# TIM图片20220321151525

# 数据预处理

## 异常值处理

Mobile\_Tag值只有1，对模型没有区分度，删除

|  |  |  |
| --- | --- | --- |
| 特征名称 | 异常值 | 替换值 |
| Client\_Income | $ | 0 |
| Loan\_Annuity | #VALUE!和$ | 0 |
| Age\_Days | # | 0 |
| Employed\_Days | # | 0 |
| Registration\_Days | # | 0 |
| ID\_Days | # | 0 |
| Credit\_Amount | $ | 0 |
| Population\_Region\_Relative | @和# | 0 |
| Score\_Source\_3 | & | 0 |

## 缺失值处理

**数值特征缺失值均值填充**

['Client\_Income', 'Car\_Owned', 'Bike\_Owned', 'Active\_Loan', 'House\_Own', 'Child\_Count', 'Credit\_Amount', 'Loan\_Annuity', 'Population\_Region\_Relative', 'Age\_Days', 'Employed\_Days', 'Registration\_Days', 'ID\_Days', 'Own\_House\_Age', 'Client\_Family\_Members', 'Cleint\_City\_Rating', 'Application\_Process\_Day', 'Application\_Process\_Hour', 'Score\_Source\_1', 'Score\_Source\_2', 'Score\_Source\_3', 'Social\_Circle\_Default', 'Phone\_Change', 'Credit\_Bureau']

**类别特征缺失值用众数填充**

['Accompany\_Client', 'Client\_Income\_Type', 'Client\_Education', 'Client\_Marital\_Status', 'Client\_Gender', 'Loan\_Contract\_Type', 'Client\_Housing\_Type', 'Client\_Occupation', 'Type\_Organization']

# 特征筛选

原始特征有37个

1. 首先根据方差筛选特征，方差越低，特征值离散程度越低，参考价值也不大。代码中将方差低于0.02的特征删除，剩余26个特征
2. 卡方检验筛选 TOP 20 特征作为待保留特征集合set1:['Type\_Organization', 'Client\_Education', 'Score\_Source\_3', 'Employed\_Days', 'Client\_Gender', 'Loan\_Contract\_Type', 'Client\_Income\_Type', 'Client\_Occupation', 'Phone\_Change', 'Workphone\_Working', 'Cleint\_City\_Rating', 'Homephone\_Tag', 'Score\_Source\_1', 'Car\_Owned', 'ID\_Days', 'Age\_Days', 'Accompany\_Client', 'Registration\_Days', 'Client\_Contact\_Work\_Tag', 'Client\_Permanent\_Match\_Tag']
3. F-检验筛选 TOP 20 特征作为待保留特征集合set2:['Score\_Source\_3', 'Score\_Source\_1', 'Age\_Days', 'Client\_Education', 'Cleint\_City\_Rating', 'Phone\_Change', 'ID\_Days', 'Client\_Gender', 'Employed\_Days', 'Client\_Permanent\_Match\_Tag', 'Registration\_Days', 'Client\_Income\_Type', 'Type\_Organization', 'Client\_Contact\_Work\_Tag', 'Loan\_Contract\_Type', 'Workphone\_Working', 'Car\_Owned', 'Homephone\_Tag', 'Client\_Occupation', 'Accompany\_Client']
4. 互信息筛选 TOP 20 特征作为待保留特征集合set3:['Score\_Source\_1', 'Age\_Days', 'Registration\_Days', 'Employed\_Days', 'ID\_Days', 'Phone\_Change', 'Score\_Source\_3', 'Type\_Organization', 'Client\_Education', 'Client\_Occupation', 'Client\_Income\_Type', 'Cleint\_City\_Rating', 'Client\_Gender', 'Client\_Permanent\_Match\_Tag', 'Client\_Housing\_Type', 'Loan\_Contract\_Type', 'Client\_Contact\_Work\_Tag', 'Client\_Marital\_Status', 'Workphone\_Working', 'Car\_Owned']
5. set1，set2，set3求交集就是最后保留的特征，总共剩下18个:['Age\_Days', 'Type\_Organization', 'Workphone\_Working', 'Car\_Owned', 'Score\_Source\_3', 'Loan\_Contract\_Type', 'Registration\_Days', 'Client\_Permanent\_Match\_Tag', 'Client\_Income\_Type', 'Client\_Gender', 'Employed\_Days', 'ID\_Days', 'Client\_Occupation', 'Cleint\_City\_Rating', 'Score\_Source\_1', 'Phone\_Change', 'Client\_Contact\_Work\_Tag', 'Client\_Education']

# 模型调参

采用五折交叉验证，逐个选择最优参数，红色代表最优参数值。调参过程和结果见excel。

## rf最优参数

n\_estimators=100,

max\_depth=18,

max\_samples=0.9,

max\_features='auto'

## lgb最优参数

learning\_rate=0.07,

max\_depth=10,

reg\_alpha=0,

reg\_lambda=0,

n\_estimators=700

## xgb最优参数

learning\_rate=0.04,

max\_depth=10,

n\_estimators=700,

reg\_alpha=0,

reg\_lambda=3

## svm最优参数

C=1,

degree=2,

kernel='poly',

max\_iter=650

# 模型效果

stacking，基模型选取随机森林，lgb和xgb，元模型选取lr。

|  |  |  |  |
| --- | --- | --- | --- |
|  | acc | auc | 混淆矩阵 |
| 随机森林 | 0.9212210733628754 | 0.758634693297603 | [[11196 2]  [ 958 30]] |
| lgb | 0.9187592319054653 | 0.7416470407888048 | [[11179 19]  [ 971 17]] |
| xgb | 0.9213031347447891 | 0.7482666619906824 | [[11186 12]  [ 947 41]] |
| SVM | 0.34186771705235514 | 0.5338505719283302 | [[3431 7767]  [ 253 735]] |
| stacking | 0.9103889709502708 | 0.7640189145979652 | [[10842 356]  [ 736 252]] |



