关于银行信用卡用户行为的评分建模及分析

作者：艾铭忠（SA16038001）

2018 年 8 月 30 日

摘要

在当前的数据信息社会中，个人信用已经变得十分重要，其关乎着生活中的出行、消费、借贷等等各个方面。商业银行和金融借贷公司为了控制其自身的风险，往往需要分析客户办理借贷后的各项行为，进而来判断客户是否有违约风险，以此作为公司的预警系统，并未后面的政策做出指导。本文通过对某银行办理信用卡后一段时间内的个人行为数据进行分析，建立预测模型对客户未来是否可能出现逾期行为进行预测，对比了多个模型的表现，最好的模型达到了99%的准确率和0.77的AUC。

按照数据挖掘的一般化流程，本文主要分为数据预处理、建立模型两个部分。其中数据预处理部分的代码和结果可以在‘Explot_data.ipynb’ 里面找到，建立模型部分的代码可以在‘Model.ipynb’里面找到。如果需要更换数据进行测试，只需要将代码里面导入数据部分的数据文件名替换一下即可。

1 数据预处理

经过简单的分析数据，很容易得知该数据集一共有148077个案例，每个案例都有262个特征。特种可以分为8类，分别是客户基本信息、账户信息、开卡、总体交易行为、分期行为、取现行为、还款行为和消费行为。其中有9个是字符型变量，剩下的253个都是数值型变量。由于数据集已经经过了简单的缺失值处理，处理的方式可以在 ’变量说明表.xlsx’ 里面找到，因此目前没有缺失值的存在。

1.1 数据清洗

根据原始数据缺失率，我们筛选出缺失率超过50%的特征分别是 ['xac-

count_addrchange', 'xaccount_empchange', 'xaccount_CON_REL', 'xaccount_stmtnum', 'tran_cyclenum_STMT6roi', 'tran_cyclenum_STMT12roi', 'mpif_num_STMT1']，因为缺失的数据过多，已经失去了数据携带的原有信息，为了不对模型产生错误的影响，选择直接删除掉这几项特征。

由于在所有的特征中既有数值类型的数据，又有字符类型的数据，而有的模型是不支持字符特征进行训练学习的，因此需要进行特征向量化。特征向量化有整数编码和哑编码两种形式。我们先对所有的字符型数据进行了整数编码，特别的，对于一些没有明显顺序的特征，例如性别、开户行、婚姻状况等，为了避免分类器将向量化后的特征认为是有序的，可以采用One-Hot编码。

不同的特征之间数据的量纲可能是不同的，并且通过前边简单的分析数据可以看到那，不同的特征之间平均值和标准差相差很大，为了防止在分析数据的时候意外抹除掉一些有效的特征，我们对数据进行归一化（Data Nomalization），在实际的操作中，我们选择的是将特征映射到[0,1]之间的maxminmap算法。

1.2 特征工程

有一句在工业界广泛流传的话：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。为了能够选择最好的特征组合，一般我们会从两个方面来考虑特征的选取：一个是特征是否发散，如果一个特征不发散，则说明样本在这个特征上基本没有差异，那么这样的特征对于样本的区分也就没有什么用；另一个是特征与目标的相关性，相关性高的特征对于样本的区分帮助越大。

本文通过多种方法探究所有特征的重要程度，可以总体分为过滤法和Embedded方法。

（1）、过滤法中分别使用了方差选择法、相关系数法和卡方检验法

在对数据进行了maxminmap之后，剔除掉方差小于0.05的特征，剩下的特征分别是

```
Index(['xaccount_stmtnum', 'xaccount_month', 'TRAN_NOW_NUMMONRIO',
       'TRAN_STMT3_NUMMONRIO', 'TRAN_STMT6_NUMMONRIO', 'TRAN_STMT12_NUMMONRIO',
       'tran_num_STMT1roi', 'tran_num_STMT3roi', 'tran_num_STMT6roi',
       'tran_num_STMT12roi', 'tran_mpamortize_num_STMT1roi',
       'tran_mpamortize_num_STMT3roi', 'tran_mpamortize_num_STMT6roi',
       'tran_mpamortize_num_STMT12roi', 'mpif_stmt_amt_STMT3roi',
       'mpif_stmt_amt_STMT6roi', 'mpif_stmt_amt_STMT12roi', 'PAYM_NOW_BILLRIO',
       'PAYM_STMT3_BILLRIO', 'PAYM_STMT6_BILLRIO', 'PAYM_NOW_CASHRIO',
       'PAYM_NOW_AUTORIO', 'PAYM_STMT3_AUTORIO', 'bill_num_curver',
       'bill_amt_curver', 'bill_dlv_amt_STMT1billrio',
       'bill_dlv_amt_STMT3billrio', 'bill_dlv_amt_STMT6billrio',
       'bill_hcrd_amt_STMT1billrio', 'bill_hcrd_amt_STMT3billrio',
       'bill_hcrd_amt_STMT6billrio', 'bill_wsale_amt_STMT1billrio',
       'bill_wsale_amt_STMT3billrio', 'bill_wsale_amt_STMT6billrio',
       'bill_net_amt_STMT1billrio', 'bill_net_amt_STMT3billrio',
       'bill_net_amt_STMT6billrio', 'bill_num_STMT1roi', 'bill_num_STMT3roi',
       'bill_num_STMT6roi', 'bill_num_STMT12roi', 'bill_daylight_num_STMT3roi',
       'bill_daylight_num_STMT6roi', 'bill_daylight_amt_STMT3roi',
       'bill_daylight_amt_STMT6roi', 'custr_organ_0', 'custr_organ_1',
       'custr_organ_2', 'custr_organ_3', 'custr_organ_4', 'custr_organ_5',
       'custr_organ_6', 'custr_organ_7', 'custr_organ_9', 'custr_gender_0',
       'custr_gender_1', 'custr_MAR_STATUS_0', 'custr_MAR_STATUS_2',
       'xaccount_addr_type_0', 'xaccount_addr_type_2', 'xaccount_ifautopay_0',
       'xaccount_ifautopay_1', 'xaccount_CON_REL_4', 'xaccount_CON_REL_5',
       'xaccount_CON_REL_7'],
      dtype='object')
```

图 1: 方差选择法提取的特征

选择pearsonr相关系数进行计算，在所有特征中相关系数的绝对值大于0.2的特征有以下几个：

```
Index(['card_active_ratio', 'bill_dlv_num_STMT1', 'bill_dlv_num_STMT3',
       'bill_dlv_num_STMT6', 'bill_dlv_num_STMT12', 'bill_dlv_amt_STMT1',
       'bill_dlv_amt_STMT3', 'bill_dlv_amt_STMT6', 'bill_dlv_amt_STMT12',
       'bill_hcrd_amt_STMT1', 'bill_hcrd_amt_STMT3', 'bill_hcrd_amt_STMT6',
       'bill_hcrd_amt_STMT12', 'bill_wsale_amt_STMT1', 'bill_wsale_amt_STMT3',
       'bill_wsale_amt_STMT6', 'bill_wsale_amt_STMT12'],
      dtype='object')
```

图 2: pearsonr系数大于0.2的特征

卡方检验适用于离散变量与目标变量之间相关性的度量，因此我们对9个字符类型的特征变量进行了卡方检验，得到的与目标变量相关程度比较高的特征分别是['bill_amt_curver', 'custr_gender', 'custr_MAR_STATUS',

'custr_EDUCA', 'xaccount_ifautopay']。

（2）、Embedded方法

Embedded方法来提取特征有利用L1、L2惩罚项和利用树模型两种方法。我们利用梯度提升树模型，选取了特征重要性排名前50的特征非别是：

```
Index(['xaccount_CRED_LIMIT', 'TRAN_NOW_RIO', 'TRAN_STMT6_AMT', 'PAYM_NOW_AMT',
       'xaccount_month', 'PAYM_STMT3_RIO', 'TRAN_STMT12_RIO',
       'TRAN_NOW_NUMMONRIO', 'xaccount_stmtnum', 'TRAN_STMT3_RIO',
       'TRAN_STMT6_RIO', 'bill_amt_STMT1', 'TRAN_STMT3_AMTMONRIO',
       'TRAN_STMT3_NUM', 'PAYM_STMT3_NUM', 'PAYM_STMT12_RIO',
       'tran_cyclenum_STMT12', 'tran_mpamortize_num_STMT6',
       'xaccount_max_cashlamt', 'xaccount_credlimitnum', 'tran_balavg_STMT6',
       'CASH_NOW_LIMITRIO', 'bill_amt_STMT3', 'CASH_NOW_TRANRIO',
       'TRAN_NOW_NUM', 'TRAN_STMT3_OVERAMT', 'bill_amt_STMT6',
       'tran_mpamortize_num_STMT3', 'CASH_STMT3_LIMITRIO',
       'tran_mpamortize_num_STMT3roi', 'tran_amt_STMT6', 'PAYM_STMT6_RIO',
       'tran_cyclenum_STMT6', 'CASH_STMT12_TRANRIO', 'CASH_STMT12_NUM',
       'TRAN_STMT3_NUMMONRIO', 'CASH_STMT3_TRANRIO', 'CASH_STMT12_LIMITRIO',
       'mpif_stmt_amt_STMT1', 'mpif_amt_STMT1', 'card_day_diff',
       'xaccount_max_billamt', 'PAYM_STMT6_NUM', 'CASH_NOW_AMT',
       'CASH_STMT6_LIMITRIO', 'bill_net_amt_STMT6', 'CASH_STMT3_NUM',
       'xaccount_credlimitrenum', 'mpif_stmt_num_STMT6', 'mpif_amt_STMT3'],
      dtype='object')
```

图 3: 提升树模型选取的重要性排名前50的特征

通过对比可以发现，不同的方法得出的特征重要性次序是不同的，因为它们进行排序的依据不同。与此同时，削减特征的数量也并不一定能够提升模型的效果，这是因为很多被削减的特征中也包含有一些有利于分类的信息，尽管信息相对来说比较少，但是依然能够提升模型的效果。因此，特征的选取是一个需要反复进行，不断实验的过程，只有结合多种特征提取方法和模型实际效果的反复比较，才可能得出针对某类数据真正有效的特征。

2 建立模型

为了比较不同模型在同一批数据中的效果，我们将这批数据按照7:3的比例划分成训练集和测试集。我们分别建立了逻辑回归、自由森林、决策

树、GBDT、XGB和Catboost等分类模型，并且用相同的数据进行训练和测试。由于训练数据存在很大程度上的正负样本不均衡问题，为了评价模型效果的好坏，我们挑选了准确率、召回率、F1-score和ROC曲线这些指标来表征模型。

经过测试，各个模型的参数表现如下（未经过调参）：

	逻辑回归	自由森林	决策树	GBDT	XGB	Catboost
准确率	0.97	0.98	0.97	0.99	0.98	0.98
召回率	0.12	0.35	0.56	0.54	0.49	0.31
F1-score	0.2	0.51	0.54	0.67	0.64	0.44
AUC	0.56	0.68	0.77	0.77	0.74	0.65

表 1: 各分类模型对相同数据的表现

得出的ROC曲线如下（排列次序参照表1的模型顺序）：

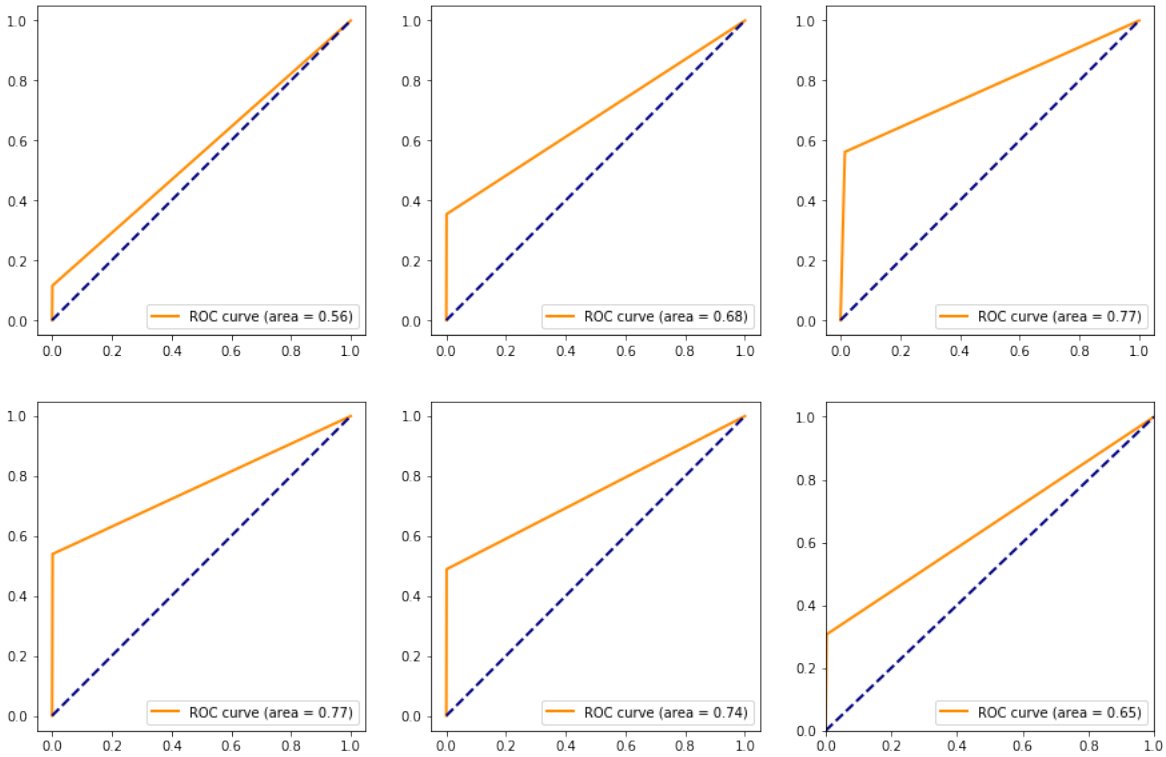

图 4: 各模型对同一批数据的ROC曲线

3 分析及优化模型

从上面的结果可以分析得知，在我们使用的6个分类模型中，准确率最高的是GBDT模型，召回率最高的是决策树模型，F1-score分数最高的是GBDT模型，AUC最高的模型有两个分别是GBDT和决策树。可以看到，在没有经过调参的情况下，表现最好的模型是GBDT，其次是决策树模型，而相对而言比较简单的逻辑回归模型的表现则显得差强人意，只比胡乱猜测（AUC=0.5）稍微好一些。

同时这些模型都是使用了全部特征并且没有调参的情况下得出的。在数据预处理一节中我们已经挑选出了很多比较重要的特征，接下来可以进行模型优化的工作可以主要从两个方面入手：

（1）、挑选在数据预处理阶段选出的特征进行模型训练；通过多项式法、对数法、指数法等方法在已有特征的基础上新增特征，进行模型的训练。

（2）、利用交叉验证、网格搜索等方法对模型的参数进行调参。

4 结论

我们通过已有的数据建立有监督的机器学习模型，通过比较各个模型的表现，发现GBDT分类模型可以在该数据集上实现较好的效果。利用训练好的模型，我们可以以很高的准确率预测客户办理借款后的行为，从而在一定程度上进行风险控制，规避公司的损失。