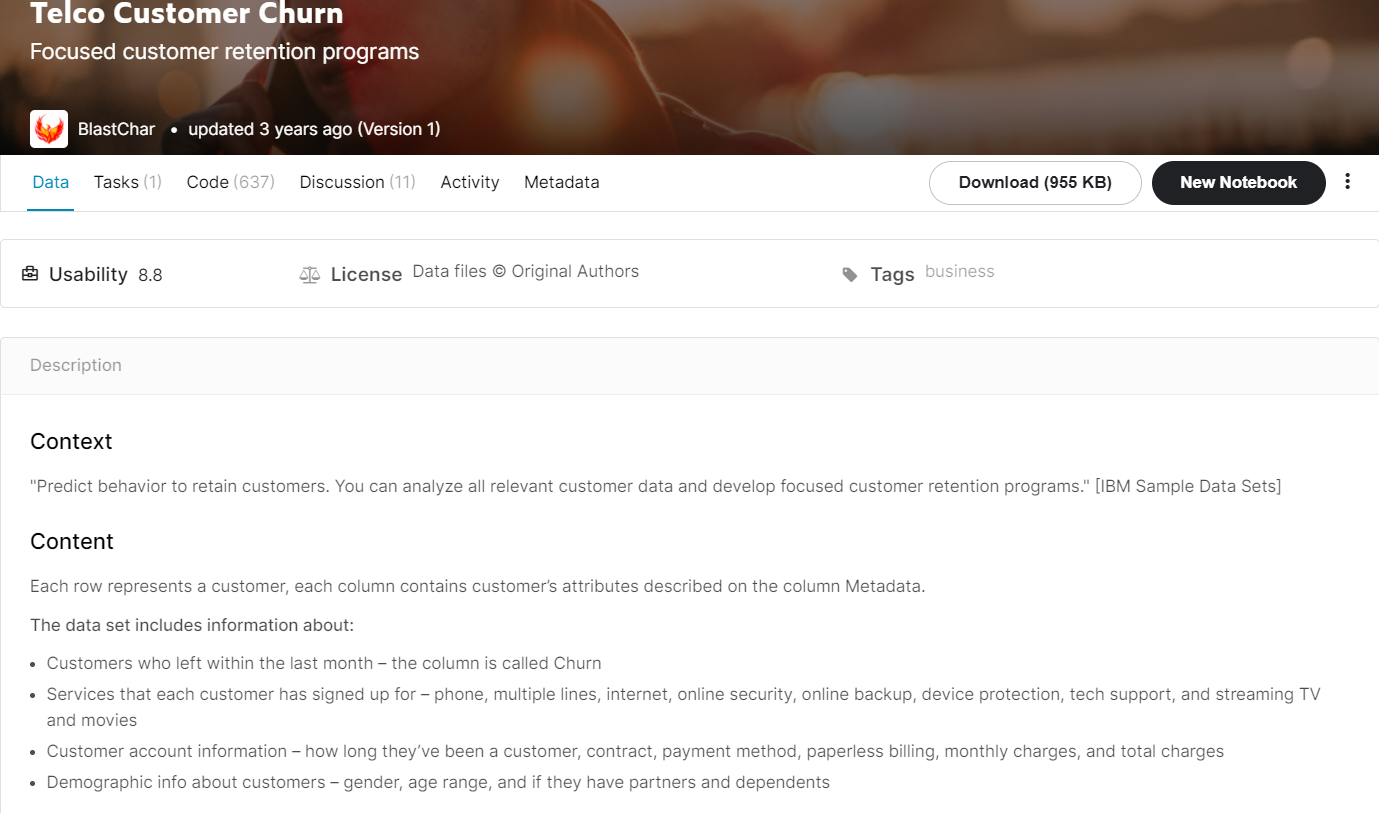
# 第四次作业设计说明书

班级：18级软件工程1班 姓名：许航

**一、数据介绍**

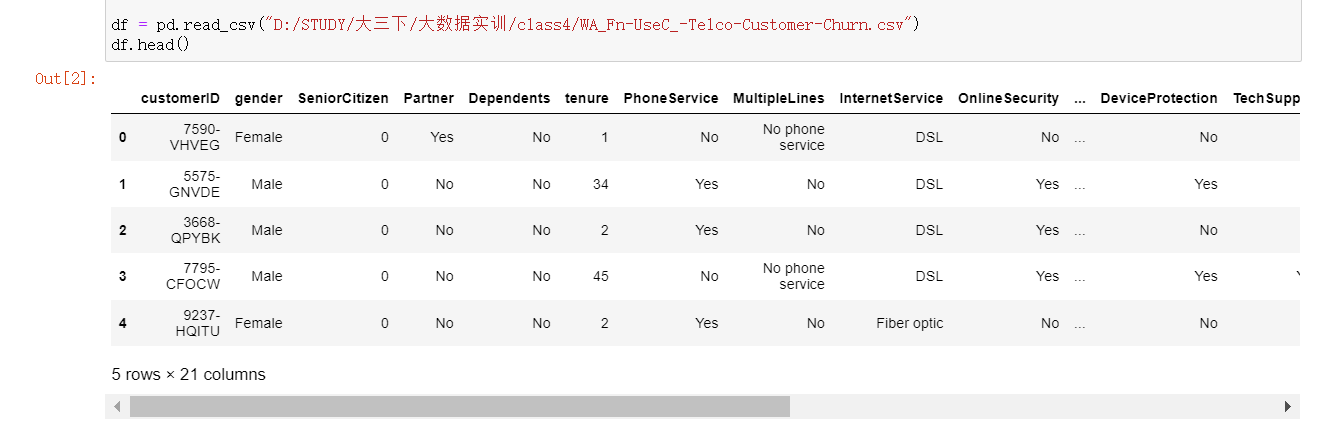


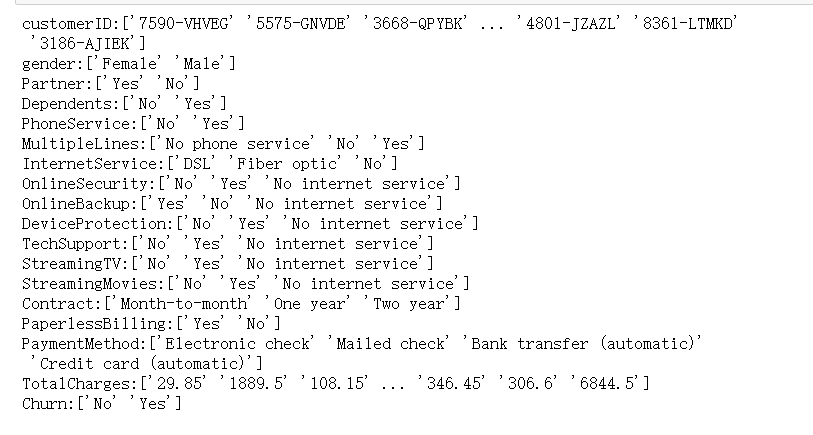
选择kaggle网站上电信客户流失分析数据集作为本次作业的练习数据集。数据集共7043条记录，21个字段。字段包括20个输入特征以及1个目标特征。目标特征为是否为流失顾客，输入特征包括：顾客Id，性别，是否老年人，是否有伴侣，是否有小孩，已使用年限，是否使用电话服务，是否使用网络服务，服务合同期限，支付方式、账单类型、月消费、总消费等一系列描述。

需要根据输入的特征预测顾客是否是流失顾客。

**二、数据清洗**

1、检查数据输入特征的数据类型



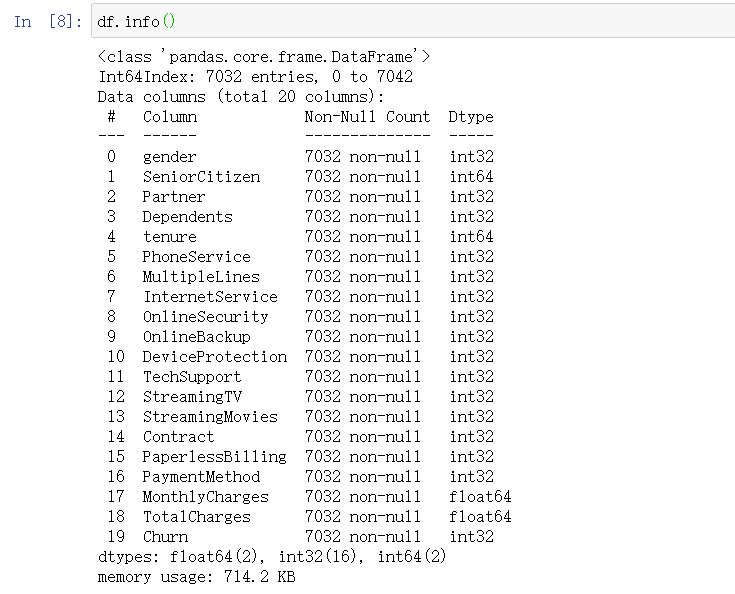


2、将非数值类数据转化成数值

通过独热编码的方式将特征只有2~4个选项的输入特征转化为数值信息。

将数据本身为数值却不为数值类型的转化为数值类型的输入特征。

顾客ID可以在训练模型的时候丢弃。

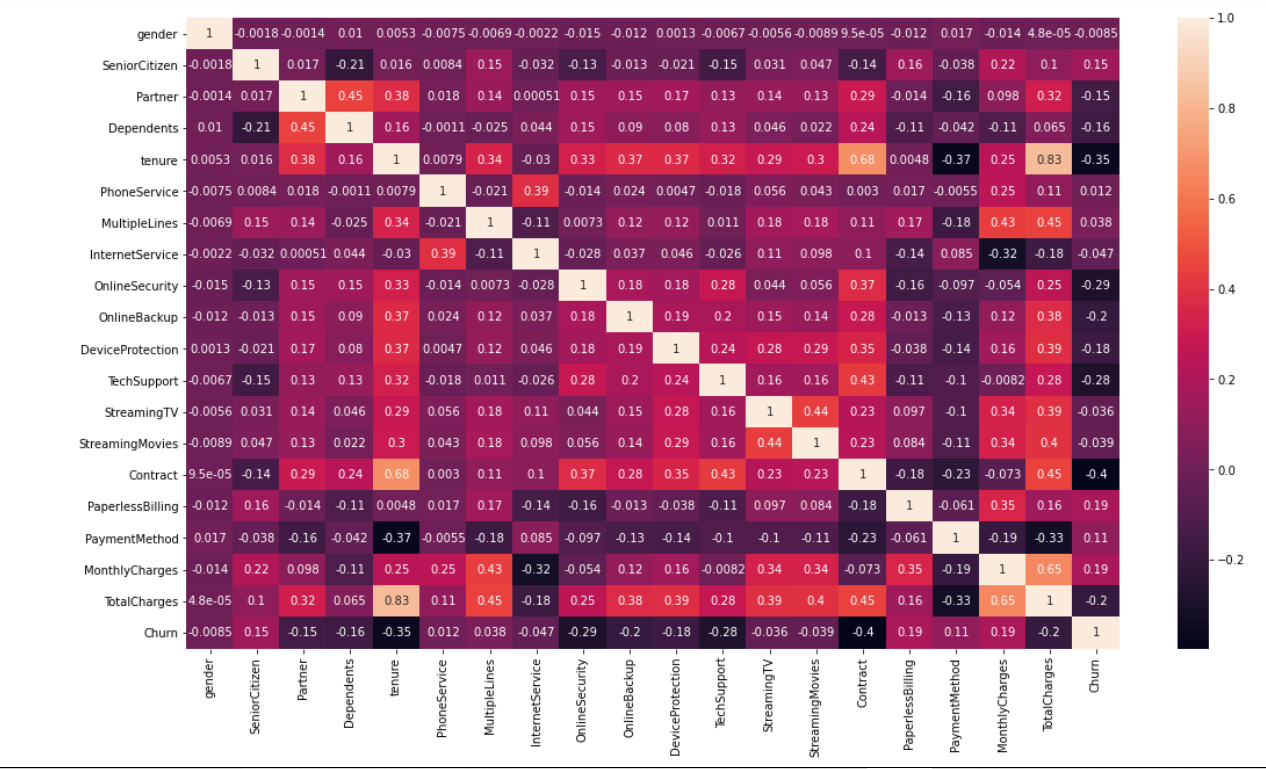


处理完成后可以看到所有20个输入特征均已转化为数值类型。

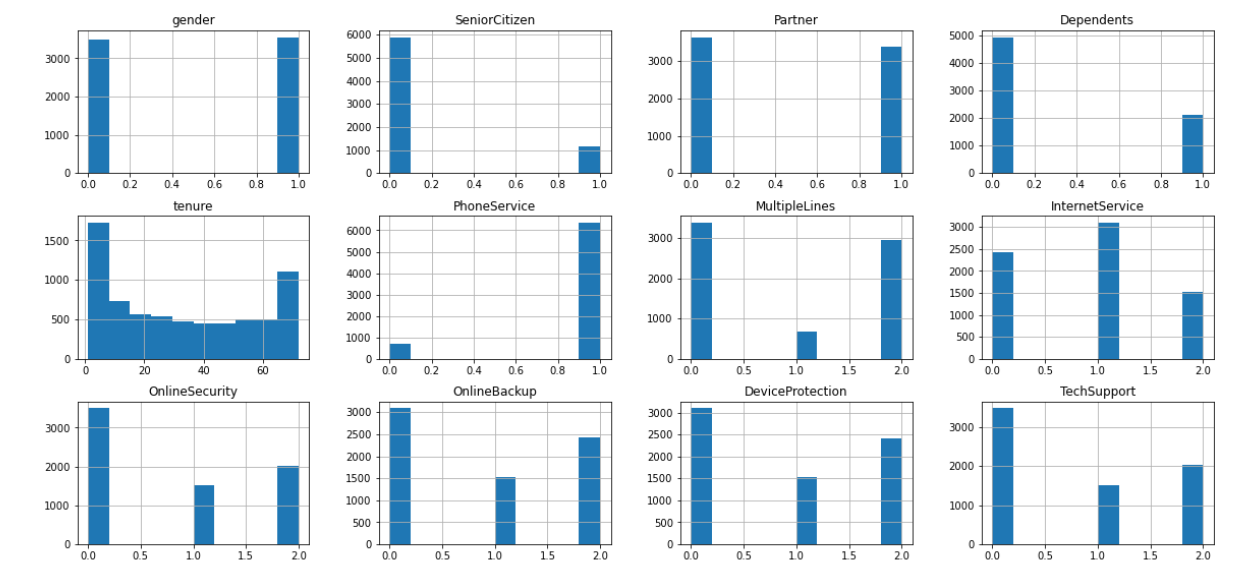
因为本数据集没有缺失值，不用做该方面的处理。

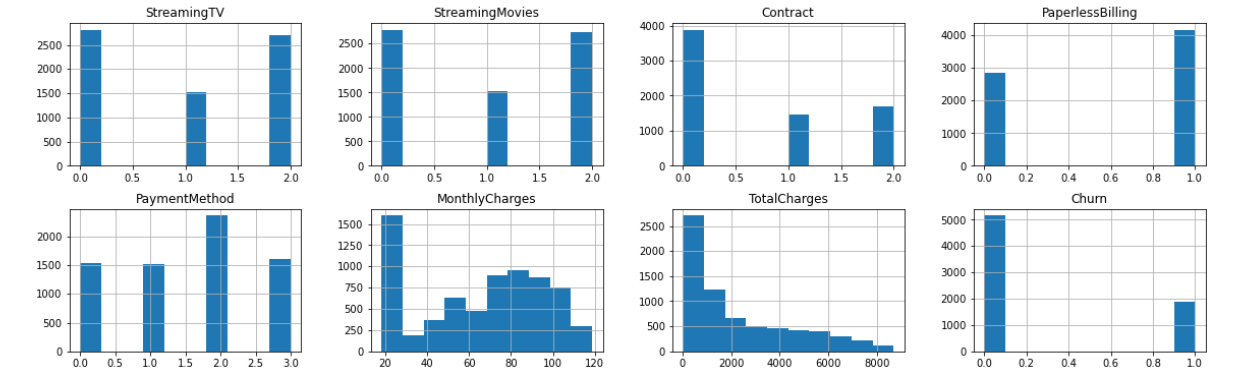
3、画出有关图像，观察主要特征

热力图



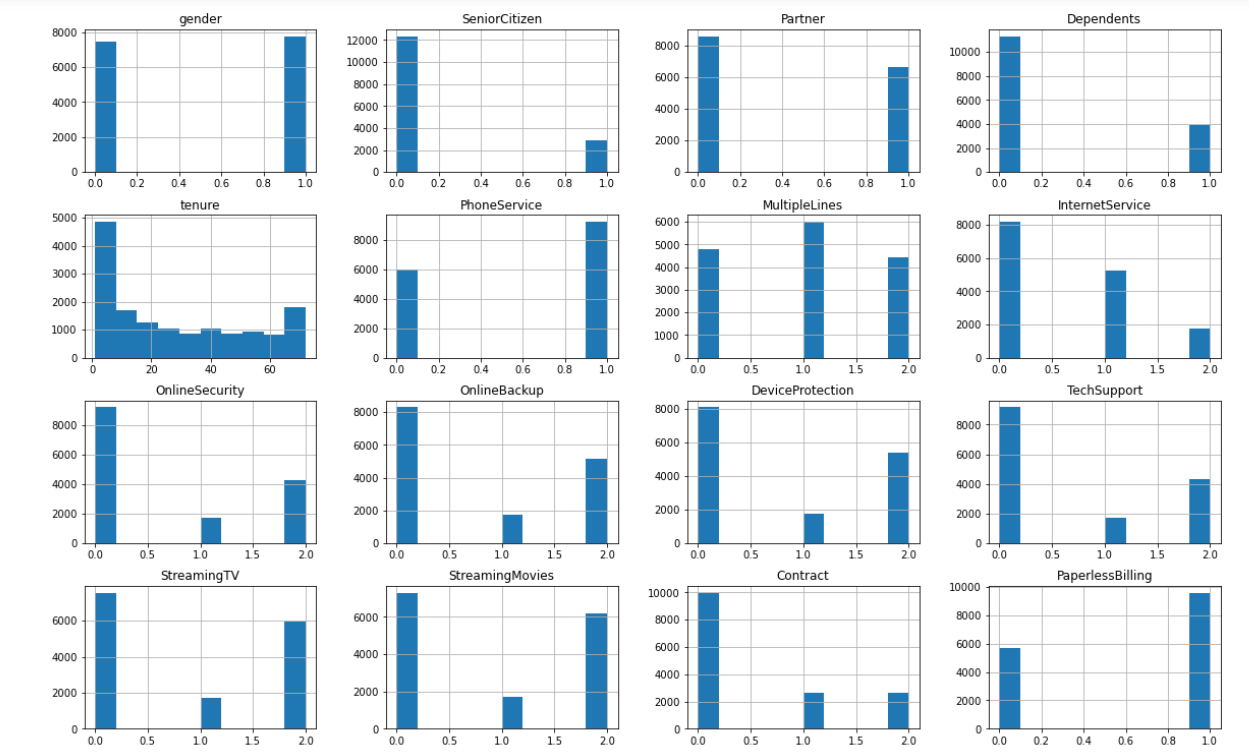
每个输入特征的分布直方图

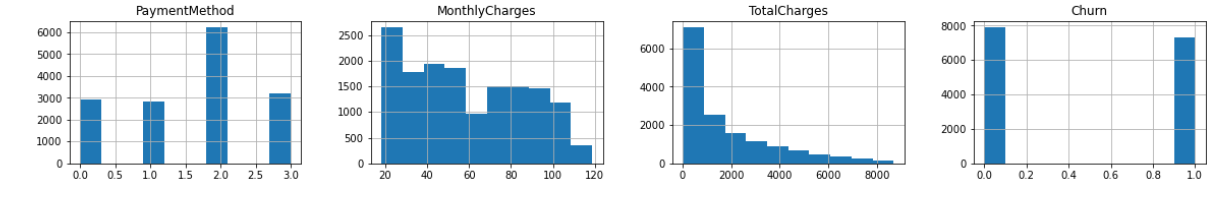




观察到churn、phoneService 数据严重不均衡，因此采用重采样的方式让其保持趋势的同时减弱不平衡。

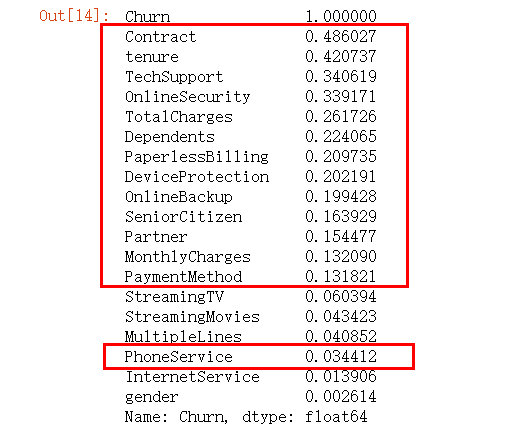
采样后的直方图分布结果：





4、选择主要特征

通过计算其他输入特征与Churn列数据之间的相关性进行排序筛选出主要特征。因为PhoneSerive经过重采样的处理，选入特征以减小误差。

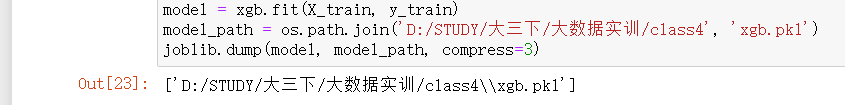


**三、训练模型**

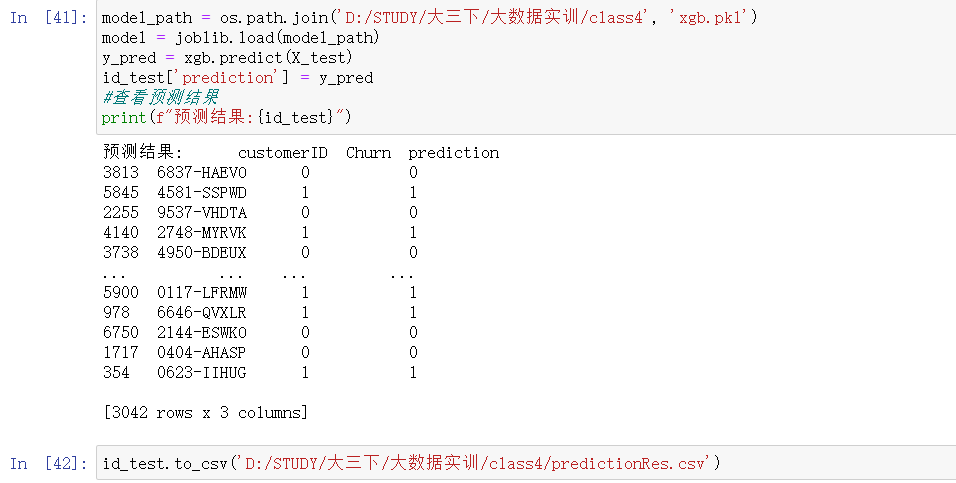
采用XGBoost训练数据。

设置模型学习率、迭代次数、树深度等参数。

生成模型保存至本地。



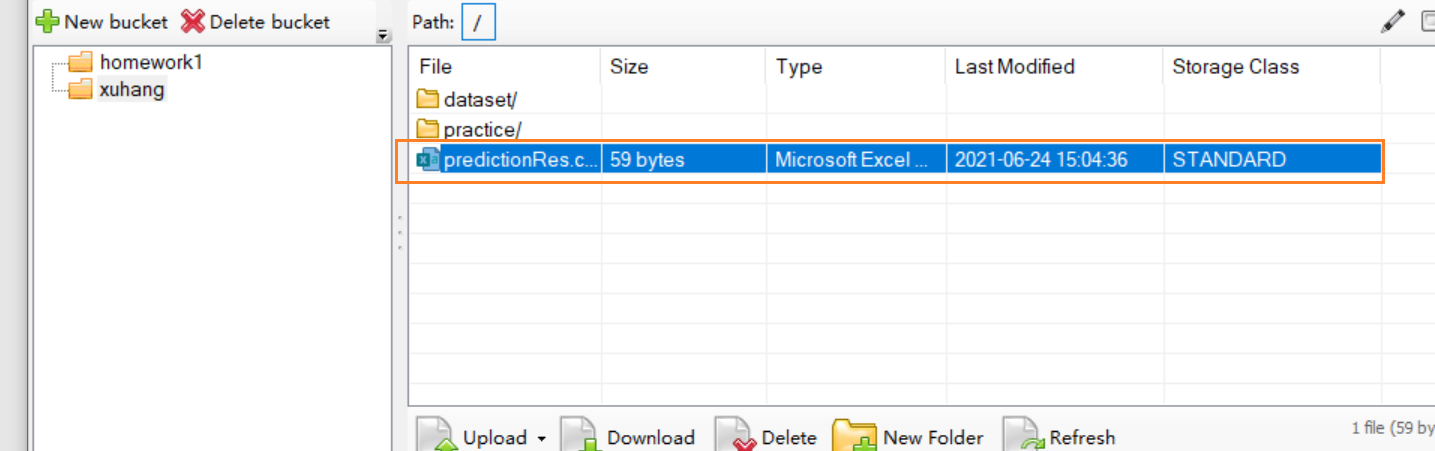
**四、模型预测**



读入训练好的模型预测结果



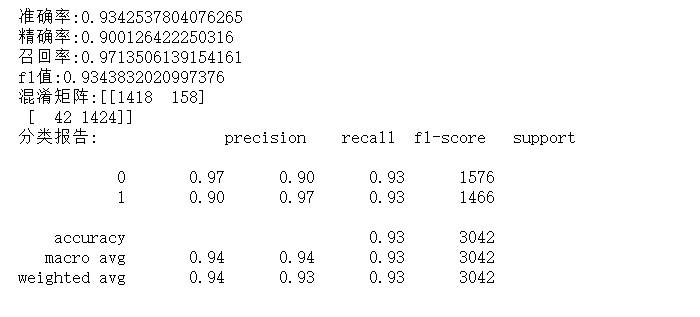
然后将预测结果存入s3中



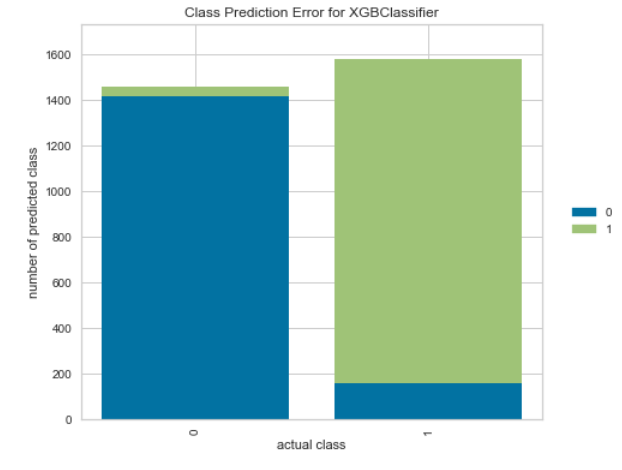
在s3中可以看到相应文件上传成功

**五、模型评估**

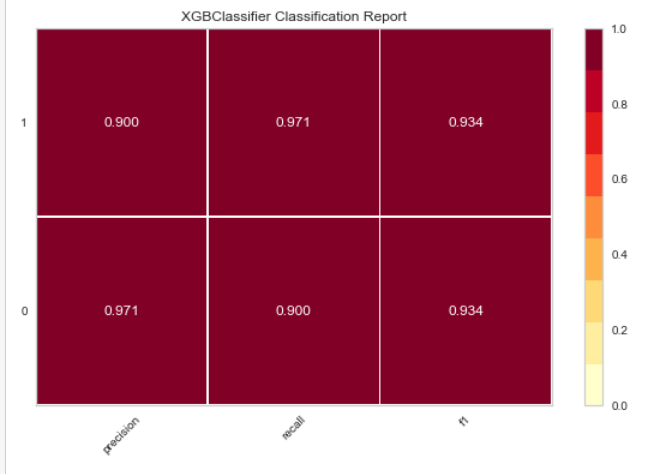
1、准确率、精确率、召回率、f1值、混淆矩阵以及分类报告



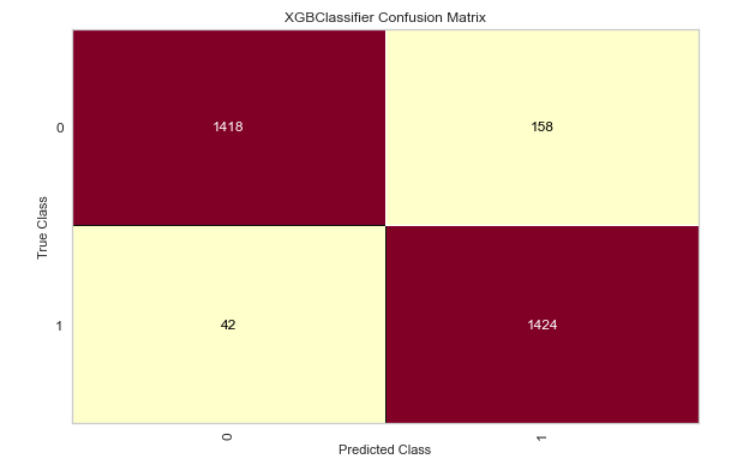
2、分类预测



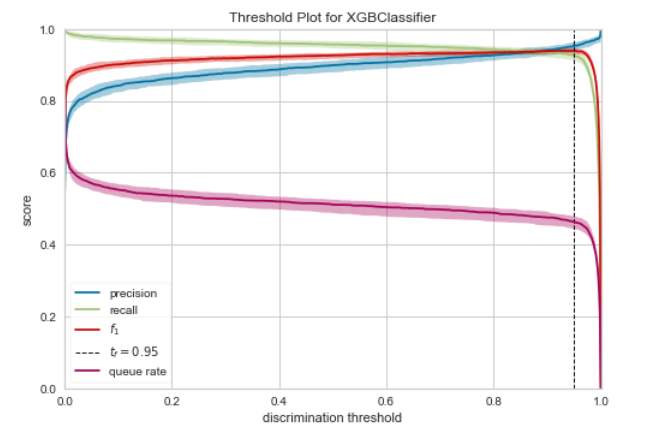
3、分类报告



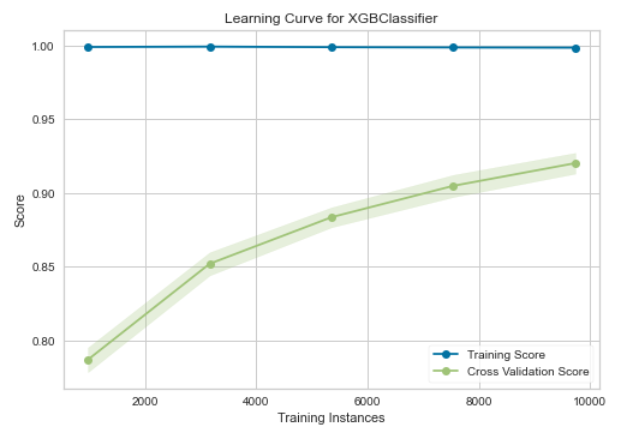
4、混淆矩阵



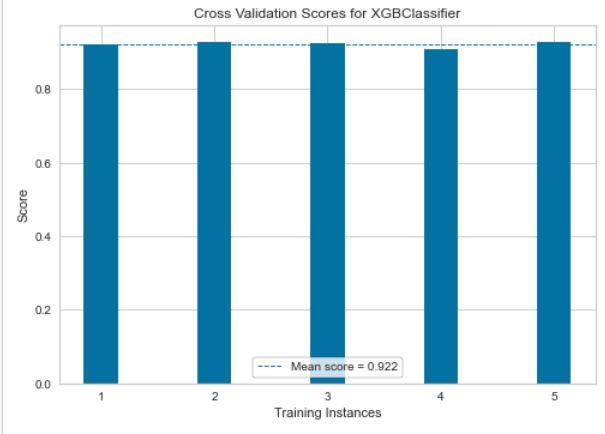
5、阈值选择



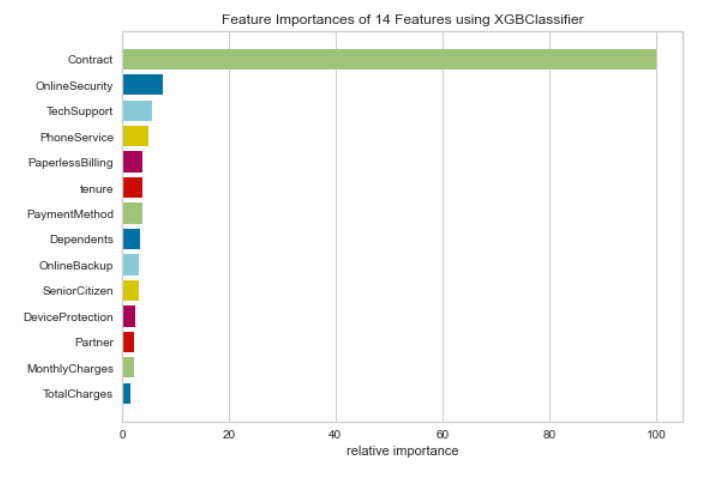
6、学习率



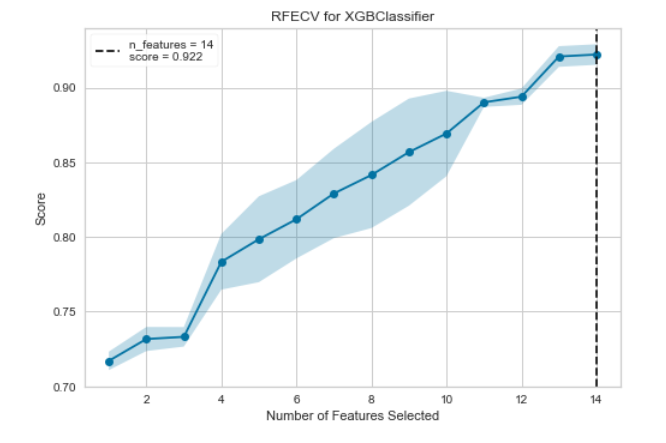
7、交叉验证



8、特征重要性



9、特征递归消减



10、特征选择

