摘要

基于小额贷款公司商户信息预测信贷会否逾期

**陈诺**

数据科学与工程学院,华东师范大学, 上海

chennuo909@163.com

2018年小额贷款公司遭遇机构数量、从业人员数、贷款余额“三降”，在巨大竞争压力下难以存活。而减少贷款出现坏账的可能性对小额贷款公司而言意义重大。本文基于收钱吧商户注册信息、信用记录及交易流水信息，利用机器学习的一些经典模型，对商户信贷是否违约（逾期30天）进行预测。

1 Introduction（数据介绍）

1.1 业务背景

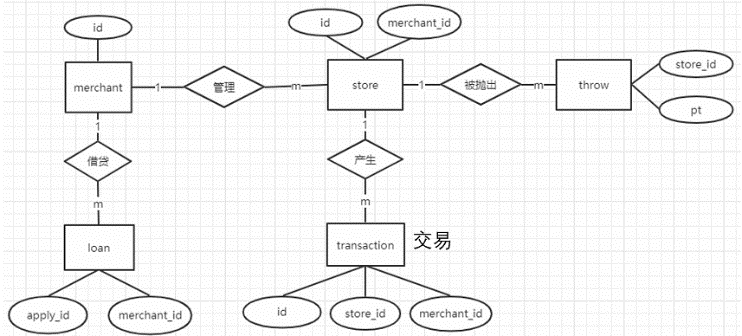
小额贷款公司自2014年始经历繁荣发展后，迎来了“寒冬期”。央行数据显示，截止2018年末，全国的小额贷款公司约有8133家，同比降幅5%；行业人员流失13149人，同比降幅13%；贷款余额降至9550亿元，同比降幅2%，遭遇“三降”[中国中央银行, 2019]。小微贷款企业竞争压力的增大，加之商家普遍存在逾期还款甚至不还款的行为，使得其必须重视商家信用，即逾期可能的掌握程度。正是因此，预测旗下商家是否会逾期对这些公司而言显得格外重要。基于该结果不仅可以减少贷款出现坏账的可能性，公司更可以调查这些商家的共性，以更好地优化自己的产品。

1.2 经验性规则

大多情况下，小额贷款公司认为逾期超过三十天的还

款视为违约[Steve Nicastro,2019]。对于逾期的预测，Donald D.Hester于1968年提出，银行准入贷款客户时，为尽量降低风险，不仅要考虑信贷数据，还需调查客户的其他信息，如日常交易信息等，以充分相信贷款的信用[Donald D.Hester,1968]。基于孟加拉村的信贷情况，Zeller和Shama提出贷款额度、贷款期限、贷款客户基本信息等因素都会影响贷款逾期时间[Zeller and Shama,1997]。而针对近年P2P经济下的小额贷款公司放出的贷款，“不良”贷款呈不断上升趋势1，我们则该着重关注那些“不良”贷款，即违约贷款，并对其进行预测。

1. 商业银行2016年末不良贷款与2015年的比较。同比YoY(year -year percent)是指本期与上年同期相比的增长率。[Wang J.et.al 2018]

1.3 数据介绍

1. 下述表格的ER图

本文基于收钱吧提供的2018.5-2018.6贷款的商户信息，贷款信息，2017.3-2018.3的历史贷款信息及2018.4-6的交易流水信息进行信贷逾期的预测，一共有如下五张表格：

* table\_datamining\_transaction 商户信息表

5599条商户ID、行业及省份信息

* table\_loan\_train 贷款信息表（用于训练 有label）

2018.5-2018.6间1200条贷款人ID、时间、金额、期数、前三十天触达次数及逾期30天次数信息

* table\_loan\_test 贷款信息表（用于测试 无label）

2018.5-2018.6间515条贷款人ID、时间、金额、期数、前三十天触达次数信息

* table\_datamining\_history\_loan 贷款历史信息表
* 2017.3-2018.3间353条贷款人ID、时间、金额、期数、理论及实际结清日及逾期30天次数信息
* table\_lend\_transaction 流水表

2018.4-6间3590059条交易类型、状态、门店商户终端ID、付款人ID、交易方式、交易金额、优惠金额、优惠渠道、交易时间等交易流水信息

2 Related Works

2.1 现有工作

现有数据挖掘方面使用硬数据的信贷风险评估的工作

大多基于银行数据，主要分为以下无个方面：判别分析[Shi X,2001]，逻辑回归[Zhang G,2005]，神经网络[Baesens B,2003],决策树[Zekic-Susac,2004]及如Bagging和Boosting的集成学习模型[Dietterich T.G,2000]，这些评估对现有P2P经济下的小额贷款公司提供的帮助有限。

2.2 本文工作及创新点

现有关于贷款逾期的工作大多为特征分析，如[Wang J.et.al 2018]等， 或基于软信息的逾期检测，如[Jiang, C.et.al 2018]等，以及上述提及的一些基于银行数据的分析。而基于硬性息的、及基于P2P经济下小额贷款公司商户数据逾期检测工作较少。2.1工作中鲜有用到如随机森林、Xgboost这样的模型进行测试，而本文中发现针对小额贷款公司的商户信息预测是否逾期的问题而言，使用上述模型比2.1中提出模型效果更优。

**Algorithm 1** XGBoost

**Input**: I ,instance set of current node.

d, feature dimension

**Output**: Split with max score

1: gain0

2:

3: **for** k=1 to m **do**

4:

5: **for** j in sorted(I,by **x**jk)

6:

7:

8:

9: **end**

10:**end**

**Algorithm 1** Random Forest

**Input**: A training set *S* := (*x*1, *y*1), . . . , (*xn*, *yn*), features *F*

**Parameter**: number of trees in forest *B and so on*

**Output**: The leingg shirned tree

1: **function** RandomForest(*S* , *F*)

2:

3: **for**  **do**

4: S(i)A bootstrap smaple from S

5:

6: **end for**

8: **return** H

9: **end function**

10: **function RandomizedTreeLearn(*S* , *F*)**

11: At eac

12: f very small subset of F

13: Split on best feature in f

14: **return** The learned tree

15: **end function**

3 Methodology

3.1 主要模型

本文主要采用两个集成学习的挖掘方法RandomForest与XGboost，并尝试对比了Adaboost，GDBT，DecisionTree，Bagging，Boosting，SVM等方法，最终准确率不如本文中使用的两个方法，伪代码如上。

3.2 模型描述

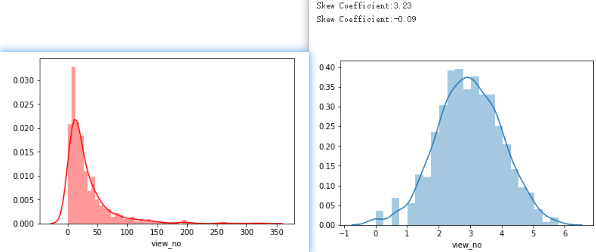
由于非集成学习的一些分类器或分类模型常常会使结果不稳定，且由于可以提取的特征较多，可能存在无关特征,对于本数据集中label不平衡（0:1的数量为7:4）问题也能得到较好的解决，所以使用集成学习的投票机制。RandomForest是一种基于Bagging的集成学习方法，XGBoost是一种基于Boosting的集成学习方法，它们的效果都可以比任意单个的分类器优异，且各有优点，故对两种方法都进行了尝试，并与单分类器进行了对比验证。

3.3 特征提取及处理

|  |  |  |
| --- | --- | --- |
| 代号 | | 变量含义 |
| X1 | 行业及省份 | |
| X2 | 某商户的贷款总额 | |
| X3 | 贷款期数 | |
| X4 | 触达次数 | |
| X5 | 历史总贷款额 | |
| X6 | 历史总贷款期数 | |
| X7 | 平均理论-实际结清日天数(还款压力) | |
| X8 | 历史平均逾期册数 | |
| X9 | 借款前30天交易总额 | |
| X10 | 前所有记录日均交易额 | |
| X11 | 日均交易额变异系数 | |
| X12 | 日均交易笔数 | |
| X13 | 交易总额 | |
| X14 | 实收金额占交易额比例及方差(优惠度) | |
| X15 | 30天内交易次数 | |
| X16 | 总交易次数 | |
| X17 | 最近一次交易距2018.6.30的天数 | |
| X18 | 笔均交易额 | |
| X19 | 信用卡支付对应30天内笔数 | |
| X20 | 储蓄卡支付对应30天内笔数 | |
| X21 | 微信余额支付对应30天内笔数 | |
| X22 | 支付宝余额支付对应30天内笔数 | |
| X23 | 花呗支付对应30天内笔数 | |
| X24 | 成功交易占比 | |
| X25 | 子门店数 | |
| X26 | 终端个数 | |
| X27 | 该商户平均折扣情况(参不参与折扣) | |
| X28 | 付款人id占所有付款人的比例 | |
|  |  | |

Table 1: Table example

由于篇幅限制，仅列出部分提出的特征。其中，X1-X4为第一类特征，是从商户信息表和贷款信息中提取出，X5-X8为第二类特征，从历史贷款信息钟提取出，X9-X18为第三类特征，属于流水中较重要的特征，X19-X23为第四类特征，反应了支付方式，X24-X28为第五类特征，反映了商户体现在流水中的一些固定的情况。

3.3.1 特征筛选

这些特征中，经过分析，有些特征不可被考虑。

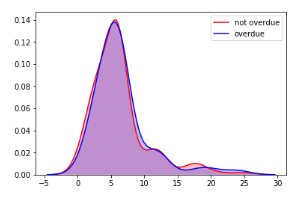
* 作弊

由于该模型会被用于实际商户信贷的过程中，一旦出现该信贷（此时贷款金额和期数并没有确定） ，只有1个月内的流水数据是值得分析的，超出一个月的数据根据经验与该笔贷款相关度不高，并且一旦考虑需浪费很多算力，故X2,X3,X13,X16特征被剔除。

* 重复、相关度大于95%

经相似度检验，X12与X16完全一致，其实际只差了一个日期，故X12被剔除。

* 与是否逾期关系不大

具体体现为，将该特征在逾期情况下雨不逾期情况下的分布图作出，结果明显重叠，这样的特征需要删去，这些特征有X63,X7,X8,究其原因，可能是因为历史数据偏少（缺失80%店铺的数据）且与贷款较为久远，与逾期与否影响不大。

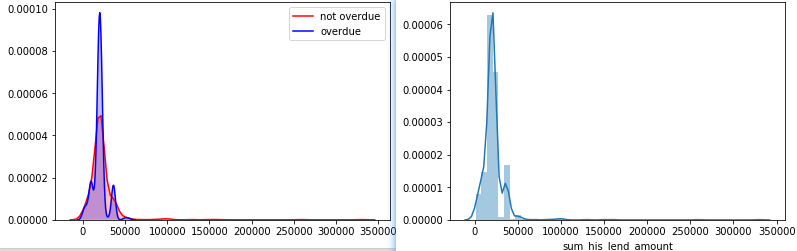
1. 历史贷款期数在逾期与不逾期下的分布图

3.3.1 特征预处理

**Categorical data**

对于如行业，省份这样的categorical data，使用one-hot处理，得到的分量l\_1,l\_2,l\_3,p\_1,p\_2,p\_3作为新的特征。

**缺失值处理**

表格中缺失值最多的为历史数据，仅353行数据，训练集中有82%的店铺没有任何记录。一种合理的方式是将训练集分为两部分，有历史记录的按有历史记录的训练，没有的按没有的训练。而个人选择了缺失值填补，具体方法是将未缺失特征的商铺所属行业及省份将某特征分为不同的集合，再用缺失数据所在集合的平均值来进行补充，这样大致确保了补充商户数据的合理性（即”物以类聚”），最后检查该特征在填充后的整体分布情况是否受到影响。如历史总贷款额，在把没有历史数据的商家的情况下，逾期的明显集中在历史总贷款额小的部分，而不逾期的则相对较分散，而补充缺失值后分布特征不变4，可见有效性。

1. 历史贷款总额在缺失值补充前后的分布

**偏态数据**

对于偏态的数据，如果不进行处理，基分类器会难以对ordinal data进行分段，故对这样的特征需要作Log Transformation处理。如下图中触达次数，经过处理后协方差从3.23降至了-0.09333Fig 4历史贷款总额在缺失值补充前后的分布5。

1. 触达次数Log Transformation前后分布对比

上图说明不同industry\_level1和province的历史贷款额是不同的。 所以，可以根据所属行业及省份将贷款额分为不同的集合，再用缺失数据所在集合的平均值来进行填充 ，并检查填充后 sum\_his\_lend\_amount ~ is\_30days\_overdue 是否受到影响。在把没有历史数据的商家防空的情况下，逾期的明显集中在历史总贷款额小的部分，而不逾期的则相对较分散，现补充缺失值的目标是补充完缺失值后分布需要尽量一致。

**类别不均衡且非对称**

本问题中，训练集与测试集中不逾期：逾期数量约为7:4，类别稍不均衡，并且由于实际中重视逾期（即label为1）的情况，故也是非对称数据，此时尝试使用过采样的方法对训练集进行处理。（欠采样也有尝试，但是效果反而差了，因为训练集本身数据就相对不足，再欠采样不合适）本文采用SMOTE进行过采样，算法如下。

**Algorithm 3 SMOTE**

1: **Start**

2: **for** i <-1 to10 (k-nearest neighbors for 10)

3: Compute k-nearest neighbors, and save the indices in the number of attribute

4: **end for**

5: **while**

6: Choose a random number between 1 and k, call it nn. Choose one of the k-nearest neighbors of 10.

7: j <-1 tonumber of attribute

8: dif = MinorityClassSample(attribute(nn)(i)) -MinorityClass Sample[i][j]

9: gap = rand() // between 0 and 1

10: NewClassSample[newindex][j] = MinorityClassSample[i][j] +gap \*dif

11: **end for**

12: newidex ++

13: **end while**

14: **End**

**归一化**

对于之后的logistic等分类模型，需要对数据归一化，消除量纲，公式如下。

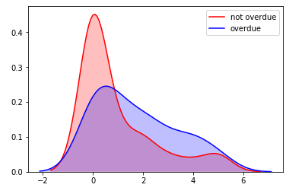
4 Result and Discussion

4.1 特征分析

由于篇幅限制，之后仅选取部分特征进行分析。

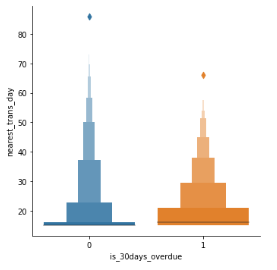
4.1.1 与label分析

**历史平均逾期次数：**

对逾期的商家来说，历史逾期次数相对较高而且平稳，对不逾期的商家，历史平均逾期次数集中在0附近，由此可见该数据集的小微贷款公司的客户相对稳定，同一商家的逾期状况也相对平稳6。

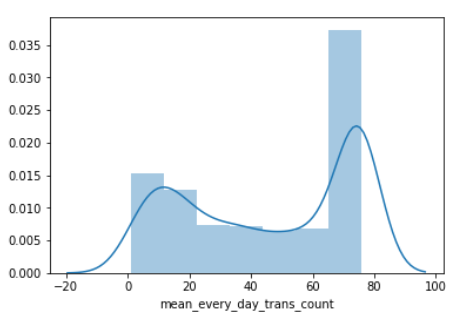
1. 历史平均逾期次数在是否逾期下的分布

**最近一次交易距某一固定时刻天数：**

****由增强箱型图可以看出，逾期商户的最近一次交易相对距该固定时刻(真实问题中可能是当天日期)较近，即说明这些店铺交易量相对较大，而不逾期商户较多当天有交易，除了当天有交易的情况外，天数间隔相对较长，即反映了不逾期的商户大多是每天有稳定交易的大商户，或是交易量很小的小商户7。

1. 最近一次交易距某一固定时刻天数的增强箱型图

4.1.2 组内分析

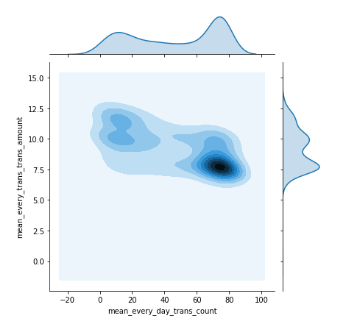
**30天内日均交易笔数：**

1. 日均交易笔数分布图

由日均交易笔数可以看出，该数据集的公司下大多公司日均交易笔数在70笔左右，由此可以反映出使用贷款业务的商家的大致情况，大多为交易相对较多的小店。

4.1.3 组间分析

**一个月内日均交易笔数与笔均交易额：**

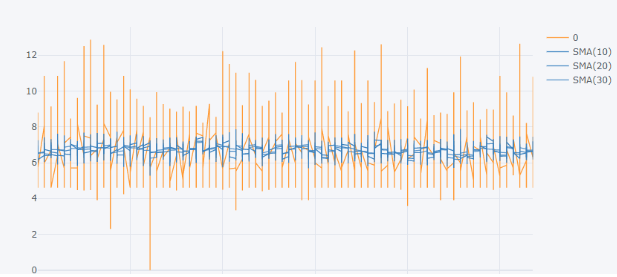
结合上述分析，由下图9可见，在30天内日均交易笔数在70笔左右占大多数商户，而这些商户的笔均交易额大约在7块钱左右，这也证实了之前大多使用贷款业务的商家为小店的推断。

1. 一个月内日均log交易额与笔均交易额的核密度图

4.1.4 补充：时间序列分析

**日均交易最多的店铺随时间log日交易额变化趋势的sma：**

SMA：前简单移动平均（simple moving average，SMA）是某变数之前n个数值的未作加权算术平均。

当有新数值加入，同时剔除旧数据，因此无需每次逐数据累加：

1. 某店铺随时间log日交易额变化趋势的sma

由于只有近一个月的数据，故连续性较差，所以不考虑时间序列。

4.2 训练及优化模型

4.2.1 交叉验证

在训练模型时，为了得到泛化误差小的学习器，一般用 测试误差来近似泛化误差，并使用交叉验证来降低泛化误差。

根据经验，由于样本大小相对正常，取K=10。

**Algorithm 3 K-fold cross validation**

**Input:** the possible value of parameter λ

**Goal:** find the best parameter λ

1: **Start**

2: **for** each k=1,2…,K:

3: fit the model with parameter λ to the other K-1 parts, giving and compute its error in predicting the kth part:

3:

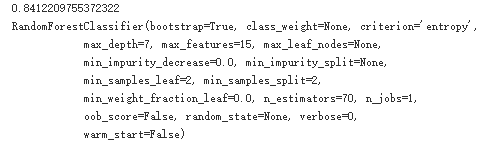
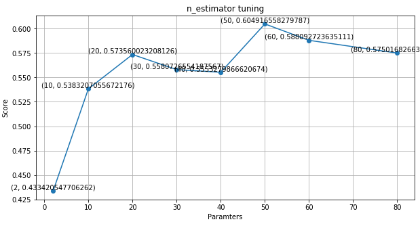
4:**end for**

5: CV(λ)=

6:choose the value of λ that make CV(λ) the smallest

7: **End**

4.2.2网格调参

网格调参，即对当前模型影响最大的参数调优，直到最优化；再拿下一个影响最大的参数调优，如此下去，直到所有的参数调整完毕（以遍历笛卡尔积）。GridSerachCV可以利用交叉验证进行网格调参，即输入一些参数λ，并自动进行上述算法第6步以给出最优参数组合。由于参数间也存在依赖关系，故该方法比调试单个参数以达到最优效果更好。

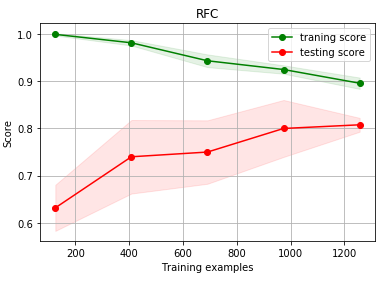
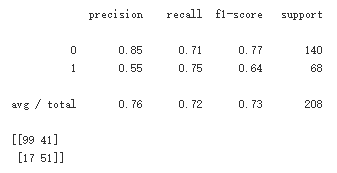
1. 逐个调参法比网格调参最优参数结果与评分

由上图11可以看出，在训练集上RandomForest使用逐

个调参的best n\_estimator为50，而网格调参结果为70，并且f1\_score为0.84,远由于逐个调参60%。

但是网格调参易于陷入局部最优，也容易过拟合（如上述f1\_score过高），故尝试绘制学习曲线，以观察过拟合的程度，也可根据模型一般的参数范围进一步确定合适的调参范围。

4.2.3学习曲线

学习曲线包括训练误差（或精度）随样例数目的变化曲线与测试误差（或精度）随样例数目的变化曲线，是判断过拟合的一种方式，同时可以判断学习器的表现。

1. 随机森林的学习曲线

观察学习曲线12，训练精度随样例数目增加而减小，测试精度则增加，过拟合程度降低。并且从学习曲线的变化趋势看，测试精度将随着训练样例的数目的增加而进一步增加。经过观察，基学习器决策树的深度以及的数目起主要的作用。

4.2.4评价标准与损失函数

由于label非对称，我们更关注于label为1，即逾期的情况，而label也存在不平衡的问题（初步由过采样解决），故选择采用label1的

f1score作为评价标准，f1score计算方式如下：

，其中P和R分别为 precision 和 recall

而由于大多模型都存在过拟合的现象，在调参时设置损失函数的正则化项，缩小特征系数，如在XGBoost中，将reg\_lambda（L2正则化）从1设为1.2，可以有效提升正确率。

通过指定的损失函数，再根据以上方法，确定随机森林网格调参范围

{'max\_depth': [4, 5, 6, 7], 'n\_estimators': [50, 60, 70], 'max\_features': [12,15,18], 'criterion': ['gini', 'entropy'], 'min\_samples\_split': [2,3,4], 'min\_samples\_leaf': [2,3,4]},

进而确定最优参数

max\_depth=7, max\_features=15, min\_samples\_leaf=2, min\_samples\_split=2, n\_estimators=70，'criterion': 'entropy'

而xgboost的范围为

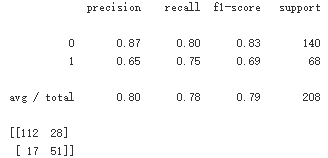
param\_grid={'n\_estimators': [150, 200, 250], 'max\_depth': [2, 3, 4], 'colsample\_bytree': [0.3, 0.4, 0.5]},

进而确定最优参数

colsample\_bytree=0.4, max\_depth=4, , n\_estimators=250

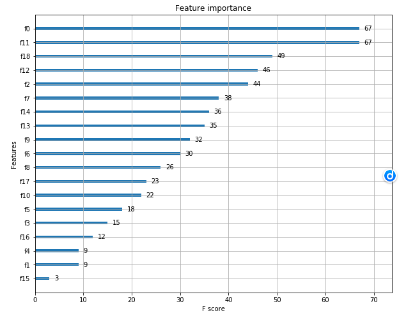
4.3 模型结果

为检测训练效果，将训练集分为80%的子训练集和20%的子测试集，对80%的子训练集使用网格调参（交叉验证）训练出最优模型，并将其应用至子测试集中。在子训练集上，RandomForest的准确率为0.84，子测试集准确率为0.64，classificationreport如Fig13所示，学习曲线如Fig12所示，ROC\_AUC分数为0.73。XGBoost 子训练集准确率为0.84，子测试集为0.69，有关图片如Fig14所示，ROC\_AUC分数为0.77。

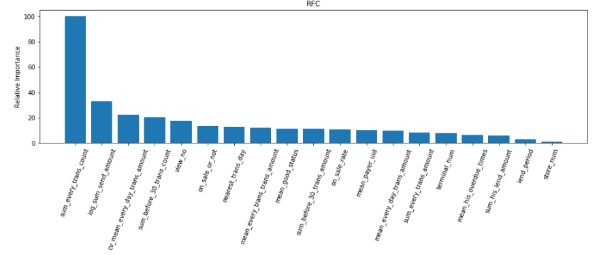
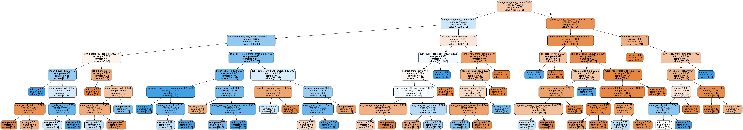
1. RandomForest ClassificationReport
2. XGBoost ClassificationReport

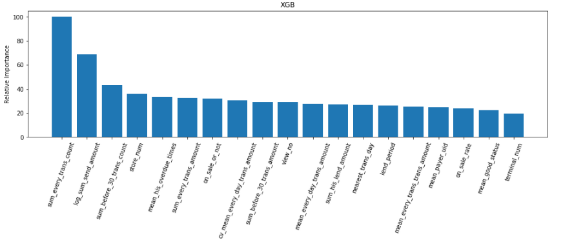
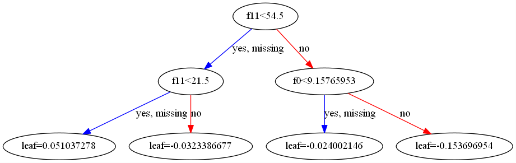
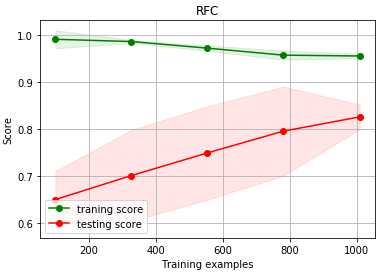
从学习曲线可以看出XGBoost训练集准确率过高，表现出明显的过拟合

4.5 特征评分与可视化

由于几乎所有模型特征评分中行业、省份及第三类特征（优惠）得分极低，故考虑将其删除，删除后最终分数几乎不变。

1. 未删除上述特征时的XGBoost特征评分

Randomforest特征评分及最优模型基分类器如Fig16所示

1. Randomforest特征评分及最优模型基分类器
2. XGBoost特征评分及最优模型

于模型调参性能而言，RandomForest运行速度较慢，而XGBoost经过对GDBT的改良，运行速度有显著一升，于结果上看XGBoost稍逊一筹，其对每个特征的评分相对均匀。

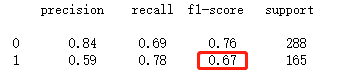
4.5 其余模型的对比

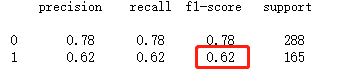
使用上述检测训练效果的方式，得出((训练集,测试集))

* DecisionTree: (0.81,0.62)
* Adaboost: (0.80,0.63)
* SVM:(0.76,0.69)kernel=linear过拟合程度最低
* Bagging: (0.80,0.66)
* GDBT(boosting): (0.80,0.59)

由此可见所有模型都存在过拟合的情况。

4.6 测试集分数

以下为真实测试集的两个模型的ClassificationReport，模型的过拟合问题并未得到很好的解决。

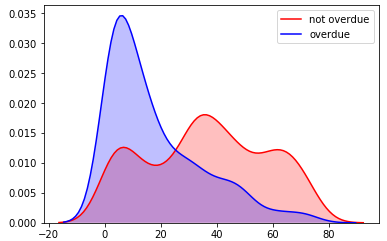
1. RandomForest 测试集ClassificationReport
2. XGBoost测试集ClassificationReport

4.6 结论及分析

虽然不同模型的正确率稍有偏差，但多个模型的特征评分都都表现出贷款前30天内日均交易笔数sum\_every\_trans\_count为第一特征，前30天内交易总额log\_sum\_send\_amount为第二特征，如图Fig16，17。

故小微贷款公司需重视门下商户短期日均交易笔数及交易总额的变化，并可以将真实逾期的这部分商户的这两特征找出，以寻找是否有短期内这两特征的共性，分覅方法于4.1节中已有介绍。如Fig20中，明显可以看出30天内交易笔数减少的店铺更易逾期，则这些商户需要被重点考虑，其可能遇到了生意淡期或发展瓶颈。同样，小微贷款公司亦可及时针对预测出即将逾期的商户使用相应措施，包括更新自己的产品，或者发放一定奖励或采取补助措施等，以及时减少出现坏账的可能性。

再者，小微贷款公司需针对这部分商家的共同需求，以更新自己的相应产品，来更好满足以及挽留、激励这部分商户贷款。相反，对于这部分商户，也可对其标上信用标记，并对即将逾期及预测即将逾期的商户进行警告，为做到及时止损，也可暂时设置进一步贷款的屏障，以免更大的损失。

若小额贷款公司使用上述随机森林模型，可以做到逾期30天f1score约大于2/3的预测成功概率。

1. 贷款前30天日均交易笔数在是否逾期下的分布

References

[中国中央银行, 2019]中国中央银行,小型贷款公司数据统计,<https://baijiahao.baidu.com/s?id=1627754882363596910&wfr=spider&for=pc>,2019

[Steve Nicastro,2019] Steve Nicastro, What Happens If I Default on a Personal Loan?,<https://www.nerdwallet.com/blog/loans/default-personal-loan/>,2019

[Donald D.Hester,1968] Donald D.Hester ,Financial Disintermediation and Policy.*Journal of Money Credit & Banking*,1969,1(3):600-617

[Zeller and Shama,1997]Repayment performance in group based credit programs in Bangladesh, *An empirical analysis,*1997,25(10):1731-1742

[Jiang, C.et.al 2018]Jiang, C., Wang, Z., Wang, R. et al. Ann Oper Res (2018) 266: 511. <https://xs.scihub.ltd/https://doi.org/10.1007/s10479-017-2668-z>

[Wang J.et.al,2018]Wang J., Zhang L., Shen P., Zhu G., Zhang Y. (2018) Preprocessing and Feature Extraction Methods for Microfinance Overdue Data. In: Xu Z., Gao X., Miao Q., Zhang Y., Bu J. (eds) Big Data. Big Data 2018. *Communications in Computer and Information Science*, vol 945. Springer, Singapore

[Shi X,2001]Shi, X., Zou, X.: The application of canonical discriminate analysis in credit risk evaluation of enterprise. Study Financ. Econ.27(10), 53–57 (2001)

[Zhang G,2005]Zhang, G., Liu, S.: Empirical study of credit risk evaluation in China’s commercial banks. J. Hebei Univ. Econ. Trade26(4), 41–45 (2005)

[Baesens B,2003]Baesens, B.: Using neural network rule extraction and decision tables for credit-risk evaluation. Manag. Sci.49(3), 312–329 (2003)

[Zekic-Susac,2004]Zekic-Susac, M., Sarlija, N., Bensic, M.: Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models. In: *International Conference on Information Technology Interfaces*, vol. 1, pp. 265–270. IEEE (2004)

[Dietterich T.G,2000]Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000).https://doi.org/10.1007/3-540-45014-9\_1