Statistics and Application 统计学与应用, 2024, 13(2), 486-495
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.132048

Hans汉斯

基于Stacking模型融合的车辆保险反欺诈识别

金天颖，张美辰，王可欣

燕山大学理学院，河北 秦皇岛

收稿日期：2024年3月28日；录用日期：2024年4月18日；发布日期：2024年4月26日

摘　要

基于Kaggle公开数据集，首先对其进行了数据清洗、数据均衡化等预处理工作，确保数据的质量和适用性。在预处理完成后，使用随机森林、XGBoost和LightGBM三种不同的机器学习算法构建了保险欺诈识别模型，并通过网格搜索法调整参数来提高其性能。采用分类模型的评估方法，对这三个模型的Pression、Recall等指标进行了对比分析，结果显示LightGBM模型在整体的分类效果上表现最好。最后，我们引入了Stacking技术，将三个单一模型作为初级分类器，Logistic Regression作为元分类器，得到了融合的的车险欺诈识别模型。这一模型结合了三个单一模型的优点，不仅具有较高的稳定性，还能提高整体的预测精度。通过模型融合，我们得到了一个更加全面、准确的欺诈识别系统，为保险公司提供了更可靠的风险管理工具。

关键词

模型融合，XGboost，LightGBM，随机森林，车险欺诈

Anti-Fraud Identification of Vehicle Insurance Based on Stacking Model Fusion

Tianying Jin, Meichen Zhang, Kexin Wang

College of Science, Yanshan University, Qinhuangdao Hebei

Received: Mar. 28th, 2024; accepted: Apr. 18th, 2024; published: Apr. 26th, 2024

Abstract

Based on the Kaggle public dataset, we first conducted preprocessing such as data cleaning and balancing to ensure the quality and applicability of the data. After preprocessing, we constructed insurance fraud detection models using three different machine learning algorithms: Random Forest, XGBoost, and LightGBM. We used grid search to tune the parameters to improve their performance. By evaluating the models using classification metrics such as Precision and Recall, we

文章引用: 金天颖, 张美辰, 王可欣. 基于 Stacking 模型融合的车辆保险反欺诈识别[J]. 统计学与应用, 2024, 13(2): 486-495. DOI: 10.12677/sa.2024.132048

found that the LightGBM model performed the best overall. Finally, we introduced the Stacking technique, combining the three individual models as base classifiers and Logistic Regression as the meta-classifier, to obtain a fused auto insurance fraud detection model. This model combines the advantages of the three individual models, exhibiting not only high stability but also improved overall prediction accuracy. Through model fusion, we obtained a more comprehensive and accurate fraud detection system, providing insurance companies with more reliable risk management tools.

Keywords

Model Fusion, XGBoost, Lightgbm, Random Forest, Automobile Insurance Fraud

1. 引言

我国成为全球最大的机动车消费市场，并实施了 2020 年的车险综改，这促使车险市场潜力巨大，同时也增加了车险欺诈问题。据报道，我国车险欺诈占据了保险欺诈总案件的 80%，每年约有 200 亿元的案件涉及金额。截至 2022 年第一季度，银保监会及其派出机构收到并转交了 2.65 万件涉及保险公司的投诉，其中机动车辆保险和普通人寿保险的投诉数量最多。车险欺诈不仅给保险公司带来了经营压力和损失，还违反了公平、诚信、正义的社会价值观念，损害了合法消费者的权益。因此，随着机器学习技术在保险领域的应用不断深入，它被认为是最有前景的欺诈识别工具之一[1] [2] [3] [4]。

自 2000 年以来，随着机器学习在各行各业的应用不断增加，它也开始在保险欺诈识别领域得到应用。Stijn Viaene 等人(2002)使用了 1993 年的马萨诸塞州的人身伤害保护索赔数据集，评估了多种算法，包括 Logistic 回归、C4.5 决策树、KNN 等。结果显示 C4.5 决策树的效果较差[5]。Clifton Phua 等人(2004)采用了反向传播(BP)、朴素贝叶斯(NB)和决策树模型进行预测，通过 bagging 技术将这些基础分类器组合，结果显示 bagging 后的模型略优于最佳分类器决策树模型[6]。Javier Muguerza 等人(2005)分析了分类树在解决汽车保险公司欺诈检测问题时的表现，并比较了合并树和 C4.5 树，同时对误差进行了更广泛的分析[7]。搨思杰(2016)比较了 AAAG 模型、BP 神经网络、EXPERT SYSTEM 等多种欺诈识别模型的优缺点，并从保险公司理赔系统的 13 个指标中筛选出了 8 个与保单欺诈显著相关的变量[8]。李亚琪(2018)通过蜂群算法优化了极限学习机和随机森林模型，并在索赔数据上验证了这种优化对模型预测准确度的提升[9]。而杜小雨(2019)则使用了 APriori 和 FP-growth 算法进行了车险理赔样本的关联分析，证实了关联分析的有效性[10]。

本文分别构建随机森林、XGboost、LightGBM 模型的保险欺诈识别模型，并分别进行调参以求获得更好性能。

2. 数据来源及预处理

2.1. 数据来源

本文使用了来自 Kaggle 数据平台的美国某保险公司的车险索赔数据集。由于国内保险公司用户信息不公开，因此选择了国外数据。该平台数据已经通过官方验证，数据可靠有保障。数据集包含不同区域

的车险索赔信息，共有 15,420 条记录，32 个特征，该数据主要描述了车辆保险用户的基本信息和事故相关情况。因变量为是否欺诈，取值为 0 或 1。示例数据见表 1，数据集变量特征见表 2。

Table 1. Sample data
表 1. 示例数据

Accident Area	Month Claimed	Sex	Marital Status	Age	…	Deductible	Driver Rating	Fraud Found
Urban	Jan	Female	Single	21	…	300	1	No
Urban	Jan	Male	Single	34	…	400	4	No
Urban	Nov	Male	Married	47	…	400	3	No
Rural	Jul	Male	Married	65	…	400	2	No
Urban	Feb	Female	Single	27	…	400	1	No

Table 2. Variable features of the dataset
表 2. 数据集各变量特征

序号	变量类型	特征名	特征含义
1	时间	Month	月份
2		Week of Month	第几周
3		Day of Week	星期几
4		Day of Week Claimed	上报星期几
5		Month Claimed	上报月份
6		Week of Month Claimed	上报周
7		Year	年份
8	汽车	Make	汽车品牌
9		Accident Area	事故发生地区(城市、郊区)
10		Vehicle Category	汽车类型
11		Vehicle Price	汽车价格
12		Age of Vehicle	汽车年龄
13	人	Marital Status	婚否
14		Age	年龄
15		Sex	性别
16		Fault	过错方
17		Policy Type	事故类型
18		Driver Rating	司机评分
19		Age of Policy Holder	投保人年龄
20		Past Number of Claims	投保人以前提交的事故数量
21		Number of Suppliments	额外索赔次数
22	与事故相关	Deductible	汽车保险可扣除的金额
23		Days_Policy_Accident	从购买保险到事故发生的天数
24		Days_Policy_Claim	从购买保险到提交索赔之间的天数
25		Base Policy	保险范围类型
26		Witness Present	证人数量
27		Agent Type	代理类型
28		Police Report Filed	该事故是否上报了警方
29		Address Change_Claim	提交地址更改的时间
30		Number of Cars	涉及到事故的总共汽车数量

续表

31	其他	Policy Number	事故类型编号
32		Rep Number	销售代表编号，1~16 之间的整数
33	目标变量	Fraud Found_P	指示索赔是否欺诈

FraudFound_P 为记录是否欺诈的标签，其余的 32 个特征本数据集将其分为五个方面进行描述，其中与人相关共 9 项变量，与车相关共 5 项变量，与时间相关 7 项变量，其他项 2 项变量，其中 RepNumber 代表处理保单人员的编号，为 1~16 间的重复数值，Policy Number 为记录用户的标签。

2.2. 数据清洗

数据清洗是为了使数据适合分析和建模，包括去除重复数据、填补缺失值、处理异常值和转换数据格式等。在本文中，经检查发现数据集无缺失值。因此，首先对数据进行了纵向无效变量的剔除以减少冗余，然后横向分析了各变量的异常值并进行了处理。

首先，我们的研究目的是对该保单是否为诈骗进行判断，而诸如保单号码、事故发生的星期、月份等对保单起标记作用的变量对判断目标变量毫无作用，因此直接将其删除。其次，变量 Policy Type 取值为“Base Policy”与“Vehicle Category”两变量取值的组合，变量冗余，因此剔除 Policy Type 变量。另外，由于数据中包含两种关于年龄的变量：一种是离散型的，另一种是类别型的。为了简化模型并提高效率，我们决定移除离散型的年龄特征，仅保留类别型的年龄信息。最后汽车品牌“Make”这个变量虽对该模型有一定的影响，但因其取值过多(19 个)，且不便于对其进行具体的分组，故将此变量删除，也便于后续的模型构建，缩短数据运行的时间。因此，我们总结得出了 21 个对车险欺诈具有显著影响的特征。

在本数据集中，客户年龄的最小值是 0，客户最大年龄达到 86。年龄的箱线图如下图 1 所示。对于年龄最小最小值为 0，毫无疑问是异常数据，考虑到查证的困难性，以及数据集 1 万多条记录的情况，本文将年龄为 0 的记录直接删除；对于年龄超过 80 岁的客户，经查证，在美国驾驶证不设年龄上限。如根据伊利诺伊州法律规定，75 至 80 岁的老人每 4 年需要进行一次路考来延长驾照，而 81 至 86 岁的老人每两年需要路考一次，87 岁以上的老人则每年需路考一次。因此，本文不对年龄上限进行调整。

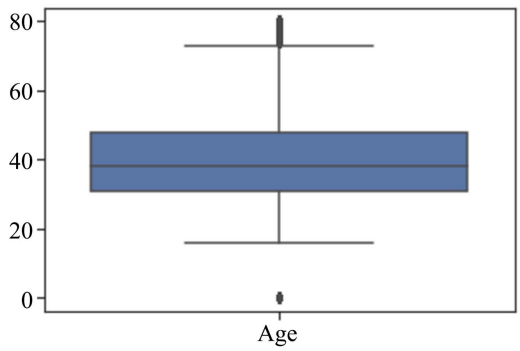

Figure 1. Age box plot
图 1. 年龄箱线图

2.3. 数据均衡化

在反欺诈问题中，由于欺诈案例较少，导致数据集不平衡。传统机器学习模型通常要求各类别样本数量均衡，因此需要不均衡样本进行处理。

该数据集包含 15,420 条索赔记录，其中 94%为正常索赔，6%为欺诈索赔，呈现明显的不平衡情况。

机器学习常使用重采样技术来平衡样本，解决不平衡的二分类问题。重采样技术通过利用已有数据和技术手段，在不收集其他数据的情况下生成新的数据分布，以实现样本的均衡。SMOTE 算法是改进的随机过采样算法，有效地解决了模型过拟合的问题。它的基本思想是对少数类样本进行分析，并根据这些样本生成新样本，从而使数据集更加均衡。因此，本文采用 SMOTE 过采样技术对数据集进行采样，以达到平衡数据集的目的，数据采样之后的训练集数据分布如表 3 所示。

Table 3. SMOTE oversampling algorithm.
表 3. 数据 SMOTE 过采样后的训练数据分布

样本类别	原始数据集样本量	采样后数据集样本量
非欺诈样本量	14,208	14,208
欺诈样本量	8892	14,208
总计	115,100	228,416

3. Stacking 模型建立

Stacking 是一种从模型本身出发的更为先进的模型融合方法，其核心思想是在第一层利用多个基学习器学习原始数据，然后将这几个基学习器的输出按照列的方式进行堆叠，构成新数据，再将新的样本交给第二层的模型进行拟合，从而得到最终预测结果。这个流程以一种有效的方式组合了多个算法，从而提高了模型的准确性和泛化能力。下图 2 为模型建立流程图：

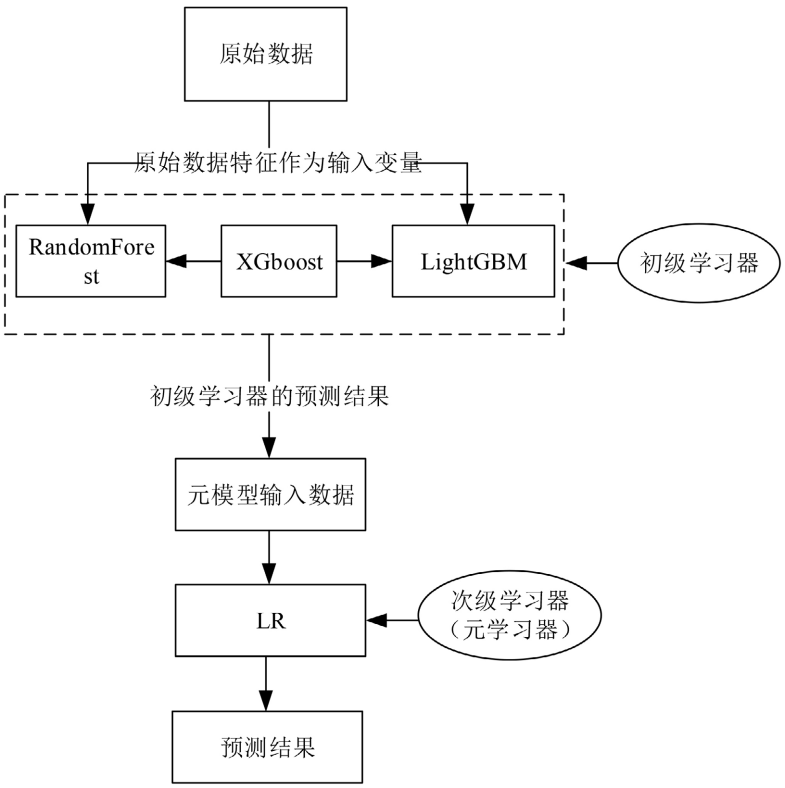

Figure 2. Diagram of model building process
图 2. 模型建立的流程图

3.1. 随机森林模型

3.1.1. 随机森林算法

随机森林(RF)是 Bagging 的一种扩展，将决策树作为基学习器，并引入了随机属性选择。通过在全样本数据集中进行有放回抽样来训练多棵随机决策树，然后通过投票决定最终预测结果。本文选用了 CART 决策树算法，利用基尼指数来判断数据的纯度。RF 在减少模型方差的同时提高了泛化能力，尤其适用于处理高维和大规模数据。

按 A 属性分裂的前后增益值，增益值最大作为特征选取原则。方法为

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{1}$$

式中 $Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$。$c$ 表示不同类别，p_i 表示类别 i 占整体的比例大小，即数据越混乱，相应 $Gini$ 系数值越大。$Gini_A(D)$ 为选取属性 A，分裂后数据集 D 的系数值，计算公式为

$$Gini_A(D) = \sum_{j}^{k} \frac{|D_j|}{|D|} Gini(D_j) \tag{2}$$

3.1.2. 随机森林建模

本文采用 Python3.7 环境下的机器学习库 sklearn 中的 ensemble 模块来建立 Random Forest 模型，以下详细介绍模型建立的过程与超参数优化的结果。

本文实验调参过程中主要对如下表 4 中参数进行试验与调整，参数列表及取也见表 4。

Table 4. Random forest network search parameter list
表 4. 随机森林网络搜索参数列表

RF 模型参数	参数含义	参数列表	参数取值
n_estimators	决策树的个数	[10, 20, 30, …, 190, 200]	100
max_depth	决策树最大深度	[1, 2, 3, 5, 7, 9, 11, 13]	13
min_samples_leaf	叶子节点含有的最少样本	[10, 20, 30, 40, 50, 100]	10
min_samples_split	节点可分的最小样本数	[80, 100, 120, 140]	80
max_features	构建决策树最优模型时考虑的最大特征数。	[3, 5, 7, 9, 11]	11

随机森林模型的混淆矩阵，如表 5 所示。

Table 5. Confusion matrix for random forest model
表 5. 随机森林模型混淆矩阵

真实标签	预测标签	
	未欺诈	欺诈
未欺诈	2100	778
欺诈	252	2554

由混淆矩阵得出的其他指标如表 6 所示。

Table 6. Random forest model classification metrics
表 6. 随机森林模型的分类指标

RF	Precision	Recall	F1_score	support
0	0.89	0.73	0.8	2878
1	0.77	0.91	0.83	2806

根据上述评估指标，我们发现模型对于非欺诈情况有更强的预测能力，Precision 较高。Recall 指标表明在欺诈样本中被正确识别的比例达到了 0.91，而欺诈检测的核心是准确识别欺诈样本，因此该模型是可接受的。整体而言，该模型对数据集的预测准确率为 0.82，预测能力一般。

3.2. XGBoost 模型

3.2.1. XGBoost 算法

XGBoost (Extreme Gradient Boosting)算法是梯度提升树(GBDT)的改进版本，通过在目标函数中引入叶子权重和 L2 正则化，有效降低模型复杂度，减少方差，从而优化模型性能。通过权衡预测误差和复杂度，XGBoost 能够迭代构建多棵树，逐步提高预测能力，使其在各种预测建模任务中表现优异。

目标函数为

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \Omega\left(f_j\right) \tag{3}$$

其中，结构风险项$\Omega\left(f_j\right)$又由两部分组成

$$\Omega\left(f_j\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

其中 T 参数为第 j 棵树叶子节点的个数，w_j^2 是叶子节点值的 L_2 范数，γ、λ 分别为惩罚系数、权重惩罚系数。

3.2.2. XGBoost 建模

构建最优模型的通常方法是通过最小化训练数据的损失函数，而 XGBoost 采用了结构风险最小化的损失函数。这个损失函数包含两个主要部分：一是衡量预测值与真实值之间差距的部分，二是正则化项，用于控制模型的复杂度，其中包括叶子节点的数量和分数。在生成树时，XGBoost 会考虑树的复杂度，从而提高模型的泛化能力。此外，XGBoost 还利用多线程技术选择最佳切分点，以提高模型的训练速度。

本文实验调参过程中主要对如下表 7 中参数进行试验与调整，参数列表及取值也见表 7。

Table 7. Random forest network search parameter list
表 7. XGBoost 网格搜索参数列表

XGBoost 模型参数	参数含义	参数列表	参数取值
n_estimators	迭代次数	[10, 20, 30, …, 190, 200]	180
max_depth	树的最大深度	[10, 11, …, 19, 20]	14
min_child_weight	子节点中最小的样本权重和	[0, 1, 2, 3, 4, 5]	4
subsample	训练模型的子样本占整个样本集合的比例	[0.6, 0.7, 0.8, 0.9, 1]	0.9
colsample_bytree	构建树时对特征随机采样的比例	[0.6, 0.7, 0.8, 0.9, 1]	0.9
reg_alpha	L1 正则化系数	[1, 2, …, 9, 10]	1
reg_lambda	L2 正则化系数	[1, 2, …, 9, 10]	8
learning_rate	学习率	[0, 0.1, …, 0.9, 1]	0.6

XGBoost 模型的混淆矩阵，如下表 8 所示。

Table 8. Confusion matrix for XGBoost model
表 8. XGBoost 模型混淆矩阵

真实标签	预测标签	
	未欺诈	欺诈
未欺诈	2483	395
欺诈	131	2675

由混淆矩阵得出的其他指标如表 9 所示。

Table 9. XGBoost model classification metrics
表 9. XGboost 模型的分类指标

XGBoost	Precision	Recall	F1_score	support
0	0.95	0.86	0.90	2878
1	0.87	0.95	0.91	2806

根据上述评估指标，由 Precision 可看出我们发现模型对于欺诈情况的识别力稍逊于非欺诈。Recall 指标表明在欺诈样本中被正确识别的比例达到了 0.95。F1_score 整体而言都比较高。该模型对数据集的预测准确率为 0.9080。

3.3. LightGBM 算法

3.3.1. LightGBM 算法介绍

LightGBM 算法使用了梯度提升树的单边采样技术(GOSS)和互斥特征捆绑技术(EFB)。GOSS 根据样本梯度对样本进行采样，减少了计算所需的样本数，提高了训练速度和内存效率。而 EFB 将互斥的特征捆绑成一个单独的特征，降低了训练样本的特征数量，进一步提高了训练效率和精度，尤其适用于处理高维稀疏数据。这些创新技术使得 LightGBM 在大规模数据训练中表现出色，成为了许多实际应用场景中的首选算法。

3.3.2. LightGBM 建模

下面应用 LightGBM 算法进行建模。本文采用 Python3.7 环境下的机器学习库 sklearn 中的 en-semble 模块来建立 LightGBM 模型，以下详细介绍模型建立的过程与超参数优化的结果。

本文实验调参过程中主要对如下表 10 中参数进行试验与调整。

Table 10. LightGBM search parameter list
表 10. LightGBM 网格搜索参数列表

LightGBM 模型参数	参数含义	参数列表	参数取值
n_estimators	迭代次数	[10, 20, …, 190, 200]	180
reg_alpha	L1 正则化系数	[1, 2, …, 9, 10]	1
reg_lambda	L2 正则化系数	[1, 2, …, 9, 10]	1
learning_rate	学习率	[0, 0.1, …, 0.9, 1]	0.7

续表

min_child_samples	单个叶子节点上的最小样本数量	[18, 19, 20, 21, 22]	22
min_child_weight	子节点的最小权重	[0, 0.1, …, 0.9, 1]	0.8
max_depth	树的最大深度	[5, 6, 7, 8, 9, 10]	9
num_leaves	一棵树上的叶节点数	[10, 20, 30, 40, 50]	70

LightGBM 模型的混淆矩阵，如下表 11 所示。

Table 11. Confusion matrix for LightGBM model
表 11. LightGBMt 模型混淆矩阵

真实标签	预测标签	
	未欺诈	欺诈
未欺诈	2497	381
欺诈	106	2700

由混淆矩阵得出的其他指标如表 12 所示。

Table 12. LightGBM model classification metrics
表 12. LightGBMt 模型的分类指标

LightGBM	Precision	Recall	F1_score	support
0	0.96	0.87	0.91	2878
1	0.88	0.96	0.92	2806

根据上述评估指标，由 Precision 可看出我们发现模型对于欺诈情况的识别力逊于非欺诈。Recall 指标表明在欺诈样本中被正确识别的比例达到了 0.96。F1_score 整体而言都比较高。该模型对数据集的预测准确率为 0.9149。

3.4. Stacking 模型融合

Stacking 模型建立过程首先将上述三种算法作为初级学习器，并引入 Logistic Regression 作为次级学习器，利用上述三种算法的输出作为其输入，最终，完成 Stacking 模型的构建。

下面是三个单一模型指标与 Stacking 融合模型指标的对比(表 13)。

Table 13. Comparison of classification metrics for various models
表 13. 各模型的分类指标对比

	准确率	精确率	召回率	F1－分数	Auc
RF	0.8188	0.7665	0.9102	0.8322	0.8199
XGBoost	0.9075	0.8713	0.9533	0.9105	0.9080
LightGBM	0.9143	0.8763	0.9622	0.9172	0.9149
Stacking	0.9152	0.8820	0.9562	0.9176	0.9329

各单一模型及 Stacking 融合模型的 ROC 曲线如图 3 所示。

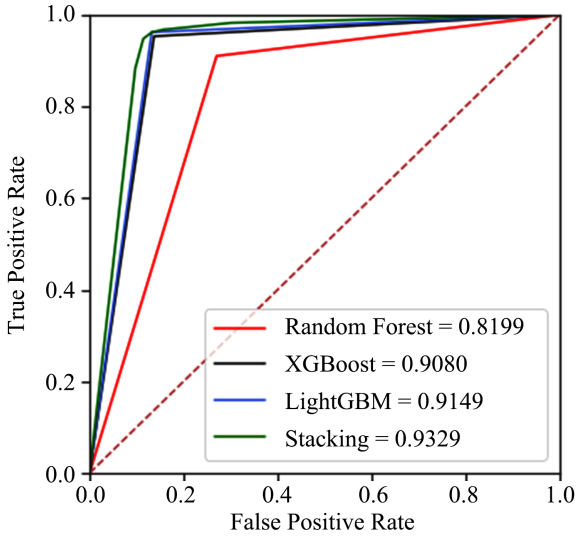

Figure 3. LightGBM ROC curve
图 3. LightGBM ROC 曲线

4. 结论

本文首先对保险欺诈识别的研究意义及研究现状进行了阐述；简要介绍了本文所用的三个模型的相关理论；将原始数据集经过数据清洗、SMOTE 经过采样等数据预处理后作为车险欺诈识别模型的输入。通过实证分析分别构建了随机森林、XGBoost、LightGBM 三种模型，并对它们的训练效果及拟合情况进行了对比分析；最后基于这三种模型利用 Stacking 法建立了融合模型，得到分类效果更好、更加稳定的模型。该融合模型对整体的预测能力比较好。

参考文献

[1] 中国银保监会. 中国银保监会关于印发实施车险综合改革指导意见的通知[EB/OL]. https://www.gov.cn/zhengce/zhengceku/2020-09/04/content_5540321.htm, 2024-02-25.

[2] 中国银保监会. 银保监会、公安部大数据反保险欺诈试点初显成效[EB/OL]. http://www.cbirc.gov.cn/cn/view/pages/ItemDetail.html?docId=947691&itemId=915&generaltype=0, 2024-02-25.

[3] Wang, Y.B. and Wei, X. (2018) Leveraging Deep Learning with LDA-Based Text Analytics to Detect Automobile Insurance Fraud. *Decision Support Systems*, **105**, 87-95. https://doi.org/10.1016/j.dss.2017.11.001

[4] Manuel, A., Mercedes, A. and Montserrat G. (1999) Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market. *Insurance*: *Mathematics and Economics*, **24**, 67-81.

[5] Stijn, V., Richard, A.D., Bart, B. and Guido, D. (2002) A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *Journal of Risk and Insurance*, **69**, 373-421. https://doi.org/10.1111/1539-6975.00023

[6] Clifton, P., Daniel, D., Vincent, L. (2002) Skewed Data in Fraud Detection: Classification of Skewed Data. *ACM SIGKDD Explorations Newsletter*, **6**, 1-19.

[7] Jesús, M.P., Javier, M., Olatz, A., Ibai, G. and José, M. (2005) Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance. *Pattern Recognition and Data Mining*, **3686**, 381-389. https://doi.org/10.1007/11551188_41

[8] 揭思杰. A 公司机动车辆保险欺诈识别指标研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2016.

[9] 李亚琪. 基于数据挖掘技术的汽车保险欺诈识别研究[D]: [硕士学位论文]. 青岛: 山东科技大学, 2018.

[10] 杜小雨. 汽车保险欺诈索赔的关联分析[D]: [硕士学位论文]. 兰州: 兰州大学, 2019.