上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

上海交通大学巴黎卓越工程师学院 SJTU Paris Elite Institute of Technology

ICT Project 3 课程报告

2021 CCF 大数据与计算智能大赛-个贷违约预测项目

王雁浩

120260910034

杨浩泽

120260910043

王炜

120260910041

王甫

120260910037

2021 年 12 月 15 日

目录

1 赛题简介 **2**
1.1 赛事 2
1.2 赛题 2

2 数据介绍 **4**
2.1 数据集 4
2.2 数据探索 4
2.3 特征工程 4

3 模型算法 **9**
3.1 基础模型 9
3.2 改进策略 9

4 结果和分析 **11**
4.1 成绩 11
4.2 分析 11

5 总结 **13**

引言

本报告为 ICT Project 3 课程的小组项目报告，本组参与的比赛为 2021CCF 大数据与计算智能大赛-个贷违约预测，为金融风险控制类项目。该报告分为 5 个部分：赛题简介、数据介绍、模型算法、结果和分析、总结。该项目由王甫、王炜、王雁浩、杨浩泽四位同学共同完成。该项目的 github 地址为：https://github.com/feihongyingxia/Loan_pre。

1 赛题简介

在本节中，我们将对赛事、赛题、赛题背景和赛题任务进行简单介绍。

1.1 赛事

CCF 大数据与计算智能大赛（CCF Big Data Computing Intelligence Contest，简称 CCF BDCI）由中国计算机学会于 2013 年创办。2021 年第九届大赛以“数引创新，竞促汇智”为主题，立足余杭、面向全球，于 9 月至 12 月举办。大赛将致力于解决来自政府、企业真实场景中的痛点、难点问题，邀请全球优秀团队参与数据资源开发利用，广泛征集信息技术应用解决方案。

图 1: 2021CCF 大数据与计算智能大赛

1.2 赛题

本组选择的赛题为：个贷违约预测，该题为中原银行提供的金融风险控制类题目。

图 2: 个贷违约预测

该项目的赛题背景为：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，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。

该项目的比赛目标/赛题任务为：利用已有的与目标客群稍有差异的另一批信贷数据，辅助目标业务风控模型的创建，两者数据集之间存在大量相同的字段和极少的共同用户。此处希望大家可以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。

该题是经典的预测任务，以 AUC 作为衡量标准；使用 ROC 曲线下面积 AUC（Area Under Curve）作为评价指标。AUC 值越大，预测越准确。

2 数据介绍

在本节中，我们将介绍该项目所使用的数据集，对数据进行的数据探索以及所使用的特征工程。

2.1 数据集

该项目的数据集由赛事官方提供，数据来源为个人贷款数据和网络信贷数据脱敏仿真，2 个训练集，一个测试集。public 训练集为 10000 条个人贷款违约记录数据，internet 训练集为 750000 条某网络信用贷产品违约记录数据。测试集为 5000 条个人贷款违约记录数据。其中，public 集和测试集包含相同的 39 个字段，internet 集包含 42 个字段，public 集与 internet 集共有 9 个不同字段。关于数据集的详细信息，请查看https://www.datafountain.cn/competitions/530/datasets。

2.2 数据探索

数据探索的目的是进一步了解数据，熟悉数据，为后续的特征工程做准备。本项目的数据探索目的如下：

- 熟悉了解整个数据集的基本情况（缺失值，异常值），对数据集进行验证是否可以进行接下来的机器学习或者深度学习建模；
- 了解变量间的相互关系、变量与预测值之间的存在关系；
- 为特征工程做准备。

考虑到 public 集与测试集字段相同，数据分析以 public 集为主，并且将 39 个字段视作 39 个不同的 feature。

本项目所使用的 public 集包含类别型特征和数值型特征（使用 pandas 库中的 info() 函数查看，详见图3），其中 7 个特征包含缺失值，并且缺失特征中缺失率没有大于 50% 的特征（详见图4）。

数据集中的数值型特征又可被划分为连续变量和分类变量。本项目将连续变量的分布可视化，如图5所示。经过对比，可以观察到 public 集和 internet 集中字段相同的特征存在分布不同的情况。

通过计算数据集中的不同特征的相关系数，我们可以得到特征的相关矩阵，如图6所示，相关系数越高，说明两个特征的相关性越强。

2.3 特征工程

特征工程，是指用一系列工程化的方式从原始数据中筛选出更好的数据特征，以提升模型的训练效果。在本项目中，使用的模型为树模型，因此不再需要对数据进行归一化，标准化处理。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 39 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   loan_id                   10000 non-null  int64
 1   user_id                   10000 non-null  int64
 2   total_loan                10000 non-null  float64
 3   year_of_loan              10000 non-null  int64
 4   interest                  10000 non-null  float64
 5   monthly_payment           10000 non-null  float64
 6   class                     10000 non-null  object
 7   employer_type             10000 non-null  object
 8   industry                  10000 non-null  object
 9   work_year                 9378 non-null   object
 10  house_exist               10000 non-null  int64
 11  censor_status             10000 non-null  int64
 12  issue_date                10000 non-null  object
 13  use                       10000 non-null  int64
 14  post_code                 10000 non-null  int64
 15  region                    10000 non-null  int64
 16  debt_loan_ratio           10000 non-null  float64
 17  del_in_18month            10000 non-null  int64
 18  scoring_low               10000 non-null  float64
 19  scoring_high              10000 non-null  float64
 20  known_outstanding_loan    10000 non-null  int64
 21  known_dero                10000 non-null  int64
 22  pub_dero_bankrup          9993 non-null   float64
 23  recircle_b                10000 non-null  float64
 24  recircle_u                10000 non-null  float64
 25  initial_list_status       10000 non-null  int64
 26  app_type                  10000 non-null  int64
 27  earlies_credit_mon        10000 non-null  object
 28  title                     10000 non-null  int64
 29  policy_code               10000 non-null  int64
 30  f0                        9502 non-null   float64
 31  f1                        9142 non-null   float64
 32  f2                        9502 non-null   float64
 33  f3                        9502 non-null   float64
 34  f4                        9502 non-null   float64
 35  early_return              10000 non-null  int64
 36  early_return_amount       10000 non-null  int64
 37  early_return_amount_3mon  10000 non-null  float64
 38  isDefault                 10000 non-null  int64
dtypes: float64(15), int64(18), object(6)
memory usage: 3.0+ MB
```

图 3: 数据信息

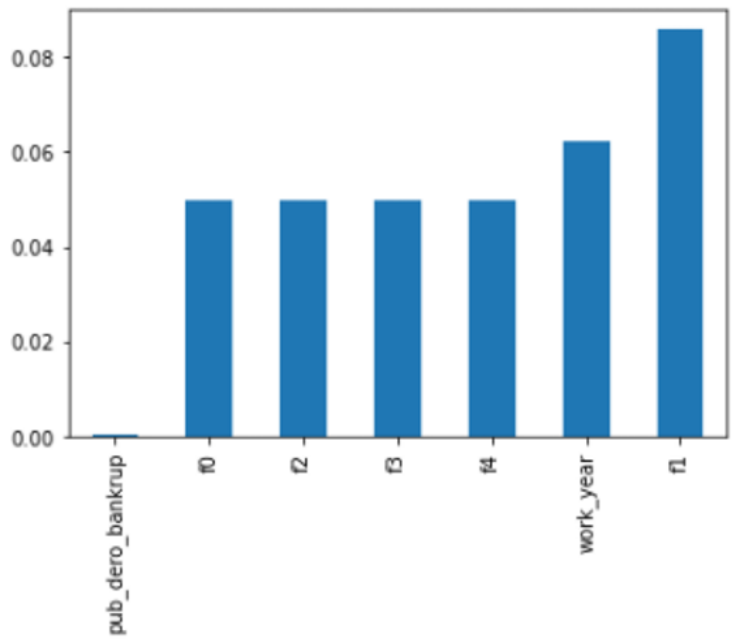


图 4: 缺失特征可视化

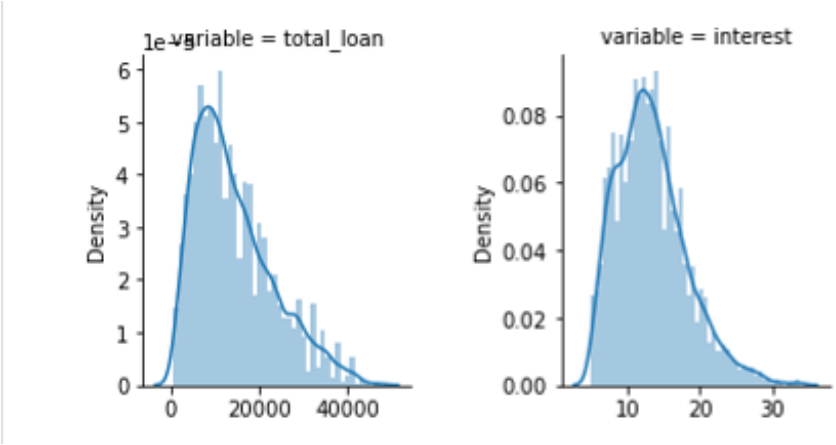
在本项目中，我们首先对数据进行了预处理，并利用一些经典的树模型，例如Catboost，进行初步的训练，在初步的训练后，如图7所示，我们得到了不同特征对模型训练的贡献度与重要性，根据初步实验的结果与所得数据，我们又对一些特征进行了更多的处理。

数据预处理包含以下步骤：

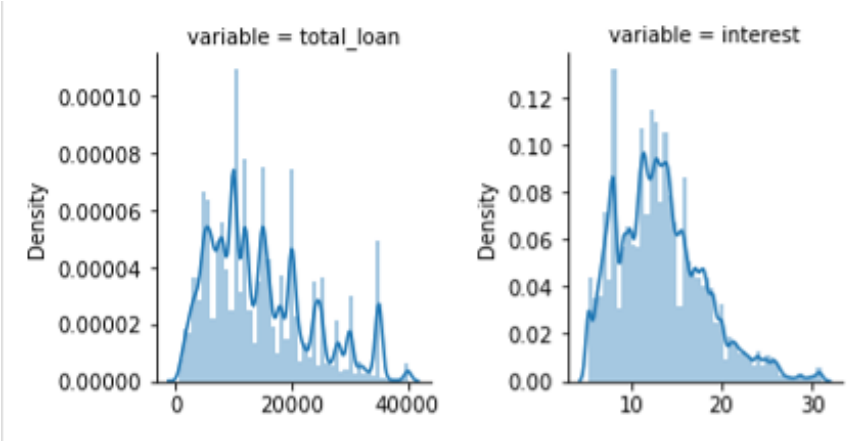
- **缺失值的填充**：经过实验，当使用列的平均值进行填充时，实验效果最好。
- **时间格式处理**：利用 pandas 库中的 to_datetime() 将时间有关特征转化为时间格式，并且计算了时间与初始时间的差值作为特征。
- **类别型特征转换到数值**：对于工作年限、贷款等级等具有优先级的特征使用自映射，对于其他数值特征，使用 sklearn 库中的 LabelEncoder 进行编码。
- **类别型特征转换到数值**：删除只包含单一值的特征。

对特征工程的改进包含以下步骤：

- **特征筛选**：根据图所示，已知存在一些相关性较高的特征，根据特征的重要性，我们删除了相关性高的部分特征与一些不重要的特征：loan_id, user_id, app_type, pub_dero_bankrup, f1, del_in_18month, initial_list_status。
- **特征计数编码**：根据特征的重要程度，我们选择重要性较强的特征对特征进行了计数编码。



(a) public 集



(b) internet 集

图 5: 连续变量分布可视化示例

- **构建交叉特征**：根据不同特征含义，我们选择一些重要的特征，构建了以下交叉特征：
  - 利息总数 pro = interest*year*year_of_loan
  - 平均每年贷款 loan_year = total_loan/year_of_loan
  - 提前还款比例 early_ratio = early_resturn_amount/total_loan
  - 年均提前还款次数 early_times_ratio = early_return/year_of_loan
  - 信贷周转余额率 recircle_ratio = recircle_bal/total_loan
  - 信贷循环额度使用 recircle_amt = recircle_util*total_loan
- **训练新特征**：在 internet 表中，存在一些不同的特征，我们选取了 4 个对贷款违约情况影响较大的特征：工作类型、房屋贷款状况、婚姻状态和子女状态。我们从 internet 表中随机采样了 10000 条数据，选取两表中相同的特征，去除分布存在明显差异的特征，最后利用 Catboost 模型进行训练，获取了 4 个新的特征。

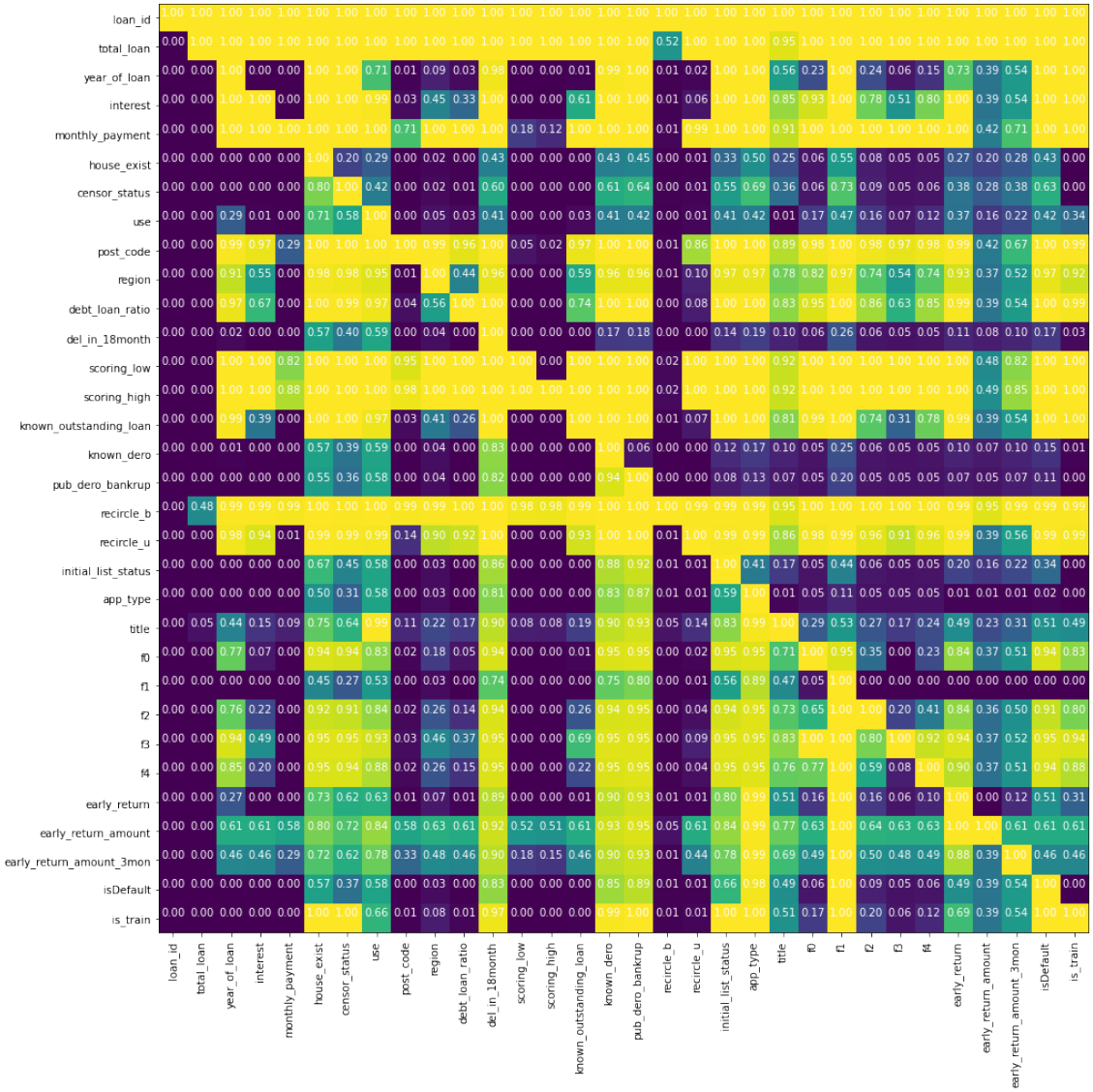

loan_id
total_loan
year_of_loan
interest
monthly_payment
house_exist
censor_status
use
post_code
region
debt_loan_ratio
del_in_18month
scoring_low
scoring_high
known_outstanding_loan
known_dero
pub_dero_bankrup
recircle_b
recircle_u
initial_list_status
app_type
title
f0
f1
f2
f3
f4
early_return
early_return_amount
early_return_amount_3mon
isDefault
is_train

图 6: 特征相关矩阵

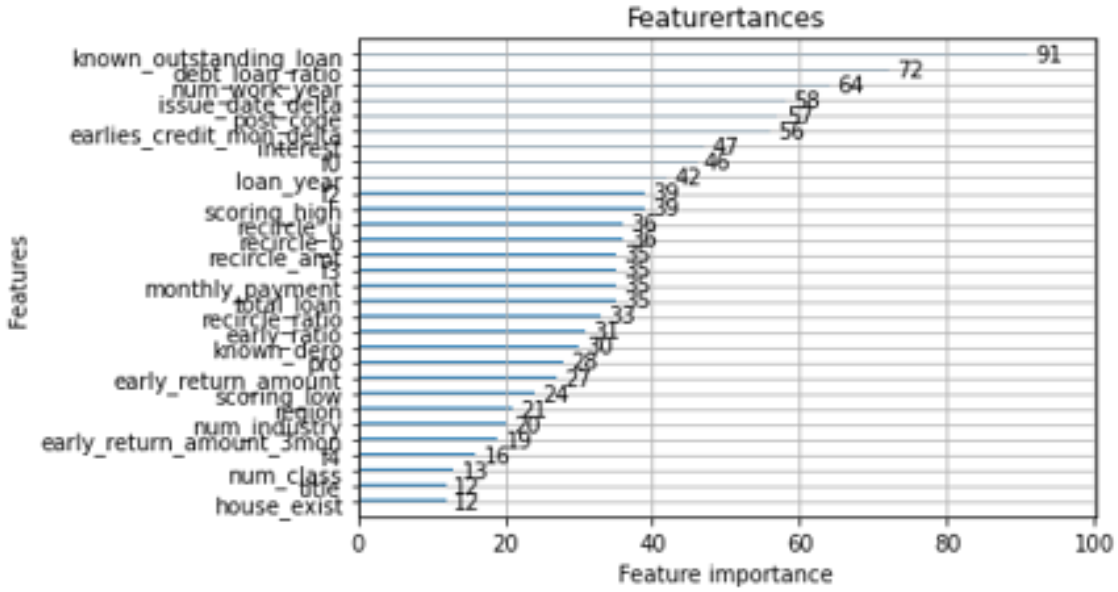


图 7: 特征重要性

3 模型算法

在本节中，我们将介绍该项目所使用的基础模型，并介绍对这些基础模型的进一步改进策略。

3.1 基础模型

考虑到比赛题目为个人贷款违约预测，该比赛对模型的可解释性要求高，并且考虑到比赛所提供的数据量与特征数均比较少，本项目选取了风控领域常用的 4 个模型作为基础模型，分别 XGBoost，Lightgbm，3 层 BP 神经网络和 Catboost。

其中 XGBoost，Lightgbm 和 Catboost 均为基于梯度提升决策树（GBDT）算法的模型。GBDT 算法是一种典型的 Boosting 算法，该算法由多棵决策树组成，将所有树的结论累加起来得到最终答案。图8简单说明了 gbdt 的训练过程，gbdt 通过多轮迭代, 每轮迭代产生一个弱分类器，每个分类器在上一轮分类器的残差基础上进行训练。

BP(back propagation) 神经网络是 1986 年由 Rumelhart 和 McClelland 为首的科学家提出的概念，是一种按照误差逆向传播算法训练的多层前馈神经网络。

3.2 改进策略

在使用了这 4 种基础模型进行训练后，我们发现 Catboost 和 Lightgbm 的效果更好，为了进一步模型效果，我们采用了交叉验证，扩充数据集，模型融合和迭代半监督的方式对模型进行改进。

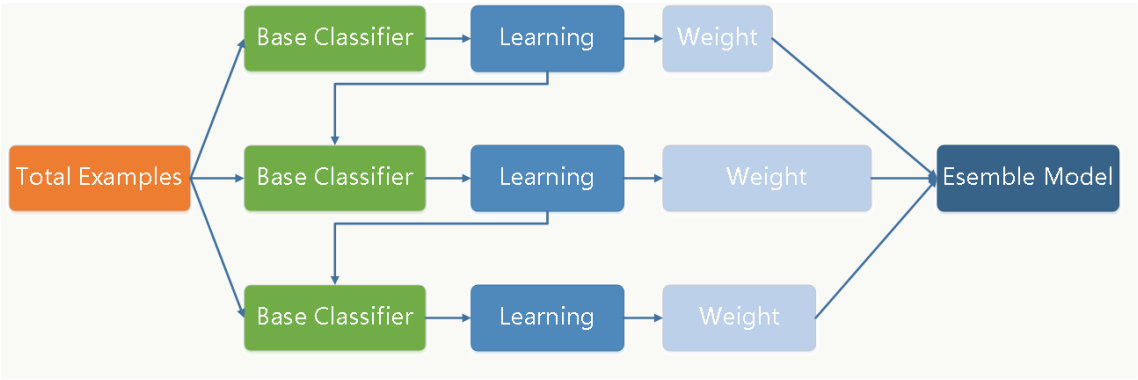


图 8: GBDT 算法原理示意图

- **交叉验证**：交叉验证主要用于防止模型过于复杂而引起的过拟合，是一种评价训练数据的数据集泛化能力的统计方法。我们使用 5 折交叉验证，将训练集随机分为 5 份每次选择 4 份作为训练集，将剩余的 1 份作为测试集，交叉验证重复 5 次，取 K 次准确率的平均值作为最终模型的评价指标。

- **扩充数据集**：赛事提供了 public 表和 internet 表两个训练集，为了更好地利用训练集，我们采取了下列方法利用 internet 表对 public 表进行扩充：

  1. 选择 Lightgbm 模型，选择 public 表作为训练集，进行训练。
  2. 利用已训练好的模型 internet 表进行预测。
  3. 将预测结果与 internte 表实际结果进行比较。
  4. 将分类置信度较高的数据添加进训练集中再次对模型进行训练。

- **模型融合**：该项目使用了平均法和学习法两种模型融合方法，平均法是指通过加权平均将不同模型的预测结果进行融合。如图9所示，学习法采用了 Stacking 方法，Stacking 先从初始数据集训练出初级学习器，然后”生成”一个新数据集用于训练次级学习器。在这个新数据集中，初级学习器的输出被当作样例输入特征，而初始样本的标记仍被当作样例标记。

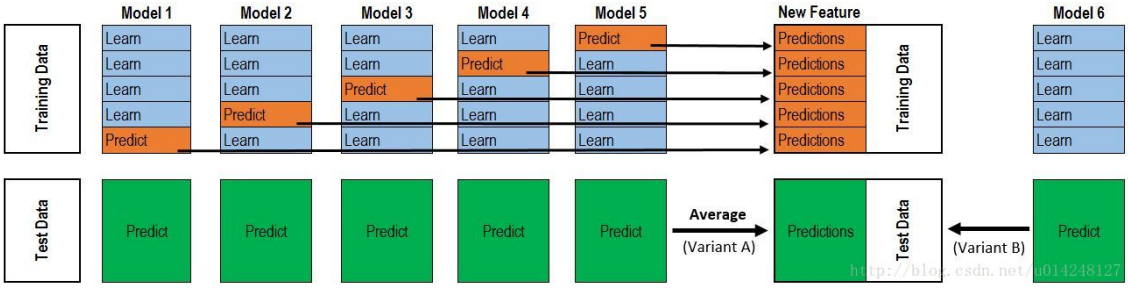


图 9: Stacking 算法原理示意图

- **迭代半监督**：迭代半监督的过程如图10所示，我们利用最好的模型预测无标签数据，并调整阈值 a，b 将测试集样本添加到训练集。

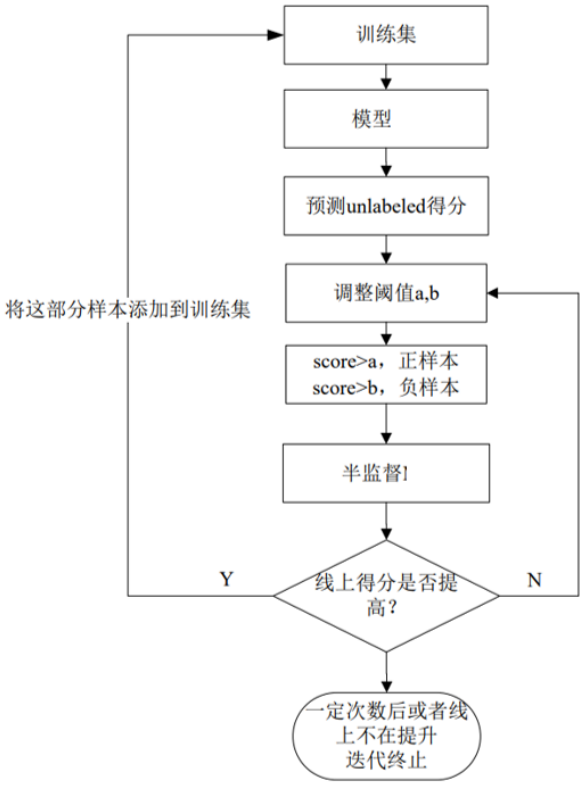


图 10: 迭代半监督方法原理示意图

4 结果和分析

4.1 成绩

通过对特征工程和模型算法不断的改进。本组的最终成绩为 0.89596998 分，名次为 211 名。

图 11: 最终成绩

4.2 分析

在实际实验的过程中，对特征工程，模型与算法的调整，会对实验结果造成不同的影响。

特征工程：赛事提供的实验数据存在数据量少，特征少等特点，这就导致对特

征的少量改动会对实验结果造成较大的影响。其中，删除不重要特征与构建新特征会使准确率提升约 0.005，而构建新特征或者删除任意 1 对相关性极高特征中的一个，就会使得实验结果直接降低约 0.01，计数编码则对实验结果无明显影响。因此，在该项目的场景下，构建新特征成为了提升模型准确率的主要方法。

模型与算法：在该项目场景下，我们尝试了多种模型与算法，并对模型参数进行了适当的改动，在基础模型中，效果最好的使 Catboost 算法，准确率能够达到 0.88050504。经过改进，准确率提升了约 0.015。在各类改进方法中，交叉验证，扩充数据集和迭代半监督的方法提升了分类的准确率，而模型融合的方法并没有起到很好的效果，图12展示了不同模型分类结果的相关矩阵，可以看到不同结果的相关性较高，造成这种结果的主要原因是：3 种基于 GBDT 的基础模型的原理相似；组内所使用的特征工程比较接近。

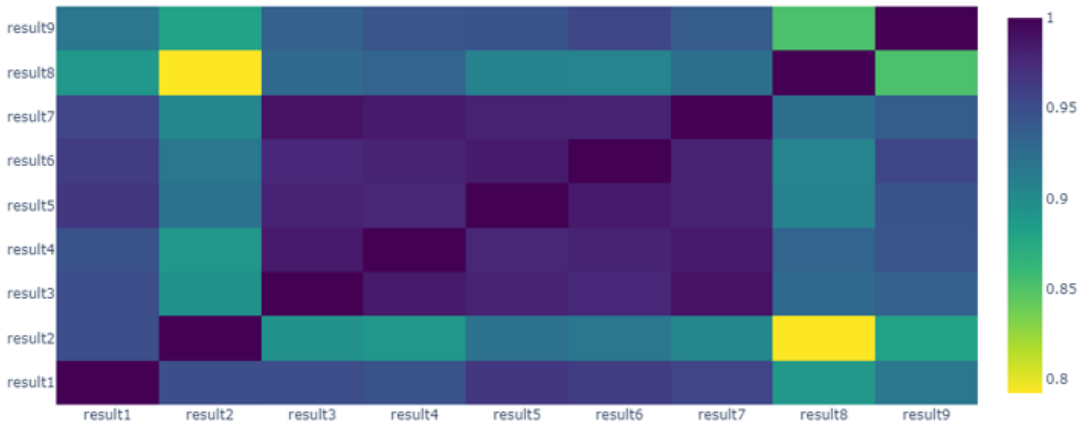


图 12: 不同模型预测结果相关矩阵

总的来说，在该项目数据量较少的场景下，对特征工程和模型的改进首先应当集中在扩充数据集方面，再进行模型结构和参数的调整。

5 总结

本次项目为本组成员第一次接触金融风控类项目，通过老师的讲解，尝试和学习，大家对特征工程有了全新的理解和更全面的掌握，能够熟练使用该领域效果最好的 Catboost 等 boosting 模型。并通过对特征工程和模型的不断改进使分数有了显著提升：分数从 0.859 最终提升至 0.896。

由衷感谢陆佳亮老师、贺阮老师和陈光烁老师在比赛过程中对我们提供的帮助和指导。