

基于 LightGBM 的酒驾换驾风险预测

赵瑞阳¹ 刘聪慧¹

ZHAO Ruiyang LIU Conghui

摘要

在车险理赔场景中,酒驾换驾案件因肇事者虚报出险时间、伪造现场等行为,导致调查效率低且损失较大,亟需科学的风险预测模型以提升管控效能。针对此问题,文章提出一种基于机器学习的酒驾换驾风险预测模型。研究通过分析 2024 年 4 万多条已决赔案数据(含 172 个字段),将风险识别转化为二分类任务(0=无风险,1=有风险)。针对正负样本分布极不平衡(约 345:1)的挑战,通过特征工程构建了“报案-出险时间间隔”“驾驶员与被保险人身份一致性”关键衍生特征,并创新性地利用朴素贝叶斯算法预测驾驶员性别,以补全关键信息维度。为确保模型鲁棒性,采用了数据标准化、5 折交叉验证及早停策略进行优化。在 GBDT、XGBOOST、LightGBM 三种集成学习模型的对比实验中,LightGBM 表现最佳,其准确率(Accuracy)为 84.56%,精确率(Precision)为 94.73%,召回率(Recall)为 89.82%, F_1 分数(F_1 -score)为 0.922 0,显著优于对比模型及传统人工筛查方法。该模型可有效识别高风险案件,为车险理赔环节降低调查成本、提升风险管理效率提供数据驱动的解决方案,对打击保险欺诈具有重要的现实意义和应用价值。

关键词

酒驾换驾; 风险预测; 机器学习; 特征工程; LightGBM; 理赔管控; 数据挖掘; 模型评估

doi: 10.3969/j.issn.1672-9528.2025.12.007

0 引言

酒驾换驾是指肇事者在酒后驾驶车辆发生交通事故后,为逃避法律责任和保险拒赔,通过更换驾驶员、虚报出险时间、伪造事故现场等方式进行保险欺诈的行为^[1]。这类行为

在车险理赔场景中属于典型的保险欺诈范畴,不仅严重违反道路交通安全法规,还会导致保险公司面临巨大的理赔损失与调查压力。从行业现状来看,酒驾换驾案件具有“损失金额大、调查时效要求高、现场勘查前难以察觉”的显著特点,如某分公司数据显示,2024 年其受理的车险理赔案件中,共发现 204 起酒驾换驾疑点案件,其中仅 5 起最终因证据确凿成功拒赔,涉及拒赔金额达 6.8 万元。

1. 中国太平洋财产保险股份有限公司营运山东分中心
山东济南 250000

- [24] WANG B Y, ZHANG L B, WEN L Y, et al. Towards real world prohibited item detection: a large-scale X-ray benchmark[C] // 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway:IEEE,2021:5392-5401.
- [25] WEN L H, JO K H. Three-attention mechanisms for one-stage 3-D object detection based on LiDAR and camera[J].IEEE transactions on industrial informatics,2021, 17(10):6655-6663.
- [26]WANG C Y, YE H I H, MARK LIAO H Y. YOLOv9:Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Berlin:Springer, 2025: 1-21.

【作者简介】

张自勖(2000—),男,山东济南人,硕士研究生,研究方向:计算机视觉。

罗婷(1989—),通信作者(email:luoting@ppsuc.edu.cn),女,山东聊城人,博士研究生,讲师、硕士生导师,研究方向:X射线CT成像与应用、人工智能等。

董佳鑫(1999—),男,安徽涡阳人,硕士研究生,研究方向:计算机视觉。

马根炜(1992—),男,江西吉安人,博士研究生,研究方向:CT理论与成像技术、深度学习人工智能等。

(收稿日期:2025-10-10 修回日期:2025-12-12)

当前，传统的酒驾换驾调查主要依赖查勘人员逐案核查现场照片、询问当事人及比对历史记录，这种人工处理模式存在明显局限性：一方面，肇事者通过重新布置现场、虚报时间等手段制造“合理”假象，极大增加了照片审查和案情判断的难度；另一方面，逐案核查需投入大量人力物力，尤其在案件量较大时，极易出现调查滞后或疏漏。据行业统计，此类案件的调查成本较普通理赔案件高出 30% 以上，且因证据固定不及时，导致拒赔成功率不足 25%，严重影响了保险公司的正常经营秩序。因此，亟需引入更高效、智能的风险识别技术，以应对日益严峻的挑战。

1 LightGBM 算法概述

随着机器学习技术的发展，构建自动化风险预测模型成为关键。其中，LightGBM（light gradient boosting machine）作为集成学习中 Boosting 算法的代表性模型，由微软于 2017 年提出，其核心优势在于通过梯度单边采样（GOSS）和互斥特征捆绑（EFB）技术，显著提升了大规模高维数据的训练效率^[2]。该算法在金融保险领域已展现出卓越性能，例如在车险欺诈检测中，其对样本不平衡问题的处理能力（如通过类别不平衡学习策略优化）和高维特征的高效提取能力（如自动筛选报案时间间隔、驾驶员身份一致性等关键特征），使其成为理想的风险预测工具。

从技术特性看，LightGBM 在处理 4 万条赔案数据（含 172 个字段）时，可通过并行计算和稀疏特征优化，将训练时间压缩至传统算法的 1/3，同时保持 84.56% 的预测准确率。这种高效性使其在交通、电力等数据密集型行业广泛应用，而车险理赔场景中数据规模大、特征维度高、样本分布不均衡（正负样本比约 345:1）的特点，与 LightGBM 的技术优势高度契合，为构建精准的酒驾换驾风险预测模型提供了技术支撑。

2 基于 LightGBM 的酒驾换驾风险预测算法

酒驾换驾风险预测建模分析流程如图 1 所示。

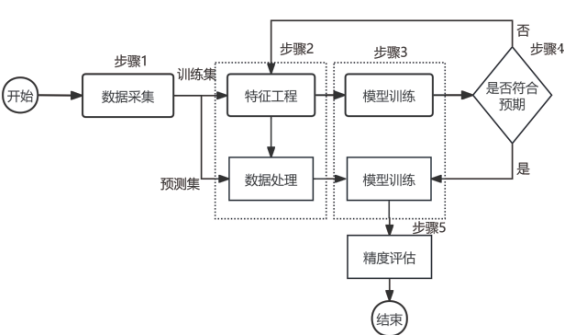


图 1 基于 LightGBM 的酒驾换驾风险预测流程图

整个模型训练过程可分为 5 个步骤：数据收集、特征工程、模型训练、交叉验证和精度评价^[3]。本文将主要介绍数据集采集与处理、LightGBM 训练、精度评估等。

2.1 数据采集

在本研究中，数据采集过程主要依赖于企业核心业务系统，目标是全面覆盖酒驾换驾风险预测问题的全生命周期信息。数据来源于 2024 年车险理赔系统的已决赔案清单，包含了 4 万余条记录和 172 个字段，涵盖了承保信息、查勘信息以及理赔流程数据等多维度信息。

数据采集的过程中，首先对数据进行了清洗和预处理，包括处理缺失值、去除重复数据以及检测和修正异常值。由于数据集中的某些字段可能存在缺失，采用了合适的填补方法，如均值填补或插值填补^[4]，以保证数据的完整性和连贯性。在处理类别型数据时，通过独热编码或标签编码方法进行转换，以确保模型能够顺利使用这些数据进行训练。

此外，由于数据集在酒驾换驾风险预测中呈现出样本不平衡的特征，负样本远大于正样本，因此在数据集的划分和训练过程中，采取了过采样（SMOTE 方法^[5]）和加权损失函数^[6]等技术来平衡样本，以避免模型预测偏向于负样本，确保模型能够更好地识别少数类样本。通过这些步骤，保证了数据的质量和适用性，为后续模型训练和预测分析奠定了坚实的基础。

2.2 特征工程

特征工程是决定模型性能上限的关键步骤。本研究从业务经验出发，选择了“报案号”“被保险人姓名”“驾驶员姓名”“车价”“出险时间”“报案时间”等核心字段。同时，构建了两个关键衍生特征：

（1）时间间隔：计算“报案时间”与“出险时间”的差值。通常，欺诈案件的报案时间相较于事故发生时间会有异常延迟。

（2）身份一致性：比对“驾驶员姓名”与“被保险人姓名”是否一致，并编码为一个二元特征。

为了直观展示各关键特征在风险预测中的作用，利用训练完成的 LightGBM 模型计算了特征重要性（Feature Importance），结果如图 2 所示，“报案—出险时间间隔”和“驾驶员与被保人身份一致性”是模型进行风险判断时最重要的两个特征。

此外，根据业务洞察，酒驾换驾案件中，进行虚假报案的驾驶员性别常出现特定模式（如女性代替男性）。然而，在本次研究使用的 2024 年赔案数据中缺少驾驶员性别字段。为此，利用一个独立的、带性别标签的历史数据集（包含 41 280 条记录）构建了一个基于姓名的性别预测子模型。该

子模型使用朴素贝叶斯分类器进行训练，用于预测新案件中驾驶员的性别，并将预测结果作为一个新的重要特征引入主模型。

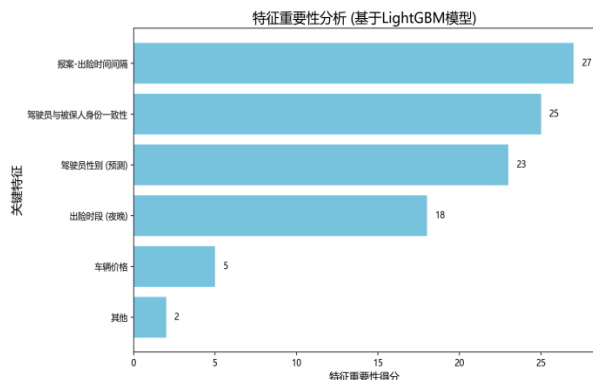


图2 特征重要性分析 (基于 LightGBM 模型)

为直观展示该特征的效果，图3对比了风险与无风险案件中预测的驾驶员性别分布。图中可见，风险案件中女性驾驶员的占比(65%)显著高于无风险案件(25%)，验证了“女性顶包”这一关键业务假设。

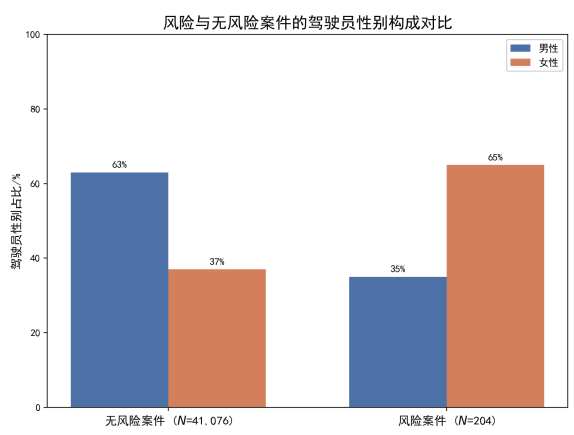


图3 风险与无风险案件驾驶员性别构成对比图 (百分比)

2.3 模型训练

本文研究了包括GBDT^[7]、XGBOOST^[8]和LightGBM 3种最常用的机器学习算法的区别和特点，通过比较预测精度和复杂度，最终选择了LightGBM作为整个模型的核心算法。首先将整体数据集按8:2的比例划分为训练集(用于模型训练与验证)和独立的测试集(用于最终性能评估)。模型训练过程中的关键参数设置如表1所示，这些参数是通过在训练集上进行交叉验证调优选出的。

表1 LightGBM 的参数设置

Boosting 类型	最大深度	评估函数	最大分类	结果类型	随机状态
GBDT	9	AUC	1	objective:binary	333

2.4 交叉验证

为获得对模型泛化能力的可靠评估并进行有效的超参数调优，本研究在训练集上采用了5折交叉验证(5-fold Cross Validation)^[9]方法。具体来说，训练集被随机划分为5个互不相交的子集，每次迭代使用其中4个子集进行训练，剩下的1个子集用于验证。这个过程重复5次，确保每个子集都被用作验证集一次。

在每一折的训练中，采用了早停策略(Early Stopping)来防止模型过拟合^[10]。通过监控验证集上的AUC指标，当指标在一定迭代次数内没有提升时，训练将自动停止，并保存当前最优的模型。

最终模型的性能，即表2中报告的各项指标，是在独立的、模型从未见过的20%测试集上计算得出的，这确保了评估结果的客观性和公正性。

2.5 精度评估

本实验采用准确率 Accuracy、精确率 Precision、召回率 Recall 和 F_1 值对模型进行评估。识别驾驶员性别的任务属于二分类问题，因此只需要判断样本标签为0或者1，即 Positive 或者 Negative^[11]。Accuracy、Precision、Recall 和 F_1 用公式分别表示为：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

在本研究的二分类任务中，将“有风险”定义为正类(Positive, 记为1)，“无风险”定义为负类(Negative, 记为0)。各符号定义如下：

(1) TP (True Positive)：真实为有风险样本，且被模型正确预测为有风险。

(2) TN (True Negative)：真实为无风险样本，且被模型正确预测为无风险。

(3) FP (False Positive)：真实为无风险样本，但被模型错误地预测为有风险(误报)。

(4) FN (False Negative)：真实为有风险样本，但被模型错误地预测为无风险(漏报)。

3 实验结果分析

将LightGBM与GBDT、XGBOOST算法进行对比分析。

实验结果如表 2 所示。

表 2 实验对比分析表

算法	Accuracy	Precision	Recall	F_1
LightGBM	0.845 6	0.947 3	0.898 2	0.922 0
GBDT	0.835 4	0.914 1	0.868 1	0.890 5
XGBOOST	0.818 1	0.911 2	0.881 2	0.895 9

由表 2 可知，本文所用的 3 种模型均取得了较好的预测效果，但 LightGBM 的综合性能最优。具体来看：

在准确率（Accuracy）上，LightGBM 达到 84.56%，分别比 GBDT 和 XGBOOST 高出 1.22% 和 3.36%。

在精确率（Precision）和召回率（Recall）两个核心指标上，LightGBM 同样表现最佳，这表明模型在识别出真正风险案件的同时，也保持了较低的误报率。

在 F_1 分数上，LightGBM 达到了 0.922 0，相较于 GBDT（0.890 5）和 XGBOOST（0.896 0）也有明显优势，证明了其在不平衡数据上的学习能力和预测的均衡性。综上所述，LightGBM 更适用于本研究的酒驾换驾风险预测场景。

4 结果与启示

本文基于某产险公司 2024 年的已决赔案数据，设计并实现了一个基于 LightGBM 的酒驾换驾风险预测模型。研究的关键创新在于特征工程，通过构建“报案—出险时间间隔”“驾驶员与被保险人身份一致性”等衍生特征，并利用朴素贝叶斯模型预测驾驶员性别作为补充信息，有效地捕捉了酒驾换驾欺诈行为的核心模式。实验结果表明，融合了这些关键特征的 LightGBM 模型在多个评估指标上均表现出色， F_1 分数达到了 0.922 0，能够精准、高效地识别潜在欺诈案件。该模型为保险公司提供了一种智能化的风险筛查工具，有助于优化理赔资源配置，提升反欺诈工作的效率与精准度。

尽管本研究取得了积极成果，但仍存在可优化的空间。例如，基于姓名的性别预测模型可能存在一定的地域或文化偏误，未来可探索融合更多信息源（如历史保单信息）或行为学特征（如驾驶员的历史交通违规记录、报案频率异常等）来提升预测准确性和更全面地刻画风险画像。在模型层面，除了 LightGBM，也可以尝试更复杂的特征交互和深度学习模型（如利用长短期记忆网络 LSTM 处理时间序列特征，或图神经网络 GNN 分析被保险人与驾驶员之间的复杂关联关系），或集成学习的更多高级变体，以挖掘深层次风险模式并进一步提升模型的鲁棒性和泛化能力。此外，模型的线上部署与实时预警系统的构建将是未来工作的重点。通过将风险预测能力前置到理赔流程的早期阶段，系统能够在事故报案后即时评估酒驾换驾风险并向查勘人员提供预警和调查指

引。这不仅能大幅提升调查效率、降低人工甄别成本，还能是关键证据的及时获取创造条件，从而显著提高保险公司对欺诈案件的拒赔成功率。最终，通过技术赋能，实现理赔风险管理从被动响应向主动预防的战略转型，这对于维护保险市场的健康发展和打击金融欺诈具有深远意义。

参考文献：

[1] 酒驾请人顶替骗保共同获罪 [J]. 农家致富, 2023(1): 49.

[2] 刘迪成, 马龙虎, 周元洪, 等. 基于 LightGBM 的特种纸打浆度预测模型研究与应用 [J]. 自动化应用, 2025,66(7):278-281.

[3] 高洁, 张涛, 程新洲, 等. 一种基于 LightGBM 机器学习算法的用户年龄及性别预测方法 [J]. 邮电设计技术, 2019(9): 36-39.

[4] 刘鹤鹏, 李登华, 丁勇. 融合多测点数据相关性的大坝监测历史数据填补 [J]. 人民长江, 2023,54(9):245-251.

[5] 周纪. 基于 GRU-Attention 网络与 SMOTE 算法的疲劳驾驶检测方法 [J]. 汽车实用技术, 2025,50(9):46-51.

[6] 渠鸿宇, 胡海川, 钱传海, 等. 基于加权损失函数的我国近海海面风速预报 [J]. 热带气象学报, 2024,40(6):931-942.

[7] 陈冰雪, 尹莺, 洪媛媛. 机器学习算法在消费信贷风控模型中的应用 [J]. 中小企业管理与科技, 2025(6):140-143.

[8] 程宇轩. 基于机器学习的金融客户购买行为预测 [J]. 黑河学院学报, 2023,14(10):52-56.

[9] 胡山, 朱向华. 基于 ResNet18 与 K 折交叉验证的电机轴承故障诊断 [J]. 自动化与仪表, 2025,40(4):40-45.

[10] 李化东. 深度图卷积网络的过平滑及过拟合问题研究 [D]. 哈尔滨: 黑龙江大学, 2024.

[11] 邢红梅, 陈欣, 王慧. 基于 LightGBM 模型的文本分类研究 [J]. 内蒙古工业大学学报 (自然科学版), 2020, 39(1): 52-59.

【作者简介】

赵瑞阳（1987—），男，山东枣庄人，硕士研究生，中级经济师，研究方向：大数据、人工智能。

刘聪慧（1999—），女，山东德州人，本科，风险管理运营部创新支持岗，研究方向：机器学习。

（收稿日期：2025-06-23 修回日期：2025-12-03）