# 西南财经大学“新网银行杯”数据科学竞赛

团队名称：baseline is here

## 一、问题概述

如何运用统计和机器学习模型对客户信用风险进行预测是金融机构风险管理关注的重要问题，如在客户贷款申请审批场景中，金融机构主要依据客户的信用评分做出是否准入、额度大小、利率高低等决策，在风险量化实践中还将面临如何处理和运用高维稀疏数据、如何充分利用无标签数据、如何将样本量充足的产品上的风控模型学习经验迁移到小样本或坏样本少的产品上等问题，本次比赛将提供真实业务场景下的脱敏数据，包含多产品（客群）的高维特征数据和表现数据（部分有标签，部分无标签），邀请参赛者对数据进行探索分析，综合利用监督和半监督机器学习算法、迁移学习算法等设计区分能力高、稳定性强的信用风险预测模型。

初赛任务：预测验证集上的客户违约概率，通过大赛网页提交预测结果。

## 二、总体思路

基于本赛题，大数据金融的违约用户风险预测，本文解决方案具体包括以下步骤：

1.对客户的信息行为数据预处理操作；

2.划分训练集数据、验证集数据；

3.对客户的信息数据进行特征工程操作；

4.建立多个机器学习模型，并进行模型融合；

5.通过建立的信用风险预测模型，预测测试集上的客户违约概率。

其中，图 1 展示了信用风险预测模型解决方案的流程图。

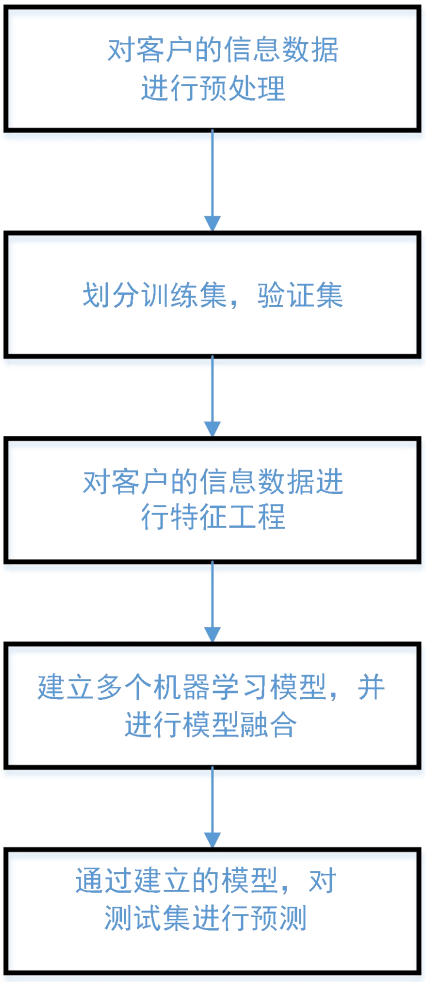



图 1 信用风险预测模型解决方案的流程图

## 三、数据清洗

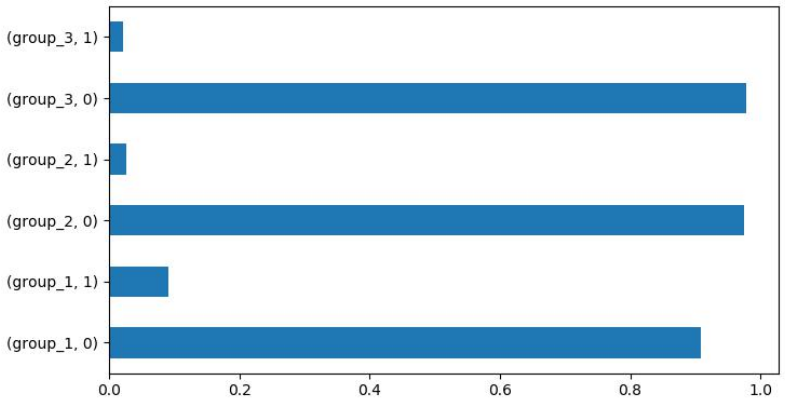
1.认识数据

在数据挖掘问题中，数据清洗是很重要的一环，特别是表格数据。首先我们要了解数据，具体来说，这次任务给出的数据集汇总如下：

| 文件名 | 用途 | 数据量 | 特征数量 | 数值变量 | 类别变量 |
|---|---|---|---|---|---|
| train_xy.csv | 训练集 | 15000 | 157 | x1-x95 | x96-x157 |
| train_x.cs | 训练集 | 10000 | 157 | 同上 | 同上 |
| test_all.csv | 测试集，线上评估 | 10000 | 157 | 同上 | 同上 |

表 1 客户信息数据集汇总表

对数据进行按组（cust_group）统计，显然数据是有偏的，其中大于 90%的用户是低风险客户，少于 10%是高风险用户。同时，客户群体 1 的低、高风险用户比例是 9:1，客户群体 2、3 的低、高用户比例大约是 98:2，这启发我们，欺诈行为在群体 1 中更容易发生，在群体 2、3 中发生概率相对较低。




| | | 数量 | 比例 |
|---|---|---|---|
| cust_group | y | | |
| group_1 | 0 | 4544 | 0.9088 |
| | 1 | 456 | 0.0912 |
| group_2 | 0 | 4871 | 0.9742 |
| | 1 | 129 | 0.0258 |
| group_3 | 0 | 4894 | 0.9788 |
| | 1 | 106 | 0.0212 |

图 2 客户群体风险情况图

2.缺失值处理

有很多的技术可以处理缺失值，简单地有随机插补，最近邻插补，稍微复杂一点的有拉

格朗日插值，蒙特卡洛插补。考虑到训练集太小了，这些或简单或复杂的插补方法都有可能造成过拟合，我们不妨将缺失值作为一个新的特征取值。

数据集中的缺失值默认是用-99（负值）进行填充的. 然而，观察数据中各个特征的分布，我们发现特征取值都是正值，并且数据缺失的情况很严重，我们没有理由将缺失值设为-99 这么一个很大的负数，这可能会人为的引入异常点，使得我们最终训练出的模型都偏向于取值为-99 的类别，造成模型泛化能力的下降.

事实上，我们只需将缺失值取为一个负值，与原有的特征取值（正值）相区分即可。不失一般性，我们将缺失值填充为-1， 实验证明，这种处理方式使得线下验证 AUC 有效地提升了 1-2 百分点。

3.关于对分类变量的处理。

赛题只告知了 x95-x157 是分类变量，却没有告知是有序变量还是无序变量，这又是一个问题。对于像{男、女}这样的无序分类变量，one-hot 奏效，对于像{高，中等，矮}这样的有序分类变量，one-hot 并不好用，而实验又证实了不处理分类变量，模型的泛化能力更好，从而，我们没有对分类变量进行特殊处理。

（1）归一化，对于树模型来说，是否归一化对模型性能的影响微乎其微，而对于像逻辑回归这样基于数值敏感的算法来说，是否归一化直接影响着模型的性能。

（2）异常点处理。对于异常点的处理，对于模型的训练是十分关键的，异常点直接影响训练出的模型的泛化能力。图 3 是通过 sklearn 中的 robust_scale+PCA 得到的，观察发现图中右侧有少量的离群点，大部分数据点集中在左侧，为了训练更好的模型，我们将最靠右的六个离群点剔除。除此之外，我们认为，一个客户的数据缺失率达到 80%以上是不正常的，应该予以剔除，进行数据统计发现这样的用户有 14 个，将这些点一并剔除。

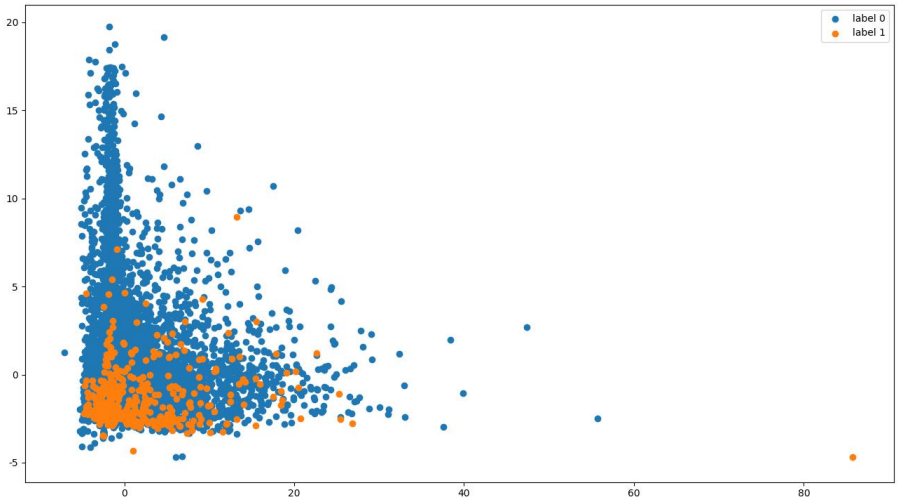



图 3 客户分布散点图

（3）类别不平衡。在风控场景中，获取到的数据常常是有偏的，即普通用户占多数类，欺诈用户占少数类。类别不平衡的问题在数据挖掘和机器学习中是一个被广泛关注的问题。现有的研究中有许多应对类别不平衡问题的技术和手段，如 SMOTE。考虑到群体 1 有更高的风险发生率，我们采取**分组 SMOTE** 的方式处理类别不平衡问题，即分别对群体 1、2、3 进行 SMOTE 采样。

（4）其他处理。观察发现，特征 x_110 和 x_112 无论在训练集 train_xy.csv 上，还是测试集 test_all.csv 均缺失，这样取值相同的特征，对于模型的训练是没有帮助的，应该予以剔除。

## 四、特征工程

在数据挖掘中，特征工程是极为重要的一环，因为“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”。特征工程的好坏取决于对特征的深入了解，特别是对于表格数据更是如此。

然而，在这一次数据挖掘比赛中，不仅数据脱敏了，顺带着连特征的含义都一同脱敏，

这样的处理给特征工程带来了极大的障碍。我们只知道特征是一堆的 x 变量，但并不知具体含义，特征有如黑箱，我们无法借助于专家意见引进先验知识以改进我们的特征质量，从而，一种可能的特征工程师：**统计分析，暴力堆叠，交叉验证。**

构造新特征：缺失率

由于数据的缺失值覆盖面积是极大的，数据缺失的原因是多样的，出于实验的目的，不妨将缺失率也作为一种新特征，加入到原特征集合中，交叉验证出一个好的模型。

构造组合特征，容易组合爆炸。

## 五、模型训练和选择

逻辑回归（以下会简称 lr）作为一种简洁有效的分类算法，在业界被广泛使用，虽然我们不认为单一模型的逻辑回归能在排行榜上取得不错的成绩，但是作为一种简单有效的方法，逻辑回归值得我们去尝试，做最后的模型融合。

Xgboost(XGB)是比赛的“大杀器”，特别对于表格数据。XGB 基于梯度提升树（GBDT），在原始的 GBDT 基础上做出了改良，训练速度很快，精度也很高。

Lightgbm 是微软推出的一款新的 GBDT 库，对于 XGB 进行了一系列的改进，在某些时候训练速度和精度会更快一些。

Catboost 是更加新颖的一种训练 GBDT 的方法，对于分类变量有更好的处理。

其他常见的分类模型如 SVM，NB，DT 等在实验过程中都存在着这样或那样的问题，从而，我们弃用这些模型。

我们的模型训练策略是先通过调参调出泛化能力最好的单模型，然后对这些模型进行融合。泛化能力的选择使用交叉验证法，对于 15000 条训练数据，使用十折交叉验证法，即使用 13500 条数据作为训练集，1500 条数据作为验证集。原理图如下：

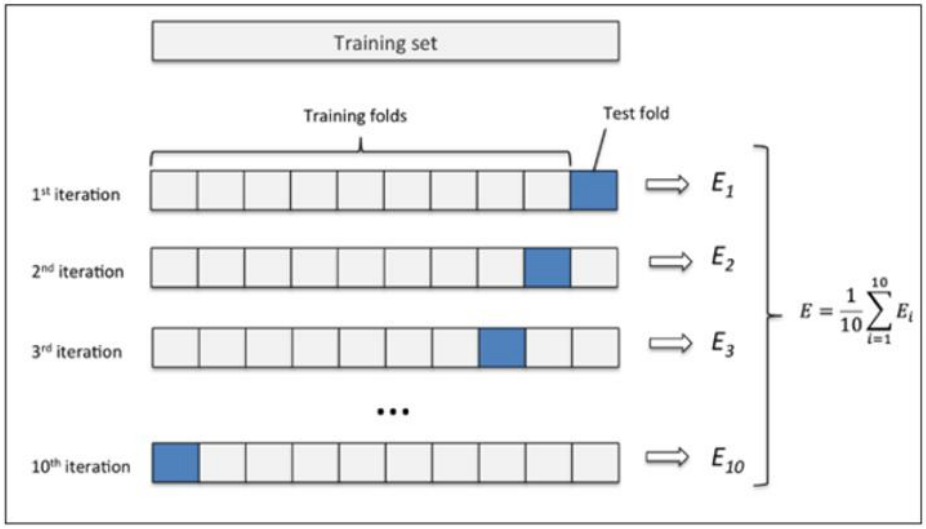



图 4 交叉验证示意图

通过交叉验证后，对单模型分别进行 voting 和 stacking 融合，得到两个融合模型，如图 5 和图 6 所示。这两个模型就是我们最终 B 榜提交的两个文件对应的模型。

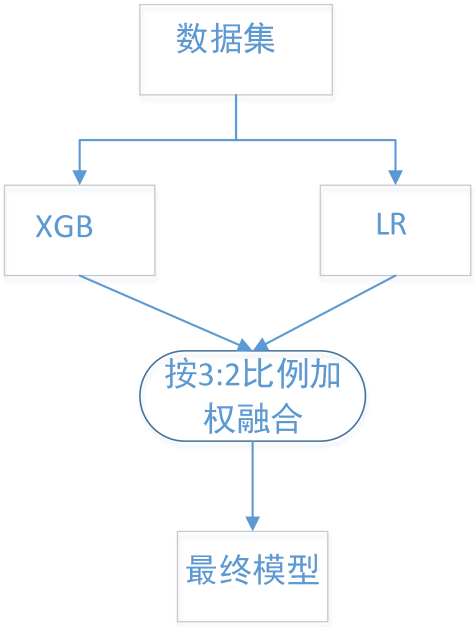



图 5 融合模型-1

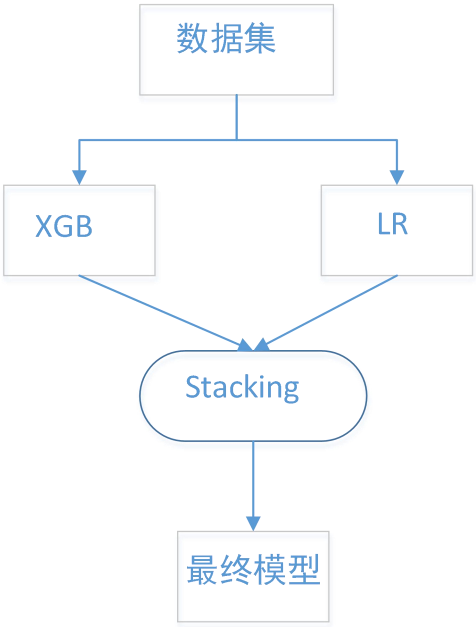



图 6 融合模型-2

## 六、创新点

1. 特征

本文将将缺失值统一替换为-1，以达到减小缺失值惩罚力度。使得训练模型具有更好的拟合性。

2. 模型

模型的创新点主要体现在模型融合上。考察指标为 AUC，融合效果很好。本文提交了两个分类器融合的方案，实际上，在线下训练中我们尝试了 3 个分类器融合达到了线上最佳的效果，在初赛最终提交时，考虑到模型的复杂性和 B 榜变化带来的风险，所以我们保守上交了相对简单模型的预测结果。可能 3 个分类器融合的模型能使预测精度更进一步。

## 七、结论与改进

- 事实证明，将缺失值从-99 改成-1 对线上提分是很有帮助的。
- 为了防范过拟合的风险，我们只是简单地融合了两个单模型。
- 由于时间等不可抗因素，很多想法没有来得及试验。
- 如果能告知特征的具体含义是什么，特征工程还有很多的东西可以挖掘。