# 基于**机器学习方法的电信用户流失的挖掘分析报告**

**目 录**

[一、项目背景 1](#_Toc185368579)

[二、运行环境 1](#_Toc185368580)

[三、技术路线 1](#_Toc185368581)

[四、数据准备 1](#_Toc185368583)

[4.1源数据描述 1](#_Toc185368584)

[4.2数据预处理 7](#_Toc185368585)

[4.3数据预处理结果展示 10](#_Toc185368586)

[4.4探索性数据分析 11](#_Toc185368587)

[五、建模分析 19](#_Toc185368588)

[5.1 数据划分 19](#_Toc185368589)

[5.2 绘制热力图观察变量之间的相关性 19](#_Toc185368590)

[5.3绘制churn与变量关系图 20](#_Toc185368591)

[5.4 特征选择 20](#_Toc185368592)

[5.5 模型选择与参数调优 21](#_Toc185368593)

[5.6 评估策略 22](#_Toc185368594)

[5.7运行结果展示与分析 22](#_Toc185368595)

[5.8分析总结 27](#_Toc185368596)

[六、创新性 29](#_Toc185368597)

[七、项目总结 29](#_Toc185368598)

# 一、项目背景

随着市场饱和度的上升，电信运营商的竞争也越来越激烈，再加上高昂的客户获取成本，流失分析就变得非常关键。流失率是一种指标，用于描述取消或未续订公司套餐的客户数量。对于客户流失率而言，每增加5%，利润就可能随之降低25%-85%。因此，如何减少电信客户流失的分析与预测至关重要。基于从客户流失分析中获得的信息，电信公司可以制定战略、瞄准细分市场，提高所提供服务的质量以改善客户体验，从而培养客户的信任度。

# 二、运行环境

操作系统：Windows11

开发环境：python3.10,JupyterNotebook

# 三、技术路线

# 

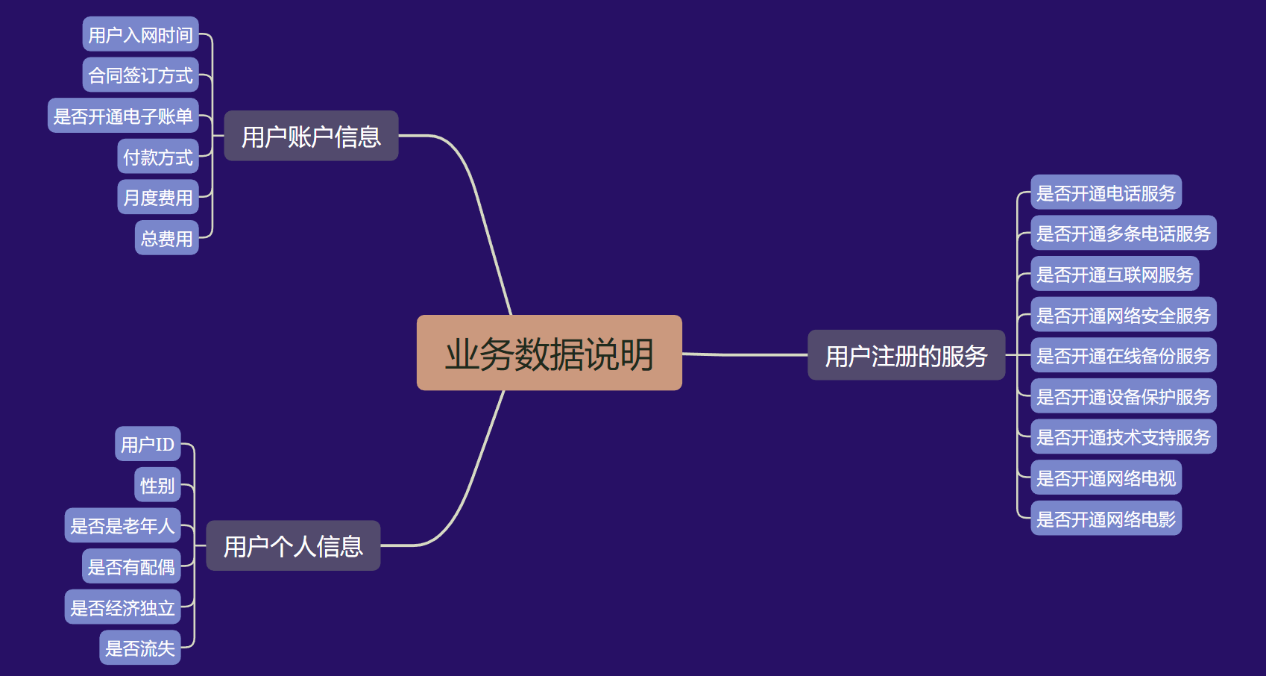
数据预测流程图

# 四、数据准备

## 4.1源数据描述

1. **数据集描述**

本数据集为电信客户流失情况，包括三个部分用户账户信息、用户个人信息、用户注册的服务



用户账户信息

tenure ：任期

MonthlyCharges：月费用

TotalCharges：总费用

Contract：签订合同方式 （按月，一年，两年）

PaperlessBilling：是否开通电子账单（Yes or No）

PaymentMethod：付款方式（bank transfer，credit card，electronic check，mailed check）

用户注册的服务

PhoneService ：是否开通电话服务业务 （Yes or No）

MultipleLines：是否开通了多线业务（Yes 、No or No phoneservice）

InternetService：是否开通互联网服务 （No, DSL数字网络，fiber optic光纤网络）

OnlineSecurity：是否开通网络安全服务（Yes，No，No internetserive）

OnlineBackup：是否开通在线备份业务（Yes，No，No internetserive）

DeviceProtection：是否开通了设备保护业务（Yes，No，No internetserive）

TechSupport：是否开通了技术支持服务（Yes，No，No internetserive）

StreamingTV：是否开通网络电视（Yes，No，No internetserive）

StreamingMovies：是否开通网络电影（Yes，No，No internetserive）

用户个人信息

Churn：该用户是否流失（Yes or No）

customerID ：用户ID。

gender：性别。（Female & Male）

SeniorCitizen ：老年人 （1表示是，0表示不是）

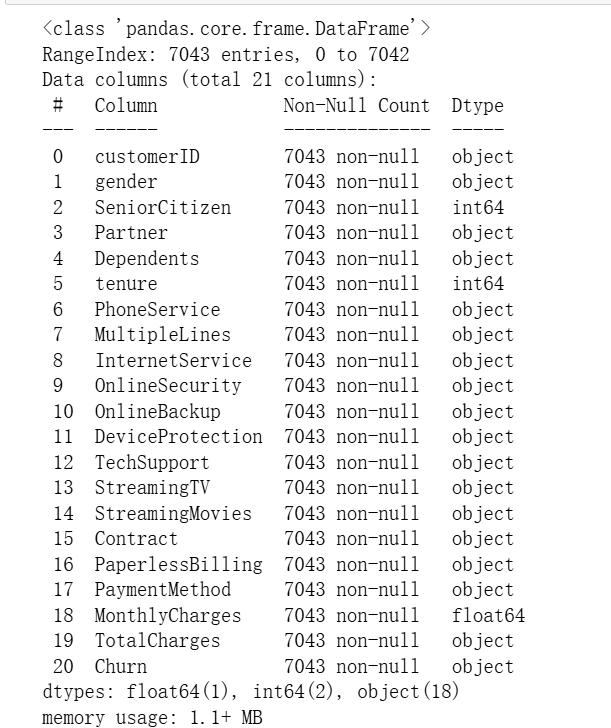
Partner ：是否有配偶 （Yes or No）

Dependents ：是否经济独立 （Yes or No）



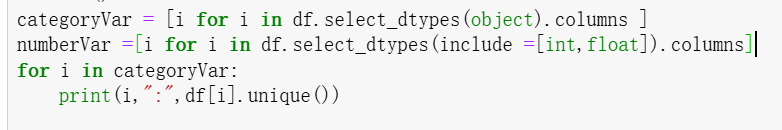
使用info函数可以看到本数据集包括7043条数据，21个变量。数据类型包括object，int64，float64

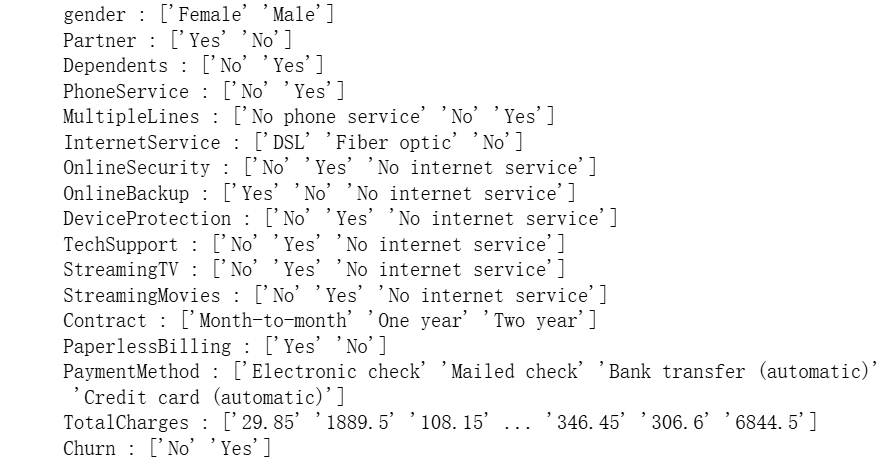
发现MonthlyCharges：月费用的数据类型错误应为浮点数类型

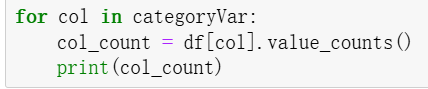


1. **数据描述**
2. **查看分类属性变量的取值情况**

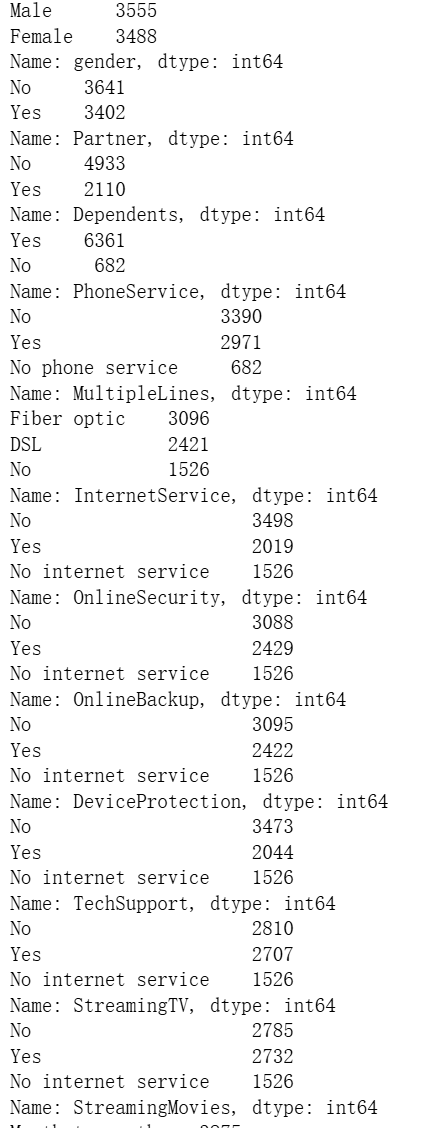
划分数值属性和分类属性

****

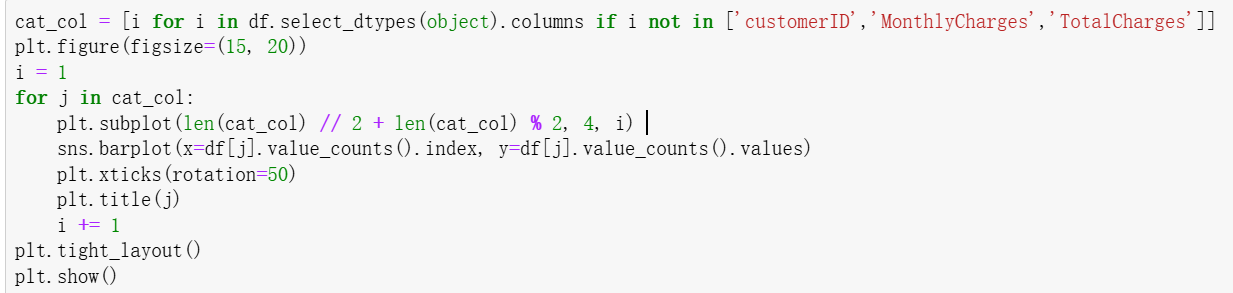
****

****

查看列值分布情况



使用条形图可视化





发现有许多的二分属性还有No internetserive的列值

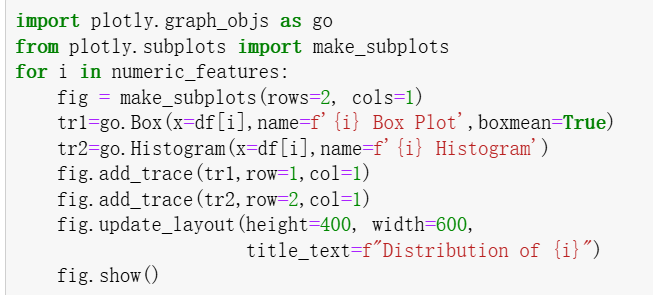
1. **连续属性的描述信息**

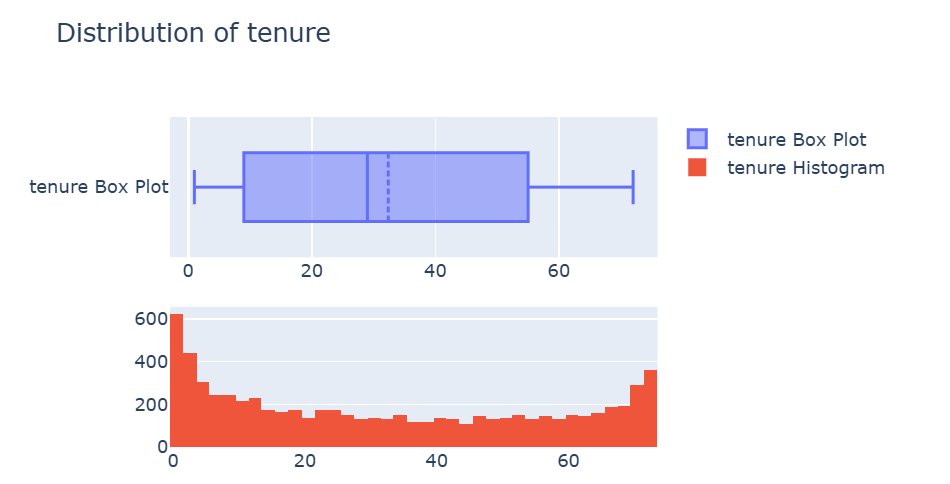
**、**

可以看出 MonthlyCharges"、"TotalCharges"两个特征跟其他特征相比，量纲差异较大

1. **连续属性数据分布情况**

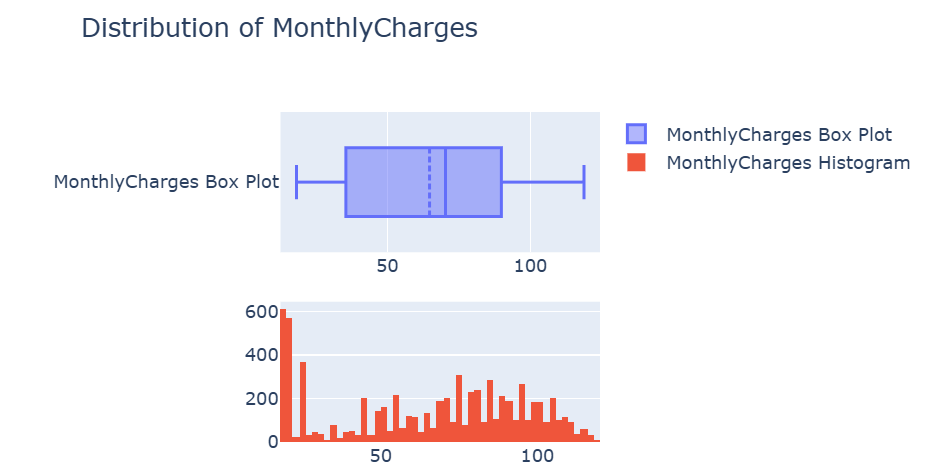
使用箱形图和直方图展现连续属性数据分布，最大值，最小值与中值

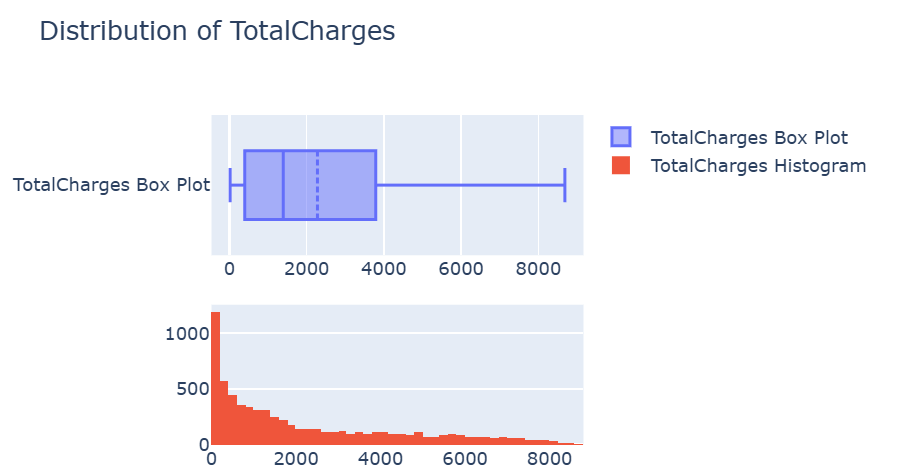
****

****

用户入网时间呈现两端高的情况，可能是入网初期有非常有吸引力的促销活动，这些优惠活动吸引了大量对价格敏感导致短期内入网人数增多。

随着时间的推移，运营商在网络质量、客户服务等方面建立了良好的口碑，对于一些对稳定性和长期服务有需求的用户来说，会选择长期保留。

****

****

消费金额所出现的波动，有可能是套餐定价策略引起

1. **预测目标变量的分布情况**

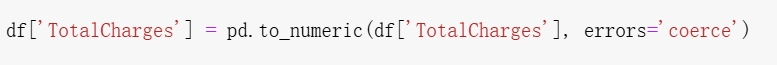
****

可以看出yes值（留存）占数据集的73.5%，NO值（流失）占数据集的26.5%存在明显的样本不均衡问题

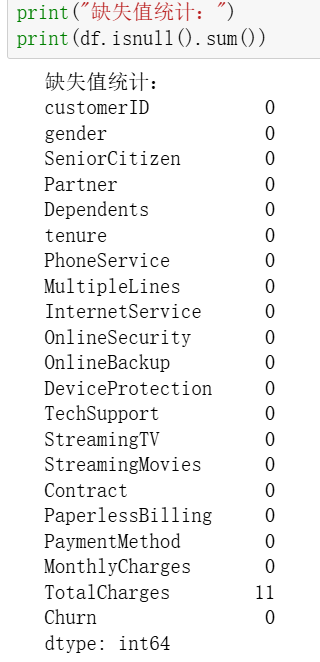
## 4.2数据预处理

**1.转换数值类型**

将MonthlyCharges月费用的数据类型转换为浮点数类型

****

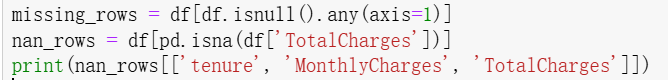
**2.缺失值处理**

****

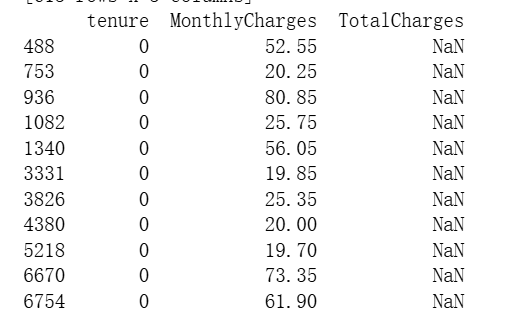
发现TotalCharges总消费额有11个用户数据缺失值为NaN

对连续属性缺失值的处理方法有多种，列入如使用中位数，众数平均数，回归填充或删除等操作

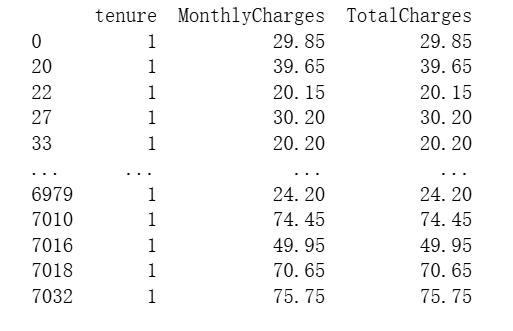
但是通过对缺失值列和其他列的观察可以发现缺失值的规律

****

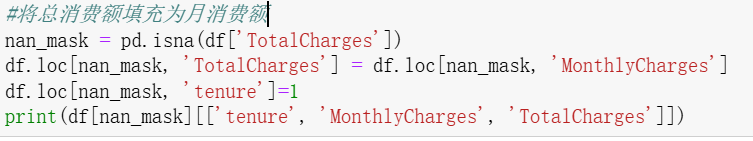
打印出包含缺失值的列，可以看出它们的tenure使用月数都为0，推测他们是本月新注册的用户。可能注册时长还没有过一个月周期，因此TotalCharge值为NaN

****

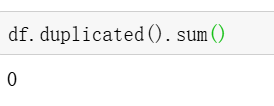
再打印出tenure使用月数为1的用户数据  
可以发现由于经过一个月周期TotalCharge变为MonthlyChargers月费用的值并且tenure变为1。可以证实我们的猜测

****

**因此我们使用MonthlyChargers的值来填充TotalCharge的缺失值并把tenure变为1**

****

**3.重复值检查**

****

**4.数据编码**

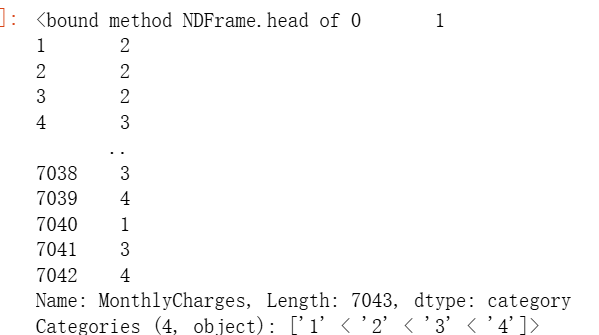
**(1).** **处理量纲差异**

“MonthlyCharges"、"TotalCharges"量纲差异大通过分析得分我们这里选择使用离散化的方式

用四分位数进行离散

df['MonthlyCharges']=pd.qcut(df['MonthlyCharges'],4,labels=['1','2','3','4'])

df['TotalCharges']=pd.qcut(df['TotalCharges'],4,labels=['1','2','3','4'])

****

**（2）分类属性编码**

由于“No internetserive”的人数占比OnlineSecurity，OnlineBackup，DeviceProtection，TechSupport，StreamingTV，StreamingTV占比一致可以判断No internetserive”不影响预测结果将 “No internetserive” 并到 “No”里面。

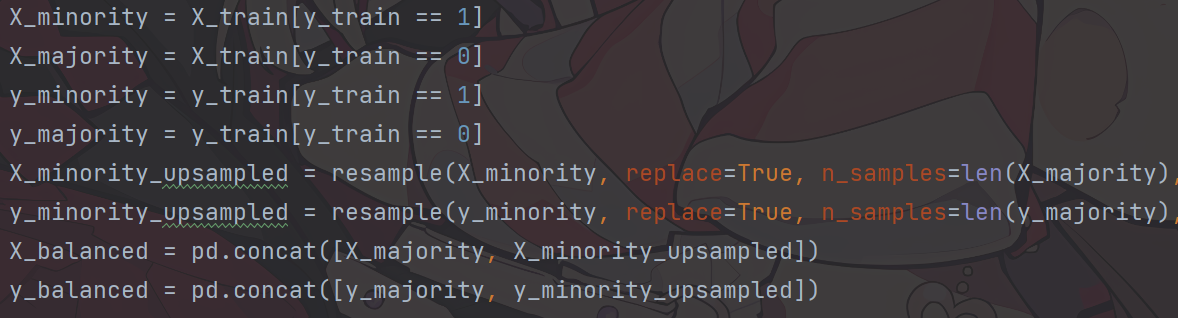
df.replace(to\_replace='No internet service',value='No',inplace=True)  
df.replace(to\_replace='No phone service',value='No',inplace=True)

对分类属性进行序列编码

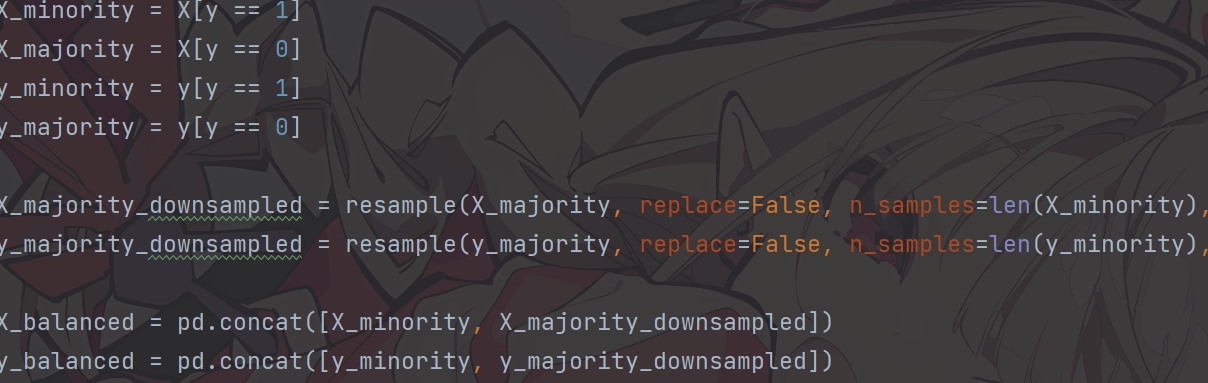
df[x] = LabelEncoder().fit\_transform(df[x])

**（3）处理样本不均**

上采样



下采样

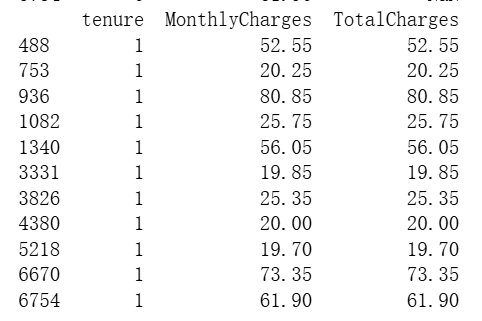


表X 数据预处理方案

|  |  |  |
| --- | --- | --- |
| 阶段 | 预处理目的 | 预处理方案 |
| 1 | 转换数值类型 | 使用to\_numeric |
| 2 | 缺失值处理 | 寻找规律填充 |
| 3 | 重复值检查 | Duplicated函数 |
| 4 | 处理量纲差异 | 四分位数离散 |
| 5 | 分类属性编码 | LabelEncoder编码 |
| 6 | 处理样本不均 | 上采样与下采样 |

## 4.3数据预处理结果展示

缺失值处理

****

11个缺失值列填充情况

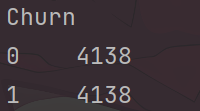


查找缺失值

编码结果：

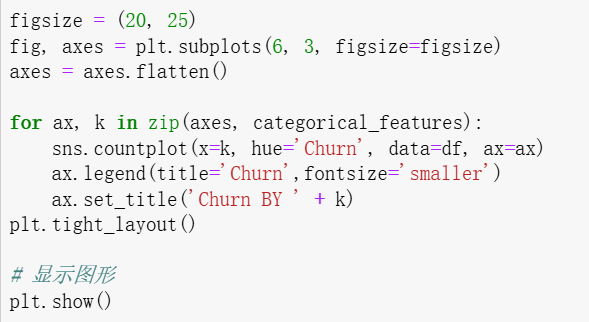


数据平衡结果



## 4.4探索性数据分析

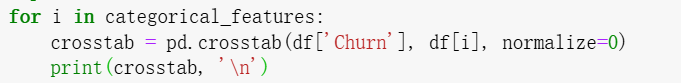
（1）查看分类属性每个标签的对应churn流失的分布情况

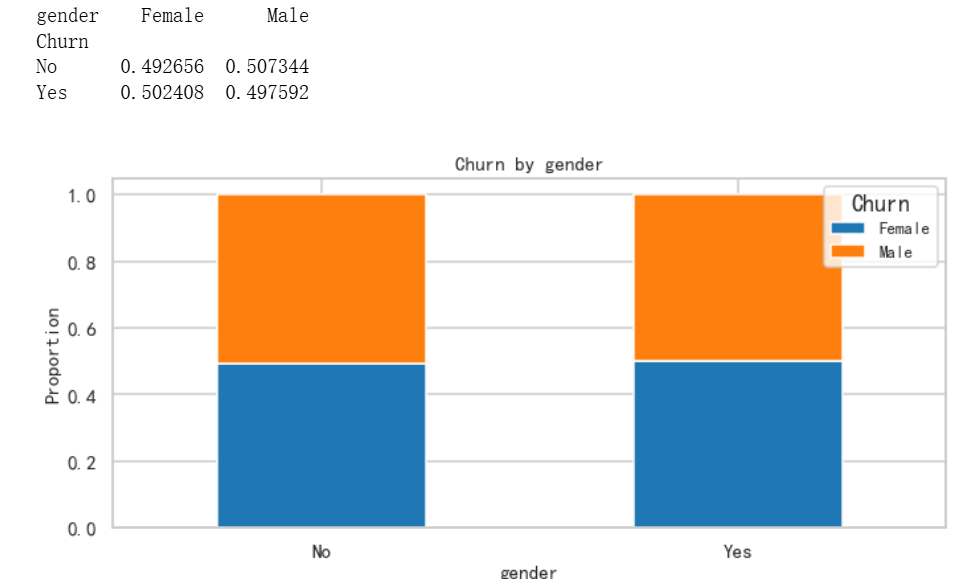




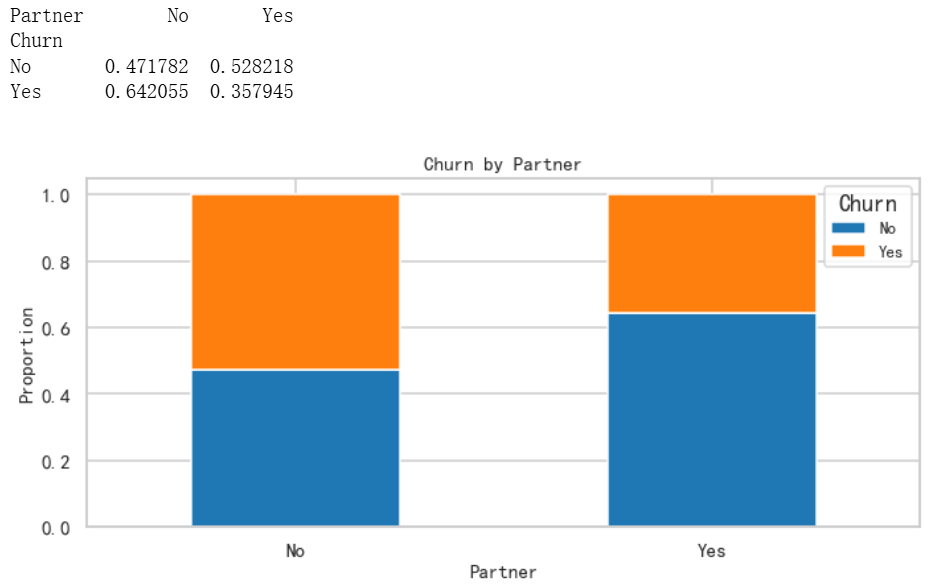
由于样本churn的用户流失比例极不平衡，很难看出维度对流失用户的作用情况

因此我们使用交叉分析的方法使用pd.crosstab()函数来创建与churn的交叉表，

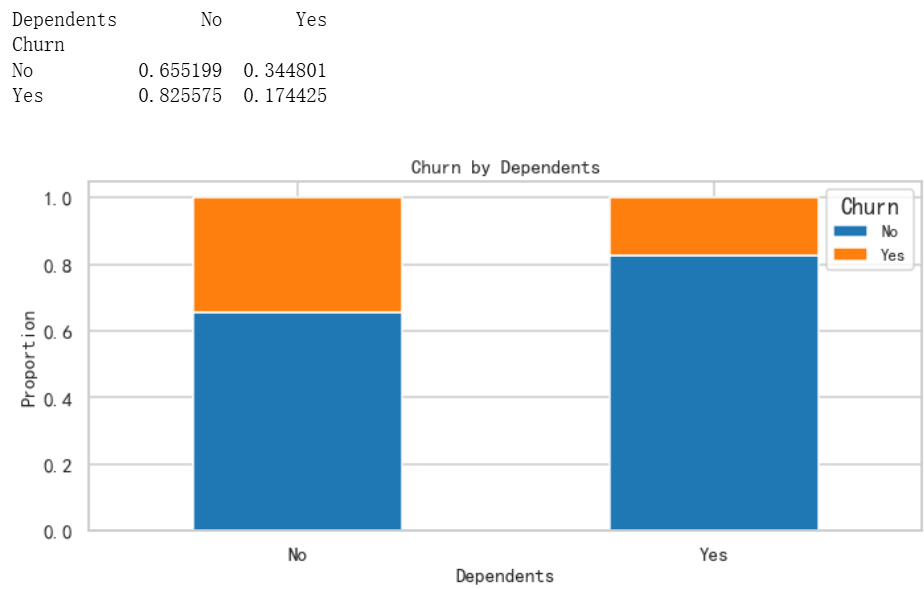




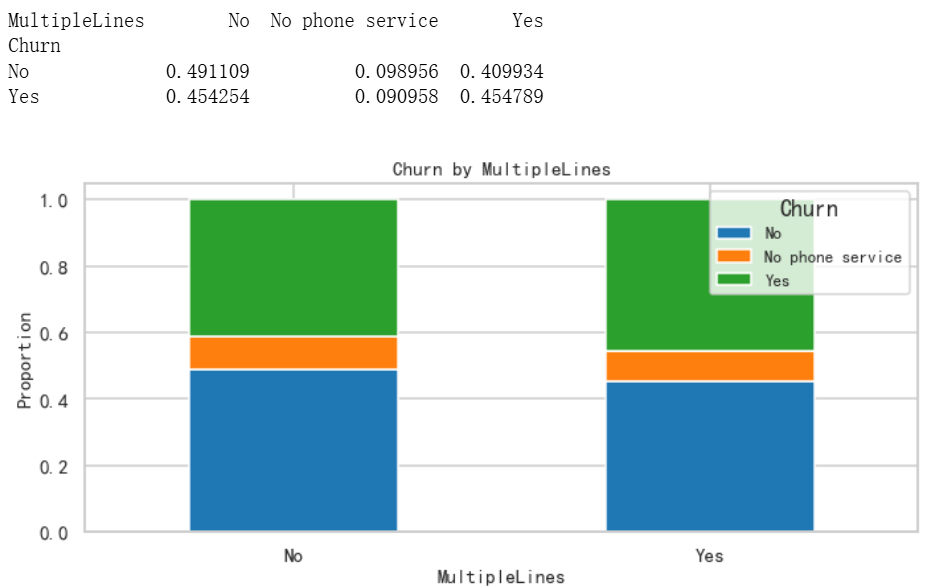
Gender：性别对用户流失留存影响不大



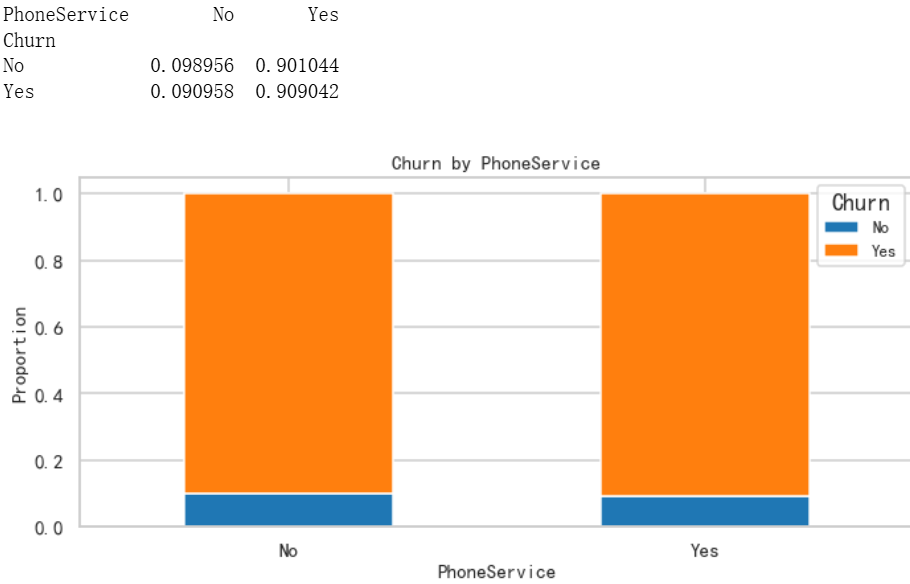
Partner：单身用户更容易流失



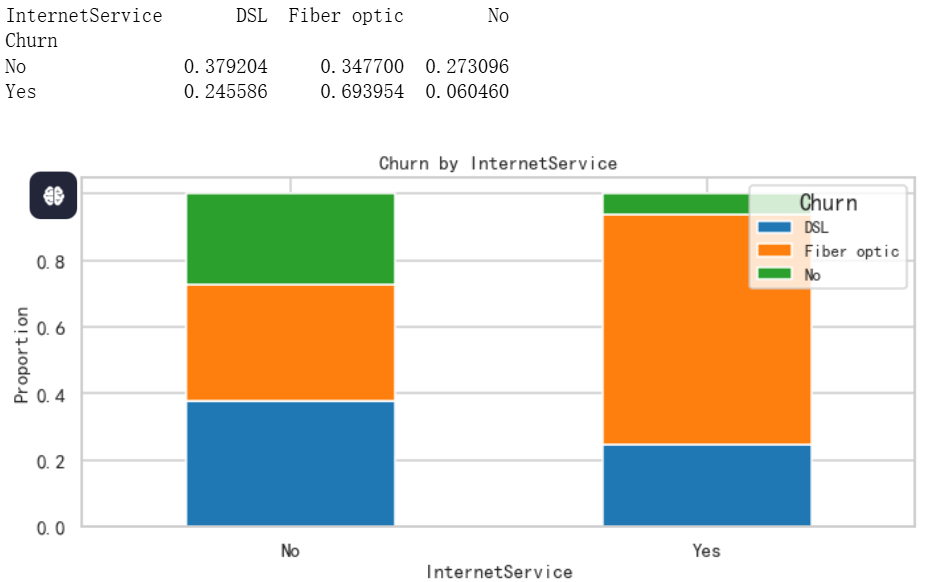
Dependents：经济不独立的用户更容易流失



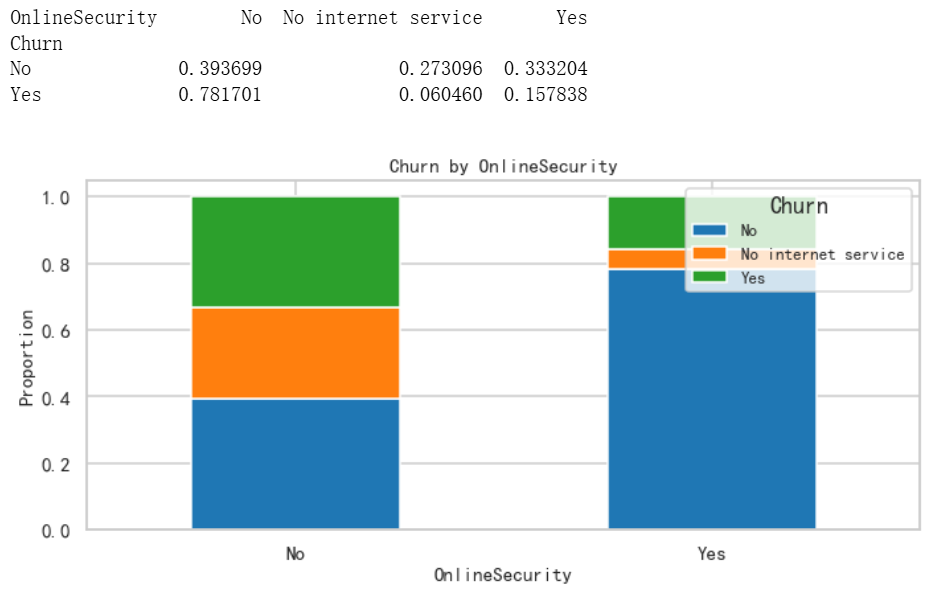
MultipleLines：是否开通MultipleLines对用户流失留存影响不大



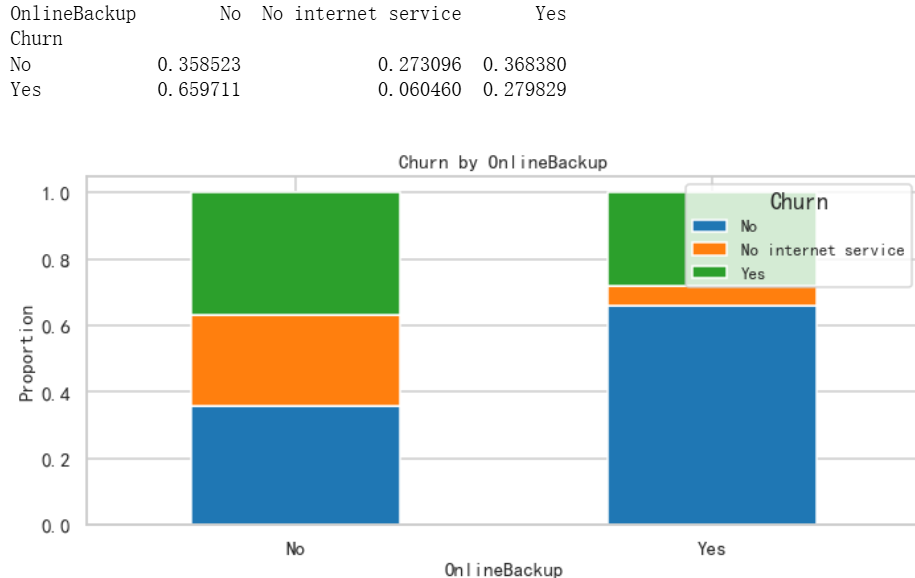
Phoneservice：是否开通Phoneservice对用户流失留存影响不大



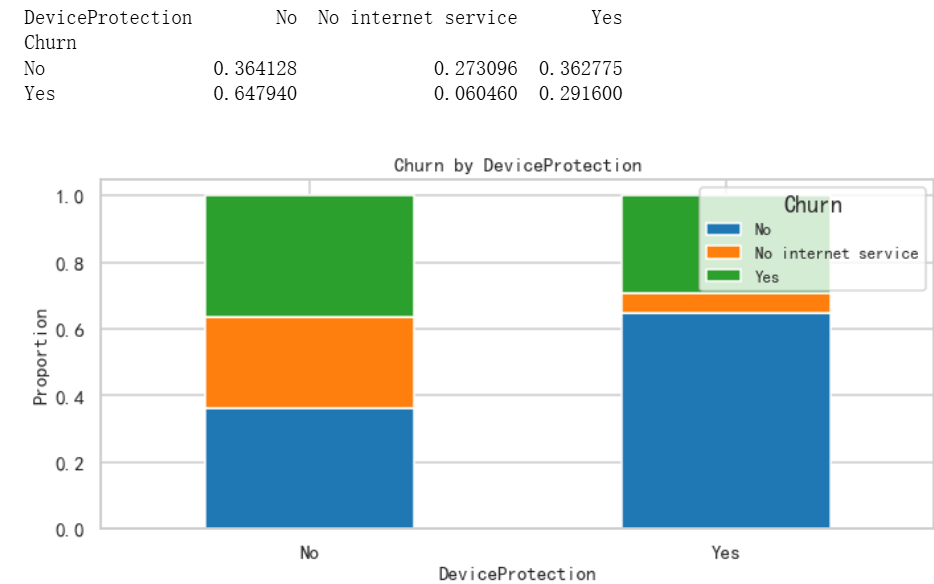
Internerservice：办理了Fiber opti的客户容易流失



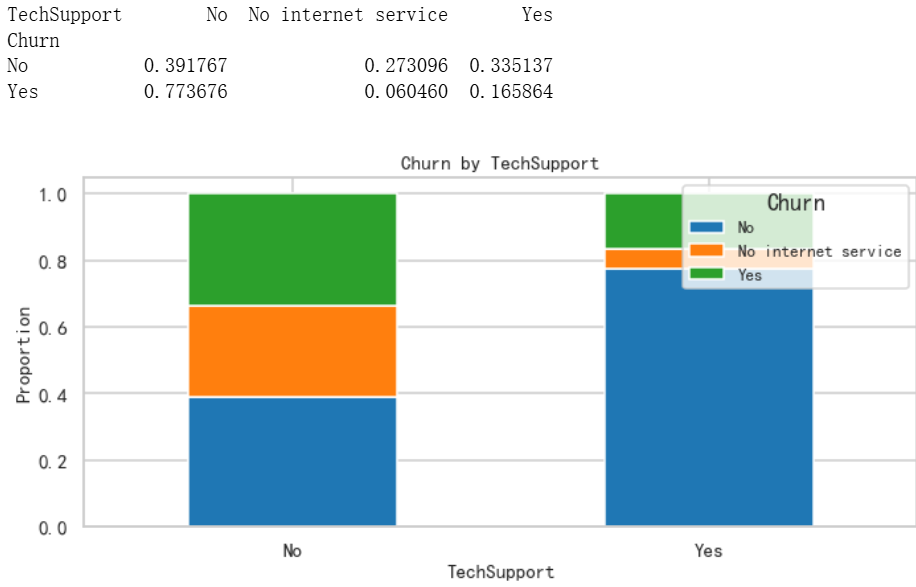
OnlineSecurity：没开通的客户容易流失。



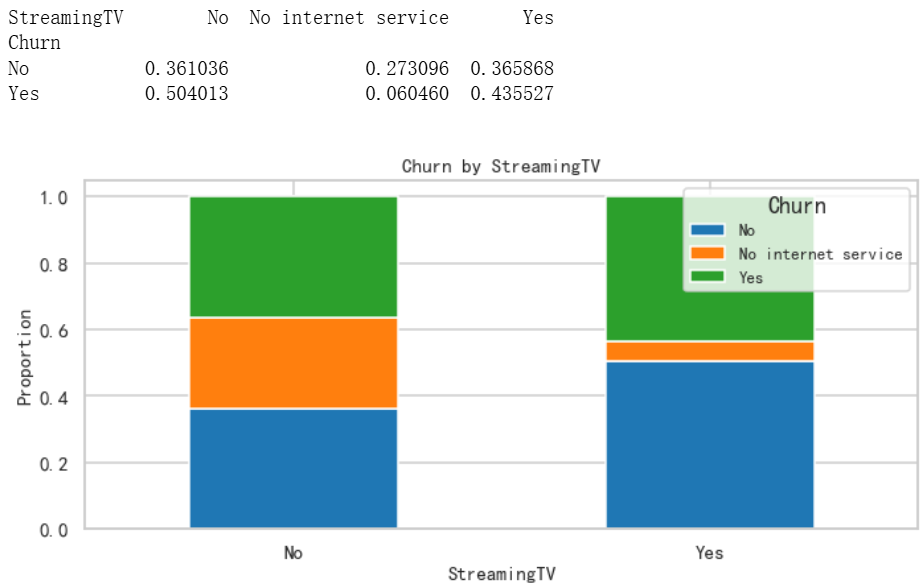
OnlineBackup：没开通在线备份服务的客户容易流失。



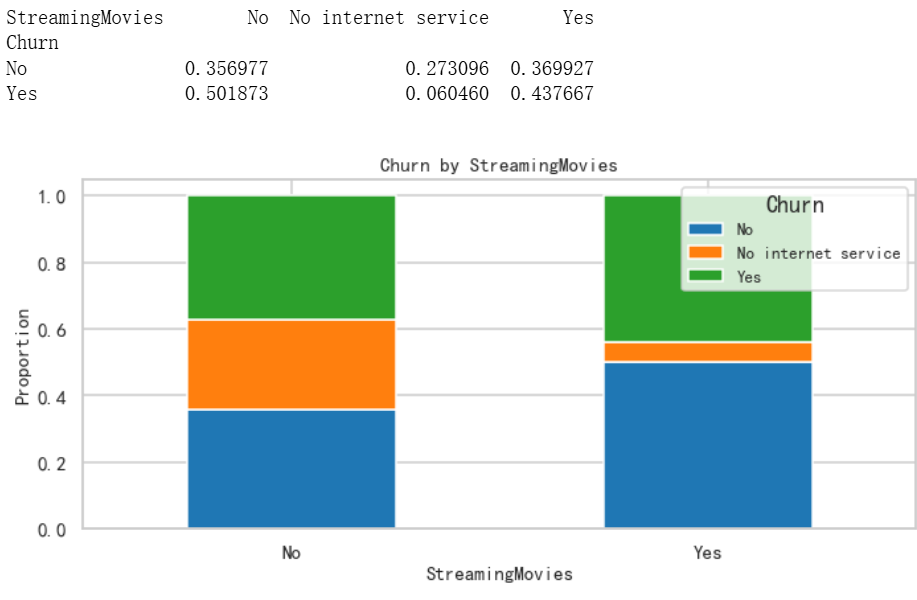
DeviceProtection：没开通设备保护业务的用户比较容易流失



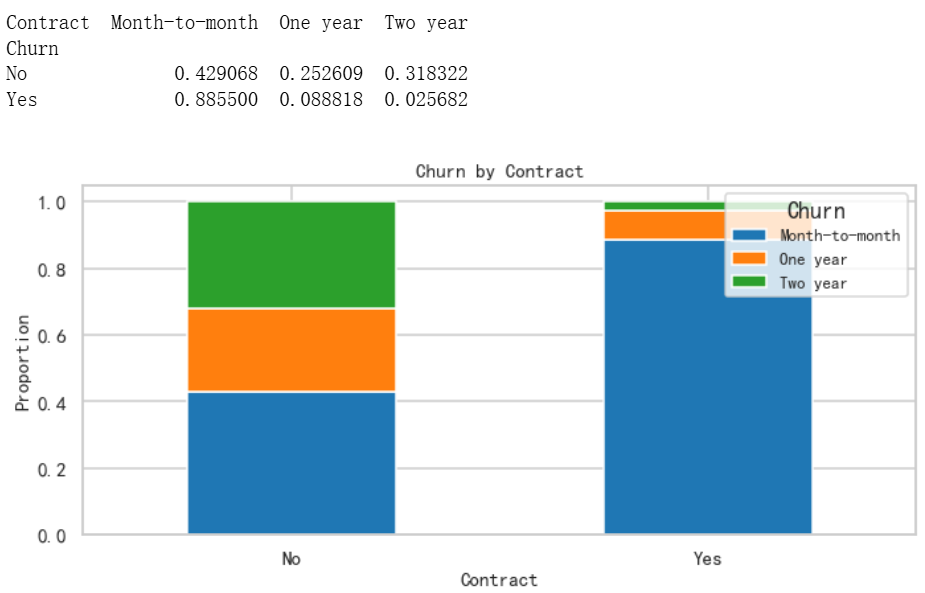
TechSupport：没开通技术支持服务的用户容易流失。



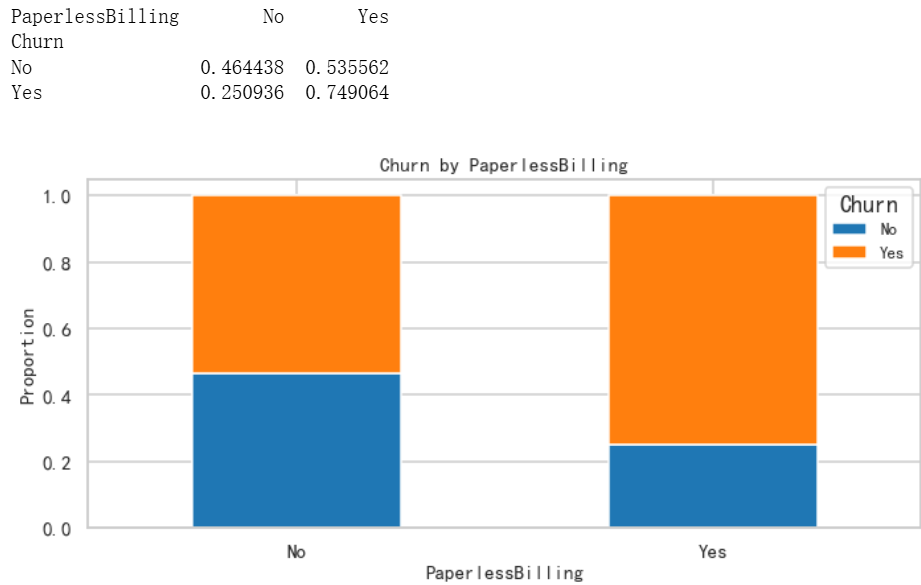
StreamingTV：没有开通网络电视服务的用户更容易流失，但较不显著



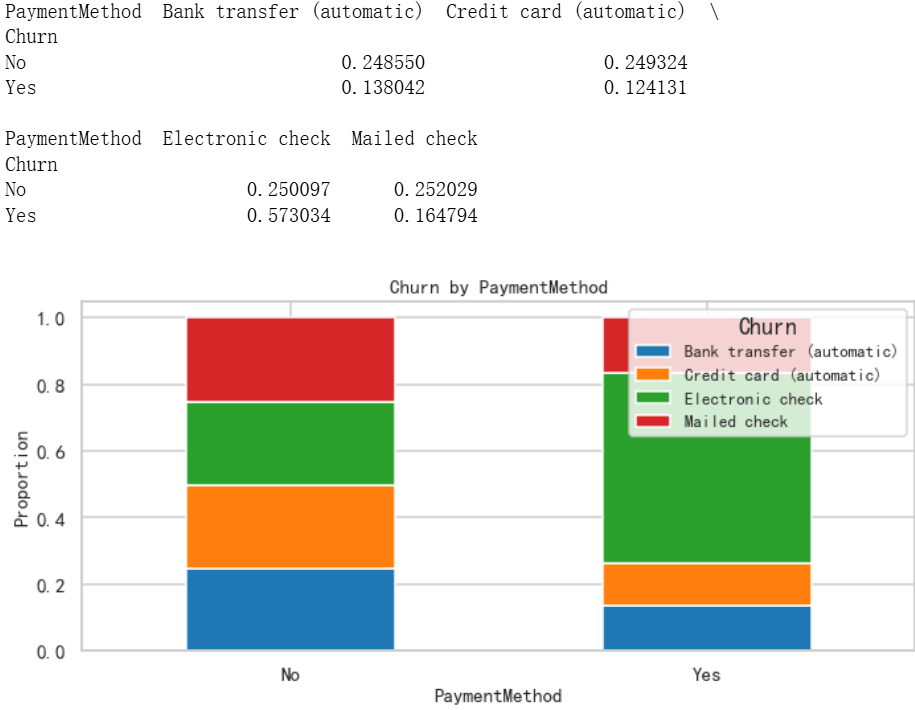
StreamingMovies：没有开通网络电视服务的用户更容易流失，但较不显著



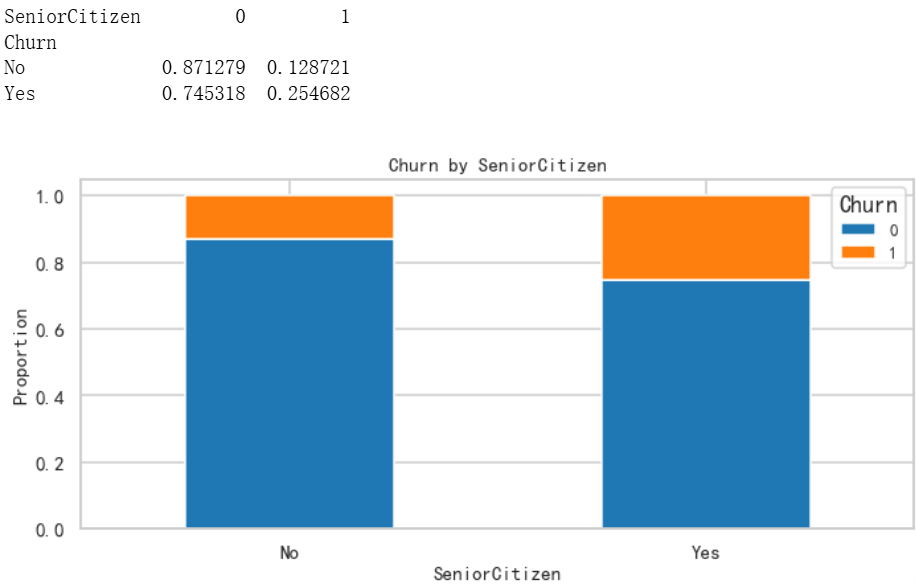
Contract：月份签订合同的用户最容易流失。签约越久越不容易流失



PaperlessBilling ：开通电子账单的用户较容易流失



PaymentMethod：使用电子支票支付的人更容易流失



SeniorCitizen 分析：年轻用户在流失、留存人数占比都高

# 五、建模分析

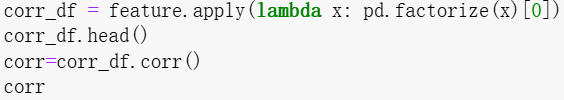
## 5.1 数据划分

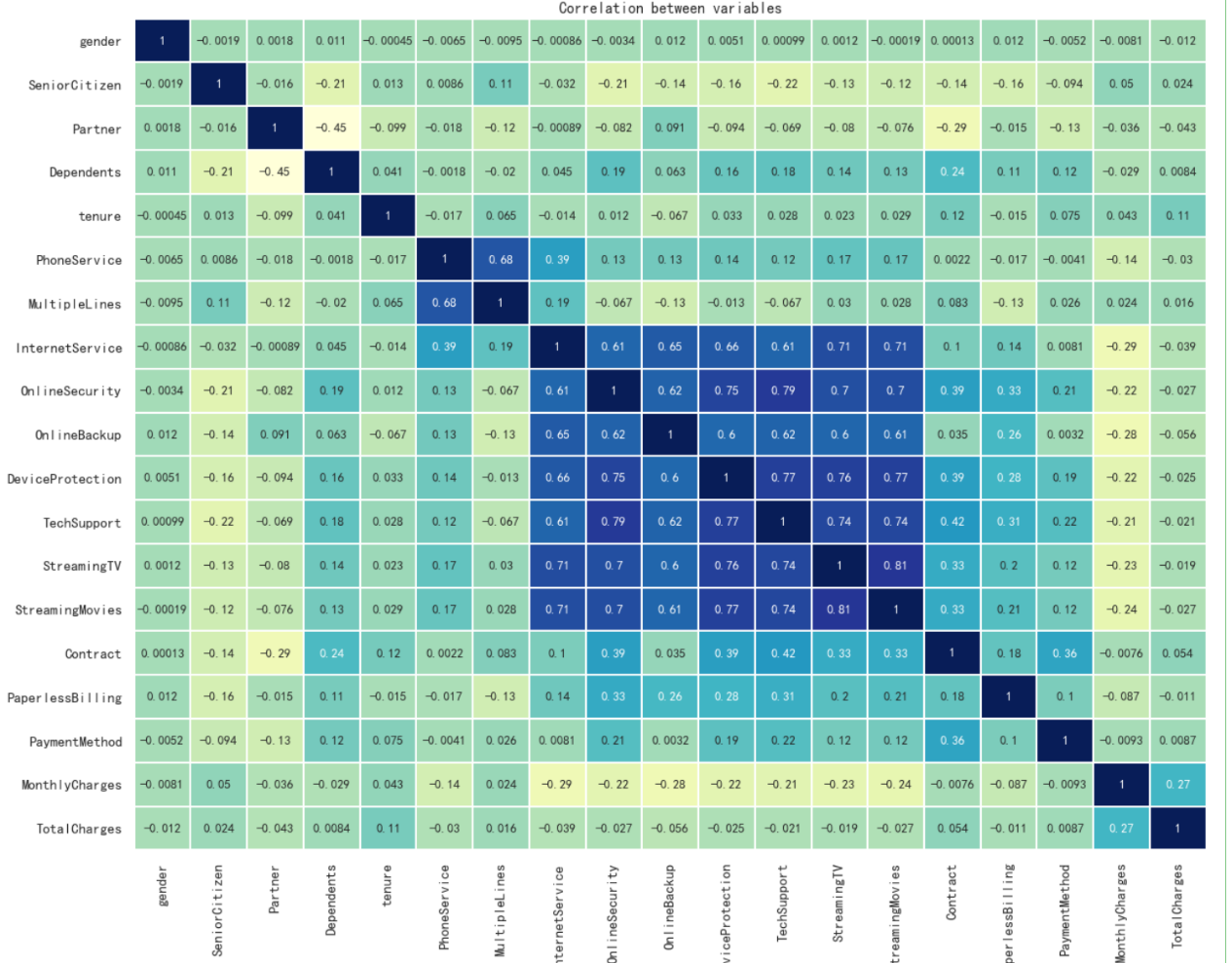
对数据集进行二八拆分为测试集与训练集

X\_train, X\_val,y\_train, y\_val = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

## 5.2 绘制热力图观察变量之间的相关性

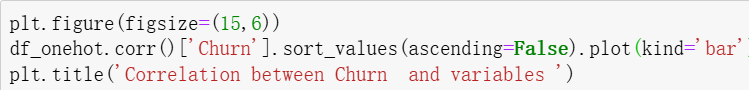
打印相关性矩阵绘制热力图

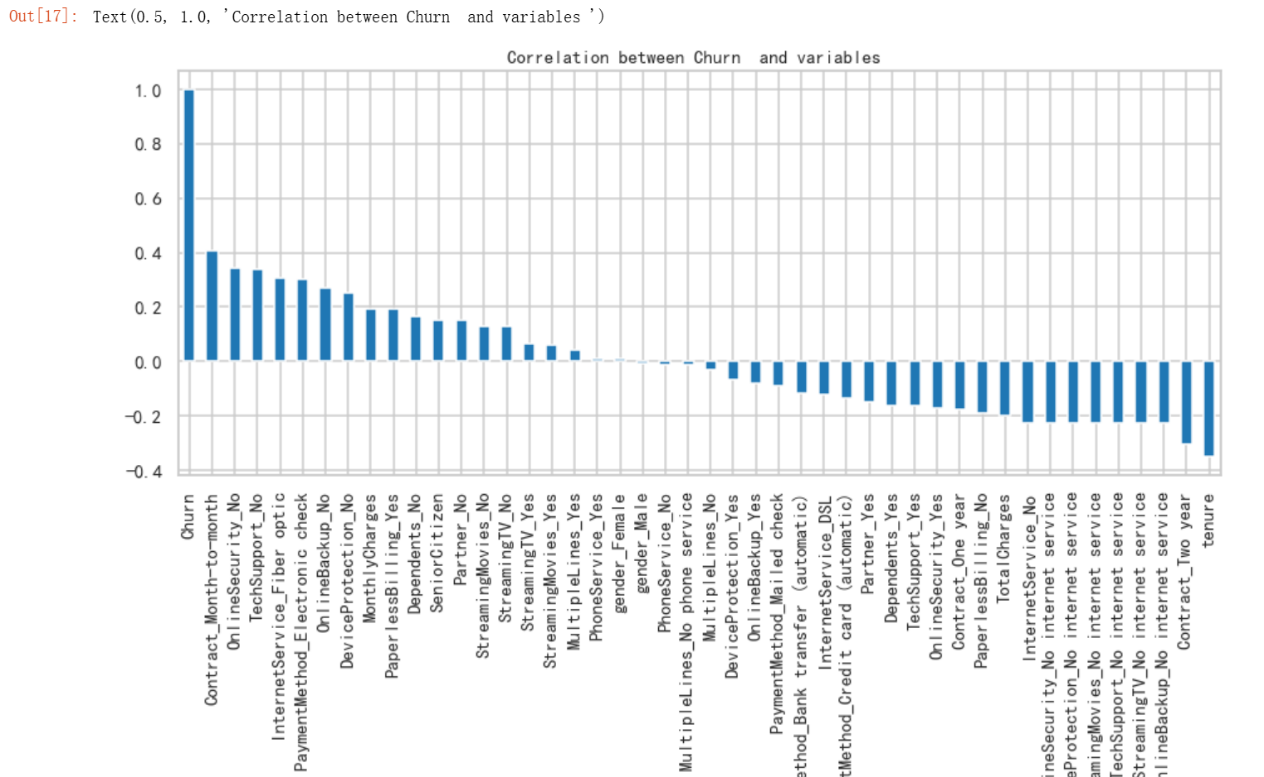




从热力图来看，互联网服务、网络安全、在线备份、设备维护服务、技术支持服务、开通网络电视服务、开通网络电影之间相关性很强，且是正相关。说明在电信业务中这些维度具有一定联系

## 5.3绘制churn与变量关系图





从图看gender（性别）、PhoneService（电话服务）相关性几乎为0，两个维度在特征选择时可以忽略。

## 5.4 特征选择

使用随机森林选择进行特征选择，并获取特征选择后的获取选择的特征索引，得到新的训练集

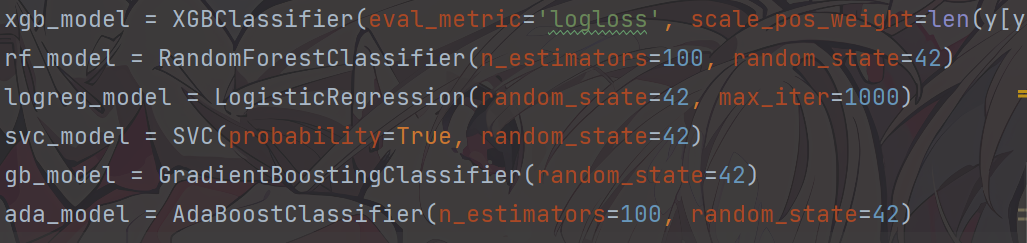
selector = SelectFromModel(RandomForestClassifier(n\_estimators=100, random\_state=42), threshold="median")  
selector.fit(X\_train\_preprocessed, y\_balanced)

特征选择结果

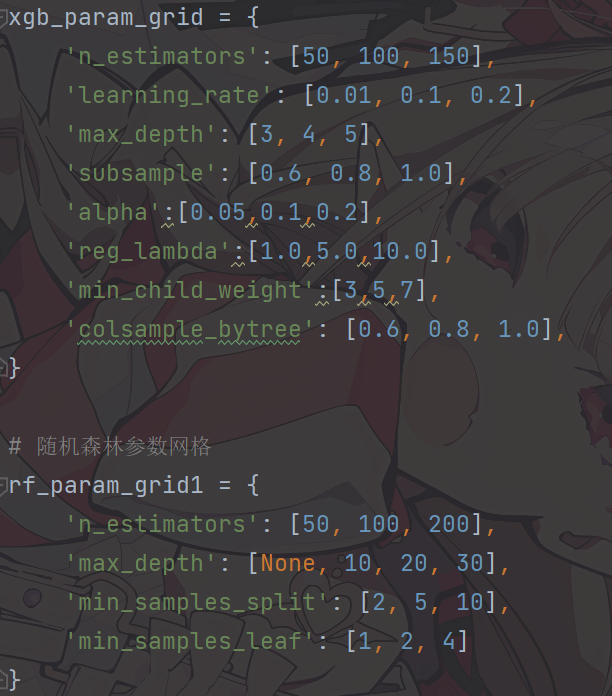
['tenure', 'MonthlyCharges', 'TotalCharges', 'gender', 'Partner', 'InternetService', 'OnlineSecurity', 'Contract', 'PaperlessBilling', 'PaymentMethod']

## 5.5 模型选择与参数调优

选择定义以下分类模型进行参数调优



然后对每模型设计参数网格进行超优调参

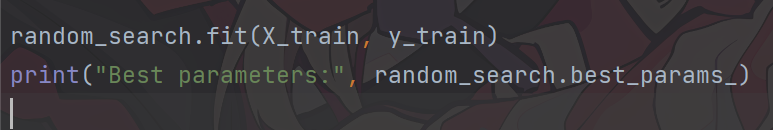


这里我使用RandomizedSearchCV随机搜索在给定的超参数范围内最优的超参数组合。

random\_search=RandomizedSearchCV(estimator=rf\_model,param\_distributions=rf\_param\_grid, n\_iter=1, cv=skf, scoring='roc\_auc', n\_jobs=1)

estimator为使用的分类模型，param\_distributions为超参数取值范围scoring='roc\_auc' 设定评估指标为 ROC AUC

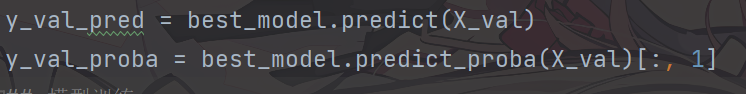
使用StratifiedKFold 进行交叉验证



对定义好的参数搜索模型进行训练并输出最优参数

## 5.6 评估策略

（1）使用训练好的最优模型对测试集进行预测

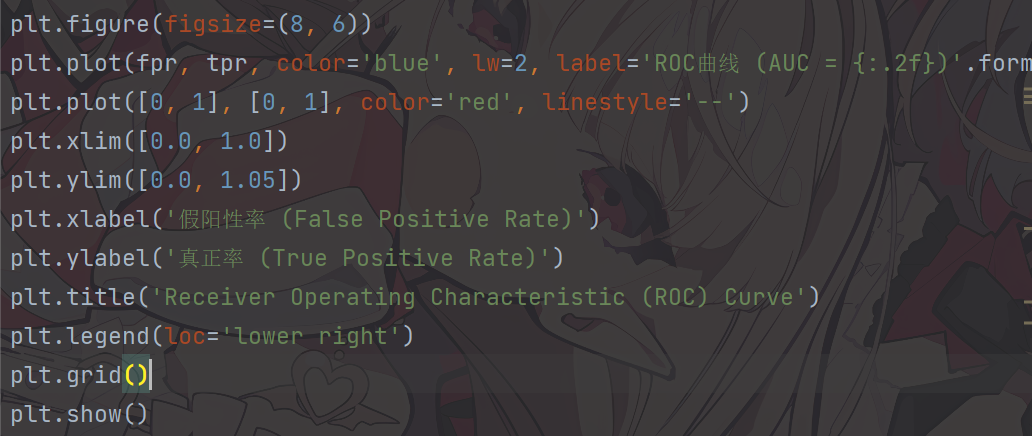


（2）调用混淆矩阵、AUC对模型进行评估

from sklearn.metrics import classification\_report, accuracy\_score, precision\_score, recall\_score, roc\_auc\_score, \  
 roc\_curve, auc

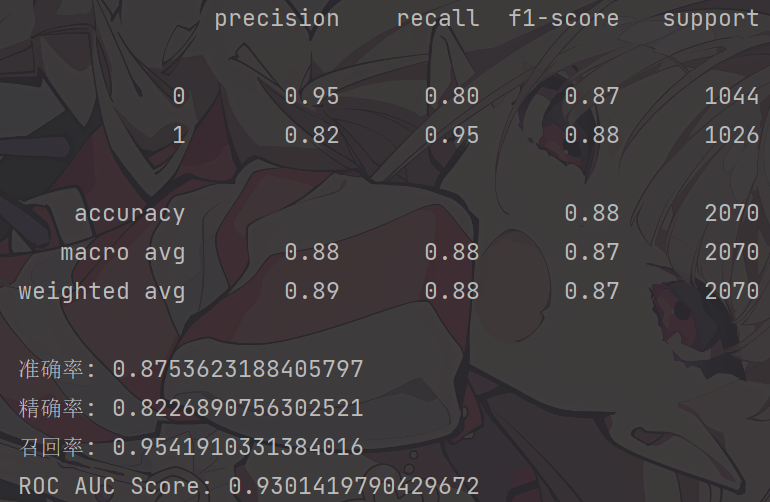
print(classification\_report(y\_val, y\_val\_pred))  
accuracy = accuracy\_score(y\_val, y\_val\_pred)  
precision = precision\_score(y\_val, y\_val\_pred)  
recall = recall\_score(y\_val, y\_val\_pred)

（3）绘制AUC曲线

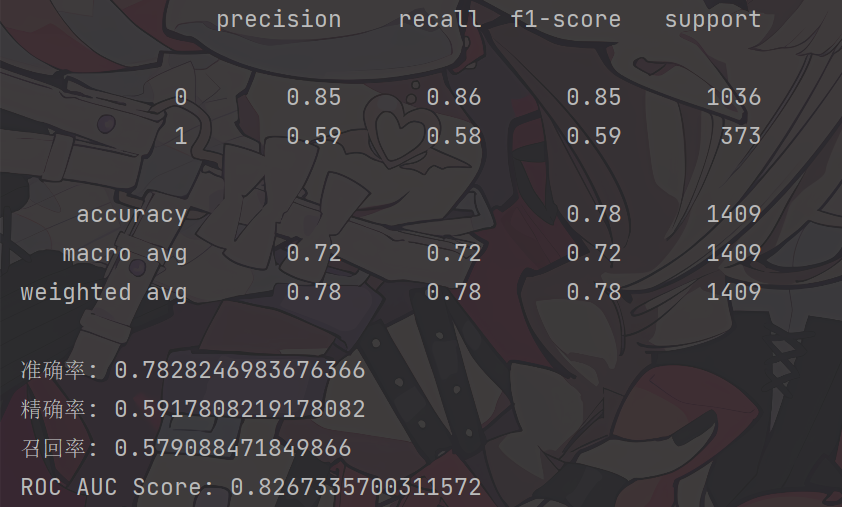


## 5.7运行结果展示与分析

（1）先分后采样与先采样后分操作对比



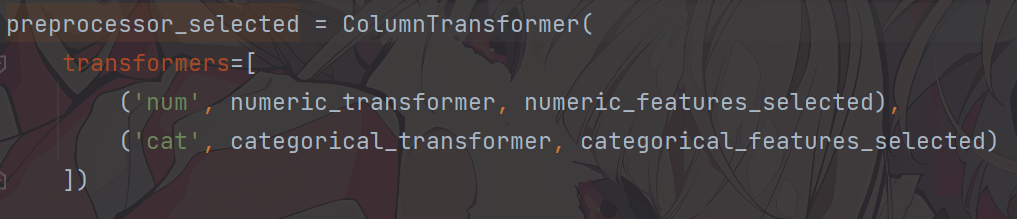
**图1先采样后分评估结果**

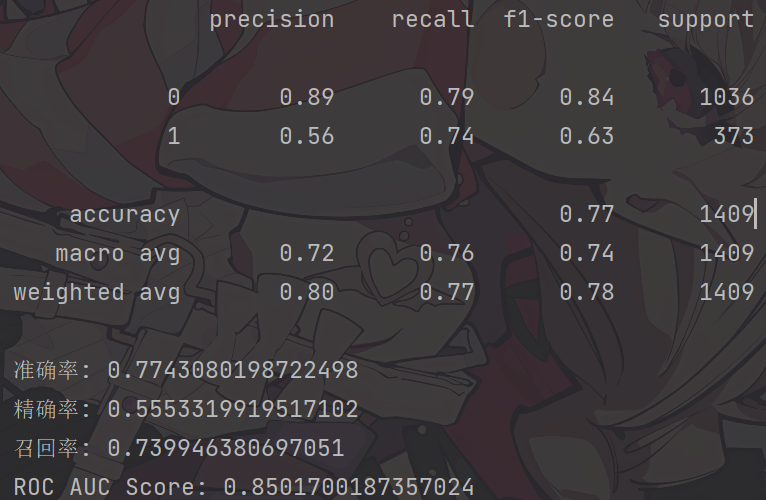


**图2先分后采样评估结果**

从图1与图2的对比中我们可以看出采样与拆分数据集的操作顺序对预测结果的影响巨大，其原因是因为先采样后拆分的数据集里包含了许多我们使用上采样方法生成的样本造成了样本的污染

（2）优化使用管道形式进行数据处理与训练（上采样）



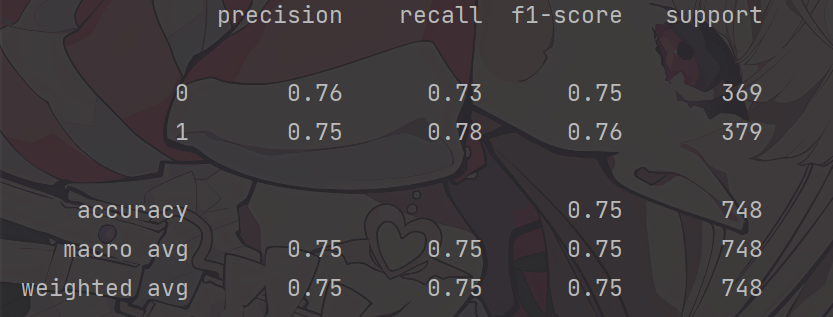




AUC曲线

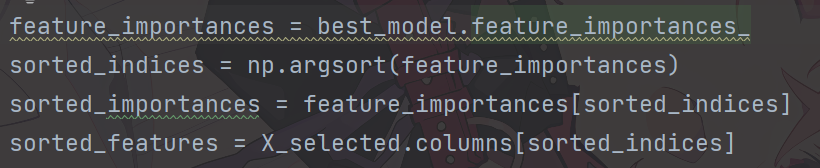
可以看出在使用道进行数据处理与训练后模型评估分数上升

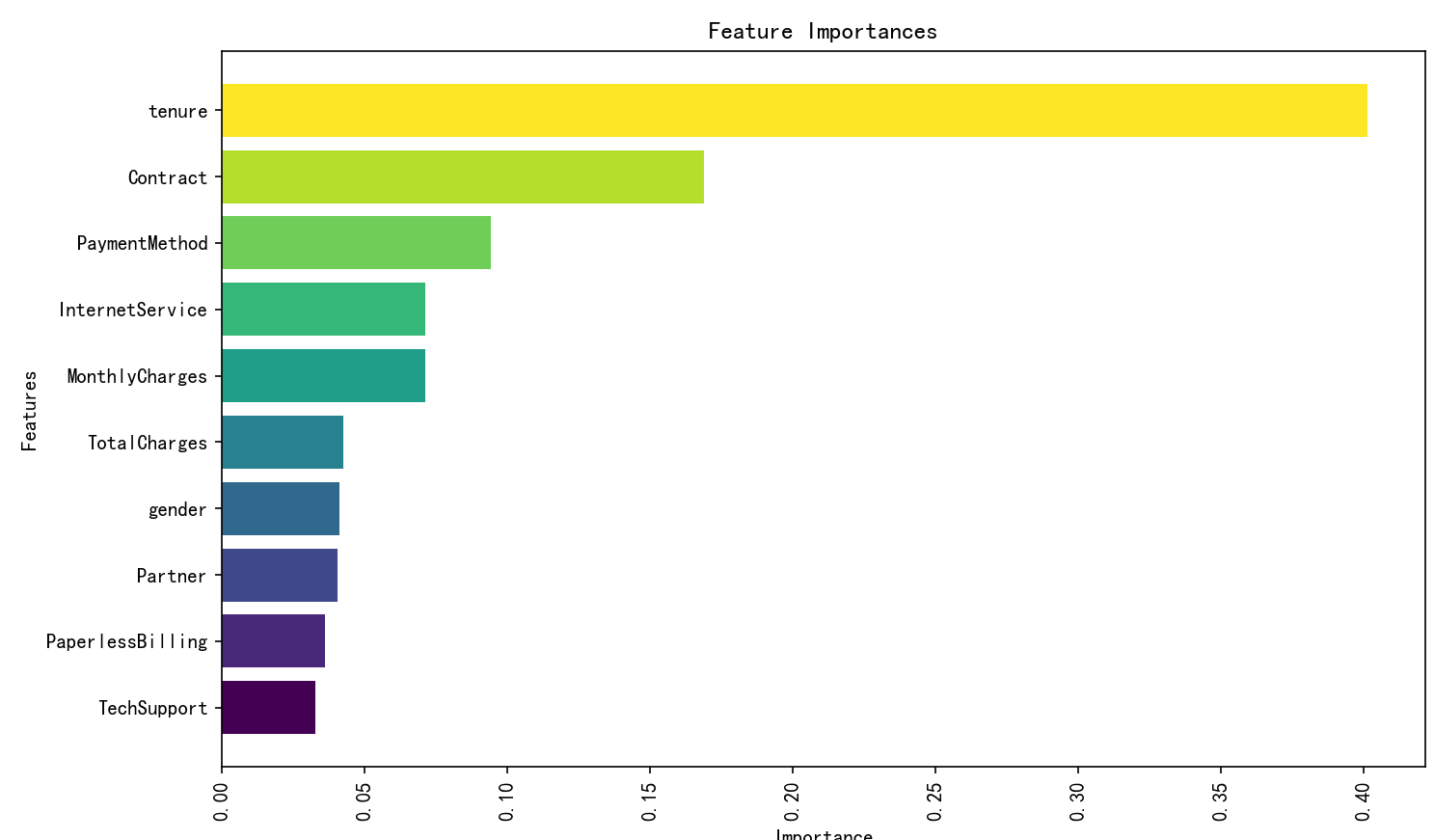
（3）数据处理与训练（下采样）



我们使用下采样对数据平衡后模型对1流失用户的识别率得到大幅提升，但对0的识别率下降

（4）使用最优模型输出特征重要性并画图

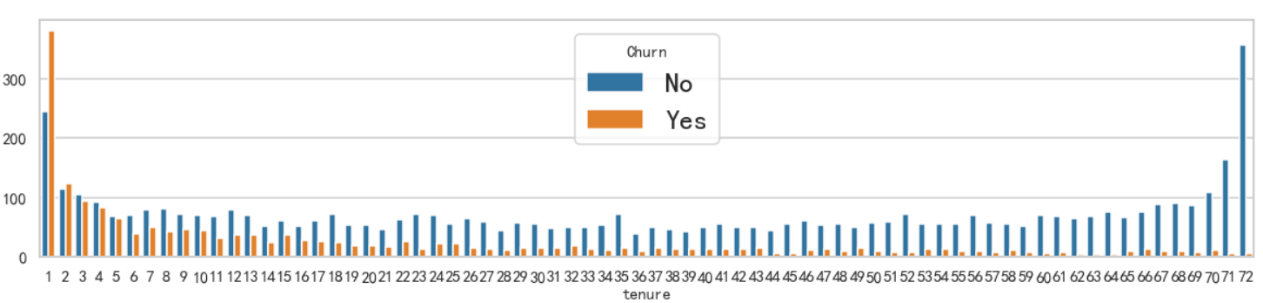




可以看出'tenure', 'MonthlyCharges', 'TotalCharges', 'Partner', 'InternetService', , 'Contract', 'PaperlessBilling', 'PaymentMethod'这些特征对客户是否流失的影响很大

（5）数值属性特征流失分布分析

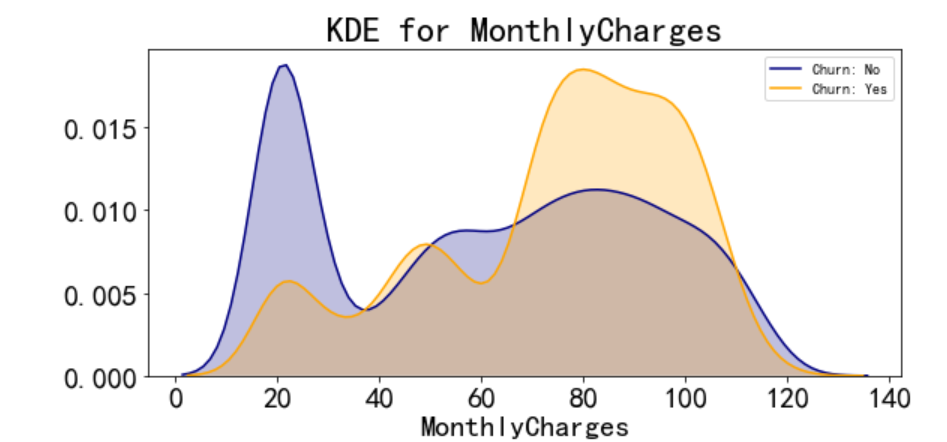
1.针对重要性最高的tenure特征查看其churn变量的分布情况



由图可以看出在任期1-5之间客户的流失量较大，也就是说刚开通电信业务1至5个月的用户容易流失，并且随着tenure任期的增加流失的概率越来越小

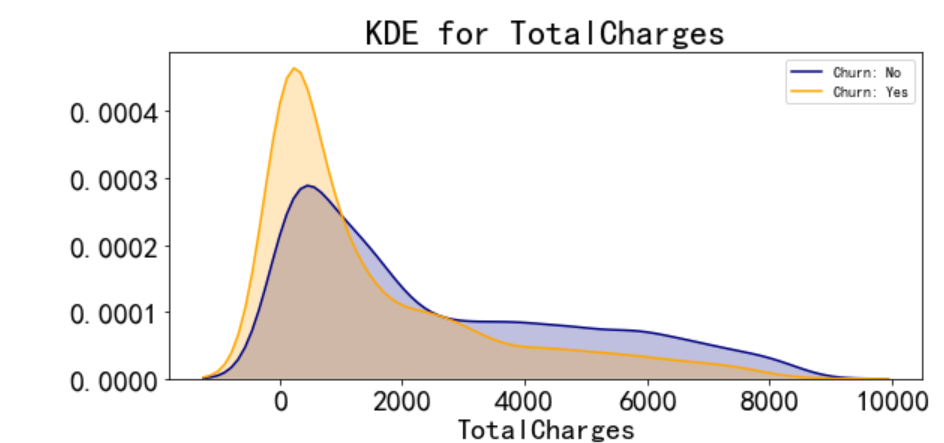
2.使用核密度估计展现churn特征的分布情况





MonthlyCharges核密度估计图

如图可以看出月消费在70至110之间的客户更容易流失



TotalCharges核密度估计图

如图可以看出总消费在1300以下的客户更容易流失

## 5.8分析总结

**一、数据分析结果**

对于分类属性：

Partner：单身用户更容易流失

OnlineSecurity：没开通的客户容易流失。

OnlineBackup：没开通在线备份服务的客户容易流失。

DeviceProtection：没开通设备保护业务的用户比较容易流失

TechSupport：没开通技术支持服务的用户容易流失。

StreamingTV：没有开通网络电视服务的用户更容易流失，但较不显著StreamingMovies：没有开通网络电视服务的用户更容易流失，但较不显著Contract：月份签订合同的用户最容易流失。签约越久越不容易流失PaperlessBilling ：开通电子账单的用户较容易流失

PaymentMethod：使用电子支票支付的人更容易流失

SeniorCitizen 分析：年轻用户在流失、留存人数占比都高

Dependents：经济不独立的用户更容易流失

Internerservice：办理了Fiber opti的客户容易流失

对于数值属性

1.任期1-5之间客户的流失量较大，也就是说刚开通电信业务1至5个月的用户容易流失，并且随着tenure任期的增加流失的概率越来越小

2.月消费在70至110之间的客户更容易流失

3.总消费在1300以下的客户更容易流失

**二、业务建议：**

1. **针对不同用户属性优化服务套餐与推广策略**

**在线安全（OnlineSecurity）、在线备份（OnlineBackup）、设备保护（DeviceProtection）、技术支持（TechSupport）服务方面**：加强对这些增值服务功能和优势的宣传推广，通过短信、APP 推送、客服介绍等多种方式向未开通用户详细说明服务能为其带来的保障，如数据安全、设备维修便利等，鼓励用户开通，可提供限时免费体验等优惠活动吸引用户尝试。

**合同签订（Contract）**：针对签订短期合同（尤其是按月签订合同）的用户，在合同即将到期前，主动联系用户，提供更有吸引力的长期合同优惠套餐，比如给予一定的费用折扣、额外的流量或通话时长赠送等，引导用户签订较长期限的合同，降低因合同短期化导致的高流失风险。

**老年用户与经济不独立用户（SeniorCitizen、Dependents）**：

针对老年用户，简化套餐内容和业务办理流程，提供专门的老年服务热线，客服人员用更耐心、通俗易懂的方式解答疑问和处理问题；推出适合老年人的优惠套餐，比如包含亲情通话时长、养生类资讯服务等，提升老年用户的使用体验和留存率。对于经济不独立的用户，设计价格亲民、性价比高的基础套餐，同时可与相关机构合作，推出学生套餐（针对学生群体等经济不独立用户），包含学习类应用流量优惠、学习资源免费获取等权益，满足其核心需求，减少流失

1. **基于用户任期和消费情况的服务优化**

**任期方面（tenure）**：对于新开通电信业务 1 - 5 个月的用户，在这个关键阶段加强客户关怀，定期发送使用指南、业务介绍、优惠活动等信息，帮助用户更好地了解和使用电信服务；设立新用户专属客服团队，快速响应和解决新用户遇到的问题，提高新用户的满意度，降低早期流失风险。随着任期增加，持续为老用户提供差异化的福利，如根据不同任期阶段给予不同等级的积分奖励、优先参与新业务体验等，激励老用户长期留存。

**消费情况方面**：

月消费（70 - 110 之间）与总消费（1300 以下）的客户：分析这类用户的消费结构，识别其主要使用的业务和未充分利用的业务，针对性地调整套餐内容，如对于月消费接近上限但流量不够用的用户，推荐升级流量包并给予一定的价格优惠；对于总消费较低的用户，推出小额消费套餐升级计划，以少量费用增加更多实用服务，提高用户消费体验，避免因性价比不高而流失。

**三、营销策略**

**精准营销活动**

基于用户画像的个性化推荐：综合用户的各类属性（如单身与否、是否开通增值服务、合同期限、年龄、消费情况等）以及行为数据，构建详细的用户画像，利用大数据算法实现精准营销。例如，向单身且月消费在 70 - 110 的年轻用户推荐包含社交流量优惠、在线娱乐服务（如 StreamingTV 等）开通优惠的套餐组合；向有经济不独立用户的家庭推荐家庭共享套餐，整合通话、流量等资源，并给予设备保护等增值服务的折扣优惠。

营销时机把握：根据用户任期情况，在用户开通业务后的第 1 个月、第 3 个月、第 5 个月等关键流失风险节点，推送有吸引力的优惠活动，如赠送话费、流量、限时免费开通增值服务等，帮助用户顺利度过高流失风险期；对于合同即将到期的用户，提前 1 - 2 个月进行针对性的续约营销，推出专属的续约优惠套餐，强调续约的好处，如费用减免、服务升级等。

**套餐组合与捆绑销售策略**

增值服务与基础套餐捆绑：将容易导致用户流失的未开通增值服务（如 OnlineSecurity、OnlineBackup 等）与基础套餐进行合理捆绑销售，设置不同档次的捆绑套餐，以整体更优惠的价格吸引用户选择，比如购买包含一定通话时长和流量的基础套餐，额外加少量费用即可开通多项增值服务，让用户感受到更高的性价比，提高用户对整体服务的依赖度。

# 六、创新性

1.在数据处理时没有使用预估值填充，而是按照业务逻辑进行分析填充

2.对比了采样与拆分数据集的操作顺序对预测结果的影响

3.使用管道来进行数据处理与训练

4．使用核密度估计展现了数值特征的样本分布情况

# 七、项目总结

深入了解了电信用户流失的关键因素，通过严谨的数据挖掘流程，清晰地识别出不同用户属性和消费行为与流失率之间的紧密联系，为精准营销和服务优化提供了坚实的数据支撑。

在数据处理和分析方法上积累了丰富经验，创新性地应用多种技术手段，如基于业务逻辑的填充方法、对比实验、管道技术以及核密度估计等，提升了项目团队的数据处理能力和分析水平，拓展了数据分析的思路和方法库。

模型的解释性方面还可以进一步加强。虽然使用了随机森林等算法进行特征选择和模型构建，但对于一些复杂的模型结果，向业务人员和非技术人员解释时仍存在一定困难，需要在后续项目中探索更直观、易懂的模型解释方法，以便更好地推动数据分析成果的落地应用。

**改进模型优化与评估**：

持续探索和尝试其他先进的机器学习和数据挖掘算法，结合本项目的业务特点和数据特征，寻找更优的模型组合和参数设置，进一步提高流失预测的准确性和稳定性。同时，加强对模型的实时监测和动态调整能力，随着业务的发展和数据的变化，及时更新和优化模型，以适应不断变化的市场环境和用户需求。

完善模型评估指标体系，除了常用的 ROC AUC 等指标外，引入更多能够反映业务实际效益和用户体验的评估指标，如用户留存率的实际提升效果、营销策略的投资回报率等，从多个维度全面评估模型的性能和价值，为项目决策提供更全面、准确的依据。