风险识别算法赛-项目说明

1. 团队介绍

1.1 公司介绍

1.1.1 长安新生（深圳）金融投资有限公司

长安新生（深圳）金融投资有限公司（以下简称“公司”），是一家大型消费金融服务商。为响应国家普惠金融政策号召，顺应国内消费金融发展大势，公司于2015年3月在深圳前海自贸区正式成立。

公司股东背景实力雄厚，其资产管理规模、综合理财能力、信托业务收益等主要指标均位居行业前列；核心团队均来自国内金融界，尤其是汽车金融和银行信贷等相关领域，从业经验丰富，既往成绩突出。

经过两年多的稳健发展，公司员工已近1500人，服务覆盖全国近30个省份、300多个城市，2000多个县，合作商户数千家，为信贷业务研究开发的“马达贷”业务平台已注册成为公司知识产权。公司智能化水平及科学的风险控制体系得到了金融业界的普遍认可，综合实力在行业内首屈一指，已进入国内汽车消费金融服务行业第一方阵。

迄今为止，公司已与国内10余家银行、券商等各类金融机构建立起资金合作关系，累计合作额度数百亿元。

1.1.2 青岛泛钛客科技有限公司

青岛泛钛客科技有限公司是长安新生（深圳）金融投资有限公司的全资子公司，是一家专注金融科技解决方案的高科技企业，国内最懂金融的科技服务商之一。目前，已与百度金融、长安信托、中航信托、中信证券、天风证券、大成律师事务所等一批行业领先的专业机构建立金融科技方面的合作，利用自身业务场景的深入理解和持续的科技创新能力，为金融、零售、出行等行业的用户提供专业服务，实现与业务的合作共赢。一直以来，泛钛客不仅专注于技术实践积累、完善产品服务、勇于迎接挑战，同时也在积极推动行业技术发展和大数据+人工智能决策战略。

1.2 FInSight团队

1.2.1 团队介绍

FInSight团队由天津大学系统工程研究所的一名博士、一名硕士和青岛泛钛客科技有限公司三名数据分析师组成，曾合作参与多项大数据风控项目与数据竞赛，在机器学习应用设计、金融数据挖掘上拥有丰富的经验。

竞赛经历：

- 中国平安前海征信“好信杯”迁移学习大数据算法竞赛 第6名，其中算法方案排名第3
- 第二届阿里云安全算法挑战赛 第16名
- 智慧中国杯金融算法资格赛——用户贷款风险预测 Top10%
- 融360“天机”金融风控大数据竞赛 Top10%
- Di-Tech 2016 算法大赛 Top20%

1.2.2 团队成员

姓名	单位	职务	天池ID
亢延哲	天津大学	博士一年级	David Kang

姓名	单位	职务	天池ID
李勇	天津大学	硕士一年级	MMKYZ
崔润邦	青岛泛钛客科技有限公司	数据团队主管	constantinecuii
刘胜旺	青岛泛钛客科技有限公司	数据分析实习生	fantake01
田璐璐	青岛泛钛客科技有限公司	数据分析实习生	fantaike02

2. 赛题分析

2.1 赛题与数据

2.1.1 赛题背景

某行业的重点业务目标是识别风险交易行为，如何在确保有限的审核比率下，提升风险交易的识别率，确保低打扰、高效能。

2.1.2 解题要求

参赛选手按照样式要求提供结果文件参加评测，结果文件格式要求见“结果数据文件”内容；参赛选手需要提供源代码和技术文档。编程语言不限（java,scala, python，r均可）

2.1.3 问题描述

某行业从日常交易明细中，进行抽样审核，对审核有风险的交易打上风险标识（即有风险时label字段值为1，否则label字段值为0），交易相关的输入为ID，V_Time，V1，V2，V3，…，V30。输入数据包括tran.csv和pred.csv，其中train.csv供训练使用，包含的字段如下图所示；pred.csv供预测使用。

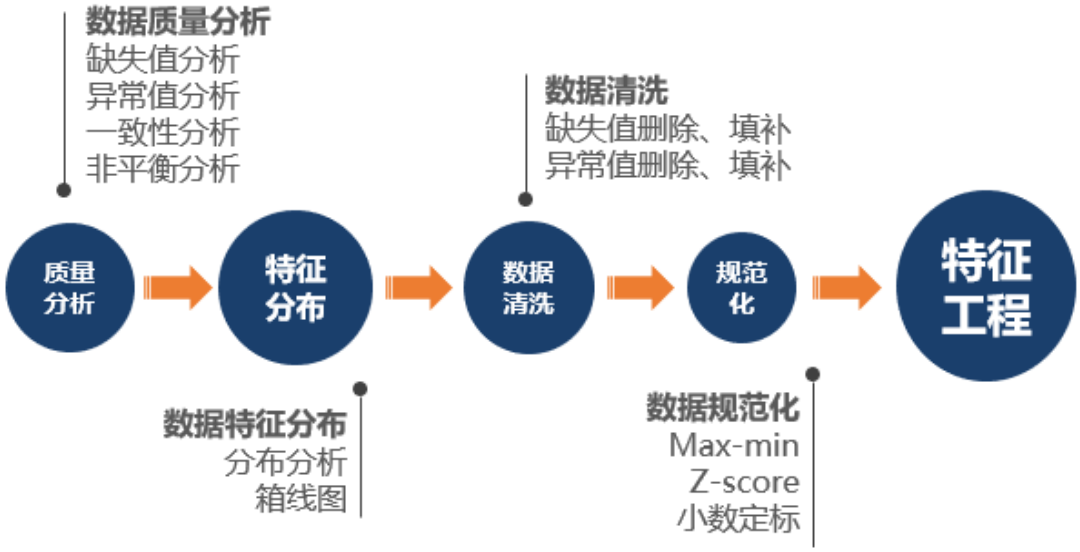
字段名	含义
ID	交易 ID 标识符
V_Time	时序
V_1，V_2，…，V_{30}	加工脱敏的交易变量
Label	Label=1 表示有风险， Label=0 表示无风险

2.2 赛题理解

通过对赛题的阅读和对数据的初步观察，可以知道该赛题要求选手设计算法方案来解决一个典型的**二分类问题**，同时二分类问题又是监督学习中一类典型问题，因此我们考虑使用经典的处理二分类问题的方法来完成对pred.csv数据集的预测。

3. 数据探索

3.1 整体流程



3.2 非平衡数据

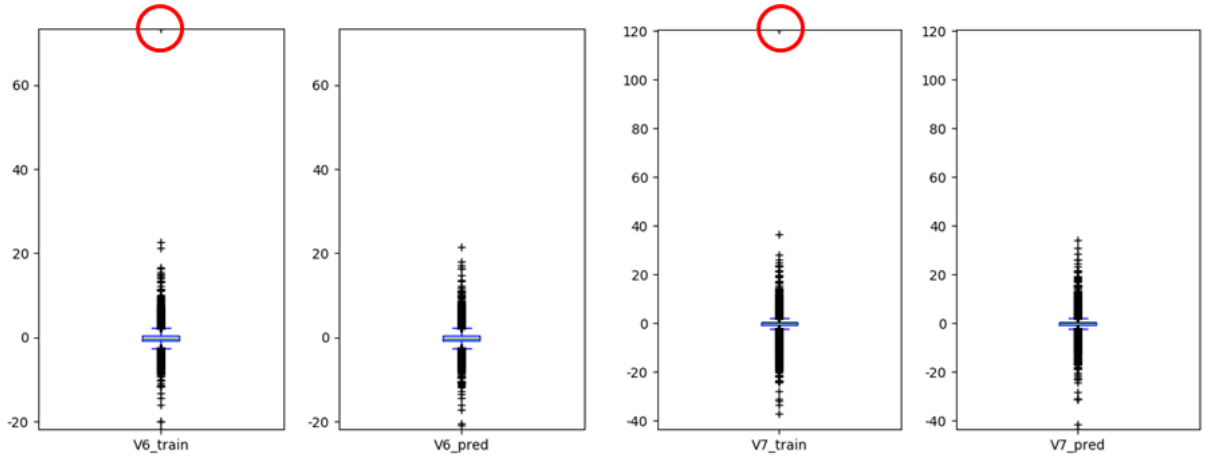
通过对train.csv的观察，发现数据具有严重的不平衡性。在训练集train.csv中，label=1的数据共计300个，而label=0的数据有99700个，因此，train.csv是一个正负类样本严重不平衡的数据集，这给模型的训练带来了很多困难。在后期的迭代测试中，我们考虑了对少数类进行过采样、对多数类进行欠采样、合成新的少数类（SMOTE算法）等方法对该数据集进行处理。

3.3 数据缺失程度

通过统计各特征的缺失值，发现train.csv各特征不存在缺失值，给数据预处理减轻了难度。

3.4 数据分布分析

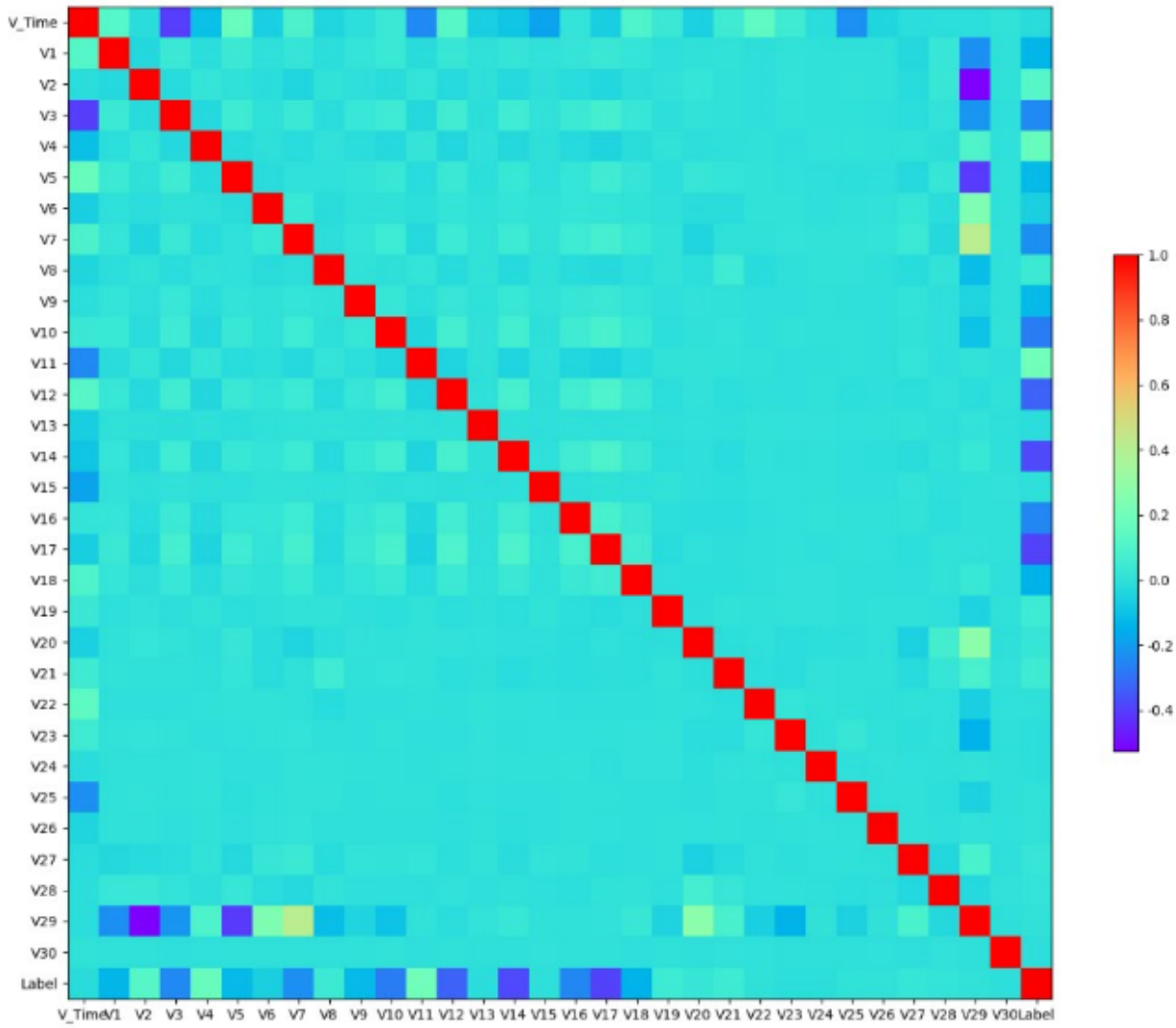
监督学习关于数据的基本假设是，输入与输出的随机变量X和Y遵循联合概率分布P(X, Y)。为了观察train.csv和pred.csv两个数据集在数据分布上的差异，我们使用箱线图来观测训练集和测试集各个特征的分布。通过观察箱线图，还可以发现异常值，对相应的数据进行删除，排除它们对模型训练的影响。



另外，还观测了训练集train.csv各特征在Label=0和Label=1两类上的分布，和训练集train.csv和测试集pred,csv在Label=0和Label=1下的分布。

3.5 互相关分析

由于赛方提供的数据集完全是脱敏数据，无法进行深入的数据探索和挖掘，因此，为了进行特征工程，我们计算了各个特征之间的互相关系数，然后选取互相关系数近似为0的重要特征做两两相乘。对于互相关性强的变量，在特征选择时考虑进行弱化。

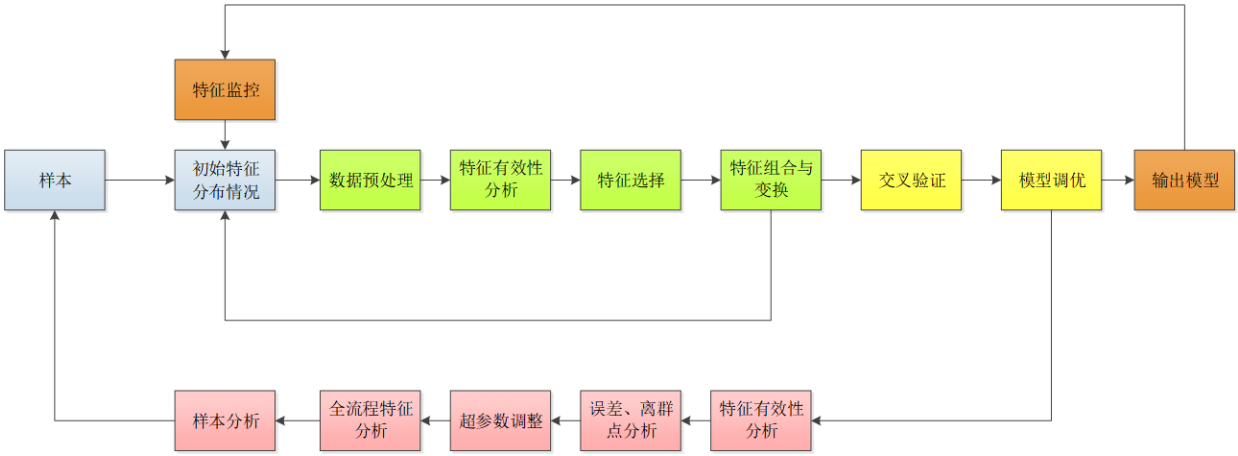


最终，加入迭代测试的新特征包括：V14×V17，V14×V4，V14×V20，V14×V7，V14×V10，V17×V4，V17×V20，V17×V7，V17×V10，V4×V20，V4×V7，V4×V10，V20×V7，V20×V10，V7×V10等等。

4. 方案设计

4.1 建模体系

FInSight团队拥有一套自己的建模体系，实现对模型的快速反馈和迭代设计。如下图所示。



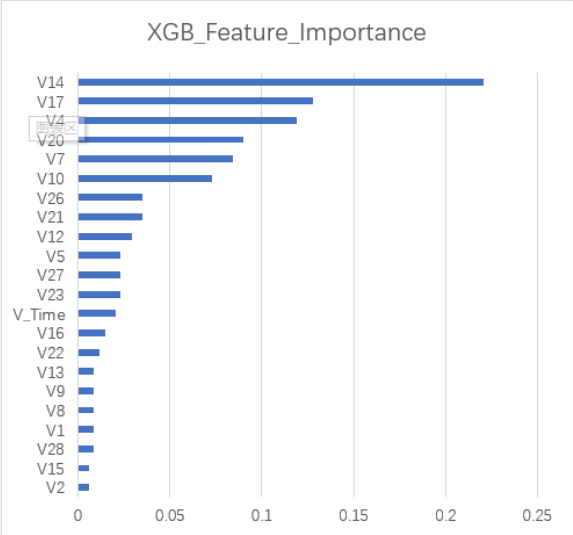
4.2 迭代测试思路总结

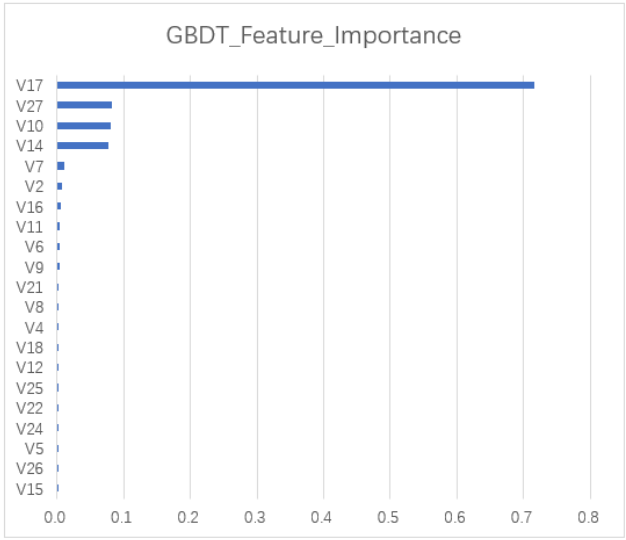
本次比赛FInSight团队自2018年1月18日开始参赛，至2018年1月31日止。在不到两周的时间内，测试过的算法思路如下：

1. 是否存在Data Leakage
2. F-score的灵敏度分析
3. 异常数据清除
4. 非平衡数据采样
5. 特征工程：统计性特征(中位数、平方、log、ln、开方、次序型、比例类、规则型)，交互型特征
6. 特征选择方法：卡方检验、GA+KNN、RF、GBDT、XGBoost、LightGBM
7. 模型训练方法：LR、SVM、KNN、RF、GBDT、XGBoost、LightGBM
8. 参数调优：网格搜索
9. 不同预测解果的比较分析

4.3 特征选择结果

通过多次测试，我们发现通过GBDT、XGBoost两种方法选择出的部分特征组合具有较强的学习潜质。特征重要性结果如下：

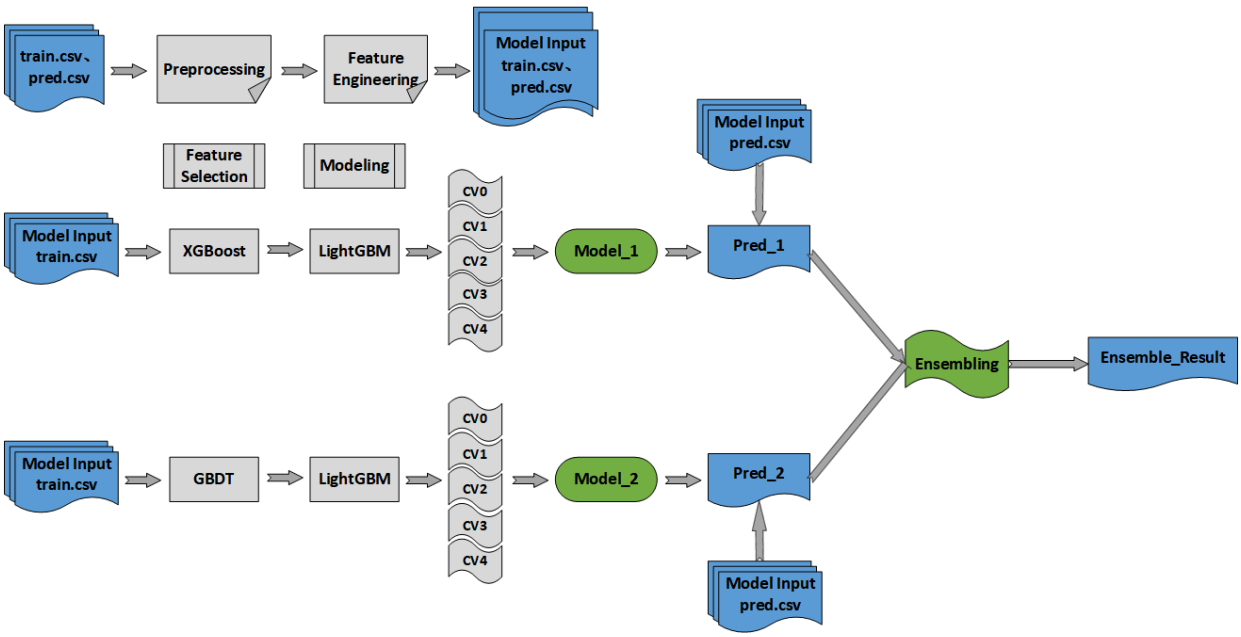




4.4 最终方案

线上f-score最高成绩0.8740对应的pred预测结果来自于对两种建模方案的融合。首先对原始的train.csv和pred.csv进行数据预处理和特征工程得到Model Input。后续方案如下：

- 方案一：基于XGBClassifier做特征选择，LGBMClassifier做模型训练，得到模型Model_1来预测pred数据集，得到label的预测结果pred_1，线上f-score为0.8760
- 方案二：基于GBDT做特征选择，LGBMClassifier做模型训练，得到模型Model_1来预测pred数据集，得到label的预测结果pred_2，线上f-score为0.8785
- 模型融合，通过对两个预测结果的对比，发现二者预测label为1的数据中，有168个的完全相同的。不同的包括：
    - pred1中ID为18650、114903和233259的三条数据； pred2中ID为18467、147575的两条数据。
    - 通过线上反馈结果可知，ID为233259和147575的两条数据是真正例，其他三条为假正例，因此对pred_1和pred_2做相应地融合，得到最终线上f-score为0.8840的结果



5. 项目代码说明

项目开发环境为Ubuntu 16.04。基于Python语言进行开发，依赖的主要开源库包括：

- Pandas https://github.com/pandas-dev/pandas
- Matplotlib https://github.com/matplotlib/matplotlib
- Scikit-learn https://github.com/scikit-learn/scikit-learn
- XGBoost https://github.com/dmlc/xgboost
- LightGBM https://github.com/Microsoft/LightGBM

5.1 code

- EDA.py
  - 探索性数据分析模块，包括前述多种数据探索的方法。
- MAIN.py
  - 主函数模块，用于运行gbdt_lgb_cv_modeling和xgb_lgb_cv_modeling两个流水线模块
- GBDT_LGB.py
  - 其中的gbdt_lgb_cv_modeling函数实现了整个机器学习建模流水线。使用GradientBoostingClassifier特征选择，然后用LGBMClassifier做训练，然后使用5折交叉验证计算mean_auc和mean_fscore，根据两个指标结果评估当前模型的性能，最后实现对train.csv的整体训练和对pred.csv的预测。
- XGB_LGB.py
  - 使用XGBClassifier特征选择，然后用LGBMClassifier做训练的流水线模块。

5.2 config

其中的chromosome.pkl文件用于存储特征选择后的特征组合。

5.3 data

用于存储加工后的train.csv和pred.csv。

5.4 eda

用于存放探索性数据分析的结果。

5.5 result

给出了线上f-score最高成绩0.8736对应的pred预测结果。