**网络空间安全学院（密码学院）**

2024-2025学年度第1学期

《机器学习》 课程设计报告

### **论文题目：电信客户流失预测：基于机器学习的多模型分析**

**学生学号及姓名：20223000347 秦涛**

**学生所在班级：大数据1班**

**论文提交时间：2024年12月 26 日**

**授课教师：赵瑶池**

**教师评阅：**

**成绩：**

**教师签名：**

**目录**

[1 选题目的 1](#_Toc22890)

[1.1 项目动机 1](#_Toc25962)

[1.2 项目目的 1](#_Toc18062)

[2 实验数据 1](#_Toc19174)

[2.1 数据来源 1](#_Toc24835)

[2.2 数据描述 2](#_Toc9711)

[2.3 数据预处理 3](#_Toc12116)

[3 特征工程 6](#_Toc220)

[3.1 特征编码 6](#_Toc8330)

[3.1.1 整数编码 6](#_Toc11358)

[3.1.2独热编码 7](#_Toc4409)

[3.2 特征选择 7](#_Toc11986)

[3.2.1 卡方检验 7](#_Toc508)

[3.2.2 交叉分析 8](#_Toc4946)

[3.3 特征处理 9](#_Toc21185)

[3.3.1标准化与离散化 9](#_Toc15947)

[3.3.2 样本平衡处理 10](#_Toc22587)

[4 所采用的机器学习方法的原理 11](#_Toc1216)

[4.1 方法概述 11](#_Toc29924)

[4.2 逻辑回归 11](#_Toc31424)

[4.3 随机森林 12](#_Toc319)

[4.4 支持向量机 12](#_Toc5309)

[4.5 决策树 13](#_Toc22467)

[4.6 朴素贝叶斯 13](#_Toc32474)

[4.7 XGBoost 14](#_Toc16140)

[5. 实验设置、实验结果及分析 15](#_Toc3139)

[5.1 模型训练过程 15](#_Toc12451)

[5.1.1初始模型性能 15](#_Toc16168)

[5.2 模型的调优与结果 16](#_Toc413)

[5.2.1 逻辑回归 16](#_Toc3424)

[5.2.2 随机森林 17](#_Toc12103)

[5.3.3 XGBoost 20](#_Toc11373)

[5.3 模型评估与对比 21](#_Toc29352)

[5.4 基于XSBoost客户流失预测分析 22](#_Toc29284)

[6 用户界面设计 24](#_Toc19918)

6.1 [模型选择： 24](#_Toc14606)

[6.2 模型训练与评估： 25](#_Toc12938)

[7. 课程设计心得体会 25](#_Toc9809)

[7.1 技术收获 25](#_Toc13429)

[7.2 模型优化的思考 27](#_Toc28430)

[8 参考文献 29](#_Toc7256)

[附录.................................................................................................................................................................30](#_Toc8309)

# 1 选题目的

## 项目动机

电信行业竞争白热化，客户流失如同失血，严重威胁公司的生存与发展。一方面，新客户获取成本持续攀升，使得客户获取难度加大；另一方面，市场份额争夺激烈，客户选择众多，稍有不慎，客户就可能转网流失。公司迫切需要一种精准有效的手段来提前洞悉客户流失倾向，以便及时干预。本项目基于客户消费习惯数据展开研究，是因为这些数据蕴含着客户行为模式和偏好的关键信息。通过机器学习技术深度挖掘这些信息，构建预测模型，公司能够提前识别高流失风险客户，进而制定个性化营销策略和服务优化方案，增强客户粘性，实现客户留存，在激烈竞争中站稳脚跟，确保业务持续稳健增长。

## 项目目的

对电信客户流失数据集进行全面的数据预处理，包括处理缺失值、异常值，对分类变量进行编码等操作，确保数据的质量和可用性。运用多种机器学习算法，如随机森林、支持向量机、逻辑回归、朴素贝叶斯、决策树和 XGBoost 等，构建客户流失预测模型结合SHAP分析影响客户流失的关键因素，实现简单用户界面展示模型效，。构建多种机器学习模型预测客户流失特征，通过可视化分析模型学习过程和预测结果。

# 2 实验数据

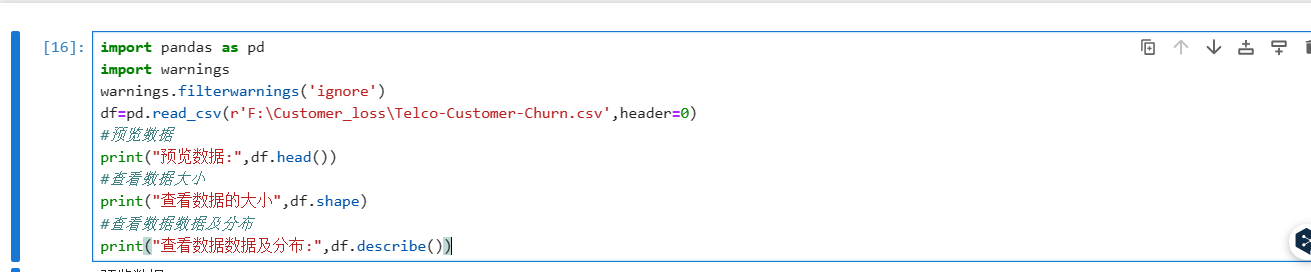
## 数据来源

数据来源：使用来自DataFountain（[https://www.datafountain.cn/datasets/35guide](https://www.datafountain.cn/datasets/35guide" \t "_blank)）的电信客户流失数据集，该数据集包含了丰富的客户信息，如客户 ID、性别、年龄、是否有伴侣、是否有家属、入网时长、电话服务、网络服务类型、在线安全服务、在线备份服务、设备保护服务、技术支持服务、流媒体电视服务、流媒体电影服务、合同类型、是否无纸化账单、支付方式、月费用、总费用以及客户是否流失等字段，能够全面反映客户的消费习惯和基本特征。

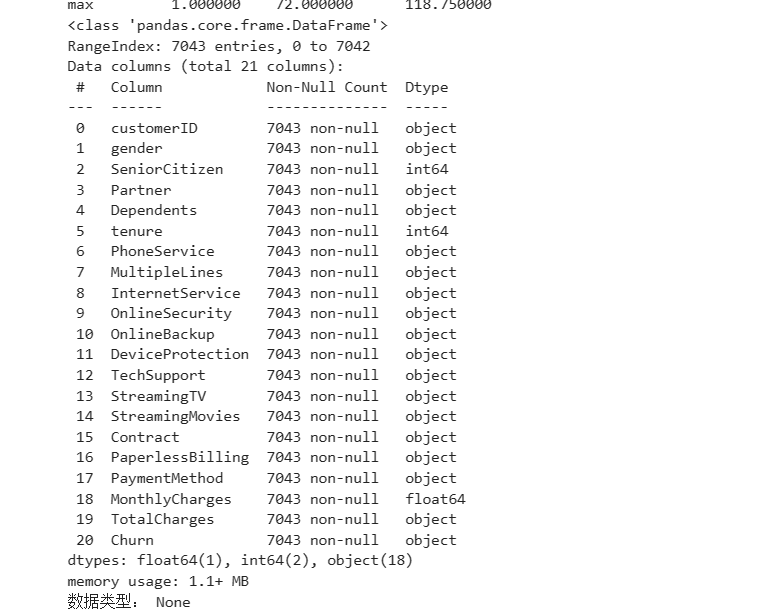
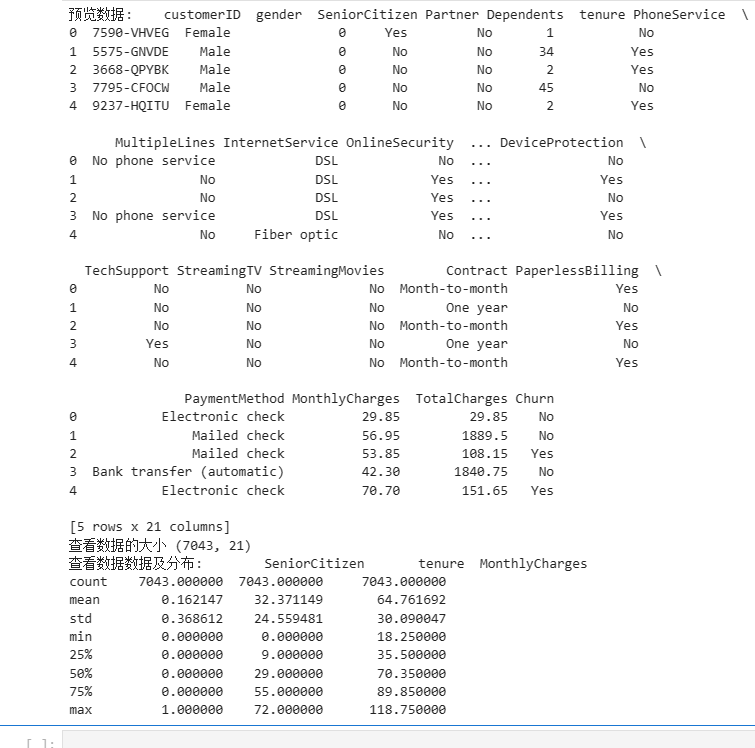
## 

## 数据描述

首先读取收集到的电信客户流失数据集并做一个初步的统计以及描述

****

**图1**

****

**图2 结果展示**

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名称** | **含义** | **类型** | **具体内容** |
| gender | 客户的性别 | object | 'Male' 或 'Female' |
| SeniorCitizen | 是否为老年客户 | int64 | 0 或 1 |
| Partner | 是否有伴侣 | object | 'Yes' 或 'No' |
| Dependents | 是否有家属 | object | 'Yes' 或 'No' |
| tenure | 客户使用公司服务的时长 | int64 | 整数 |

**表1 用户行为**

**表2 用户行为**

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名称** | **含义** | **类型** | **具体内容** |
| PhoneService | 是否有电话服务 | object | 'Yes' 或 'No' |
| MultipleLines | 是否有多线服务 | object | 'Yes' 或 'No' |
| InternetService | 客户选择的互联网服务类型 | object | 'DSL'、'Fiber optic'或'No' |
| OnlineSecurity | 是否有在线安全服务 | object | 'Yes' 或 'No' |
| OnlineBackup | 是否有在线备份服务 | object | 'Yes' 或 'No' |
| DeviceProtection | 是否有设备保护服务 | object | 'Yes' 或 'No' |
| TechSupport | 是否有技术支持服务 | object | 'Yes' 或 'No' |
| StreamingTV | 是否有在线电视服务 | object | 'Yes' 或 'No' |
| StreamingMovies | 是否有在线电影服务 | object | 'Yes' 或 'No' |
| Contract | 客户签订的合同类型 | object | 'Month-to-month'、'One year'或'Two year' |
| PaperlessBilling | 是否采用电子账单 | object | 'Yes' 或 'No' |
| PaymentMethod | 客户的付款方式 | object | 'Electronic check'、'Mailed check'、'Bank transfer’或'Credit card' |
| MonthlyCharges | 客户每月支付的费用 | float64 | 以美元为单位 |
| TotalCharges | 客户历史总支付费用 | object | 以美元为单位 |
| Churn | 客户是否流失 | object | 'Yes' 或 'No' |

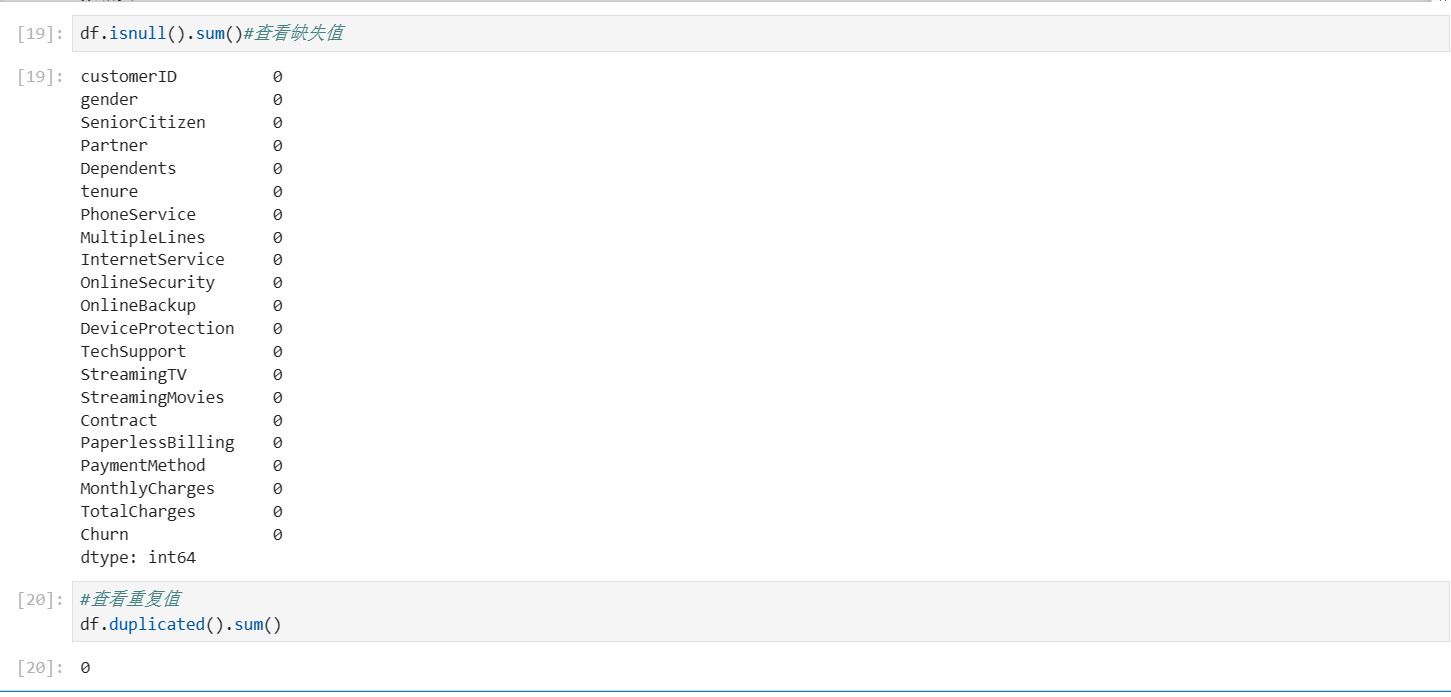
**表3 研究对象**

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名称** | **含义** | **类型** | **具体内容** |
| Churn | 客户是否流失 | object | 'Yes' 或 'No' |

## 2.3 数据预处理

**1）缺失值,重复值处理**

通过df.isnull().sum()查看缺失情况，df.duplicated().sum()查看重复值

****

**图3 结果展示**

可以发现该数据集并没有数据缺失以及重复值

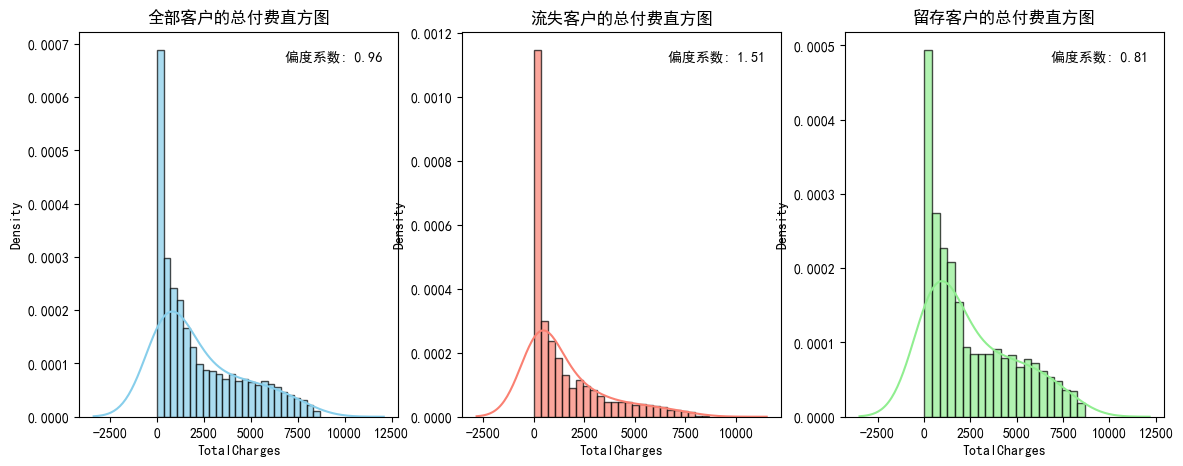
**2）数据转换并删除无用列(cutomerID)**

由图2以及表2可以看出，“TotalCharages”总费用是object需要转换成float64，将 TotalCharges 列转换为数值型。参数 errors='coerce'会将无法转换的非数值值替换为NaN，方便后续缺失值处理。



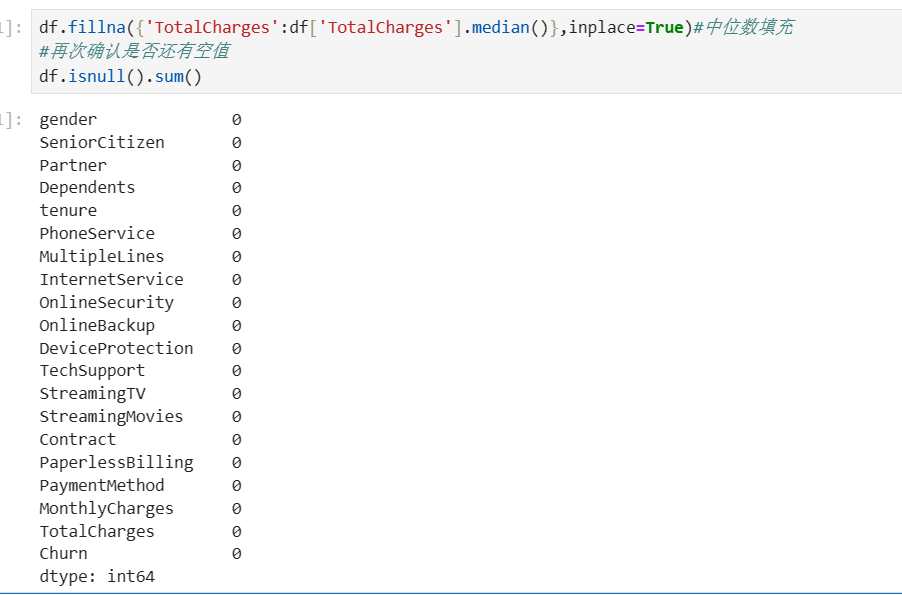
**图4 结果展示**

可以发现转换后出现了11个缺失值，为了选择合适的方法处理缺失值，需要了解数据的分布情况，且对于分类模型来说，不同的数据分布会影响模型的选择。



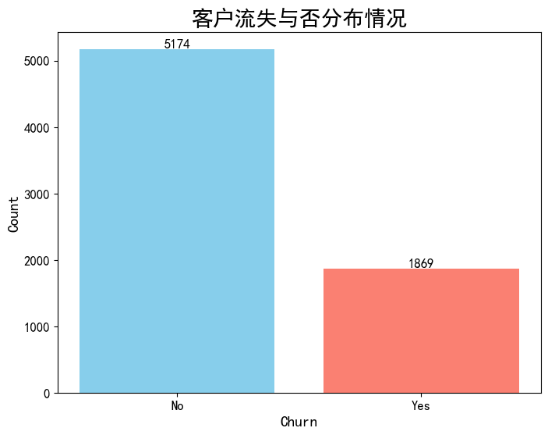
**图5 数据分布图**

由图可见，偏度系数均大于0，无论是全部客户、流失客户还是留存客户，其总付费数据都呈现右偏态。这表明在客户群体中，低付费客户占比较大，故用中位数填充。



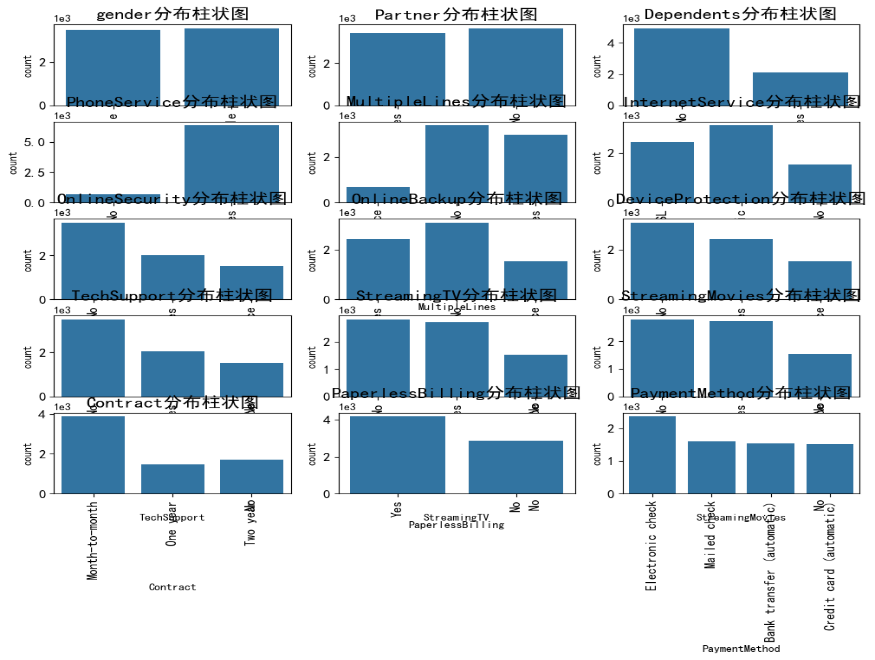
**图6 缺失值填充**

3）数据初步可视化

****

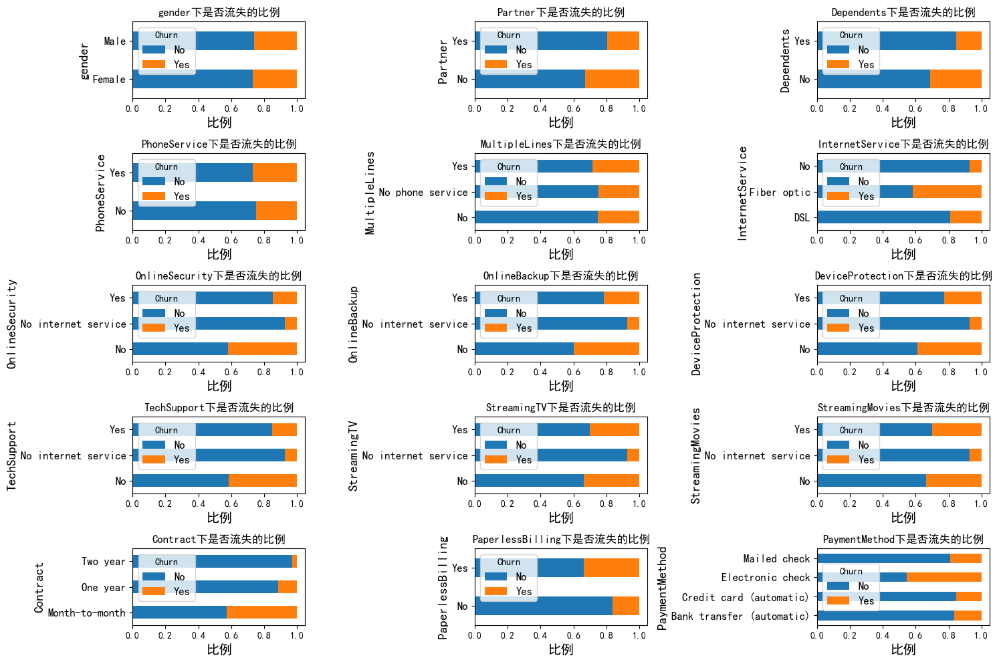
**图7 客户流失柱状图**

图7可以直观地展示出客户流失与未流失的数量差异，发现客户流失数量约占整体25%，作为样本存在不均衡，对于后续的分类模型有影响，需要平衡样本数量

****

**图8** 分类变量的分布情况

图8展示了多个分类变量的分布情况，充分展示了不同分类特征对于客户的吸引程度。在（但是不准确，由于样本不均衡），不能初步判断。

****

**图9 不同特征客户流失与否情况**

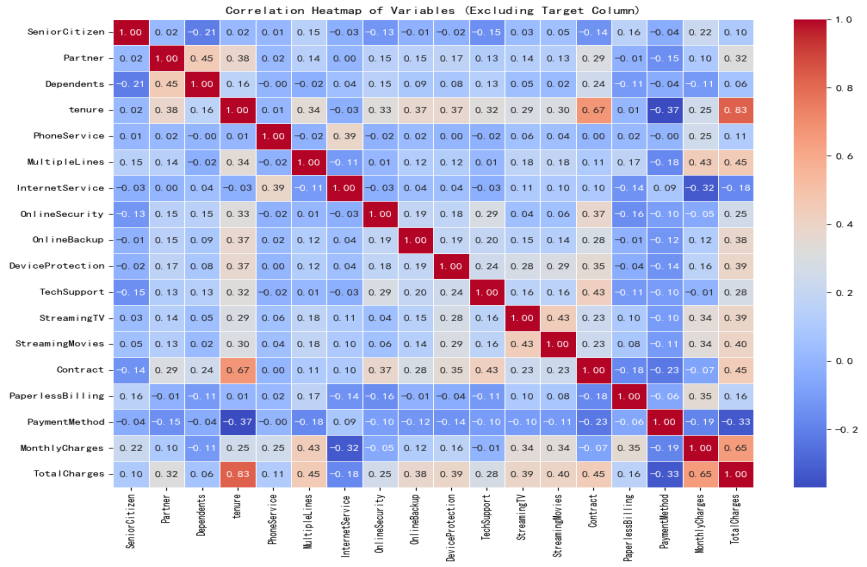
图9展示了多个分类特征下客户流失（Churn = Yes）与未流失（Churn = No）的比例情况可以初步看出一些特征与客户流失有较强的相关性

# 3 特征工程[1]

## 3.1 特征编码

### 3.1.1 整数编码

为了查看不同变量之间的相关性，我们进行重新整数编码，并用热力图展示结果

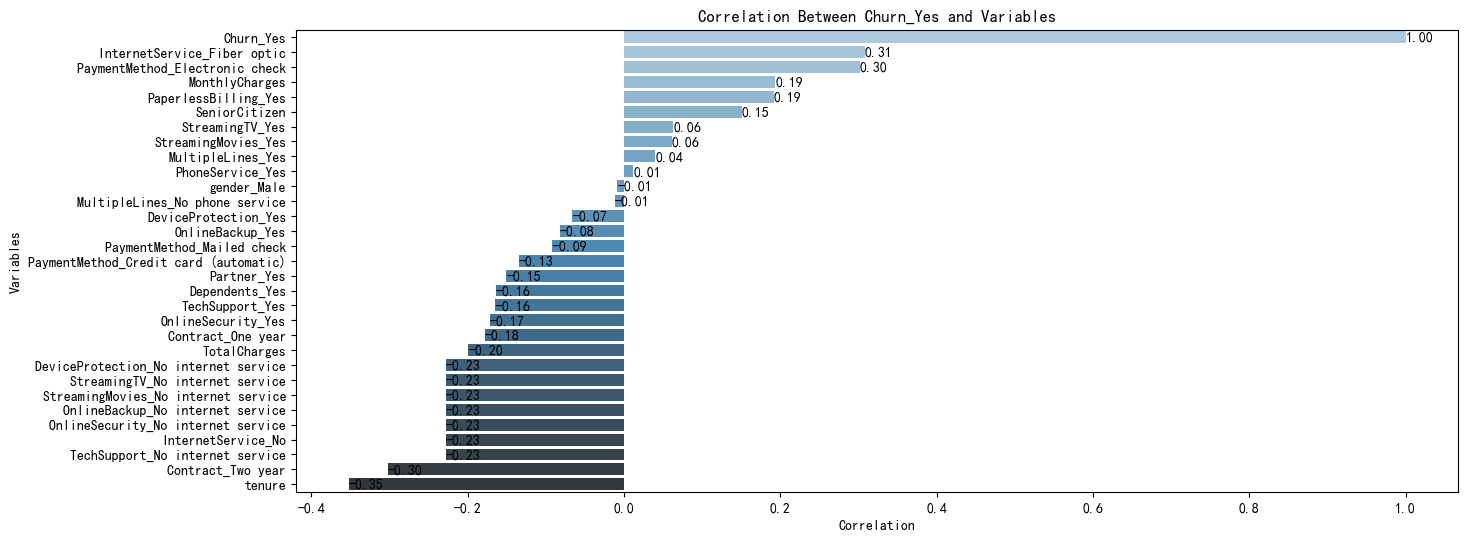


**图10 变量相关性**

可以发现MonthlyCharges和TotalCharges(相关性:0.65)Contract和tenure(相关性:0.67)OnlineSecurity和TechSupport(相关性:0.66)StreamingTV和StreamingMovies (相关性:0.43)tenure和TotalCharges(相关性:0.83)呈现高相关。

### 3.1.2 独热编码

为了能让模型更准确地捕捉到每个类别对结果的影响我们采用One-Hot Encoding编码，计算流失率查看"Churn"与其他变量下的标签相关性。



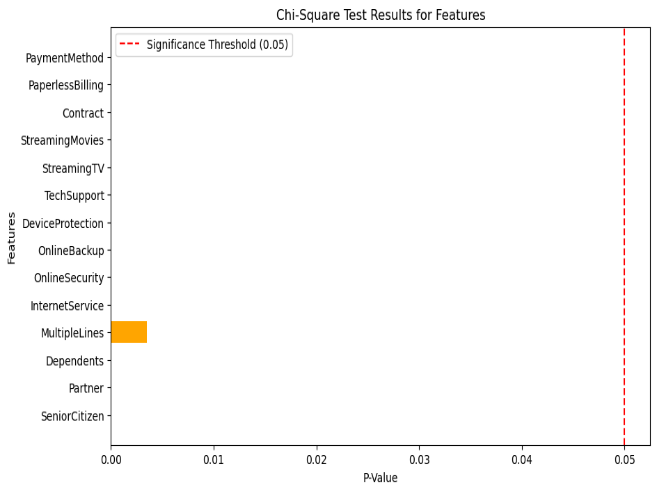
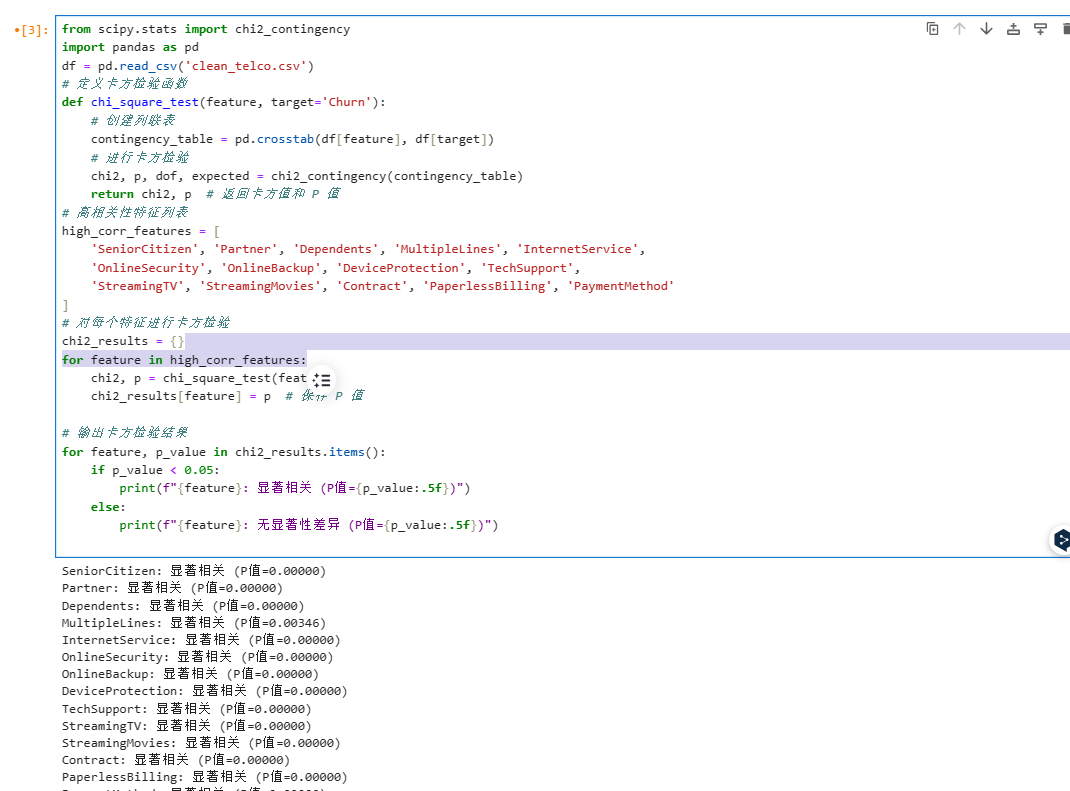
**图11 Churn与其他变量下的标签相关性**

从图看gender（性别）、PhoneService（电话服务）流失率几乎为0，故两个维度可以考虑[pendents,MultipleLines,InternetService,,OnlineSecurity,OnlineBackup,DeviceProtection,TechSupport,StreamingTV,StreamingMovies,Contract,PaperlessBilling,PaymentMethod]都有较高流失率。

## 3.2 特征选择[2]

### 3.2.1 卡方检验

由图7到图9的分析可以发现，由于样本的不均衡，不可以直接从柱形图去判断对哪个特征对流失客户的影响不能，因此需要进行交叉分析挖掘出潜在特征对于客户流失的影响。在此之前得先进行显著性检（卡方检验）确保交叉分析有意义，同时剔除掉无显著差异的特征。

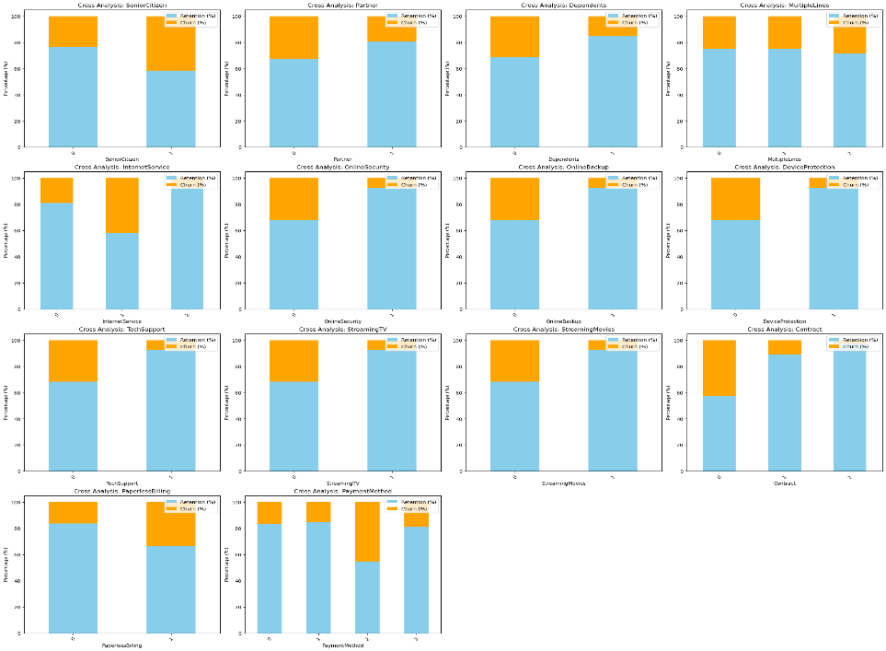


**图12 卡方检验图**

可以发现特征都呈显著相关。根据卡方检验的结果，筛选出显著相关的特征，接下来进行交叉分析。

### 3.2.2 交叉分析

通过交叉分析，计算留存率和流失率我们可以清楚的得出哪些特征对于顾客流失的影响大，具有哪些特征的用户容易流失。



**图13 交叉结果展示**

**表4 部分结果展示**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | SeniorCitizen | Retention (%) | Churn (%) |
| SeniorCitizen | 0 | 76.39 | 23.61 |
| SeniorCitizen | 1 | 58.32 | 41.68 |
| Partner | 0 | 67.04 | 32.96 |
| Partner | 1 | 80.34 | 19.66 |

（完整见combined\_cross\_analysis\_results.xlsx文件）

可以发现具有如下特征的顾客容易流失：

**表5 容易流失顾客特征**

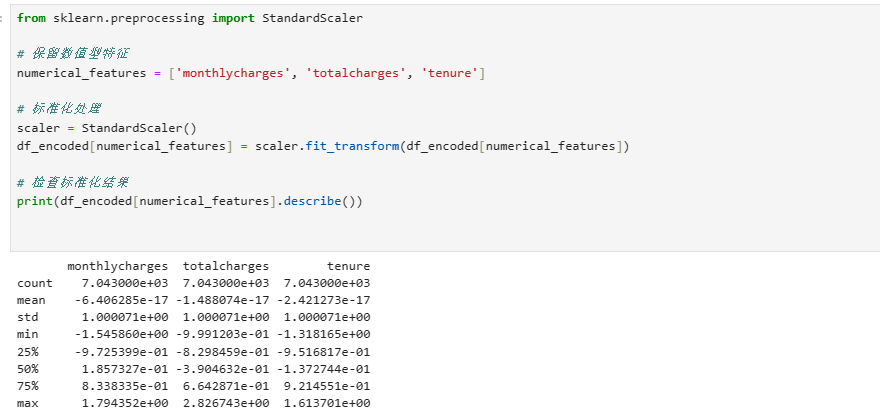
|  |  |  |
| --- | --- | --- |
| 特征 | 特征值 | 流失率 (%) |
| SeniorCitizen | 是（1） | 41.68 |
| Partner | 无伴侣（No） | 32.96 |
| Dependents | 无经济依赖（No） | 31.28 |
| InternetService | 开通Fiber optic（1） | 41.89 |
| OnlineSecurity | 未开通（No） | 41.77 |
| OnlineBackup | 未开通（No） | 39.93 |
| DeviceProtection | 未开通（No） | 39.13 |
| TechSupport | 未开通（No） | 41.63 |
| Contract | 按月签订（Month-to-month） | 42.71 |
| PaperlessBilling | 开通（Yes） | 33.57 |
| PaymentMethod | 电子支票（Electronic check） | 45.29 |

3.3 特征处理

3.3.1标准化与离散化

**1）特征的标准化**

在将分类变量转化为数值特征（独热编码）后，对数值型数据（'monthlycharges', 'totalcharges', 'tenure'）进行标准化用于后续。标准化适用于连续型特征，使其均值为 0，标准差为 1，便于模型（逻辑回归和支持向量机）训练。



**图14 数值标准化**

**2）特征的离散化**

通过四分位数将连续特征转化为分类特征，便于树模型和交叉分析随机森林、XGBoost

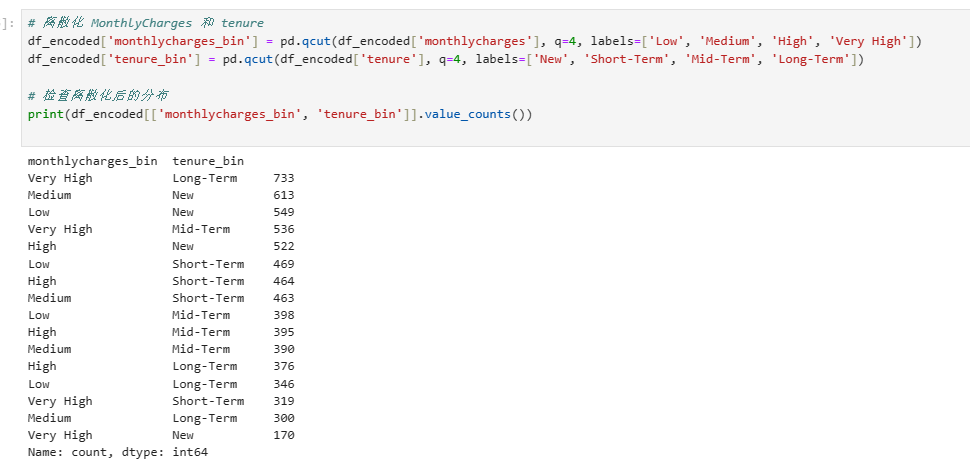
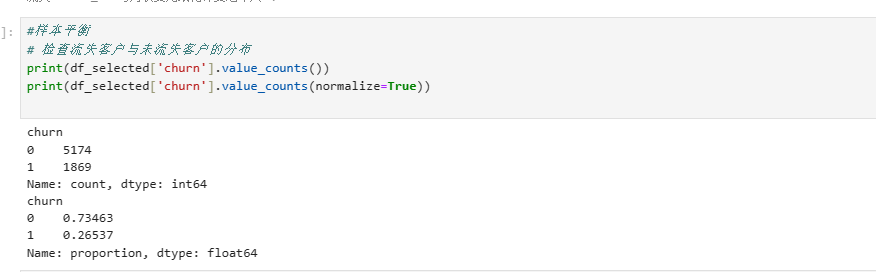


图15 特征离散化

3.3.2 样本平衡处理

Step1）检查样本的分布



**图16 查看数据分布**

**Step2）过采样（SMOTE）**

由于多数类别样本较多的情况，我们采用通过生成新的少数类别样本来平衡数据分布。**过采样的核心思想：**是通过重复或生成新样本来增加少数类别的样本数量，使得少数类别与多数类别的样本数量相近。

**SMOTE是**通过少数类别样本之间的插值生成新样本最终平衡数据集，数据的多样性也得到了一定程度的提升。SMOTE 的公式核心为线性插值：



****

图17 过采样

Step3）保存为balanced\_tarin\_data.ccsv文件

将采样后的数据保存，便于后续的模型训练。

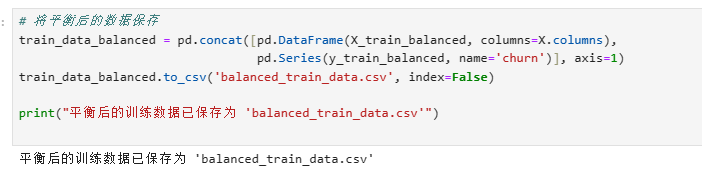


图18 保存数据

# 4 所采用的机器学习方法的原理

4.1 方法概述

机器学习方法的核心思想是根据训练数据构建一个模型,使得在给定输入时，模型预测的输出尽可能接近真实值。

每种方法的差异主要体现在假设空间、目标函数和优化方法上：

假设空间：模型拟合数据的函数形式及其复杂度。

目标函数：定义优化目标的损失函数或代价函数。

优化方法：通过数学或算法手段求解最优模型参数。

4.2 逻辑回归

**函数的假设空间**：逻辑回归假设目标变量,其预测值由输入特征通过 sigmoid 函数映射到区间(0,1)。模型形式为：



其中，W是权重向量，是偏置。

**目标函数：**采用对数似然函数作为目标函数，通过最大化似然估计得到最优参数：



其中，是模型的预测值。

**优化方法：**通过梯度下降算法更新参数：



其中：是学习率；是目标函数对参数的偏导数。

## 4.3 随机森林

**假设空间：**随机森林是决策树的集合，假设空间为一组不同的决策树模型。其输出是所有树的平均值(回归任务)或多数投票(分类任务)



**目标函数：**随机森林的目标函数是减少基模型的加权平均误差。基于树的分裂准则(Gini不纯度或信息增益)优化单棵树。以分类任务为例，Gini不纯度定义为：



其中，是类别的比例。

**优化方法：**随机森林通过引入数据子采样和特征随机选择，减少单棵树的过拟合。训练时采用以下步骤：① 随机从训练数据中抽取样本（Bootstrap）。②随机选择部分特征进行分裂。

4.4 支持向量机

**假设空间：**支持向量机的目标是通过引入核函数将数据从原始空间映射到高维特征空间，并找到一个能够最大化分类间隔的超平面。假设空间由所有线性可分超平面构成： 

其中:为核函数,是拉格朗日乘子,是类别标签

**目标函数：**在软间隔支持向量机中，目标是优化以下凸二次规划问题：



subject to:, 

其中：是正则化项，确保间隔最大化；是松弛变量，允许一定程度的分类错误，

**优化方法：**通过对偶问题引入拉格朗日对偶优化，将目标函数转化为：



subject to: 利用序列最小优化算法(SMO)高效求解该问题。

4.5 决策树

**假设空间：**决策树的假设空间是由所有可能的分裂规则构成的树集合。每个叶节点对应一个类别或数值预测值。对于输入,决策树可以表示为：



是叶节点的输出值；是指示函数。

**目标函数：**决策树的分裂准则通过最大化信息增益或最小化 Gini 不纯度优化。以信息增益为例：



是熵；是类别在数据集中的比例。

**优化方法**：使用贪心算法逐层分裂节点，选择分裂使得信息增益最大，直到达到预设条件（如树深度限制或最小节点样本数）。

4.6 朴素贝叶斯

**假设空间：**朴素贝叶斯假设特征在给定类别的条件下相互独立，联合概率分布为：

其中：是类别的先验概率；是给定类别条件下特征的条件概率。

**目标函数：**分类通过最大化后验概率实现：

****

**优化方法：**基于极大似然估计(MLE)计算和



4.7 XGBoost

**假设空间：**假设是多个基学习器的累加

****

其中：是 CART 树的集合;是第

**目标函数：**由损失函数和正则化项构成：

****

其中：,是叶节点权重平方和；为正则化参数

**优化方法：**通过二阶泰勒展开沂似目标函数，更新规则为：



利用贪心算法构建树分裂。

# 5 实验设置、实验结果及分析

# 5.1 模型训练过程

# 5.1.1初始模型性能

经过特征工程后，接下来进行对6个模型的初步的训练，并通过表格展示数据以及可视化展示AUC曲线初步对比模型的性能。

表6 初步模型性能

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | F1 Score | AUC |
| Logistic Regression | 0.786569 | 0.802553 | 0.861036 |
| Random Forest | 0.821527 | 0.827402 | 0.894017 |
| Decision Tree | 0.764489 | 0.766636 | 0.764264 |
| Naive Bayes | 0.736431 | 0.775558 | 0.827539 |
| Support Vector Machine | 0.630635 | 0.604239 | 0.714376 |
| XGBoost | 0.822447 | 0.832310 | 0.8948 |

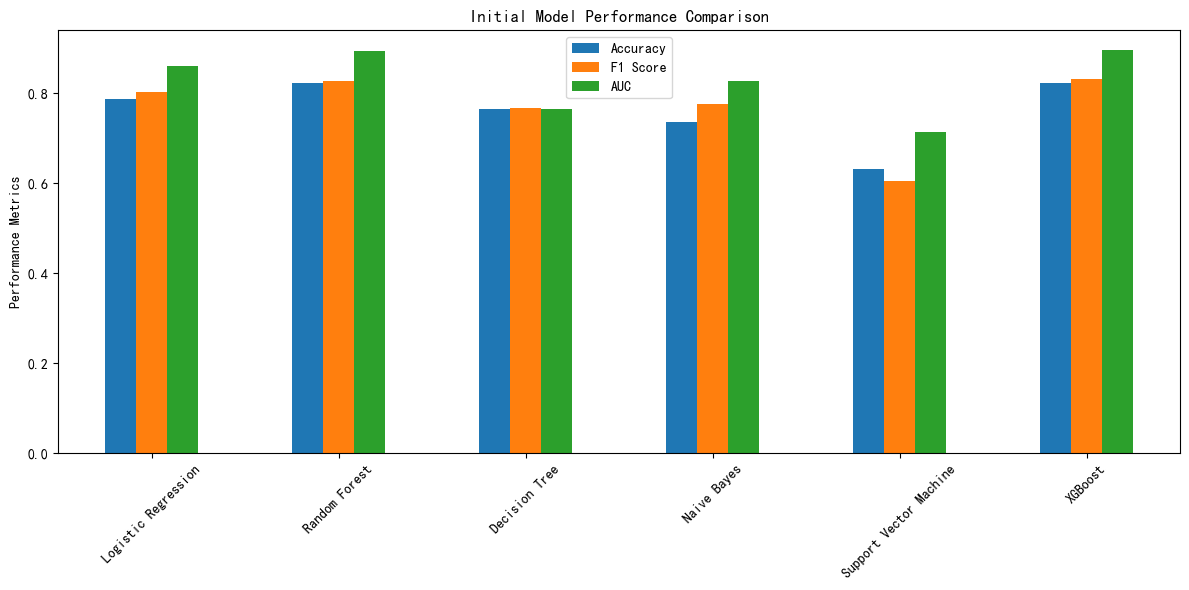
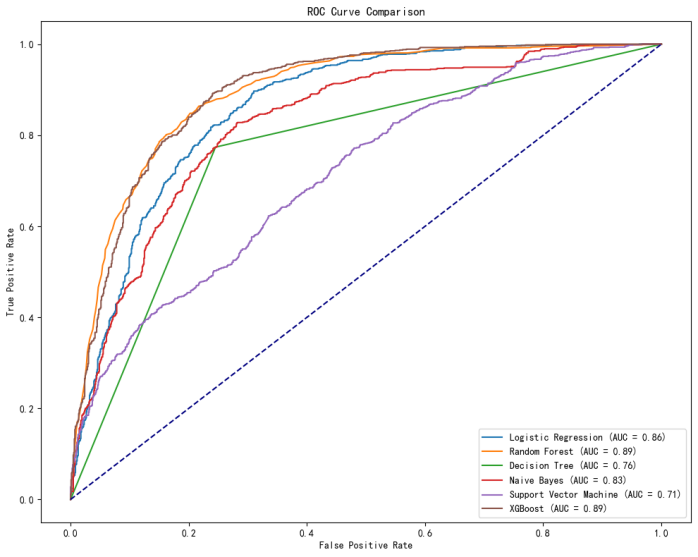


图19 模型初步性能对比图



**图20 AUC曲线**

通过初步的模型训练可以看出随机森林和XGBoost 在各项性能指标和ROC曲线下面积上都表现突出，是较为优秀的模型。支持向量机在所有评估指标上表现都不佳，决策树在 AUC 方面表现一般。

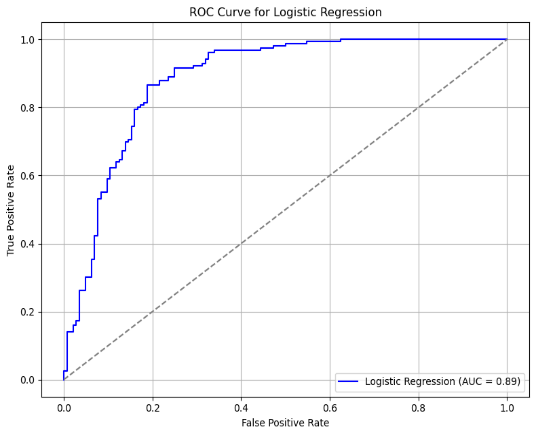
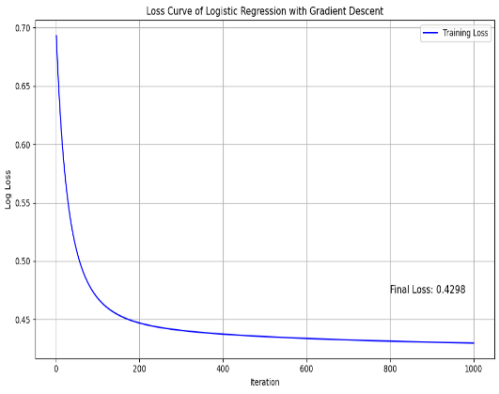
对模型的综合性能要有较高要求，而随机森林和XGBoost，逻辑回归模型性能良好，但是支持向量机在这个数据集上表现不佳。所以为了预测顾客流失问题，我们通过对初始模型训练最好的三种模型进行参数调优，来提高准确度

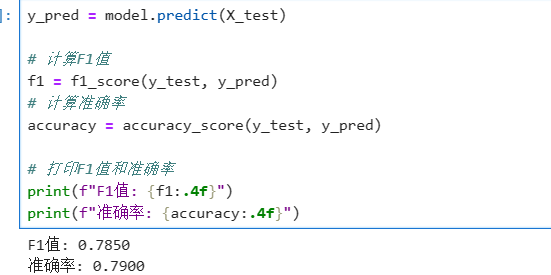
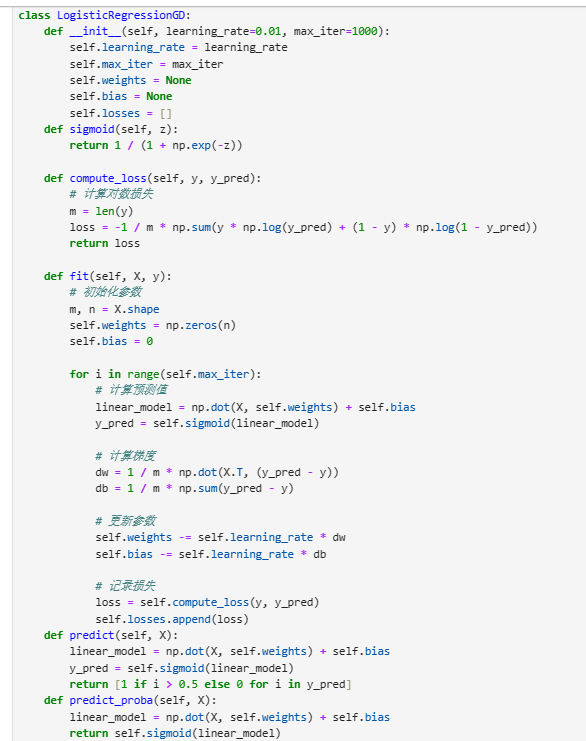
## 5.2 模型的调优与结果

# 5.2.1 逻辑回归[3]

1）模型调优结果

通过梯度下降优化后的模型性能。输出训练损失输出为最后一次迭代的log loss。可视化展示AUC值并基于ROC曲线计算的分类能力指标。





**图22 梯度下降算法跌打**

通过自定义梯度下降算法，对逻辑回归进行参数调优。梯度下降通过逐步减少损失函数的值，成功优化了逻辑回归模型，损失曲线收敛表明优化过程稳定。不断迭代得出最终损失值为0.4298，AUC调优到0.89，F1值:0.7850准确率:0.7900。

2）优缺点分析

优点：逻辑回归模型是一种线性模型，公式简单，优化方法成熟（如梯度下降）。对于数据预处理和特征选择要求较低，易于快速实现和解释。训练速度快，适合大规模数据。

缺点：无法处理非线性关系：逻辑回归只能学习线性决策边界，无法处理复杂的非线性关系。对于特征与目标变量之间有复杂非线性关系的问题，模型性能受限且特征之间存在多重共线性会影响模型的稳定性。

3）改进方向

虽然AUC值较高，但逻辑回归无法捕捉数据中可能存在的非线性关系，可能导致分类能力受限所以用更好的模型替代。使用具有更强表达能力的模型替代逻辑回归随机森林能捕捉非线性关系以及XGBoost处理复杂关系和大规模数据效果。

# 5.3.2 随机森林[4]

随机森林通过Bootstrap采样（有放回地从训练集中随机抽样）生成多个训练子集，训练多棵决策树。在构建每棵树时，随机搜索选择特征子集，用于分裂节点，降低了不同树之间的相关性，增强了模型的鲁棒性。

**1)调整以下超参数优化随机森林**

n\_estimators: 决策树的数量。更多的树通常会提高性能，但也增加了训练时间。

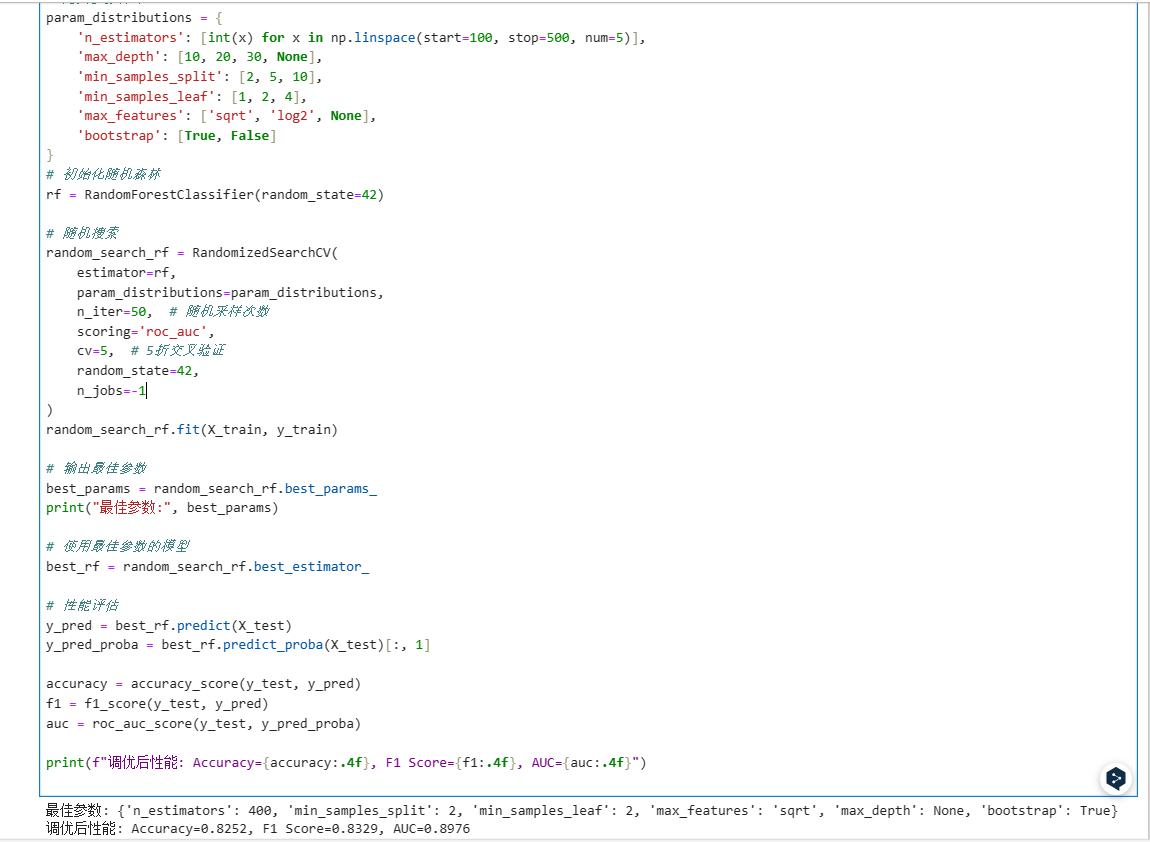
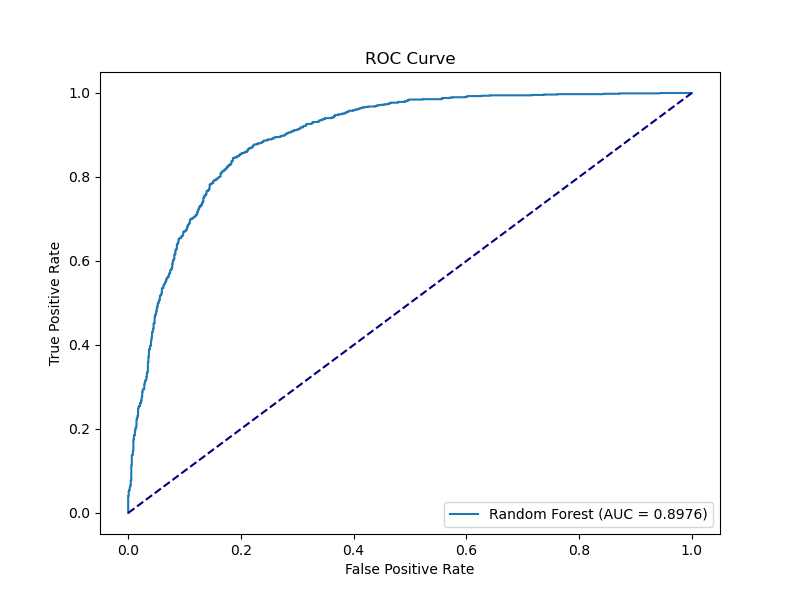
max\_depth: 单棵树的最大深度，限制深度可以减少过拟合。

min\_samples\_split: 内部节点再分裂所需的最小样本数。较大的值有助于控制模型复杂度。

min\_samples\_leaf: 叶子节点的最小样本数，防止生成过小的叶子节点。

max\_features: 分裂时使用的最大特征数，控制特征随机性。

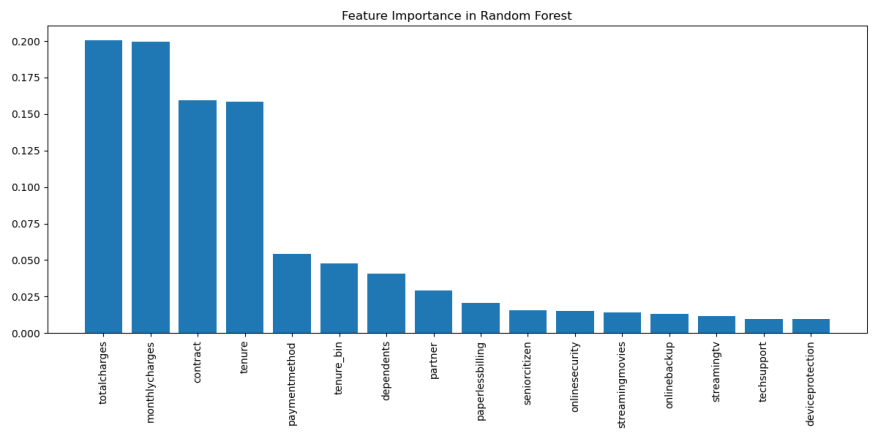
bootstrap: 是否启用样本的随机抽样。

****

**图23 随机搜索调优结果 图24 ROC曲线**

求得最优参数的各类值，最佳参数: {'n\_estimators':400,'min\_samples\_split':2,'min\_samples\_leaf':2,'max\_features':'sqrt', 'max\_depth': None, 'bootstrap':True}

调优后性能: Accuracy=0.8252,F1 Score=0.8329,AUC=0.8976



**图25 顾客流失特征图**

随机森林特征重要性分析显示，客户流失主要受费用高低、合同类型和服务时长的影响，高额总费用和月费用显著增加流失风险，而长期合同和较长服务时长则有助于降低流失概率。同时，支付方式（电子支票）与流失存在一定相关性，优化支付体验可能有效减少流失

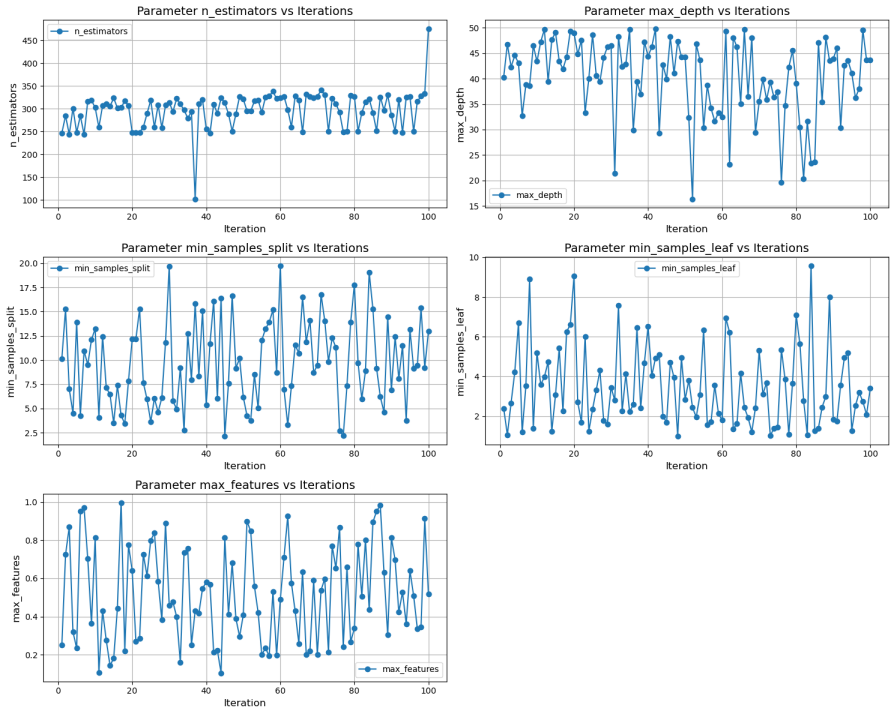
**2）改进方向**

由Boostrap采样后进行随机搜索调优，发现随机森林的性能都有提升但是不高，可能是搜索的参数范围不够大，如果调大参数范围以及深度，会让运行的时间长。

随机森林虽然是强大的模型，但在某些非线性复杂关系中，其表现可能不如梯度提升类模型所以在后续替换为梯度提升决策树（GBDT）-XGBoost。

**3）贝叶斯优化随机森林**

使用贝叶斯优化来调优随机森林超参数，可以在更少的迭代次数内找到更优的参数组合，降低过拟合的风险，提高每棵树分裂的特征多样性，增强泛化能力同时显著提升模型的AUC和其他性能，并展示参数的迭代过程。



**图26 参数迭代图**



**图27 各参数值**

可以看出最佳参数:{'n\_estimators':327, 'max\_depth':18, 'min\_samples\_split':2, 'min\_samples\_leaf':1,'max\_features':0.1892329315025182},Final Performance: Accuracy=0.8284, F1 Score=0.8328, AUC=0.9024

迭代100次个参数的过程，型超参数的变化趋势和 AUC 的优化过程但各参数似乎并没有收敛稳定，说明迭代次数不够，但对于最后结果可以看出各项性能都有提高。

5.3.3 XGBoost[5]

改进的优点：①XGBoost 通过梯度提升决策树（GBDT）的框架，优化了损失函数，比起随机森林来说对流失客户和非流失客户的区分能力更强。②为防止过拟合且支持子采样（对样本和特征进行随机采样），XGBoost能自动生成交互特征，尤其对复杂的消费习惯数据表现优异结合贝叶斯优化进一步提高模型的泛化能力在树模型中

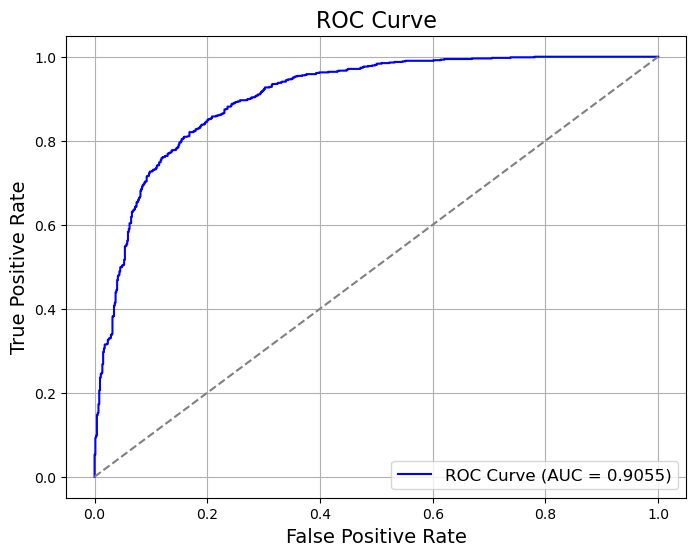
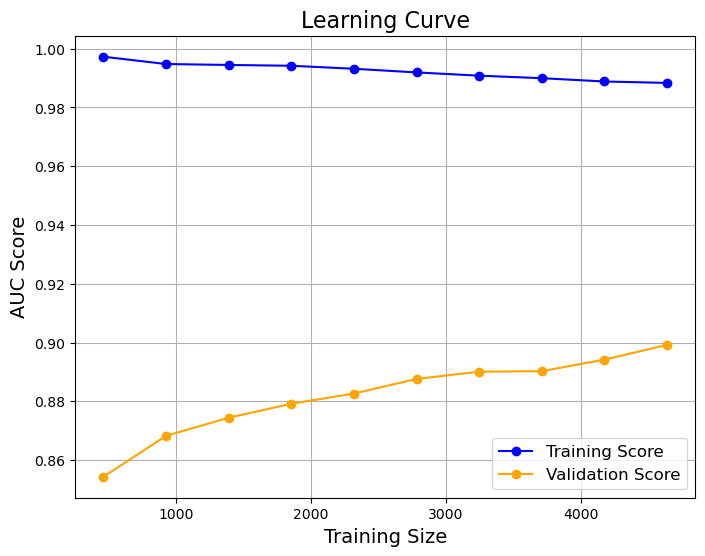
① 在优化的过程中，贝叶斯优化[5]动态调整了XGBoost模型的以下参数：

n\_estimators: 决策树的数量，优化模型的学习能力。

max\_depth: 决策树的最大深度，控制模型的复杂度和防止过拟合。

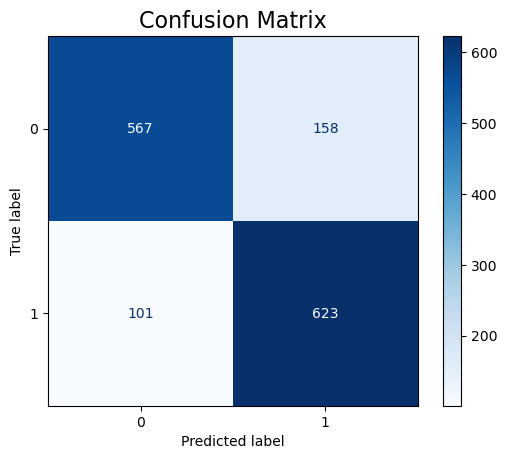
learning\_rate: 学习率，影响模型的学习步幅。

subsample: 子采样比例，减少过拟合的可能性。colsample\_bytree: 每棵树的特征采样比例，进一步控制特征空间。



**图28 模型学习曲线**  **图29 ROC曲线**

由模型学习曲线可以看出训练曲线和验证曲线收敛较好，且验证分数达到较高的水平，表明模型在训练和验证集上都有良好的表现。同时由ROC曲线可以得出AUC为0.9055相对于随机森林有进一步提高



**图30 混淆矩阵**

由混淆矩阵可以计算得出其准确率为0.849，F1 score为0.828。且得出最佳参数: {'colsample\_bytree': 0.6652619567813226, 'gamma': 0.4153798702870832, 'learning\_rate': 0.02825421700024073, 'max\_depth': 11.083496755677036, 'n\_estimators': 455.62416698403825, 'subsample': 0.8602760335347635

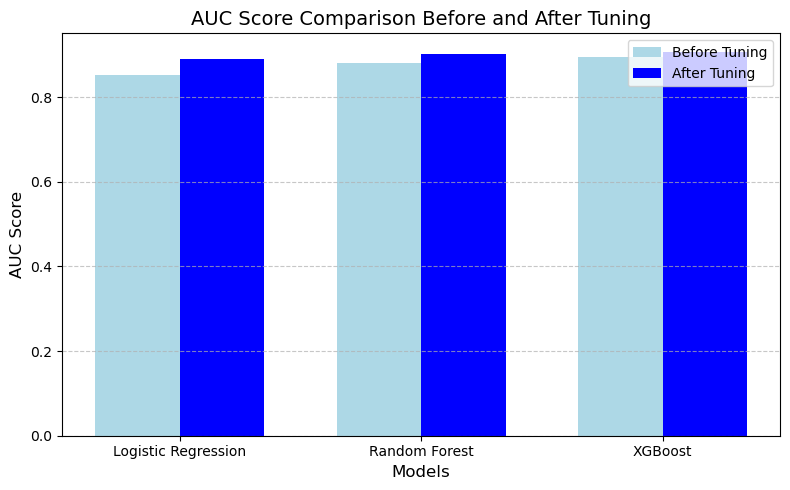
5.3 模型评估与对比

**1）模型评估**

本次实验中，我们初步计算训练了六种模型，选择初步平均性能最好的前三种对逻辑回归、随机森林和XGBoost进行了调优，并评估了其在测试集上的表现。通过分析模型的精度 (Accuracy)、F1值(F1 Score)、AUC值(ROC-AUC)，总结各模型性能如下：

表7 调优模型对比

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 调优后准确率 | 调优后F1 Score | 调优后AUC |
| 逻辑回归 | 0.7900 | 0.7850 | 0.8900 |
| 随机森林 | 0.8284 | 0.8328 | 0.9024 |
| XGBoost | 0.8490 | 0.8280 | 0.9300 |



**表31 模型AUC对比柱状图**

调优后，XGBoost模型凭借其复杂的特征处理能力和高效的参数优化方法，在三种模型中表现最佳，最终AUC达到**0.9055**，优于逻辑回归和随机森林。XGBoost的调优成功得益于贝叶斯优化对超参数的细致调整和强大的非线性建模能力。这表明 XGBoost 是客户流失预测问题中的优选方法。

**2）模型调优过程对比**

表8 调优过程与方法对比

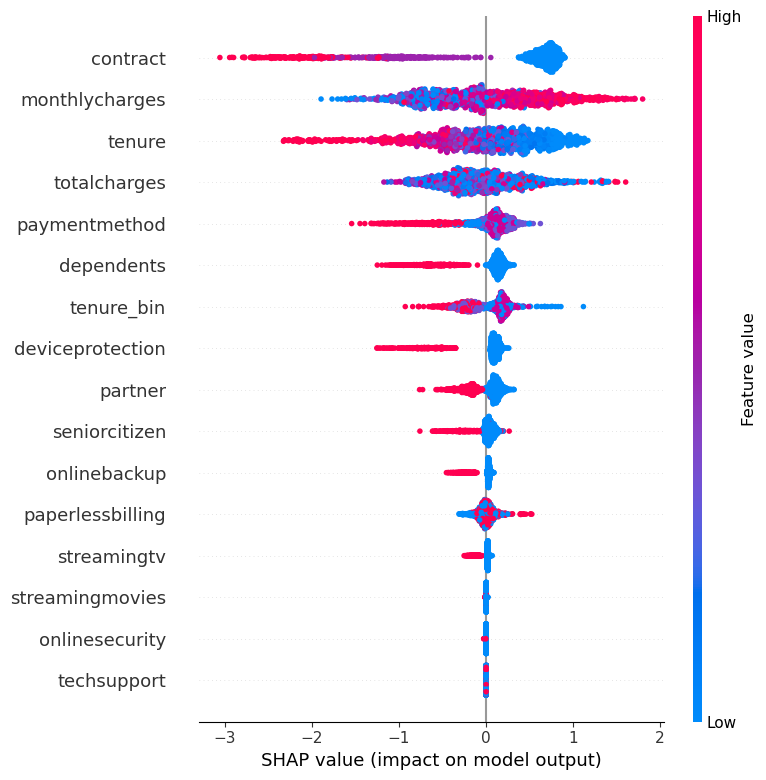
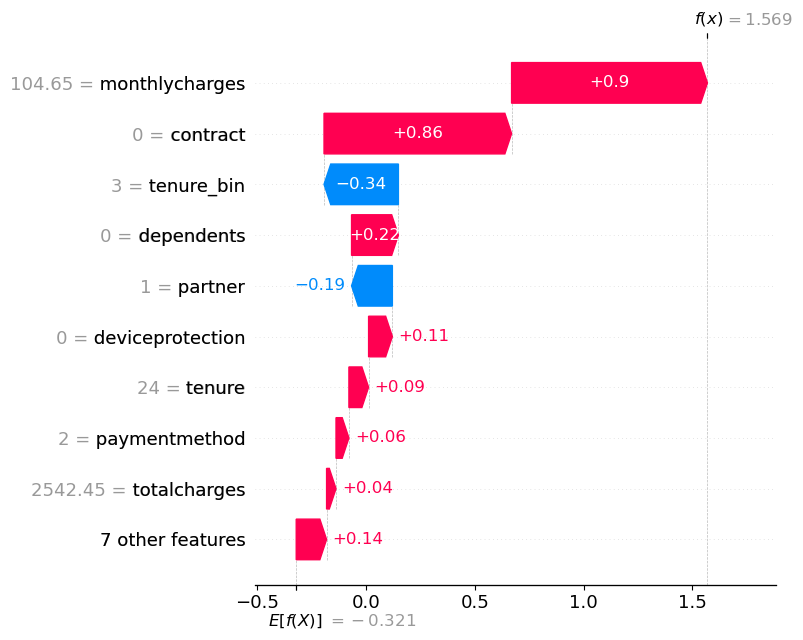
|  |  |  |
| --- | --- | --- |
| 模型 | 调优方法 | 调优过程说明 |
| 逻辑回归 | 网格搜索，特征工程 | 用梯度下降法进行调优;提高了模型对复杂非线性关系的拟合能力。 |
| 随机森林 | 贝叶斯优化 | 自动搜索最佳参数组合，如n\_estimators（树数量）、max\_depth（树深度）、min\_samples\_split（最小分裂样本数）等，显著减少过拟合，提高泛化能力。 |
| XGBoost | 贝叶斯优化早停策略 | 对学习率(learning\_rate)、最大深度(max\_depth)、子采样率 (subsample) 等进行了调优，并加入早停防止过拟合；最终参数组合显著提升了模型表现。 |

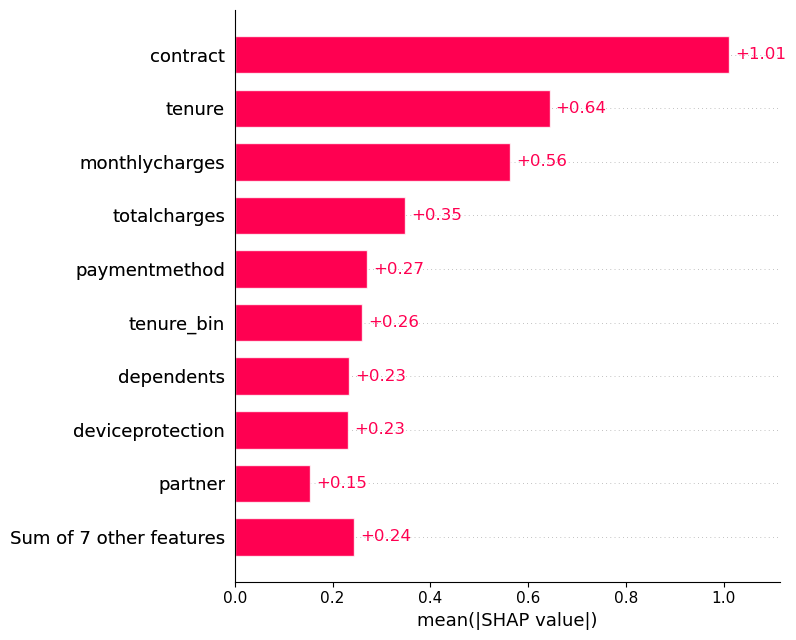
5.4 基于XSBoost客户流失预测分析

根据前面模型的评估以及对比，选择平均性能更高的XGBoost模型根据客户的消费习惯分析是否会流失。

**1）SHAP[5]值分析**

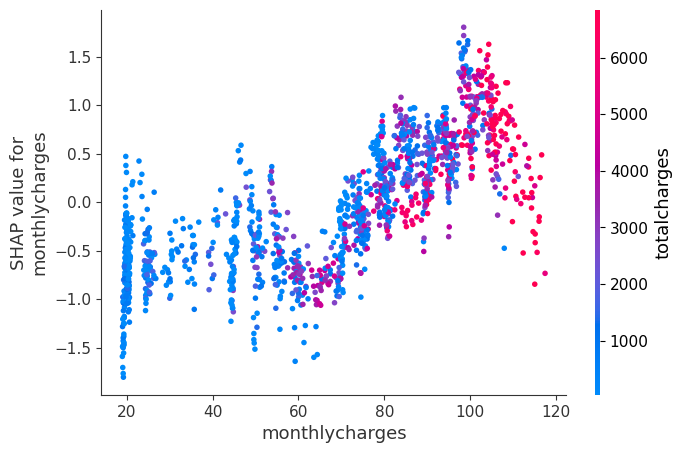
采用解释能力更强的SHAP图对重要的客户特征进行基础分析。总结哪些特征对客户流失预测影响最大。确定正向或负向影响对特定客户的预测结果进行解释，找出哪些特征是决定性因素。



**图32 SHAP分析图**

根据shap分析图，可以看出合同类型 (contract) 是全局范围内影响流失的首要因素。较长的服务时长 (tenure\_bin) 和有伴侣 (partner) 减少流失概率。短期合同和高消费是高度相关的流失风险人群。



**图34 月消费与总消费shap图**

有图可以分析到monthlycharge和totalcharge 每月消费和总消费都对流失有较大的正向影响，MonthlyCharges 在 60-100 且 TotalCharges 在 1000-3000流失最严重

基于XGBoost模型和SHAP分析得出流失客户有以下特征。

表9 顾客流失特征特点

|  |  |
| --- | --- |
| 特征名称 | 高流失风险情况 |
| Contract (合同类型) | 按月签订合同的客户流失风险最高。 |
| Tenure (客户使用时长) | 时长较短（<12个月）的客户流失风险显著增加。 |
| MonthlyCharges (每月账单金额) | 月账单金额在 60-100的中高消费客户流失风险较高。 |
| TotalCharges (累计账单金额) | 累计消费金额较低1000-3000的客户更容易流失。 |
| PaymentMethod (支付方式) | 使用电子支票支付的客户流失风险较高。 |
| Tenure\_bin (时长分段) | 时长分段为新客户和（短期客户的客户流失率最高。 |
| Dependents (是否有家属) | 无家属的客户（独居或单身）流失风险较高。 |
| DeviceProtection (设备保护服务) | 未开通设备保护服务的客户流失风险较高。 |
| Partner (是否有配偶) | 没有配偶的客户流失风险较高。 |
| 增值服务（StreamingTV、OnlineBackup） | 未订阅这些增值服务的客户流失风险较高。 |

# 5 用户界面设计

用户可以通过界面选择数据集文件进行加载，支持CSV格式。加载完成后，在界面上显示数据集的前几行（头部），方便用户快速预览数据的基本结构。

自动识别数据集中的特征和目标变量。



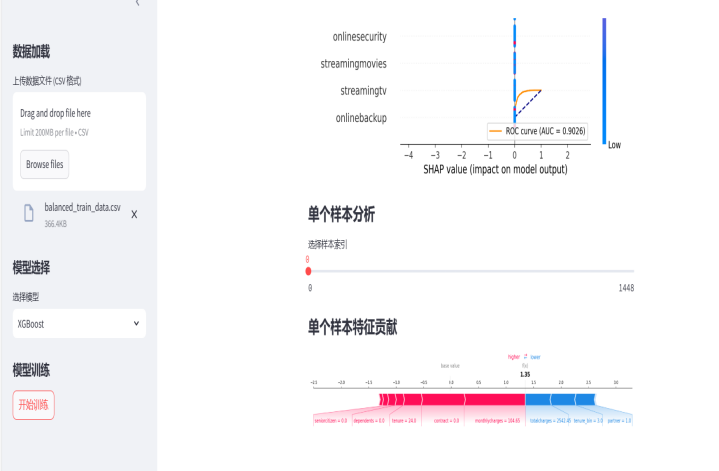
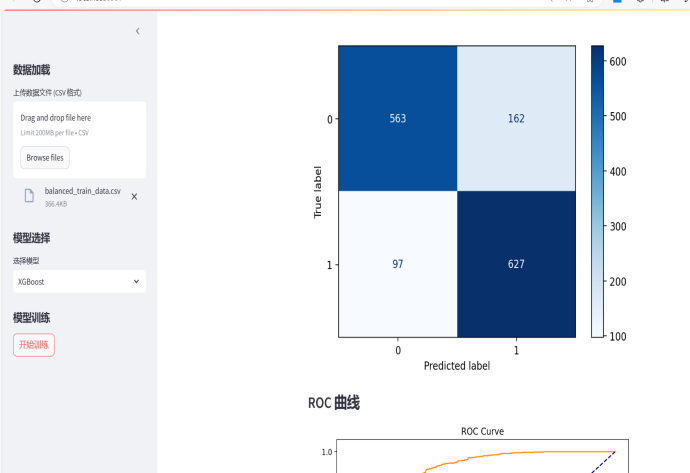
**1）模型选择**

包含逻辑回归、随机森林、支持向量机、决策树、朴素贝叶斯、XGBoost 六种模型。



**2）模型训练与评估**

自动评估模型的 准确率和AUC，显示混淆矩阵和ROC曲线。对支持特征重要性的模型显示特征贡献。SHAP 可视化：对XGBoost 模型支持全局特征重要性和单个样本的特征贡献分析



# 6. 课程设计心得体会

## 6.1 技术收获

1. **算法深度剖析**

**逻辑回归**：深入学习逻辑回归的过程中，理解其通过线性假设与sigmoid函数将特征映射到概率区间进行分类的原理，是对基础理论的一次有力巩固。在项目实践中，手动实现梯度下降算法来优化模型参数，真切体会到了模型从初始状态逐步收敛至较优解的过程。这一实践让我深刻理解了模型训练中参数更新的动态机制，以及学习率、迭代次数等超参数对模型收敛速度和稳定性的影响。例如，当学习率设置过大时，模型可能会在最优解附近震荡甚至发散；而过小的学习率则会导致训练过程漫长。通过不断调试这些参数，我找到了一个相对合理的组合，使模型在电信客户流失数据集上达到了一定的预测准确率。

**随机森林**：随机森林作为一种强大的集成学习算法，其通过构建多个决策树并综合它们的预测结果来提高模型性能的思想给我留下了深刻印象。在实践中，调整超参数的过程让我明白了各个参数（如决策树数量、树深度、节点样本数、特征数等）如何影响模型的复杂度和泛化能力。例如，增加决策树数量可以增强模型的稳定性，但过多可能导致过拟合且增加计算成本；合理限制树深度能有效防止模型过度学习训练数据中的细节。通过随机搜索和贝叶斯优化等方法寻找最优超参数组合，我观察到模型性能在准确率、F1值和AUC等指标上的显著提升。

**支持向量机**：尽管在本次项目中支持向量机在处理电信客户流失数据集时表现欠佳，但对其原理的深入研究让我拓展了机器学习算法的视野。理解其通过核函数将数据映射到高维空间寻找最优分类超平面的方法，以及软间隔概念的引入以处理实际数据中的噪声和异常点，是一次对复杂算法理论的深度探索。这让我明白不同算法在不同数据分布和特征情况下的适用性差异，以及如何根据数据特点选择合适的算法。

**决策树**：决策树的构建过程基于贪心算法，通过选择最优分裂特征来逐步形成树结构，这一过程让我清晰地看到数据是如何被分类和决策的。在项目中，我意识到决策树容易出现过拟合的问题，尤其是在处理复杂数据时，树的深度过深会导致模型过于复杂，对训练数据过度拟合，从而降低在测试集上的泛化能力。

**朴素贝叶斯**：朴素贝叶斯基于贝叶斯定理和特征条件独立假设进行分类预测，其简洁的模型结构和高效的计算速度在某些特定场景下具有优势。然而，实际数据中特征往往并非完全独立，这一假设的局限性在一定程度上影响了模型的准确性。在电信客户流失预测项目中，虽然朴素贝叶斯能够给出一个初步的预测结果，但与其他模型相比，其性能相对较弱。

**XGBoost**：XGBoost 作为一种先进的梯度提升决策树算法，其在处理复杂数据关系和特征交互方面表现出色。深入理解其目标函数由损失函数和正则化项构成，以及通过二阶泰勒展开近似目标函数来优化模型的方法，让我感受到了算法设计的精妙之处。在实践中，利用贝叶斯优化对超参数进行动态调整，并结合早停策略有效防止过拟合，使模型在客户流失预测任务中取得了优异的性能表现。。

1. **特征工程的关键作用**

**数据预处理**：数据预处理是整个机器学习流程中不可或缺的重要环节。在处理电信客户流失数据集时，从最初的数据读取、查看数据结构和基本信息，到处理缺失值和重复值，每一步都需要细致入微。例如，将“TotalCharges”列的数据类型转换为数值型时，遇到了部分数据无法转换的问题，通过合理设置参数将其转换为缺失值后，再根据数据分布选择中位数填充，确保了数据的完整性和合理性。数据可视化在这个过程中发挥了关键作用，通过绘制各种图表，我能够直观地观察数据的分布特征、各特征之间的相关性以及客户流失情况在不同特征下的表现。

**特征选择**：特征选择直接关系到模型的性能和可解释性。通过卡方检验和交叉分析，我学会了从众多特征中筛选出与客户流失显著相关的特征，去除冗余和无关特征，从而降低模型的复杂性，提高模型的训练效率和预测准确性。卡方检验为特征选择提供了一种基于统计学原理的方法，能够快速判断特征与目标变量之间的相关性是否显著。而交叉分析则进一步深入挖掘了特征与客户流失之间的具体关系，明确了哪些特征值对客户流失具有较大影响。

**特征处理**：特征处理是将原始数据转换为适合模型训练的形式的重要步骤。对数值型特征进行标准化处理，使其符合特定模型的要求，能够加速模型训练过程并提高模型性能。通过将连续特征离散化，例如利用四分位数将其转化为分类特征，有助于树模型更好地捕捉数据中的非线性关系，提高模型的分类效果。针对样本不平衡问题采用过采样技术（如 SMOTE），平衡了数据集的类别分布，使模型在训练过程中能够更加关注少数类样本，提高了对客户流失这一少数类别的预测能力。

## 6.2 模型优化的思考

**逻辑回归**：在逻辑回归中，超参数调整是优化模型的重要途径。学习率和迭代次数的不同取值对模型收敛速度和准确性有着显著影响。较小学习率虽能保证稳定收敛，但可能导致训练时间过长；过大学习率则可能使模型无法收敛或错过最优解。迭代次数过少会使模型欠拟合，过多可能引发过拟合。通过不断尝试不同组合，我深刻体会到在实际应用中需根据数据特点和计算资源合理选择超参数，以达到最佳性能。

**随机森林**：随机森林的超参数众多且相互关联。决策树数量的增加可提升模型稳定性，但过多会增加计算成本和过拟合风险；树深度影响模型复杂度，过深易导致过拟合，过浅则可能欠拟合；节点样本数和特征数的调整也需谨慎权衡。通过随机搜索和贝叶斯优化寻找最优超参数组合的过程，以实现模型性能的最大化。

**XGBoost**：XGBoost 的超参数调整更为复杂且关键。决策树数量、学习率、树深度、子采样比例和特征采样比例等参数共同影响模型性能。例如，学习率与决策树数量需协同调整，以平衡模型学习能力和防止过拟合。通过贝叶斯优化和早停策略，我发现即使微小的参数调整也可能对模型性能产生较大影响。

**不同模型特性与适用性对比**：对比逻辑回归、随机森林、XGBoost 等多种模型在电信客户流失预测中的表现，清晰认识到各模型的特性和适用场景差异。逻辑回归适用于线性关系明显、数据规模大且对可解释性要求较高的情况；随机森林和 XGBoost 在处理非线性关系和复杂数据时优势明显，但解释性相对较差。在实际业务中，需根据数据特点、业务需求和对模型性能与解释性的要求，综合权衡选择最合适的模型。

**集成学习的优势与挑战分析**：集成学习通过组合多个基模型提高性能，但也面临诸多挑战。随机森林和XGBoost作为集成学习模型，其性能提升得益于基模型的多样性和组合方式。然而，选择合适的基模型、确定最佳组合权重以及控制计算成本都是需要解决的问题。

# 7 参考文献

1. 李海,熊升华,孙鹏.基于特征工程的S-FCN火灾图像检测方法[J].中国安全科学学报,2024,34(09):191-201.DOI:10.16265/j.cnki.issn1003-3033.2024.09.2063.
2. 孙苑苑.基于特征工程的运营商数据在泛金融行业的应用[J].江苏通信,2024,40(03):60-64.
3. 王琼,马恩点,胡潇杰,等.基于随机森林和逻辑回归的舟山梅雨期暴雨关键指标研究[J].科技通报,2024,40(09):7-12.DOI:10.13774/j.cnki.kjtb.2024.09.002.
4. 何雪峰,何厚华,杨蕾,等.基于随机森林模型的烟草销量影响因素分析[J].信息技术,2024,(11):147-153.DOI:10.13274/j.cnki.hdzj.2024.11.022.
5. 杨光飞, et al."基于贝叶斯优化XGBoost的Ⅰ类切口预防用抗菌药物多标签处方点评."中华医院感染学杂志 23(2024):3644-3649.
6. 彭白雪,陈清华,季家东.基于XGBoost和SHAP的制冷系统故障分析[J].低温与超导,2024,52(07):89-96.DOI:10.16711/j.1001-7100.2024.07.014

# 附录

**用户交互界面app代码（完整代码见Customer\_loss）**

import pandas as pd  
import numpy as np  
import streamlit as st  
from sklearn.model\_selection import train\_test\_split, cross\_val\_score  
from sklearn.linear\_model import LogisticRegression  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.svm import SVC  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.naive\_bayes import GaussianNB  
from xgboost import XGBClassifier  
from sklearn.metrics import accuracy\_score, roc\_auc\_score, confusion\_matrix, classification\_report, ConfusionMatrixDisplay, roc\_curve  
from sklearn.preprocessing import LabelEncoder  
import shap  
import matplotlib.pyplot as plt  
import seaborn as sns  
import warnings  
warnings.filterwarnings('ignore')  
  
*# 设置 Streamlit 标题*st.title("客户流失预测系统")  
  
*# 数据加载模块*st.sidebar.header("数据加载")  
uploaded\_file = st.sidebar.file\_uploader("上传数据文件 (CSV 格式)", type=["csv"])  
  
if uploaded\_file is not None:  
 *# 加载数据* data = pd.read\_csv(uploaded\_file)  
 st.write("### 数据预览")  
 st.write(data.head())  
  
 *# 特征和目标分离* X = data.drop(columns=['churn'])  
 y = data['churn']  
  
 *# 编码类别型特征* categorical\_columns = X.select\_dtypes(include=['object']).columns  
 encoder = LabelEncoder()  
 for col in categorical\_columns:  
 X[col] = encoder.fit\_transform(X[col])  
  
 *# 拆分训练集和测试集* X\_train, X\_test, y\_train, y\_test = train\_test\_split(  
 X, y, test\_size=0.2, random\_state=42, stratify=y  
 )  
  
 *# 模型选择模块* st.sidebar.header("模型选择")  
 model\_name = st.sidebar.selectbox(  
 "选择模型",  
 ["逻辑回归", "随机森林", "支持向量机", "决策树", "朴素贝叶斯", "XGBoost"]  
 )  
  
 *# 模型初始化* if model\_name == "逻辑回归":  
 model = LogisticRegression(max\_iter=1000, random\_state=42)  
 elif model\_name == "随机森林":  
 model = RandomForestClassifier(random\_state=42)  
 elif model\_name == "支持向量机":  
 model = SVC(probability=True, random\_state=42)  
 elif model\_name == "决策树":  
 model = DecisionTreeClassifier(random\_state=42)  
 elif model\_name == "朴素贝叶斯":  
 model = GaussianNB()  
 elif model\_name == "XGBoost":  
 model = XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss', random\_state=42)  
  
 *# 模型训练* st.sidebar.header("模型训练")  
 if st.sidebar.button("开始训练"):  
 model.fit(X\_train, y\_train)  
  
 *# 模型性能评估* st.write(f"### 模型：{model\_name}")  
 y\_pred = model.predict(X\_test)  
 y\_pred\_proba = model.predict\_proba(X\_test)[:, 1]  
  
 *# 准确率与AUC* accuracy = accuracy\_score(y\_test, y\_pred)  
 auc = roc\_auc\_score(y\_test, y\_pred\_proba)  
 st.write(f"准确率: {accuracy:.4f}")  
 st.write(f"AUC: {auc:.4f}")  
  
 *# 混淆矩阵* st.write("### 混淆矩阵")  
 cm = confusion\_matrix(y\_test, y\_pred)  
 disp = ConfusionMatrixDisplay(confusion\_matrix=cm)  
 disp.plot(cmap='Blues', values\_format='d')  
 st.pyplot(plt)  
  
 *# ROC 曲线* st.write("### ROC 曲线")  
 fpr, tpr, \_ = roc\_curve(y\_test, y\_pred\_proba)  
 plt.figure(figsize=(8, 6))  
 plt.plot(fpr, tpr, label=f"ROC curve (AUC = {auc:.4f})", color='darkorange')  
 plt.plot([0, 1], [0, 1], linestyle='--', color='navy')  
 plt.xlabel('False Positive Rate')  
 plt.ylabel('True Positive Rate')  
 plt.title('ROC Curve')  
 plt.legend(loc="lower right")  
 st.pyplot(plt)  
  
 *# 特征重要性 (仅适用于支持特征重要性的模型)* if model\_name in ["随机森林", "决策树", "XGBoost"]:  
 st.write("### 特征重要性")  
 feature\_importances = model.feature\_importances\_  
 importance\_df = pd.DataFrame({  
 'Feature': X.columns,  
 'Importance': feature\_importances  
 }).sort\_values(by='Importance', ascending=False)  
 st.bar\_chart(importance\_df.set\_index('Feature'))  
  
 *# SHAP 分析 (仅适用于支持 SHAP 的模型)* if model\_name == "XGBoost":  
 st.write("### SHAP 分析")  
 shap.initjs()  
 explainer = shap.TreeExplainer(model)  
 shap\_values = explainer.shap\_values(X\_test)  
  
 *# 全局特征重要性* st.write("#### 全局特征重要性")  
 shap.summary\_plot(shap\_values, X\_test, show=False)  
 st.pyplot(plt)  
  
 *# 单个样本分析* st.write("#### 单个样本分析")  
 sample\_index = st.slider("选择样本索引", 0, X\_test.shape[0] - 1, 0)  
 st.write("#### 单个样本特征贡献")  
 shap.force\_plot(  
 explainer.expected\_value,  
 shap\_values[sample\_index, :],  
 X\_test.iloc[sample\_index, :],  
 matplotlib=True  
 )  
 st.pyplot(plt)