**预处理前数据量:175 3622条**

**预处理后数据量:163 2432条**

特征选取：

X:

'M\_TR\_CURRENCY'：直接转换成类别，数据无空值

'M\_SETTLEMENT\_AMOUNT'：空值12 1190条，是同一个商户的，可以去掉

'M\_TR\_CARDSCHEME\_ID'：直接转换成类别，数据无空值

'M\_CARDTYPE' ：空值23 7475条，用’NoType’填充

'M\_TR\_CARDBIN'：空值60 9952条，用’0000 0000’填充

'M\_CI\_IPCOUNTRY'：空值3001条，用’NoIpcountry‘填充

'M\_CI\_IPCITY' ：空值22 9554条，用’NoIpcity’填充

'M\_MS\_SHOPPER\_FACTOR'：空值1336条，用’1’填充

'M\_PAY\_DURATION'：空值40 2717条，用求得的均值填充

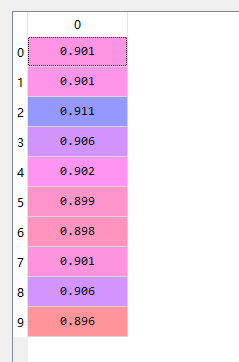
'M\_CI\_PLATFORM'：空值44 2738条，’NoPlatform’填充

Y:

'M\_ISFRAUD\_CB'

1. **训练集选择：18年1月1日-18年7月30日**

用决策树模型，选取100W数据进行训练，经过交叉验证得到的模型性能：





模型准确率：90.2%



**测试集选择：18年7月31日-18年10月30日**

第一次分类（决策树）：

|  |  |  |
| --- | --- | --- |
| 实际 预测 | 非欺诈 | 欺诈 |
| 非欺诈 | 实际非欺诈，预测非欺诈  的有501653条 | 实际非欺诈，**预测欺诈**  **的有130685条** |
| 欺诈 | 实际欺诈，预测非欺诈  的有9条 | 实际欺诈，**预测欺诈**  **的有85条** |

第二次分类就是拿**第一次预测出来的欺诈订单**（13 0685+85）作为测试集

第二次分类（lightGBM）：

|  |  |  |
| --- | --- | --- |
| 实际 预测 | 非欺诈 | 欺诈 |
| 非欺诈 | 实际非欺诈，预测非欺诈  的有78389条 | **实际非欺诈，预测欺诈**  **的有52296条** |
| 欺诈 | 实际欺诈，预测非欺诈  的有14条 | **实际欺诈，预测欺诈**  **的有71条** |

结论：

1. 放入**63 2432**条数据作为测试集，测试集中的欺诈订单有**93**条
2. 最终分类测结果中，预测为欺诈的有**5 2367**条，也就是拿5 2367笔订单去开启3D（占总的订单的8%）
3. **93条欺诈订单中的71条**进入了3D验证（范围占比76 %）

3.

**测试集选择：18年9月6日-18年10月30日**

第一次分类（决策树）：

|  |  |  |
| --- | --- | --- |
| 实际 预测 | 非欺诈 | 欺诈 |
| 非欺诈 | 实际非欺诈，预测非欺诈  的有246606条 | **实际非欺诈，预测欺诈**  **的有185770条** |
| 欺诈 | 实际欺诈，预测非欺诈  的有1条 | **实际欺诈，预测欺诈**  **的有55条** |

第二次分类就是拿**第一次预测出来的欺诈订单**（18 5770+55）作为测试集

第二次分类（LightGBM）：

|  |  |  |
| --- | --- | --- |
| 实际 预测 | 非欺诈 | 欺诈 |
| 非欺诈 | 实际非欺诈，预测非欺诈  的有109980条 | **实际非欺诈，预测欺诈**  **的有75790条** |
| 欺诈 | 实际欺诈，预测非欺诈  的有9条 | **实际欺诈，预测欺诈**  **的有46条** |

结论：

1. 放入**43 2432**条数据作为测试集，测试集中的欺诈订单有**56**条
2. 最终分类结果中，预测为欺诈的有**75836**条，也就是拿75386笔订单去开启3D（占总的订单的17%）
3. **56条欺诈订单中的46条**进入了3D验证（范围占比82 %）