Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 688-700
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144197

Hans汉斯

基于集成学习的汽车保险反欺诈预测研究

蔚全爱

青岛大学数学与统计学院，山东 青岛

收稿日期：2025年3月22日；录用日期：2025年4月15日；发布日期：2025年4月24日

摘　要

近年来，我国保险行业快速发展，但随之而来的车险欺诈问题日益突出，严重威胁企业运营和市场稳定。本文基于车险欺诈数据，结合事故、赔付和客户特征，分别构建随机森林、GBDT、LDA和LGBM四种单模型进行预测分析，其中LGBM模型表现最优。在此基础上，首先采用传统的Stacking集成学习方法融合上述模型，提升了整体预测性能。随后，为进一步缓解可能存在的过拟合问题，本文借鉴残差网络的思想，将第一层模型输出的预测概率与原始特征直接拼接作为第二层模型输入，实验发现模型性能有一定波动，表现不稳定。最终，本文利用PCA技术对融合特征进行优化，并动态调整融合权重，实现了模型的深度融合，最终预测准确率提高至0.870，显著提升了车险欺诈行为的识别效果，为保险公司的风险控制提供了有效支持。

关键词

集成学习，保险欺诈预测，机器学习

Research on Anti-Fraud Prediction of Auto Insurance Based on Ensemble Learning

Quanai Yu

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Mar. 22nd, 2025; accepted: Apr. 15th, 2025; published: Apr. 24th, 2025

Abstract

In recent years, China's insurance industry has experienced rapid growth; however, automobile insurance fraud has concurrently become increasingly prevalent, posing significant threats to business operations and market stability. In this study, automobile insurance fraud data, including accident information, compensation records, and customer characteristics, are used to develop four individual prediction models: Random Forest, GBDT, LDA, and LGBM. Among these, the LGBM model

文章引用: 蔚全爱. 基于集成学习的汽车保险反欺诈预测研究[J]. 应用数学进展, 2025, 14(4): 688-700.
DOI: 10.12677/aam.2025.144197

demonstrated superior performance. Subsequently, a traditional stacking ensemble approach was adopted to integrate these base models, achieving enhanced prediction accuracy. To further address potential overfitting issues, this study initially borrowed from residual network methodology, directly concatenating original features with predicted probabilities from the first-layer models as input for the second-layer model. However, experiments revealed inconsistent improvements, with performance fluctuations. Ultimately, PCA-based dimensionality reduction and a dynamic weighting strategy were introduced, achieving a deeper fusion of features. This improved ensemble strategy raised prediction accuracy to 0.870, significantly enhancing automobile insurance fraud detection performance and providing valuable support for risk management within insurance companies.

Keywords

Ensemble Learning, Insurance Fraud Prediction, Machine Learning

1. 引言

近年来，保险行业的高速发展与大数据、人工智能技术的深度融合，为解决保险领域复杂问题提供了新的路径。随着生活水平的提高，人们越来越注重相关保险的购买，我国保险行业蓬勃发展，已成为世界第二大保险市场[1] [2]。但与此同时，保险欺诈问题显露出严峻的形势。每年保险行业都面临着保险欺诈带来的重大损失，其严重影响了市场秩序，而随着我国逐年增长的汽车持有量，车险欺诈成为了不容忽视的问题[3]。由于数据壁垒和有效反欺诈工具缺乏，车险企业难以有效利用数据开展欺诈防控，导致严重经济损失和保费上涨，威胁市场健康发展。因此，开展有效的车险欺诈识别研究，对降低企业风险、维护市场秩序和社会稳定具有重要意义。

目前，借助大数据和机器学习来应对保险行业中日益凸显的问题，已成为备受关注的研究焦点。例如，Artis 等人采用了 Logistic 回归模型来识别机动车辆保险欺诈行为[4]；Nabraw 等人则利用利用机器学习方法识别医疗保险索赔欺诈的模型[5]。闫春等人提出了一种基于群优化算法的随机森林组合分类器，用于提取机动车辆保险欺诈识别的关键特征。与传统的单一随机森林算法相比，该方法显著提高了模型在欺诈识别任务中的精确度与稳健性[6]。

在上述研究的基础上，本文使用真实保险欺诈数据，借助机器学习算法，通过数据预处理、特征工程以及多种 Stacking 集成模型构建了一个预测准确的保险反欺诈模型。该模型不仅能够有效识别保险欺诈行为，还能进一步分析关键影响因素，从而为保险公司和每一位投保人提供可靠的决策支持。

2. 数据处理

2.1. 数据来源

本文数据来源于阿里云天池平台，包含某保险公司客户信息和索赔记录共 700 条、38 个特征，涵盖保单、被保人、事故、索赔等信息类别，事故敏感信息已脱敏处理。其中，“policy_csl”表示汽车保险中的组合单一赔付限制，如“500/1000”代表单次事故赔付上限为 500,000 美元，保单期间累计赔付上限为 1,000,000 美元。部分特征介绍表 1。

Table 1. Table of instructions for some data indicators
表 1. 部分数据指标说明表

变量名称	变量解释	变量名称	变量解释
policy_id	保险编号	umbrella_limit	保险责任上限
age	年龄	insured_zip	被保人邮编
customer_months	成为客户的时长，以月为单位	insured_sex	被保人的性别
policy_bind_date	保险绑定日期	insured_education_level	被保人的学历
policy_csl	组合单一限制	capital-gains	资本收益

2.2. 特征工程

首先对数据缺失值进行分析，如图 1 所示，缺失集中在“collision_type”(<20%)及“property_damage”“police_report_available”(均 >30%)三个字段。通过构建缺失指示变量，经卡方检验判断这些缺失值符合完全随机缺失(MCAR)特性，即缺失的发生与数据本身或其他变量无关，直接删除可能导致样本分布偏差[7]。也就是说，缺失数据的概率在所有观测中都是相同的，不存在任何系统性模式，直接删除可能导致样本分布偏差；同时，相关性分析显示三者与目标变量相关性显著，因此最终采用众数填充缺失值，以平衡数据完整性与特征有效性。

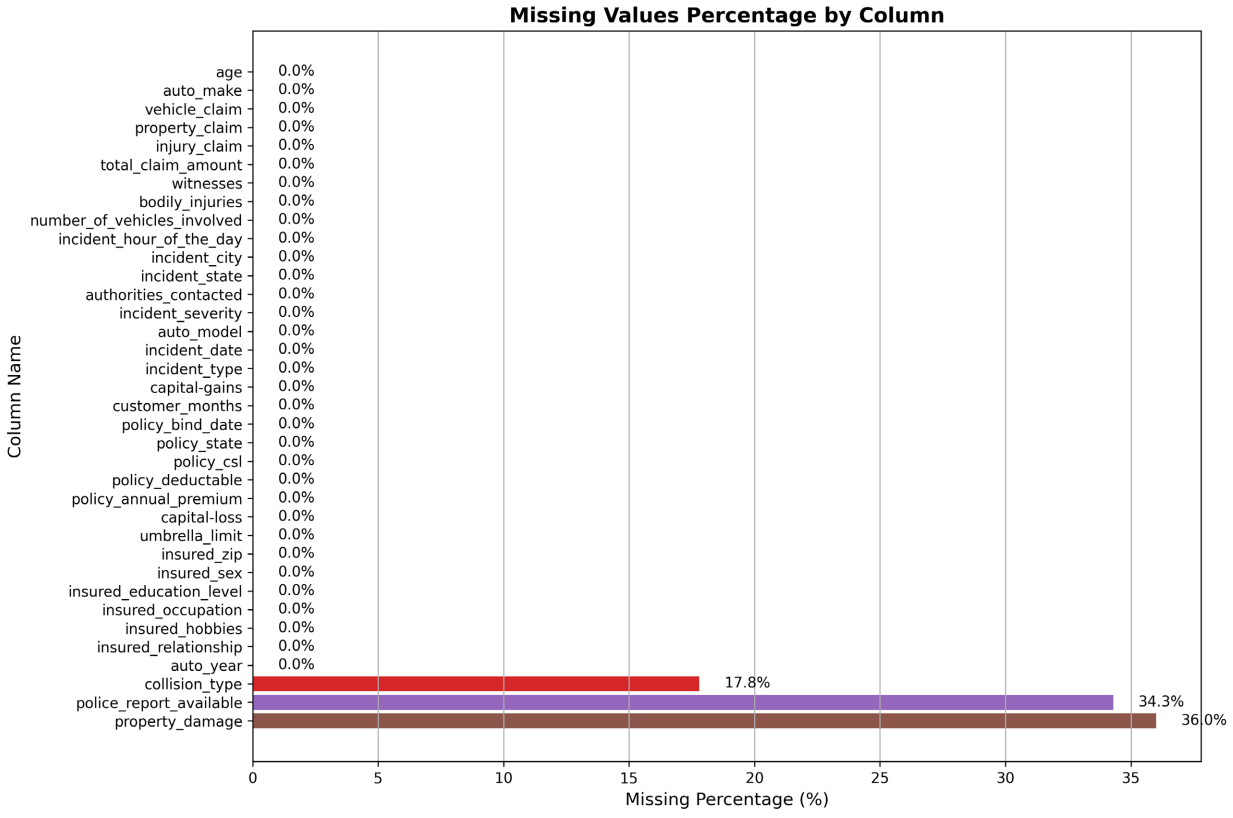

Figure 1. Visualization of missing values for all variables
图 1. 全变量缺失值可视化

在完成缺失值填补后，为了进一步提高预测和分析的准确性，我们对原始数据进行了特征工程与优

化处理。首先，针对汽车保险责任限制特征“policy_csl”，利用“\”分隔符将其拆分为“policy_BI”和“policy_PD”两个新特征，分别表示身体伤害赔付限额和财产损失赔付限额，以便更细致地探讨各责任限额与目标变量之间的相关性。其次，为了将类别型变量转化为数值型变量，我们综合考虑了变量中类别数的多寡：对于具有十个以上不同类别的特征，采用均值编码；其余特征则采用整数编码，从而实现了不同编码方式的合理搭配。最后，针对时间特征“policy_bind_date”和“incident_date”，我们计算了事故发生日期与保险绑定日期之间的差值，生成新变量“delta_time”，并从“incident_date”中提取出月份信息生成“picked_month”，以捕捉季节性变化。完成这些操作后，剔除原有的“policy_bind_date”和“incident_date”，避免变量的高相关性。

在上述数据处理后，简单通过相关性分析筛选出高相关性变量见图 2。从通过相关性分析(见图 3)，发现“age”与“customer_months”高度相关，保留“customer_months”能更直观反映客户生命周期价值；同时，“total_claim_amount”整合了各项赔付数据，保留总金额可避免多重共线性问题。

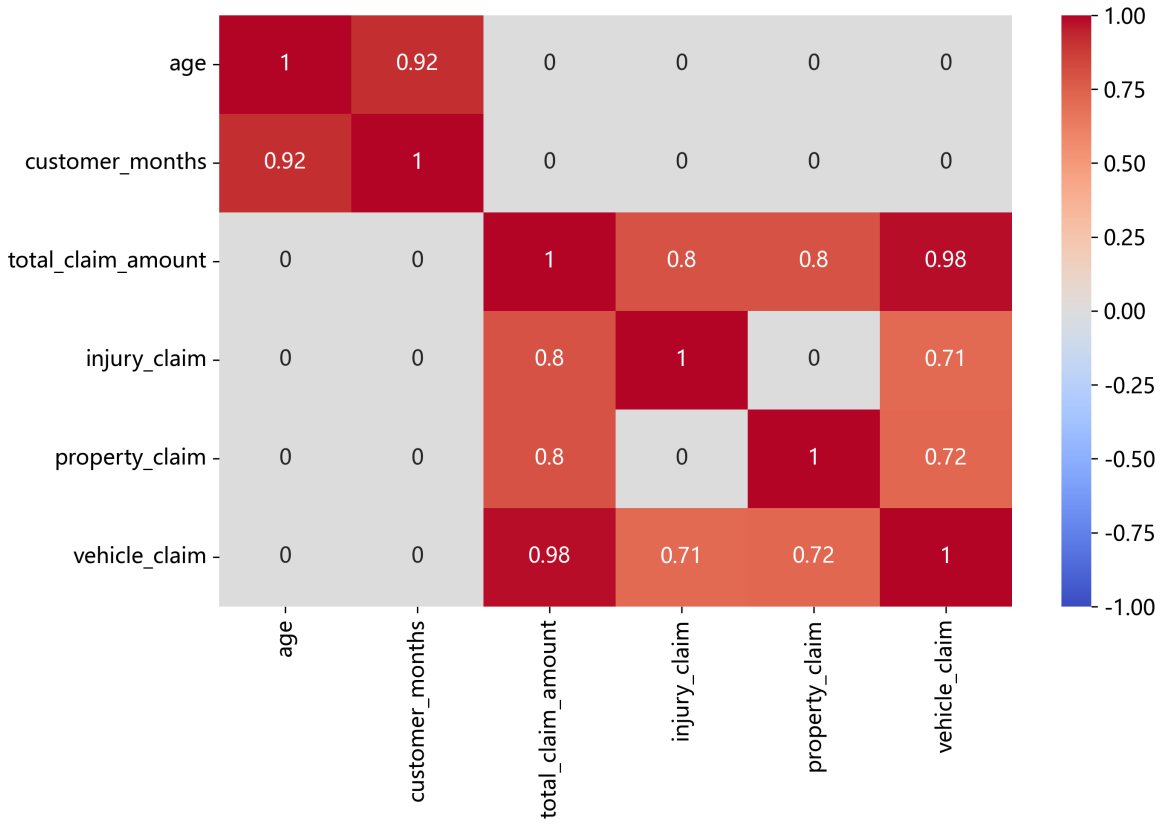

Figure 2. Heat map of highly correlated variables
图 2. 高相关变量热力图

经过上述特征工程处理后，我们获得了 40 个特征变量。图 3 展示了保险欺诈标签的分布情况，其中欺诈客户仅占 25.9%，而未欺诈客户达到 74.1%。由于类别严重不平衡，直接基于原始数据训练模型可能会使其偏向于多数类，忽略少数类信息。为解决这一问题，我们采用 ADASYN 采样方法[8]。ADASYN 方法通过自适应生成新的少数类样本，来扩充欺诈客户的数据量，从而在保持整体样本数量基本不变的前提下实现类别平衡。通过 ADASYN 采样方法，我们获得一个平衡的数据集，其中欺诈客户样本数量为 560 个，非欺诈客户样本数量为 519 个。

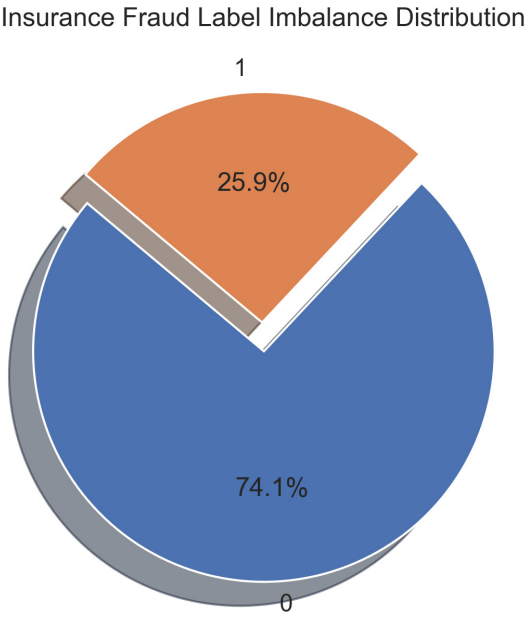

Figure 3. Distribution of insurance fraud labels
图 3. 保险欺诈标签分布图

2.3. 特征分析

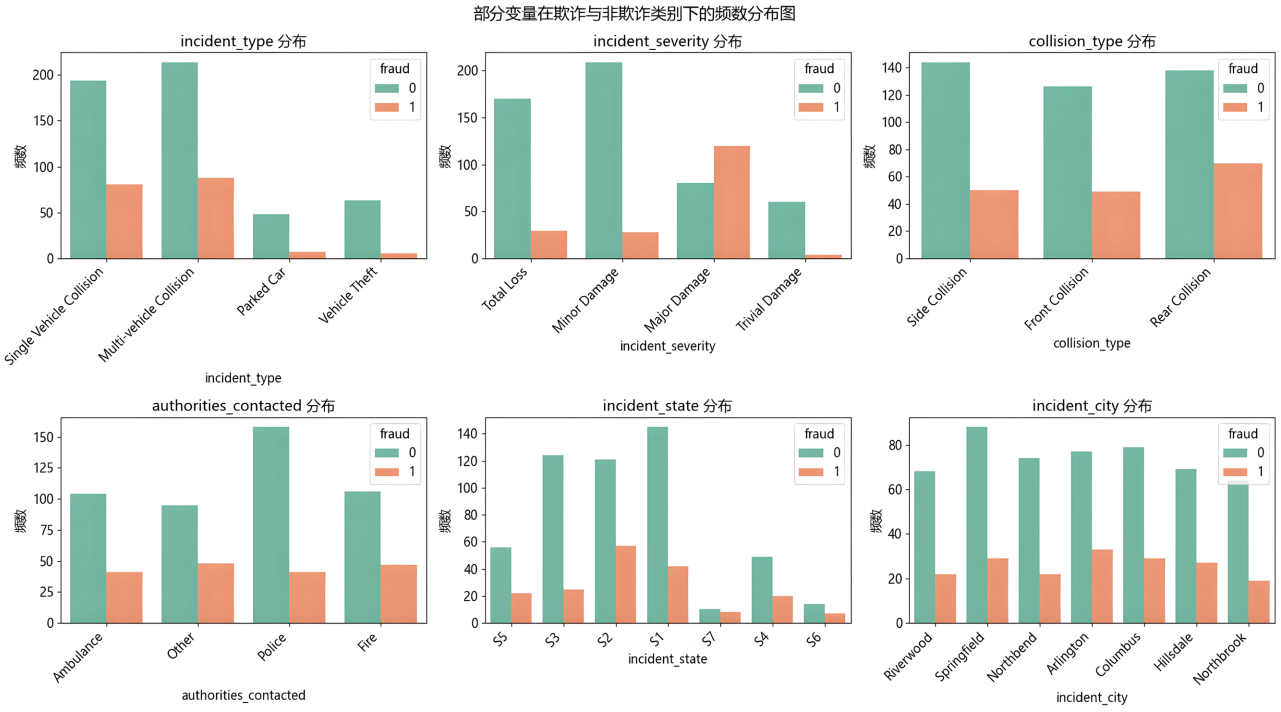

Figure 4. Frequency distribution diagram of some discrete variables under the categories of fraud and non-fraud
图 4. 部分离散变量在欺诈与非欺诈类别下的频数分布图

针对部分离散型特征(包括“incident_type”“incident_severity”“collision_type”“authorities_contacted”“incident_state”“incident_city”)，本研究通过频数分布对比分析，重点考察了欺诈(fraud = 1)与非欺诈

(fraud = 0)类别的分布差异。观察上图 4，在“incident_severity”(事故严重性)特征 major_damage (重大损害)中，欺诈类别下的频数高于非欺诈类别下频数，表明欺诈者可能更倾向于选择重大事故来进行欺诈行为，这种情况下，保险公司可能需要支付高额的赔偿金，为防止此类情况发生，在面对此类型事故时，保险公司需多调查，多处理，防止欺诈行为的发生。

3. 模型建立

3.1. 评价指标

为了比较不同模型之间的表现差异，选择以下几种评价指标，为便于描述，相应定义见表 2。

Table 2. Confusion matrix
表 2. 混淆矩阵

		真实值	
		正例	负例
预测值	正例	TP	FP
	负例	FN	TN

在此基础上，给出如下评价指标定义：

准确率(Accuracy)：所有正确预测样本占总样本的比例，反映整体判别能力。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

精确率(Precision)：预测为正例的样本中实际为正例的比例，衡量模型误判成本。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

召回率(Recall)：实际为正例的样本中被正确识别的比例，衡量模型漏检风险。

$$\text{Recall} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

F1 分数：精确率与召回率的调和平均数，综合评估模型在类别不平衡数据中的稳健性。

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC 曲线反映了模型在不同阈值下的表现，其横轴为假阳性率(FPR)，纵轴为真正阳性率(TPR)。理想的情况是 TPR 越大、FPR 越小，即期望 TPR = 1 且 FPR = 0，因此 ROC 曲线越接近左上角(远离 45˚对角线)表明模型效果越好。AUC 则表示 ROC 曲线下的面积，数值越大说明模型在区分正负样本方面性能越优。

3.2. 简单预测模型

本文集中研究预测是否存在保险反欺诈，基于此，我们建立了如下简单预测模型，主要构建了随机森林分类模型、LDA 分类模型、GBDT 模型以及 LightGBM 模型。

1) 随机森林模型

随机森林(Random Forest, RF)是一种基于决策树的集成学习方法，通过构建多棵彼此独立的决策树并

结合它们的预测结果来提高模型的准确性和泛化能力[9]。其核心思想是“集体智慧”，即多个弱学习器(决策树)通过投票或平均的方式共同决策，从而超越单一模型的性能。

2) LDA 模型

线性判别分析(Linear Discriminant Analysis, LDA)是一种监督学习的降维技术，也是一种特征提取算法[10]。LDA 模型的基本思想是将样本从高维空间投影到低维空间，使得类间方差最小，类外方差最大。根据投影点的位置确定新样本的类别。对于二分类线性判别分析，假设投影直线为向量 w，对于任意样本 x_i，其在 w 上的投影为 $w^{\mathrm{T}}x_i$。两个类别的中心点 μ_0，μ_1 经过投影后的点为 $w^{\mathrm{T}}\mu_0$，$w^{\mathrm{T}}\mu_1$。因为线性判别分析是想让类外间距尽可能的大，即最大化 $\left\|w^{\mathrm{T}}\mu_0 - w^{\mathrm{T}}\mu_1\right\|_2^2$，并且其类内间距尽可能的小，可以转化成最小化 $w^{\mathrm{T}}\Sigma_0 w + w^{\mathrm{T}}\Sigma_1 w$，根据上面的分析，LDA 模型的优化目标为：

$$\underbrace{\operatorname{argmax}}_{w} J(w) = \frac{\left\|w^{\mathrm{T}}\mu_0 - w^{\mathrm{T}}\mu_1\right\|_2^2}{w^{\mathrm{T}}\Sigma_0 w + w^{\mathrm{T}}\Sigma_1 w} = \frac{w^{\mathrm{T}}(\mu_0-\mu_1)(\mu_0-\mu_1)^{\mathrm{T}}w}{w^{\mathrm{T}}(\Sigma_0+\Sigma_1)w} = \frac{w^{\mathrm{T}}S_b w}{w^{\mathrm{T}}S_w w}$$

其中，$S_b = w^{\mathrm{T}}(\mu_0-\mu_1)(\mu_0-\mu_1)^{\mathrm{T}}$ 为类间散度矩阵；$S_w = \Sigma_0 + \Sigma_1$ 为类内散度矩阵。

3) GBDT 模型

GBDT (Gradient Boosting Decision Tree，梯度提升决策树)是一种基于 Boosting 的集成学习模型，通过迭代地构造一组弱学习器(决策树)来逐步优化预测结果。每一棵新树拟合的是前一棵树的残差(即损失函数的负梯度)，并将多棵树的预测结果累加作为最终输出。

4) LightGBM 模型

LightGBM (Light Gradient Boosting Machine)是一种基于 GBDT 的高效实现，显著提高了计算效率[11]。与传统 GBDT 不同，LightGBM 采用基于直方图的决策树算法，将连续数据离散化为直方图，从而大幅降低计算复杂度和内存消耗。此外，其采用基于叶子分割(leaf-wise)的树生长策略，比传统的按层生长更能降低训练误差，使得模型在大规模数据集上能更快收敛。通过梯度提升的方式，LightGBM 在每次迭代中都对上一次的残差进行拟合，从而不断改进模型预测。

3.3. 基于传统 Stacking 的集成学习模型

Stacking 是一种集成学习策略，通常采用分层嵌套的模型架构。最底层(特征提取层)中，同时运行多个不同的基模型，例如支持向量机、随机森林和逻辑回归等，每个模型都会生成预测结果。随后，这些预测输出被整合成新的特征集合，并作为输入供顶层模型训练使用。顶层模型利用这些特征对数据标签进行最终预测，从而构建出完整的 Stacking 集成学习框架。下图 5 展示了这一学习流程的示意图：

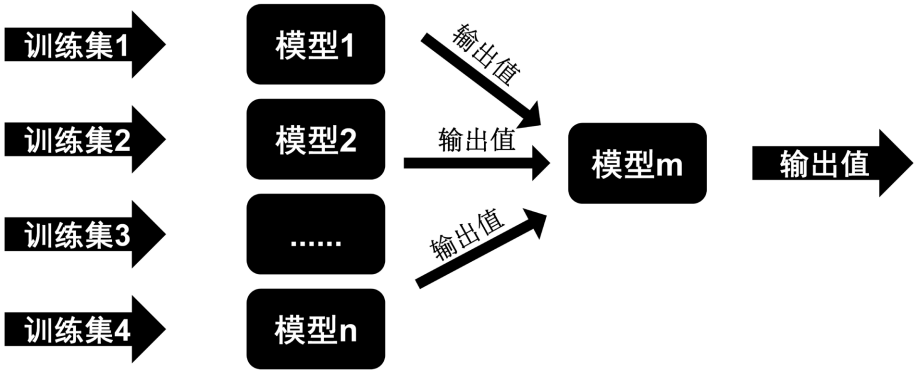

Figure 5. The Stacking ensemble learning framework
图 5. Stacking 集成学习框架

3.4. 仿照 res-net 改进传统 Stacking 的集成学习

传统的 Stacking 的集成学习是将每个模型的输出作为下一层的输入，但是这样不可避免的会造成过拟合，从 res-net (残差网络)构造得到启发，在传统基学习器输出传递的基础上，通过跨层残差连接将原始输入特征与模型预测结果融合，形成双向信息传递通道。该方法将基学习器生成的概率特征与原始特征拼接后共同输入元学习器，既保留了数据的底层分布信息以缓解梯度衰减，又通过特征复用增强非线性表达能力。第二层采用逻辑回归、KNN 等低复杂度模型对融合特征进行决策，利用简单模型结构抑制过拟合风险，同时维持了模型选择的灵活性。该设计通过残差结构的梯度短路效应，有效解决了深层集成中信息丢失与训练不稳定的核心矛盾。下图 6 是其学习示意图：

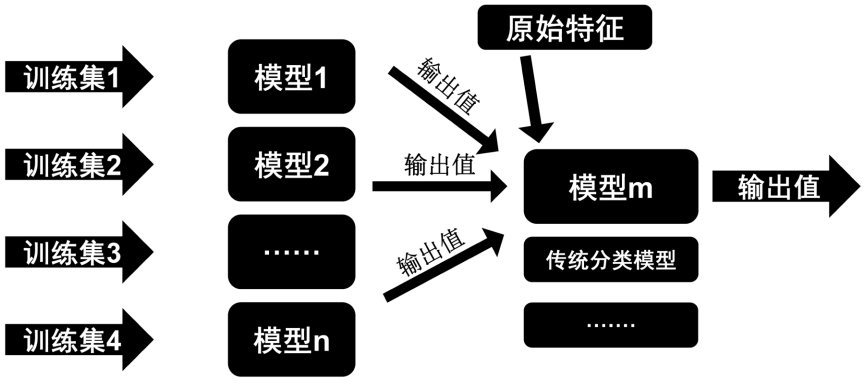

Figure 6. Improve the traditional Stacking ensemble learning framework by imitating the res-net
图 6. 仿照 res-net 改进传统 Stacking 的集成学习框架

4. 模型的评价与分析

4.1. 简单预测模型的评价与分析

本研究旨在预测保险反欺诈的可能性。为此，我们构建了一个简单的预测模型体系，包括随机森林分类模型、LDA 分类模型、LGBM 集成模型和 GBDT 模型。我们通过网格搜索确定各模型的最优参数，并将数据集划分为训练集和测试集，在训练集上采用 K 折交叉验证对用户是否存在保险欺诈行为进行预测。随后，我们整合对比了这四个模型在五个评价指标下的表现。模型对比结果见表 3。

Table 3. Performance metrics of four simple prediction models
表 3. 四种简单预测模型的性能指标

模型	Accuracy	precision	Recall	F1 分数	AUC
随机森林	0.797107	0.790169	0.797107	0.785970	0.858699
GBDT	0.805642	0.809511	0.805642	0.805919	0.858798
LDA	0.804217	0.796984	0.804217	0.791506	0.835980
LightGBM	0.812813	0.814727	0.812813	0.811280	0.853405

我们对比了四种模型在汽车保险反欺诈预测中的表现，包括随机森林、GBDT、LDA 和 LightGBM。实验结果表明 LightGBM。在准确率(0.813)、精确率(0.815)、召回率(0.813)和 F1 分数(0.811)上均表现最优，表明 LightGBM 在对车险欺诈和非欺诈识别方面具有显著优势，其高精确率意味着模型误判正常保单为欺诈的比例较低，可减少因错误拦截带来的客户投诉；而较高的 F1 分数则反映其在漏检(未识别真

实欺诈)与误判之间的平衡能力较强，这对保险业务中控制风险损失和客户体验的权衡至关重要。但总体而言，模型之间差距不大。

4.2. 传统 Stacking 集成模型的评价与分析

在 Stacking 集成框架中，第一层基模型(随机森林、GBDT、LDA 和 LightGBM)输出的预测概率(即样本被判定为欺诈类别的概率)被构造为元特征矩阵，作为第二层元模型的输入。具体而言，每个样本经过第一层模型推理后，将生成四维概率向量

$$\boldsymbol{p} = \left[p_{RF}, p_{GBDT}, p_{LDA}, p_{LGB} \right]$$

作为元特征，供第二层的元模型(meta-model)进一步学习和决策。生成的部分新指标见表 4。

Table 4. New indicators of the training set based on four simple prediction models
表 4. 基于四种简单预测模型的训练集新指标

	p_{RF}	p_{GBDT}	p_{LDA}	p_{LGB}
0	0.17	0.000189	0.031800	0.029088
1	0.16	0.002398	0.019811	0.022291
2	0.17	0.003263	0.042782	0.036061
3	0.95	0.996266	0.987562	0.951375

将上述指标打入二层的元模型中，二层元模型我们考虑了随机森林模型、朴素贝叶斯(Naive Bayes)、LDA、逻辑回归、支持向量机(SVC)、K 近邻(KNN)、GBDT、AdaBoost、LightGBM 和极端随机树(ExtraTrees)。这些模型分别利用各自独特的算法优势对第一层简单模型的输出进行融合，从而捕捉不同模型间的互补信息。性能表现见表 5。

Table 5. Performance evaluation of conventional Stacking models
表 5. 传统 Stacking 模型性能指标

模型	Accuracy	precision	Recall	F1 分数	AUC
随机森林	0.854938	0.827027	0.910714	0.866856	0.919395
Naive Bayes	0.848765	0.825137	0.898810	0.860399	0.931891
LDA	0.851852	0.822581	0.910714	0.864407	0.933761
逻辑回归	0.848765	0.825137	0.898810	0.860399	0.933646
SVC	0.854938	0.820106	0.922619	0.868347	0.862599
KNN	0.864198	0.848315	0.898810	0.872832	0.908387
GBDT	0.848765	0.825137	0.898810	0.860399	0.922447
AdaBoost	0.802469	0.788889	0.845238	0.816092	0.800824
LGBM	0.842593	0.834286	0.869048	0.851312	0.910981
ExtraTrees	0.854938	0.823529	0.916667	0.867606	0.918536

我们仅用第一部分生成的 4 个指标作为下一层的输入，第二层的模型分别选用 10 种模型进行比较，从上表中可以看出，集成学习后的结果普遍比单一模型下的效果要好，特别是 KNN 结果表现最好，且一些线性模型(逻辑回归和 Naive Bayes)在 AUC 指标上显示出较强的区分能力。说明我们的集成学习是有

意义的，可以在第二层采用较为简单的模型就可以达到较好的结果，但是也不可避免的存在过拟合问题。

4.3. 仿照 res-net 改进传统 Stacking 的集成学习

为解决过拟合问题，借鉴残差网络的思想，我们不仅将基础模型结果作为指标引入第二层，同时引入原指标，共 44 项指标，并通过堆叠融合构建第二层元模型，来捕捉基础模型预测误差与原始特征之间的残差信息。这样的设计既保留了原始数据的丰富信息，又利用残差学习机制对基础模型的不足进行校正，从而进一步提升整体预测精度。为进一步改进集成模型，从残差网络得到启发，构建 Stacking 融合模型。最终模型性能表现见表 6。

Table 6. Performance metrics of the traditional Stacking improved by imitating res-net
表 6. 仿照 res-net 改进传统 Stacking 的性能指标

模型	Accuracy	precision	Recall	F1 分数	AUC
随机森林	0.842593	0.803109	0.922619	0.858726	0.926702
Naive Bayes	0.620370	0.596567	0.827381	0.693267	0.609165
LDA	0.854938	0.823529	0.916667	0.867606	0.935325
逻辑回归	0.595679	0.591133	0.714286	0.646900	0.575511
SVC	0.521605	0.520124	1.000000	0.684318	0.514843
KNN	0.617284	0.601852	0.773810	0.677083	0.664339
GBDT	0.854938	0.823529	0.916667	0.867606	0.934982
AdaBoost	0.817901	0.797814	0.869048	0.831909	0.815934
LGBM	0.851852	0.815789	0.922619	0.865922	0.932082
ExtraTrees	0.854938	0.820106	0.922619	0.868347	0.930899

从表 5 和表 6 的对比来看，改进方案对模型的性能影响呈现两极分化：

优势模型：LDA、GBDT、ExtraTrees 在 AUC 和部分指标上提升显著；劣势模型：Naive Bayes、逻辑回归、SVC、KNN 性能大幅下降。

在此基础上，我们对简单的特征拼接方式进行了改进。具体而言，对原始特征和新特征分别进行标准化以保证融合特征时的尺度一致性。通过对原始特征采用 PCA (主成分分析)降维，基于以下考虑：一是维度对齐需求。将原始特征降维至 4 维，与第一层模型输出的 4 维概率向量 $\boldsymbol{p}=[p_{RF}, p_{GBDT}, p_{LDA}, p_{LGB}]$ 保持维度匹配，从而实现更加有效的动态融合计算。二是经 PCA 分析可知，前 4 个主成分累计方差贡献率已超过 90%，表明降维后的特征已保留了绝大部分的原始信息，选择 4 维是合理且有效的。我们将降维后的原始特征记为 X_{PCA}，并将原始训练数据及对应的新特征划分为训练子集和验证集，寻找最优融合权重 α，然后利用下式构建新的融合特征[12]-[14]：

$$X_{new} = X_{PCA} + \alpha \boldsymbol{p}$$

最终模型性能表现见表 7。

Table 7. Performance metrics of PCA-based dynamic weighted residual fusion Stacking model
表 7. 基于 PCA 的动态加权残差融合 Stacking 模型性能指标

模型	Accuracy	precision	Recall	F1 分数	AUC
随机森林	0.861111	0.851429	0.886905	0.868805	0.904647

续表

Naive Bayes	0.861111	0.836066	0.910714	0.871795	0.905601
LDA	0.870370	0.846154	0.916667	0.880000	0.904151
逻辑回归	0.861111	0.839779	0.904762	0.871060	0.903083
SVC	0.861111	0.832432	0.916667	0.872521	0.901213
KNN	0.839506	0.818681	0.886905	0.851429	0.881563
GBDT	0.870370	0.853933	0.904762	0.878613	0.907853
AdaBoost	0.864198	0.844444	0.904762	0.873563	0.900183
LGBM	0.851852	0.844828	0.875000	0.859649	0.901442
ExtraTrees	0.848765	0.836158	0.880952	0.857971	0.906174

对原始特征和新特征(基模型概率输出)分别进行标准化，消除特征间量纲差异，使不同来源的特征在融合时处于同一数值量级。表 7 中原本因量纲敏感而表现波动的模型(如 Naive Bayes、逻辑回归)准确率显著提升，验证了标准化对模型稳定性的增强作用。比较结果见图 7。

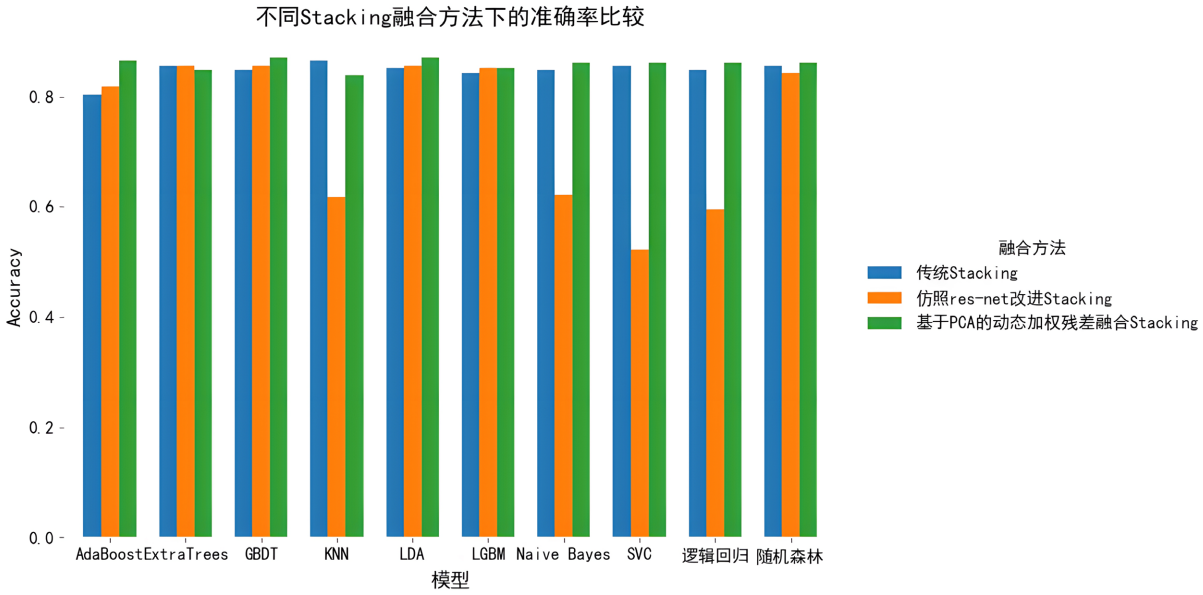

Figure 7. Performance comparison under different Stacking fusion methods
图 7. 不同 Stacking 融合方法下的性能比较

图 7 展示了不同 stacking 方法的性能表现，除了仅采用简单拼接特征的改进 stacking 方式导致部分模型性能下降外，其余集成方法均优于传统简单模型(见表 3)，且通过 PCA 降维后进行残差融合的方法，在大部分模型中均显著提高了准确率，相较于传统集成方法表现更为优越。

4.4. 特征重要性

特征重要性衡量的是每个特征在决策树分裂过程中带来的信息增益，信息增益越大，特征越重要。通过计算特征重要性，我们可以直观地了解哪些特征对预测结果影响最大，从而为模型解释提供定量依据。

通过筛选前十个特征的重要性排序图(图 8)可以看出，总索赔金额是最核心的指标，该指标反映了保

险客户提出的总赔偿金额，数值越高，意味着赔付金额可能异常。除此之外，时间间隔衡量事故发生相对于保险生效时间的延迟情况，事故发生时间距保险购买时间越近，越可能涉及欺诈行为等。

基于上述特征重要性分析可知，总索赔金额、事故时间间隔、客户所在地和事故严重程度等因素对保险欺诈行为的识别具有关键作用。高额理赔金额、短期投保即发生事故、夜间发生的案件以及特定区域或客户群体存在更高的欺诈风险。保险公司应利用这些关键特征建立客户风险画像，优化理赔审核机制，加强事前预警和风险筛查能力，从而有效降低欺诈风险，保障公司稳健运营并促进保险市场的健康发展。

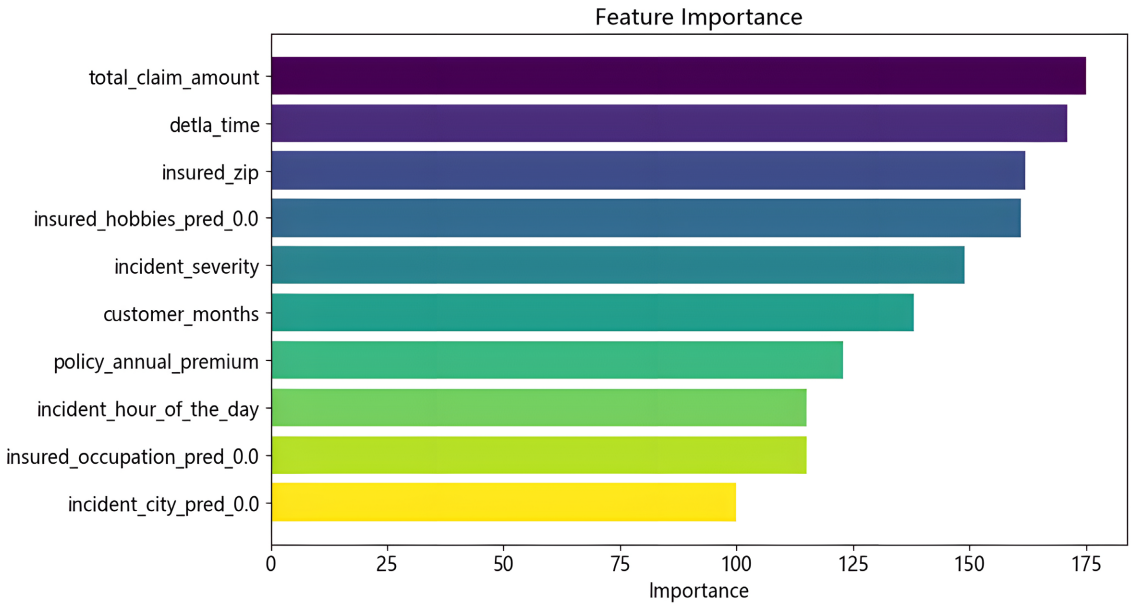

Figure 8. Feature importance ranking
图 8. 特征重要性排序

5. 结论

本文基于车险反欺诈数据，综合事故信息、理赔记录以及客户特征，建立了随机森林、GBDT、LDA和 LGBM 四个单模型进行预测分析，其中 LGBM 表现最佳。为进一步提高欺诈识别的准确性，我们采用 Stacking 集成学习方法，将基础模型输出的预测概率作为元特征输入次级模型，有效提高了整体模型性能。但传统 Stacking 存在一定过拟合风险，为此我们尝试引入残差网络思想，直接拼接基础模型预测概率与原始特征，然而该方法表现不稳定，预测效果时有波动。最终，我们原始特征进行 PCA 降维处理，采用动态权重融合策略与原特征融合，成功解决了模型性能不稳定问题，显著提升了预测的稳健性与准确性，可以更有效地识别欺诈行为。

通过模型分析，可识别高赔付金额与异常时间间隔等关键风险指标。对此，为保障保险公司的稳健经营，建议保险公司关注高风险指标：如高赔付金额和异常时间间隔等，同时通过模型预测，结合实时数据，构建全流程预警体系。

致　谢

在本文的写作过程中，得到了许多人的无私帮助以及支持，对此我们深表感激。正是由于他们的指导与鼓励，本文才能顺利完成。

参考文献

[1] 陈辉, 董洪斌. 中国保险业可持续发展报告 2022 [M]. 北京: 中国经济出版社, 2022.

[2] 孙祁祥. 中国保险业发展报告[M]. 北京: 北京大学出版社, 2019.

[3] 魏丽, 戴稳胜, 陈泽. 现代保险制度建设[M]. 北京: 中国人民大学出版社, 2024.

[4] Artís, M., Ayuso, M. and Guillén, M. (1999) Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market. *Insurance*: *Mathematics and Economics*, **24**, 67-81. https://doi.org/10.1016/s0167-6687(98)00038-9

[5] Nabrawi, E. and Alanazi, A. (2023) Fraud Detection in Healthcare Insurance Claims Using Machine Learning. *Risks*, **11**, Article No. 160. https://doi.org/10.3390/risks11090160

[6] 闫春, 李亚琪, 孙海棠. 基于蚁群算法优化随机森林模型的汽车保险欺诈识别研究[J]. 保险研究, 2017(6): 114-127.

[7] Heymans, M.W. and Twisk, J.W.R. (2022) Handling Missing Data in Clinical Research. *Journal of Clinical Epidemiology*, **151**, 185-188. https://doi.org/10.1016/j.jclinepi.2022.08.016

[8] Brandt, J. and Lanzén, E. (2021) A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification.

[9] Hu, J. and Szymczak, S. (2023) A Review on Longitudinal Data Analysis with Random Forest. *Briefings in Bioinformatics*, **24**, bbad002. https://doi.org/10.1093/bib/bbad002

[10] Yu, D. and Xiang, B. (2023) Discovering Topics and Trends in the Field of Artificial Intelligence: Using LDA Topic Modeling. *Expert Systems with Applications*, **225**, Article ID: 120114. https://doi.org/10.1016/j.eswa.2023.120114

[11] Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., *et al*. (2021) LightGBM: Accelerated Genomically Designed Crop Breeding through Ensemble Learning. *Genome Biology*, **22**, 1-24. https://doi.org/10.1186/s13059-021-02492-y

[12] Kurita, T. (2021) Principal Component Analysis (PCA). In: Ikeuchi, K., Ed., *Computer Vision*: *A Reference Guide*, Springer International Publishing, 1013-1016. https://doi.org/10.1007/978-3-030-63416-2_649

[13] Yu, W., Zhang, Q. and Li, W. (2025) High-Dimensional Projection-Based ANOVA Test. *Journal of Multivariate Analysis*, **207**, Article ID: 105401. https://doi.org/10.1016/j.jmva.2024.105401

[14] 王宏漫, 欧宗瑛. 采用 PCA/ICA 特征和 SVM 分类的人脸识别[J]. 计算机辅助设计与图形学学报, 2003, 15(4): 416-420+431.