中图分类号：　　密级：

学科分类号：　　论文编号：192104010

山东财经大学

硕 士 学 位 论 文

（学术学位）

人工智能在保险反欺诈中的应用研究

作者姓名：周鹤鸣

学科专业：风险管理与精算硕士

指导教师：闫庆悦（教授）

培养学院：保险学院

二〇二二年三月十三日

Research on the application of artificial intelligence in insurance anti-fraud

A Dissertation Submitted for the Degree of Master

Candidate：Zhou Heming

Supervisor：Prof. Yan Qingyue

School of Insurance

Shandong University of Finance and Economics

中图分类号： **密级：**

学科分类号： **论文编号：192104010**

硕 士 学 位 论 文

人工智能在保险反欺诈中的应用研究

作 者 姓 名： 周鹤鸣 申请学位级别：硕士

指导教师姓名：闫庆悦 职 称：教授

学 科 专 业： 风险管理与精算硕士 研 究 方 向：保险反欺诈

学 习 时 间： 自 2019 年 9 月 1 日 起至 2022 年 6 月 30 日 止

学位授予单位：山东财经大学 学位授予日期：2022 年 6 月

山东财经大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得山东财经大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：[signature]　　日期：2022年 5 月 23 日

山东财经大学学位论文使用授权声明

本人完全同意山东财经大学有权使用本学位论文(包括但不限于其印刷版和电子版)，使用方式包括但不限于:保留学位论文，按规定向国家有关部门(机构)送交学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名：[signature]　　日期：2022年 5 月 23 日

指导教师签名：[signature]　　日期：2022 年 5 月 23 日

摘 要

保险欺诈是保险业面临的一大现实难题，严重扰乱保险市场秩序。风险存在于人们生活工作的方方面面，这就决定了保险公司的经营范围需要囊括各个方面，其具有广阔的业务对象。然而由于保险公司不能完全获取关于投保人的相关信息，因此在保单的订立过程中，保险人处于相对弱势的地位，造成保险诈骗活动日益频繁。为进行保险反欺诈，保险学界针对保险反欺诈行为识别进行了很多种不同的尝试，例如使用 PRIDIT 模型、Probit 模型、专家规则系统等，但类似这些方法仅仅只能处理维度相对较低的数据，在处理规模较大并且维度较高的数据时，能力受到一定的限制。信息技术的快速发展，人工智能化的广泛应用，保险数据也会呈现出多维度的发展趋势，但是不管是人工识别又或者是 Logistic 回归、SVM、ELM 等技术都不能识别处理该类新兴数据，因此保险业对新的技术渴求迫在眉睫，人工智能技术就是一个非常合适的选择，因此本文围绕人工智能在保险反欺诈过程中的作用进行研究。

本文利用保险公司往期保险理赔数据，借助大数据分析、机器学习模型以及深度学习模型分别对真实欺诈事件数据进行识别，以此研究了人工智能技术对保险公司反欺诈的作用。本文首先通过训练不同模型，并对各训练后模型的预测准确率高低进行对比，寻找出对保险反欺诈的识别能力相对最高的模型，然后再并在该模型的基础上进行进一步优化，得到最优模型。最后根据最优模型的预测结果提出相应的结论。本文主要结论如下：

第一，在模型横向比较中，基线模型在设定提调率条件下的精准率指标均优于本组其他模型指标。各种深度学习模型的性能基本相同，其中 DCN 模型是其中的代表模型之一。

第二，在模型纵向比较中，最终模型在设定提调率条件下的精准率指标均优于基线模型，中间模型相较于基线模型预测能力有所下降，但是作为中间环节，其预测能力的强弱不具备较强的参考性。

第三，最终模型在设定准确率基准的情况下，提调率指标的表现远远优于基线模型。最终模型的单条数据测试响应时间指标表现一般，说明模型还存在去粗存精的可能性。

综上，本文证实了人工智能对保险反欺诈具有促进作用，并且构建出了具有良

好保险反欺诈识别能力的模型，成功找到能够进行保险反欺诈的应用方式，因此保险公司应该大力提高人工智能水平，以期在未来可以用机器取代人工，一方面，可提高自身保险反欺诈识别能力，减少保险理赔欺诈率以及保险公司损失，另一方面，也可以倒逼欺诈者减少欺诈行为，帮助构建良好的保险业生态。

关键词：人工智能；保险反欺诈；特征工程； XGBoost 模型

Abstract

Insurance fraud is a real problem faced by the insurance industry, which seriously disrupts the order of the insurance market. Risks exist in all aspects of people's life and work, which determines that the business scope of insurance companies needs to cover all aspects, and it has a wide range of business objects. However, because insurance companies cannot fully obtain relevant information about policyholders, the insurers are in a relatively weak position in the process of formulating insurance policies, resulting in increasingly frequent insurance fraud activities. In order to conduct insurance anti-fraud, the insurance academia has made many different attempts to identify insurance anti-fraud behaviors, such as using PRIDIT model, Probit model, expert rule system, etc., but similar methods can only deal with relatively low-dimensional data, When dealing with large-scale and high-dimensional data, the ability is limited. With the rapid development of information technology and the wide application of artificial intelligence, insurance data will also show a multi-dimensional development trend, but neither manual identification nor Logistic regression, SVM, ELM and other technologies can identify and process such emerging data. Therefore, the insurance industry is imminent for new technologies, and artificial intelligence technology is a very suitable choice. Therefore, this paper focuses on the role of artificial intelligence in the process of insurance anti-fraud.

This paper uses the insurance claims data of insurance companies in the past, and uses big data analysis, machine learning models and deep learning models to identify real fraud event data, so as to study the role of artificial intelligence technology in anti-fraud of insurance companies. In this paper, by training different models and comparing the prediction accuracy of each model after training, we find the model with the highest identification ability of insurance anti-fraud, and then further optimize the model based on the model to obtain the best insurance anti-fraud recognition ability. Excellent model. Finally, the corresponding countermeasures and suggestions are put forward according to the prediction results of the optimal model. The main conclusions of this paper are as follows:

First, in the horizontal comparison of models, the accuracy rate indicators of the baseline model under the condition of setting the adjustment rate are better than other model

indicators in this group. The performance of various deep learning models is basically the same, and the DCN model is one of the representative models.

Second, in the longitudinal comparison of models, the final model's accuracy index under the condition of setting the adjustment rate is better than that of the baseline model. Compared with the baseline model, the prediction ability of the intermediate model is lower, but as an intermediate link, its prediction ability is higher than that of the baseline model. Strength is not a strong reference.

Third, the final model far outperforms the baseline model on the upscaling rate metric when the accuracy benchmark is set. The performance of the single data test response time index of the final model is average, indicating that the model still has the possibility of removing the rough and saving the fine.

To sum up, this paper confirms that artificial intelligence can promote insurance anti-fraud, and builds a model with good insurance anti-fraud identification ability, and successfully finds countermeasures against insurance fraud. Therefore, insurance companies should vigorously improve the level of artificial intelligence, with a view to In the future, machines can be used to replace labor. On the one hand, it can improve its own insurance anti-fraud identification ability, reduce insurance claims fraud rate and insurance company losses, on the other hand, it can also force fraudsters to reduce fraud and help build a good insurance industry ecology .

Key words: Artificial intelligence; Insurance anti-fraud; Feature engineering; XGBoost model

目 录

第 1 章　绪论

1.1　研究背景与意义

自从改革开放至今，我国的保险行业经历了飞速的发展阶段，分析 2022 年的保险业发展报告（发布于保监会）可知，上一年，中国保险行业的原保险保费收益累积已达到 4.49 万亿，已成为世界第二大保险缴费最高的国家。虽然我国的保险业发展趋势欣欣向荣，然而保险欺诈现象却时有发生，保险损失的数额也逐年增加。在上个世纪末，西方等资本主义国家曾出现因为保险诈骗活动的频繁发生导致保险行业的市场秩序混乱的局面出现。首先，由于欺诈行为的时有发生，保险公司被迫提升保费，消费者因此需要支付更多的金钱。其次，保险公司正常生产过程中的经营管理难度也因为欺诈行为的出现而有所提升，投入的管理成本因此加大。按照《新金融》报道，至 2018 年 11 月，财险寿保事后稽查结束后，因保险诈骗致使公司损失的总额额已经超过了八十亿元，按照此概率进行估算，那么整个保险行业的总减损金额应该超过 200 亿，并且这个总减损仅占没有识别的保险欺诈中的很小一部分。按照相关数据推测，国内商业保险欺诈损失率已经超过 10%，车险的损失能够超过 20% [1]，部分省份甚至能够达到 30%[2]，但是依据 IAIA 等国际组织的相关数据可知，在总赔付额当中，保险欺诈数额的占比已达到了 10%～20%。基于此，保监会于 2018 年 2 月份专门颁布了《反保险欺诈指引》，以指引保险机构对反欺诈采取一定的举措，维护自身的权益。保险欺诈发生频率如此频繁，严重影响了国内保险业的稳定发展，更有甚者，造成了金融体系的动荡。

为了规制保险欺诈行为，理论界和实务界采取了各种手段和方法，像专家系统、Probit 模型、PRIDIT 模型等，然而这些处理方法对大规模的数据处理产生的作用微乎其微。随着电子信息的迅猛发展以及大数据时代的到来，不管是人工识别又或者是 Logistic 回归、SVM、ELM 等技术都不能识别处理该类多维度的复杂数据，保险业对新的技术渴求迫在眉睫，人工智能是一个非常合适的手段。人工智能技术可以通过识别保险理赔中的欺诈风险来帮助保险公司更好的进行理赔。例如可以通过图像处理及分析技术来识别客户上传虚假照片的行为，避免保险公司被客户所使用的的网络图片或处理照片进行欺诈，骗取保险赔付。或者通过对成千上万个客户数据的分析可以识别各个客户之间的潜在联系，从而在内部构建起全面的涉嫌欺诈案件

人员社会关系网络，让保险公司在投保环节之前得到风险预警；二从长期的角度看，保险公司可以使用生物技术来对客户的身份进行识别核验，从而避免出现客户身份被顶替的骗保现象。此外，利用语音识别技术还可以通过分析客户的情绪特征等指标实现欺诈指数的测算等。

因此，本文通过对保险公司以往的保险理赔经验数据进行分析，探究人工智能技术在保险公司反欺诈过程中作用，一方面可以为现有的保险反欺诈理论研究提供实证分析层面上的数据支撑，另一方面也能为保险公司提高自身反欺诈手段提供一些借鉴。

1.2 研究思路与方法

1.2.1 研究思路

本文根据文献研究提出假设，然后进行理论分析以及实证分析验证可操作性，最后得出研究结论以及可以改进之处。

本文主要研究人工智能在保险反欺诈中的应用。首先简要说明当前科技水平下保险业的反欺诈业务发展的情况，即保险公司对保险反欺诈识别主要依靠业务人员的经验以及人工总结的专家规则，其次简要介绍了人工智能在优化保险公司反欺诈业务流程能够起到的作用，给出文章的主要创新点及不足之处。结合上述理论分析和现状分析，本文提出人工智能可以深刻影响保险业反欺诈的判断机制且大数据技术和机器学习技术在其中发挥重要作用的假设。

在实证部分将主要介绍数据来源和相关数据的处理及应用过程，采用国内某财险公司截至 2020 年 4 月的车险理赔数据，首先通过特征工程对数据进行一系列的预处理，而后运用机器学习和深度学习分别进行模型训练，首先在结论与建议部分按照有关数据对全文进行总结。本文方法的应用可行性以及现阶段的不足之处，之后提出相关展望。

1.2.2 研究方法

本文主要通过文献回顾法、实证分析法和描述性研究法对所关注的问题进行学习梳理和分析。

第一，文献回顾法。通过回顾保险欺诈的现象、人工智能的现状以及关于人工智能对保险反欺诈的影响作用及其对策的文献，发现我国商业保险公司普遍面临保

险欺诈的问题，而且这一问题正受到越来越多的人关注，研究此话题具有一定的现实意义。

第二，描述性研究法。理解目前的商业保险反欺诈模式、保险欺诈的特性以及人工智能各种模型的理论，对其给予叙述和解释，从而提出人工智能在保险反欺诈中的应用的问题，并进一步通过数据处理和检验的结果描述现象，最后得出结论。

第三，实证分析法。探究人工智能对保险反欺诈的影响，并进一步寻求其中的影响机制，以及通过回归树模型、深度学习模型以及特征深度衍生和模型优化较为全面的研究人工智能在保险反欺诈中的应用，对进一步提高保险公司反欺诈能力有重要意义。

1.3 研究内容与框架

1.3.1 研究内容

为研究人工智能在保险反欺诈中的应用，本文研究内容如下：

第一部分为绪论。对本文的选题背景和意义进行阐述，得出我国存在商业保险欺诈率高，人工识别反欺诈效果不佳、成本过高的现状，通过研究人工智能对保险反欺诈的作用可以得出人工智能在保险反欺诈领域具有良好应用前景。与此同时，本文第一部分对本文的研究方法与思路进行详细介绍，同时指出本文的不足和创新之处。

第二部分为文献综述。通过分析现有的文献资料，简要概括了机器学习、深度学习研究现状，以及人工智能在保险反欺诈中的作用，为后续研究的进行提供了思路和指引。

第三部分为相关理论基础。介绍了与保险欺诈相关的的保险概念，以及保险欺诈成因的信息不对称理论。通过文献综述和相关理论学习，得出了人工智能在保险反欺诈业务中的应用逻辑，为后续数据分析提供有力支持。

第四部分为模型理论。重点介绍了 GBDT 模型理论以及进一步的 XGBoost 模型理论，即本文的主要模型理论，并通过数学公式推导阐述了模型的基本原理；其次简要介绍了深度学习模型理论，其中利用数学推导过程稍作介绍了基础模型 WDL 模型的原理，以及简要介绍了另五种在 WDL 模型基础上变化的深度学习模型。

第五部分为本文的实证部分，利用国内某财险公司截至 2020 年 4 月的脱敏后车

险理赔数据对文章提出的假设进行数据分析、多种模型训练和模型结果比较，并通过对模型的进一步优化得到的最终结果得出人工智能可以强化保险公司的保险反欺诈识别能力的结论。

第六部分为结论和展望。该部分对前文的理论分析以及实证分析结果进行了充分的总结，并基于此提出了人工智能在保险反欺诈中的应用展望。

1.3.2 研究框架

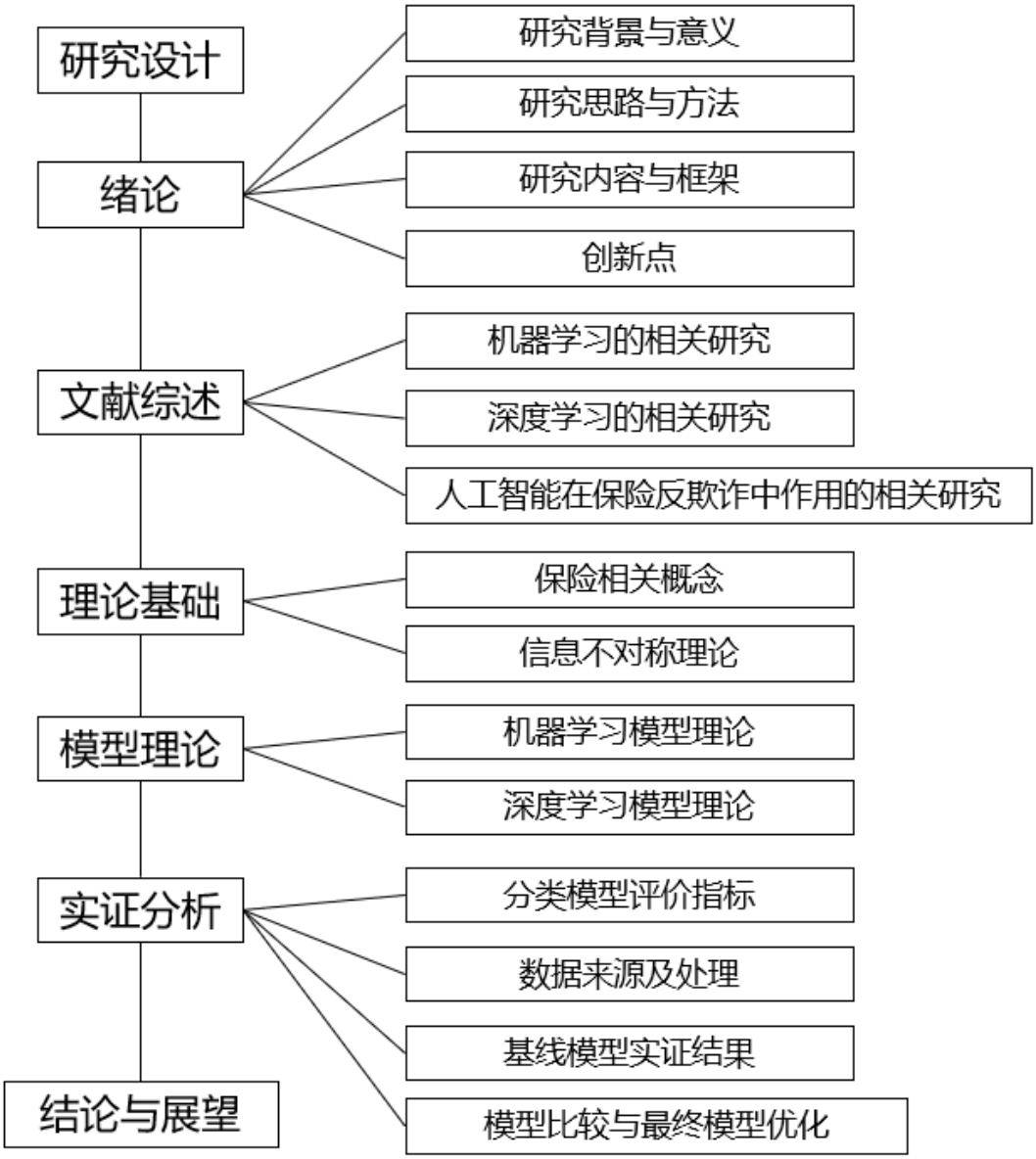


图 1-1 研究框架

1.4 创新点

（1）以往关于保险反欺诈的文献中，主要研究的是人工智能对保险行业的整体影响。由于保险反欺诈涉及到的保险公司核心数据较难获取，当前国内外学者对于人工智能和保险反欺诈的研究较少，且大多运用定性分析的方法或仅进行理论知识推理。本文首次利用保险公司的非公开数据进行研究。

（2）在模型测试过程中由于预设不同的提调率会得到不同的模型准确度，于是本文将提调率添加为模型的指标之一结合准确率进行分析，线性分段取值，能更合理地观测人工智能模型在保险反欺诈识别中具备的能力。

（3）在模型调整过程中，本文对特征工程模型进行了进一步优化，最终形成了 SAFE 半自动数据特征分类+SF 特征深度衍生+XGBoost 模型的格局，创新特征工程模型的构建方法，使其数据处理能力更加适配机器学习模型。

第 2 章　文献综述

在金融经济学术界，保险反欺诈问题历来是一大研究热点，自 1979 年起，国内的保险产业渐渐开始恢复并稳健发展至今，现已发展了 40 几年，然而与欧美国家近百年的保险发展史相比，仍差距甚远。不仅如此，欧美国家的保险行业市场化程度较高，较早的繁荣也使得保险欺诈现象猖獗，因此西方各类学者对保险反欺诈问题的研究也早已开始。相对来说，国内的相关研究较晚才开始。此章首先对国内、国外的有关文献资料进行整理，并对现有研究成果进行总结，找出其中存在的不足，然后对此次研究的目的做出说明。

2.1　机器学习的相关研究

Jerome Friedman（1999）提出梯度提升决策树模型（GBDT）[3]。进入新世纪之后，在机器学习被运用于各大领域的同时，它在保险期许识别领域也得到了应用。Stijn Viaene 等人（2002）使用 Logistic 回归、KNNC4.5 决策树等多种算法对 993 年马萨诸塞州的人身伤害保护索赔数据集的数据进行测量对照，研究结果表明 C4.5 决策树结果与实际情况不太相符[4]。Clifton Phua 等人（2004）又对决策树模型、BP(反向传播)、(NB）朴素贝叶斯依次做了预测，再利用基础分类工具与 bagging 技术结合在一起，结果表明和最佳分类器决策树模型比较来说，bagging 后模型的性能比相对高一些，而且通过组合三种模型能够有效地节约数据成本[5]。Wavier Muguerza 等人（2005）采用分类树模型对汽车保险公司欺诈行为进行预测，并将其和合并树以及 C4.5 树进行准确性以及结构稳定性（解释能力）的对比，此外，还大量分析了 RQC 曲线、召回率等的误差[6]。LI. Bermudez 等人（2008）运用 Gibbs 抽样、数据放大的办法，经过对 logit 模型进行改造来来解决由于信息不对称造成的保险欺诈，相关数据表明，正确分类的案例数量在模型估计后显著提高。我国研究者萧超武等人（2014）运用德国信誉信息对比分析了随机森林、KNN 等模型，结果发现，与其他模型相比，随机森林模型与 GBDT 模型的预测具有更高的稳定性、精准性，并且与 GBDT 模型相比，随机森林模型的平均精准率要高出 1.82%[7]。之后 T. Q. Chen（2016）等人以 GBDT 为基础，研究出了 Xtrerne Gradient Boosting（即 XGBoost）算法，这一算法不仅对 GBDT 模型做了大量优化，还很好的将 GBDT 的全部功能延续了下来，之后，其模型也在信誉风险评估领域获得了大量的运用[8]。 Yufei Xia（2017）等人

针对商业银行借贷业务的信誉，借助 XGBoost 模型完成了相应的风险评估，得出的结果显示，在 AUC 以及精准性等方面，该模型的表现非常不错[9]。王重仁（2019）等人利用贝叶斯参数对 XGBoost 模型做了改进，并且经实证分析证实，相比神经网络、Logistic 回归等模型， XGBoost能够得到更好的预测结果[10]。2020 年，叶成、程云辉等人依据 Bagging 自助采样（在随机森林模型之中），提出了全新的一个预测之法，即利用自助采样法的 Stacking 集成法，此法能够实现数据的多次采样，随后把通过采样获得的信息适合运用训练基分类器副本，让其副本依次投票，进而确定其最终决策，在测量多次后，得出的研究结论说明，改造后的集成学习方法与同一架构的Stacking集成法相比，能够获得更理想的F1、查准率以及精准率[11]。Swetha P、Dayananda B 在 2021 年提出的 Improvized-XGBoost 模型中自带特征函数，他运用该模型来对电信业的顾客流失做预测，他先利用这个模型与 XGBoost 模型结合完成特征函数的创建，然后再利用迭代之法来完成损失函数的创建，构建特征函数，结果显示 Improvised-XGBoost 模型的数据与其他模型相比具有更良好的召回率、准确率、查准率等，其效率水平比较高，能够适用于收集和处理复杂数据[12]。

2.2 深度学习的相关研究

在大数据时代，深度学习成了网络技术研究领域的一大热点，而各种保险的理赔场景正好会产生数量庞大数据，是非常适用于实践深度学习的一个场景，现如今在保险反欺诈中运用深度学习可以说是该领域热度极高的研究方向之一。

谷歌团队在 2016 年提出了全新的一种互联网结构，用途主要是推荐谷歌应用市场上的软件，即 Wide& Deep 网络框架。该框架的目的是提高 App 推荐算法的精准程度，并在提高精准度的同时兼顾推荐系统的可扩展性[13]。谷歌第一次将深度模型融入到传统特征项目当中，借助模型 Wide 层来对大量的简单关联性特点进行学习，运用其 Deep 层来完成深度网络的建立以融合特征。借助此方法实现的高维疏松特征减维法以及持续特征处理，时至今日仍为人所津津乐道。Wide& Deep 模型是典型的深度推荐模型之一，后续的不少模型都是对其进行优化后得到的。关于 WDL 模型的交叉形式（特征工程），在 2016 年，微软所提出的 Deep Crossing 将人工特征的出现彻底隔绝了，在获取每个特征之下的详细特征，以特定任务为依据自行检索最理想组合并将其实现（Embedding初始特征之后，在神经网络层中将其直接输入，并由模型来完成对处理全部特征交叉的工作）。而新出现的 DIN 更是使分段拟合思想得到了

深化，attention 机制的加入了使其实现了 LocalActivation，从而能够实时学习吸引用户的 embedding 向量。DCN 同样是近些年的主流模型之一。K+1 层的特征重叠经由 Wide 对 k 层来实现，同时还能够用于具体特征重叠的实现。Deep 层的功能是对高阶重叠特征进行捕捉。在 2010 年， Vincent(外国学者)提出了不存在监督的一种神经网络模型，是由单层 DAE 进行堆叠从而形成的一种深度学习模型，且在图像类型划分、识别行为等领域已获得了大量运用，在预处理模型信息、提升模型泛化水平方面，其表现非常出色。过去几年，在模型的影响领域逐渐扩大，众多学者渐渐研究出了其在各大领域场景中的应用。Zheng 等人在检测电力盗窃方面，运用了 CNN 模型（其网络结构为 Wide& Deep），首先针对 Deep 模型，就输入数据，他们利用 3 层卷积神经网络对其做了更深入的特征提取操作，之后把提取出来的特征与一层的全连接层和池化层相连，最终获得了输出结果，得出的结果能够适用于电力盗窃问题的解决，帮助电力供应局降低亏损。[14]。Niu 等在专利和论文分类的任务中使用 Wide& Deep 网络框架，在 Deep 模型中，他们率先运用 embedding 层完成了对初始文本信息的量化操作，然后把理化过的信息录入至一层卷积神经网络中又提取信息特征，再运用最大程度池化这一方式来减少特征数量，最后获得 Deep 模型输出的数据，随后拼接输出和 Wide 模型的特征，再将有关结果输入到 sigmoid 激活函数，从而得出数据的分类，帮助专利和论文的相关工作人员对能够相对迅速的将信息类型识别出来 [15]。对于病症识别，Nguyen 等人同样运用了 Wide & Deep 模型结构，对于 Deep 模型，他们运用的是双层完全相连的网络形式，随后把上述结果输入 ReLU 激活函数最后，得到 Deep 模型的数值，接着拼接输出和 Wide 模型的特征，再在 ReLU 的激活函数中将以上结果代入得出最后数据，从而使模型预测糖尿病患情况的水平再次得到提高 [16]。Bastani 等人在 P2P 借贷模式中运用 Wide& Deep 互联网模型来研究预测贷款者出现逾期的机率，该模型和 Cheng 等的模型存在相同之处，采用考察五个方面的数据，分别是贷款特征以及借款人的个人信息、信用记录、信用评级和债务 [17]。

2.3 人工智能在保险反欺诈中作用的相关研究

保险欺诈的存在严重危害整个保险行业发展。业内普遍认为欺诈出现的原因是因为赔付管控存在薄弱环节，才产生的套保、骗保行为。最开始对于保险反欺诈的研究大体有以下两种：:1、识别欺诈性索赔的发展态势，2、对行业的综合数据进行

研究，分析其对经济系统的产生的作用[18]。上个世纪末以来，在计量经济学和电子信息技术的迅猛发展的情况下，科研人员开始运用计量经济模型对保险个人欺诈相关的数据展开研究分析。利用回归框架，Weissberg An Derrig（1993）对某些欺诈指标（如事故情况、保险特征、医疗等）进行了筛选，此外还依照欺诈指标数分析了保单存在欺诈的可能性[19]。Patrick L. Brockett 等人（1995）以 65 个指标为依据利用神经网络模型对车险诈骗的可能性做了预测。在先辈研究成果的基础上，Dionne、Belhadji（1997）研制出了用以帮助保险机构的理算师预判投保人的行为是否为欺诈的新二元 logistic 模型[20]。Mercedes Ayuso 等人（1998）利用效用模型研究了西班牙国内一家保险机构的索赔数据，得到的结果显示多项式 logit 模型的预测性与之前的模型相比更加稳定[18]。通过对以上数据进行分析，我们能够发现当时普遍运用统计建模技术，例如运用分析识别、逻辑回归等方式来识别保险诈骗。

近现代对于保险欺诈的识别问题产生了很多种方法，对这些方法进行分类主要可以分为两大类，一种是辅助学习方法，另一种是非辅助学习方法。在这之中，辅助学习方法是指在数据中随机挑选的一部分训练样本，通过归纳样本中数据特征与因变量（即该样本是否属于欺诈案件）的内在联系，而获取可以作为因变量的识别因子的方法，具体来说，辅助学习方法主要包括离散选择模型[21-23]、其他标准计量模型[24-26]以及专家规则系统[27-28]等。

而另一种方法，即非辅助学习方法与辅助学习法有根本上的不同。这种方法不需要依靠现有既存的关于因变量的信息来进行对于欺诈的判定与识别，而是直接从自变量中独立挖掘识别因子。以现有的保险欺诈识别文献来说，他们中的大部分都使用了非辅助学习方法。具体来说，费辅助学习方法主要包括聚类分析方法、Brocketti 等（1998）使用的非辅助的神经网络方法[29]以及其他种类的数据挖掘方法等[30-32]。Major＆Riedinger（2002）对美国某保险公司的关于理赔案件欺诈行为进行识别的欺诈电子识别系统进行了全面分析，并总结出了一个包含财务、物流等识别指标在内的分析框架[45]。

2.4 简要述评

通过对相关文献的梳理，可以发现欺诈控制模型逐渐受到保险理论家的广泛关注，近年来，随着科学技术水平的高速发展，各种新科技、新手段对保险行业产生了巨大影响，科技可以重构保险业的全环节，人工智能时代必然来临，势不可挡，

而科学技术水平的提升将成为保险行业反欺诈业务能力提升的核心驱动力。以往关于人工智能对保险反欺诈的研究主要集中在理论领域，由于保险公司对于各自数据的不公开也导致保险反欺诈相关的实证研究较少。

因此，本文试图在国内外已有文献的基础上，利用机器学习和深度学习等人工智能手段，结合大数据挖掘与数据分析，探讨人工智能对保险反欺诈的作用，并寻找出具有较强的在保险反欺诈领域中的识别能力的人工智能模型。希望通过提升大数据分析能力以及人工智能的水平，提高模型的保险反欺诈识别能力，为保险公司反欺诈提供解决方案，尽可能消除道德风险以及黑色产业链造成的影响，促进保险市场健康稳定发展。

第 3 章　相关概念与理论基础

本章首先根据现存文献的学习介绍保险相关概念以及保险反欺诈理论，从源头上分析保险欺诈产生的逻辑过程，再结合人工智能的应用特性得出人工智能对保险反欺诈的作用机理。

3.1　欺诈与保险欺诈

从法律角度来说，世界上大部分的国家都做出了如下规定，当行为人有义务说明真实情况时，故意不做出说明或保持沉默，均被认为是欺诈行为。我国的《民法通则》、《合同法》等法律法规以及最高人民法院均对“欺诈”的定义及导致的后果做出了规定或说明。按照《民法通则》的有关规定，一方使用不正当手段，比如欺诈、胁迫等，造成另一方处于弱势地位从而达成合同的的民事法律行为无效。《合同法》又进一步详细规定，在上述情形下订立的合同能够采取变更或者撤销的手段维护无过错方的利益；《最高人民法院关于贯彻有<中华人民共和国民法通则>若干问题的意见（试行）》第 68 条提到：当事人故意将实情隐瞒或将虚假情况提供给对方，诱使另一方当事人做出与其有利竟未表示者，可判定其行为为欺诈。

保险欺诈在国际领域通常也被称作是保险犯罪，但严格说来，与保险犯罪相比，保险欺诈的含义更广。保险交易过程中当事人双方都有可能出现保险欺诈。进行保险交易时，投保人只要存在与诚信原则不符的情况，例如刻意不将与保险标的有关的实情告知对方，从而成功诱使交易另一方承保，或者是针对保险合约规定，投保人以捏造保险案例、制造虚假案例的行为使保险机构蒙受经济损失，从而得到保险赔偿的情况，皆为投保方的保险欺诈。除此之外，若保险人在其本人偿付实力欠缺的情况下，或未获得相关部门批准前提下经营相应业务，同时借助其自行制定保险合同条款、保费费率等机会，或者有夸大保险责任范围、收益等从而诱导、欺骗投保人进行投保的，均属于保险人欺诈行为。保险欺诈行为一经实施，则必然造成严重的危害结果，因此有必要对保险欺诈行为严加防范。

3.2　信息不对称理论

M.Spence、George Arthur Akerlof, Joseph Eugene Stiglitz 三位美国经济学家得了 2001 的诺贝尔经济学奖，他们三人对“不对称信息市场”的研究为经济学的发展做

出了重要贡献。他们指出，所谓信息不透明，指的是在市场经济行为当中，经济主体所了解的信息的深度有所不同，处于信息优势地位的一方会利用自己的优势地位损害另一方当事人的合法利益。

进行商业活动时，当事人双方所掌握的信息通常都是不对称的。掌握更多有效信息的主体具备损人利己的特点，再加上世人对获利和损利所持心态有别，为获取更多利益而不惜使不具信息优势的一方利益受损从而实现自身目的。关于信息不对称， James Alexander Mirrlees 的解释是:在交易当中，其中一方当事人知道另一方当事人不知道的消息。他把不对称信息的问题分为以下两种:首先是外生的不对称信息，像当事人的兴趣爱好、健康状态等比较隐私的信息，这种信息和当时为的行为联系不大，是一种事前的信息。詹姆斯把该类信息叫做隐藏知识，隐藏知识的存在有可能会带来逆向选择问题。这种隐匿性信息通常产生于合同未生效以前，例如，一家公司在招录新人的过程中，雇主并不能准确了解某一特定应聘者的各项能力，而该应聘者对自身情况的了解程度远远超过雇主的了解程度，因此雇主为了获取可以用以判断某一应聘者能力的信息有必要采取一些考核的方式，或者通过其他方式来诱使劳动者主动将与其本人实力相关的信息公开，如此方能使雇主做出与自身最有利的合同安排。另一种是内生的不对称信息，该种情况通常在双方订立合同后产生，主要是由于一方当事人不能对另一方当事人进行适当监督产生的。一方当事人无法预测对方当事人的某种行为，该行为便是隐藏行动，这种行动的存在会引发道德风险。而此种风险概念最早可见于海上保险，其在经济学领域的运用则最早可见于 20 世纪 60 年代的西方社会。现如今，在社会中，道德风险问题这一问题大量存在于医疗、政治等各大领域，并且受到了人们的高度关注，但它无法被彻底消除。例如，在订立劳动合同之后，在工作中，雇员态度可以是积极的，也可以是消极的，在购入车险之后，车主驾驶时会更小心还是更放心大胆等等。这种情况的存在就引发了经典的激励问题:究竟要实施怎样的激励机制方能使对方的行动与自己的预期一致，比如公司要制定什么样的职工考核制度才能使员工认真工作。

因此，笔者认为，所谓信息不对称指的是进行市场交易的当事人所掌握与交易有关信息存在差异，掌握信息较少的一方在交易中不具优势，而另一方则在交易中占据优势地位。

Arrow（1963）所做研究表明，在保险经济学领域，信息不对称这个问题非常关键，并且这也是使保险制度稳健实施受影响的一个重要问题。在保险领域，信息不

对称现象的存在非常广泛，并且引起了非常严重的问题。在各类保险市场中，各个不同市场主体之间比较都存在着一定的信息不对称情况，例如保险人和保险代理人、投保方和保险人之间、、保险监管机构和保险人之间、保险人和保险经纪人之间都存在各种形式的信息不对成。从保险人的角度出发，他无法获取和自己进行交涉的保险经纪人的职业道德素养和专业技术能力的相关信息，与此同时，保险经纪人对于和自己签订保险合同的保险人的经营水平、资金周转能力的相关信息也无法有效获得，由于信息不对称始终存在于双方，因此保险人始终无法获得最优质的保险服务。虽然保险监管机构把保护保险消费者权益为最主要的目标，并且对保险人的运营行为进行现场监管和非现场监管从而加以监督规范，但由于保险人存在提交虚假资料的可能性，对于保险人服务质量是否提升的问题保险监管机构也无法完全确定。

在保险市场上，信息不对称问题主要以以下三种方式存在：

首先，保险人以及投保方(投保人、受益人、被保险人)之间的信息并不对称，并且保险人在交易过程中由于信息掌握较多而处于优势地位。在此类情形下，由于投保人、保险人双方在地位、信息沟通等各方面存在差异，因此在进行保险交易时，他们之间是存在信息不透明问题的，这重点表现在保险类型、保险流程、主客体性质等多个方面。由于在保险领域，保险人扮演着卖方角色，他必然掌握着与其所售产品有关的大量信息，基于此，与投保人，即买方来所掌握的信息来说，毫无疑问其所掌握的信息更占优势，双方信息不透明的问题就此产生。详细说来，在信息方面，保险人具有的优势重点表现在:其一，保险人对自身保险产品信息的了解比投保人更深，而投保人则难以达到全面深入了解保险人机构以及其产品的情况，基于此可知，在这方面，投保人所掌握的信息是不具优势的;其二，因保险产品的合同条款是由保险人来制定的，因此对于投保人来说，它更具优势，再加上保险条款内容中存在大量高度专业的措辞，这又将导致投保人难以全面准确获得关于保险条款的具体信息；其次，保险人由于自身的专业性和技术性，在进行保险类型的推销以及进入理赔流程的相关信息具有得天独厚的优势。假如保险人为了自身的经济利益诉求，从而选择欺瞒投保方，导致投保方的利益受到侵害，就产生了欺诈行为。

另一种信息不对称则存在于保险人与其代理人之间，并且在此种情况下，信息劣势方则变成了保险人。保险人与其代理人之间的关系是授权代理性质，保险人难以全面获悉其代理人在运营当中有否全面履责，会否出现协同投保人骗保的现象，或者在理赔时有否存在以人情关系为依据实行理赔等不利于保险人的各种情况，在

这种情况下保险人的信息相对来说不够全面。

第三种是保险人和投保方(受益人、投保人、被保险人)并不处于平等的信息获取地位，此时投保方的优势地位更加凸显。保险人在不同的合同订立过程中都处于优势地位是不可能的，因为投保人能够拥有和控制既定的保险标的，相较于保险人来说，其对保险标的的情况更为了解,，并且对方很难真正掌握其选择投保的真正动机。比如投保方可能会通过故意制造事故或将事故损失夸大等各种操作来骗保，在这种情况下，保险人变成了信息劣势一方。因保险与大数法定律相符，所以投保人的骗保行为会导致全体并不骗保的保险人、投保人权益受损，因此保险人必须实施各种反欺诈措辞来维护公平的市场环境。

在现实当中保险欺诈主要是指投保方利用信息优势骗取保险金，即第三种信息不对称情况。若可以借助种种措施来使此种信息不透明问题得到缓解，则必将能够有效遏制保险欺诈现象的发生。

3.3 人工智能对保险反欺诈的作用机制

根据 2006 年《预防、发现和纠正保险欺诈指引》由国际保险监督官协会特别巩固的，从保险公司角度来划分保险欺诈行为，则分为内部欺诈、中介欺诈以及保单持有人欺诈。

内部欺诈指的是保险公司的内部人员与外部人员进行勾结来对保险公司进行欺诈，其中包括中高层管理人员与外部人员或基层从业者与外部人员。内部欺诈的行为主要表现形式主要为侵占保险公司的财产，例如现金、客户信息、或设备等，此类表现形式属于内部欺诈中最传统的欺诈行为，实际上也在其他公司中普遍存在。而保险公司中独有的内部欺诈主要是通过赔付获取额外的赔偿，这种额外的赔偿非常有可能是超额的赔偿或者保险公司本不应该赔偿的部分。

中介欺诈是指中介作为投保人和保险公司之间搭建的桥梁，其在保险交易过程中的定位使得他们有机会接触到保险公司的一些重要交易的具体信息和一些关键节点，因此中介欺诈的问题值得保险公司进行一些关注。

保单持有人欺诈是指投保人以个人的方式或者多人串通的方式，在购买保险产品时实施针对保险公司的不正当保险行为或在执行过程中实施针对保险公司的赔款欺诈行为。

本文根据以上理论内容概括了人工智能对保险反欺诈的作用机制。

影响机制一：人工智能能够使保险欺诈行为暴露机率加大，进而降低持有保单者以及中介发生欺诈的机率，这也是人工智能在保险反欺诈业务中的主要应用方面。

影响机制二：人工智能可以通过对海量数据的训练与模拟，模型的结果不以个人意志为转移，完善保险反欺诈流程的客观性和独立性，削弱了个人在职务行为中产生的信息优势，从而降低了内部欺诈的可能性。

第 4 章　模型理论

4.1　GBDT 模型及其改进算法 XGBoost 模型

4.1.1 GBDT 模型原理

GBDT 是基于梯度提升理念来构建单棵决策树的一种模型，它利用 Boosting 算法来组合弱分类器（即 CART 回归树模型）。此外，它还利用了迭代理念，在第一次进行迭代时都会在上一棵决策树减少残差趋向生成新的一棵决策树，为的是持续减少模型的残差值。模型最后输出的值是将所有决策树结果相加后得出的。也就是说，假定该模型在上一轮迭代过程中，强学习器的累加所得为 $f_{t-1}(x)$，而损失函数是 $L(y_i, f(x_i))$，则搜索到一个弱分类器 $f(x)$ 就是进行这一轮迭代需要实现的目的，即致使此次获得的损失函数 $L(y_i, f(x_i)+f_{t-1}(x_i))$ 实现最小化[34]。

4.1.2 GBDT 的负梯度拟合

对于回归问题的求解，GBDT 所用的损失函数通常是平方误差函数，这种函数的残差即为预估值与实际值相减的结果。为了在分类问题求解中运用 GBDT，Freidman 提出，残差近似值可选为损失函数的负梯度，以此来近似本轮的损失，进而完成一棵 CART 回归数的构建。假定在进行第 t 轮训练时，对于其第 I 个样本损失函数的负梯度用以下公示表示:

$$r_{ti} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{t-1}(x)} \tag{4.1}$$

运用样本残差和样本，$(x_i, r_{ti})(i=1,2,...,m)$ 能够完成一棵 CART 回归树的构建，假定这棵回归树的全部叶片结点是 $R_{tj}, j=1,2,...,J$ 当中的 J 代表叶节点数。则关于每个叶片结点的样本，进行训练的目的是尽可能的减小其输出的损失函数，假定这时该叶片节点输出的最佳数值是 c_{tj}:

$$c_{tj} = \arg\min \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i)+c) \tag{4.2}$$

那么本轮的决策树拟合函数为:

$$h_t(x) = \sum_{j=1}^{J} c_{tj} I(x \in R_{tj}) \tag{4.3}$$

那么本轮得到的强学习器为:

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I(x \in R_{tj}) \tag{4.4}$$

4.1.3 二元 GBDT 分类算法

若在分类问题中运用 GBDT 模型，则因输出值为离散值而导致模型不可能直接运用第一轮的输出值来对输出误差进行拟合。为使此问题得到解决，进而在分类问题的解决中更好的运用 GBDT 模型，可运用分类问题求解过程中对 Logistic 回归思想的运用，例如关于二分类问题，可构建各种梯度提升树来完成关于对数几率 $\ln\frac{p}{1-p}$ 的拟合，进而获得相应的各种 CART 回归树。这时，可用以下公式来表示其分类模型:

$$P(y=1 \mid x) = \frac{1}{1+e^{-\sum_{m=1}^{M} h_m(x)}} \tag{4.5}$$

其中 $h_m(x)$ 就是学习到的决策树。损失函数为：

$$L(y, f(x)) = \log(1+\exp(-yf(x))) \tag{4.6}$$

其中 $y \in \{-1,1\}$。此时负梯度误差为：

$$r_{ti} = \frac{y_i}{1+\exp(y_i f(x_i))} \tag{4.7}$$

对于生成的 CART 回归树，各个叶节点的最佳负梯度拟合的近似值为:

$$c_{ti} = \frac{\sum_{x_i \in R_{tj}} r_{ti}}{\sum_{x_i \in R_{tj}} |r_{ti}|(1-|r_{ti}|)} \tag{4.8}$$

4.1.4 XGBoost 模型原理

能够把弱学习器进行转型升级为强学习器的算法为 Boosting 一族。虽然每种弱分类器的总体预测正确率在处理分类数据时有较大的可能性并不理想，但其有几率会在数据的特定方向上有着不错的正确预测率。若想产生一个总体预测正确率高、效率强大的强分类器[35]，则可通过某种方式将多个局部较高预测正确率的弱分类器进行整理。

XGBoost 全名为（extreme gradient boosting），意译为极端梯度增强算法，是陈

天奇等提出的基于集成思想的机器学习算法，这种学习算法是有监督的，它包含了多棵决策树，能够使分类、回归等问题得到解决。与传统集成学习有区别，XGBoost 模型是以降低模型误差来使其性能得到提高的。作为一种学习办法，XGBoost 模型在很多工程应用方面取得了良好的效果，例如网络入侵检测、社区网络状态监控等。其整体构思便是在当前模型中添加不同的另一个模型，从而使得组合模型具有比当前模型更好的效果。

假设有数据集$T=\{(x_i,y_i)\}$,其中$x_i\in R^m, y_i\in R, i=1,2,...,n$。象征特征向量的维度为 m 的$x_i$,，样本的标志为$y_i$。假如 K 棵树组成了 XGBoost 模型，则可将模型定义为:

$$\hat{y}_l = F_K(x_i) = F_{K-1}(x_i) + f_K(x_i) \tag{4.9}$$

如果第 K 棵决策树为$f_K(x_i)$。那么样本特征都由各个决策树来反映，进一步让这棵决策树的那些叶子节点能够承载所有的样本。一个权重分数ω会分散到各个叶子节点上，某叶子节点上所持有的样本数值在这棵决策树上的表征预测值就是权重分数。对每棵树进行叶子节点的累加，即累加全部决策树叶子节点的权重分数ω，从而获取样本预测值在 XGBoost 模型上的数值。

目标函数（以 XGBoost 为模型）为:

$$Obj = \sum_{i=1}^{n} L(y_i + \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4.10}$$

用来比较样本显示的真实值和由模型预测出的数值的相差值为目标函数的首项，同时将正则化项引入 Obj 中，进而规避错综复杂类模型的再现，以便获得更好的拟合成效。正则化项的表达式为:

$$\Omega(f) = \lambda T + \frac{1}{2}\lambda \| \omega \|^2 \tag{4.11}$$

树的复杂度由系数λ和树的叶子节点数 T 组成，用来确定叶子节点的权重分数为 L2 的正则项$\frac{1}{2}\lambda \| \omega \|^2$，使得各个叶子结点最大概率的拥有相同的重要程度。用旧时的办法很难在欧氏空间中对目标函数实现改良，XGBoost 通过差不多相同的办法使难题得到解决。最终的函数可写出：

$$Obj^{(s)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(s-1)} + f_s(x_i)) + \Omega(f_s) \tag{4.12}$$

样本x_i的第 s-1 轮模型预测值为$\hat{y}_i^{(s-1)}$，新子模型在第 s 轮中是$f_s(x_i)$。

通过泰勒公式对 XGBoost 模型的目标函数来拟合，使函数简洁化。函数中某点数值信息可以用泰勒公式来表达该临近点的函数值数据，如果简洁后的函数拥有圆滑的曲线，则就能够来构建一个由来自该函数点的各阶导数值组成的多项式，拟合该数据点邻域的函数值[52]。把最后目标表达函数中的$\hat{y}_i^{(s-1)}+f_s(x_i)$看作 x，$\hat{y}_i^{(s-1)}$看作$x_0$，使用泰勒公式将其展开可得到:

$$Obj^{(s)}=\sum_{i=1}^{n}L(y_i,\hat{y}_i^{(s-1)})+g_i f_s(x_i)+\frac{1}{2}h_i f_s^2(x_i)+\Omega(f_s) \tag{4.13}$$

可以得出，该损失函数所具有的一阶梯度的统计表达式是$g_i=\frac{\partial L(y_i,\hat{y}_i^{(s-1)})}{\partial \hat{y}^{(s-1)}}$，其二阶梯度统计表达式为$h_i=\frac{\partial^2 L(y_i,\hat{y}_i^{(s-1)})}{\partial \hat{y}^{(s-1)}}$·常数项则无任何表达意义，故将常数项$L(y_i,\hat{y}_i^{(s-1)})$删除，再进行目标函数的简化，并在目标函数中将$\Omega(f_s)$的表达式代入，可得:

$$Obj^{(s)}\cong\sum_{i=1}^{n}[g_i f_s(x_i)+\frac{1}{2}h_i f_s^2(x_i)]+\lambda T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{4.14}$$

$f_s(x_i)$代表的是第 s 棵树，即属于 XGBoost 模型之下的子模型，ω_j代表是其叶片节点权重，经整合可得:

$$Obj^{(s)}=\sum_{i=1}^{n}[g_i\omega_{q(x_i)}+\frac{1}{2}h_i\omega_{q(x_i)}^2]+\lambda T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{4.15}$$

第 s 棵树得到的 1 到 T 中的某个叶子节点的单个权重值就是$\omega_{q(x_i)}$。

放弃样本遍历而用叶子节点遍历来计算损失函数，此时有:

$$Obj^{(s)}=\sum_{i=1}^{n}[(\sum_{j\in I_j}g_i)\omega_j+(\sum_{j\in I_j}h_i)\omega_j^2]+\lambda T \tag{4.16}$$

来自树的第 j 个决策树的叶子节点的样本数据合集就是I_j，叶子节点获得样本数据则是通过第 s 个树模型$f_s(x_i)$，理论分析得该叶子节点中样本的权重分数ω。因此，当$i\in I_j$时，$f_s(x_i)$可被ω_j取代。可将该目标函数视作是自变量为ω_j的函数，则就固定树架构来说，j 叶片节点取值的最佳结果是：

$$\omega_j^* = -\frac{\sum_{j\in I_j} g_i}{\sum_{j\in I_j} h_i + \lambda} \tag{4.17}$$

这时，目标函数的最佳结果是：

$$Obj^{(s)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{j\in I_j} g_i)^2}{\sum_{j\in I_j} h_i + \lambda} + \lambda T \tag{4.18}$$

若将 $\sum_{j\in I_j} g_i$ 记为 G_j，将 $\sum_{j\in I_j} h_i$ 记为 H_j，可得:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{4.19}$$

$$Obj^{(s)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j + \lambda} + \lambda T \tag{4.20}$$

现在，可得一个评价函数 $Obj^{(s)}$ 来评价决策树，评分和样本呈现负相关性。

理论分析上，通过筛选在训练每颗树的所获得的评分，就能得到最优树模型，也就是得分最差的模型。但事实上，由于备选树数量的无穷性，所有的评分使不可得的。对此情形，贪心算法就是来处理这个难题，在节点分裂时采取当前最有的分裂策略而非全局最优。

节点分裂就是 XGBoost 模型在创建不同的树的模型过程中一定会有的步骤。如果 j 为目前分裂节点，则可得到该节点在目标函数中的贡献为:

$$Obj_j = -\frac{1}{2}\frac{G_j}{H_j + \lambda} + \gamma \tag{4.21}$$

现在，该节点对树模型复杂度的贡献只剩下 γ 这一个节点，基于此，在 j 点分裂之后，左右子节点对目标函数的奉献是:

$$Obj_s = -\frac{1}{2}(\frac{G_{jL}^2}{H_{jL} + \lambda} + \frac{G_{jR}^2}{H_{jR} + \lambda}) + 2\gamma \tag{4.22}$$

来自节点分裂后的目标函数的改变最终就可以被得到，其表达式为：

$$Obj_{split} = \frac{1}{2}(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}) - \gamma \tag{4.23}$$

当中，G 代表的是现节点的一阶梯度统计和，而 H 则代表的是二阶梯度统计和；G_R、H_R 分别代表的是现节点分裂所得右侧子节点样本集的一阶梯度统计和以及其

二阶梯度统计和，而G_L、H_L则代表的是现节点分裂所得左侧子节点样本集的一阶梯度统计和以及其二阶梯度统计和。$\frac{G_L^2}{H_L+\lambda}$表示左子树得分，$\frac{G_R^2}{H_R+\lambda}$表示右子树得分，$\frac{G^2}{H+\lambda}$表示未分割时的得分，若$Obj_{split}$值大于 0 则可以分裂，反之则不行。

节点分裂之前与其分裂之后所产生的的两个树模型仅在结构上有着两个新的节点的区别。所以，其函数的差就是Obj_{split}。使Obj_{split}最为显著的方案既目前在每个节点的分裂中的最优解，在此点进行切割。从最初的那个节点开始对一个个的节点实施最优的分裂，就可以得到一个完整的树模型。

经过对 XGBoost 模型原理的分析，得出了 XGBoost 模型的具有以下几个优势：

（1）可同时运用一阶与二阶导数，并且模型还可对损失函数做自定义，先决条件是需要先求导损失函数的二阶与一阶，并得到相应的结果；

（2）正则项进入了模型并控制了其复杂程度使过度拟合的情况难以发生；

（3）为防止过度拟合情况的出现，采用随机森林的方式；

（4）采用了一种模拟近似法来搜索最佳切割点，极高地提升了效率，并在同一时刻处理了稀疏数据集，样本缺失值等数据；

（5）允许并行；

（6）通过近似直方图算法以更高的效率产生备用分割数据点；

（7）实施了多重改善为提高模型的运行算法效率以及实现算法效率，在运行内容不足的情况下，模型用到了多线程协作、分块等理念。

4.2 深度学习模型

本文将特征工程处理后的数据与多种深度学习模型进行训练测试，得到在不同提调率下与的精准率指标，并于前文 XGBoost 模型的结果进行横向比较，以凸显本文主模型的横向比较优势。具体测试模型如表 4-1 所示：

表 4-1 测试模型列表

模型类别	模型名称	名称缩写
机器学习模型	eXtreme Gradient Boosting	XGBoost
深度学习模型	Attentional Factorization Machine	AFM
深度学习模型	Deep & Cross Network	DCN
深度学习模型	deep Factorization Machines	deepFM
深度学习模型	Neural Factorization Machine	NFM
深度学习模型	Probabilistic Neural Network	PNN
深度学习模型	Wide & Deep Learning	WDL

由于本文使用深度学习模型的目的仅仅是为了与主模型进行横向比较，凸显主模型的优势，故对深度学习模型的理论不做过于详细的介绍，仅做简单概述。

本文使用的六种深度学习皆以 WDL 模型为基础。具体关系如图 4-1 所示：

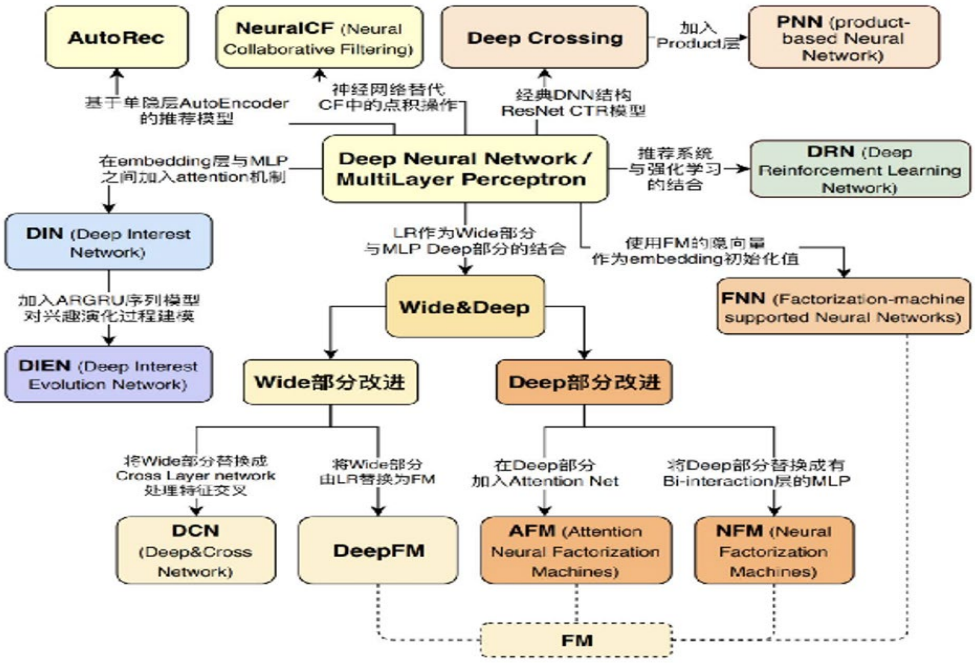


图 4-1 六模型关系图

4.2.1 WDL 模型原理

WDL 模型，中文名为宽度&深度模型，是深度学习模型的一种，其结构为深浅层式学习。该模型经对深度与线性两大模型的结合，使其具有了泛化、记忆两方面的优势，同时运用联合训练法加以改进[36]。WDL 模型于 2016 年由谷歌团队所提出，其用途主要增强谷歌应用市场的软件推荐功能，进而使软件推荐算法更具准确性，与此同时还能够兼顾到推荐体系整体的可扩展性。

宽度模型通常指的是各种传统的机器学习模型，由于这类模型的输入特征通常是较高维的数据，从而叫宽度模型。分散型、连续型特征以及离散型特征融合产生的交叉特征或是经过代码编码后的数据向量都能够成为宽度模型的输入特征。可以获取最初输入特征间的交互信息的叉特征还可以使模型更多的具有一些非线性的数据结构，同时还可以扩大模型的数据输入特征的整体维度数量。

深度模型是指深度学习模型，因模型是以多层线性和非线性结构叠加而成而得名，深度学习模型可以从原始数据中直接对关键特征进行捕捉，并将其用于各类机器学习的任务当中。基于此，可知在包含非结构性初始信息特征的学习中运用该模型是非常合适的。深度学习能够从非结构性信息中将关键特征提取出来并对其进行学习，然后再经由非线性、多层线性转化后找到极少出现或者未曾在历史信息中出现过的新特征组，鉴于此可知，深度模型的泛化性极其强大，它能够迅速完成学习又获得有效的、全新的特征组，进而使模型整体的预测效果得到提高。

WDL 模型，一个广泛应用的泛化线性模型的宽度部分：

$$Y_{wide} = W_{wide}^{T} X_{wide} + b \tag{4.24}$$

我们将 Y_{wide} 设定为预测值，把 d 维特征向量设为 $X_{wide} = [x_1, x_2, ..., x_d]$，模型参数 $W_{wide} = [w_1, w_2, ..., w_d]$ 为模型的参数，b 为偏置函数。

这意味着深度部分即为深度学习框架，它包含的主要是隐匿层、输入以及输出层。对于未知特征和稀疏特征组，这部分的模型能够进行低维嵌入式处理，进而确保整个模型具有记忆功能、泛化功能。可用以下公示来表示第一个隐匿层：

$$a^{(l+1)} = f(W^{(l)} a^{(l)} + b^{(l)}) \tag{4.25}$$

在这当中，l 表示层数，$f(W^{(l)} a^{(l)} + b^{(l)})$ 为模型激活函数，$a^{(l)}$ 为第 l 层的输出结果，$b^{(l)}$、$W^{(l)}$ 分别表示第 l 层的偏置和权重占比。模型整体的表示如下：

$$Y = W_{wide}^{T} X + W_{deep}^{T} a^{(lf)} + b \tag{4.26}$$

其中，W_{wide}^{T}、X 分别是宽度部分的权重向量和输入特征向量，$a^{(lf)}$、W_{deep}^{T} 代表的是深度部分最后一层激活函数的输出向量以及其权重向量。

模型训练过程中对深度与宽度部分同时进行优化的参数。

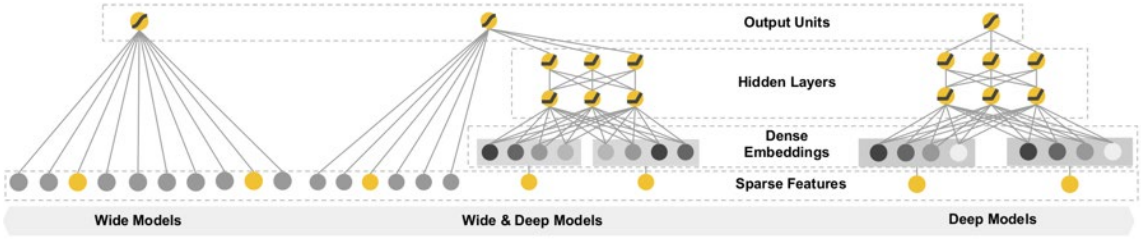


图 4-2 WDL 模型网络架构示意图

4.2.2 其他模型原理

PNN 模型是一种基于点积的神经网络模型，是深度学习模型的一种。与 MLP 这一传统模型相比，该模型借助嵌入层来学习分类信息的分布式表示，接着以增加点积层（一层）又用于获取来自域间种类的交互形式，还可以深层研究高阶特征交互[37]在全连接层中的使用。

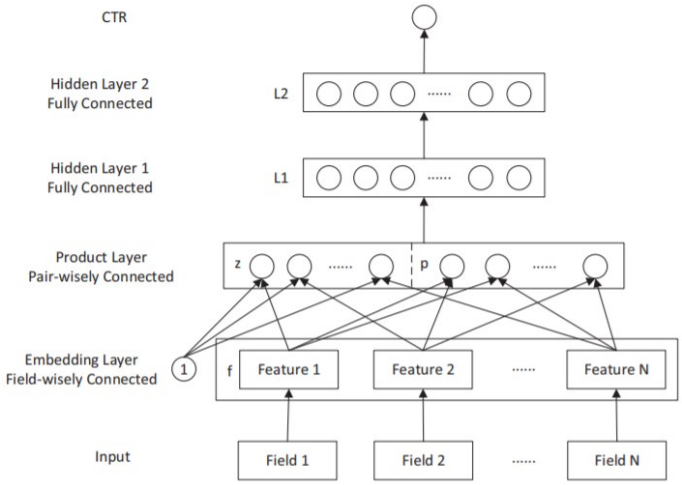


图 4-3 PNN 模型网络架构示意图

Deep FM 模型同样是深度学习模型的一种,它也是因子分解机。它同时具有深度学习以及分解机制两种能力，前者用于特征学习，后者则用于推荐推荐，它使 Wide 部分(在 WDL 模型中)得到了优化，利用 LR 取代了因子分解机（即 FM）从而达到了自行构建二阶特征的目的[38]。

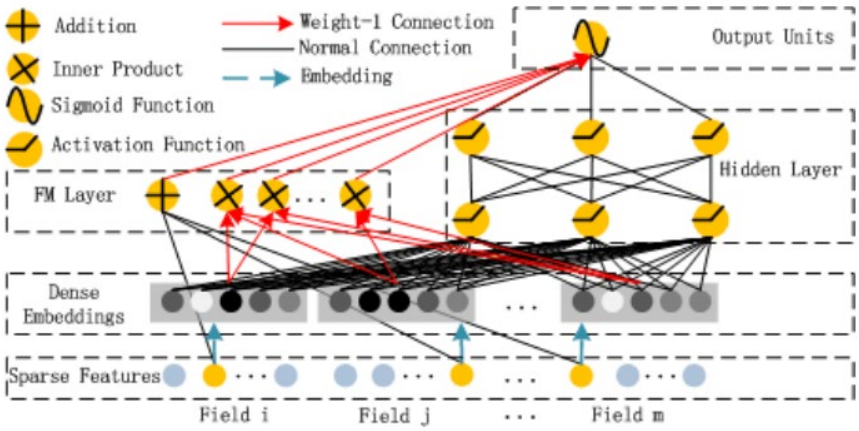


图 4-4 DeepFM 模型网络架构示意图

DCN 模型是对 WDL 模型中 Wide 加以优化后的结果，该模型可以将每层应用特征重叠显示出来，它可以自动构建有限高阶的重叠特征并学习相应的权重[39]。

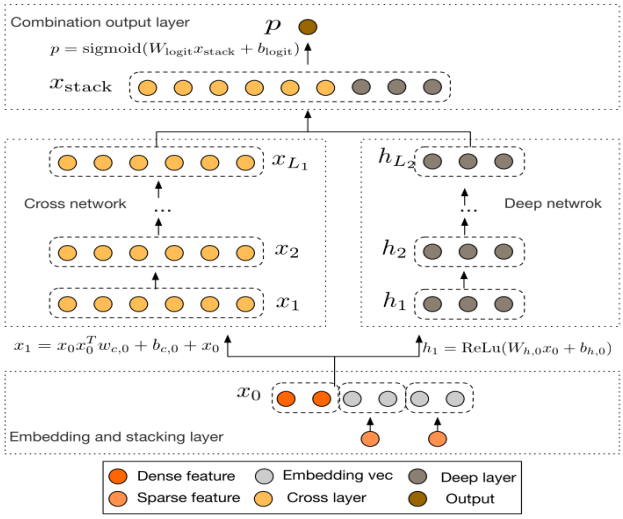


图 4-5 DCN 模型网络架构示意图

NFM 模型是对 WDL 模型中 Deep 加以优化，把 FM 的二阶重叠项输入至 Deep 模型中，以隐匿层的添加来提高其性能，它是神经因子分解机[40]。

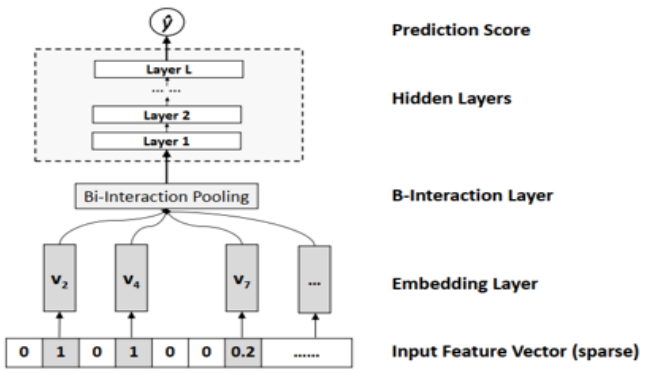


图 4-6 NFM 模型网络架构示意图

AFM 模型为注意力因子分解机。其对 WDL 模型中的 Deep 做了改进，将专注力机制加入了进来从而达到了区分各种重叠特征重要作用的目的[41]。

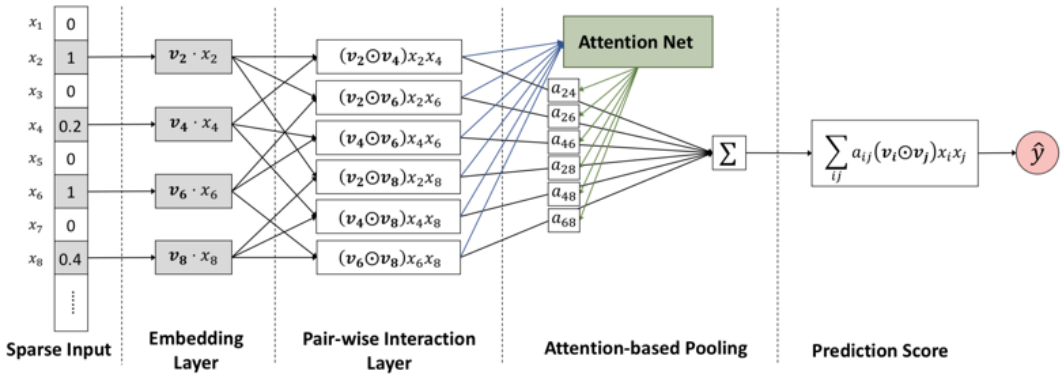


图 4-7 AFM 模型网络架构示意图

第 5 章　实证分析

5.1　分类模型评价指标

5.1.1 混淆矩阵

以达到使得正确率或错误率能够甄别不同种类的样本的错误分类程度在数据分类中的目标，我们使用了以混淆矩阵为工具来创建的方法。用该矩阵行中的数值来表示分类数据模型的估计类别，列中的数值表示样本数据的真正类型。混淆矩阵中将各列数据相加的结果即为模型对全部样本中进行预测后所得结果中相应类别的真实样本数量;将混淆矩阵中将各行数据相加的结果即为模型对全部样本中进行预测后所得结果中相应类别的真实样本数量。

对于二元分类问题，通常将两类研究对象分别记成正类和负类。文中注意的是保险数据是否为欺诈数据，因此欺诈数据被记录为正，非欺诈数据被记录为负。其混淆矩阵详见表 5-1 所示。

表 5-1 二分类问题的混淆矩阵

真实 类别	预测结果	
	1	0
1	TP	FN
0	FP	TN

对于这一问题，样本的标记值与模型预测的结果之间存在下述几种情况:

TP:样本实际类型是正，模型预做正确，样本经预测后结果是正，即类型真实。

FN:样本的真实类别为正，模型预测错误。预测样本为负，即假负类别。

FP:样本的真实类别为负，模型预测错误，样本预测为正，即假正类别。

TN:样本的真实类别为负，模型预测是正确的，样本预测为负，即真负类别。

在前文已概述混淆矩阵相关的基础上，现将评估其他 4 种分类模型的指标导出。

精准率指的是在模型预测过程中，在所有样本中，预测正确类样本的占比，其公式为:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

精确率指的是在模型预测过程中，在所有预测是正类的样本中，预测结果正确

类样本的占比，其公式为：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

召回率是指模型预测正确的样本在所有阳性样本中的比例，表达式见下方：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

F1 得分为模型召回率和模型正确率的调和平均值，满足以下条件：

$$\frac{2}{F_1} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \tag{5.4}$$

$$F_1 = \frac{2TP}{2TP+FP+TN} \tag{5.5}$$

5.1.2 ROC 曲线和 AUC

ROC 曲线，也就是受试者工作特征曲线。曲线的纵轴是真实的等级率，即前文提及的召回率。其横轴代表的是特异性（即假阳性率），是被模型预测为正的样本类然而样本实际为负类样本数在全部负类样本当中的数量占比。二分类问题中，模型为预测每一个样本的类型会界定一个阈值。如果模型输出的某一个样本结果比阈值要大，则该样本为正，反之则为负。若将阈值减少，自然能够预测出数量更多的正类样本，模型也将具有更高的真实类率，然而也会有更多实际为负的样本被预测为正，进而导致模型假正类比率就会得到提升。以此为基础，就需要引入 ROC 曲线来描述模型分类性能随阈值变化的过程，进一步对模型进行全局评价。

在大部分的实际问题中，样本分布不均匀和误分类代价不一致有可能会对评价结果产生影响，ROC 曲线可以很好地解决这一问题。这是由于当绘制 ROC 曲线的时候,所有样本的输出值排列从小型到大型,他们作为阈值,然后真正的类率和假阳性类的分类模型在每个阈值计算,这样每个阈值将对应于一个点的坐标系统。

将以上点连接得到的曲线为 ROC 曲线。特别地，当分类模型的阈值为 0 时，模型预测所有样本为正类，模型的真实类率和误报类率为 1。如果分类模型的阈值为 1，则模型预测所有样本为负类，此时模型的真正类率与假正类率皆为 0。位于左上方

的点，所对应模型的假正类率小，而真正类率大。由此可知，ROC 曲线位于左上方时，模型对数据的分类性能较为良好，且 ROC 曲线的图示越向左上方凸则说明模型分类的能力越好。

AUC 是一个特定值，该值与坐标轴与 ROC 所构成图形的面积相同。理论上来说，其上、下限值分别为 0、1 若 ROC 曲线即为直线 y = x，则各种阈值之下，模型的真实类率等同于虚假类率。这意味着，样本预测为正和为负的机率是相同的，即模型的类别区分功能等同于以抛硬币方式来区分的结果。基于此，于分类模型而言，其 ROC 曲线通常比 Y=X 这一直线要高，也就是具有分类性能的模型，它的 AUC 值应大于 0.5。另外，AUC 值越趋向于 1，则意味着该模型识别各类样本的功能亦越强。

5.2 数据来源及处理

由于数据的初始形式直接应用于 Xgboost 模型和深度学习模型效果不佳，需要对数据先进行一系列的预处理，故本文先使用 python 语言设计了一个特征工程模型，实现了将 58 个原数据表合成为一个总表，再根据不同数据特征进行划分并对一些数据特征进行编码处理，然后投入 Xgboost 模型和深度学习模型进行训练，得出最后的结果。本文的实证分析在 Jupyter Notebook（anaconda3）上利用 Python 编程语言进行。

特征工程分为特征分类和特征衍生，特征分类指基于特征值的分布情况将原始特征进行分类，特征衍生指基于分类后的特征进行特征合成，获得更加丰富的特征组合。本文设计特征工程模型的主要作用有两个：一是提高数据与后续模型的适配程度，二是增加数据特征的数量，提升数据维度，以满足后续模型的训练需要。这两个方面最终都能影响并提升模型的最终结果。

5.2.1 数据来源

本文所采用的数据来源于国内某财险公司车险部门截至 2020 年 4 月的经过脱敏处理的车险理赔数据。数据共有 29 个不同数据表种类，例如主表、投保车信息表、车辆维修费用表等，再根据是否欺诈分成共计 58 个 excel 表格。共计约 417000 条车险理赔数据。

5.2.2 数据清洗

数据清洗，或者成为数据清理，用于检测和纠正（或删除）记录集，表或数据库中的不准确或损坏的记录。广义上讲，数据清除或清理它指的是识别到一些不正确完整、不相关不准确或有其他问题的(“脏”)数据部件，然后去替换、修改或着删除这些脏数据。

本文的数据由于只经过了脱敏处理，极大的保留了原数据的真实性，这也意味着数据的格式化程度很低（很“脏”），具有非常多的数据内容或者格式错误，并且原数据中包含欺诈标签（"is_fraud"），需要预先去除，不然会很大程度上影响模型的最终预测结果。所以进行数据清理处理掉无用途的部分（不完整，不影响结果的数据），这是一个有价值的过程，可以帮助后续环节，节省时间并提高效率。

本文数据清理过程一共包含三个部分：

第一，删除名称为"is_fraud"的数据特征列，此列数据为原始数据的欺诈标签，即这列数据已经能够说明案件是否为欺诈，此数据列为保险公司事后添加的数据特征且该数据列具有对模型结果的强相关性，在后续实证分析过程中无法使用此列数据特征作为信息来源，因此需要对此列数据特征进行隔离。

第二，对数据进行去重，即在完全相同的两条或两条以上的相同数据列中只保留其中一列，从而降低了数据维度提高模型运算速度，并且保证每条不同数据的重要性相当。

第三，对剩余每列特征进行数据内容的格式检测，错误值与空值处理，由于原始数据可能存在投保人或保险单录入工作人员的误填和漏填，导致出现数字特征列中存在某一个数据为文本格式的数据或者直接为空值的情况，由于本文无法获取额外的信息与资料来源来补全，无法还原真实情况，因此最好的办法是对错误数据进行删除处理，所以本文统一对错误值进行整行删除处理，对空值进行填 0 处理，以期在维持数据量的同时尽可能减小处理后数据对模型结果真实性的影响。

第一步于整个实证过程最开始进行，而后两步清洗过程贯穿于特征工程模型全程，并且多次进行确保有效性。

5.2.3 多表横向合并

原数据共有 58 个表，无法同时输入模型进行训练且每个表中不同的数据特征可能存在重复或者对应关系，需要归类并处理，因为事故数据总量是固定的，所以需

要将所有表合并成一张总表，然后将总表根据特有方式进行拆分，以满足后续模型训练需要。

首先进行多表合并成总表的过程，利用 Python 编程语言中的“concat”函数对欺诈数据和非欺诈数据的相同类型表格采取纵向合并，得到 29 张不同类型的表格。而对于 29 张不同类型表格，采取添加数据特征但不添加数据量的方式进行多表横向合并，本文优先考虑将其余表格合并到主表中，横向合并需要找到多个表格相同的数据特征，以数据特征为锚，一一对应来进行多表的合并。

首先提出奇异值的概念，奇异值为某一列数据特征中包含的不同数据的个数。以主表为例，主表中奇异值最大的数据特征为“REGISTNO”，可以作为锚定数据特征。同理，对 29 个表分别进行此操作，得到结果如表 5-2 所示。

表 5-2 锚定数据特征表

表名	锚定数据特征
主表，索赔表等 12 个表	“REGISTNO”，“ACCIDENTNO”
车辆损失表、车辆维修费用表等 11 个表	“ACCIDENTNO”
保单信息表等 6 个表	“REGISTNO”

基于表 5-2，本文设计多表横向合并的并表过程为先进行三类同一级别的表格的横向合并，再进行不同等级的子母表的合并。

在并表过程中，由于原始数据中存在前表一行对后表多行的情况，例如一次事故中车辆有多个部件受损，在主表中仅有一条数据一个事故号，但在车辆损失表中有多个相同事故号对应不同的部件损失情况，所以在并表过程中需要进行数据聚合。

数据聚合是指合并来自不同数据源的数据，通常指的是转换数据，是每一个数组生成一个单一的数值。常见的数据聚合操作，例如 sum 函数、mean 函数和 count 函数，这些函数均是操作一组数据，得到的结果只有一个数值。本文对于不同的数据特征特性采用不同的聚合方式，具体问题具体分析：涉及可以相加的数据特征例如金额等采用求和，对于无法相加的数据特征例如性别、年龄等采用求简单算数平均值（数据表中性别的表示方法为 0 代表男性 1 代表女性），对于姓名、事故地点等文本数据直接舍弃整列。最终利用 Python 编程语言中“eda_dataframe”函数得到得到总表的分布式数据集如图 5-1 所示。

```
In [24]: df_features = eda_dataframe(df_target)# 展示特征的奇异值个数，奇异值的内容，奇异值使用量的占比；
         df_features.sort_values('nunique', ascending=False) # 依照某个字段中的数据进行排序，这里的字段是 “nunique”；
```

Out[24]:

	feature	dtype	nunique	vunique	cunique	dunique
0	REGISTNO	object	416904	[605012018110000000051, 605012018110000000053,...	{'605112018140000055479': 1, '6050720191500000...	{'605112018140000055479': 0.0, '60507201915000...
4	CLAIMNO	object	403504	[505012018110108000023, 505012018110111000002,...	{nan: 13400, '505112018129700003898': 1, '5050...	{nan: 0.03, '505112018129700003898': 0.0, '505...
44	CLAIMCONFIRMCODE	object	398175	[85CLPC021800000000012708390721, 85CLPC0218000...	{nan: 18729, '55GPIC330019001551967289331393':...	{nan: 0.04, '55GPIC330019001551967289331393':...
45	MODIFYDATE	object	396560	[2018/5/11 17:30:55, 2018/5/24 17:00:13, 2018/...	{'2018/10/31 19:19:08': 5, '2018/8/3 21:18:27'...	{'2018/10/31 19:19:08': 0.0, '2018/8/3 21:18:2...
3	POLICYNO	object	395298	[805012016110108020645, 805012016110111002497,...	{'805072018220105001718': 8, '8051120182201050...	{'805072018220105001718': 0.0, '80511201822010...
...	...	...	...	...	...	...
31	FLAG	float64	0	[nan]	{nan: 416904}	{nan: 1.0}
15	SUBBUSINESSNATURENAME	float64	0	[nan]	{nan: 416904}	{nan: 1.0}
42	CUSTOMERCHECK	float64	0	[nan]	{nan: 416904}	{nan: 1.0}
62	THREEFLAG	float64	0	[nan]	{nan: 416904}	{nan: 1.0}
14	SUBBUSINESSNATURE	float64	0	[nan]	{nan: 416904}	{nan: 1.0}

65 rows × 6 columns

图 5-1 总表的分布式数据集

其中，“feature”代表数据特征列的列名，“dtype”为该列数据的数据存储格式，“nunique”为该列的奇异值总数，“vunique”为该列各个奇异值的具体内容，“cunique”为该列各个不同奇异值的各自总数，“dunique”为该列各个奇异值的各自总数占比。

5.2.4 数据特征分类

由于原始数据并表合成的总表直接置入机器学习模型进行训练效果不佳，因此不能直接置入机器学习模型当中进行训练，需要对合并后的总表进行特征工程处理，包括总表的重新拆分以及数据衍生。

首先是对总表的重新拆分，即数据特征分类。本文根据后续模型的对于数据特征的不同需要，首先对并表合成的总表的所有数据特征列进行合理的分类，将总表重新拆分成多张匹配机器学习模型训练方式的表。使得分类后的数据表能够更契合机器学习模型的训练，从而提高模型的预测能力。流程如图 5-2 所示。

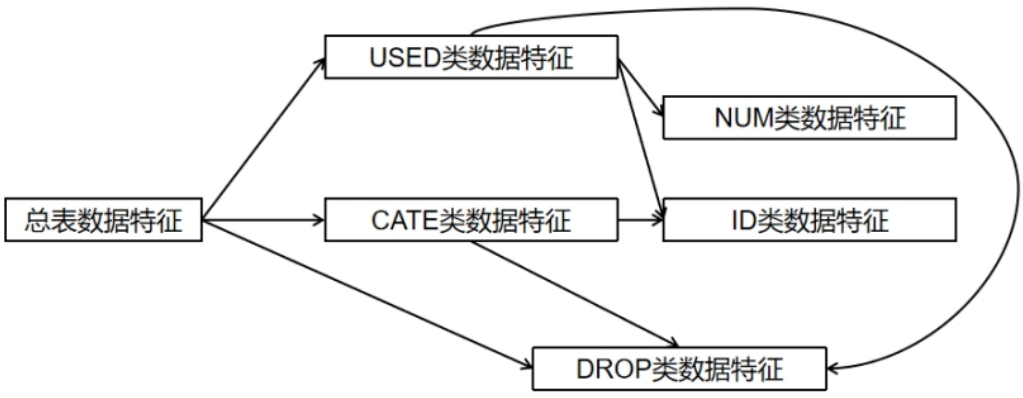


图 5-2 总表数据特征的处理流程

如图 5-2 所示，本文根据总表数据特征的奇异值数量对总表数据特征进行分类，在分类时设置分类的阈值为 1 和 5，将奇异值总数小于等于 1 的数据特征归入舍弃类数据特征（记为 DROP 类），因为奇异值为一代表整列数据皆相同，于是该列数据对模型的最终结果，即是案件否欺诈完全不产生任何的影响；将奇异值总数大于 1 且小于等于 5 的数据特征归入枚举型数据特征（记为 CATE 型），因为枚举型特征可以进行后续的特征衍生，选择阈值为 5 是出于对实际运算机器硬件条件限制，考虑到阈值每往上提高 1，特征衍生过程中对运算算力要求提高非常多，阈值设置过高则无法正常进行该特征衍生过程；将余下所有特征，即奇异值总数大于 5 的数据特征记为使用型数据特征（记为 USED 类），使用型数据特征还需要再进一步进行细化分类。

在完成初步分类之后形成了三种数据特征类，分别是舍弃类数据特征、枚举型数据特征和使用型数据特征。在二次分类中对枚举型数据特征和使用型数据特征分别进行进一步的数据分类调整。

首先对枚举型数据特征进行进一步的识别与分类：当数据特征名称后缀为“NO”或“CODE”等，这些数据特征名称表示该列数据的含义为编号或者代号，将此类数据特征归入 ID 类数据特征，因为此类数据特征仅代表一个编号，所包含的信息与案件本身无关；然后利用“crosstab”函数对奇异值个数相同并且奇异值总数比例相似的两个不同数据特征进行检测，若该函数输出的表格结果中每一行与每一列均只有一个值不为零，则说明被检测的两个数据特征仅数据特征列名有区别，该列数据所包含的信息重复，于是可以将其中一列数据特征归入舍弃类数据特征。

对 USED 类数据特征的细化分类时，第一步是对 USED 类数据特征中的每列数据进行数据格式的判断，由于机器学习模型不能处理非结构化数据，例如文本数据、图像数据等，所以在格式判断完成后将非结构化的数据特征列与时间型数据特征归入舍弃类数据特征。第二步则是与对枚举型数据特征的二次分类相同，在 USED 类数据特征中也会存在一些数据特征名称后缀为“NO”或“CODE”等的数据特征，这些数据特征名称仅仅表示该列数据的含义为编号或者代号，因此也需要将此类数据特征归入 ID 类数据特征，因为此类数据特征所包含的信息仅为一个编号，所包含的信息与案件本身无关。其余数据特征包括金额数据等单独归入数值型数据特征（记为 NUM 类）。

经过以上两步数据特征分类后，可以得到四个分类完成后的表格，分别为 ID 类

数据特征表、数值类数据特征标、枚举类数据特征表、舍弃类数据特征表。其中 ID 类数据特征表仅用作寻找案件对应关系、舍弃类数据特征表中的数据没有用处，只有数值类数据特征表和枚举类数据特征表中的数据为原始数据经过分类后的包含有效信息的数据，其中数值类数据特征表中有 3 个数据特征，枚举类数据特征表中有 22 个数据特征。如图 5-3 与图 5-4 所示。

```
In [45]: #特征信息汇总
df_features_CATEGORICAL = eda_dataframe(df_target[list_features_CATEGORICAL])# 展示特征的奇异值个数，奇异值的内容，奇异值使用量的占比；
df_features_CATEGORICAL.sort_values('nunique', ascending=False) # 依照某个字段中的数据进行排序，这里的字段是 “nunique”；
```

Out[45]:

	feature	dtype	nunique	vunique	cunique	dunique
0	NODESTATUS	int64	5	[0, 1, 4, 5, 2]	{0: 402544, 1: 10192, 2: 3188, 4: 545, 5: 435}	{0: 0.97, 1: 0.02, 2: 0.01, 4: 0.0, 5: 0.0}
2	CASEFLAG	int64	5	[1, 2, 5, 3, 7]	{1: 402586, 2: 14262, 3: 54, 7: 1, 5: 1}	{1: 0.97, 2: 0.03, 3: 0.0, 7: 0.0, 5: 0.0}
5	CASEMANAGERFLAG	int64	4	[0, 2, 3, 1]	{0: 413917, 2: 2874, 1: 110, 3: 3}	{0: 0.99, 2: 0.01, 1: 0.0, 3: 0.0}
9	CATASTROPHECODE1	float64	3	[nan, 6.0, 2.0, 7.0]	{nan: 416706, 7.0: 152, 6.0: 39, 2.0: 7}	{nan: 1.0, 7.0: 0.0, 6.0: 0.0, 2.0: 0.0}
1	TPFLAG	int64	3	[1, 0, 2]	{1: 312267, 0: 57200, 2: 47437}	{1: 0.75, 0: 0.14, 2: 0.11}
10	WHOLELOSS	float64	3	[0.0, 2.0, 1.0, nan]	{0.0: 415845, 2.0: 944, 1.0: 113, nan: 2}	{0.0: 1.0, 2.0: 0.0, 1.0: 0.0, nan: 0.0}
11	CHEATFLAG	float64	3	[nan, 1.0, 0.0, 2.0]	{nan: 404642, 0.0: 9284, 1.0: 2968, 2.0: 10}	{nan: 0.97, 0.0: 0.02, 1.0: 0.01, 2.0: 0.0}
6	TRANSFERFLAG	int64	3	[2, 0, 1]	{2: 415960, 0: 880, 1: 64}	{2: 1.0, 0: 0.0, 1: 0.0}
7	PLSWITCH	int64	2	[1, 0]	{1: 416902, 0: 2}	{1: 1.0, 0: 0.0}
8	COINSFLAG	int64	2	[0, 3]	{0: 415804, 3: 1100}	{0: 1.0, 3: 0.0}
4	MAJORLEVEL	int64	2	[0, 1]	{0: 416749, 1: 155}	{0: 1.0, 1: 0.0}
3	CLAIMTIMESFLAG	int64	2	[0, 1]	{1: 289190, 0: 127714}	{1: 0.69, 0: 0.31}
12	CLAIMTYPE	float64	2	[nan, 1.0, 0.0]	{nan: 374710, 1.0: 38611, 0.0: 3583}	{nan: 0.9, 1.0: 0.09, 0.0: 0.01}
13	ISCHECKEND	float64	2	[3.0, 2.0, nan]	{3.0: 416408, 2.0: 487, nan: 9}	{3.0: 1.0, 2.0: 0.0, nan: 0.0}
14	SELFHELPCLAIMFLAG	float64	2	[0.0, nan, 1.0]	{0.0: 415896, nan: 1002, 1.0: 6}	{0.0: 1.0, nan: 0.0, 1.0: 0.0}
15	CANSELFHELPCLAIMFLAG	float64	2	[0.0, 1.0, nan]	{0.0: 352692, 1.0: 61716, nan: 2496}	{0.0: 0.85, 1.0: 0.15, nan: 0.01}
16	ISBZPROPQUICKFLAG	float64	2	[0.0, nan, 1.0]	{0.0: 416633, 1.0: 264, nan: 7}	{0.0: 1.0, 1.0: 0.0, nan: 0.0}
17	ISBZPERSONQUICKFLAG	float64	2	[0.0, nan, 1.0]	{0.0: 416851, 1.0: 46, nan: 7}	{0.0: 1.0, 1.0: 0.0, nan: 0.0}
18	OPENCASEFLAG	float64	1	[nan, 1.0]	{nan: 416588, 1.0: 316}	{nan: 1.0, 1.0: 0.0}
19	STOREFLAG	float64	1	[1.0, nan]	{1.0: 409571, nan: 7333}	{1.0: 0.98, nan: 0.02}
20	CANCELCASEFLAG	float64	1	[nan, 1.0]	{nan: 415947, 1.0: 957}	{nan: 1.0, 1.0: 0.0}
21	AUDITLOCKFLAG	float64	1	[nan, 0.0]	{nan: 412338, 0.0: 4566}	{nan: 0.99, 0.0: 0.01}

图 5-3 枚举类数据特征的分布式数据集

```
In [46]: #特征信息汇总
df_features_NUMERICAL = eda_dataframe(df_target[list_features_NUMERICAL])# 展示特征的奇异值个数，奇异值的内容，奇异值使用量的占比；
df_features_NUMERICAL.sort_values('nunique', ascending=False) # 依照某个字段中的数据进行排序，这里的字段是 “nunique”；
```

Out[46]:

	feature	dtype	nunique	vunique	cunique	dunique
0	SUMPAID	float64	42330	[135.0, 550.0, 0.0, 4640.0, 62900.0, 14500.0, ...	{0.0: 106413, 2100.0: 43219, 2000.0: 42638, 10...	{0.0: 0.26, 2100.0: 0.1, 2000.0: 0.1, 1000.0: ...
1	SUMREALPAID	float64	41785	[135.0, 550.0, 0.0, 4640.0, 62900.0, 14500.0, ...	{0.0: 106417, 2100.0: 46445, 2000.0: 43299, 10...	{0.0: 0.26, 2100.0: 0.11, 2000.0: 0.1, 1000.0:...
2	SUMPREPAID	float64	3147	[0.0, 100000.0, 45896.0, 10900.0, 45358.3, 200...	{0.0: 403036, 2000.0: 2685, 100.0: 1824, 10000...	{0.0: 0.97, 2000.0: 0.01, 100.0: 0.0, 10000.0:...

图 5-4 数值类数据特征的分布式数据集

5.2.5 数据特征衍生

由于枚举型数据特征和数值型数据特征的总和，即有效数据特征的数量对于机器学习模型训练需求量来说仍然不够丰富，同时在对数据进行编码时不论何种人为编码方式都有可能会改变原始数据的所带有的一些信息，从而影响模型的预测能力

[42]，因此本文对枚举类数据特征进行额外的 One-Hot 编码，进行特征衍生，获得更加丰富的特征组合，以期提高后续模型的训练效果。

One-Hot 编码，也叫单热编码，该编码方式使用 n 位状态寄存器分别对一列数据特征中的 n 种不同状态进行编码。每个被编码的状态都由一个独立的寄存器存储，并且在任何时候仅有一位是有效的。

One-Hot 的编码方法是把分类变量编码成二进制向量表示。这个过程首先需要将分类值逐个映射为整数值。然后，将每个映射的整数值表示为一个二进制向量，除了整数的索引线之外，所有位置都是 0，标记为 1。下图 5-5 是 One-Hot 编码的效果示意图。

数据序号	xx特征
1	0
2	1
3	2
4	3

One-Hot 编码 →

数据序号	xx特征_0	xx特征_1	xx特征_2	xx特征_3
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

图 5-5 One-Hot 编码示意图

由图中可知，本文使用 One-Hot 编码时将新产生的数据特征列进行编号，方式为原特征名加后缀“_n”，其中 n 为第 n 个从原始特征中编码形成的新特征。

One-Hot 编码不仅良好解决了分类器无法对属性数据进行合适处理的问题，同时该编码方式也在一定程度上也起到了对整体数据特征进行扩充的作用。One-Hot 编码是将类别变量通过二进制形式的编码转换为机器学习算法能够读取的一种形式的过程。表 5-3 为 Python 编程语言中 One-Hot 编码的主要参数表。

表 5-3 One-Hot 编码主要参数表

参数名称	数据类型	参数描述	默认值
data	array-like， Series，or DataFrame	输入的数据	无
prefix	string	get_dummies 转换后，列名的前缀，默认为 None	None
columns	无	指定需要实现类别转换的列名。否则，转换所有分类列	None

表 5-3 One-Hot 编码主要参数表（续表）

参数名称	数据类型	参数描述	默认值
dummy_na	bool	新添一列用来表示空缺值，如果 False 就忽略空缺值	False
drop_first	bool	获得 k 中的 k-1 个类别值，去除第一个，防止出现多重共线性	False

本文使用Python编程语言中的“get_dummies”函数对枚举类数据特征进行One-Hot 编码，参数设置为“prefix=f， dummy_na=True”，其余为默认值，将原本去除的“is_fraud”数据特征加到枚举类数据特征列表中同步进行编码，编码前枚举类数据特征表为一个 426904 行*23 列的表格，将编码结果横向并入数值类数据特征表后得到一个 426904 行*78 列的新数据特征表，其中 3 列为原数值类数据特征，75 列为编码后的枚举类数据特征，相较于编码前数据特征数量增加了 52 个。

5.3 XGBoost 模型参数设置及初步结果分析

对原始数据进行的一系列处理完成后，下一步是将处理后的数据表置入 XGBoost 模型中进行模型训练，而在训练之前首先要对 XGBoost 模型的参数进行设置。

表 5-4 为 Python 编程语言中 XGBoost 模型的的主要参数介绍。

表 5-4 XGBoost 模型主要参数表

参数名称	参数描述	默认值
booster	有两种模型可以选择，gblinear 使用线性模型进行提升计算，gbtree 使用基于树的模型进行提升计算	无
objective	定义学习任务及相应的学习目标。	无
eval_metric	对于验证数据所需的评价指标，不同的目标函数都会有默认的评价指标。	None
eta	取值范围[0,1]。通过减小步长，提升计算过程更加保守。	0.3
max_depth	数的最大范围，取值范围为[1，∞]。	6
colsample_bytree	在建立树时对特征采样的比值，范围为（0，1]。	1
subsample	训练模型的子样本在整个样本集中所占的比例，可以防止过度拟合。取值范围为(0,1]	1

表 5-4 XGBoost 模型主要参数表（续表）

参数名称	参数描述	默认值
min_child_weight	孩子节点中最小的样本权重和。取值范围为[0，∞]。	1
seed	随机数	0
silent	当选择 0 时，将输出运行时信息;当选择 1 时，运行时信息将以静默方式打印出来	None

本文使用 Python 编程语言设置的 XGBoost 模型参数代码为“'booster': 'gbtree'，'objective': 'binary:logistic'，'eval_metric': 'auc'，'eta': 0.02，'max_depth': 5，'colsample_bytree': 0.7，'subsample': 0.7，'min_child_weight': 1，'seed': 1111，'silent': 1”。

其中“'booster': 'gbtree'”表示设置 XGBoost 模型使用基于树的模型进行提升计算；学习目标为“'objective': binary:logistic”，为二分类的逻辑回归问题，输出为概率；“'eval_metric': 'auc'”表示模型的训练结果利用 AUC 面积作为评价指标；：“'eta': 0.02”表示收缩步长为 0.02；“'max_depth': 5”表示树的最大深度为 5，控制训练过程的复杂程度；“'colsample_bytree': 0.7”表示特征采样比例为 0.7；“'subsample': 0.7”表示在所有投入模型的数据总量中，用于训练的数据占 70%其余 30%用于训练完成后的测试，即训练集与测试集的比例为 0.7：0.3；“'min_child_weight': 1”表示孩子节点中最小样本权重和为 1，“'seed': 1111”表示随机数种子为 1111，此处可以随机选择，都是随机数；“'silent': 1””表示模型的运行方式为缄默运行，不打印关于模型训练与测试的运行过程中的信息。

在训练完成后使用 Python 编程语言的“print”函数打印该模型的 AUC 面积、准确率指标、精度指标、召回率指标以及 F1 度量指标等各项结果。如表 5-5 所示：

表 5-5 XGBoost 模型在测试集上的表现

AUC	Accuracy	Precision	Recall	F1-Score
0.7982	0.9699	1	0.002387	0.004762

对表 5-5 结果进行分析，可知 XGBoost 模型 AUC 面积表现良好，说明该模型在二分类问题中具有较强的处理能力。此外该模型在精度与查准率方面表现良好，而召回率及于召回率相关的 F1 度量表现不佳，深入分析后发现原因在于该模型未考虑到模型不同提调率的设置对于模型各个指标会产生不同的影响。就召回率指标而言，如果提调率相对模型真实欺诈数据比例过小则会导致较多欺诈案件被判别为正常案件，从而出现准确率指标优秀而召回率指标非常低的情况。同时提调率设置的偏小

导致模型识别出的欺诈数据总量较小，结合分析召回率的定义可知，模型的召回率对提调率的变化敏感程度较高，不同的提调率的微小变化会对召回率产生极大的影响。

另一方面，由于欺诈数据在总数据中的占比较低，因此会存在大量正常数据被判别为正常数据的情况，在这种情况下，XGBoost 模型将欺诈数据判别为正常数据的部分占所有预测正确的比例也小、即正常数据判别正确的数量远远大于欺诈数据被判别为正常数据的数量，所以在提调率极小或提调率极大甚至为 1 的情况下（模型将所有数据全部判为欺诈），精度指标的变动程度较小，可以依然保持优秀（在提调率为 1 时精度等于正常数据占总数据的比例），同时在真实世界保险反欺诈的识别过程中也存在着欺诈数据远远少于正常数据的情况，所以在判断关于保险反欺诈数据的识别能力时，精度指标很难充分体现出不同模型对欺诈数据识别能力的强弱。

将指标的定义结合保险反欺诈的业务特性，通过上述分析后本文认为模型精度指标和模型召回率指标，这两者对于判断模型对反欺诈识别能力的强弱不具备较高的参考价值，因此后文不再关注模型精度、召回率与 F1 度量这三个指标，而是引入提调率指标，将重点放在对比不同模型的不同提调率下模型的准确率指标的表现。

5.4 横向对比

在完成模型初步分析以及确定相应指标、明确后续实证方向后，本文将特征工程处理后的数据置入包含 WDL 模型等在内的多种深度学习模型中，分别对数据进行深度学习模型的训练与测试，最后得到不同深度学习模型在不同提调率下与的精准率指标。将深度学习模型的这些结果与前文 XGBoost 模型的结果进行横向对比，以凸显本文主模型、即 XGBoost 模型在处理二分类问题时的横向比较优势。

笔者通过与保险业务人员的交流，同时结合本文样本数据情况，认为设定提调率的参考区间为 0-0.2 较为合理，并且以提调率 0.01 为间隔，得到每个模型的 20 个一一对应指标。

经过上述测试，可得在不同提调率条件下的 XGBoost 模型、AFM 模型、DCN 模型、DeepFM 模型、NFM 模型、PNN 模型 、WDL 模型各自的预测精准率结果，每个模型的 20 个指标结果如表 5-6 所示：

表 5-6 多模型提调率-准确率数据对比表

提调率	准确率						
	XGBoost	AFM	DCN	DeepFM	NFM	PNN	WDL
0.01	78.26%	65.22%	52.17%	65.22%	65.22%	60.87%	52.17%
0.02	80.43%	52.17%	63.04%	60.87%	60.87%	60.87%	60.87%
0.03	78.26%	50.72%	57.97%	57.97%	57.97%	59.42%	57.97%
0.04	74.73%	50.55%	54.95%	53.85%	53.85%	59.34%	62.64%
0.05	71.93%	50.88%	54.39%	51.75%	51.75%	57.02%	58.77%
0.06	69.34%	50.36%	56.93%	51.09%	51.09%	56.20%	54.74%
0.07	68.75%	50.94%	55.35%	51.57%	51.57%	54.72%	51.57%
0.08	69.78%	51.10%	55.49%	52.75%	52.75%	55.49%	50.55%
0.09	68.29%	51.22%	54.63%	52.68%	52.68%	54.63%	49.76%
0.10	67.98%	51.54%	54.19%	50.22%	50.22%	53.30%	48.02%
0.11	67.60%	49.20%	54.00%	50.40%	50.40%	51.60%	48.80%
0.12	65.57%	48.35%	51.65%	48.35%	48.35%	50.55%	48.72%
0.13	64.53%	47.97%	50.00%	47.97%	47.97%	50.34%	47.97%
0.14	62.70%	47.17%	49.37%	48.43%	48.43%	49.69%	48.74%
0.15	61.58%	46.04%	49.56%	47.21%	47.21%	48.39%	47.80%
0.16	60.16%	46.70%	48.35%	46.43%	46.43%	48.35%	46.70%
0.17	60.21%	46.89%	48.45%	45.60%	45.60%	48.19%	46.89%
0.18	59.17%	46.94%	48.17%	44.50%	44.50%	47.19%	47.19%
0.19	58.56%	46.99%	47.22%	44.91%	44.91%	46.99%	46.53%
0.20	58.24%	45.81%	46.04%	44.93%	44.93%	46.26%	46.26%

在表 5-6 中，第一列为设置好的不同间隔的提调率指标，本文样本数据的正样本比约 10%，基于现实考量，笔者认为取值范围在小于等于 20%之间的提调率所对应的模型准确率指标具有较高的参考价值。因此，在提调率-模型准确率数据表中，本文选取了 0-0.2 范围内提调率的模型表现。

第二列为 XGBoost 模型的准确率指标，第三列至第八列为六种不同深度学习模型的准确率指标。上述测试方案包含两大类：机器学习模型和深度学习模型，其中机器学习模型为 XGBoost 模型，该模型的数据是基于自动化特征分类方法和自动特征深度衍生方法；深度学习模型包含其余 6 种主流深度学习模型，这些模型使用的特征分类方法也是基于自动化特征分类方法和自动特征深度衍生方法。而通过定性观察上表，得到 XGBoost 模型+自动化特征分类+自动化特征衍生的模型组合在表中

设定不同提调率条件下的 20 个精准率指标均优于横向对比下的其他深度学习模型指标。由于特征分类过程没有区别，各个模型输入的数据也没有区别，因此可以确定 XGBoost 模型的准确率指标优于其他是源自自身模型的优势，因此应当选用该模型为本文后续进行保险反欺诈识别应用研究的最佳模型。此外，上述各种深度学习模型的性能差异较小，其中 DCN 模型为深度学习模型的代表模型。从模型结果的对比不难发现 XGBoost 模型相较于其余六种深度学习模型在二分类问题中的强大分析能力。因此，本文优先选择 XGBoost 模型+自动化特征分类+自动化特征衍生的模型组合为本文数据样本的基线模型，并对其进行了模型改良与优化，进一步提升了该模型对本文数据中欺诈案件的识别能力。

5.5 模型优化与纵向对比

5.5.1 模型优化

通过前文横向对比得出 XGBoost 模型+自动化特征分类+自动化特征衍生的模型组合为较优的保险反欺诈识别模型。由于此模型的识别能力只是相对更优，但客观上识别效果仍然不够优秀，因此本文要对此模型进行优化，提高对欺诈数据的识别能力。

在优化过程中，由于对 XGBoost 模型进行参数调整对于预测的准确率结果影响较小，可知在优化模型的过程中对原数据的不同处理方式对最后的模型预测结果有较大影响，因此本文对特征分类过程进行了进一步的深入改良。

本文所使用的数据主要分为两类，分别是结构化数据与非结构化数据。结构化数据指具有良好数据分类属性的数据，常见的结构化数据包含数值型数据，BOOL 型数据，枚举型数据，时间属性数据等，非结构化数据指文本数据，图像数据等。分类结构化数据是应用 AI 模型到现实场景业务的必经之路。前文使用的数据特征分类方法有基于特征属性值的自动化分类，自动化特征类别识别和特征衍生。其中基于特征属性值的自动化分类的局限性在于识别的准确性不足，类别型与数值型数据特征容易混淆，文本型数据难以识别等问题；而自动化特征类别识别和特征衍生的局限性在于数值型与枚举型数据分类模糊，文本数据难以识别，特征维数指数级增长等。如何将场景知识嵌入数据特征分类方案，实现场景数据特征的合理分类，是提升本文模型预测能力的关键。

本文的基线模型属于原始数据特征自动化分类：基于特征实例的统计学特征，通过自动识别特征实例的属性，获得数值型数据特征集合、枚举型数据特征集合、ID 型数据特征结合、和舍弃型数据特征集合，然后通过使用门限设置的方式来实现类别特征的再分类调整进而实现数据特征分类的二次调整。根据笔者对保险反欺诈业务流程的调研，本文提出了新的特征分类方法及场景深度特征衍生算法，特征工程的方法列举如表 5-7 所示：

表 5-7 新特征工程方法

特征工程名称			功能说明
特征分类	SAFE	semi-auto feature engineer	基于个人经验和场景理解，对数据特征进行分类。
特征深度衍生	SF	deep features by scene	基于场景规则，合成深度特征。

SAFE 是一种半自动数据特征分类方法，是对前文原始数据特征自动化分类模型的强化，可以应用于数据原始特征分类，可识别的类别有数值型、枚举型、时间型、BOOL 型等类型，并支持类别扩展。可自动或半自动的对原始特征进行分类，支持特征的统计结果展示和数据特征类别人工选择，本文在使用 SAFE 特征分类方式的基础上对分类后数据进行更细致的人工分类。SAFE 相对于前文原始数据特征自动化分类模型一方面增加了对时间数据的处理，另一方面加入人工处理步骤。该方法一方面增加了包含有效信息的数据特征的数量，另一方面通过人工分类优化了原始数据的数据特征分类情况。

特征工程方法 SAFE 是构造深度特征 SF 的基础，基于 SAFE 特征分类结果，本文构造了深度场景特征。

SF 是一种基于字段分类结果，通过笔者对保险欺诈业务的深度调研，与场景业务技术人员的深度探讨，从而构建出场景因子并基于场景因子分析，最终生成深度场景特征的一种特征深度衍生方法。该方法类似于在 XGBoost 模型的训练过程中以通过添加一些既定专家规则的方式来提高模型的训练效果。场景特征继承了场景业务逻辑，更加真实的反应了场景业务的经验、逻辑。

5.5.2 前后结果比较

本文对基线模型、SAFE 模型以及 SF 模型对数据进行二分类问题识别的预测能

力进行纵向比较。

基于结构化数据的自动化特征工程算法的综合性能（ROC 曲线）如图 5-6 所示：

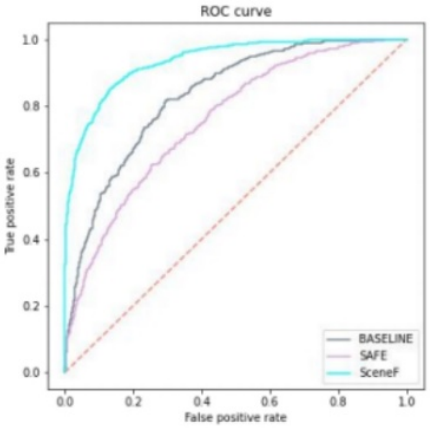


图 5-6 三种模型的 ROC 曲线对比图

从图 6-1 中可以看出，从左上往右下三条分别代表 SF 模型 ROC 曲线、基线模型 ROC 曲线以及 SAFE 模型 ROC 曲线；SAFE 模型相对于基线模型的 ROC 曲线下偏，综合性能相对基线模型降低；SF 模型对于二分类问题数据的识别的综合性能最强，ROC 曲线位于基线模型上方。

基线模型、SAFE 模型以及 SF 模型在不同提调率条件下的模型预测精准率结果纵向对比如表 5-8 所示：

表 5-8 多模型提调率-准确率数据对比表

提调率	基线模型	中间模型		最终模型	
	准确率	准确率	提升率	准确率	提升率
0.01	78.26%	91.30%	16.67%	100.00%	27.78%
0.02	80.43%	78.26%	-2.70%	100.00%	24.33%
0.03	78.26%	81.16%	3.70%	100.00%	27.78%
0.04	74.73%	70.33%	-5.88%	100.00%	33.82%
0.05	71.93%	66.67%	-7.32%	98.76%	37.30%
0.06	69.34%	64.96%	-6.32%	98.96%	42.72%
0.07	68.75%	64.78%	-5.77%	98.67%	43.52%
0.08	69.78%	62.09%	-11.02%	98.44%	41.08%
0.09	68.29%	59.51%	-12.86%	97.58%	42.89%
0.10	67.98%	59.03%	-13.17%	96.26%	41.60%
0.11	67.60%	58.80%	-13.02%	95.20%	40.83%
0.12	65.57%	58.61%	-10.61%	95.08%	45.00%

表 5-8 多模型提调率-准确率数据对比表（续表）

提调率	基线模型	中间模型		最终模型	
	准确率	准确率	提升率	准确率	提升率
0.13	64.53%	56.42%	-12.57%	94.50%	46.44%
0.14	62.70%	55.03%	-12.22%	93.33%	48.86%
0.15	61.58%	53.96%	-12.38%	91.08%	47.90%
0.16	60.16%	53.30%	-11.42%	89.11%	48.11%
0.17	60.21%	52.07%	-13.51%	87.00%	44.49%
0.18	59.17%	52.32%	-11.57%	85.12%	43.86%
0.19	58.56%	50.93%	-13.04%	83.28%	42.21%
0.20	58.24%	50.66%	-13.02%	81.15%	39.34%

分析上表可知，经过两步优化的最终模型（XGBoost 模型+SAFE 半自动数据特征分类+SF 特征深度衍生）在设定提调率条件下的精准率指标均优于本组其他模型指标，且在 0-0.2 的提调率条件下模型准确率都大幅提升。最终模型在 0-0.2 的提调率条件下的准确率指标最低为 81.15%，均高于 80%；最终模型在 0-0.15 的提调率范围内都能够达到 90 以上的准确率，甚至最终模型在 0-0.04 的较小提调率情况下出现了 100%完全准确的情况。

只经过一步优化的中间模型（XGBoost 模型+SAFE 半自动数据特征分类）相较于基线模型预测能力有所下降，但是半自动数据特征分类是特征深度衍生必不可少的中间环节，作为中间环节他的预测能力强弱不具备重要性。

5.6 模型其他指标

本文基于对预测模型在保险反欺诈业务中的作用、对保险反欺诈业务的调研以及场景业务技术人员的探讨，认为模型在一定准确率标准下能够识别出欺诈案件的数量多少也具有实际意义，也就是模型的提调率与准确率之间的关系可以互换，即在既定的准确率要求下，将模型在准确率标准下的提调率作为模型的性能指标之一。故对基线模型与最终模型进行重新测试，结果如表 5-9 所示：

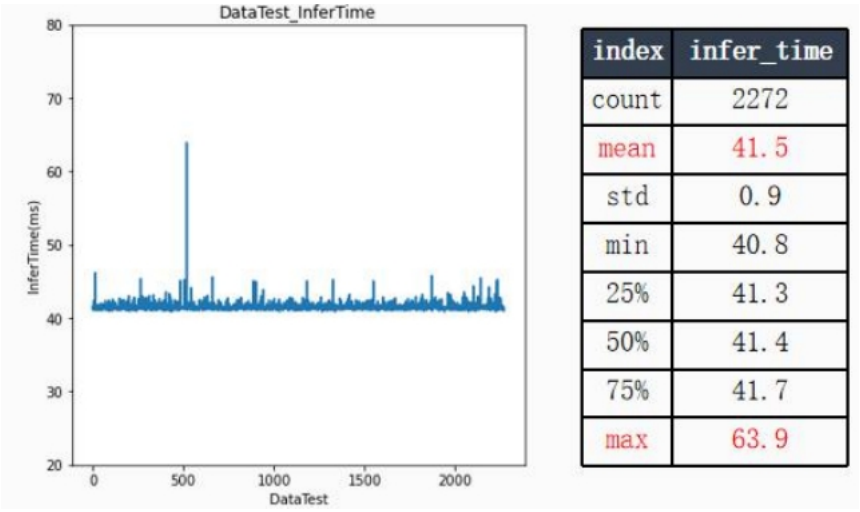
表 5-9 基线模型与最终模型准确率-提调率对比表

准确率	提调率		提调率提升
基线模型	基线模型	最终模型	
78.26%	0.01	0.219	2090.0%
80.43%	0.02	0.204	920.0%
78.26%	0.03	0.219	630.0%
74.73%	0.04	0.236	490.0%
71.93%	0.05	0.254	408.0%
69.34%	0.06	0.267	345.0%
68.75%	0.07	0.273	290.0%
69.78%	0.08	0.265	231.3%
68.29%	0.09	0.277	207.8%
67.98%	0.10	0.280	180.0%
67.60%	0.11	0.283	157.3%
65.57%	0.12	0.297	147.5%
64.53%	0.13	0.304	133.8%
62.70%	0.14	0.319	127.9%
61.58%	0.15	0.328	118.7%
60.16%	0.16	0.339	111.9%
60.21%	0.17	0.338	98.8%
59.17%	0.18	0.351	95.0%
58.56%	0.19	0.355	86.8%
58.24%	0.20	0.357	78.5%

表 5-9 中第一列为基线模型的准确率，第二列为在第一列基线模型的准确率下基线模型的提调率，第三列为最终模型在第一列准确率标准下的提调率，第四列为最终模型相对于基线模型的提条率提升程度。从表 5-9 中不难看出，第一列与第二列即为上文基线模型的测试结果，只不过互换了一下位置。此外，最终模型相对于基线模型在相同准确率情况下的提调率提升非常大，在最高的准确率标准 80.43%时提调率提升了 920%，在不同的准确率标准下，最终模型的提调率指标表现皆远优于基线模型，这意味着在保险实务中，当保险人通过对模型识别的准确率进行设定最低标准，，最终模型能够最大限度的将疑似欺诈案件之数显示出来，进而降低保险机构的经济损失。

当然，除了提调率-准确率指标的互换，还有一个在实务中较为重要的指标，即

该模型在对数据进行预测所需要的时间，若是花费时间过长则会降低该模型的实际应用价值。本文模型测试硬件配置为：CPU（Intel（R） Core（TM）i5-8265U @1.60GHz, 1.80GHz ）RAM（8.00GB）。最终模型测试单条数据所需的时间如图 5-7 所示：



index	infer_time
count	2272
mean	41.5
std	0.9
min	40.8
25%	41.3
50%	41.4
75%	41.7
max	63.9

图 5-7 最终模型单条数据测试时间

从图中可以看到，模型共计测试 2272 条数据，平均每条数据的测试所需时间为 41.5 毫秒，其中测试一条数据最低用时 40.8 毫秒，最高用时 63.9 毫秒，可知模型进行数据识别的时间耗费指标表现一般。

第 6 章　研究结论与展望

6.1 主要研究结论

本文采用国内某财险公司截至 2020 年 4 月的经过脱敏处理的车险理赔数据研究了人工智能在保险反欺诈中的应用，利用XGBoost、WDL和PNN等模型，将模型主要变量设置为模型提调率、模型预测准确率以及ROC曲线，分析人工智能模型在保险反欺诈业务中的作用。此外，本文还进行了保险反欺诈业务需求的分析，对提调率和模型反应速度也进行了测试。研究结论如下：

第一、在模型横向比较中，本文对比了机器学习模型和深度学习模型，机器学习模型为基线模型（XGBoost 模型+自动化特征分类+自动化特征衍生），深度学习模型包含 6 种主流深度学习模型，可得基线在设定提调率条件下的精准率指标均优于本组其他模型指标。各种深度学习模型的性能差异较小，其中 DCN 模型为深度学习模型的代表模型。

第二、在模型纵向比较中，最终模型（XGBoost 模型+SAFE 半自动数据特征分类+SF 特征深度衍生）在设定提调率条件下的精准率指标均优于基线模型，从 ROC 曲线图也可以直观的看出最终模型的综合性能优于基线模型；另一方面，只经过一步优化的中间模型（XGBoost 模型+SAFE 半自动数据特征分类）相较于基线模型预测能力有所下降，但是半自动数据特征分类是特征深度衍生必不可少的中间环节，作为中间环节他的预测能力强弱不具备太强的参考性。

第三、通过保险反欺诈业务需求分析，观察提调率和模型反应速度的测试结果，本文发现最终模型在设定准确率基准的情况下，提调率指标的表现远远优于基线模型。最终模型的单条数据测试响应时间指标表现一般，说明模型还存在去粗存精的可能性。

6.2 不足与展望

本文以中国人寿财产保险股份有限公司截至 2020 年 4 月的脱敏车险理赔数据为基础，建立了不同类型的预测模型，并比较了各模型的分类性能。最后发现 XGBoost 模型+SAFE 半自动数据特征分类+ SF 特征深度推导模型的组合预测效果较好，但本文仍有许多不足之处:

首先，我们可以尝试更多的机器学习算法。在第一次模拟考试中，我们使用XGBoost 和 6 种深度学习算法来构造一个单一的模型。机器学习和第一次模拟考试的算法是多种多样的。有很多算法可以用于保险反欺诈。稍后，我们可以尝试使用其他算法参与构建单个模型。

其次，我们可以尝试更多的模型组合。在构建复合模型时，本文仅将特征工程模型与 xgboost 模型相结合。第一次模拟考试是先构建多台单机模型，然后选择合适的模型参与组合模型的构建。而且，第一次模拟考试是第一次模拟考试。在这个过程中，我们不仅可以尝试用不同的模型来建立一个组合模型，还可以研究不同模型对组合模型的影响。

最后，寻找能够有效利用图像数据以及文本数据的算法模型。本文仅仅利用了原始数据中的结构化数据，对于文本数据以及图像数据利用率较低，可以开发例如知识图谱、图像识别算法等手段，再与结构化数据模型相结合，进一步提高总体模型的反欺诈识别能力。

参考文献

[1]叶明华. 基于 BP 神经网络的保险欺诈识别研究——以中国机动车保险索赔为例[J]. 保险研究，2011(3):79-86.

[2]刘坤坤. 车险保险欺诈识别和测量模型实证研究——基于广东省车险历史索赔数据[J]. 暨南学报：哲学社会科学版，2012，34(8):5.

[3] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[4] Viaene S, Derrig R A, Baesens, Bet al. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection[J]. Journal of Risk and Insurance, 2002, 69.

[5] Clifton Phua, Damminda Alahakoon, Vincent Lee. Minority report in fraud detection: classification of skewed data[J]，ACM SIGKDD Explorations Newsletter,2002, 6 (June 2004:1-19.

[6]JM Pérez, Muguerza J, Arbelaitz O, et al. Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance[J]. Springer-Verlag, 2005.

[7]郭元凯. 基于 XGBoost 的混合模型在股票预测中的应用研究[D]. 兰州理工大学，2020.

[8] Kia A. Using MLP and RBF Neural Networks to Improve thePrediction of Exchange Rate Time Series with ARIMA[J]. International Journal of Information and Electronics Engineering, 2012.

[9] Mallqui D C A, Fernandes R A S. Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques[J]. Applied Soft Computing, 2018.

[10]李超. 人民币汇率波动对我国国际贸易的影响[J]. 投资与创业，2018(1):2.

[11]郑红、叶成、金永红、程云辉. 基于 Stacking 集成学习的流失用户预测方法[J].应用科学学报，2020，38(6):11.

[12] Lalwani P, MK Mishra, Chadha J S, et al. Customer churn prediction system: a machine learning approach[J]. Computing, 2021(8):1-24.

[13] CHENG H T, KOC L, HARMSEN J, et al. Wide & deep learning for recommender

systems[C]//Deep Learning for Recommender Systems．2016:7-10．

[14] Zheng Z, Yang Y, Niu X, et al. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids[J]. IEEE Transactions on Industrial Informatics, 2017.

[15]NIU M,CAIJ．A Label Informative Wide & Deep Classifier for Patents and Papers[C]//In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP)．2019:3419-3434．

[16] Nguyen B P, Pham H N, Tran H, et al. Predicting the Onset of Type 2 Diabetes using Wide and Deep Learning with Electronic Health Records[J]. Computer Methods and Programs in Biomedicine, 2019, 182:9.

[17] Bastani K, Asgari E, Namavari H. Wide and Deep Learning for Peer-to-Peer Lending[J]. Papers, 2018.

[18]M ArtíS, Ayuso M, M Guillén. Modelling different types of automobile insurance fraud behaviour in the Spanish market[J]. Insurance Mathematics & Economics, 1999, 24(1–2):67-81.

[18] Herbert Weisberg. Quantitative methods for detecting fraudulent automobile bodily injury claims[J]。Researchgate, 1998, 10 (March 26 1998):1-19

[20]楚宵莹．基于机器学习的机动车辆保险的欺诈识别研究[D]．山东大学.

[21]M ArtíS, Ayuso M, M Guillén. Modelling different types of automobile insurance fraud behaviour in the Spanish market[J]. Insurance Mathematics & Economics, 1999, 24(1–2):67-81.

[22] Belhadji E B, Dionne G, Tarkhani F. A Model for the Detection of Insurance Fraud*[J]. The Geneva Papers on Risk and Insurance - Issues and Practice, 2000, 25(4):517-538.

[23] Caudill S B, Mercedes Ayuso and Montserrat Guillén. Fraud Detection Using a Multinomial Logit Model with Missing Information[J]. The Journal of Risk and Insurance, 2005, 72(4):539-550.

[24] Weisberg H I, Derrig R A. Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachussetts[J]. Journal of Insurance Regulation, 1991.

[25]Weisberg，H．I.，and R.A．Derrig，1995，“Identification and Investigation of

Suspicious Claims，in: AIB Cost Containment /Fraud Filing”，Boston，Mass: Automobile Insurers Bureau of Massachusetts.

[26]Weisberg，H. I.，and R.A. Derrig，1998，“Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims”， Risques，35，75-99.

[27] Bordoni S, Emilia R, Facchinetti G. Insurance Fraud Evaluation - A Fuzzy Expert System[C]// Fuzzy Systems, 2001. The 10th IEEE International Conference on. IEEE, 2001.

[28] John, A, Major, et al. EFD: A hybrid knowledge/statistical-based system for the detection of fraud[J]. International Journal of Intelligent Systems, 1992.

[29] Brockett P L, Xia X, Derrig R A. Using Kohonen's self [J]. Journal of Risk & Insurance, 1998, 65(: n2):245--274.

[30] Kou Y, Lu C T, Sirwongwattana S, et al. Survey of fraud detection techniques[C]// IEEE International Conference on Networking. IEEE, 2004.

[31] Yamanishi K, Takeuchi J I, Williams G, et al. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms[J]. Data Mining & Knowledge Discovery, 2000, 8(3):275-300.

[32] Bhowmik R. Detecting Auto Insurance Fraud by Data Mining Techniques[J]. Journal of Emerging Trends in Computing & Information ences, 2011, 2(4).

[33] John, A, Major, et al. EFD: A hybrid knowledge/statistical-based system for the detection of fraud[J]. International Journal of Intelligent Systems, 1992.

[34]马海花. 随机森林和 XGBoost 模型在个人信用风险评估中的应用. 中央民族大学，2021.

[34]张孟迪. 基于 Logistic 回归和 XGBoost 的银行信用卡客户流失预测. 山东大学，2021.

[36] Wang R, Fu B, Fu G, et al. Deep & cross network for ad click predictions[M] //Proceedings of the ADKDD'17. 2017: 1-7

[37] He X, Chua T S. Neural Factorization Machines for Sparse Predictive Analytics[J]. ACM SIGIR FORUM, 2017, 51(cd):355-364.

[38] Xiao J, Ye H, He X, et al. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks[J]. 2017.

[39] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender

systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10.

[40] Guo H, Tang R, Ye Y, et al. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction[J]. 2017.

[41] Qu Y, Cai H, Ren K, et al. Product-based Neural Networks for User Response Prediction[J]. 2016..

[42]邓拓. 基于 LSTM 神经网络的机场能见度预测. 山东大学, 2019.

致谢

转眼间，三年的研究生时间即将过去，回顾这三年的时光，感受颇多，我将珍惜这三年学到的知识和经验并带着它们向未来出发。

第一我想感谢我的导师。有一句古话说得很好，授人以鱼不如授人以渔，导师正是这样教导我们，他更多地在学习方法和学习习惯上给予我们指导和帮助。同时，在我论文写作过程中，我的导师从问题意识的培养到论文的成稿，都给予了我悉心的指导，他在学术上的造诣和对待工作的认真态度也令我敬佩，是我终身学习的榜样。此外，还要感谢保险学院所有的老师们，为我传授了受益匪浅的专业知识，是他们的谆谆教导促使我不断进步。在此，愿我的导师以及所有的老师们万事胜意。

第二我想感谢我的父母这么多年对我的支持和鼓励，在他们的保护下我可以快乐无忧地长大，爸爸妈妈总说：学术上我不能给你帮助，但是其他任何方面，我俩永远是你坚强的后盾。他们无时无刻的关心和帮助让我可以没有任何后顾之忧地去追求我的未来，我也希望自己可以尽快成长去报答父母的养育之恩和对我不计回报的付出。

第三我想感谢我的舍友熊，他随时随地地督促我帮助我在学术上的困惑和不解，他乐于助人又对自己严格要求，在学习和生活上都能做到一丝不苟，既是我学习的榜样又是我的好伙伴，希望他前程似锦。

第四我想感谢我的应老师，感谢你一直以来的陪伴，有你的帮助我才能成为现在的我，对于我的缺点和错误，你没有一味地纵容，而是包容和正确的引导，希望未来的日子我们还能一起进步。

第五我想感谢我师门的伙伴们，没想到我们能成为这么亲密的朋友，平时的吵闹和打趣不会影响我们在关键时刻相互陪伴相互支持，以后虽然我们要生活在祖国的不同地方，但是相信我们的友谊不会变温。

最后，衷心感谢在百忙之中抽出时间评阅本文、参加论文答辩的各位专家和教授！