# 首届“桂林银行杯”数据建模大赛解决方案

1. **赛题分析与理解**

本赛题提供acct、cust、tran、cards的训练集和测试集等脱敏数据，构建模型并预测同样特征下的卡号的0~1连续数值。该赛题为二分类问题，可用机器学习、统计建模等多种方式对其进行预测。根据赛题提供的脱敏数据，可以大胆推测：四个表分别对应着账号表、客户表、交易表、卡表。进一步推测的话，预测结果可能为客户名下的银行卡风险预测或者信用卡欺诈检测。

1. **建模流程设计**

建模流程设计粗略地分为五个部分：

（1）数据观察、数据分析：对重要数据表进行可视化辅助分析。当数据量过大时，人工通常难以挖掘出数据的规律。需要使用到可视化工具分析数据规律，了解数据的全貌。

（2）数据预处理：在评估的基础上对可用的原始数据进行一系列清洗和转换，保证使用的数据尽可能赶紧，使之达到建模的需求。

（3）特征工程：数据和特征决定了机器学习的上限，而特征工程的重要性在学术界和工业界都得到充分地验证。特征工程是最大限度地从原始数据提取和组织特征以供算法和模型使用。

（4）模型方面：模型的选取以及调参对多大程度能逼近数据和特征所决定的上限是至关重要的。而好的模型融合能将差异较大的模型组织形成1+1>2的效果。

（5）线下模型评估：

建模流程如图2.1所示：

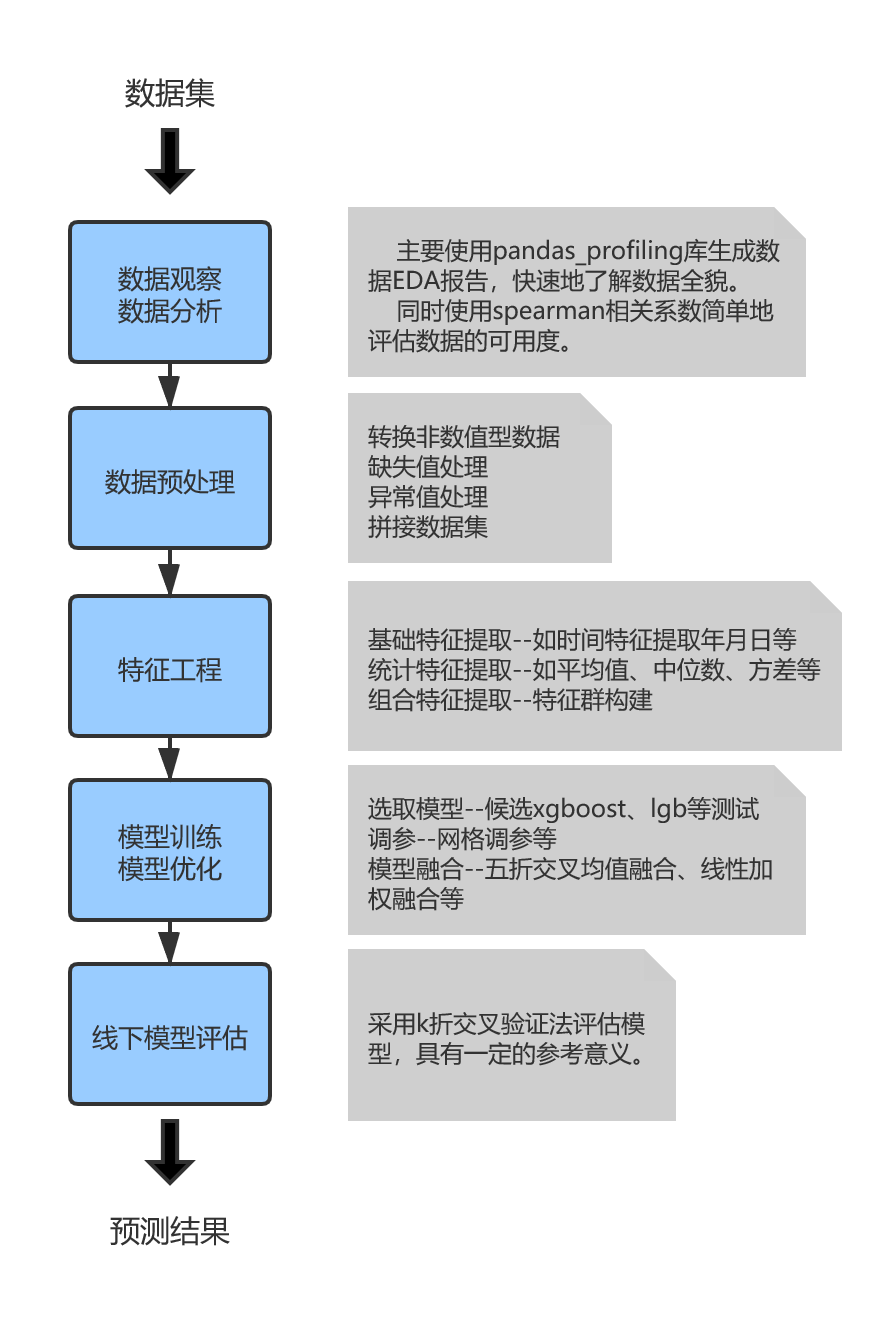


图 2.1 建模流程图

1. **特征工程**
   1. **数据分析**

赛题中提供了四个表，分别为acct、cust、tran、cards表，并依次对各表使用pandas\_profiling生成的数据EDA报告进一步的分析。

以下数据分析以acct表为例，其他表的分析也按照该流程进行，就不赘述了。

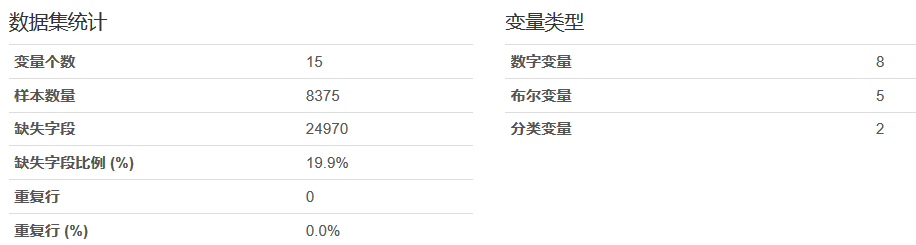


图 3.1.1 acct表概略信息

（1）表的总览分析：从图3.1.1中得知acct表的变量类型，缺失比例和重复比例，从而在数据预处理就要处理缺失值，删除重复行的操作就没有必要了。

|  |  |
| --- | --- |
|  |  |
| 图 3.1.2 acct\_x\_cat4统计信息 | 图 3.1.3 acct\_x\_cat4的长度频率 |

（2）单变量的分析：以acct\_x\_cat4为例，由图3.1.2和图3.1.2得知，cat4不仅缺失值比例高，同时还是一个高基数类别特征，由此可以判断将acct\_x\_cat4剔除。

|  |  |
| --- | --- |
|  |  |
| 图 3.1.4 acct表缺失值直方图 | 图 3.1.5 acct表Spearman's ρ图 |

（3）对相关性、缺失值等分析：通过缺失值直方图可以直观地剔除一些特征，如图3.1.4中可以直观地看到date1、date2、date4缺失值比例高，应该予以剔除。由图3.1.5Spearman's ρ图中得出，date1与label的相关性很高，衡量缺失值与相关性下，应判断保留date1，观察后续的重要性。

* 1. **数据预处理**
     1. **缺失值处理**
     2. **非数值型数据转换**
  2. **特征提取**
  3. **特征筛选**

1. **模型选取与训练**
2. **模型评估**
3. **建模过程中的发现与创新点**