上海财经大学

Shanghai University of Finance and Economics

期 末 论 文

题目：基于集成学习的个人贷款违约预测

课程名称 机器学习

任科教师 李文东

姓　　名 马靖淳

院　　系 统计与管理学院

专　　业 数据科学与大数据技术

学　　号 2020111235

基于集成学习的个人贷款违约预测

摘要

个人贷款违约问题是许多银行以及小型借贷机构都会面临的问题，如何收集过往用户数据并进行分析决策，从而最小化贷款风险显得至关重要。随着机器学习领域的快速发展，机器学习算法的引入成功帮助银行及其他机构成功预测出许多违约客户，并且准确率较高。本文聚焦于个人贷款违约数据，基于集成学习算法在处理贷款违约问题预测精度的优越性，使用了 LightGBM、XGBoost 以及 CatBoost 三种集成学习算法进行预测分析，并对三个模型进行融合获得了优于三种模型单独分析的结果，最终通过三种方法的特征重要性分析提出对银行等小型贷款机构的建议。

关键词：集成学习，个人贷款违约，LightGBM，XGBoost，CatBoost

一、引言

(一) 研究背景

经济增长是一个国家得以持续发展的重要力量，能够避免因经济下行或经济停滞所带来的各种社会稳定问题。自 20 世纪 80 年代以来，为了维持国家经济的高速健康发展，许多发达国家盲目追求高增长，推动国民进行超前消费解决自身经济问题，地方政府以及企业都存在过高负债的情况，债务膨胀逐渐变得无序化，从而进入了 21 世纪的债务时代。聚焦我国，截至 2022 年上半年，我国居民的存款达到了 112.83 万亿元；如果按照最新人口普查、全国 14.12 亿人来进行计算，则中国人均存款约为 7.99 万元左右。

在国民负债如此高的情况下，相应产生许多信贷违约问题，对于许多银行以及小型借贷机构来说，如何收集过往贷款人相关数据并进行分析决策，从而最小化贷款风险显得至关重要。随着机器学习领域的快速发展，机器学习算法的引入成功帮助银行及其他机构成功预测出许多违约客户，机器学习算法的引入成功帮助银行及其他机构成功预测出许多违约客户，并且准确率较高。对于商业银行来说，这一部分的工作显得至关重要。

(二) 文献综述

针对信贷违约这一问题，国内许多学者对此进行了相关分析。

刘鹏、张小乐（2020）利用 Logistic 回归模型分析了银行贷款违约数据，研究发现逻辑回归具有不错的分类能力。柯孔林（2009）通过建立粗糙集与支持向量机进行集成，对企业贷款违约相关数据进行分析，对违约与否进行了判别，获得了比多元判别分析更好的判别效果。刘思蒙（2021）使用了随机森林及决策树两种分类算法对美国贷款违约平台爬取的数据进行预测，并进行了综合比较分析，最终随机森林算法在该不平衡数据集上获得的精度结果更高。金鑫（2019）使用 XGBoost 算法对 P2P 网贷违约数据进行了预测，并基于 XGBoost 模型获得相关变量重要性排名，最终对我国 P2P 行业提出相关建议。张丽颖（2022）通过比较单一模型和集成模型在同一贷款违约数据上的分类准确度，研究发现集成学习模型在预测精度上具有优越性。

(三) 研究内容与创新

基于以上研究成果，本文聚焦于个人贷款违约数据，首先对数据集中各变量进行相关数据分析，探究各变量与预测变量的相关性。然后，基于集成学习算法在处理贷款违约问题预测精度较高的优越性，本文使用 LightGBM、XGBoost 以及 CatBoost 三种集成学习算法进行预测分析，进行预测结果的对比。最终，对三个模型的预测结果进行融合并分析预测效果。

针对创新层面，一是研究内容上，现如今国内许多针对国内信贷的研究内容集中于企业贷款违约以及 P2P 网贷违约，针对个人贷款违约分析相对较少，本文通过对个人贷款违约数据集的分析，

获得相关影响个人违约的因素分析。二是研究方法上，使用了 CatBoost 模型对贷款违约数据进行预测，并对几个集成学习模型获得结果进行融合，最终得出了预测效果更好的方法。

二、使用方法介绍

(一) 算法介绍

1. XGBoost 算法介绍

梯度增强决策树（GBDT）是机器学习领域的一个重要模型，其优点在于精度高并且不容易发生过拟合。XGBoost 全称是 Exterme Gradient Boosting（极限梯度增强），是由 GBDT 发展而来的一种算法。

XGBoost 模型在目标函数上选择了二阶泰勒展开并加入正则化项的方式。其生成决策树的过程是逐步寻找最佳分裂点的过程：首先使用预排序算法，即对所有的特征按照特征的数值进行排序，然后遍历所有特征上的所有分裂点，并计算按照这些候选分裂点分裂后的全部样本的目标函数的优化程度，最后选择优化程度最大的特征及其对应的分裂点位进行分裂，这样一层一层的分裂过程就完成了一棵决策树的生成过程。在 XGBoost 模型的整个训练过程中都聚焦于残差，每次会通过聚焦于残差训练出一棵决策树，最终的预测结果是将所有树的预测结果进行加和从而组合成强学习器。其优点在于鲁棒性强，相较于其他的一些模型不需要特别精细的调参就能取得一个不错的结果。

2. LightGBM 算法介绍

LightGBM 也是实现 GBDT 的一种框架。因 GBDT 模型在获得最优模型的方式主要是利用决策树来进行迭代训练，所以对内存具有较高的要求，倘若通过反复读写训练数据来解决内存不够的问题，又会存在耗费时间较长的缺点。LightGBM 算法的提出解决了 GBDT 模型的上述缺点，LightGBM 模型能够兼具处理海量数据及高速训练的优良性能。

LightGBM 模型是在 XGBoost 模型后提出的，其相比 XGBoost 做出了一些优化。在上述介绍中可以看出 XGBoost 模型在决策树的训练过程中对最优分裂点的选择采用了遍历的方式，即对每个可能的分裂点都进行计算，这样的计算过程很浪费时间和空间。LightGBM 就是从这个问题入手，主要考虑了三个方面去解决这个问题：一是采用直方图算法减少分裂点数量，二是采用梯度抽样算法减少样本数量，三是采用互斥特征捆绑算法减少特征的数量。因此其相较于 XGBoost 训练速度上有了很大的提升。

3. CatBoost 算法介绍

CatBoost 是 categorical boosting（分类提升）的简写，其是一种能够很好的对类别型特征的梯度进行提升的算法。该模型的一个重要特征是对定性变量的处理，在输入编码好的类别变量数据后，CatBoost 会对其进行重新编码，首先其对样本进行随机排序，每个样本的某一定性变量的编码都是

基于计算这个样本以前的该定性变量的原编码结果的均值获得，同时在这个均值的计算过程中加入了优先级，并对优先级赋予不同的权重系数，这样做的结果是可以定性变量中某一类型总数较小带来的噪声，可以缓解因数据不平衡导致预测精度的下降。另外，因使用这种编码策略将类别特征转化为了数值型特征，会影响特征之间的交叉，CatBoost 模型使用贪心策略来进行特征交叉，在生成决策树的第一次分裂，CatBoost 不使用任何交叉特征。但在后续结点分裂过程中，CatBoost 会使用生成决策树过程中所用到的全部原始特征和交叉特征，并与数据集中的全部类别特征进行交叉。

在内存应用方面，CatBoost 模型与 LightGBM 模型具有同样的效果，但其优势在于处理类别型变量的方式不同，从而在处理类别型变量较多的分类问题上预测精度更高。

(二) 模型评估指标

1. AUC 值及 ROC 曲线评估方法

ROC 曲线中文名称叫做受试者工作特征曲线，通常情况下，横坐标一般是假阳率（FPR），纵坐标一般是真阳率（TPR）。通过计算某一模型在测试样本的预测结果，可以与实际结果获得一个 TPR 和 FPR 的点对，通过调整分类器分类的阈值，最终就会获得一条经过原点与 (1,1) 的曲线，即为 ROC 曲线。

AUC 值在机器学习领域中是一种比较常见的模型评估指标，其全称是 area under the curve，计算的是 ROC 曲线下的面积，面积越大代表预测结果越好。

2. logLoss 评估方法

logLoss 中文名称为对数损失函数，也叫二元交叉熵损失函数，是基于概率的一个非常重要的分类度量方法。计算公式如下：

$$L(Y, P(Y|X)) = -logP(Y|X)$$

对于给定的问题，较低的对数损失函数值往往代表更好的预测结果。

三、数据介绍与分析

(一) 数据选择

本文选择的数据集是个人贷款违约预测数据集，其中包含 13 个字段如表 3-1，其中训练集中包含 168000 个样本，测试集中包含 84000 个样本，其中 is_married、house_ownership、car_ownership、label 为定性变量，其余均为定量变量。

表 3-1: 个人贷款违约数据字段说明

id	数据序号
income	收入
age	年龄
experience_years	从业年限
is_married	婚姻状况
city	居住城市
region	居住地区

current_job_years	现任职位工作年限
current_house_years	在现房屋的居住年数
house_ownership	房屋类型
car_ownership	是否拥有汽车
profession	职业
label	是否存在违约

(二) 数据预处理

首先进行定性变量进行数据类型转化，因三种方法模型需要的输入数据的格式不同，所以先进行统一编码后输入。将 is_married、house_ownership、car_ownership 三个字符型变量进行标签编码，编码规则如表 3-2：

表 3-2: 编码对应说明

编码	is_married	house_ownership	car_ownership
0	married	norent_noown	no
1	single	owned	yes
2		rented	

数据预处理的第二步是对缺失值进行处理，对各个变量缺失值进行统计，发现各列均没有缺失值，可以继续对数据集进行分类研究。

(三) 数据分析

首先针对训练集各样本进行统计，其中发生违约的总人数为 20675，占比 12.3%，未发生违约的人数为 147325，占比 87.7%，未发生违约人数大约为违约人数的 7 倍左右，样本存在不均衡问题。

接着针对违约人群与未违约人群进行对比分析，先对三个分类变量 is_married、house_ownership、car_ownership 进行统计，因根据上述分析违约与非违约样本存在不均衡问题，所以绘制比例图更容易看出个数据之间的区别，获得相关可视化结果如下：

对婚姻状况进行统计可以发现，单身贷款人数几乎是已婚贷款人数的 8.8 倍，从图 3-1 可以看

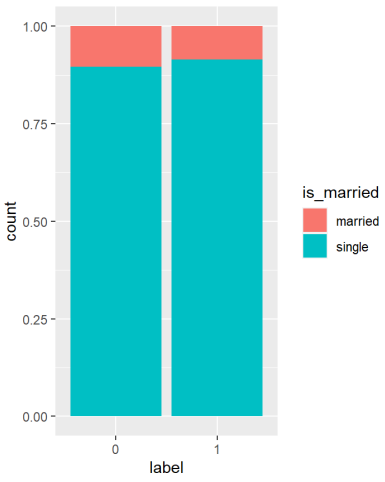

图 3-1: 婚姻状况对比图

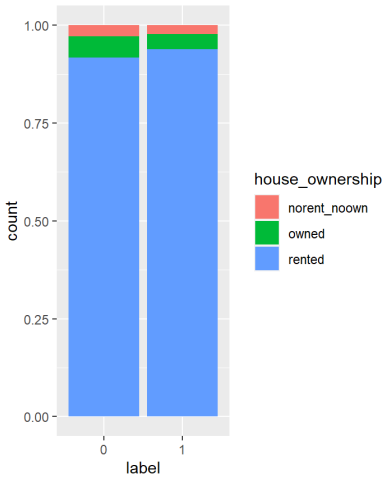

图 3-2: 房屋类型对比图

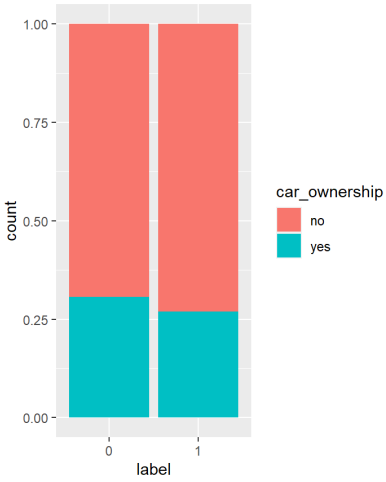

图 3-3: 拥有汽车与否对比图

出，违约群体（1）相比正常偿还贷款群体（0）单身人数占比更大，根据统计发现，单身人群违约占比相较于已婚人群高 2.24%。对房屋类型进行统计可以发现，租房人数占比最大，从图 3-2 分析可以看出，违约群体中租房人数占比更大，而通过统计也发现，拥有住房人群相比租房人群违约占比高 3.45%。对拥有汽车与否进行统计可以发现，无车人群相对较多，从图 3-2 分析可以看出，违约群体中无车人群占比更大，而通过统计也发现，无车人群相比有车人群违约占比高 1.85%。

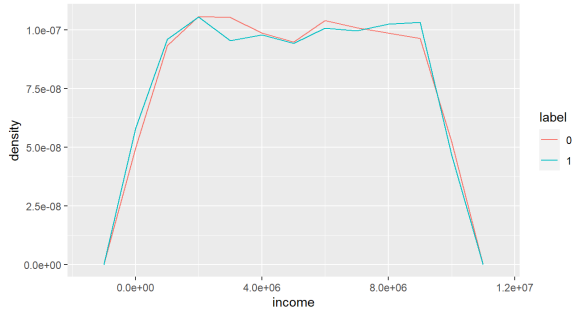

图 3-4: 收入对比变化图

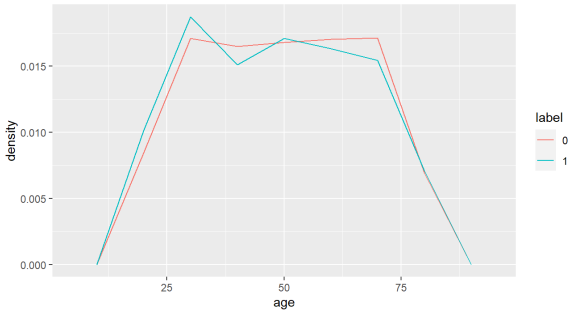

图 3-5: 年龄对比图

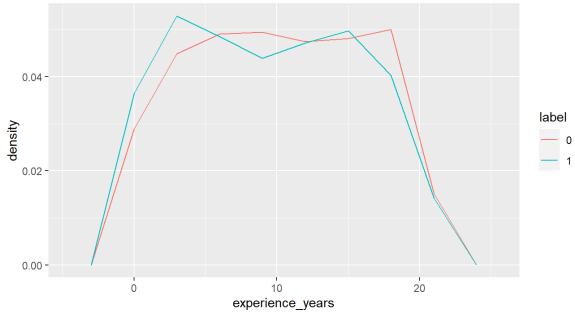

图 3-6: 从业年限对比图

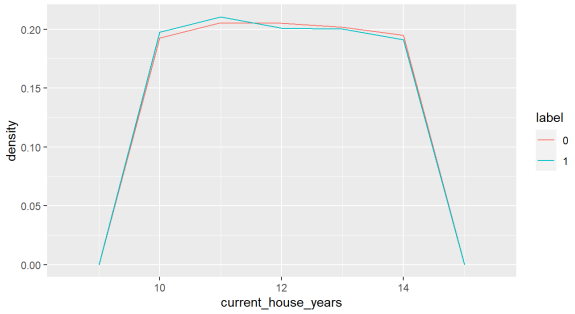

图 3-7: 现有房屋居住年限对比图

对收入、年龄、从业年限、现有房屋居住年限几个变量进行密度图绘制如图 3-4 到图 3-7，可以发现蓝色线（违约人群）相比红色线（未违约人群）在左侧密度值更大，代表意义是收入越低，年龄越小，从业年限及居住年限都越少，这一部分群体往往不稳定性更强，更容易发生贷款违约。

针对地区、城市以及职业这几个变量，由于收集数据时考虑到个人的隐私性，没有注明具体内容，而是采用数字进行表示。获得统计结果，地区共有 29 个，其中编号为 25 的地区违约人数最多，共 2221 人，而编号 21 的地区违约人数最少，共 15 人。针对城市进行统计，城市共 317 个，其中编号为 160 的地区违约人数最多，共 159 人，而编号 98 的地区违约人数最少，共 13 人。针对职业进行统计，职业共 51 个类型，其中编号为 43 的地区违约人数最多，共 517 人，而编号 49 的地区违约人数最少，共 265 人。以上三个变量各个分类之间均具有比较大的差异性，对响应变量的影响程度不同，可以进行进一步的分析。

从上述分析结果可以看出，数据集中各个字段均对个人贷款违约产生影响，所以可以对这些变量进行模型的训练。

四、模型建立

(一) K 折交叉验证

使用 K 折交叉验证的目的是为了得到更加稳定可靠的预测结果，在这里选择 k=5 进行分折，每次采用训练集中 4/5 的数据进行训练，剩余 1/5 的数据划分为验证集，在每一折训练过程中保留验证集数据拟合效果最好的模型对测试集数据进行拟合，并保留每次测试集数据拟合后获得的结果，最终模型对测试集的预测结果为 5 次预测结果的平均值。

根据前文分析，训练集样本统计结果表明，未发生违约人数与违约人数的比值大约为 7:1，样本存在不均衡问题。直接使用 K 折交叉验证随机分折会影响模型的性能，因此在分折过程中，R 语言分析过程中选择了 createDataPartition 函数代替 createFolds 函数进行分折处理。Python 语言的分析过程中选择了 StratifiedKFold 进行分折操作，在 KFold 的基础上，StratifiedKFold 加入了分层抽样的思想，使得测试集和验证集有相同的数据分布，与 createDataPartition 效果相同。以上两个函数都可以自动从标签的各个取值中随机取出等比例的数据来组成验证集，使得训练集和验证集具有相似的数据分布。

(二) LightGBM 模型实验过程

首先采用 LightGBM 模型进行训练，因模型参数过多，仅考虑对一些重要参数进行调整，其余均采用默认值，调整后的结果如下表 4-1 所示。在对 LightGBM 模型进行调参过程中，重点考虑决策树的最大深度以及叶子节点数，因为这两个参数是需要调整的核心参数，其对模型的预测精度以及泛化能力起着决定性的作用。在调参过程中调整这两个参数对精度提升最为明显，通过反复调整参数最终确定了 max_depth 为 9 以及 num_leaves 为 29。

在调参过后模型最终模型获得的 AUC 值为 91.2%，logLoss 值为 0.21。绘制 ROC 曲线图如图 4-1 以及混淆矩阵如图 4-2。绘制特征重要性排序图如下图 4-3，从特征重要性图中可以看出根据 LGB 模型进行预测，收入和居住城市对最终标签预测结果影响较大。

表 4-1: LightGBM 模型参数调整

参数	设置原因
learning_rate = 0.1	默认
n_estimators = 10000	设置了一个较大的值，配合 early_stopping_round=200 让模型自主选择最好的迭代次数
max_depth = 9	采用这个树深与 num_leaves = 29 配合调整树的形状获得了不错的精度和泛化能力
subsample = 0.7	在每次迭代时重采样部分数据，防止过拟合

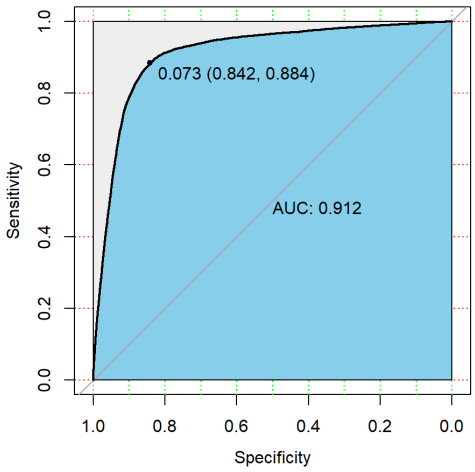

图 4-1: LightGBM 模型 ROC 曲线

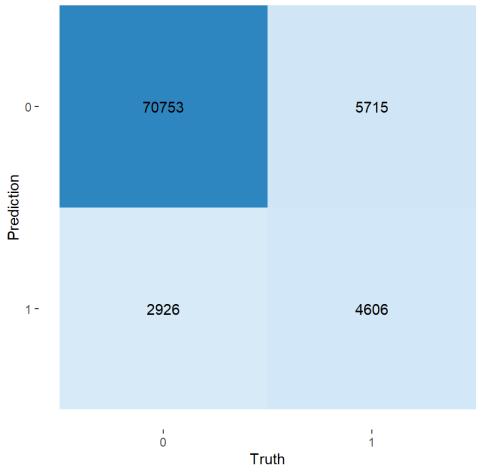

图 4-2: LightGBM 混淆矩阵

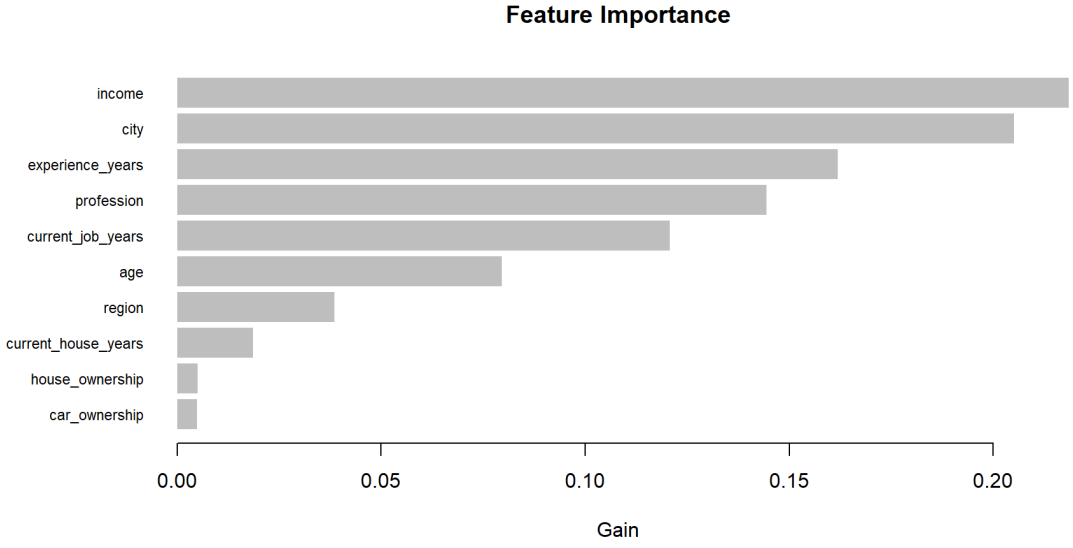

图 4-3: LightGBM 模型特征重要排序图

(三) XGBoost 模型实验过程

XGBoost 模型与 LightBoost 模型在 R 中需要的输入形式不同，需要将数据先用 sparse.model.matrix 函数处理为稀疏矩阵，在此过程中，该函数对类别变量重新进行了 One-hot 编码。对 XGBoost 模型进行了类似的调参操作，对一些重要参数调整后的结果如下表 4-2 所示。对 XGBoost 模型进行训练过程中发现其相比 LightGBM 在这一数据集上更容易发生过拟合，因此选择了相比 LightBoost 选择较小的树深，并且提高了参数 min_child_weight 来防止过拟合的发生。

表 4-2: XGBoost 模型参数调整

参数	设置原因
learning_rate = 0.1	默认
n_estimators = 10000	设置了一个较大的值，配合 early_stopping_round=400 让模型自主选择最好的迭代次数
max_depth = 5	相比 LightBoost 选择较小的树深
min_child_weight = 3	默认值为 1，提高这一数值用于避免过拟合
subsample = 0.7	根据 LightGBM 训练结果在初始化时即选择了相同的参数

在调参过后最终模型获得的 AUC 值为 91.4%，logLoss 值为 0.206，在测试集上的预测结果相比 LightGBM 模型略有提升，绘制 ROC 曲线图如图 4-4 以及混淆矩阵如图 4-5。整个 XGBoost 模型的训练时长是 LightGBM 模型的几倍，也可以看出 LightGBM 模型的训练速度相较于 XGBoost 模型有了很大的提升。

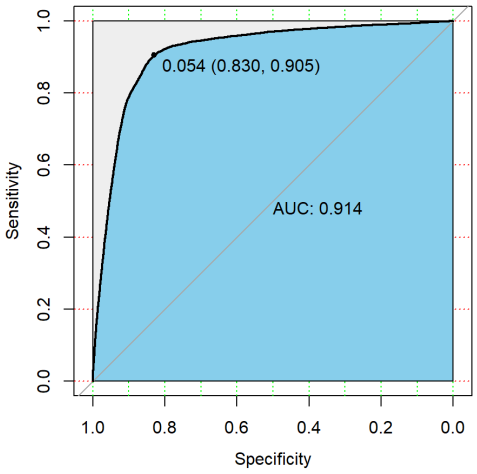

图 4-4: XGBoost 模型 ROC 曲线

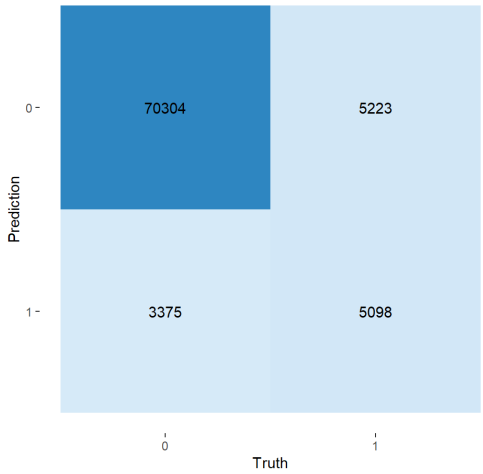

图 4-5: XGBoost 混淆矩阵

绘制特征重要性排序图如下图 4-6，XGBoost 模型选择的重要特征前两名仍为收入和城市，但收入这一变量相比 LightGBM 模型的训练结果，其重要程度明显提升，另外第三名第四名的特征发生变化，提取的特征为职业和年龄，前五个标签的总重要性程度超过 70%，以上几个变量均在贷款

违约预测中发挥着比较重要的作用。从这张图中也可以看出 sparse.model.matrix 重新进行 One-Hot 编码生成了新的列，比如排名倒数第一的 is_marriedsingle 变量。

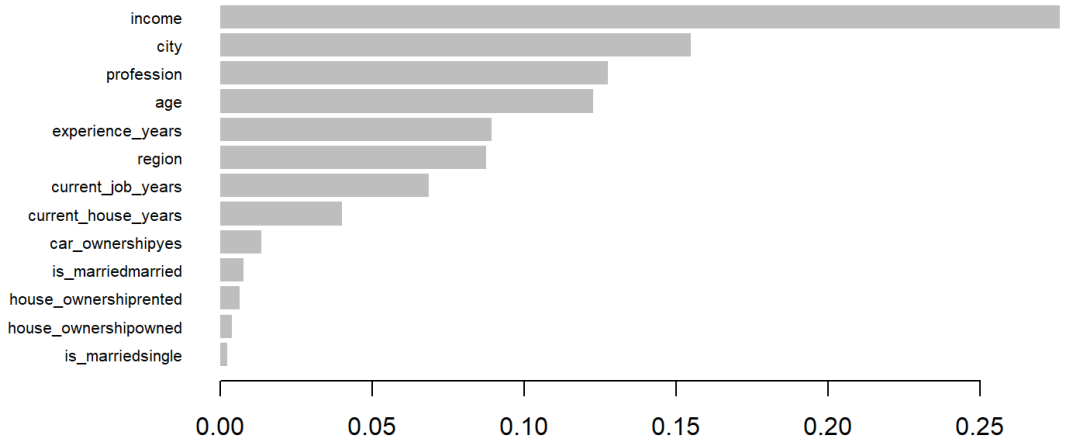

图 4-6: XGBoost 模型特征重要排序图

(四) CatBoost 模型实验过程

对 CatBoost 模型进行调参，对一些重要参数调整后的结果如下表 4-3 所示。因为不同模型的输入参数各不相同，所以避免过拟合的方法也不尽相同，在 CatBoost 模型中主要选择了增大正则化项来避免过拟合，最终选择的 l2 参数为 10。

表 4-3: CatBoost 模型参数调整

参数	设置原因
learning_rate = 0.1	默认
depth = 9	深度与 LightGBM 相同
l2_leaf_reg = 10	较大的正则化项，防止发生过拟合
od_wait = 50	在迭代之后以最佳度量值继续训练的迭代次数，与过拟合检测器 Iter 配合，防止发生过拟合

最终训练模型获得的 AUC 值为 92.6%，相较于前两个模型的最终训练结果有了较大的提升，绘制 ROC 曲线图如图 4-7 以及混淆矩阵如图 4-8。CatBoost 模型相较于 XGBoost 模型的训练速度也快很多，但会略慢于 LightGBM 模型。

绘制特征重要性排序图如下图 4-9，采用 python 算法绘制的重要性图没有进行排序处理相对 R 的绘图结果比较杂乱。不过还是能够清晰地看出收入与城市是最重要的两个特征，但相比前两个模型其重要性程度更加接近，并且年龄和职业两个变量的重要性也有提升。再次对比三张图可以发现，是否结婚与房屋拥有状况两个变量的重要性程度一直最小，说明其对响应变量的影响程度相对较小。

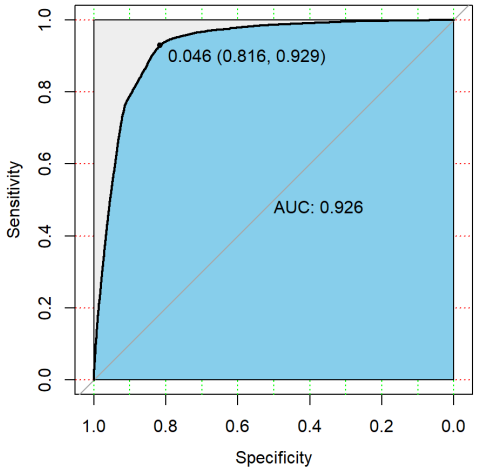

图 4-7: CatBoost 模型 ROC 曲线

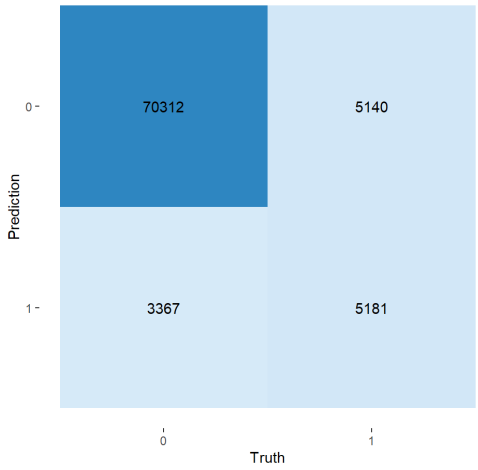

图 4-8: CatBoost 混淆矩阵

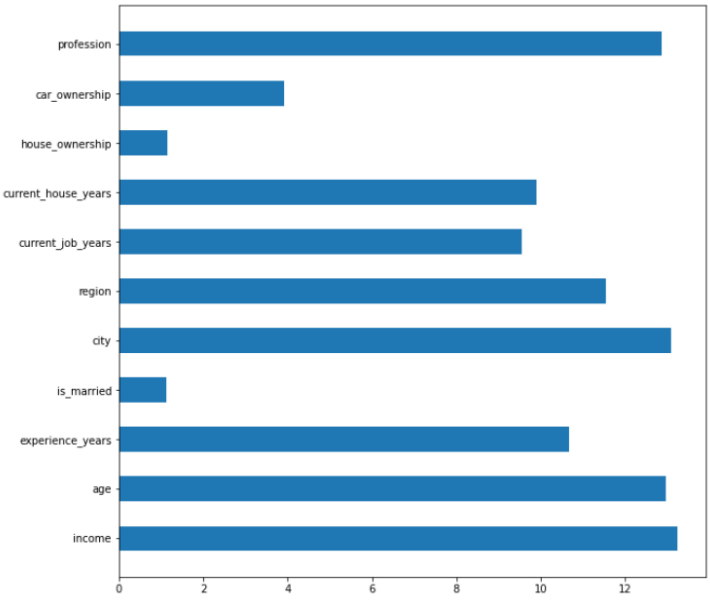

图 4-9: CatBoost 模型特征重要排序图

(五) 模型融合

模型融合是先产生一组学习器，再用某种方法将这些学习器结合起来，从而达到加强模型效果的方法。针对上述三个模型，对其预测结果处理为相同的格式，首先对三个模型的预测结果进行直接平均，获得的 AUC 值为 92.58%，相较于 CatBoost 模型的预测结果 92.6% 反而有所下降。

考虑到 CatBoost 模型相较于前两个模型的预测效果好很多，采用加权平均的方式，对其赋予一个较大的权重，这里设为 0.8，对前两个模型 LightGBM 和 XGBoost 分别赋予 0.1 和 0.2 的权重，最终获得的 AUC 值为 92.83%，获得了比三个模型单独预测更好的结果。

模型融合与三个模型的对比结果如下：

模型名称	LightGBM	XGBoost	CatBoost	模型融合
AUC 值	91.2%	91.4%	92.6%	92.83%

倘若三个模型单独预测的结果与上述结果不同，也可以参照类似的方法，对预测准确率最高的模型赋予最高的权重，然后对模型的预测结果进行加权平均，应该可以得出类似的结论。

五、结论与建议

通过上述分析，对比三个模型，可以发现 CatBoost 模型在处理个人贷款违约数据集的效果要明显好于 LightBoost 模型和 XGBoost 模型。在训练速度上，LightBoost 模型的训练速度最快，CatBoost 模型其次，两个模型的训练速度都会远远优于 XGBoost 模型。另外，尽管 CatBoost 模型的训练精度相较于前两个模型已经有较大提升，但通过加权平均的方式将三个模型进行融合，仍然获得了更好的效果。

根据本文的研究结果，针对银行以及小型借贷机构，可以采用集成学习的方法对贷款人进行未来会违约与否的预测。另外，根据模型的特征重要性分析，在进行放款决策前可以着重考虑贷款人的收入，这一点也很符合常理，因为收入水平很大程度上决定了贷款人的偿还能力。贷款人的年龄及职业也是比较重要的考虑因素，相较之下，贷款者的婚姻状况以及房屋拥有状况对违约的影响相对较小，可以作为后续的考虑因素。

参考文献

[1] 金鑫. P2P 贷款违约预测模型的实证分析 [D]. 上海财经大学, 2020.

[2] 柯孔林. 基于粗糙集与支持向量机的企业短期贷款违约判别 [J]. 控制理论与应用, 2009,26(12):1365-1370.

[3] 程朋媛. 基于大数据及机器学习方法的贷款违约风险评估 [J]. 营销界, 2021(26):98-99.

[4] 张佳倩, 李伟, 阮素梅. 基于机器学习的贷款违约风险预测 [J]. 长春理工大学学报 (社会科学版), 2021,34(04):105-111.

[5] 刘思蒙. 基于决策树与随机森林的个人网络贷款违约行为研究 [D]. 中国地质大学 (北京), 2020.