个贷违约预测

雅俗共赏

翟亚雷  
交通运输工程&研三  
上海工程技术大学  
中国-上海  
2542174006@qq[.com](mailto:email@email.com)

团队简介

雅俗共赏团队只有一个成员翟亚雷。

翟亚雷，现就读于上海工程技术大学城市轨道交通学院交通运输工程专业，研究方向是能量回收。

凭借对数据科学的热爱，2020年开始相继学习了《机器学习》、《深度学习》等相关课程。2021年陆续参加各大平台举办的数据挖掘竞赛，其中有【科大讯飞】举办的《线下商店销量预测挑战赛》、《广告点击率预估》、《恶意软件分类》、《移动设备用户年龄和性别预测》；阿里云【天池】举办的《工业蒸汽量预测》、《O2O优惠券》、《二手车交易价格预测》等10余场，成绩均在前列。

摘要

随着银行信贷用户的不断扩展，利用银行现有信贷行为数据进行迁移学习辅助创建目标业务风控模型来服务新场景、新客群变得至关重要。本文提出了利用CatBoost过滤历史信贷行为数据的数据过滤方法和利用K-means聚类构建特征的特征构建方法。通过CatBoost对历史信贷行为数据的过滤，筛选出59633条精确数据，通过K-means聚类对public表构建了重要特征sub\_class，加入目标编码及交叉特征后，组成74633条样本、56个特征，建立XGBoost算法模型，最终B榜得分0.91216。

关键词

迁移学习，CatBoost，数据过滤，K-means聚类

0 背景

为进一步促进金融普惠的推广落地，金融机构需要服务许多新的客群。银行作为对风险控制要求很高的行业[1]，因为缺乏对新客群的了解，对新的细分客群的风控处理往往成为金融普惠的重要阻碍。如何利用银行现有信贷行为数据来服务新场景、新客群成了一个很有价值的研究方向，利用迁移学习捕捉不同业务中用户基本信息与违约行为之间的关联，帮助实现对新业务的用户违约预测是其中一个重要方法。

1 AUC简介

AUC[2]是用来评价模型分类器分类性能的指标，与混淆矩阵和ROC[3]曲线有着密切的关系。

1.1 混淆矩阵

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例(true positive)、假正例(false positive)、真反例(true negative)、假反例(false negative)四种情形，令TP、FP、TN、FN分别表示其对应的样例数。分类结果的“混淆矩阵”(confusion matrix)如表1所示。

表1：混淆矩阵

|  |  |  |
| --- | --- | --- |
|  | 预测值=1 | 预测值=0 |
| 真实值=1 | TP | FN |
| 真实值=0 | FP | TN |

1.2 ROC曲线与AUC

ROC曲线是以真正例率(True Positive Rate,简称TPR)为纵轴，假正例率(False Positive Rate,简称FPR)为横轴所绘制的一条曲线，其中真正例率(TPR)表示真实值里面预测正确的量和所有真实值的量的比值，其定义如公式1所示。假正例率(FPR)表示虚假值中预测错误的量和所有虚假值的量的比值，其定义如公式2所。

 (1)

 (2)

ROC曲线示意图如图1所示，曲线越靠近左上方，表示模型分类性能越好。AUC表示ROC曲线下方的面积，即ROC曲线在[0,1]上的积分。AUC取值范围为[0,1]，AUC为1表示分类预测效果绝对准确。

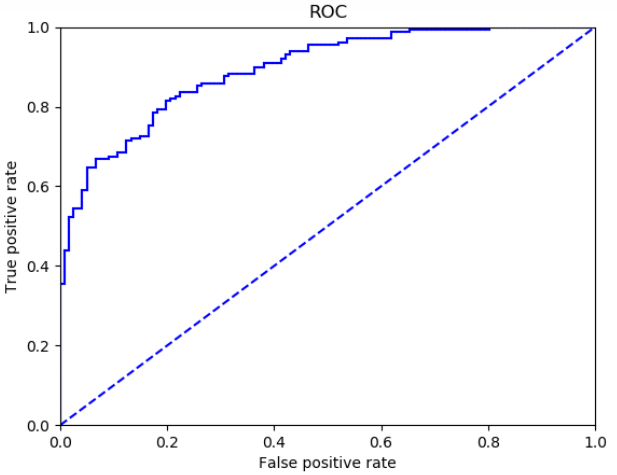


图1：ROC曲线图

2 数据预处理

数据预处理部分主要做了缺失值处理、文本类型特征处理、日期类型特征处理以及异常值处理，原始数据经过预处理之后可以直接放入模型进行训练。

2.1 缺失值处理

public表和internet表缺失部分大致相同，internet表缺失情况如图2所示：

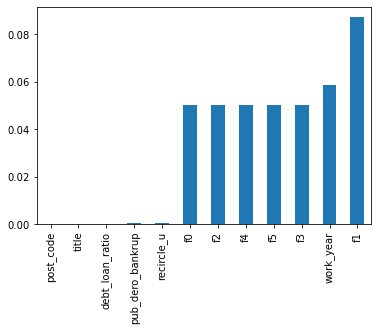


图2：train\_internet表缺失率

对于缺失数据做了以下处理：

(1) 对f1~f5和recircle\_u特征用均值填充。

(2) 对特征pub\_dero\_bankrup用前一个值进行了填充。

(3) 对debt\_loan\_ratio、title、post\_code存在少量缺失值的特征，将缺失的样本删除。

2.2文本类别型特征处理

文本分类型特征有work\_year、class、work\_type、employer\_type、industry，其类别数如表2所示，对文本类别型特征做了标签编码(LabelEncoder)处理。

表2：文本类别型特征含义及种类

|  |  |  |
| --- | --- | --- |
| 特征 | 特征含义 | 种类 |
| work\_year | 就业年限(年) | 11 |
| class | 网络贷款等级 | 7 |
| work\_type | 工作类型 | 5 |
| employer\_type | 所在公司类型 | 6 |
| industry | 工作领域 | 14 |

2.3 日期类型特征处理

日期类型特征有issue\_date和earlies\_credit\_mon。其中对issue\_date提取了年份、月份以及所在的周。对特征earlies\_credit\_mon首先转化为标准的年月日，再提取年份和月份。

2.4 异常特征处理

通过箱线图可以发现，debt\_loan\_ratio、house\_exist等特征存在着异常离群点，如下图3、图4所示：

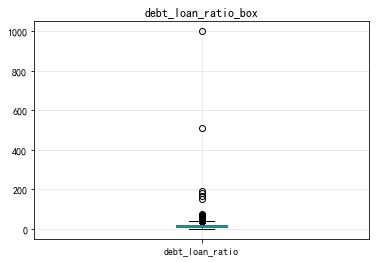


图3：debt\_loan\_ratio箱线图

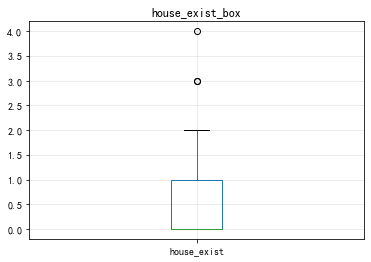


图4：house\_exist箱线图

对该类特征，以一定阈值删除了异常值所在的样本，具体如表3所示：

表3：异常特征删除阈值

|  |  |
| --- | --- |
| 异常特征 | 删除阈值 |
| debt\_loan\_ratio | <= 43.34 |
| house\_exist | <= 2 |
| total\_loan | <= 38000 |

3 双表分析及整合

如表4所示，根据官方提供的baseline，将internet全量数据加入public表中，用LightGBM[4]进行预测，AUC得分为0.8025；若使用public单表进行预测AUC达到了0.8763。参考<python机器学习专栏>提供的baseline，在internet表中筛选出比较精确的数据加入public表，此时AUC得分达到了0.891。

表4：双表数据结合方式效果对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Baseline来源 | train\_internet表加入量 | 数据量 | 模型 | AUC |
| 官方 | 全量加入 | 76w | Lgb | 0.8025 |
| 官方 | 不加入 | 1w | Lgb | 0.8763 |
| <python机器学习专栏> | 阈值小于0.5的加入 | 63w | Lgb | 0.891 |

由此可以看出：

(1) public单表的预测效果远比internet单表的预测效果要好。

(2) internet表虽然含有大量数据，但是由于含有很多不准确数据，全量加入public表中会导致预测结果降低。

得出结论：合理的利用好双表的数据以及特征，以合适的方式结合两表就应该能提升模型的预测准确度。

3.1 双表特征分析

public表和internet表不是完全重合，两表特征对比情况如表5所示。

表5：双表不同特征对比

|  |  |
| --- | --- |
| public表独有的特征 | internet表独有的特征 |
| known\_outstanding\_loan | marriage |
| app\_type | offsprings |
| known\_dero | house\_loan\_status |
|  | sub\_class |
|  | f5 |
|  | work\_type |

由表可知，public表中缺失了6个特征，且其中有4个是多分类特征。查阅个贷相关资料[5]可知，不同贷款等级之间偿还债务能力不同。等级越高，说明偿还债务的能力越强，是否还款越不容易受经济环境的影响，银行对用户越信任，违约的概率越小[6]。因此，网络贷款等级(class)及网络贷款等级之子集(sub\_class)与是否违约密切相关，而public表中缺失了特征sub\_class。考虑在public表中构建这一特征。

3.2 构建特征

由internet表可知，sub\_class有7种类别，因此，在public表中对class中的每一种类别再分为7类，这个过程采用K-means[7]聚类。

构建好特征之后，查看其特征重要性，sub\_class排到了第一，如图5所示。证实了分析的准确性以及特征构建的有效性。

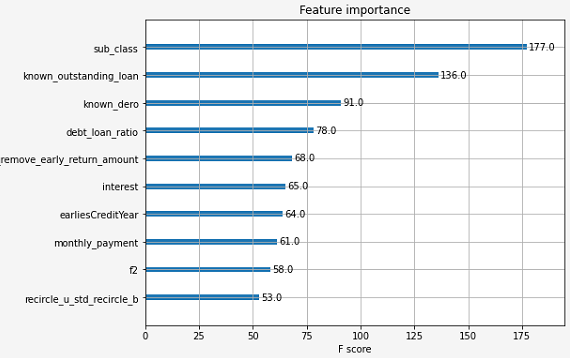


图5：特征重要性

3.3 数据过滤

为了合理的利用双表数据，将原训练集扩充到一个合理的数据量，以public为训练集、internet表作为测试集，对internet表进行预测。筛选出模型预测准确的一部分与public表合并共同构成最终训练集。

这里算法模型选用CatBoost，原因在于其可以有效解决数据不平衡时的过拟合问题，适合本数据集的特点。

数据过滤的过程中阈值的选择决定了最终使用的数据量。阈值代表违约的概率，过滤的目的是筛选出小于这个概率不违约的部分。选择不同阈值筛选出的数据量的变化如图6所示。本方案希望在阈值尽量低的情况下筛选出尽量多的数据，所以选择阈值为0.07，该阈值下过滤出59633条样本。

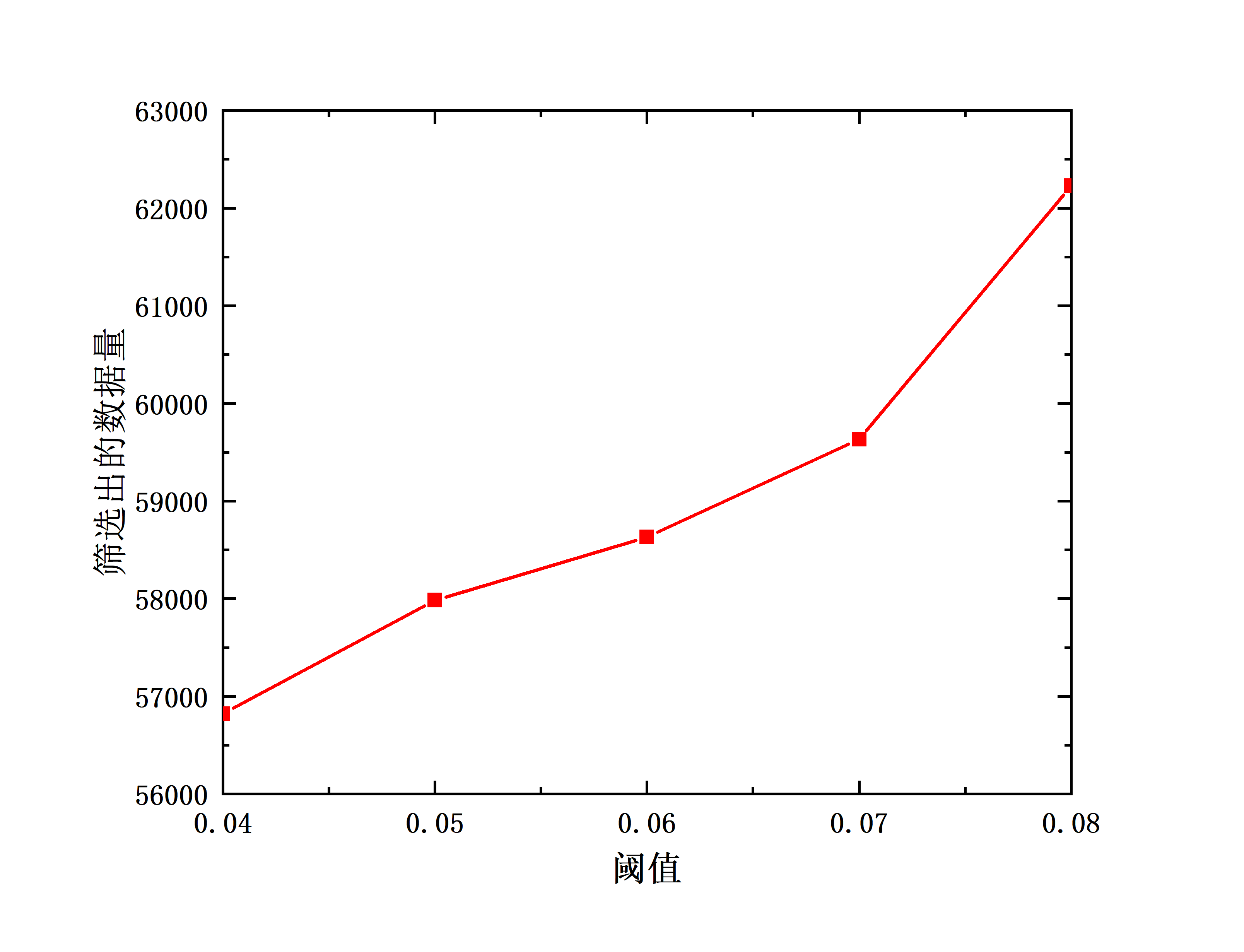


图6：不同阈值过滤出的不同数据量

4 特征工程

特征工程部分主要分为目标编码和交叉特征构造。

4.1 目标编码

目标编码是用标签的统计量对类别特征进行编码，即根据目标变量进行有监督的特征构造。编码过程中为了避免信息泄露的问题，采用K折交叉统计法，即对其中的每一份数据，采用其他K-1份数据来计算每个类别对应目标变量的统计值，即未知数据在已知的数据里面取特征。

本赛题目标编码部分依据house\_exist、debt\_loan\_ratio、industry、title四个特征统计了标签的均值，AUC得分提升了千分之一。

4.2 交叉特征

交叉特征可以拟合特征与标签之间更深层次更细粒度的关系，合理的构造交叉特征可以使模型得到更好的表达，一定程度上可以提高预测准确率。

本文构建交叉特征，依据线上提分以及特征重要性选出了以下6种组合方式，可以有效提升模型的表现力：

(1) 依据industry的不同类别，统计了f1~f4的均值。

(2) 依据post\_code、industry、employer\_type的不同类别，统计了interest的均值。

(3) 依据recircle\_u的不同类别，统计了recircle\_b的标准差。

(4) 对early\_return\_amount与early\_return做了比值。

(5) 对policy\_code做了kmeans聚类预测。

根据线上得分，交叉特征使模型预测效果提升了千分之一。

5 算法模型构建及结果预测

完整的数据和重要的特征决定了预测上限，模型只是去逼近这个上限。然而不同的模型构造原理不同，处理类别型特征的方式不同，在数据不均衡情况下对数据的处理方式不同，防止过拟合的程度不同，因此对相同数据的表现能力也不尽相同。本赛题由于数据量较小，选择常见的集成树算法(XGBoost[8]、LightGBM[9]、CatBoost[10])作为备选模型。如表6所示，LightGBM采用leaf\_wise生长策略，效率高但容易过拟合；XGBoost使用leave\_wise按层分裂的生长策略，一定程度上降低了过拟合风险；CatBoost由于其对类别型变量独特的编码方式，可以很好的解决数据不平衡时带来的过拟合问题。

表6：算法模型特点比较

|  |  |
| --- | --- |
| 模型 | 特点 |
| LightGBM | 采用leaf\_wise生长策略，效率高但容易过拟合，需要控制树的深度。 |
| XGBoost | 采用level-wise生长策略，不容易过拟合但效率低，带来了很多没必要的开销。 |
| CatBoost | 支持类别型变量，通过过拟合检测器来防止过度拟合模型，而且其特殊的编码方式能够很好的解决数据不平衡带来的过拟合问题。 |

经过线上试验，XGBoost的AUC得分最高，CatBoost居中，LightGBM得分最低。最后选择不容易过拟合的XGBoost和CatBoost作为本方案使用的算法模型，选用CatBoost进行过滤，XGBoost进行预测。CatBoost过滤线下AUC得分0.9016，XGBoost预测线下AUC得分0.9611。B榜最终得分0.9121。

6 总结与思考

本文方案使用不容易过拟合的CatBoost模型从internet表中过滤出小量而精确的数据，将训练数据扩充到一个合适的数据量，达到了模型预测所需要的数据量的基本要求。使用双表构建了重要特征，达到了模型预测所需要的特征基础。算法模型简单，不容易过拟合，能够做出较为准确的预测。

致谢

感谢CCF中国计算机协会创办本届比赛，也感谢DataFountain提供的大赛平台，感谢专家评委以及工作人员的耐心指导。感谢竞赛界的大佬们提供的baseline解决方案，没有他们的启发就不会有后来分数的提升。

参考

[1] 周翔，张文宇，江业峰.个人信贷违约预测模型的研究[J]. 辽宁科技大学学报，2020,43(03):233-230.

[2] Yao W, Li Z, Graubard B I. Estimation of ROC curve with complex survey data[J]. Statistics in Medicine, 2015, 34(8):1293-1303.

[3] Repp P, Fayzrakhmanov R and Kulikov A. The Difference Between Precision-recall and ROC Curves for Evaluating the Performance of Credit Card Fraud Detection Models[C]. 6th International Conference on Applied Innovations in IT(ICAIIT), 2018:7-22.

[4] 微软亚洲研究院《开源：LightGBM:三天内收获GitHub1000+星》[EB/OL]. https://www.sohu.com/a/123480446\_133098.

[5] 温小霓, 蔡瑞媛. 基于保险的互联网金融网贷平台风险控制[C]//陕西省保险学术优秀论文集(2013-2014). [出版者不详], 2014:236-247.

[6] 邱伟栋. 基于LightGBM模型的P2P网贷违约预测研究[D]. 江西财经大学，2020.

[7] Anil K J. Data clustering:50 years beyond K-Means[J]. Pattern Recognition Letters, 2010, 31(8):651-666.

[8] CHEN, T.&C. Guestrin. Xgboost:A scalable tree boosting system[C]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 15(6):785-794.

[9] Ke, Meng, Finley et al. Lightgbm: a highly efficient gradient boosting decision tree[C]. 31ST Annual Conference on Neural Information Processing Systems, 2017:3149-3157.

[10] Liu W, Deng K, Zhang X, et al. A Semi-Supervised Tri-Catboost Method for Driving Style Recogmition[J]. Symmetry, 2020, 12(3):336.