



西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

2023 届 本科毕业论文（设计）

论文题目： 不平衡样本下基于集成学习的
车险欺诈风险预测方法研究

学生姓名： 邓 嘉 怡

所在学院： 金融学院

专 业： 保险学(财务与会计双语实验班)

学 号： 41905216

指导教师： 张 婷 婷

成绩：

2023 年 03 月

西南财经大学

本科毕业论文原创性及知识产权声明

本人郑重声明：所呈交的毕业论文是本人在导师的指导下取得的成果，论文写作严格遵循学术规范。对本论文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本毕业论文引起的法律结果完全由本人承担。

本毕业论文成果归西南财经大学所有。

特此声明

毕业论文作者签名：邓嘉怡

作者专业：保险学（财务与会计双语实验班）

作者学号：41905216

2023年4月18日

摘要

随着我国经济水平发展，汽车产业发展和机动车保有量不断攀升，机动车保险已成为财产险的第一大险种。与此同时，不法分子利用车险欺诈获取巨额资金，造成严重的社会危害。2020 年车险综合改革提出了“降价、增保、提质”的目标，如何运用智能化风险评估手段加强车险欺诈风险识别，是财产保险公司风险控制的一个重要方面。本文利用机器学习方法对开源理赔数据进行分析，提出了能够有效识别车险欺诈风险的模型方法，助力车险风控。

首先，通过对理赔数据的探索性分析，探究不同数据特征对欺诈风险的影响，给出重要特征的可视化分析。进一步，针对数据的缺失特性和特征共线性，给出缺失值的处理策略。同时根据特征含义构建衍生特征，利用包装式特征选择方法对数据进行降维，构建起具有区分车险欺诈能力的特征工程。最后，分别利用支持向量机、决策树、随机森林、XGBoost、LightGBM 和 CatBoost 六个机器学习模型建模，比较集成模型和基准模型的优劣，同时利用 Stacking 方法进行模型融合。此外，考虑到数据不平衡对模型性能具有较大的影响，将集成模型和数据采样算法进行结合，进一步提高了模型整体性能。

研究表明：最终的最优模型是经过采样算法优化的 XGBoost 和 CatBoost 的融合模型，在 AUC 指标上达到 0.80433，得到具有较强的车险欺诈风险识别能力的模型。因此，机器学习模型能够有效助力车险反欺诈，但在实践中还需提高车险公司数据维度和质量，开发知识图谱、小样本学习等新技术，从数据中挖掘更多价值；促进多方数据共享，在信息安全的前提下实现信息的互融互通。这对于我国车险行业健康发展具有积极意义。

关键词：机动车辆保险；保险欺诈风险；不平衡数据；机器学习

Abstract

With the development of China's economic level, the improvement of the automobile industry and the continuous increase in the number of motor vehicles, automobile insurance has become the largest type of property insurance. At the same time, criminals take advantage of auto insurance fraud to obtain huge amount of money, causing serious social harm. The comprehensive reform of auto insurance in 2020 put forward the goal of "reducing prices, increasing insurance, and improving quality". How to use intelligent risk assessment methods to strengthen the identification of auto insurance fraud risks is an important aspect of risk control for property insurance companies. This paper uses machine learning methods to analyze open source claims data, and proposes a model that can effectively identify the risk of auto insurance fraud to help auto insurance risk control.

First, through exploratory analysis of claims data, the impact of different data features on fraud risk is explored, and a visualization of important features is given. Furthermore, according to the missing characteristics and feature collinearity of the data, a strategy for handling missing values is given. At the same time, the derived features are constructed according to the meaning of the features, and the dimensionality reduction of the data is carried out by using the Wrapper selection method, so as to construct the feature engineering with the ability to distinguish auto insurance fraud. Finally, six machine learning models of support vector machine, decision tree, random forest, XGBoost, LightGBM and CatBoost are used for modeling, and the advantages and disadvantages of the ensemble model and the benchmark model are compared, and the Stacking method is used for model fusion.

In addition, considering that data imbalance has a great impact on model performance, the ensemble model and SMOTE algorithm are combined to further improve the overall performance of the model.

The research results show that: the final optimal model is the fusion model of XGBoost and CatBoost optimized by the sampling algorithm, and the AUC score reaches 0.80433, so the model with strong identification ability of auto insurance fraud risk is obtained. Therefore, the machine learning model can effectively help the anti-fraud of auto insurance. But in practice, it should also improve the data dimension and quality of more claims data, develop new technologies such as knowledge graphs and small sample learning, and mine more value from data. Promoting multi-party data sharing under the premise of information security, information interoperability and intercommunication can be realized. This is of positive significance to the healthy development of China's auto insurance industry.

Keywords: auto insurance; insurance fraud risk; unbalanced data; machine learning

目录

- 1. 引言 1
 - 1.1 研究背景及意义..... 1
 - 1.2 文献综述..... 2
 - 1.3 研究内容与技术路线..... 4
 - 1.4 研究创新点..... 5
- 2. 数据来源与数据准备 5
 - 2.1 数据来源与样本说明 5
 - 2.2 数据探索性分析 7
 - 2.2.1 正负样本探索分析 7
 - 2.2.2 特征分布探索分析 9
 - 2.2.3 测试集与训练集探索分析 10
 - 2.3 数据清洗..... 10
 - 2.3.1 缺失值识别和处理 10
 - 2.3.2 共线性识别和处理 11
 - 2.4 特征工程..... 12
 - 2.4.1 特征衍生 12
 - 2.4.2 特征选择 13
 - 2.4.3 特征标准化 15
- 3. 优化样本不平衡的车险欺诈风险预测模型实证分析 16
 - 3.1 模型评价指标..... 16

3.2 基于集成学习的车险欺诈风险预测模型实证分析	17
3.2.1 支持向量机(SVM)模型构建	17
3.2.2 决策树模型构建	18
3.2.3 随机森林模型构建	18
3.2.4 XGBoost 模型构建	19
3.2.5 LightGBM 模型构建	20
3.2.6 CatBoost 模型构建	20
3.2.7 模型训练结果比较	20
3.3 样本不平衡优化方法及其实证分析	23
3.4 车险欺诈风险预测融合模型实证分析	26
4. 研究结论与政策建议	30
4.1 研究结论	30
4.2 政策建议	31
4.2.1 增强数据治理，提高数据质量	31
4.2.2 创新技术手段，高效打击欺诈	31
4.2.3 整合各方资源，完善共享平台	31
参考文献	33
附录	35
致谢	36

1.引言

1.1 研究背景及意义

随着汽车生产量和保有量的增长，汽车保险逐渐成为与全球经济增长和人民生活息息相关的重要行业，是我国财产保险的重要支柱。车险能够负责承担自然灾害和人为造成车辆事故的损失费用，其中包括汽车保险和第三方机动车责任保险。如图 1-1 所示，从 2017 年到 2022 年，中国车险原保费收入除 2021 年以外均表现为逐年上升的趋势。

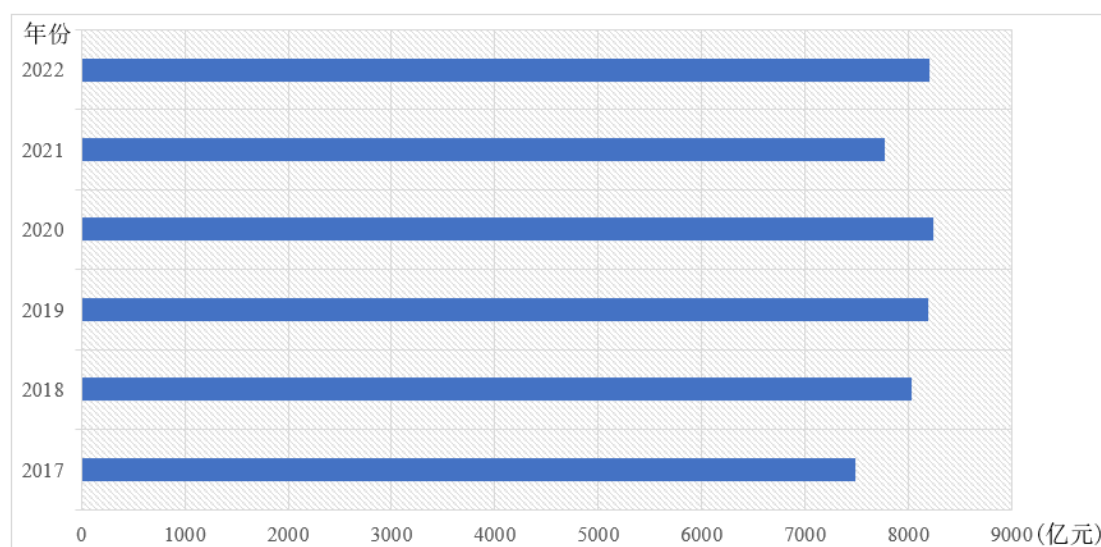


图 1-1 2017-2022 年中国车险原保费收入条形图¹

随着行业和社会对车险正增长的信心增强，更多的资本将加入车险市场，从而使车险公司之间的竞争激烈程度增加。因此，车险公司将更加专注于降低成本，并领先于竞争对手。但在我国，保险欺诈占保险费用的很大一部分。车险行业中，欺诈金额占理赔金额的 20%²。保险欺诈不仅降低了保险业务的收益，导致重大损失，

¹ 资料来自中国保险行业协会 <http://www.iachina.cn>

² 数据来源于中国保险学会与金融壹帐通联合发布的《2019 年中国保险行业智能风控白皮书》

从长远看还会影响保险公司的价格策略和社会经济价值。马亮亮（2021）指出，从保险行业每年发布的十起保险欺诈典型案例中可以看出，车险欺诈手段层出不穷，有诸如伪造交通事故、夸大财产损失、团伙分工骗保等手段。因此车险公司有必要加大对欺诈风险的识别和打击力度，从而降低其社会危害性。

但车险欺诈风险的控制和降低十分困难，主要体现在以下三个方面：（1）车险索赔量巨大，因此产生了海量针对车险理赔的业务数据，使用人工验证每项索赔以进行欺诈风险检测是不可行的。因此，全球领先的财产保险公司都在尝试建立机器学习、数据挖掘和深度学习预测保险欺诈风险的系统。但目前由于技术手段的限制，系统性、自动化的保险欺诈风险预警、控制和防范仍处于探索阶段，保险公司对大数据的利用率仍有很高的提升空间。（2）大多数车险公司，特别是中小型公司，难以收集、清洗并治理海量保险欺诈数据，因此难以形成建立欺诈风险预测系统的数据基础。在这种情况下，车险欺诈风险预测形成了较高的门槛，在实际业务中难以普及。（3）车险欺诈数据是典型的不平衡数据，即由于保险欺诈产生的赔付案件数量占有所有车险赔付案件数量的比例远小于 50%。若不针对该问题做特殊处理，将会降低机器学习和数据挖掘算法预测车险欺诈风险的准确性。而不平衡问题的处理方法尚处于探索阶段，能够应用到车险欺诈风险识别和预测的方法数量十分有限。

考虑到上述问题，本文分析了数据挖掘策略，以建立用于检测车险欺诈风险的预测模型；同时针对数据不平衡问题，引入采样算法提高模型整体性能。从理论层面来讲，本文基于美国数据治理较成熟的保险公司所披露的车险理赔数据，将数据挖掘方法和车险欺诈风险相结合，并建立模型进行实证研究，目的在于探索出基于决策树的集成算法中适用于车险欺诈风险预测问题的最优模型，为保险公司优化车险反欺诈提供理论支撑。从现实意义来说，本文在探索最优预测模型的基础上加入过采样算法，缓解了数据不平衡的现实问题，为车险公司降本增效，减少车险欺诈的社会危害，助力我国车险行业达到更高的反欺诈水平。

1.2 文献综述

保险欺诈的研究最早起源 20 世纪 60 年代，美国学者 Arrow（1963）利用经济

信息学理论探索研究医疗保险欺诈问题。他认为代理人会在与医疗委托人出现利益冲突时，利用信息不对称来保证自己的利益最大化，但同时损害了医疗委托人的利益。而国内学者对保险欺诈风险成因的分析研究大多从博弈论的角度出发。刘家养（2008）分析了利用扩大保险事故损失行为进行欺诈的案例，从博弈论的角度得出结论，保险公司的调查概率与其成本是高度相关的。赵桂芹（2010）通过实证分析，认为由于投保人在保险金额降低时更加注重风险防范，道德风险的激励效应会更加显著。

在对车险反欺诈的实证研究中，研究者常用的两种方法分别是回归分析和机器学习算法。在传统的专家系统和回归模型应用方面，Artis（1999）基于 1995-1996 年西班牙车险市场的理赔数据，利用 Probit 和 Logit 等回归模型对欺诈风险进行检测和分析，利用 AGG 保险欺诈风险模型成功识别了近 95% 的车险欺诈案件。刘兴跃（2017）采用问卷统计收集大量多维度数据，运用 Probit 模型对理赔案件中的欺诈因子进行实证分析，结果发现：驾驶员与被保险人是否有亲属关系、是否出具交警事故责任认定书、被保险车辆使用年限、车辆维修厂家类型等指标为重要欺诈因子。

在利用机器学习技术识别车险欺诈的领域，不少专家学者革新模型算法，旨在提高模型性能和准确度。Viaene（2005）将神经网络分类器与贝叶斯分类结合，用于车险索赔欺诈案件的检测。闫春（2017）基于随机森林模型，加入蚁群算法融合出新分类器，通过随机森林算法产生的特征重要程度排名，挖掘欺诈因子的特征组合，为车险公司提供了建立欺诈模型和底层数据指标的重要参考。王海巍（2016）提出为保险公司建立基于 Hadoop 大数据平台的动态风险因子聚类分析，从而实现自动化识别高危主体，降低道德风险。Yan（2020）利用关联规则挖掘车险欺诈规则，并将基于规则剪枝的最近异常值检测方法应用于此领域，发现关联规则改进后的车险欺诈风险识别算法具有准确度更高、复杂度更低、对聚类算法 K 值影响更小等优点。

基于大数据技术挖掘车险欺诈风险的理论研究，有的专家学者探究了其在实际业务中的应用。廖新年（2010）研究了保险公司使用大数据技术挖掘欺诈风险的主要手段，认为应该由行业协会或监管部门建立大数据反欺诈平台，实现数据共享，

推动业内公司反欺诈工作的开展。段冉（2018）从法律的角度研究了保险行业的反欺诈现状，认为目前缺乏专门的政策和法律法规来约束大数据的使用，国家相关部门应尽快完善相关法律法规。

综上，在现有文献中，国内外学者从博弈分析、信息不对称、道德风险等方面进行了深入研究，探索出保险欺诈产生的诸多因素，为保险反欺诈研究提供了理论基础。不少学者从专家经验、问卷调查、统计分析、数据挖掘角度进行了实证理论研究，提出了许多针对车险反欺诈数据建模的创新方法。也有研究者结合业界实践经验，从法律监管、平台实施、防范措施等角度提出建设性意见。但笔者认为，目前针对车险反欺诈的数据不平衡性问题的解决方法尚不成熟，且容易被研究者忽略，因此本研究在建立机器学习模型时，在模型寻优的基础上将重点解决数据不平衡问题，提高模型整体性能。

1.3 研究内容与技术路线

第一章为引言。主要介绍研究背景、文献综述、研究内容和创新点。

第二章为数据来源与数据准备。首先介绍了本文研究的数据来源及形式，对数据进行探索性分析，然后进行数据清洗，最后通过特征衍生、特征选择、特征标准化的手段建立特征工程。

第三章为优化样本不平衡的车险欺诈风险预测模型实证分析。首先介绍本文选用的多维模型评价指标，然后分别构建支持向量机、决策树、随机森林、XGBoost、LightGBM、CatBoost 六种模型，预测车险欺诈风险。然后将采样算法用于模型训练的交叉验证中，解决数据不平衡问题，通过实验寻找每种模型适用的最优采样算法。最后使用 Stacking 模型融合方法，对六个模型进行排列组合，寻找预测欺诈风险最优融合模型，作为本次研究的最终模型。

第四章为研究结论与政策建议。提出研究的主要结论，并结合业界实际情况和政策，提出本文政策建议。

本文研究的技术路线如图 1-2 所示。

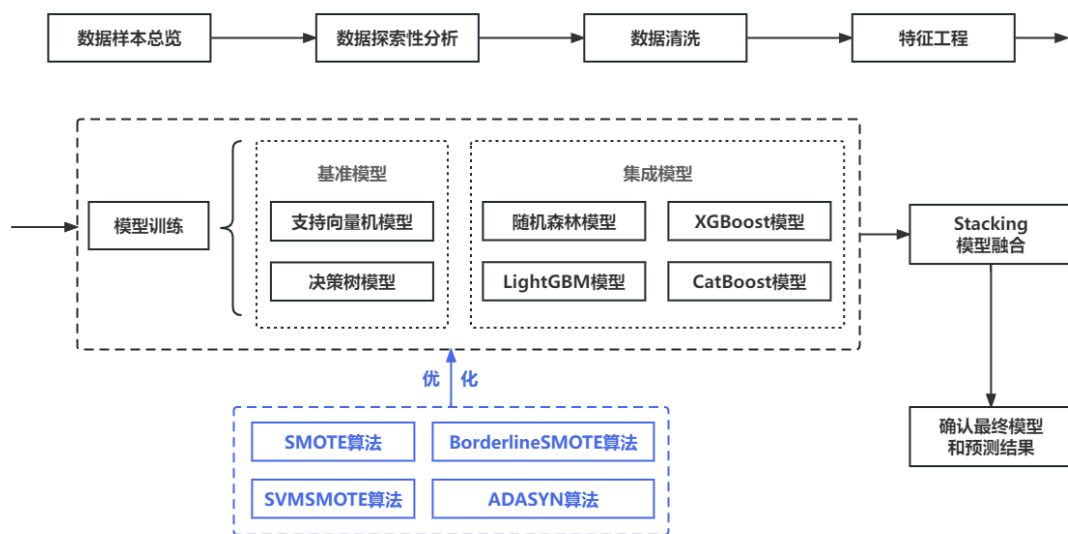


图 1-2 技术路线图

1.4 研究创新点

本文主要创新点在于：一是考虑到车险欺诈问题数据不平衡的自然性质，结合机器学习模型和采样算法优化模型整体性能。该操作使车险欺诈风险预测模型在数据集平衡后进行训练，使正样本和负样本的数量相当，从而使模型预测准确度达到最优。二是对车险反欺诈问题的最优模型进行详细的实验探索，得出使用常见机器学习模型解决该问题的最优方式。通过探索六大机器学习模型与各自最适用的采样算法结合后的模型，探索发现预测准确度最高的单个模型；再通过将单个模型使用 Stacking 方法融合，探索最能识别车险欺诈风险的融合模型。

2.数据来源与数据准备

2.1 数据来源与样本说明

本文选取 Kaggle 数据科学社区的公开脱敏数据集作为建立模型的基础，其内

容是美国某车险公司在 2020 年的部分理赔数据，共计 1000 条理赔记录，39 个特征变量。数据集的特征变量主要包括保单信息、被保险人个人信息、事故现场信息、事故车辆信息等。该数据集的优点一是数据整齐、缺失值少、维度丰富，对客户和交通事故的信息保留完整，为模型预测准确性建立起良好的基础；二是经过脱敏处理后的数据，隐藏了客户的关键信息，最大程度保护了客户的隐私，同时不因数据脱敏降低模型的性能。

数据的所有特征变量可分为两类：数值型特征和分类型特征。其中部分数值型特征如表 2-1 所示。

表 2-1 部分数值型特征的描述性统计

特征变量	平均数	最大值	最小值	中位数	标准差
保单持有人年龄(age)	39.10	64.00	20.00	38.00	9.16
涉事车辆数 (number_of_vehicles_involved)	1.87	4.00	1.00	1.00	1.03
目击证人数量(witnesses)	1.47	3.00	0	1.00	1.11
伤害索赔金额(injury_claim)	7463.25	21450.00	0	6820.00	4890.61
财产索赔金额(property_claim)	7324.14	23670.00	0	6685.00	4778.11
汽车索赔金额(vehicle_claim)	37688.28	79560.00	70.00	41760.00	18695.68
保单年保费 (policy_annual_premium)	1245.97	2047.59	433.33	1252.28	248.31

从上表可知，该车险公司所披露的理赔数据中保单持有人的平均年龄约 39 岁，年龄段从 20 岁到 64 岁不等。在交通事故现场，涉事车辆数从 1 辆到 4 辆不等，但大多数事故仍只有一辆涉事车辆；目击证人数量从 0 人到 3 人不等，总体看来该公司统计到的目击证人数量十分有限。从索赔情况来看，人身伤害和财产损失造成的平均索赔金额都为 7300 余美元，标准差超过 4700 美元；所有赔付中车辆索赔金额整体最大，平均索赔金额达到 37688 美元，标准差达到 18695 美元，因此车辆索赔金额的波动较大，可能存在大量离群值。数据样本中每年保费的均值为 1245.97 美元，标准差 248.31 美元，由此可知各保单的保费波动远小于索赔金额，与实际情况相符。从表中信息可知，由于各数值型特征的均值和标准差的差异过大，即数值型特征取值范围差异较大，后续数据清洗时须对数值型变量进行标准化处理。

部分分类型特征如表 2-2 所示。

表 2-2 部分分类型特征的描述性统计

特征变量	类别数	最大频数变量	最大频数
被保险人性别 (insured_sex)	2	FEMALE	537
被保险人兴趣爱好 (insured_hobbies)	20	exercise	57
是否有警察记录报告 (police_report_available)	2	NO	343
汽车品牌(auto_make)	14	Suburu	80
事故城市(incident_city)	7	Springfield	157
事故严重程度 (incident_severity)	4	Minor Damage	354
是否有财产损失 (property_damage)	2	NO	338

从上表可知，向保险公司报告索赔的保单中，被保险人为女性的超过 50%，被保险人兴趣爱好最多的是体育锻炼。被保车辆品牌一共有 14 中，数量最多的是斯巴鲁汽车。该数据集中的交通事故分布于 7 个城市，最常发生的城市是美国春田市。交通事故的严重程度分为完全报废(Total Loss)、严重伤害(Major Damage)、较小伤害(Minor Damage)、轻微伤害(Trivial Damage)四个等级，大多数交通事故的严重程度为较小伤害。另外值得注意的是，大多数的索赔案例没有警察记录报告，也没有发生财产损失，但这两个特征变量的最大频数并未超过 500，说明这两个特征存在大量缺失值，需要在数据清洗时按一定规则填充。

2.2 数据探索性分析

首先将数据集按 7:3 的比例随机分为训练集和测试集，后续将在训练集和测试集上进行相同的数据清洗操作，使用训练集训练机器学习模型，使用测试集预测结果，并对其结果进行评价。

2.2.1 正负样本探索分析

该数据集样本的正负样本分布情况如图 2-1 所示。

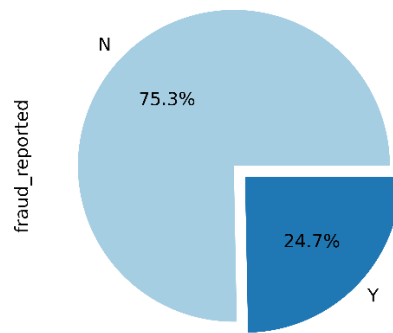


图 2-1 车险欺诈情况的分布

由图 2-1 可知，样本标签含义为是否报告为欺诈案例(fraud_reported)，可分为是和否两类，因此本研究要解决的是一个二分类问题，将正样本(Y)标记为 1，负样本(N)标记为 0。从标签的扇形图可知，正样本仅占比 24.7%，远不足样本总量的 50%，该数据集存在样本标签不平衡的问题，后续将应用采样算法进行优化。

当各个特征变量在正样本和负样本上的分布存在异质性时，机器学习模型能够更好地通过该变量区别正负样本，从而进一步提高模型性能。部分特征变量在正负样本上的分布比较如图 2-2 所示。

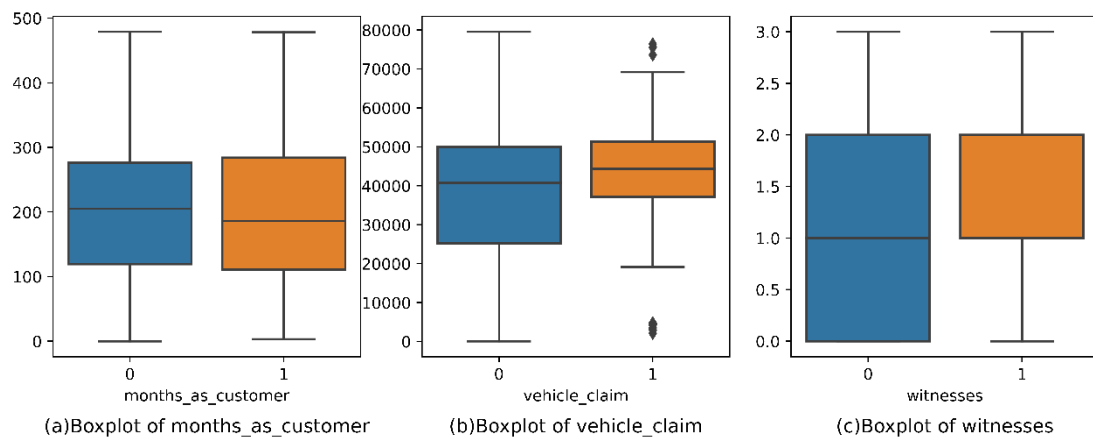


图 2-2 部分特征在正负样本上的分布图比较

由图 2-2 可知，该数据集的特征变量能够在正负样本上的分布产生异质性。车辆索赔金额(vehicle_claim)在正样本上的分布明显偏高，意味着发生车险欺诈的案例在索赔时对于车辆索赔金额往往有夸大的嫌疑。目击证人数量(witnesses)同样也在正样本上的分布更偏高，表明目击证人更多时可能更易识别出发生车险欺诈。而客户签约时长(months_as_customer)则在正负样本上的分布趋于一致，该变量不易

识别是否存在车险欺诈风险。

2.2.2 特征分布探索分析

分类型特征共 20 个，部分分类型特征的分布情况如图 2-3 所示。

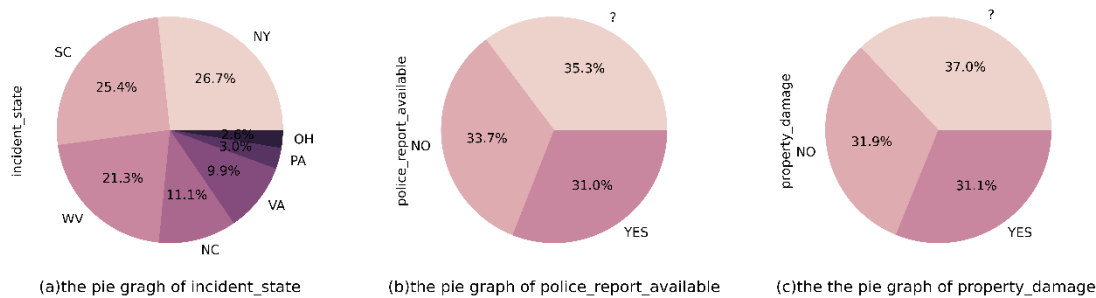


图 2-3 部分分类型特征的分布图

图 2-3 展示了一个多分类变量和两个二分类变量的分布情况。其中事故发生所在州(incident_state)是多分类变量，可知该数据集中包含的索赔案例分布在美国 7 个不同的州，发生索赔最多的州是纽约州。是否有警察记录报告(police_report_available)和是否有财产损失(property_damage)是二分类变量，但上述两个变量分别存在 35.3%和 37.0%的缺失值，将在数据清洗时做填充处理。

数值型特征共 19 个，部分数值型特征的分布直方图如图 2-4 所示。

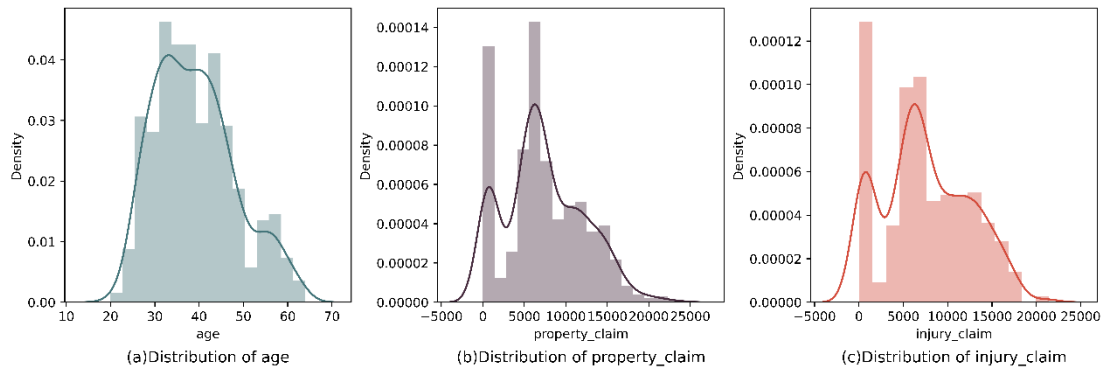


图 2-4 部分数值型特征的分布图

图 2-4 展示了三个数值型变量的分布直方图。其中保单持有人年龄(age)接近于正态分布，可能因为该变量是客户的自然属性；而财产索赔金额(property_claim)和伤害索赔金额(injury_claim)服从于长尾分布，可能是由于上述变量是与保险索赔

直接相关的社会属性。

2.2.3 测试集与训练集探索分析

能够在机器学习模型中取得较高预测准确度的数据集，其训练集和测试集须满足以下条件：各特征变量在训练集和测试集上的分布须大致相同，而不产生变量分布的异质性。部分特征在训练集和测试集上的分布比较如图 2-5 所示。

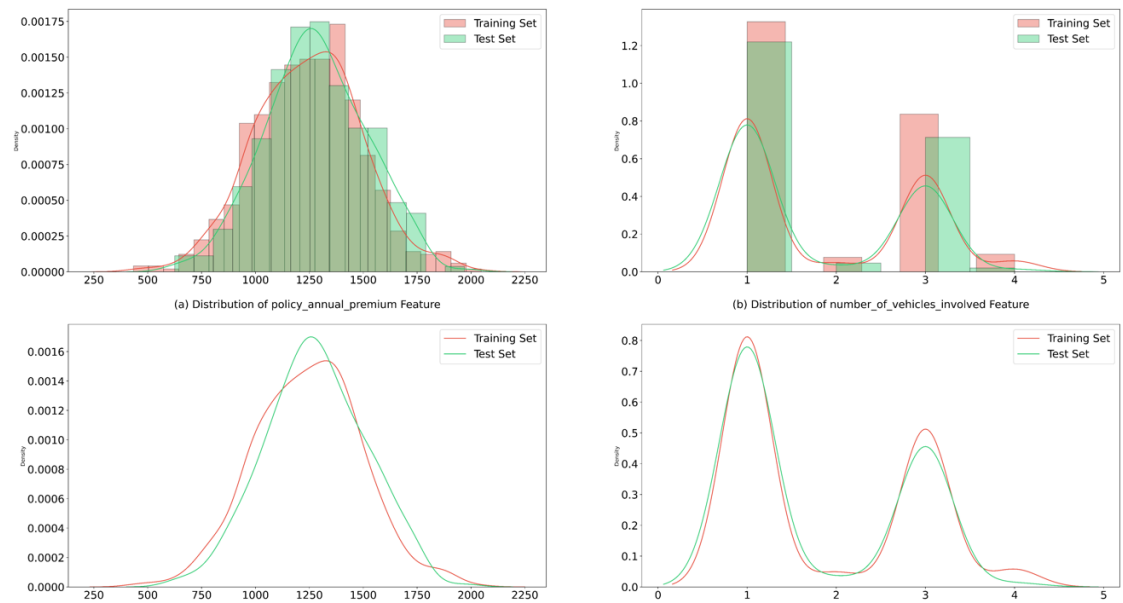


图 2-5 部分特征在训练集和测试集上的分布比较

由图 2-5 可知，部分特征如保单年保费(policy_annual_premium)和涉事车辆数(number_of_vehicle_involved)的密度分布图在训练集和测试集上表现大致相同，因此机器学习模型能很好地从训练集迁移至测试集。

2.3 数据清洗

2.3.1 缺失值识别和处理

数据缺失通常是在数据收集过程中认为操作失误或问卷填写缺失造成，使得所形成的数据集信息不完整，导致数据截断或分类不合理等情况，对后续数据建模造成影响。数据缺失值识别总览如图 2-6 所示。

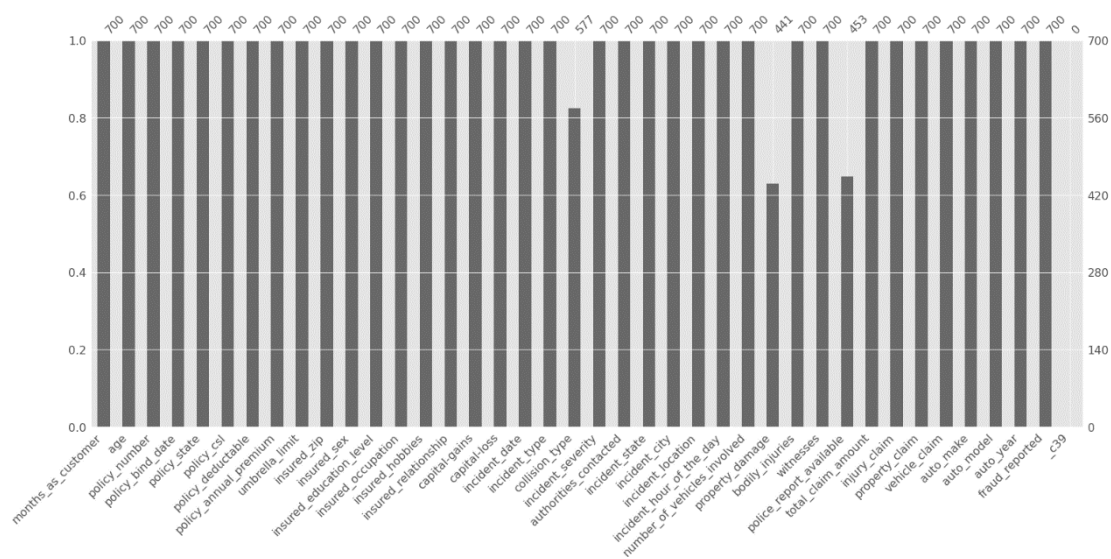


图 2-6 数据缺失值总览

由图 2-6 可知，共有 4 个特征变量出现缺失现象，其中备注信息(_c39)完全缺失，其信息对车险欺诈风险的预测贡献度为 0，此处直接删除该变量。另外，碰撞类型(collision_type)、是否有财产损失(property_damage)和是否有警察记录报告(police_report_available)的缺失值不足 40%，考虑对其进行填充。上述三个变量均为分类型变量，为了最大程度保留有效信息，使用最大频数类别进行填充。

2.3.2 共线性识别和处理

多重共线性问题容易导致小样本条件下，标准误差较大，从而更容易犯第二类错误。本数据集属于大样本情况，并且机器学习模型不需要做参数的假设检验，因此多重共线性并不会降低模型预测的准确度。但共线性问题过大会使机器学习模型产生过拟合，不能很好地将模型从训练集迁移至测试集。

特征变量中较有代表性的 20 个变量的相关系数图如图 2-7 所示。

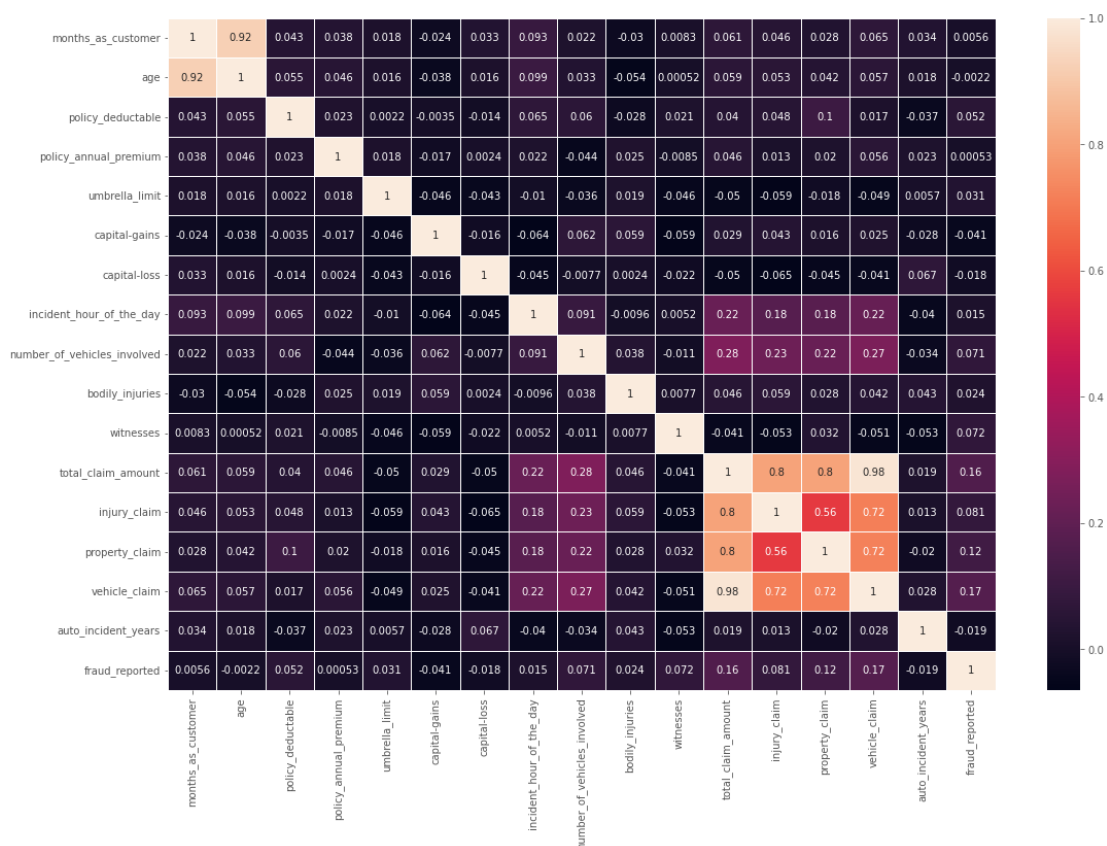


图 2-7 相关系数图

由图 2-7 可知，保单持有人年龄(age)与客户签约时长(months_as_customer)的相关系数高达 0.92，总索赔金额(total_claim_amount)与三个其他变量的相关系数都超过 0.8，存在较大的共线性。经过实验发现，删除保单持有人年龄(age)和总索赔金额(total_claim_amount)两个变量后，交叉验证后各模型训练集的平均 AUC 值整体提高，因此删除上述两个变量，从而降低多重共线性和模型过拟合程度。

2.4 特征工程

2.4.1 特征衍生

特征衍生是要再原有特征的基础上，根据经济含义和实践经验构建新的特征，从而提高模型预测能力和变量的可解释性，同时可加强线性模型对于非线性特征的学习能力。本文将保单生效日期(policy_bind_date)与事故发生日期(incident_date)

作差，构建了事故发生距离保单生效的天数，公式如下：

$$bind_incident_days = incident_date - policy_bind_date \quad (2-1)$$

通过该特征，模型能够通过识别交通事故距离保单开始生效的时间长度，来识别车险欺诈风险是否会发生。二者的关系可能呈正比或反比，也可能是非线性影响关系，因此该特征可能为线性特征或非线性特征。

本文还将事故发生年份，即从事故发生日期(incident_date)变量中提取的年份，与汽车购买年份(auto_year)作差，构建了事故发生时车辆已使用年数的变量，公式如下：

$$auto_incident_years = YEAR(incident_date) - auto_year \quad (2-2)$$

通过该变量，模型能够通过识别事发时车辆已使用的时间长度，或者是车辆的折旧程度，来识别车险欺诈风险。与事故发生距离保单生效天数(bind_incident_days)相似，事故发生时车辆已使用年限(auto_incident_years)也可能是线性变量或非线性变量。

2.4.2 特征选择

经过数据清洗，删除无效特征和产生共线性的特征共 3 个；经过特征衍生，增加新特征 2 个，剩余 38 个特征变量。特征变量之间可能相互干扰，容易造成模型训练难度高、信息冗余以及预测准确度低等问题。因此需要对模型再次降维，减少冗余信息，提升模型的泛化能力。本文采用包装方法(Wrapper)对特征进行选择，通过迭代不同的特征数量获得模型所需要的最佳特征数，然后基于随机森林模型求出数值型特征的重要程度，从而进一步实验重要性较低的特征是否会降低模型预测准确度。

包装法(Wrapper method)在特定分类器下生成并评估特征子集。本文应用 RFE (Recursive feature elimination，递归式特征)，消除结合 SVM 模型来进行多轮训练，每轮训练结束后，消除若干权值系数对应的特征，再基于新的特征集进行下一轮训练，直到满足要求。迭代后筛选出贡献不突出的特征如表 2-3 所示，共删去 9 个特征变量，删去特征大多为分类类别过多的分类型变量，达到特征选择的目的。

表 2-3 特征删除

序号	被删除变量	序号	被删除变量
1	保单编号 (policy_number)	6	事故发生所在州 (incident_state)
2	保单生效时期 (policy_bind_date)	7	事故发生城市 (incident_city)
3	保单生效州 (policy_state)	8	汽车品牌 (auto_make)
4	被保险人邮编 (insured_zip)	9	汽车购买年份 (auto_year)
5	事故发生地点 (incident_location)		

删除上述 9 个特征后,基于随机森林模型计算所有数值型变量的特征重要度。在训练随机森林模型时,每一棵决策树从训练集中进行有放回的抽样,训练集约 2/3 的数据进入决策树,剩余 1/3 未进入决策树的数据被称为袋外数据(out of bag, OOB)。首先通过袋外数据计算每一棵决策树的误差记为 err_{OOB1} , 然后对袋外数据中单独的一个特征加入随机噪音干扰,并将误差记为 err_{OOB2} 。最后将二者作差再除以树的个数,如公式 2-3 所示:

$$\sum \frac{err_{OOB2} - err_{OOB1}}{N} \quad (2-3)$$

其中 N 表示随机森林中决策树的数量。本文计算出数值型特征的重要度结果如图 2-8 所示。

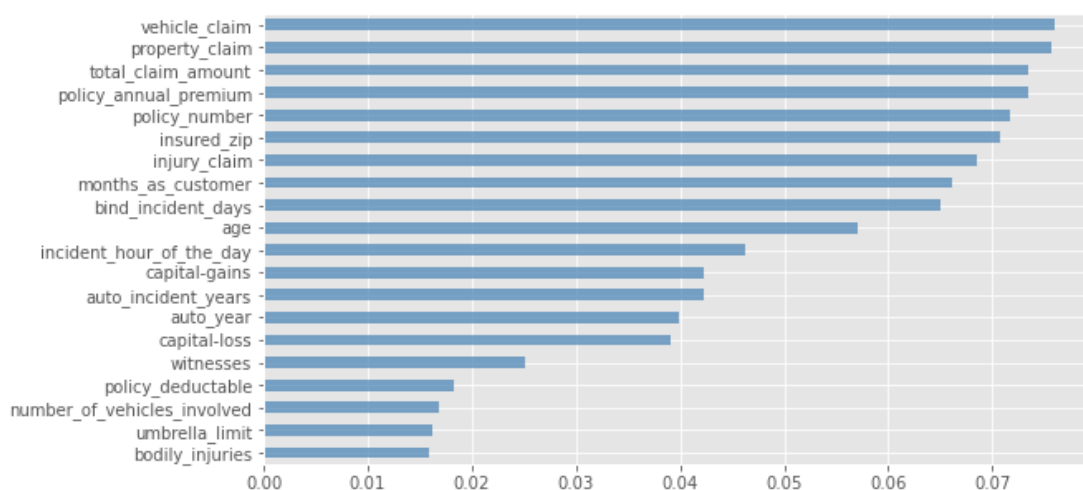


图 2-8 数值型特征的重要度

由图 2-8 可知，所有数值型变量的重要度均超过 0.015，且经过实验发现删去任何特征都将降低模型对车险欺诈风险识别的准确程度，因此保留所有数值型特征，包括本文衍生出来的特征。经过特征选择后的变量共 29 个。

2.4.3 特征标准化

（1）数值型特征标准化

在 2.2 节的数据探索性分析中发现，本数据集中的数值型变量取值范围差异性较大，也即不同特征的量纲相差较大，会干扰部分模型的预测准确度，降低模型性能。为解决该问题，在数据建模之前对数值型特征进行标准化处理。本文采用 *z-score* 标准化方法，使原始数据经过处理后符合均值为 0、标准差为 1 的标准正态分布，其转化公式如下：

$$x^* = \frac{x - \mu}{\sigma} \tag{2-4}$$

公式(2-4)中 μ 为样本数据的变量的均值， σ 为变量的标准差。

（2）分类型特征标准化

为使分类型变量转化为模型能够识别的数字，本文对分类型变量采用独热 (one-hot) 编码。One-hot 编码通过模拟 N 位状态寄存位对 N 个分类状态进行编码，每种类别有其独特的寄存位，并且任意时刻只有一位有效。

One-hot 编码可以将所有分类型变量表示成二进制向量的形式。首先将各分类类别映射到整数值，然后将所有整数值表示为二进制向量，形成以该向量位数为数量的若干特征。每个类别占其中一位二进制位标记为 1，其余二进制位标记为 0。One-hot 编码的图示如图 2-9 所示。

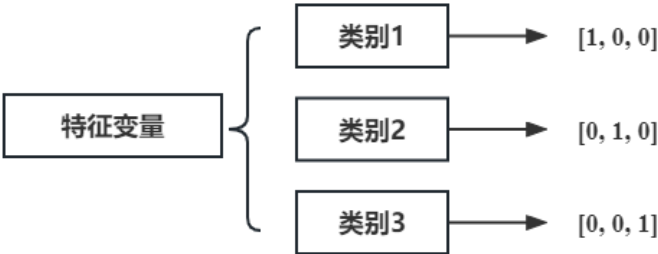


图 2-9 独热 (one-hot) 编码示意图

3. 优化样本不平衡的车险欺诈风险预测模型

实证分析

3.1 模型评价指标

本文采用多维度的评价指标来衡量模型的学习效果。由于车险欺诈风险存在数据不平衡的自然性质，也即不发生欺诈的场景远多于发生欺诈的场景，因此整体样本的准确率存在与实际情况的偏差。本文选取 4 种不同的指标衡量模型的分类性能：AUC(Area Under Curve)、准确率(Accuracy)、精度(Precision)、召回率(Recall)。

(1) 准确率、精度、召回率

根据实际数据风险标签和模型预测风险标签可以将结果分为四类：该样本是欺诈样本且也被模型判定为欺诈样本，为真阳性(TP)；该样本不是欺诈样本但被模型判定为欺诈样本，为假阳性(FP)；该样本不是欺诈样本也不被模型判定为欺诈样本，为真阴性(TN)；该样本是欺诈样本但未被模型识别为欺诈样本，为假阴性(FN)。准确率(Accuracy)表示预测正确的样本数占总样本数的比例，公式如下：

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3-1)$$

在车险欺诈中，正负样本数量不均衡，accuracy 不能很好的反映模型性能。欺诈风控场景中模型方法是否能识别出欺诈案例十分重要，所以增加精度(Precision)、召回率(Recall)作为性能评价指标。两者的计算公式如下：

$$precision = \frac{TP}{TP + FP} \quad (3-2)$$

$$recall = \frac{TP}{TP + FN} \quad (3-3)$$

(2) AUC 值

区别正样本和负样本的能力越强，说明模型的效果越好。AUC 值是常用的衡量模型区别正负样本能力的指标。

首先需要明确真阳性率(TPR)和假阳性率(FPR)，计算公式如下：

$$TPR = \frac{TP}{TP + FN} \times 100\% \quad (3-4)$$

$$FPR = \frac{FP}{FP + TN} \times 100\% \quad (3-5)$$

FPR 和 TPR 之间的关联曲线称为 ROC 曲线，其中 x 轴为 FPR， y 轴为 TPR，两者的取值范围均为 0 到 1。AUC 值则是 ROC 曲线下方的面积，衡量模型区别正负样本的能力，AUC 值越接近 1 模型性能越好。

3.2 基于集成学习的车险欺诈风险预测模型实证分析

3.2.1 支持向量机(SVM)模型构建

支持向量机(SVM)是一种简单的机器学习模型，作为本文的基准模型之一。SVM 的主要思想是求解点到超平面的最大距离，也即正负样本之间的最大距离，如图 3-1 所示。

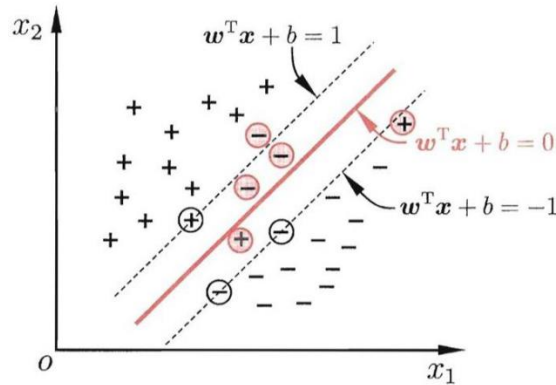


图 3-1 SVM 原理图示

SVM 将点到超平面的距离定义为间隔 (margin)，一方面 SVM 通过间隔最大化 (如上图的红色斜线) 的方式将正负二分类的点区分开；另一方面，对于介于两者之间的点，也要通过足够大的置信度将其区分，使其具有更好的泛化能力。通俗地讲：在车险欺诈识别模型当中，SVM 根据所输入的训练集找一个最优的划分诚信和欺诈样本的平面，并以此为依据区别是否为欺诈样本。

本文采用 sklearn 第三方库中的支持向量分类器，设置超参数如表 3-1 所示。

表 3-1 SVM 模型超参数设置

超参数名	设置值	超参数名	设置值
kernel	'rbf'	gamma	'auto'
degree	3	cache_size	200

3.2.2 决策树模型构建

决策树也是一种简单的机器学习模型，作为本文的基准模型之二。本文采用经典的决策树算法——CART 算法。CART 在分枝时采用“分类基尼系数”衡量分类集合的熵值。基尼系数计算公式为：

$$Gini(D) = 1 - \sum_{k=1}^K \left(\frac{|C_k|}{|D|} \right)^2 \quad (3-6)$$

其中 C_k 是 D 中属于 k 类子集的数量。基尼系数反应的是特征划分样本 D 的不确定程度，因此在对决策树剪枝时，采用基尼系数小的特征。

通过网格搜索 GridSearch 方法，寻找决策树的最优参数，如表 3-2 所示。

表 3-2 决策树模型超参数设置

超参数名	设置值
max_depth	8
min_samples_leaf	4
min_samples_split	3

3.2.3 随机森林模型构建

随机森林是由不同决策树组成的集成模型，基于决策树中 CART 算法的良好性能，在随机森林中，本文将运用结合 CART 算法和 Bagging 算法的随机森林。随机森林是从原始训练集中有放回随机抽取样本，并从所有特征中随机选择特征，生成新的训练集构建决策树的方法。不同决策树最终组合形成一个随机森林模型。本文应用随机森林，设置超参数如表 3-3 所示。

表 3-3 随机森林模型超参数设置

超参数名	设置值
n_estimators	200
min_samples_leaf	8
min_samples_split	6

3.2.4 XGBoost 模型构建

XGBoost 同样是基于决策树的集成模型，采用 Boosting 方法迭代生成。XGBoost 以梯度提升树为参照进行了一系列优化，全称为 Xtreme Gradient Boosting，即极端型梯度提升，在常规的分布式梯度提升上进行优化，能够实现高效、高灵活性且可借鉴。它在梯度提升树的基础上，使用正则化项的限制，降低过拟合的可能；在对计算目标函数方差时，XGBoost 引入二阶泰勒展开式。同时，XGBoost 采用一种可以针对贪心算法优化的可以并行的近似最优解算法，降低了计算过程的复杂度。

极端梯度提升是一个迭代训练的加法模型。对于每一次迭代的决策树模型，其拟合的对象是上一次迭代后拟合值与真实值的残差，因此将所有决策树对某一样本的拟合值相加就得到该样本的极端梯度提升模型的拟合值。对第 t 棵树的第 i 个样本，模型的拟合值可以表示为：

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (3-7)$$

其中， $\hat{y}_i^{(t)}$ 表示第 t 次迭代之后样本 i 的预测结果， $f_t(x_i)$ 是第 t 棵树的预测结果， $\hat{y}_i^{(t-1)}$ 是第 $t-1$ 棵树的预测结果。

本文对 XGBoost 模型的超参数设置如表 3-4 所示。

表 3-4 XGBoost 模型超参数设置

超参数名	设置值	超参数名	设置值
eta	0.005	objective	'binary_hinge'
max_depth	10	eval_metric	'auc'
subsample	0.8	silent	True
colsample_bytree	0.8	nthread	4

3.2.5 LightGBM 模型构建

LightGBM 是一个实现梯度提升算法的框架，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式的优点。利用 GOSS (Gradient-based One-Side Sampling, 基于梯度的单边采样) 与 EFB (Exclusive Feature Bundling, 互斥特征捆绑) 两种方法，使用样本采样代替样本点计算梯度，将某些特征进行捆绑在一起而不是扫描所有特征寻找切分点来降低特征的维度。

对 LightGBM 模型进行超参数调整，缓解模型过拟合问题。综合网格搜索结果，最后确定的参数组合如表 3-5 所示。

表 3-5 LightGBM 模型超参数设置

超参数名	设置值	超参数名	设置值
max_depth	5	min_data_in_leaf	20
num_leaves	20	lambda_l1	0.1
bagging_fraction	0.8	min_child_samples	30

3.2.6 CatBoost 模型构建

CatBoost 是一种能够很好地处理类别型特征的梯度提升机器学习库，同样是一种集成模型。CatBoost 具有两大明显优势，其一，它在训练过程中支持类别型变量，无需对非数值型特征进行预处理；其二，选择树结构时，计算叶子节点的算法可以减少过拟合。上述两个优势是本文使用 CatBoost 的主要原因。

对于 CatBoost 模型最终的超参数设置如表 3-6 所示。

表 3-6 CatBoost 模型超参数设置

超参数名	设置值	超参数名	设置值
depth	5	learning_rate	0.05
subsample	0.6	reg_lambda	3

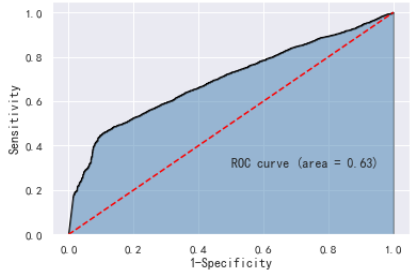
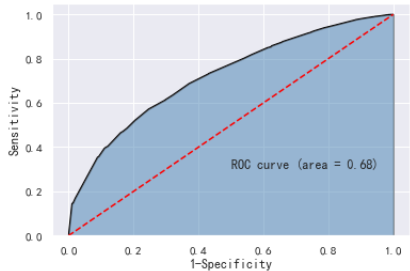
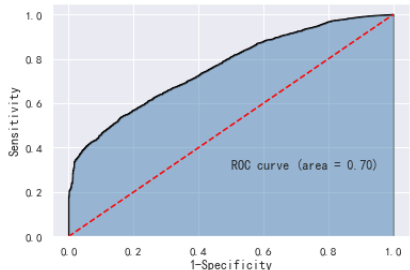
3.2.7 模型训练结果比较

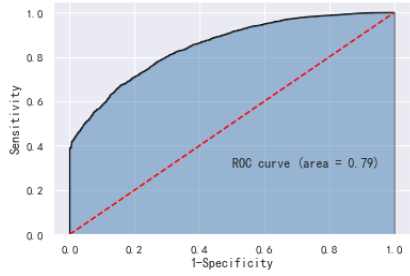
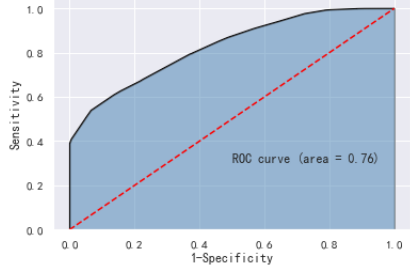
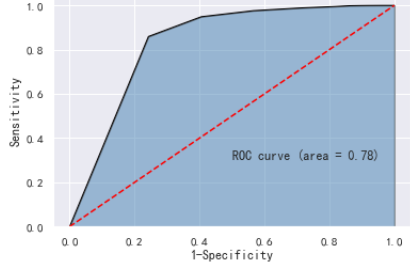
为了更好地比较集成模型和基准模型的分类效果，本文将采用五折交叉验证的方式训练前述的 6 个模型。五折交叉验证将投入模型进行训练的数据集分为训

训练集和验证集，80%的数据为训练集，20%的数据为验证集。共进行五次训练，每次训练时使用训练集来训练模型，验证集来预测检验结果。在此基础上可以得到多组不同的训练集和验证集，某次训练集中的某样本在下次可能成为验证集中的样本，即所谓“交叉”。交叉验证的优点是所有数据都会参与到训练和验证中，能够有效避免过拟合。本节所展示的所有评价指标均是由五折交叉验证后计算的平均指标。

本文通过 Python 3.9 编程语言实现模型构建。支持向量机(SVM)、决策树、随机森林、XGBoost、LightGBM、CatBoost 六个模型的 AUC 值、准确率、精度、召回率和 ROC 曲线如表 3-7 所示。

表 3-7 六个模型的评价结果

模型	AUC	Accuracy	Precision	Recall	ROC 曲线
SVM	0.629	0.741	0.127	0.500	
决策树	0.677	0.750	0.547	0.516	
随机森林	0.700	0.771	0.354	0.598	

XGBoost	0.786	0.830	0.890	0.619	
LightGBM	0.764	0.819	0.608	0.663	
CatBoost	0.778	0.823	0.718	0.640	

ROC 曲线下方面积为 AUC 值，其值越大表明模型预测车险欺诈风险的准确度就越高，分类就越精确。从表中可知，XGBoost 模型的 AUC 值最大，CatBoost 模型次之，而支持向量机的 AUC 值最小。所有集成模型的 AUC 值均大于基准模型的 AUC 值，以 Bagging 方法为基础的随机森林的预测准确度只略高于基准模型，不如以 Boosting 方法为基础的 XGBoost、LightGBM 和 CatBoost，这体现出 Boosting 族模型在预测车险欺诈风险问题上的明显优势。其中，XGBoost 模型不仅在 AUC 值上表现优秀，其准确率(Accuracy)和精度(Precision)都是所有模型中最高的，召回率(Recall)排在第三名，说明在模型返回的所有正样本中有高达 89%的样本预测都是正确的。相比之下支持向量机作为基准模型，只能达到 12.7%的精度，也即支持向量机模型所识别出的车险欺诈案例当中只有 12.7%为真正的欺诈案例，正样本的识别准确度较低。

综上，在使用单个模型预测车险欺诈风险时，XGBoost 模型识别准确度最高。

3.3 样本不平衡优化方法及其实证分析

本文所使用的数据集所包含的训练集正样本仅占 24.7%，针对类别不平衡问题，使用数据采样的方法进行处理，SMOTE 算法是采用平衡各类别样本量的方法重构数据集，人工合成新的少数类样本添加到数据集中。具体算法步骤如下：

第一步，对于少数类中每一个样本 x ，计算该点与少数类中其他样本点的距离，得到最近的 k 个近邻；

第二步，根据样本不平衡比例设置一个采样比例以确定采样倍率，对于每一个少数类样本 x ，从其 k 近邻中随机选择若干个样本，假设选择的近邻为 x' ；

第三步，对于每一个随机选出的近邻 x' ，分别与原样本按照公式构建新的样本数据，从而实现数据平衡。

本文应用 SMOTE 算法及其衍生算法 BorderlineSMOTE、SVMSMOTE 和自适应综合过采样方法 ADASYN 对数据集进行优化。对于每个模型，将四种采样方法应用到交叉验证的训练集。

对于支持向量机，模型训练的优化结果和采样算法最优参数如表 3-8 所示。

表 3-8 支持向量机的优化结果和采样算法最优参数

采样算法	最佳采样比例	K 近邻个数	AUC	Precision	Recall
未采样			0.62920	0.12707	0.50000
SMOTE	0.50	3	0.68880	0.35911	0.57522
BorderlineSMOTE	0.79	3	0.71096	0.47513	0.59310
SVMSMOTE	0.86	2	0.69606	0.46961	0.56666
ADASYN	0.73	2	0.70002	0.45303	0.57746

由表可知，BorderlineSMOTE 算法对支持向量机的性能提升最大，不仅提高了 8.18% 的 AUC 值，还将精度和召回率分别提高了 34.8% 和 9.31%，大大提高了该模型识别车险欺诈案例的能力。最重要的是，在训练集进行采样后，模型所识别的欺诈案例确认为真欺诈案例的概率大大提升。

对于决策树，模型训练的优化结果和采样算法最优参数如表 3-9 所示。

表 3-9 决策树的优化结果和采样算法最优参数

采样算法	最佳采样比例	K 近邻个数	AUC	Precision	Recall
未采样			0.67710	0.54696	0.51562
SMOTE	0.50	3	0.72538	0.66850	0.57345
BorderlineSMOTE	0.50	3	0.71382	0.63535	0.56097
SVMSMOTE	0.50	3	0.70002	0.61325	0.54146
ADASYN	0.90	4	0.71800	0.62983	0.57000

由表可知，SMOTE 算法在重采样比例为 0.5，K 近邻个数为 3 个的时候，能对决策树的性能提升最大，将 AUC 值提升了 4.83%，精度和召回率分别提高了 12.2% 和 5.78%。

对于随机森林，模型训练的优化结果和采样算法最优参数如表 3-10 所示。

表 3-10 随机森林的优化结果和采样算法最优参数

采样算法	最佳采样比例	K 近邻个数	AUC	Precision	Recall
未采样			0.70041	0.35359	0.59813
SMOTE	0.50	3	0.74464	0.50276	0.65000
BorderlineSMOTE	0.75	3	0.74243	0.53591	0.63815
SVMSMOTE	0.70	3	0.74732	0.57458	0.63803
ADASYN	0.85	4	0.74332	0.56906	0.63190

由表可知，SVMSMOTE 算法对随机森林的性能提升最大，其最优参数为重采样比例 0.7 和 K 近邻个数 3 个，此时可将 AUC 值提升 4.69%，精度和召回率分别提高 22.1%和 3.99%。

对于 XGBoost，模型训练的优化结果和采样算法最优参数如表 3-11 所示。

表 3-11 XGBoost 的优化结果和采样算法最优参数

采样算法	最佳采样比例	K 近邻个数	AUC	Precision	Recall
未采样			0.78688	0.88950	0.61923
SMOTE	0.49	2	0.79355	0.87845	0.63600
BorderlineSMOTE	0.71	3	0.79402	0.88397	0.63492
SVMSMOTE	0.51	3	0.79402	0.88397	0.63492
ADASYN	0.60	3	0.79402	0.88397	0.63492

由表可知，有三种算法均能将 XGBoost 的 AUC 值提升至最大值 0.79402，此处保留计算速度最快的 ADASYN 算法。ADASYN 算法能将 XGBoost 的召回率提高 1.57%，但使其精度下降了 0.5%。此时的最佳采样比例是 0.6，最优 K 近邻个数

为 3 个。

对于 LightGBM，模型训练的优化结果和采样算法最优参数如表 3-12 所示。

表 3-12 LightGBM 的优化结果和采样算法最优参数

采样算法	最佳采样比例	K 近邻个数	AUC	Precision	Recall
未采样			0.76434	0.60773	0.66265
SMOTE	0.85	3	0.78987	0.82320	0.64782
BorderlineSMOTE	0.83	3	0.78698	0.79005	0.65296
SVMSMOTE	0.60	3	0.77748	0.71823	0.65656
ADASYN	0.85	4	0.79047	0.80110	0.65610

由表可知，ADASYN 算法对 LightGBM 的性能提升最大，其最优参数为重采样比例 0.85 和 K 近邻个数 4 个，此时可将 AUC 值提升 2.61%，精度提高 19.3%，但召回率下降了 0.6%。

对于 CatBoost，模型训练的优化结果和采样算法最优参数如表 3-13 所示。

表 3-13 CatBoost 的优化结果和采样算法最优参数

采样算法	最佳采样比例	K 近邻个数	AUC	Precision	Recall
未采样			0.77848	0.71823	0.64039
SMOTE	0.70	3	0.78132	0.80110	0.63876
BorderlineSMOTE	0.84	2	0.79834	0.84530	0.65665
SVMSMOTE	0.81	2	0.78760	0.80662	0.64888
ADASYN	0.94	3	0.79015	0.82872	0.64655

由表可知，BorderlineSMOTE 算法在重采样比例为 0.84，K 近邻个数为 2 个的时候，能对 CatBoost 模型的性能提升最大，将 AUC 值提升了 1.99%，精度和召回率分别提高了 12.7%和 1.63%。

综上所述，采样算法对六个模型均有不同程度的性能提升，其中对支持向量机模型的 AUC 提升最大，AUC 值提高了 8.18%；对随机森林模型的精度和召回率提升最高，分别提高了 22.1%和 3.99%。相比于较为复杂的 Boosting 族模型，采样算法更能提升简单模型和 Bagging 集成模型的分类准确度。经过采样算法的优化后，CatBoost 模型成为预测车险欺诈风险的最优模型，其 AUC 值达到 0.79834。采样算法与机器学习模型的结合可以再次体现集成模型的整体优越性。精度和召回率的明显提升使得模型的预测结果更符合实际，提高了模型的实用性。采样算法在预测车险欺诈风险的问题上，能够普遍提升简单机器学习模型与集成模型的预测准

确度和性能，后续模型融合时将保留采样算法。

3.4 车险欺诈风险预测融合模型实证分析

Stacking 融合是一种分层模型集成框架。Stacking 融合的概念是学习几个不同的弱学习器，并通过训练一个元模型来组合它们，然后基于这些弱模型返回的多个预测结果输出最终的预测结果。以两层为例，第一层由多个基学习器组成，其输入为原始训练集，第二层的模型则是以第一层基学习器的输出作为特征加入训练集进行再训练，从而得到完整的 Stacking 融合模型。原理图示如图 3-2。

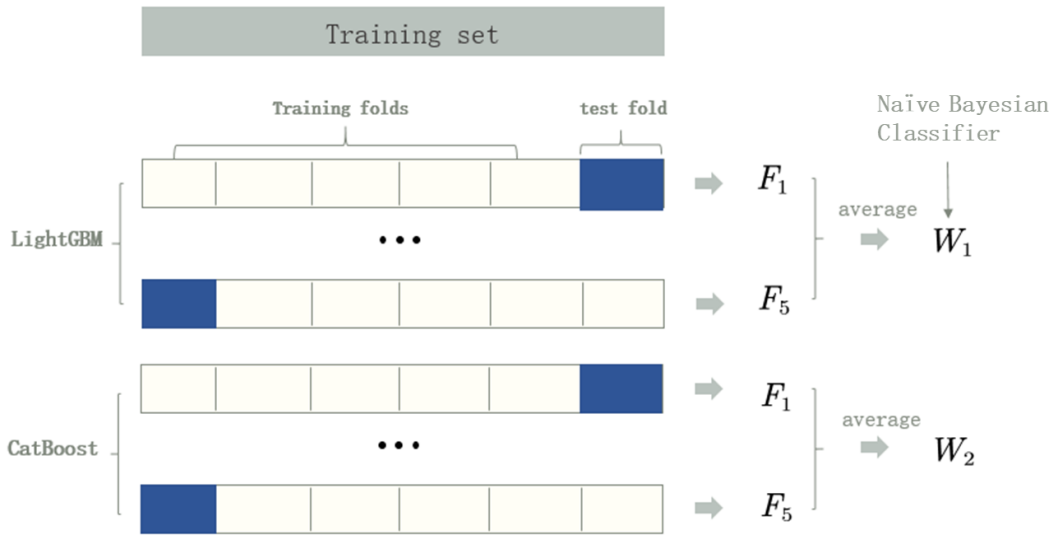


图 3-2 Stacking 融合方法图示

对于上一节中训练出的结合最优采样算法的六个模型，采用排列组合的方式，分别选用 2 至 6 个基模型，以基于伯努利朴素贝叶斯分类器的 Stacking 方法融合成新的模型，并与基模型的最高 AUC 值进行比较。分别有以下五种情况：

将任意 2 个模型进行融合，共 $C_6^2 = 15$ 种融合模型；

将任意 3 个模型进行融合，共 $C_6^3 = 20$ 种融合模型；

将任意 4 个模型进行融合，共 $C_6^4 = 15$ 种融合模型；

将任意 5 个模型进行融合，共 $C_6^5 = 6$ 种融合模型；

将 6 个模型进行融合，共 $C_6^6 = 1$ 种融合模型。

下面依次探索比较上述五种融合方式的结果：

选取任意 2 个模型进行融合的结果如表 3-14 所示。

表 3-14 任意 2 个模型融合结果

模型 1	模型 2	AUC	AUC 变化
支持向量机	决策树	0.71204	下降
	随机森林	0.72420	下降
	XGBoost	0.75337	下降
	LightGBM	0.76900	下降
	CatBoost	0.77146	下降
决策树	随机森林	0.75835	上升
	XGBoost	0.77488	下降
	LightGBM	0.78260	下降
	CatBoost	0.77931	下降
随机森林	XGBoost	0.74732	下降
	LightGBM	0.74826	下降
	CatBoost	0.75370	下降
XGBoost	LightGBM	0.79047	下降
	CatBoost	0.80433	上升
LightGBM	CatBoost	0.79149	下降

由表 3-14 可知，大多数融合模型的 AUC 介于基模型的两个 AUC 值之间，因此没有超过所选 2 个模型中最高的 AUC 值。输入层模型为决策树和随机森林、XGBoost 和 CatBoost 融合时效果最佳，AUC 值相较于所选单个模型时都表现为上升。XGBoost 和 CatBoost 融合模型为 2 个单模型融合的最优模型，AUC 值高达 0.80433，相较于单个模型最高 AUC 值(0.79834)提升了 0.599%。

选取任意 3 个模型进行融合的结果如表 3-15 所示。

表 3-15 任意 3 个模型融合结果

模型 1	模型 2	模型 3	AUC	AUC 变化
支持向量机	决策树	随机森林	0.74877	上升
		XGBoost	0.78013	下降
		LightGBM	0.77582	下降
		CatBoost	0.78105	下降
	随机森林	XGBoost	0.75958	下降

		LightGBM	0.75960	下降
		CatBoost	0.75410	下降
	XGBoost	LightGBM	0.79110	下降
		CatBoost	0.79862	上升
	LightGBM	CatBoost	0.78562	下降
决策树	随机森林	XGBoost	0.77778	下降
		LightGBM	0.77953	下降
		CatBoost	0.77626	下降
	XGBoost	LightGBM	0.79009	下降
		CatBoost	0.79667	下降
	LightGBM	CatBoost	0.79513	下降
随机森林	XGBoost	LightGBM	0.79135	下降
		CatBoost	0.79982	上升
	LightGBM	CatBoost	0.78716	下降
XGBoost	LightGBM	CatBoost	0.80416	上升

由表 3-15 所示，相比于单个基准模型或集成模型，将 3 个模型进行 Stacking 融合后，有 4 种组合的 AUC 值都得到了提升。其中将 Boosting 族三个模型，即 XGBoost、LightGBM、CatBoost 融合后的 AUC 最高，其数值达到 0.80416，相较于单个模型最高 AUC 值(0.79834)提升了 0.582%。

选取任意 4 个模型进行融合的结果如表 3-16 所示。

表 3-16 任意 4 个模型融合结果

模型 1	模型 2	模型 3	模型 4	AUC	AUC 变化
支持 向量机	决策树	随机森林	XGBoost	0.78339	下降
			LightGBM	0.78316	下降
			CatBoost	0.78171	下降
		XGBoost	LightGBM	0.78516	下降
			CatBoost	0.80008	上升
		LightGBM	CatBoost	0.78408	下降
	随机森林	XGBoost	LightGBM	0.78960	下降
			CatBoost	0.79982	上升
		LightGBM	CatBoost	0.78716	下降
	XGBoost	LightGBM	CatBoost	0.80067	上升
决策树	随机森林	XGBoost	LightGBM	0.79110	下降
			CatBoost	0.80132	上升
		LightGBM	CatBoost	0.78856	下降
	XGBoost	LightGBM	CatBoost	0.79958	上升
随机森林	XGBoost	LightGBM	CatBoost	0.79634	下降

由表 3-16 可知，4 个模型融合后 AUC 值提升的比率高，有 5/15 的模型效果得到了提升。存在 3 种组合的融合模型的 AUC 值超过了 80%，其中将决策树、随机森林、XGBoost、CatBoost 四个模型融合后的 AUC 最高，其数值达到 0.80132，相较单个模型最高 AUC 值(0.79834)提升了 0.298%。

选取任意 5 个模型进行融合的结果如表 3-17 所示。

表 3-17 任意 5 个模型融合结果

模型 1	模型 2	模型 3	模型 4	模型 5	AUC	AUC 变化
支持 向量机	决策树	随机森林	XGBoost	LightGBM	0.79241	下降
				CatBoost	0.79513	下降
			LightGBM	CatBoost	0.78562	下降
		XGBoost	LightGBM	CatBoost	0.79688	下降
	随机森林	XGBoost	LightGBM	CatBoost	0.78891	下降
决策树	随机森林	XGBoost	LightGBM	CatBoost	0.79834	不变

由表 3-17 可知，当模型融合的数量达到 5 个时，融合模型的 AUC 值不再提升，总体呈现出下降或不变的趋势。这种情况下，所有融合模型相较于预测能力最弱的单个模型支持向量机(0.71096)有所提升，但 AUC 均未超过最优单个模型 CatBoost(0.79834)。

选取 6 个模型进行融合时，AUC 值达到 0.79835，相较于单个最优模型 CatBoost(0.79834)提升了 0.001%，提升效果十分微弱。

综上所述，第一，考虑六个模型的所有排列组合，以 XGBoost 和 CatBoost 为融合模型输入层时为最优组合，其 AUC 值相比于支持向量机、决策树、随机森林、XGBoost、LightGBM、CatBoost 分别提升了 9.377%、7.895%、5.701%、1.031%、1.386%和 0.599%。这种组合能成为最优组合可能是因为 XGBoost 和 CatBoost 本身单个模型的效果已经相对较好，又由于两模型的原理差异，融合后增加了模型的泛化能力。第二，Stacking 融合时选取 2 个或 3 个单模型时模型提升效果最好，而选取 5 个以上的模型时已经几乎不再有提升，说明预测车险欺诈风险时并非基模型数量越多越好。

4. 研究结论与政策建议

4.1 研究结论

本文利用 Kaggle 数据科学社区的开源数据集，基于美国某车险公司的理赔与欺诈案例数据，首先通过支持向量机、决策树、随机森林、XGBoost、LightGBM、CatBoost 六种模型，分别单独训练和预测车险欺诈风险。然后将机器学习模型与采样算法结合，解决数据不平衡问题，通过实验寻找每种模型适用的最优采样算法。最后使用 Stacking 融合方法，将六模型进行排列组合，通过实验寻找预测欺诈风险的最优融合模型，将其作为最终模型。

基于上述研究内容，本文的研究结论如下：

（1）在车险欺诈风险预测前期，对数据集进行探索性分析发现：对于分类型特征变量，在收集时容易由于类别过多产生冗余信息，降低模型性能；数值型特征变量的量纲差异较大，在模型训练之前须进行标准化处理。产生较高相关系数的特征变量容易带来共线性，最好做删除处理。

（2）在对比单个机器学习模型时，发现 XGBoost 模型的预测准确度最高，其 AUC 值、准确率和精度都在所选模型中排名第一，召回率排名三。由于该模型优越的精度和召回率，使用 XGBoost 模型预测欺诈风险能够最有效地识别真欺诈案例，因为模型识别返回的欺诈案例中有高达 89% 的样本能够被认定为真欺诈案例。

（3）在将机器学习模型与采样算法结合以解决数据不平衡问题时，发现不同模型适用的采样算法和最优参数不同。支持向量机和 CatBoost 模型与 BorderlineSMOTE 算法结合最好，决策树模型最适用于 SMOTE 算法，随机森林模型与 SVM SMOTE 算法结合最优，而 XGBoost 和 LightGBM 模型最适用于 ADASYN 算法。经过采样算法优化后的训练集，均能提升所有模型的预测准确度，优化后 CatBoost 成为最优模型，AUC 值达到 0.79834。

（4）在保留采样算法，并用 Stacking 方法将单个机器学习模型融合为新模型时，发现融合模型能够进一步提升预测准确度，在将 XGBoost 和 CatBoost 融合时

使 AUC 值最高达到 0.80433。在选取要进行融合的模型的个数时，无需一味追求模型的丰富性，因为模型数量越多，并不意味着融合后必定提升预测效果。

4.2 政策建议

4.2.1 增强数据治理，提高数据质量

由于国内保险公司无一提供开源数据集，本文使用美国车险公司提供的数据集进行探索研究。首先，国内保险公司竞争激烈，为避免关键数据和客户隐私泄露，选择不公开理赔数据。但更重要的是，保险公司的理赔数据量巨大，并且维度较高，大多数保险公司并不能对数据进行良好地治理，一定程度上降低了业务效率。若国内保险公司能保证数据完整性，在国内外研究结论的基础上扩充数据维度，并成立专门的数据治理部门提升公司的数据质量，将会有更多的保险公司能够建立起自己的欺诈风险预警系统，从而进一步提升车险行业的风控能力。

4.2.2 创新技术手段，高效打击欺诈

使用和优化机器学习、深度学习方法，能够有效提高车险公司对欺诈风险的预测效率和能力。随着新兴模型和技术的发展，保险公司可以开发和应用更多数据挖掘技术。在未来的研究中，保险公司可以聚焦于小样本学习，用于解决二分类和多分类问题，同时可以很好地缓解车险欺诈风险预测问题中正样本稀缺的现状。保险公司也可以投入到知识图谱、联邦学习等方法的研究，致力于欺诈团伙的检测。同时，在数据丰富和实践经验足够时，保险公司可以基于专家规则，研究开发案因回溯检测方法，在交通事故的勘测阶段即可识别欺诈风险，从而提高打击欺诈的效率。

4.2.3 整合各方资源，完善共享平台

目前我国有部分保险公司共同建立了车险信息共享平台，实现了部分资源整合。美国和部分欧洲国家建立了跨国的保险反欺诈平台，形成较为完善的反欺诈系统。我国可以由政府牵头，建立起由保险公司、专家学者、交警部门等联合组成的

车险反欺诈组织，对关键数据和隐私数据脱敏后，在组织内共享数据和技术资源，提高欺诈风险识别能力。以共享高维数据为原料，以大数据和人工智能技术为动力，在政府带头作用下，促进车险反欺诈发展，为保险公司降本增效，维护社会公共安全。

参考文献

- [1] 马亮亮. 大数据背景下的车险反欺诈策略研究[D]. 上海财经大学, 2021.
- [2] 刘家养. 保险欺诈的博弈论与市场分析[J]. 商场现代化, 2008, 000(023):267-267.
- [3] 赵桂芹, 吴洪. 汽车保险市场中存在道德风险吗?——来自动态续保数据的分析[J]. 金融研究, 2010(06):175-188.
- [4] 车险反欺诈联合课题组. 车险欺诈与反欺诈问题研究及监管建议[J]. 保险研究, 2021, No.398(06):3-10.
- [5] 刘兴跃. 基于保险公司视角的车险理赔反欺诈研究[D]. 山东大学, 2017.
- [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).
- [7] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 10-18.
- [8] 王海巍. 我国险企运营中道德风险甄别问题研究——以大数据 Hadoop 聚类分析技术为视角[J]. 保险研究, 2016 (2).
- [9] 廖新年. 国际反保险欺诈经验借鉴与思考[J]. 上海保险, 2010(03): 22-24+44.
- [10] 段冉. 大数据时代下我国互联网保险的法律规制研究[D]. 江西财经大学, 2018.
- [11] 汤俊, 莫依雯. 基于数据挖掘技术的车险反欺诈系统构建[J]. 上海保险, 2013(11):39-42.
- [12] 徐徐, 王正祥, 王牧群. 基于深度学习技术的机动车辆保险欺诈识别模型与实证研究[J]. 上海保险, 2019(8):53-58.
- [13] 袁幕琴. 保险欺诈与大数据时代下的防范对策[J]. 金融与经济, 2015(7):77-80.
- [14] 焦清宇. 机器学习助力机动车辆保险反欺诈研究[D]. 辽宁大学, 2022.
- [15] 黄章杰. 基于机器学习的信贷违约预测研究[D]. 重庆工商大学, 2022.
- [16] Arrow, K.J. Uncertainty and the Welfare Economics of Medical Care[J]. American

Economic Review 1963(53):941-973.

- [17] Artis M , Ayuso M , Guillen M . Modeling Different Types of Automobile Insurance Fraud Behavior in the Spanish Market[J]. Insurance Mathematics and Economics, 1999, 24(1-2):67-81.
- [18] YAN C, LI Y, LIU W, et al. An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification[J]. Neurocomputing, 2020, 393:115-125.
- [19] Viaene, Stijn, Derrig, R, Baesens, Bart, et al. New developments in insurance fraud detection modeling: a comparison of state-of-the-art classification techniques for expert automobile insurance fraud detection[J]. Wharton School Penn State University, 2005.
- [20] KAŠĆELAN L, KAŠĆELAN V, NOVOVIĆ- BURIĆ M. A data mining approach for risk assessment in car insurance: evidence from Montenegro[J]. International Journal of Business Intelligence Research, 2014, 5(3):11-28.

附录

本文使用数据集的所有字段及其说明如下：

字段	说明	字段	说明
policy_number	保险编号	incident_severity	事故严重程度
age	年龄	authorities_contacted	联系了当地的哪个机构
months_as_customer	成为客户的时长，以月为单位	incident_state	事故所在省份
policy_bind_date	保险绑定日期	incident_city	事故所在城市
policy_state	上保险所在地区	incident_location	事故所在地点
policy_csl	组合单一限制 Combined Single Limit	incident_hour_of_the_day	出事所在的小时
policy_deductable	保险扣除额	number_of_vehicles_involved	涉及的车辆数
policy_annual_premium	每年的保费	property_damage	是否有财产损失
umbrella_limit	保险责任上限	bodily_injuries	身体伤害
insured_zip	被保险人邮编	witnesses	目击证人
insured_sex	被保人性别	police_report_available	是否有警察记录的报告
insured_education_level	被保险人学历	total_claim_amount	整体索赔金额
insured_occupation	被保险人职业	injury_claim	伤害索赔金额
insured_hobbies	被保险人兴趣爱好	property_claim	财产索赔金额
insured_relationship	被保险人关系	vehicle_claim	汽车索赔金额
capital-gains	资本收益	auto_make	汽车品牌
capital-loss	资本损失	auto_model	汽车型号
incident_date	出险日期	auto_year	汽车购买的年份
incident_type	出险类型	_c39	备注
collision_type	碰撞类型	fraud_reported	是否欺诈

致谢

感谢我的论文导师张婷婷老师对我毕业论文的帮助。从开题到论文定稿，张老师学术上的一针见血和语言上的不断鼓励使我受益匪浅，在此向老师表示衷心的感谢，愿老师身体健康，工作顺利。

感谢父母家人为我提供了坚实的后盾，让我在大学中自由闯荡。物质、精神与爱，我会用余生予以回报。

感谢我的辅导员杜老师和我的朋友们与我大学四年的陪伴。一千多个欢声笑语的日子就要结束，愿大家展翅高飞，我们顶峰相见。

感谢大学四年的自己，不断探索、目标清晰、不畏挑战。未来我将不忘初心，在新的生活中闯出一片天地。

最后祝愿各位平安喜乐，万事顺意！

