨大的维度成本。因此，本实验采取了较为保守的策略即不做处理。

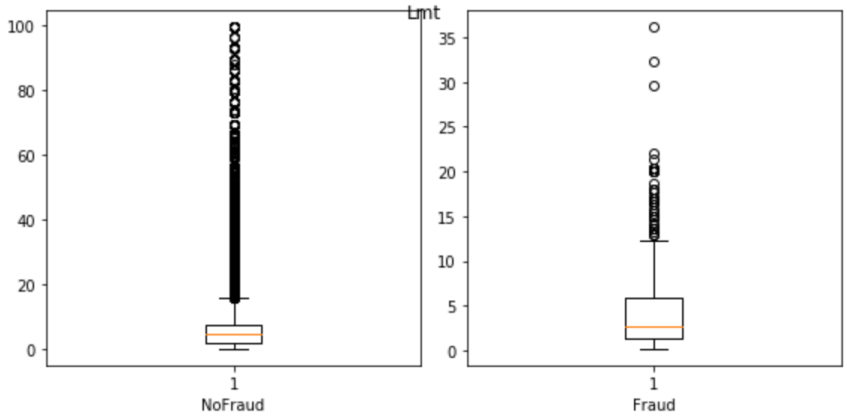

图 4-6 预授信金额箱线图

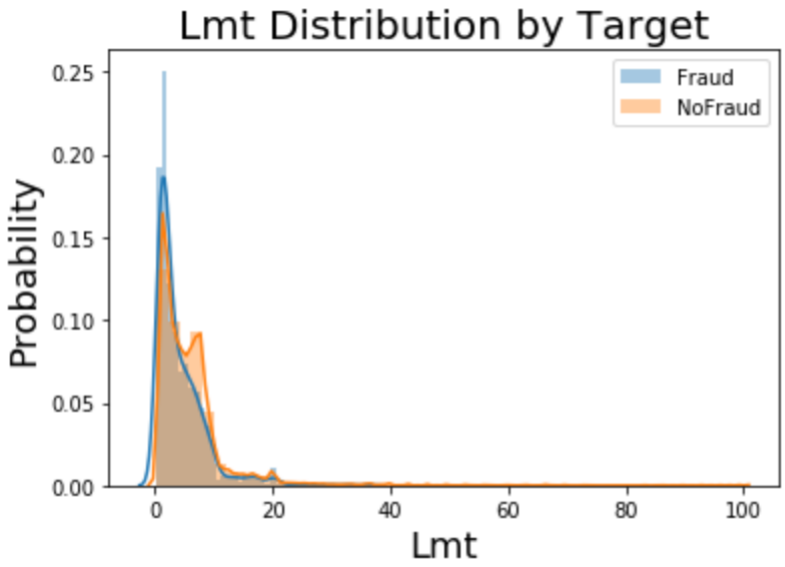

图 4-7 预授信金额分布图

对于像预授信金额这样的数值型特征，本次实验对其进行标准化处理，也即将该特征数据减去其均值除以标准差。

以上便是本次实验数据预处理的主要步骤，基于经过预处理后的数据，本实验 还进行了特征工程，期望挖掘更多有用的特征从而有助于提升算法性能。

考虑到一些交叉特征具有更强的表征能力，本实验将一些特征进行两两组合然后再进行编码处理。例如性别和工作两个特征单独时具有其实际意义，而如果将其进行组合，那么所获得新的特征就可以表示为该客户从事某种类型工作且为某种性别，显然能够使模型捕捉到更多的信息。

除此之外，本次实验还围绕一些类别特征对预授信金额进行分组聚合，从而提取一些诸如最大值、最小值和平均值等统计特征。

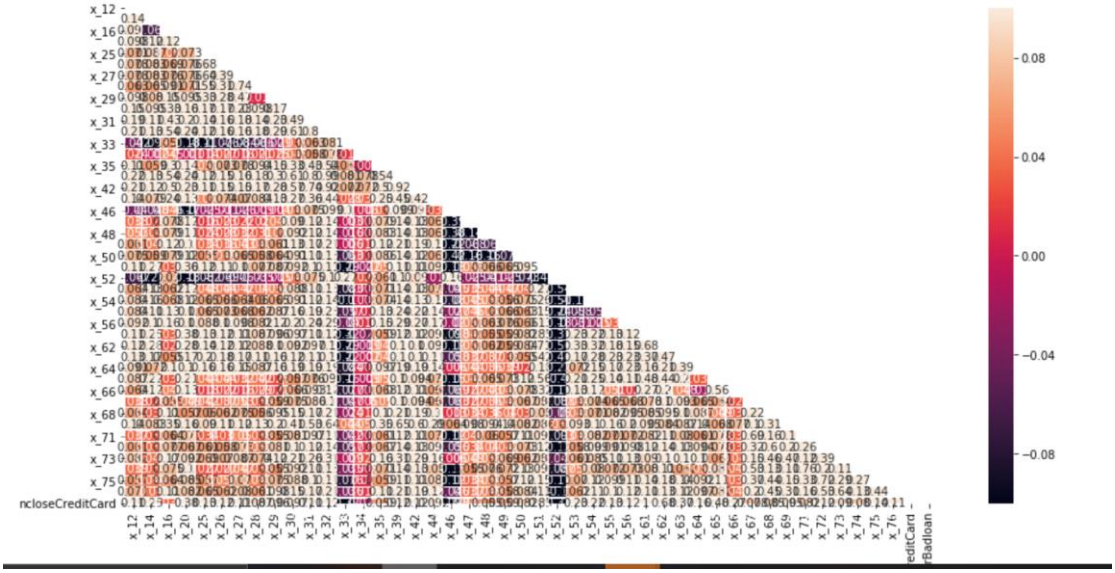

图 4-8　匿名特征相关系数矩阵

基于以上特征工程，本次实验采用一种基于决策树的算法 XGBoost 进行实验。不同于 Bagging 框架下的随机森林算法，XGBoost 是一种 Boosting 算法，支持并行计算。该算法是对梯度提升算法的改进，求解损失函数极值时使用了牛顿法，将损失函数泰勒展开到二阶，另外损失函数中加入了正则化项。训练时的目标函数由两部分构成，第一部分为梯度提升算法损失，第二部分为正则化项。

考虑到正负样本比例的失衡问题，本实验使用分层五折交叉验证对样本数据集进行分割。

```
print(train_x.shape)
xgb_model = xgb.XGBClassifier( learning_rate=0.01, n_estimators=10000, max_depth=6 ,
     tree_method = 'gpu_hist',subsample=0.9, colsample_bytree=0.7, min_child_samples=5,eval_metric = 'auc',random_state=128
    )
xgb_model.fit(train_x, train_y, eval_set=[(train_x, train_y),(test_x, test_y)],early_stopping_rounds=500, verbose=2)
oof[valid_index] = xgb_model.predict_proba(test_x)[:,1]
pred+=xgb_model.predict_proba(test[feat_col])[:,1]
score.append(xgb_model.best_score)
feat_imp['skf'+str(index)] = xgb_model.feature_importances_
```

图 4-9　模型相关参数

本次实验通过网格搜索法来寻找一组较优参数。最终模型所使用的 XGBoost 算法参数为学习率为 0，01，树的棵树为 10000 棵，最大深度为 6，随机采样比例为 0.9，列采样比例为 0.7。

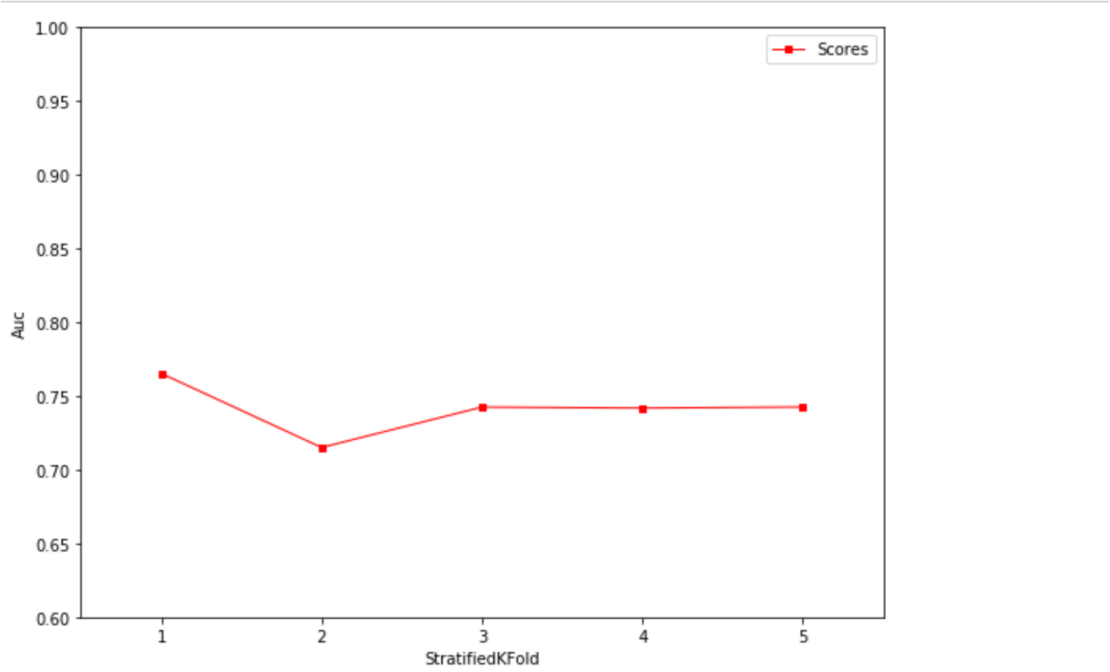

图 4-10　每折交叉验证所对应 auc 值

本次实验在分层五折交叉验证中 auc 最高值为 0.77，最低值为 0.72，其余三次集中在 0.74 至 0.75 之间。平均 auc 值为 0.74。使用训练所得模型对线上测试集进行测试其 auc 值可以达到 0.78，因此模型泛化效果较好，模型训练结果较为理想。

五、实验小结

本实验针对样本数据集中的类别不平衡问题，一方面使用分层抽样的手段进行五折交叉验证，另一方面使用通过特征工程的方式构建有效的特征变量使得所用的模型能够更好地去逼近数据的上限。考虑到问题的复杂性，所作尝试还是比较成功的。

除此之外，本实验还有许多值得改进之处。例如模型的精度还有进一步提升的空间，特征维度较高可以可以考虑采取一些适当的降维方法来降低数据维度等。