客户流失判断

1. 数据预处理

数据属性共15个，分别为：

- ID：编号
- Contract：是否有合同
- Dependents：是否有家属
- DeviceProtection：是否有设备保护
- IntenrnetService：是否有互联网服务
- MonthlyCharges：月度费用
- MultipleLines：是否有多条线路
- Partner：是否有配偶
- PaymentsMethod：付款方式
- PhoneService：是否有电话服务
- SeniorCitizen：是否为老年人
- TVProgram：是否有电视节目
- TotalCharges：总费用
- gender：性别
- tenure：任期年数

在训练集中，增加了标签属性：

- Label：用户是否流失

1.1 数据摘要

训练集和测试集数据样例个数分别为5227和1307，在训练集中客户流失比例约占37%。

各属性的数据类型如下：

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                5227 non-null   int64
 1   Contract          5227 non-null   object
 2   Dependents        5227 non-null   object
 3   DeviceProtection  5227 non-null   object
 4   InternetService   5227 non-null   object
 5   MonthlyCharges    5227 non-null   float64
 6   MultipleLines     5227 non-null   object
 7   Partner           5227 non-null   object
 8   PaymentMethod     5227 non-null   object
 9   PhoneService      5227 non-null   object
 10  SeniorCitizen     5227 non-null   int64
 11  TVProgram         5227 non-null   object
 12  TotalCharges      5227 non-null   float64
 13  gender            5227 non-null   object
 14  tenure            5227 non-null   int64
 15  Label             5227 non-null   object
dtypes: float64(2), int64(3), object(11)
```

查看各属性取值发现如下：

```
(ID                 5227
 Contract              3
 Dependents            2
 DeviceProtection      3
 InternetService       3
 MonthlyCharges     2620
 MultipleLines         3
 Partner               2
 PaymentMethod         4
 PhoneService          2
 SeniorCitizen         2
 TVProgram             3
 TotalCharges       5016
 gender                2
 tenure               73
 Label                 2
 dtype: int64,
 ID                 1307
 Contract              3
 Dependents            2
 DeviceProtection      3
 InternetService       3
 MonthlyCharges      900
 MultipleLines         3
 Partner               2
 PaymentMethod         4
 PhoneService          2
 SeniorCitizen         2
 TVProgram             3
 TotalCharges       1289
 gender                2
 tenure               72
 dtype: int64)
```

经观察后初步发现：

- PhoneService为No的MultipleLines为No phone service，二者存在关联
- InternetService为No的DeviceProtection，TVProgram为No internet service，存在关联
- tenure为0的8行数据对应的TotalCharfes为2283.300441，在TotalCharges中为最大值，考虑可能是为tenure为0的用户总费用设为最大值。
- 大部分属性的取值个数为2-4个，且数值类型为object，需要进行处理。

1.2 数据清洗

主要任务包括：

- 二值属性转为0/1，其中yes为1,No为0
- 多值属性用one-hot，去掉第一个，将处理后数据附加在数据后
- 包含No xx service的多值属性，one-hot值去掉该列
- 删除处理前的属性列

将数据类型为object，取值为yes/no的属性Dependents','Partner','PhoneService','Label'替代为1/0

```
collist = ['Dependents','Partner','PhoneService','Label']
# map函数
def binary_map(x):
    return x.map({'Yes':1,"No":0})
custmer_train[collist] =
custmer_train[collist].apply(binary_map)
col = ['Dependents','Partner','PhoneService']
custmer_test[col] = custmer_test[col].apply(binary_map)
```

多值属性如'Contract','InternetService','PaymentMethod','gender'修改为One-hot属性并去掉第一个新属性。

```
# train one-hot
# 3个属性，新增2*2+3*1+1*1
dummy1 =
pd.get_dummies(custmer_train[['Contract','InternetService'
,'PaymentMethod','gender']],
                      drop_first=True)

# 拼接到源数据
custmer_train = pd.concat([custmer_train,dummy1],axis=1)

# test
dummy2 =
pd.get_dummies(custmer_test[['Contract','InternetService',
'PaymentMethod','gender']],
                      drop_first=True)
custmer_test = pd.concat([custmer_test,dummy2],axis=1)
```

针对一些具有no xx service属性值的数据如‘DeviceProtection’，‘MultipleLines’，‘TVProgram’，考虑到该类取值较少且对结果影响较小，one-hot处理后去掉该属性。

将与上述操作相关的源数据删除后的数据类型如下：

```
RangeIndex: 5227 entries, 0 to 5226
Data columns (total 23 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   ID                                    5227 non-null   int64
 1   Dependents                            5227 non-null   int64
 2   MonthlyCharges                        5227 non-null   float64
 3   Partner                               5227 non-null   int64
 4   PhoneService                          5227 non-null   int64
 5   SeniorCitizen                         5227 non-null   int64
 6   TotalCharges                          5227 non-null   float64
 7   tenure                                5227 non-null   int64
 8   Label                                 5227 non-null   int64
 9   Contract_One year                     5227 non-null   uint8
 10  Contract_Two year                     5227 non-null   uint8
 11  InternetService_Fiber optic           5227 non-null   uint8
 12  InternetService_No                    5227 non-null   uint8
 13  PaymentMethod_Credit card (automatic) 5227 non-null   uint8
 14  PaymentMethod_Electronic check        5227 non-null   uint8
 15  PaymentMethod_Mailed check            5227 non-null   uint8
 16  gender_Male                           5227 non-null   uint8
 17  DeviceProtection_No                   5227 non-null   uint8
 18  DeviceProtection_Yes                  5227 non-null   uint8
 19  MultipleLines_No                      5227 non-null   uint8
 20  MultipleLines_Yes                     5227 non-null   uint8
 21  TVProgram_No                          5227 non-null   uint8
 22  TVProgram_Yes                         5227 non-null   uint8
```

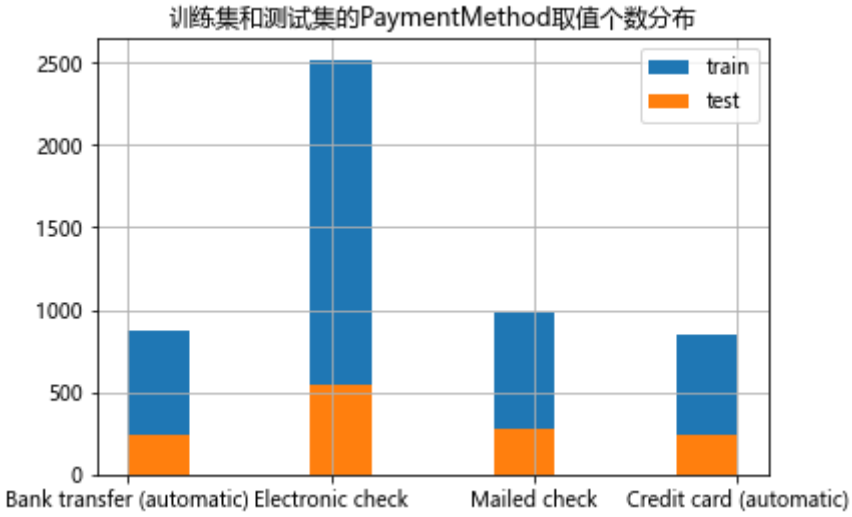
查看多值属性的相关数值统计信息如下：

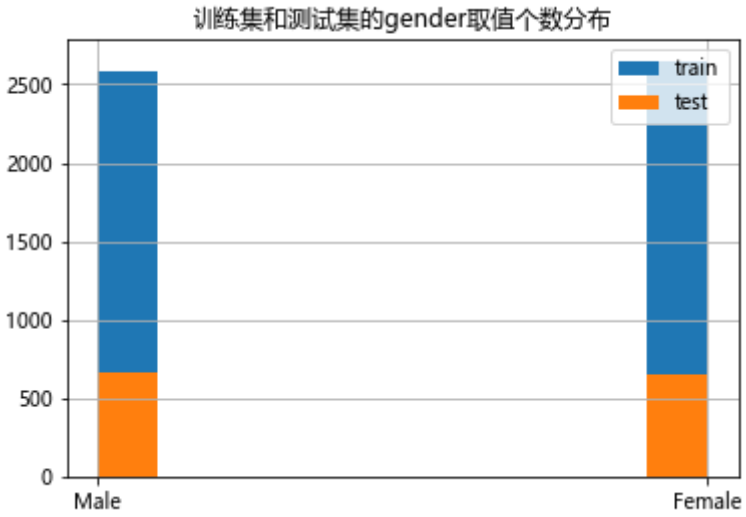
	tenure	MonthlyCharges	SeniorCitizen	TotalCharges
count	5227.000000	5227.000000	5227.000000	5227.000000
mean	28.775971	66.823765	0.118615	2084.477153
std	24.293077	28.862749	0.323366	2183.825066
min	0.000000	18.250000	0.000000	18.800000
25%	5.000000	45.000000	0.000000	292.979609
50%	23.000000	74.200000	0.000000	1218.650000
75%	51.000000	89.900000	0.000000	3373.825000
90%	67.000000	101.395283	1.000000	5683.670000
95%	71.000000	106.196180	1.000000	6632.480000
99%	72.000000	114.037000	1.000000	7937.495000
max	72.000000	118.600000	1.000000	8564.750000

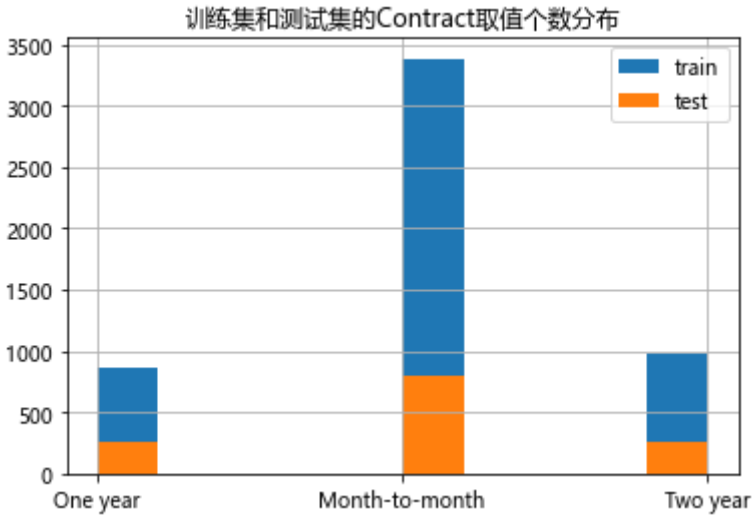
看到以上四种数据分布范围较广，需要对其进行特征缩放处理，其他的数据清洗操作如特征选择等将在最终报告中进行展示。

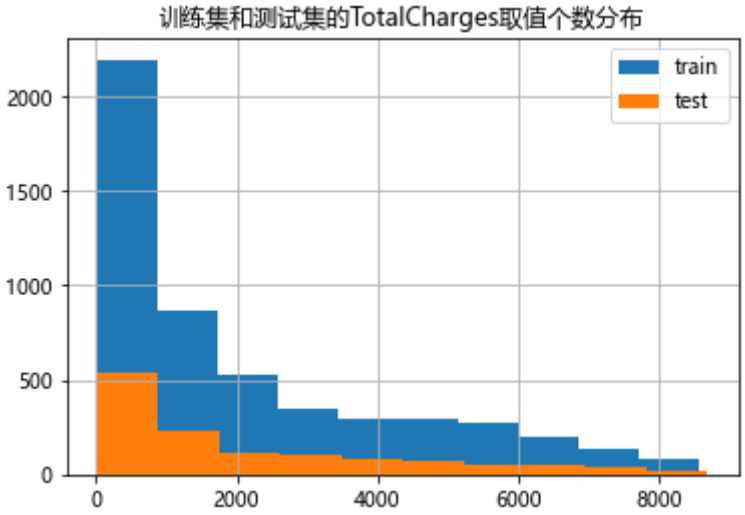
2. 数据可视化

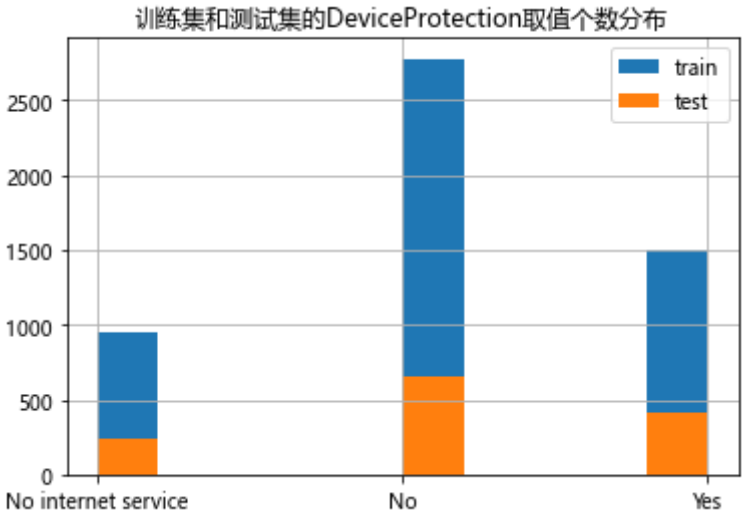
查看各属性在训练集和测试集的取值分布

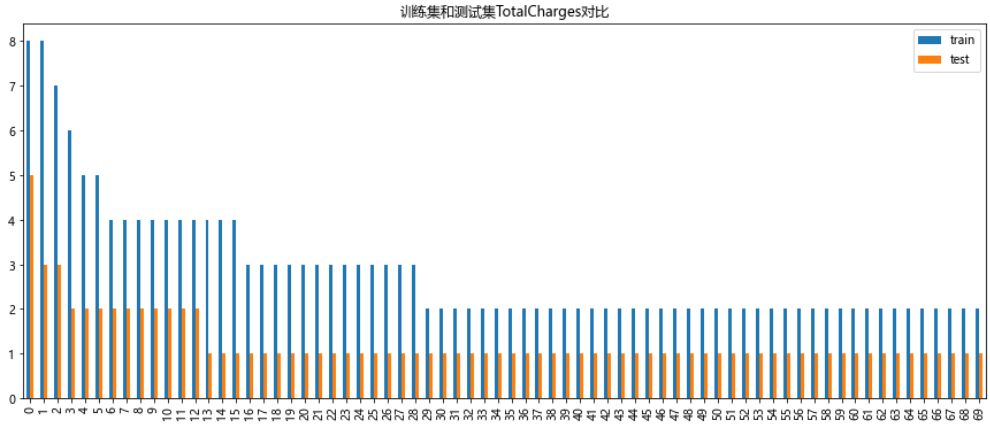

训练集和测试集的PaymentMethod取值个数分布
train
test
2500
2000
1500
1000
500
0
Bank transfer (automatic)
Electronic check
Mailed check
Credit card (automatic)


训练集和测试集的gender取值个数分布
train
test
2500
2000
1500
1000
500
0
Male
Female

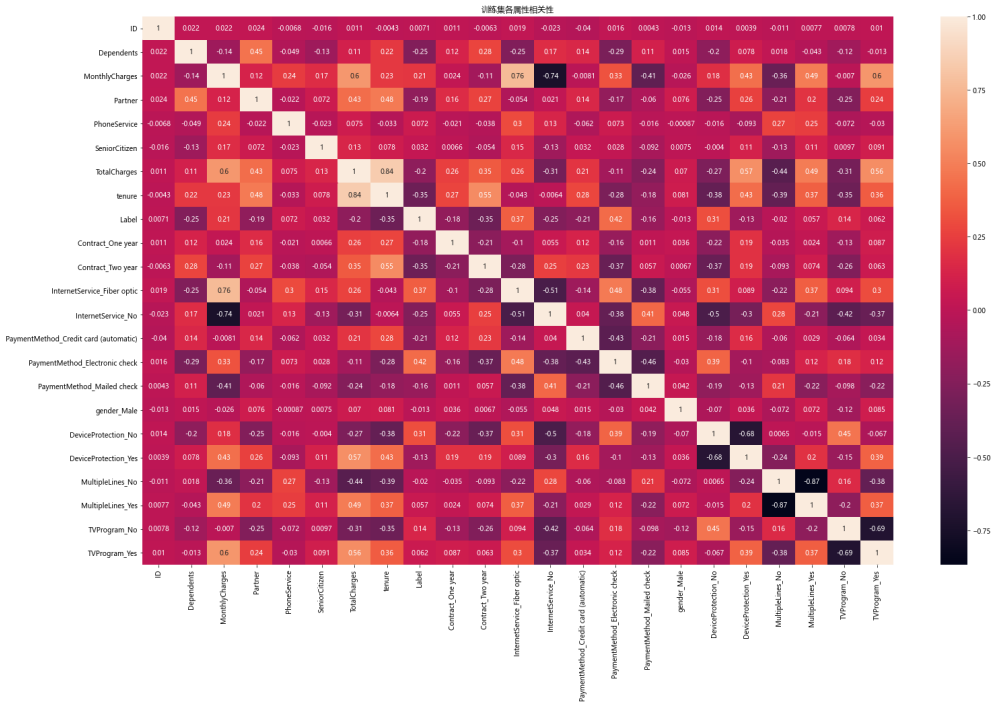

训练集和测试集的Contract取值个数分布
train
test
3500
3000
2500
2000
1500
1000
500
0
One year
Month-to-month
Two year







经上述分析，可以看出训练集和测试集在数据分布比例上基本一致。



从属性相关性图中可看出有一些属性是强相关的，在特征选取时需要将其删除，避免影响模型效果。

在最终报告中将阐述更详尽的可视化结果。

3. 模型选取

与 项目提出时预选的模型一致，我们将用数据集训练各种机器学习及神经网络模型，用准确率评价各模型的结果，并比较模型的效果。

4. 实验结果

在进行特征选取后使用模型进行客户流失判断，使用GBDT并进行微调后的预测准确率在76.7%，进一步的优化，所有模型的结果及分析将在最终报告中展示。

5. 存在的问题

特征选取不当可能对模型有一定影响，考虑优化特征工程，尝试改进二分类算法提高准确率。

6. 下一步工作

优化特征工程，选取代表性强的特征进行训练。

完善模型，对模型参数及结构，验证方法进行改进。

7. 任务完成情况

根据项目提出时的任务分配情况，我们针对每个任务进行了详尽的讨论，小组各成员的任务目前进展顺利，后期将致力于优化过程和提高准确率的任务。