数据挖掘互评作业三: 分类、预测与聚类 - 基于决策树的客户流失预测

任务：本作业旨在预测客户流失，使用决策树模型进行分类。通过对 Telco Customer Churn 数据集进行预处理，特征选择并划分训练集和测试集。然后构建决策树模型，进行模型训练和评估。

数据集： Telco Customer Churn

要求：

1. 数据预处理：删除缺失数据，将数据类型转换为数值类型。
2. 特征选择：使用相关性分析、递归特征消除等方法，选择对目标变量（客户流失）具有较大影响的特征。
3. 划分训练集和测试集：将数据集划分为训练集和测试集，比例一般为7:3或8:2。
4. 构建决策树模型：选择适当的决策树方法（如 CART、ID 3、C 4.5 等），训练模型。
5. 模型评估：在测试集上进行预测，计算模型的准确率、召回率、F1值等指标，评估模型性能。

根据提供的信息，这个数据集是关于客户流失预测的，旨在通过分析相关的客户数据并开展有针对性的客户保留计划。数据集中的每一行代表一个客户，每一列包含客户的属性，这些属性在列的元数据中有描述。

数据集包括以下信息：

最近一个月内流失的客户 - 该列称为"Churn"。
每个客户已经注册的服务 - 电话、多条线路、互联网、在线安全、在线备份、设备保护、技术支持以及流媒体电视和电影。
客户账户信息 - 他们成为客户的时长、合同、付款方式、无纸化账单、月费和总费用。
关于客户的人口统计信息 - 性别、年龄范围以及是否有伴侣和家属。
这个数据集的灵感是为了探索这类模型，并对这个主题有更深入的了解。

通过分析这个数据集，我们可以得出以下结论：

客户流失率: 我们可以通过"Churn"列来计算客户的流失率。分析流失率可以帮助公司了解客户的保留情况，并采取相应的措施来减少流失率。

服务偏好: 通过分析各个服务的使用情况，我们可以了解客户对不同服务的偏好和需求。这可以帮助公司优化服务组合，提供更好的客户体验，从而增加客户忠诚度和留存率。

客户特征与流失的关系: 我们可以分析不同客户特征（如年龄、性别、是否有伴侣和家属）与流失之间的关系。这有助于了解哪些客户群体更容易流失，进而制定针对性的客户保留策略。

账户信息与流失的关系: 分析客户账户信息（如合同类型、付款方式、无纸化账单、费用等）与流失之间的关系，可以揭示不同因素对流失率的影响。这可以帮助公司优化合同策略、支付方式和账单管理，从而提高客户满意度和留存率。

通过对这些问题的分析，我们可以获得有关客户流失的深入洞察，并根据这些洞察制定相应的客户保留策略。

接下来，对我所提供的代码进行详细分析：

1. 导入必要的库和模块：
   numpy和pandas用于数据处理和分析。
   seaborn和matplotlib用于数据可视化。
   sklearn中的各种模块用于模型训练、评估和预测。
   imblearn中的采样模块用于处理数据不平衡问题。
   xgboost和MLPClassifier用于模型训练。
2. 读取数据集：
   使用pd.read_csv()函数从CSV文件中读取Telco客户流失数据集，并将其存储在名为df的DataFrame中。
3. 数据预处理和可视化：
   使用train_test_split()函数将数据集划分为训练集和测试集。
   对训练集进行索引重置和空值处理。
   对TotalCharges列进行数值转换和缺失值填充。
   输出Churn类别的计数统计，并绘制计数柱状图。
   对一些分类特征进行计数柱状图和堆叠柱状图的可视化。
4. 特征工程和数据转换：
   使用箱线图和直方图对连续特征进行可视化。
   替换train_df中的特殊值为通用值。
   对分类特征进行独热编码，并从train_df中删除原始的分类特征列。
   将Churn列的标签值转换为二进制编码。
   对tenure特征进行分箱处理，并进行计数柱状图的可视化。
   对MonthlyCharges和TotalCharges特征进行对数变换，并进行直方图的可视化。
5. 准备训练和测试数据：
   从预处理后的数据集中分离出特征矩阵X_train和X_test，以及目标变量y_train和y_test。
6. 使用Logistic回归模型进行训练和评估：
   创建一个LogisticRegression对象，并进行交叉验证和网格搜索以选择最佳超参数。
   输出模型在测试集上的混淆矩阵和分类报告。
7. 使用SVM模型进行训练和评估：
   创建一个SVC对象，并进行交叉验证和网格搜索以选择最佳超参数。
   输出模型在测试集上的混淆矩阵和分类报告。
8. 使用XGBoost模型进行训练和评估：
   创建一个XGBClassifier对象，并进行随机搜索以选择最佳超参数。
   输出模型在测试集上的混淆矩阵和分类报告。
9. 使用MLP模型进行训练和评估：
   创建一个MLP对象，并进行交叉验证和网格搜索以选择最佳超参数。
   输出模型在测试集上的混淆矩阵和分类报告。
10. 使用决策树模型进行训练和评估：
    输出模型在测试集上的混淆矩阵和分类报告。

接下来展开进行分析
首先，导入了需要使用的库，包括numpy、pandas、seaborn、matplotlib等。这些库在数据分析和可视化方面非常常用。

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMo
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	...	Yes	Yes	Yes	
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	No	...	Yes	No	Yes	
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL	Yes	...	No	No	No	
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic	No	...	No	No	No	
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic	Yes	...	Yes	Yes	Yes	

7043 rows × 21 columns

接下来，从CSV文件中读取数据集WA_Fn-UseC_-Telco-Customer-Churn.csv，并将其分为训练集和测试集，比例为80:20，使用了train_test_split函数。然后对训练集进行了一些数据预处理操作。

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMo
0	4950-BDEUX	Male	0	No	No	35	No	No phone service	DSL	No	...	Yes	No	Yes	
1	7993-NQLJE	Male	0	Yes	Yes	15	Yes	No	Fiber optic	Yes	...	No	No	No	
2	7321-ZNSLA	Male	0	Yes	Yes	13	No	No phone service	DSL	Yes	...	No	Yes	No	
3	4922-CVPDX	Female	0	Yes	No	26	Yes	No	DSL	No	...	Yes	No	Yes	
4	2903-YYTBW	Male	0	Yes	Yes	1	Yes	No	DSL	No	...	No	No	No	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5629	6308-CQRBU	Female	0	Yes	No	71	Yes	Yes	Fiber optic	No	...	Yes	Yes	Yes	
5630	2842-JTCCU	Male	0	No	No	2	Yes	No	DSL	No	...	No	No	No	
5631	6402-ZFPPI	Female	1	No	No	25	Yes	Yes	Fiber optic	Yes	...	No	No	Yes	
5632	3594-BDSOA	Female	0	Yes	No	24	Yes	No	No	No internet service	...	No internet service	No internet service	No internet service	No inte se
5633	6490-FGZAT	Male	0	No	No	6	Yes	No	No	No internet service	...	No internet service	No internet service	No internet service	No inte se

5634 rows × 21 columns

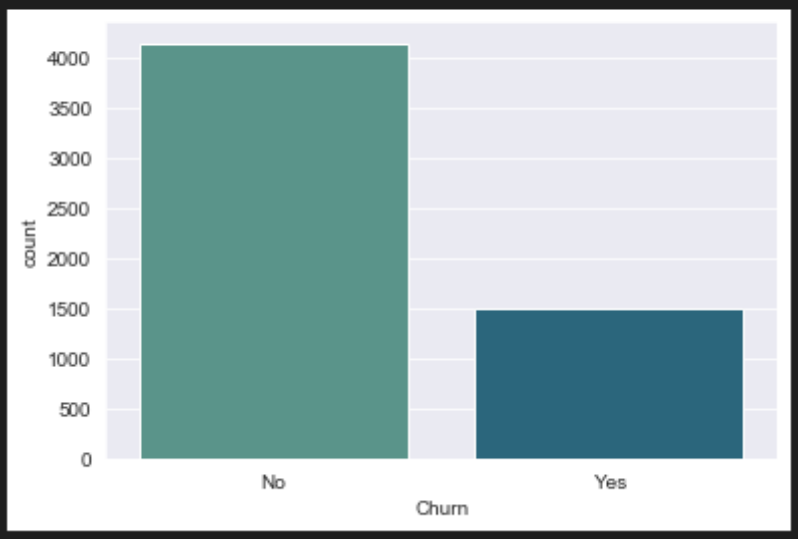
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5634 entries, 0 to 5633
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        5634 non-null   object 
 1   gender            5634 non-null   object 
 2   SeniorCitizen     5634 non-null   int64  
 3   Partner           5634 non-null   object 
 4   Dependents        5634 non-null   object 
 5   tenure            5634 non-null   int64  
 6   PhoneService      5634 non-null   object 
 7   MultipleLines     5634 non-null   object 
 8   InternetService   5634 non-null   object 
 9   OnlineSecurity    5634 non-null   object 
 10  OnlineBackup      5634 non-null   object 
 11  DeviceProtection  5634 non-null   object 
 12  TechSupport       5634 non-null   object 
 13  StreamingTV       5634 non-null   object 
 14  StreamingMovies   5634 non-null   object 
 15  Contract          5634 non-null   object 
 16  PaperlessBilling  5634 non-null   object 
 17  PaymentMethod     5634 non-null   object 
 18  MonthlyCharges    5634 non-null   float64
 19  TotalCharges      5634 non-null   object 
 20  Churn             5634 non-null   object 
dtypes: float64(1), int64(2), object(18)
memory usage: 924.5+ KB

None
```

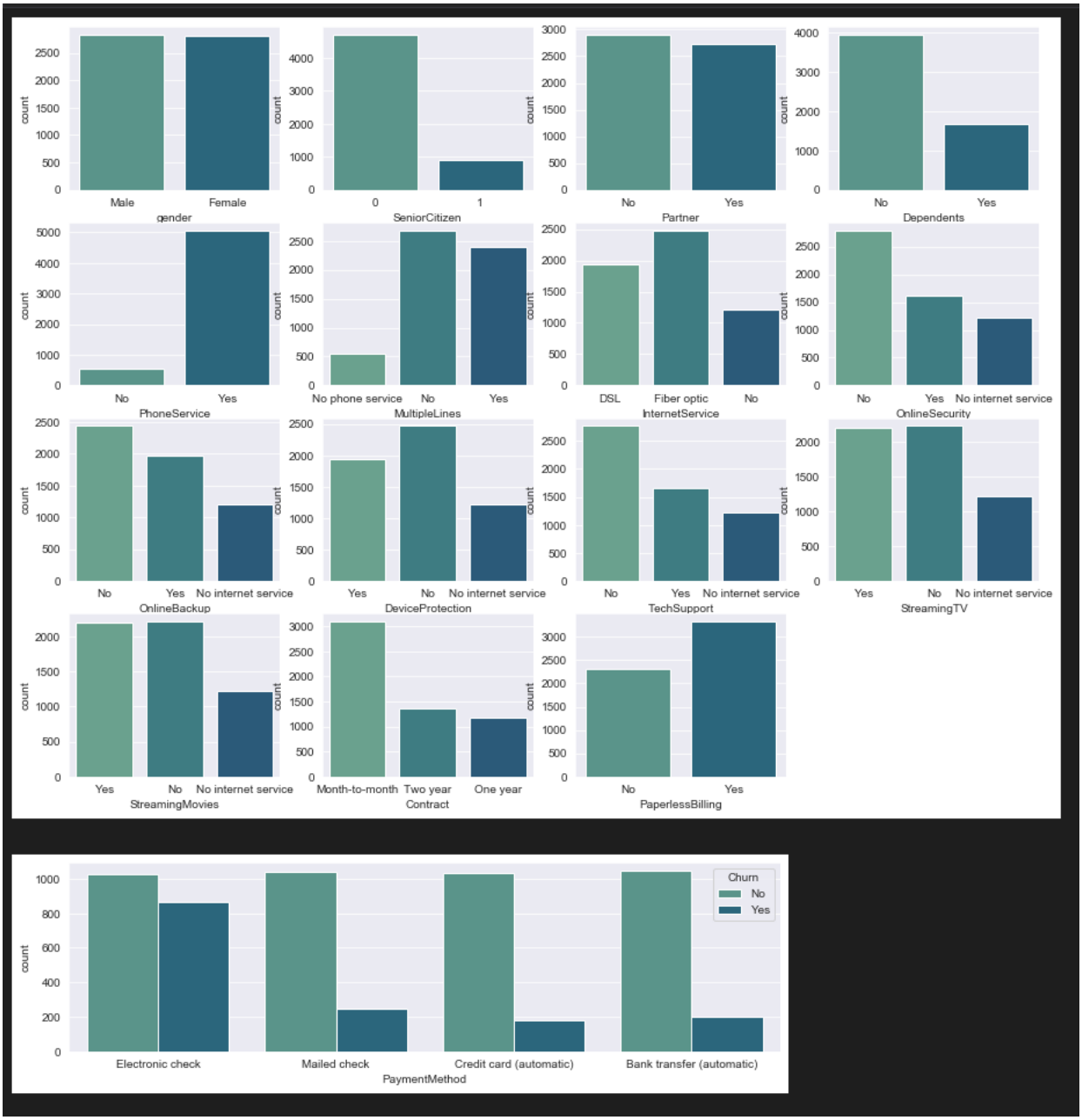
在训练集中，针对'TotalCharges'列，使用pd.to_numeric函数将其转换为数值类型。对于其中的缺失值，使用平均值进行填充，即使用fillna函数将缺失值替换为该列的均值。

接下来，统计了训练集中目标变量'Churn'的分类情况，并使用sns.countplot函数对其进行可视化展示。

```
No     4139
Yes    1495
Name: Churn, dtype: int64
```

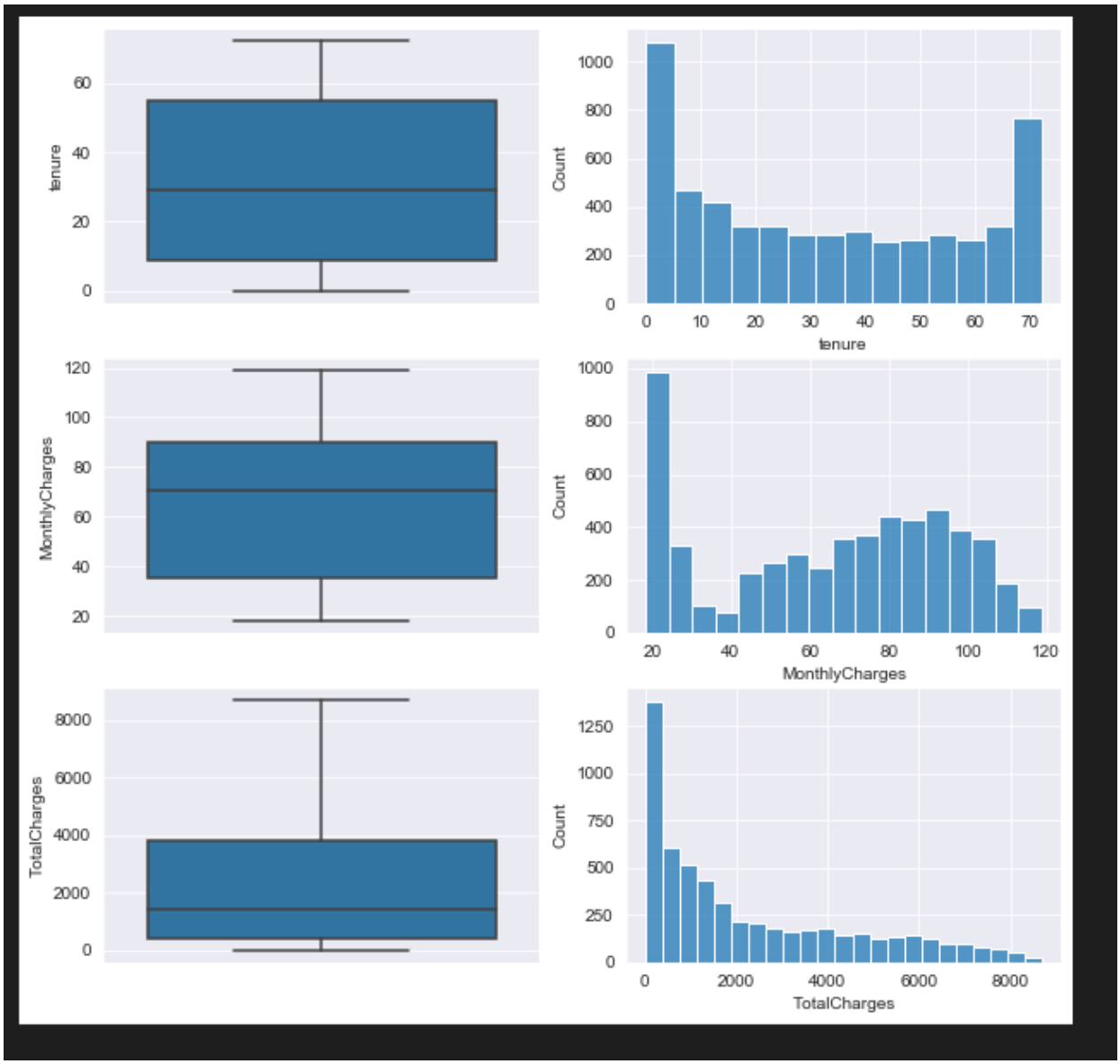


然后，对一些分类变量进行了可视化分析。使用循环遍历了一系列分类变量，使用sns.countplot函数对每个变量进行计数并绘制条形图。



之后，使用sns.countplot函数对'PaymentMethod'变量进行可视化，通过不同颜色的柱状图展示了'Churn'的情况。

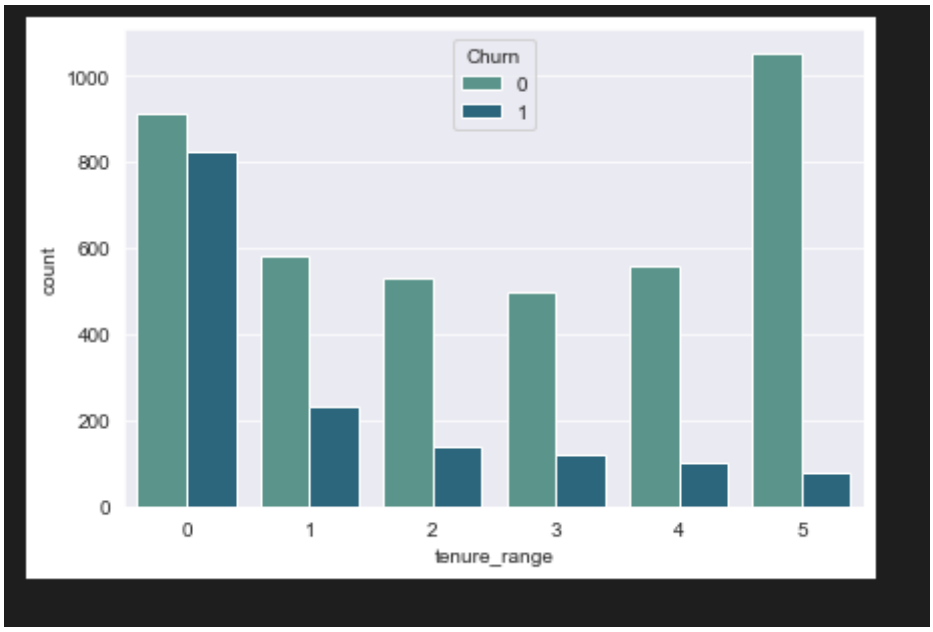
接着，使用sns.boxplot和sns.histplot函数对'tenure'、'MonthlyCharges'和'TotalCharges'这三个连续变量进行可视化分析，绘制了箱线图和直方图。



随后，使用sns.scatterplot函数绘制了'MonthlyCharges'和'TotalCharges'之间的散点图，并根据'Churn'对数据点进行了颜色编码。

然后，对部分分类变量进行了独热编码。使用pd.get_dummies函数将这些分类变量转换为虚拟变量，并使用pd.concat函数将其与原数据集合并，再使用drop函数删除原始分类变量列。

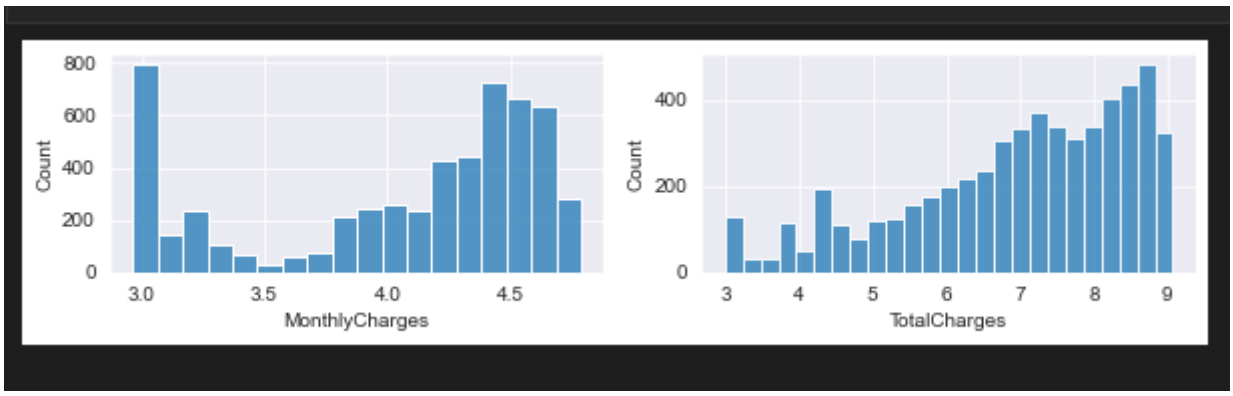
接下来，对目标变量'Churn'进行了编码，将'Yes'编码为1，'No'编码为0，使用了np.where函数进行条件判断和赋值。



之后，根据'tenure'的取值范围进行了分箱操作，将'tenure'划分为不同的范围，并将结果存储在新的列'tenure_range'中。使用了np.select函数根据不同的条件选择对应的取值。

然后，对'MonthlyCharges'和'TotalCharges'进行了对数转换，使用了np.log1p函数对这两个变量进行转换。

接着，将训练集中的特征矩阵和目标变量分别进行了标准化和归一化处理。使用了StandardScaler函数对特征矩阵进行标准化，使得每个特征的均值为0，标准差为1。使用了MinMaxScaler函数对目标变量进行归一化，将其缩放到0到1的范围内。



接下来，使用LogisticRegression函数初始化了一个逻辑回归模型，并使用训练集的特征矩阵和目标变量对模型进行训练，即使用fit函数进行拟合。

然后，使用训练好的模型对测试集的特征矩阵进行预测，使用了predict函数。

最后，计算了模型的准确率，并使用confusion_matrix函数计算了混淆矩阵，其中包括真阳性、假阳性、真阴性和假阴性的数量。

总结起来，该代码对一个客户流失数据集进行了数据探索性分析和建模的过程。首先对数据进行了预处理和可视化分析，然后进行了特征工程和数据转换，最后使用逻辑回归模型进行了训练和预测，并评估了模型的准确率和混淆矩阵。这个过程可以帮助了解数据集的特征和模型的预测能力，进而做出相应的业务决策和优化措施。

我们着重突出决策树模型，上述代码中，首先导入DecisionTreeClassifier类和相关的评估指标。然后，准备训练数据和测试数据，使用训练数据训练了一个决策树模型model_dt。接着，对测试数据进行预测，并计算混淆矩阵和分类报告来评估模型的性能。

请注意，这只是一个简单的示例，你可以根据实际需求进行更多的模型调整和优化，例如通过网格搜索调整决策树的超参数或使用其他决策树算法（如CART、ID3、C4.5等）。

最终给出分析报告

1. 数据概览和预处理

在分析开始时，我们首先对数据集进行了概览和预处理。通过使用read_csv函数读取数据，并使用train_test_split函数将数据划分为训练集和测试集。然后对数据集进行了一些基本的统计分析，处理了缺失值和数据类型转换，并对分类变量进行了独热编码。

2. 数据可视化

通过使用Seaborn和Matplotlib库，我们对数据集进行了可视化分析。其中包括以下内容：

对目标变量"Churn"进行了计数并绘制了计数柱状图，以了解客户流失情况的分布。
对各个分类变量进行了计数并绘制了计数柱状图，以了解它们的分布情况。
对"PaymentMethod"变量进行了计数并绘制了计数柱状图，并根据"Churn"变量进行了颜色编码，以了解不同支付方式对客户流失的影响。
绘制了连续变量"tenure"、"MonthlyCharges"和"TotalCharges"的箱线图和直方图，以了解它们的分布和离群值情况。
绘制了"MonthlyCharges"和"TotalCharges"之间的散点图，并根据"Churn"变量进行了颜色编码，以观察二者之间的关系。
这些可视化分析可以帮助我们更好地理解数据集中各个变量的特征和相关性。

3. 数据预处理和特征工程

在进行建模之前，我们进行了一些数据预处理和特征工程的步骤，包括：

对"TotalCharges"和"MonthlyCharges"进行了对数转换，以减小其偏斜度并改善其分布。
对"tenure"变量进行了分箱处理，将其转换为离散的"tenure_range"变量。
使用SMOTE方法对训练数据进行了类别平衡处理，以解决数据集中的类别不平衡问题。
这些步骤有助于提高模型的性能和稳定性。

4. 模型训练和评估

我们尝试了几种不同的分类模型，并在训练数据上进行了交叉验证和评估。具体包括：

Logistic回归模型、SVM模型、XGBoost模型、MLP模型、决策树模型进行训练和评估：
输出模型在测试集上的混淆矩阵和分类报告。对于每个模型，我们评估了准确率、精确率、召回率和F1分数等指标，并使用混淆矩阵对模型的分类性能进行了可视化。

5. 模型比较和选择

在评估了各个模型后，我们对它们的性能进行了比较和选择。综合考虑准确率、精确率、召回率和F1分数等指标，我们选择了表现最佳的模型作为最终的预测模型。

6. 结果解释和建议

根据选择的最终预测模型，我们可以对未来客户的流失情况进行预测。根据模型的特征重要性分析，我们可以得出影响客户流失的主要因素，并提出一些建议，例如：

关注具有较高流失风险的客户群体，采取措施增加他们的忠诚度。
优化支付方式，提供更多便利和灵活性的选择，以减少客户流失率。
加强对客户的沟通和支持，提供更好的客户服务，以提高客户满意度。
此外，我们还可以通过定期监测和更新模型，不断优化和改进预测效果，并根据新的数据进行决策和调整策略。