主流机器学习方法识别车险欺诈效果的比较研究

陈　凯　李斌杰

[摘　要]　近年来，我国车险市场巨大的体量也催生了许多车险欺诈案件，然而传统的车险欺诈识别手段效率不佳，本文采用机器学习的方法，基于包含中国在内的四个数据集进行了实证分析，以比较六种主流机器学习方法对车险欺诈的预测表现以及预测表现的稳健性。本文对四个原始数据集进行数据分割，使原数据集分为训练集和测试集，训练集用于构建机器学习模型，测试集用于评估机器学习模型的效果，从而评估各机器学习方法的预测表现以及预测表现的稳定性。首先基于特征空间采用 SMOTE 采样法，使训练集中的欺诈样本数与非欺诈样本数达到平衡。之后采用 10 折交叉验证法选取最佳的参数组合来确定机器学习中的最优调节参数，并采用 ROC 曲线及曲线下方的面积 AUC 作为模型预测效果的评估标准，以避免主观选取截断点造成的影响。最终，研究发现极端梯度提升决策树模型和随机森林模型的预测表现以及预测表现的稳定性较好。

[关键词]　汽车保险；机器学习；SMOTE 采样；ROC 曲线

[中图分类号]　F840；TP181　　[文献标识码]　A　　[文章编号]　1004－3306(2022)12－0090－13

DOI: 10.13497/j.cnki.is.2022.12.006

一、引　言

近年来，我国车险市场巨大的体量也催生了许多车险欺诈案件。目前保险公司主要依靠查勘专家判断来识别车险欺诈，这种方法成本高、效率低，并不能很好地解决车险欺诈问题。随着机器学习方法的普及，各行各业已经开始运用机器学习方法做各种预测分析。保险公司也可以运用机器学习方法来识别车险欺诈，从而大大降低保险公司的费用支出。

从理论层面来看，机器学习方法的种类繁多，不同的机器学习方法对不同问题和不同数据集的有效性不同。但是当前的理论研究大多仅基于单一数据集运用机器学习方法建立机器学习模型，来比较各种机器学习方法对车险欺诈识别的效果，所得结果的稳健性得不到保证。因此，本文将基于国内外四个数据集分别建立六个机器学习模型：Logistic 回归、决策树(Decision Tree)、K 近邻(KNN，K Nearest Neighbors)、支持向量机(SVM，Supporting Vector Machine)、随机森林(Random Forests)以及极端梯度提升决策树(Extreme Gradient Boosting Decision Tree)。我们将对比不同的机器学习模型在同一个数据集中表现的差异以及同一个机器学习模型在不同的数据集中表现的差异，通过这种交叉分析的方法，研究机器学习方法在车险欺诈识别中的应用价值。

[作者简介]　陈　凯(通讯作者)，北京大学经济学院副教授，E-mail：chen.kai@pku.edu.cn；李斌杰，北京大学经济学院硕士研究生。

从现实层面来看，我国许多保险公司都在推动车险理赔的线上化和智能化，如中国平安的“智能闪赔”、阳光保险的“阳光 E 赔”等。显然，车险理赔的线上化与智能化需要车险欺诈识别技术作为支撑，否则保险公司会因理赔审核不严而遭受亏损。同时，车险理赔的线上化将使保险公司获取更多保单持有人的数据，这十分有利于机器学习方法的实施。因此，本文在分析中也采用了 2017 年中国某保险公司的车险索赔数据集，来探究机器学习方法对车险欺诈识别的效果，尤其是在中国市场的效果，是具有现实意义的。

目前车险欺诈的相关文献主要包括理论研究和实证研究。理论研究主要基于博弈论和信息不对称来探究保险欺诈形成的原因及可能的对策；实证研究则主要侧重于运用各种机器学习方法对保险欺诈进行识别。保险欺诈属于道德风险的一种，而最早将道德风险这一概念引入经济学领域的学者是 Arrow(1971)，他用道德风险这一概念解释了投保人的欺诈行为，即投保人购买保险后会改变自身的行为方式以获得保险利益，如故意制造交通事故。Holmstrom(1979) 以及 Spence 和 Zeckhauser(1971) 分别研究了事前道德风险和事后道德风险，事前道德风险即投保人在有了保险作为保障后会倾向于放任自身的危险行为从而导致保险事故发生的可能性上升，事后道德风险即在保险事故发生后投保人可能会通过掩盖真相、夸大损失等手段获取更多的保险赔偿金。毛钦(2008) 提出在信息不对称的环境下，投保人与保险人之间存在博弈关系，并针对我国车险欺诈问题提出尽快建立健全激励制度的建议。陈翠霞(2014) 同样基于博弈论分析车险欺诈行为，发现如果博弈进行的次数足够多，那么投保人和车险公司双方就会倾向于通过合作来获取长期的好处，并在此之上提出我国保险业需要建立消费者保险信誉体系。

在实证研究方面，目前文献基本都是基于某一个数据集运用各种机器学习方法来搭建机器学习模型，以评估所搭建的模型识别车险欺诈的效果。例如，德国某保险公司公开了 1994 年至 1996 年的索赔数据，目标变量是一个二元变量，记录了索赔的两种状态——正常索赔、欺诈索赔。其中，6% 的索赔是欺诈索赔，94% 的索赔是正常索赔。Phua 等(2004) 提出运用机器学习方法识别车险欺诈时要特别注意数据不平衡问题，即欺诈索赔样本远小于正常索赔的问题。Xu 等(2011) 使用了神经网络(Neural Network) 来搭建机器学习模型；Badriyah 等(2018) 使用了 K 近邻来搭建机器学习模型；徐徐等(2018) 对比了卷积神经网络(CNN，Convolution Neural Network)、Logistic 回归模型、K 近邻、支持向量机的预测效果，发现卷积神经网络的预测效果最好。另外一个数据集是马萨诸塞州汽车保险局(AIB，Automobile Insurers Bureau) 1993 年的车险索赔数据，目标变量是一个整数变量，取值范围为整数 0 到 10，记录了索赔的可疑程度——0 为正常保单、10 为可疑程度最高的保单，数字越大可疑程度越大。Brockett 等(2003) 运用了 PRIDIT(Principal Component Analysis of Relative to an Identified Distribution Unit) 方法，达到了较好的分类效果，但该方法的成本较高。由于该数据集并非公开数据集，所以在 Brockett 等后运用该数据集进行的研究较少。Francis(2016) 对该数据集进行了特征工程分析，删除了冗余变量，并生成了一个公开的模拟数据集供有需要的学者们研究，并用 PRIDT 和随机森林聚类(Random Forest Clustering) 对该数据集进行了分析。李秀芳等(2019) 使用了 Logistic 回归模型、支持向量机、决策树、K 近邻和朴素贝叶斯来搭建机器学习模型，并比较了 Bagging、Random Subspace 以及 Random Patches 三种 Bagging 方法改进基学习器的效果，发现 Bagging－决策树的表现最好；陈思迎(2019) 使用了 K 近邻和 K 均值聚类算法(K－means Clustering) 来搭建机器学习模型，得到了较好的预测效果。

总结来说，目前有学者比较过不同的机器学习方法识别车险欺诈的效果，这些研究都论证了其使用

的机器学习方法对车险欺诈有较好的识别效果。然而,由于公开的车险欺诈数据集很少,目前已有的文献均仅基于单一的数据集进行实证,所以结论的稳健性并不能得到保证,即某个机器学习方法在一个数据集的预测表现很好但在另一个数据集的预测表现并不一定也很好。然而,在研究机器学习方法识别车险欺诈的效果时,结论的稳健性十分重要。如果结论不稳健,那么保险公司直接选用研究中表明预测效果最佳的机器学习方法来开展保险欺诈识别工作就不一定能达到最满意的预测效果。

因此,本文分别基于四个数据集进行实证分析,以衡量六种主流机器学习方法对车险欺诈的识别效果和稳健性。六种主流机器学习方法分别为:Logistic 回归、决策树、K 近邻、支持向量机、随机森林以及极端梯度提升决策树。其中 Logistic 回归是广义线性模型的代表,决策树是非距离方法的代表,K 近邻是样本学习的代表,支持向量机是非线性方法的代表,随机森林是决策树经过 Random Patches 集成后的模型,极端梯度提升决策树是决策树经过 XGBoost(Extreme Gradient Boosting)集成后的模型。除了前文提到的 1994 年至 1996 年德国某保险公司索赔数据集、1993 年美国马萨诸塞州汽车保险局(AIB)模拟索赔数据集以外,本文还选取了 2015 年美国七州交通事故索赔数据集。这个数据集是 Bunty Shah 在 kaggle 上发布的公开数据集,据我们所知,目前尚未有基于该数据集的学术研究。这些数据集覆盖了欧洲和美国不同州的情况,可以在研究机器学习方法稳健性时控制地域的变量。同时,为了更好地分析机器学习方法是否符合中国的情况,本文还加入了 2017 年中国某保险公司的车险索赔数据集。

本文的主要创新在于对多个数据集而非单个数据集进行了实证检验,以比较不同的机器学习方法对车险欺诈的识别效果并且着重讨论识别效果的稳健性。如果仅基于单一的数据集进行实证,可能存在某个机器学习方法在一个数据集的预测表现很好但在另一个数据集的预测表现并不一定也很好的情况。而采用不同数据集可以讨论在不同情景下,控制地区因素和数据质量因素,从而针对车险数据的机器学习方法的有效性和表现水平,提高结论稳定性。另外一个次要贡献是本文对两个较新的数据集进行了分析,得到的结果和之前相比也比较稳健。

二、数据预处理

本文所使用的数据集包括德国某保险公司索赔数据集、美国马萨诸塞州 AIB 模拟索赔数据集、美国七州交通事故索赔数据集以及中国某保险公司 2017 年的车险索赔数据集。首先,需要对这四个数据集分别进行数据预处理,包括数据描述、数据分割以及 SMOTE 采样。

(一)数据描述

1. 德国某保险公司索赔数据集(以下简称“德国数据”)

该数据集为德国某保险公司 1994 年至 1996 年的索赔数据,包含 33 个变量(1 个目标变量和 32 个解释变量)和 15420 条样本。其中,目标变量是一个二元变量,记录了索赔的两种状态——正常索赔、欺诈索赔。正常索赔和欺诈索赔各占 94%、6%。在 32 个解释变量中,为避免相似变量,重复删除 PolicyType、AgeOfPolicyHolder、Year、PolicyNumber。此外,Make、DayOfWeekClaimed 、Days. Policy. Accident、Days. Policy. Claim、AddressChange. Claim 以及 NumberOfCars 均为类别型变量且都含有样本数量过少的水平,对预测效果不利,因此在建立机器学习模型前需要对样本过少的水平进行整合。

2. 美国马萨诸塞州 AIB 模拟索赔数据集(以下简称“美国 AIB 数据”)

该数据集为 Louis A Francis(2016)基于原数据集生成的公开模拟数据集。目前已有许多学者基于该模拟数据集开展了相关研究。该数据集的原数据集为马萨诸塞州汽车保险局(AIB,Automobile Insur-

ers Bureau) 收集的该州 1993 年的车险索赔数据，包含 100 多个变量。经 Francis 对原数据集进行特征工程分析、删除冗余变量后，该模拟数据集仅包含 27 个变量(1 个目标变量和 26 个解释变量) 和 1500 条样本。其中，目标变量是一个二元变量，记录了索赔的两种状态——正常索赔、欺诈嫌疑索赔。正常索赔和欺诈嫌疑索赔各占 69%、31%。在 26 个解释变量中，ID 记录了索赔编号，其提供的有效信息有限，因此删除。

3. 美国七州交通事故索赔数据集(以下简称“美国七州数据”)

该数据集为 2015 年美国 7 个州的交通事故索赔数据集，由 Bunty Shah 在 kaggle 网上公开发布，目前暂时没有基于该数据集的学术研究。该数据集包含 39 个变量(1 个目标变量和 38 个解释变量) 和 1000 条样本。其中，目标变量是一个二元变量，记录了索赔的两种状态——正常索赔、欺诈索赔。正常索赔和欺诈索赔各占 75%、25%。在 33 个解释变量中，policy_number 和 insured_zip 提供的有效信息有限，因此删除; incident_location 虽然记录了交通事故发生地的道路名，但由于数据集中的事故发生地过于分散，对预测效果不利，因此亦删除。此外，collision_type、property_damage、police_report_available 三个变量含有大量的缺失值，同样删除。最后，policy_bind_date、incident_date 和 months_as_customer 存在重复，故删除。

4. 中国某保险公司 2017 年索赔数据集(以下简称“中国数据”)

该数据集为 2017 年中国某保险公司的车险索赔数据集，包括 2017 年全年全国的拒赔数据，同时，随机抽样了全国正常结案的数据，将正常结案和拒赔案件混合后产生了数据样本。目标变量是一个二元变量，记录了两种状态——正常结案、拒赔案件。数据中原有 33 个相关变量(1 个目标变量和 32 个解释变量) ，其中被保险人姓名、被保险人年龄、身份证号、保单号、驾驶员姓名、驾驶证类型提供有效信息有限，故删除。出险时间段和报警时间段提供信息类似，进行合并。根据常见的欺诈因子特征，选出了 28 个相关变量(1 个目标变量和 27 个解释变量) 。再对数据进行清洗，剔除了结案金额小于 100 的案件①，最终得到有效拒赔案件 6548 条，正常结案数据 13827 条，共 20375 条。

(二) 数据特点

针对这四个数据集的特点，从以下三个维度分析了四个数据集之间的区别。

1. 欺诈样本占比

德国数据集的欺诈样本较少，另外三个数据集欺诈样本较多。其中美国 AIB 数据集的欺诈类型更为详细，包含硬欺诈和软欺诈(硬欺诈主要是指蓄意诈赔，软欺诈的典型例子是酒驾出险为了能获赔找人顶包) ，另外三个数据没有明确说明欺诈的类别。中国数据的总体数据量较大，相对欺诈样本的占比不大，但可以通过采样办法选择欺诈比例。

2. 数据质量

美国 AIB 数据集其实是模拟数据集，是数据集作者基于源数据集进行脱敏并且删除了冗余变量后的数据集，所以质量比其他三个数据集都好。德国数据集和中国数据集中规中矩，与赔案无关信息较少。美国七州数据集有很多从逻辑上来看就和欺诈无关联的变量，比如驾驶者家里的邮政编码，数据质量会比较差。

从结果来看，前两个数据质量比较好的数据集都是极端梯度模型表现最好，美国七州数据逻辑回归

① 结案金额较小时，欺诈可能性较小，拒赔案件较少，但数据量偏大，会误导欺诈分析结果。

表现最好。这是符合预期的，因为极端梯度模型拟合度比较好，对于质量好的数据集，该模型可以充分利用样本的信息，从而达到更好的预测效果。而对于质量差的数据集，用逻辑回归这种相对来说“浅拟合”的方法会更好，因为模型不会“过度解读”无关自变量与因变量之间的关系。

3. 样本量

德国数据和中国数据的数据样本量都较大，可以做较为复杂的分析。美国 AIB 数据为模拟数据，虽然样本量没有很大，但质量更好。美国七州数据集的样本量最少，叠加上面提到的数据质量差的问题，是四个数据集中比较特殊的一个。

(三) 数据分割

为了评估机器学习模型的表现，数据集需要被划分成训练集和测试集：训练集用于构建机器学习模型，测试集用于评估机器学习模型的效果。为了使训练集和测试集具有代表性，本文采取分层抽样方法抽取原数据集中 70% 的样本作为训练集以及 30% 的样本作为测试集。采用 R 语言 caret 包中的 createDataPartition 函数实现该操作。表 1 是四个数据集数据分割的结果，可以看出四个数据集拆分后的训练集和测试集中欺诈索赔的占比均与原数据集十分接近，因此四个数据集拆分后的训练集和测试集均具有代表性。

原数据集、训练集及测试集简况

表 1

德国数据	欺诈索赔占比	非欺诈索赔占比	样本量
原数据集	5.986%	94.014%	15420
训练集	5.994%	94.006%	10795
测试集	5.968%	94.032%	4625
美国 AIB 数据	欺诈索赔占比	非欺诈索赔占比	样本量
原数据集	31%	69%	1500
训练集	31.018%	68.982%	1051
测试集	30.958%	69.042%	449
美国七州数据	欺诈索赔占比	非欺诈索赔占比	样本量
原数据集	24.7%	75.3%	1000
训练集	24.679%	75.321%	701
测试集	24.749%	75.251%	299
中国数据	欺诈索赔占比	非欺诈索赔占比	样本量
原数据集	32.137%	67.863%	20375
训练集	32.139%	67.861%	14263
测试集	32.133%	67.866%	6112

(四) SMOTE 采样

从表 1 中可以看出，4 个训练集中的欺诈索赔占比与非欺诈索赔占比均失衡。过少的欺诈索赔样本不利于机器学习模型学习和捕捉欺诈索赔的特征，从而影响模型的预测效果。因此，训练集需要人为地重新进行采样，从而使训练集中欺诈索赔占比与非欺诈索赔占比相当。我们采用 SOMTE 采样法来改善这一不平衡性，使两者基本达到平衡。

1. 德国数据集

该数据集的原训练集中包含 647 个欺诈索赔样本和 10148 个非欺诈索赔样本，共 10795 个样本。由于该数据集的样本量较大，因此在每个欺诈索赔样本周围选择的近邻可以稍多。令每个欺诈索赔样本在其 20 个邻近（欧氏距离）索赔样本中随机选取 7 个样本，并在该样本与随机选取的样本的连线上随机选取一点作为新样本加入训练集。此时训练集中含有 5176（647 + 647 * 7）个欺诈索赔样本和 10148 个非欺诈索赔样本，再随机删除 10148 个非欺诈样本中的 4529 个样本，剩余 5619 个非欺诈样本。

2. 美国 AIB 数据集

该数据集的原训练集中包含 326 个欺诈嫌疑索赔样本和 725 个非欺诈索赔样本，共 1051 个样本。由于该数据集的样本量较小，因此在每个欺诈索赔样本周围选择的近邻不必过多。令每个欺诈嫌疑索赔样本在其 10 个邻近（欧氏距离）索赔样本中随机选取 1 个样本，并在该样本与随机选取的样本的连线上随机选取一点作为新样本加入训练集。此时训练集中含有 652（326 +326* 1）个欺诈嫌疑索赔样本和 725 个非欺诈索赔样本，再随机删除 725 个非欺诈样本中的 73 个样本，剩余 652 个非欺诈样本。

3. 美国七州数据集

该数据集的原训练集中包含 173 个欺诈索赔样本和 528 个非欺诈索赔样本，共 701 个样本。该数据集的样本量也较小，令每个欺诈索赔样本在其 10 个邻近（欧氏距离）索赔样本中随机选取 1 个样本，并在该样本与随机选取的样本的连线上随机选取一点作为新样本加入训练集。此时训练集中含有 346（173 +173* 1）个欺诈索赔样本和 528 个非欺诈索赔样本，再随机删除 528 个非欺诈样本中的 173 个样本，剩余 355 个非欺诈样本。

4. 中国数据集

该数据集的原训练集中包含 4584 个拒赔样本和 9679 个正常索赔样本，共 14263 个样本。该数据集的样本量同样较大，而且拒赔样本较多，占比达到 32. 14%。因此，维持原拒赔样本，随机删除 5095 个正常索赔样本。

表 2 是对训练集进行 SOMTE 采样后的结果，可以看出相比原训练集，经 SMOTE 采样后的训练集的平衡性大大改善。

SMOTE 采样结果

表 2

德国数据	欺诈索赔占比	非欺诈索赔占比	样本量
原训练集	5. 994%	94. 006%	10795
SMOTE 采样训练集	47. 948%	52. 052%	10795
美国 AIB 数据	欺诈索赔占比	非欺诈索赔占比	样本量
原训练集	31. 018%	68. 982%	1051
SMOTE 采样训练集	50%	50%	1304
美国七州数据	欺诈索赔占比	非欺诈索赔占比	样本量
原训练集	24. 679%	75. 321%	701
SMOTE 采样训练集	49. 358%	50. 642%	701
中国数据	欺诈索赔占比	非欺诈索赔占比	样本量
原训练集	32. 139%	67. 861%	14263
SMOTE 采样训练集	50%	50%	9168

三、机器学习模型搭建

在上述数据清洗、分割和 SMOTHE 采样的过程后，搭建 Logistic 回归模型、K 近邻模型、支持向量机模型、决策树模型、随机森林模型以及极端梯度提升决策树模型，并利用交叉验证法来确定各模型的调整参数。

(一) Logistic 回归模型

首先尝试使用所有变量进行 logistic 回归模型的搭建，四个数据集的回归结果的 AIC 值如表 3 所示。

一般来说，将所有变量引入 Logistic 回归模型的搭建并不是最好的选择。选取的解释变量越多，机器学习模型越有可能过拟合，从而造成模型在训练集中的表现非常好而在测试集中的表现欠佳，因此需要删除一些预测能力不强的解释变量，从而达到更好的预测效果。选用后向筛选法(Backward Selection)基于 AIC 指标来逐步剔除预测能力不强的解释变量，结果如表 4 所示。

Logistic 回归模型结果

表 3

Logistic 回归 AIC 值		Logistic 回归 AIC 值	
德国数据集	10624.00	美国七州数据集	633.68
美国 AIB 数据集	337.83	中国数据集	832.23

采用后向筛选法后的 Logistic 回归模型结果①

表 4

Logistic 回归 AIC 值		Logistic 回归 AIC 值	
德国数据集	10621.00	美国七州数据集	576.30
美国 AIB 数据集	332.17	中国数据集	837.81

注：参数估计结果和显著性水平可登录保险研究官网 https://bxyj.cbpt.cnki.net/WKH/WebPublication/index.aspx?mid=BXYJ，下载本文附录查看。

(二) K 近邻模型

K 近邻法由 Cover 和 Hart 于 1967 年提出，它是懒惰学习(lazy learning)的著名代表。K 近邻模型的算法简单直观。需要先确定近邻 K 的值，K 值的选取会在很大程度上影响模型的预测效果。K 值过小会导致模型过拟合，从而使机器学习模型的方差过大，影响预测效果；而 K 值过大会导致模型欠拟合，从而使机器学习模型的偏差过大，影响预测效果。为避免主观意愿影响 K 值的选取，我们在训练集中采用 10 折交叉验证确定 K 值，并进行交叉验证，如图 1 所示。

德国数据集的样本量较大，K 值的选取可以稍大。令 K 在 11 至 20 间取值，可以看出，当 K 取 12 时交叉验证误差最小。美国 AIB 数据集的样本数较小，K 值的选取不必过大。令 K 在 5 至 15 间取值，可

① 经过后向筛选后，德国某保险公司索赔数据集中 WitnessPresent 和 AgentType 两个解释变量被移除；美国马萨诸塞州 AIB 模拟索赔数据集 Inj02、Clt07、NumProv 以及 NumTreat 四个解释变量被移除；美国七州交通事故索赔数据集 auto_model、incident_city、incident_type、auto_make、insured_occupation、police_report_available、insured_education_level、incident_state、number_of_vehicles_involved、capital-gains、property_claim、total_claim_amount、policy_deductable、insured_sex 以及 umbrella_limit 共 15 个解释变量被移除。

以看出，当 K 取 9 时交叉验证误差最小。美国七州数据集的样本数较小，K 值的选取不必过大。令 K 在 5 至 15 间取值，可以看出，当 K 取 5 时交叉验证误差最小。中国数据集的样本量较大，K 值的选取可以稍大。令 K 在 11 至 20 间取值，可以看出，当 K 取 11 时交叉验证误差最小。

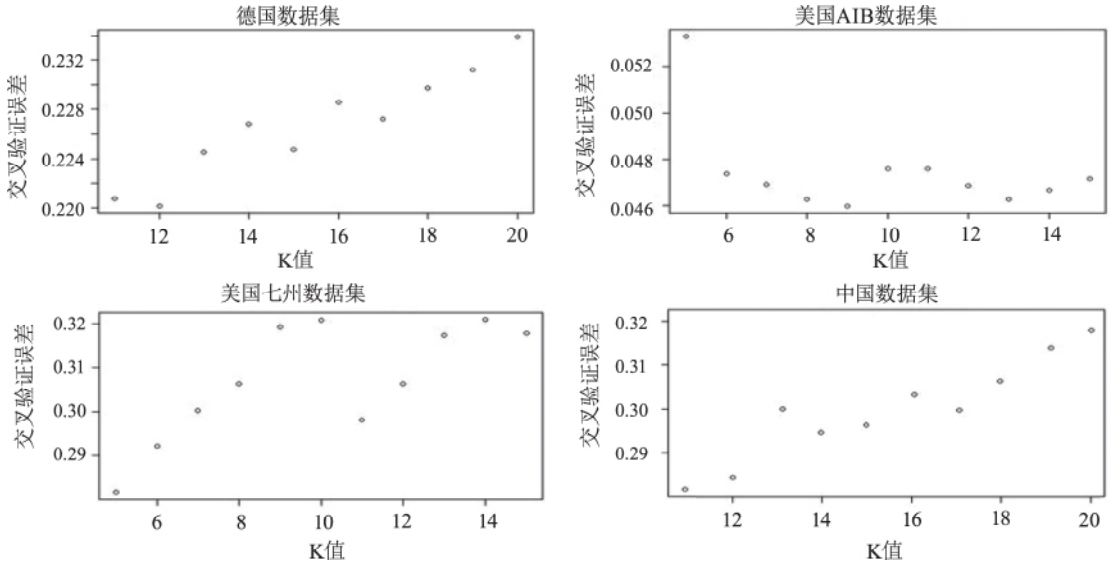

图 1 K 近邻模型 10 折交叉验证散点图

(三) 支持向量机模型

支持向量机通过引入核函数(kernel)的概念泛化了支持向量分类器的线性分界面，从而使非线性分界面的引入成为了可能。它是一种有坚实理论基础的新颖的适用小样本学习方法，基本上不涉及概率测度及大数定律等，简化了通常的分类和回归等问题。在搭建支持向量机模型之前，需要先确定核函数(kernel function)、常数 C 以及核函数对应的参数。这里同样采用 10 折交叉验证来确定这些参数的值，在交叉验证中，尝试多项式核函数(polynomial kernel function)以及径向核函数(radial kernel function)作为核函数。对于多项式核函数，分别考虑 1、2、3、4 作为多项式的次数；对于径向核函数，分别考虑 0.001、0.01、0.1、1 作为参数 γ 的值。对于两个核函数，分别考虑 1、10、100、1000 作为常数 C 的值。为方便展示，交叉验证结果如表 5，显示了两种核函数的最优交叉检验结果。

1. 德国数据集

对于多项式核函数，最小的交叉验证误差为 0.1199；对于径向核函数，最小的交叉验证误差为 0.1109，小于前者。因此，取径向核函数作为核函数、取 10 作为常数 C 的值、取 0.1 作为参数 γ 的值搭建支持向量机模型。

2. 美国 AIB 数据集

对于多项式核函数，最小的交叉验证误差为 0.02148；对于径向核函数，最小的交叉验证误差为 0.02224，大于前者。因此，取多项式核函数作为核函数、取 10 作为常数 C 的值、取 3 作为多项式次数搭建支持向量机模型。

3. 美国七州数据集

对于多项式核函数，最小的交叉验证误差为 0.1440；对于径向核函数，最小的交叉验证误差为 0.1340，小于前者。因此，取径向核函数作为核函数、取 100 作为常数 C 的值、取 0.01 作为参数 γ 的值

搭建支持向量机模型。

4. 中国数据集

对于多项式核函数，最小的交叉验证误差为 0.1041；对于径向核函数，最小的交叉验证误差为 0.1028，小于前者。因此，取径向核函数作为核函数、取 10 作为常数 C 的值、取 0.1 作为参数 γ 的值搭建支持向量机模型。

支持向量机模型 10 折交叉验证结果

表 5

德国数据集					
多项式核函数			径向核函数		
常数 C	多项式次数	交叉验证误差	常数 C	γ	交叉验证误差
1000	3	0.1199	10	0.1	0.1109
美国 AIB 数据集					
多项式核函数			径向核函数		
常数 C	多项式次数	交叉验证误差	常数 C	γ	交叉验证误差
10	3	0.02148	10	0.1	0.02224
美国七州数据集					
多项式核函数			径向核函数		
常数 C	多项式次数	交叉验证误差	常数 C	γ	交叉验证误差
1000	2	0.1440	1000	0.01	0.1340
中国数据集					
多项式核函数			径向核函数		
常数 C	多项式次数	交叉验证误差	常数 C	γ	交叉验证误差
1000	3	0.1041	10	0.1	0.1028

（四）决策树模型

决策树模型的基本思想是通过一系列拆分（splits）来把训练集中的样本分到不同的组别里，这些拆分基于解释变量的值。对二分类问题而言，测试集中样本系某类的预测概率值即为其所在组别内训练集中该类样本的占比。由于不加限制地搭建决策树很有可能造成过拟合问题，因此在搭建决策树模型时，我们添加了一些限制：决策树的终节点（terminal nodes）的最低样本数；决策树的最大深度（depth）；设定决策树每次拆分（split）至少需要提高的节点纯度（node purity，以基尼指数衡量），并利用 10 折交叉验证确定该值。不同数据集中决策树模型的 cp 值的交叉验证结果见表 6。

决策树模型 10 折交叉验证结果

表 6

德国数据集		美国 AIB 数据集		美国七州数据集		中国数据集	
cp 值	交叉验证误差	cp 值	交叉验证误差	cp 值	交叉验证误差	cp 值	交叉验证误差
0.001474	0.1902	0.006684	0.05217	0.01168	0.2024	0.001352	0.1832

（五）随机森林模型

随机森林模型是将集成学习方法 Random Patches 用于决策树模型得到的机器学习模型，该模型可

以有效地降低决策树模型预测值的方差从而达到更好的预测效果。我们运用自展法(bootstrap) 有放回抽样的次数,这个次数越大越能降低预测值的方差,从而达到更好的预测效果,考虑到算力,将该值设为 2000。仍然执行 10 折交叉验证确定备选解释变量的个数。不同数据集中随机森林模型的备选解释变量个数的交叉验证结果见表 7。

随机森林模型 10 折交叉验证结果

表 7

德国数据集		美国 AIB 数据集		美国七州数据集		中国数据集	
备选解释变量数	交叉验证误差	备选解释变量数	交叉验证误差	备选解释变量数	交叉验证误差	备选解释变量数	交叉验证误差
18	0. 1043	6	0. 01918	22	0. 08713	15	0. 1327

(六) 极端梯度提升决策树模型

极端梯度提升决策树模型是将集成学习方法 XGBoost(Extreme Gradient Boosting) 用于决策树模型得到的机器学习模型,该模型可以进一步降低决策树模型预测值的偏差从而达到更好的预测效果。在参数设置上,我们尝试每个基决策树的最大深度分别为 1、3、5; XGBoost 算法的迭代次数分别为 2000、3000、4000; 学习速率分别为 0. 005、0. 01、0. 05、0. 1。

德国数据集的交叉验证结果显示①最优结果为学习速率取 0. 01、决策树最大深度取 5、基决策树备选解释变量比例取 0. 4、基决策树使用样本比例取 0. 6、迭代次数取 4000,此时可以达到最小交叉验证误差 0. 06318。美国 AIB 数据集的交叉验证结果显示最优结果为学习速率取 0. 01、决策树最大深度取 5、基决策树备选解释变量比例取 0. 4、基决策树使用样本比例取 0. 8、迭代次数取 2000 时,此时达到最小交叉验证误差 0. 01610。美国七州数据集的交叉验证结果显示最优结果为学习速率取 0. 005、决策树最大深度取 5、基决策树备选解释变量比例取 0. 6、基决策树使用样本比例取 0. 8、迭代次数取 2000 时,此时达到最小交叉验证误差 0. 09567。中国数据集的交叉验证结果显示最优结果为学习速率取 0. 01、决策树最大深度取 3、基决策树备选解释变量比例取 0. 4、基决策树使用样本比例取 0. 6、迭代次数取 4000 时,此时达到最小交叉验证误差 0. 04936。

四、预测效果比较分析

灵敏度(sensitivity) 与特异度(specificity) 都是衡量预测效果的指标,灵敏度表示真实为阳的样本中被机器学习模型正确识别为阳的比例,而特异度表示真实为阴的样本中被机器学习模型正确识别为阴的比例。一个好的机器学习模型需要同时取得高的灵敏度和特异度。但这两者都需要事先设定截断点(cutoff) ,改变截断点会影响最终结果,因此,为了避免主观意愿影响评估结果,我们使用 ROC(Receiver Operating Characteristic Curve) 曲线(每一个截断点的值对应着 ROC 曲线上的一点,点的连线即 ROC 曲线) 以及曲线下方的面积 AUC(Area Under the Curve) 来衡量预测效果。

ROC 曲线越往左上方凸说明模型的预测效果越好,如图 2 所示,但凸度的大小不够直观。因此,采用 AUC 值来反映 ROC 曲线下方的具体面积数值。之后,对不同数据集的 AUC 值做了排序,以此来比较不同机器学习方法的效果和稳健度。

① 交叉验证结果过长,这里仅给出最优的参数组合及其对应的交叉验证误差。

1. 德国数据集

在这个数据集中，极端梯度提升决策树模型的表现最好，其 AUC 为 0.8198。随机森林模型的 AUC 排名第二，相对 AUC 为 99.37%。对于此数据集而言，使用极端梯度提升决策树模型和随机森林模型都能达到优异的预测效果。

2. 美国 AIB 数据集

在这个数据集中，仍然是极端梯度提升决策树模型的表现最好，其 AUC 为 0.992。随机森林模型的 AUC 排名第二，相对 AUC 为 99.99%。对于此数据集而言，使用极端梯度提升决策树模型和随机森林模型都能达到优异的预测效果。

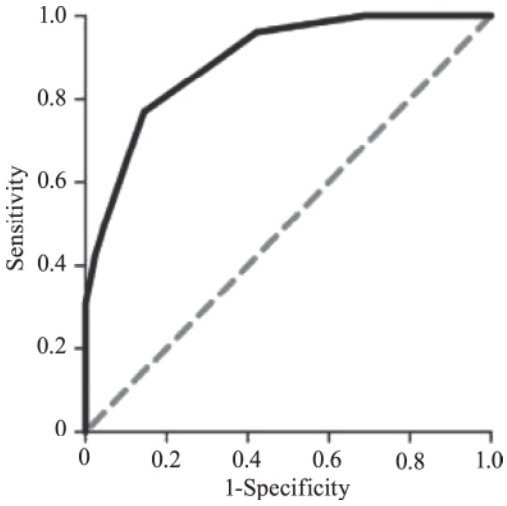

图 2　ROC 曲线示例

3. 美国七州数据集

这个数据集的表现与以上两个数据集略有不同，逻辑回归模型的表现最好，其 AUC 为 0.849。随机森林模型的 AUC 排名第二，相对 AUC 为 99.60%。对于此数据集而言，使用逻辑回归模型和随机森林模型都能达到优异的预测效果。极端梯度提升决策树模型的表现排名第三，相对 AUC 为 98.09%，差距不算太大。

4. 中国数据集

在这个数据集中，极端梯度提升决策树模型的表现最优，其 AUC 为 0.859。此外，随机森林模型的 AUC 排名第二，相对 AUC 为 98.60%。对于此数据集而言，使用极端梯度提升决策树模型和随机森林模型都能达到优异的预测效果。

表 8 汇总了四个实证数据集中各模型的预测效果。虽然没有一个机器学习模型能在四个数据集中都表现最优，但极端梯度提升决策树模型和随机森林模型的整体表现较好。前者在三个数据集中的预测表现排名第一，仅在美国七州数据集中的预测表现排名第三（与该数据集中最高 AUC 的差额约为 2%）；后者在四个数据集中的预测表现都排名第二，且与各数据集中最高 AUC 的差额都在 2% 以内。由此可见，相同的机器学习模型在不同的数据集中的预测表现略有差异，这说明数据本身的质量会对机器学习的效果产生一定影响。但总体来看，极端梯度提升决策树模型和随机森林模型对于不同国家和地区的实证预测结果都不错，可以作为识别车险欺诈的基本模型。

四个实证数据集中各模型预测效果汇总

表 8

	德国数据集			美国 AIB 数据集			美国七州数据集			中国数据集		
模型	AUC	相对 AUC①（%）	排名	AUC	相对 AUC（%）	排名	AUC	相对 AUC（%）	排名	AUC	相对 AUC（%）	排名
Logistic 回归	0.763	93.09	4	0.981	98.86	3	0.849	100.0	1	0.670	78.00	6
K 近邻	0.649	79.17	6	0.973	98.07	4	0.591	69.64	6	0.702	81.72	5
支持向量机	0.762	92.90	5	0.96	96.80	5	0.753	88.75	4	0.794	92.43	3
决策树	0.791	96.49	3	0.953	96.03	6	0.736	86.75	5	0.758	88.24	4
随机森林	0.814	99.35	2	0.992	99.99	2	0.846	99.60	2	0.847	98.60	2
极端梯度提升决策树	0.82	100.0	1	0.992	100.0	1	0.833	98.09	3	0.859	100.0	1

① 相对 AUC 即各机器学习模型 AUC 占最大 AUC 的比例。

五、结论与展望

随着我国保险市场的发展，保险欺诈案件越来越频繁。在各险种中，车险的欺诈问题尤为严重，已经严重扰乱了保险公司的经营、侵害了各方利益。近年来，车险理赔的线上化浪潮又对保险公司监测保险欺诈问题提出了新的挑战。由于传统车险欺诈识别手段的效率不佳，越来越多的文献和保险公司开始尝试使用机器学习这一新技术来识别车险欺诈。过往文献在比较各种机器学习方法识别车险欺诈的效果时均仅基于单一数据集进行实证检验，结论的稳健性难以保证。因此，本文基于四个数据集进行了实证检验，以比较六个主流机器学习方法识别车险欺诈的效果以及稳健性。本文分别建立了 Logistic 回归模型、K 近邻模型、支持向量机模型、决策树模型、随机森林模型以及极端梯度提升决策树模型。机器学习模型建立后，利用 ROC 曲线及其对应的 AUC 比较各模型的预测效果。

研究发现虽然没有一个机器学习模型能在四个数据集中都表现最优，但极端梯度提升决策树模型和随机森林模型的整体表现较好。前者在三个数据集中排名第一，仅在美国七州数据集中的预测表现排名第三；后者在四个数据集中的预测表现都排名第二，且与各数据集中最高 AUC 的差额都在 2% 以内。而且在针对中国车险数据集时，极端梯度提升决策树模型的表现要明显优于其它方法。

由于没有一个机器学习模型能在四个数据集中都表现最优，保险公司在运用机器学习方法预测车险欺诈时，需要选择表现较稳定的随机森林模型或极端梯度提升决策树模型或二者择优。如果资源允许或对预测效果有极高要求，保险公司也可以选择定期重新建立多个机器学习模型并从中挑选表现最佳的模型用于预测车险欺诈问题。

此外，我们也发现同一个机器学习模型在不同的数据集中表现的略有差异，这说明数据本身的质量会对机器学习的效果产生一定影响。从理论上来说，对于数据质量较好的数据集应该采用使用极端梯度决策树这类的“深拟合”模型。这样可以充分利用样本的信息，从而达到更好的预测效果。这四个数据集中，德国数据、美国 AIB 数据和中国数据都属于质量较好的数据，因此极端梯度提升决策树模型也发挥了其算法优势。对于数据质量较差的数据集，用逻辑回归这种相对来说“浅拟合”的方法会更好，因为这类模型不会过度解读无关自变量和因变量之间的关系。例如美国七州数据，简单的 Logistic 回归模型达到了最佳的效果。在实践中，保险公司如果决定用机器学习识别欺诈，在初期数据质量应该是比较差的，建议采用逻辑回归、随机森林这种类型的机器学习方法。在保险公司积累了一定经验，逐步改善数据质量之后，就可以考虑使用极端梯度决策树这类模型。

在其他影响因素方面，数据集中解释变量对目标变量的解释能力以及数据采集的准确性也会对车险欺诈识别的效果产生影响。通过分析不同国家和地区的数据集，可以在一定程度上控制该变量，但未来还需要进一步控制其他变量来分析相对最优的机器学习方法。

目前国内外已有文献研究了如何选取机器学习模型以达到更好的车险欺诈识别效果，但鲜有文献涉及数据质量对车险欺诈识别效果的影响。未来的研究可以聚焦于寻找对车险欺诈解释力强的解释变量，从而从改善数据质量的角度进一步加强机器学习对车险欺诈的识别效果。机器学习在保险欺诈识别中的应用仍处于探索阶段，保险公司在尝试使用机器学习方法预测车险欺诈时需要保持谨慎，特别是在应用的初期，保险公司不可全盘放弃传统方法，而应循序渐进地提高机器学习方法在车险欺诈识别中的比重。

[参考文献]

[1] 陈翠霞. 我国机动车辆保险欺诈骗赔的博弈分析[J]. 保险职业学院学报,2014,28(4):4.

[2] 陈思迎. 大数据背景下机动车辆保险欺诈风险及其防范研究[D]. 西南财经大学,2019.

[3] 李秀芳,黄志国,陈孝伟. Bagging 集成方法在保险欺诈识别中的应用研究[J]. 保险研究,2019,(4):66-84.

[4] 毛 钦. 汽车保险欺诈骗赔的博弈分析[J]. 商品储运与养护,2008,(9):47-49.

[5] 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.

[6] Arrow,K. J.,Insurance Risk and Resource Allocation[M]. Berlin,Springer Netherlands,1992,14:220-229.

[7] Badriyah T,Rahmaniah L,Syarif I. Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance[C]// 2018 International Conference on Applied Engineering(ICAE). 2018. doi:10.1109/INCAE.2018.8579155.

[8] Brockett P L,Derrig R A,Golden L L,et al. Fraud Classification Using Principal Component Analysis of Ridits[J]. Social Science Electronic Publishing,2003.

[9] Cover,T. M.,and P. E. Hart,Nearest Neighbor Pattern Classification[J]. IEEE Transactions on Information Theory,1967,13(1):21-27.

[10] Francis,L. Application of Two Unsupervised Learning Techniques to Questionable Claims:PRIDIT and Random Forest. In E. Frees,G. Meyers,& R. Derrig(Eds.),Predictive Modeling Applications in Actuarial Science(International Series on Actuarial Science,pp. 180-207). Cambridge:Cambridge University Press. 2016. doi:10.1017/CBO9781139342681.008

[11] Holmstrom B. Moral Hazard and Observability[J]. CORE Discussion Papers RP,1979,10(1):74-91.

[12] Phua C,Alahakoon D,Lee V. Minority Report in Fraud Detection:Classification of Skewed Data[J]. Acm Sigkdd Explorations Newsletter,2004,6(1):50-59.

[13] Spence M,Zeckhauser R. Insurance,Information,and Individual Action[J]. Uncertainty in Economics,1971,61(2):380-387.

A Comparative Study on the Effectiveness of Machine Learning Methods in Auto Insurance Fraud Identification

CHEN Kai,LI Bin-jie

Abstract: The magnitude of China's auto insurance market has induced a large amount of auto insurance frauds. However,the traditional auto insurance fraud identification methods are not effective. This paper uses machine learning methods and makes an empirical analysis based on four data sets to compare the prediction performance and robustness of six mainstream machine learning methods on auto insurance fraud detection. We split all four original data sets into training set and test set. The training set is used to build the machine learning model,and the test set is used to evaluate the effect of the machine learning model. Together,we evaluate the prediction performance of each machine learning method and the robustness of the prediction performance. Firstly,we use SMOTE sampling method to generate new data,in order to balance the number of fraud samples and non-fraud samples in the training set. We then use the 10-fold cross validation method to select the best parameter combination to determine the optimal adjustment parameters in machine learning. We use the Receiver Operating Characteristic Curve and the Area Under the Curve as the evaluation standard of the prediction effect of the model. Finally,we find the prediction performance and robustness of the stochastic forest model and extreme gradient lifting decision tree model are better.

Key words: auto insurance; machine learning; SMOTE Sampling; ROC Curve

[编辑:施 敏]