Frontiers of International Accounting 国际会计前沿, 2024, 13(6), 964-975
Published Online December 2024 in Hans. https://www.hanspub.org/journal/fia
https://doi.org/10.12677/fia.2024.136125

Hans汉斯

保险反欺诈预测

李欣朵，余　旺

重庆理工大学理学院，重庆

收稿日期：2024年10月23日；录用日期：2024年11月28日；发布日期：2024年12月11日

摘　要

保险行业不断发展，保险欺诈现象在社会上越来越常见，不仅给保险公司带来巨大的损失，而且对社会发展产生了极其不利的影响。因此需要建立保险欺诈预测模型，从一定程度上遏制保险欺诈行为。利用700条保险反欺诈数据，其中特征有39个，标签为是否保险欺诈。首先，对数据进行缺失值、重复值、日期数据转化、类别数据离散化等数据预处理，以便于后续模型的建立。同时考虑到特征之间可能存在共线性问题，对特征进行方差分析、相关性分析筛除掉特征之间相关性大的多余特征，以及与标签相关性小的无用特征。对上述数据处理和特征筛选后的数据建立机器学习模型，来预测保险欺诈行为。选择使用用于分类预测的常见模型：logistics回归、knn以及Bagging集成学习随机森林模型和Boosting集成学习的LightGBM。从各模型对测试集的预测结果评估可以发现LightGBM模型的整体的模型预测性能最好，预测准确率达到88%，可以作为保险欺诈预测的参考模型。而logistics回归、knn模型存在对保险诈骗的预测准确率较低，将大部分数据预测为非保险诈骗数据，因此实际应用性较差。

关键词

保险欺诈，方差分析，相关性分析，机器学习预测

Insurance Anti-Fraud Forecasting

Xinduo Li, Wang Yu

College of Science, Chongqing University of Technology, Chongqing

Received: Oct. 23rd, 2024; accepted: Nov. 28th, 2024; published: Dec. 11th, 2024

Abstract

With the continuous development of the insurance industry, insurance fraud has become more and more frequent in society, which not only brings huge losses to insurance companies, but also has an extremely adverse impact on social development. Therefore, it is necessary to establish an insurance

文章引用: 李欣朵, 余旺. 保险反欺诈预测[J]. 国际会计前沿, 2024, 13(6): 964-975.
DOI: 10.12677/fia.2024.136125

fraud prediction model to curb insurance fraud to a certain extent. 700 insurance anti-fraud data are used, of which there are 39 features and the label is whether it is insurance fraud. First, the data is preprocessed by missing values, duplicate values, date data conversion, and category data discretization to facilitate the establishment of subsequent models. At the same time, considering the possible collinearity problem between features, variance analysis and correlation analysis are performed on the features to screen out redundant features with a high correlation between features and useless features with a low correlation with labels. A machine learning model is established for the data after the above data processing and feature screening to predict insurance fraud. Common models for classification prediction are selected: logistics regression, knn, Bagging ensemble learning random forest model and Boosting ensemble learning LightGBM. From the evaluation of the prediction results of each model on the test set, it can be found that the overall model prediction performance of the LightGBM model is the best, with a prediction accuracy of 88%, which can be used as a reference model for insurance fraud prediction. However, the prediction accuracy of logistics regression and knn models for insurance fraud is low, and most of the data are predicted as non-insurance fraud data, so the actual application is poor.

Keywords

Insurance Fraud, Variance Analysis, Correlation Analysis, Machine Learning Prediction

1. 引言

保险行业的发展伴随着保险欺诈行为的上升。保险是一种商业行为，从经济、法律、社会和风险管理角度来看，保险起到了分摊意外损失、提供经济保障和管理风险的作用。然而，保险欺诈严重扰乱了正常的保险经营，损害了保险公司和消费者的权益，阻碍了行业的健康发展。保险欺诈是指以骗取保险金为目的，通过虚构保险标的、制造保险事故等手段非法获取赔偿的行为。保险欺诈的产生有多种原因：一是保险公司内部控制薄弱，缺乏风险识别和管理机制；二是法律法规不完善，对保险欺诈行为的处罚力度不足；三是社会信用体系不健全，对保险欺诈行为的宽容态度助长了不法分子的动机[1]。

保险欺诈通常表现为虚假材料、故意制造事故、虚高损失等形式。提前通过参保人数据建立反欺诈模型分析和预测这些欺诈行为有助于减少损失。保险欺诈影响了保险市场的公平性和透明度，破坏了保险行业的信任基础。因此，有效防范和打击保险欺诈不仅能保护保险公司和消费者的权益，还能促进保险行业的健康发展。

2. 数据预处理与分析

2.1. 数据预处理

保险反欺诈预测数据训练集有 700 行数据，测试集有 300 行数据，都有 38 个字段，其中自变量特征 37 个，因变量特征为是否保险欺诈(fraud)。处理数据时将训练集和特征集合并在一起进行处理。对数据的重复值进行检查，发现无重复值情况。同时检查数据缺失值情况，发现数据也无缺失值。绘制因变量(fraud)正负样本比例，如图 1 所示。

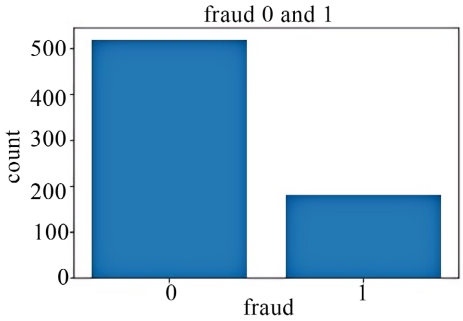

Figure 1. Scale map of the positive and negative samples of the dependent variable
图 1. 因变量正负样本比例图

从图中可以发现，正样本即没有保险欺诈数据有 500 条，负样本有保险欺诈数据有 200 条，正负样本之比为 5:2，故不需要进行数据的不均衡处理。

2.1.1. 异常值处理

进一步，对数据的数值性特征进行异常值检验，首先绘制箱线图观察是否存在异常值情况。绘制不同特征下的箱线图如图 2 所示。

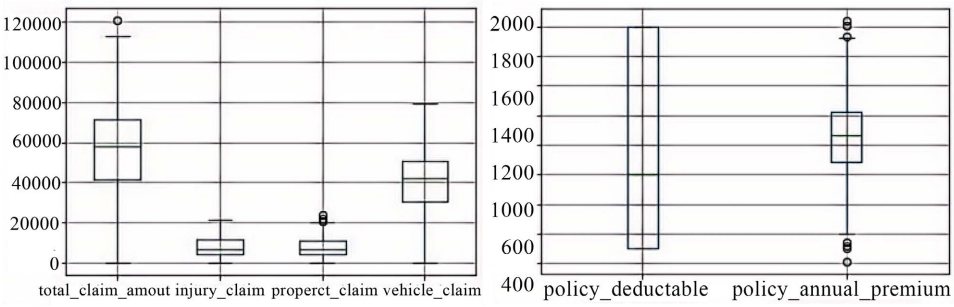

Figure 2. Characteristic boxplot
图 2. 特征箱线图

从箱线图中可以发现特征“policy_annual_premium”，“total_claim_amount”，“property_claim”可能存在异常值情况，进一步使用 3σ 原则对上述三个特征的异常值进行检验发现特征“policy_annual_premium”存在 5 个异常值，“property_claim”存在 1 一个异常值。考虑到异常值总共只有 6 行，对比于总数据量 1000 行，所占比例很小，故直接将异常值删除。

2.1.2. 日期值处理

数据中存在“policy_bind_date”和“incident_date”两个日期型特征。考虑到两个特征都精确到日，相比于日期本身，日期的差值即时间间隔更有价值。因此查出两个特征下数据的最小日期，再将两列特征数据与最小日期做差值转化为数值型特征。同时考虑到“policy_bind_date”和“incident_date”日期之差为发生事故时间和参保时间之间的时间间隔对因变量是否保险欺诈可能具有关联性，因此做两列特征的差值生成一列新的特征，将该特征用于后续模型建立。

2.1.3. 离散特征编码

数据中存在 17 个离散类别特征，为了后续建立预测模型，需对离散特征进行数值化编码。离散型数据编码的原则是没有大小意义的特征使用 One Hot 编码，有大小意义的对不同类别使用数值映射。同时考虑到有些特征分类类别过多，使用 One Hot 编码会产生过多特征列，因此选择进行数值映射。

基于上述考虑，对特征“incident_state”，“insured_sex”，“policy_state”进行 One Hot 编码，对剩余的离散类别特征使用 python 中的 OrdinalEncoder()函数，将分类特征映射为数字编码，并且还会保留原来的顺序。通过离散特征编码将字符串型特征转化为离散数值型特征，便于后续预测模型的建立。

2.2. 特征筛选

2.2.1. 单特征变量分析

方差是衡量一个变量的离散程度(即数据偏离平均值的程度大小)；变量的方差越大，我们就可以认为它的离散程度越大，也就意味着这个变量对模型的贡献和作用会更明显，因此要保留方差较大的变量，反之，要剔除掉无意义的特征。我们在这里通过特征本身的方差来筛选特征，得到的各连续性特征的方差都大于 0.5，我们进一步将阈值设置为 1 后，仅筛选出了 bodily_injuries [2]。但我们目前不知道该特征是噪声还是有效特征，因此需要进一步地将其投入模型中验证该特征的存在对模型结果的影响。此处，我们简单建了一个 KNN 模型，得到的精度如下表。所以还是选择保留 bodily_injuries 特征[3]。是否考虑特征 bodily_injuries 下的精度如表 1 所示。

Table 1. Precision comparison
表 1. 精度比较

是否考虑了 bodily_injuries	精度
是	0.691
否	0.687

2.2.2. 相关性分析

为了计算特征之间的相关性，及其正负影响关系，计算特征间的相关系数，观察哪些特征具有和是否保险欺骗具有较高的相关性。从相关系数的结果可以发现 age 与 customer_months 之间，total_claim_amount、injury_claim、property_claim、vehicle_claim 相关性极高。进一步观察上述特征的具体相关关系，发现 age 与 customer_months、total_claim_amount 与 vehicle_claim 几乎呈现出完全的线性关系。而其中 vehicle_claim 相较 total_claim_amount 而言与 fraud 的相关性更小， customer_months 相较 age 而言相关性更小，故删去 customer_months 及 vehicle_claim。

3. 模型预测分析

3.1. 建立模型

3.1.1. 逻辑回归

Table 2. Confusion matrix table (Where 0 is not fraudulent and 1 is fraudulent)
表 2. 混淆矩阵表(其中 0 为未欺诈，1 存在欺诈)

预测真实	无保险诈骗	保险诈骗
无保险诈骗	96	42
保险诈骗	1	0

建立逻辑回归模型，对处理后的数据进行分析。以数据中保险欺诈特征作为因变量，以其是否欺诈作为自变量[4] [5]。将整体样本按照 8:2 的比例拆分成训练集和测试集。接下来建立模型并进行评估。得出模型的准确度为 0.6906474820143885，混淆矩阵的结果，如表 2 所示。

结果显示，真实数据为“0”的样本有 96 + 42 = 138 个，真实数据为“1”的样本有 1 个，预测结果为“0”的有 97 个，预测结果为“1”的有 42 个，从数据上看，“0”的预测正确的个数比“1”的个数多。由于“1”的样本量只有一个，导致“1”的预测准度很低。原因为拆分训练集和测试集时，模型自动把“1”的样本量减少了，重新尝试多次但“1”的样本量仍然很低。接下来是模型算出的预测精确度、召回率、F1 值，如下表 3 所示。

Table 3. Logical regression model evaluation
表 3. 逻辑回归模型评估

	准确率	召回率	f1 分数	支持度
无保险诈骗	0.99	0.70	0.82	138
保险诈骗	0.00	0.00	0.00	1
整体	0.49	0.35	0.41	139

表中结果显示，“0”的预测精确度为 99%，“1”的预测精确度为 0。由于拆分训练集和测试集时，模型自动把“1”的样本量减少了，得到了这样的结果。由此可以看出模型整体的预测结果的正确率和可信度比较差。但单独对于“0”的预测准确度非常高。接下来绘制 ROC 曲线，如图 3 所示。

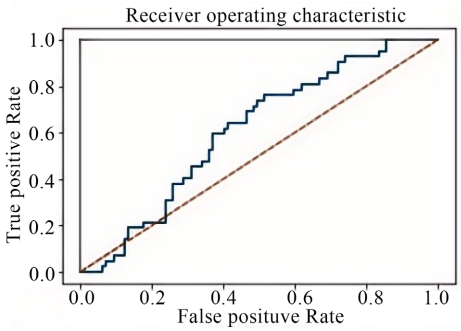

Figure 3. The ROC curve
图 3. ROC 曲线

由于拆分数据不当，导致了一系列的问题，ROC 曲线也十分不理想。综上所述，不适合用逻辑回归模型来预测是否欺诈。

3.1.2. KNN

在建立模型前，先选择一组二维特征空间，并绘制出该特征与是否保险欺诈二维空间下的分散聚集情况，如图 4 所示。

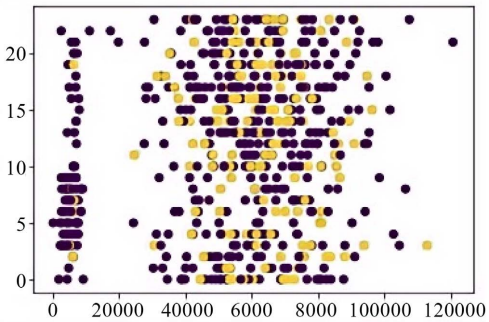

Figure 4. The frmean distribution in the two-dimensional feature space
图 4. 二维特征空间中 fraud 分布

可以明显查看到，0 和 1 在这组二维数据中分布十分散乱，甚至距离非常近，尝试多组特征空间，皆有 0、1 散乱、近距离等特点。因此做好该模型性能不好甚至差的准备。

建立 KNN 模型前，首先将模型按照 8：2 的比例拆分成训练集和测试集。此处为了保证样本的公平性，引入 for 循环，将其设置为多次随机拆分。由于每次拆分样本的随机性都不一样，得不到一个更加中肯的评价结果，故在 for 循环中引进随机种子，以保证打乱样本顺序的一致性，进而增强评价的公平性。样本顺序保持一致后，保证了样本的公平性，因此样本拆分比例如 2：8，1：9 两者采用的样本顺序是一致的不再是随机的了，可以横向比对[6] [7]。

KNN 的参数只有一个决定 k 值的 n_neighbors 参数，即要选取最邻近样本实例的个数。k 值取奇数，防止因为平票而无法分类的情况且一般情况下 k 值的选取应小于样本数量的平方根。所以我们选取 $k \in [1, 23]$，建立一个步长为 2 的 for 循环，横向对比各 k 值下模型的得分。将各 k 值下模型的得分可视化，得到结构如图 5，图 6 所示。

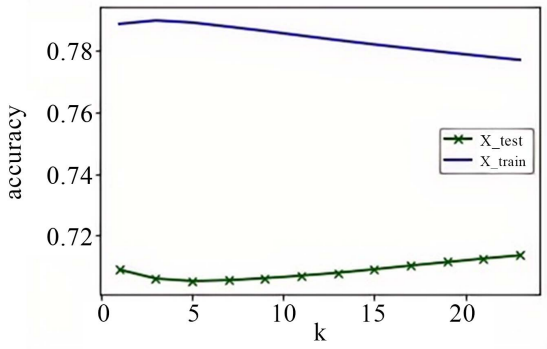

Figure 5. Model score
图 5. 模型得分

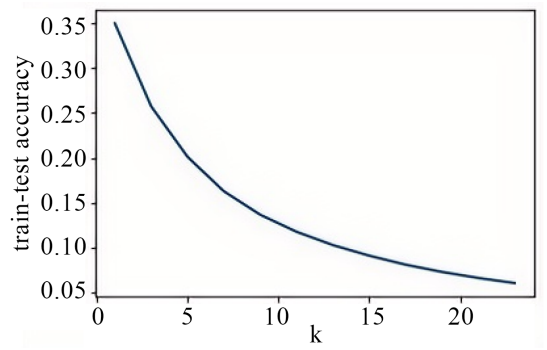

Figure 6. The difference between the scores of the training set and the test set
图 6. 训练集与测试集得分之差

Table 4. Confusion matrix
表 4. 混淆矩阵

		预测结果	
		无保险诈骗	保险诈骗
实际结果	无保险诈骗	87	28
	保险诈骗	17	7

如图 5，k 取 23 模型得分为最佳，正确率可达到 0.744。但就测试集预测结果来看，对 1 预测的正确度极低。考虑到该建模建立的目的是为了对保险欺诈达到一个预测，故最终还是选择对 1 预测正确率最高的 k 为 3 的模型。根据该模型预测结果形成的混淆矩阵如表 4 所示。

可以观察出模型的测试结果很不好，对测试集中无保险欺诈的预测结果准确性低，虽然 115 个无保险诈骗样本预测出 87 个无保险诈骗样本，但 24 个造假样本中预测正确的只有 7 个样本，模型在造假样本的预测结果上准确率较低。进一步得到模型测试的各个评估值如下表 5 所示。

Table 5. Prediction evaluation of the KNN model
表 5. KNN 模型的预测评价

类别	正确率	精确率	召回率	F 值	支持度
无诈骗样本	0.88	0.84	0.76	0.79	115
诈骗样本	0.2	0.2	0.29	0.24	24
总体	0.68	0.73	0.68	0.7	139

从表中可以看出 KNN 模型对无诈骗样本的各个评估指标都很差，由此可以看出模型的预测结果的正确率和可性度比较差。绘制 KNN 模型的 ROC 曲线图，如图 7 所示。

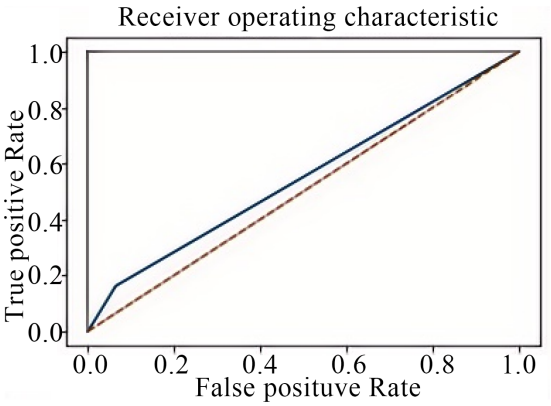

Figure 7. The ROC curve of the LightGBM model
图 7. LightGBM 模型 ROC 曲线

从图 7 中可见 ROC 曲线距离纵轴远，AUC 为 0.51，模型性能并不好。

3.1.3. 随机森林

建立随机森林模型对 994 个保险样本进行了分析判断。以第三章所选保险欺诈特征作为因变量，以是否欺诈作为自变量。将整体样本按照 8:2 的比例拆分成训练集和测试集[8]。

Table 6. The confusion matrix of the random forest model
表 6. 随机森林模型的混淆矩阵

		预测结果	
		不造假	造假
实际结果	不造假	96	4
	造假	37	2

设置参数“max_depth” = 10,“n_estimators”=1000,“min_samples_split” = 2,“min_samples_leaf” = 1,“random_state” = 1000，最后在拟合模型后将预测值存储到“rf_yhat”中。将调参后的模型应用于数据集上，使用训练集进行模型训练。将模型应用于测试集上，得到测试集结果用混淆矩阵表示如表 6 所示。

从表中可以观察出模型的测试结果一般，对测试集中不造假样本的预测结果准确性很高，100 个不造假样本中预测出 96 个不造假样本，只有 4 个不造假样本预测错误。39 个造假样本中预测正确 2 个样本，在造假样本的预测结果上准确率较低。由此可以看出模型的预测结果的正确率和可性度比较差。

通过混淆矩阵的测试结果我们可以得到模型测试的各个评估值如下表 7 所示。

Table 7. Random forest assessment values
表 7. 随机森林评估值

类别	正确率	精确率	召回率	F 值	支持度
无诈骗样本	0.96	0.96	0.72	0.82	133
诈骗样本	0.05	0.05	0.33	0.09	6
总体	0.71	0.71	0.71	0.79	139

由混淆矩阵得到支持向量机模型的评估指标，模型性能一般，具有良好的可信度。

绘制 Random forest 的 ROC 曲线图，如图 8 所示。

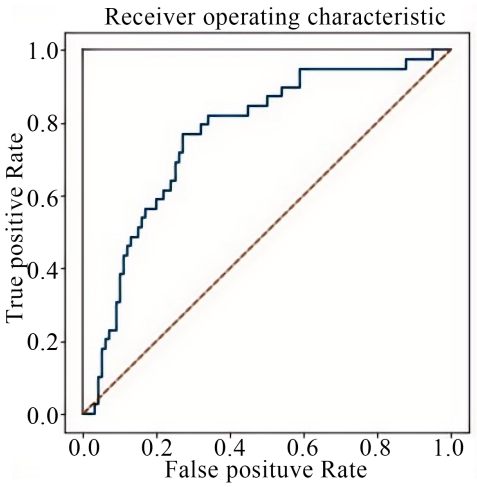

Figure 8. Random forest ROC curve
图 8. 随机森林 ROC 曲线

从图中可见 ROC 曲线较为靠近纵轴，AUC 为 0.767，模型性能良好。

3.1.4. LightGBM

在 695 条保险反诈骗数据的基础上建立 LightGBM 模型，首先将整体样本按照 8:2 的比例拆分成训练集和测试集。LightGBM 模型可调整的参数主要有六个，分别为 objectiv：模型应用的类型；num_leaves 调节决策树的复杂程度；reg_alpha：L1 正则化系数；reg_lambda：L2 正则化系数；max_depth：树的深度；min_date_in_leaf：取值取决于训练数据的样本个数和 num_leaves。

对于特征 reg_alpha 和 reg_lambda 其值从(0.05, 1)以 0.05 为步长进行遍历，采用 score 函数得到的评

分结果作为参数值好坏的评估标准。绘制出其学习曲线，观察出 reg_alpha = 0.1，reg_lambda = 0.25 时模型的评分最高，达到 0.865。再对 max_depth 从(−5, 5)以步长为 1 进行遍历，得到最优参数为−5。进一步在对 leaning_rate 从(0.05, 1)以 0.05 的步长进行遍历，得到最优参数为 0.85。最后对 num_leaves 进行调参，得到最优参数为 31。

调参后，模型的评分从默认参数下的 0.71 提升到 0.87，模型性能有显著提升。基于上述学习曲线，结合在所选范围内的最佳值，得到 LightGBM 模型选择的拟合参数如表 8 所示。

Table 8. LightGBM hyperparameters
表 8. LightGBM 超参数

参数名称	参数取值	参数名称	参数取值
objective	“binary”	reg_lambda	0.25
num_leaves	31	max_depth	−1
reg_alpha	0.25	min_date_in_leaf	3

其中，首先选择较高的学习率，在 0.1 附近，能够加快收敛速度。调节决策树基本参数 max_depth 和 num_leaves，是提高精确度的重要参数。min_date_in 进行正则化调参，防止过度拟合。将调参后的模型应用于数据集上，使用训练集进行模型训练。将模型应用于测试集上，得到测试集结果用混淆矩阵表 9 所示：

Table 9. The LightGBM confusion matrix
表 9. LightGBM 混淆矩阵

		预测结果	
		无保险诈骗	保险诈骗
实际结果	无保险诈骗	95	9
	保险诈骗	8	27

从表中可以观察出模型的测试结果很好，对测试集中无保险欺诈的预测结果准确性很高，104 个无保险诈骗样本预测出 95 个无保险诈骗样本，只有九个无诈骗样本预测错误。35 个造假样本中预测正确 27 个样本，模型在造假样本的预测结果上准确率较高。虽然造假样本的预测准确率低于不造假样本，但差别不明显，整体正确率较高。由此可以看出模型的预测结果具有很高的正确率和可性度。通过混淆矩阵的测试结果我们可以得到模型测试的各个评估值如表 10 所示。

Table 10. LightGBM assessed values
表 10. LightGBM 评估值

类别	正确率	精确率	召回率	F 值	支持度
无诈骗样本	0.92	0.91	0.92	0.93	104
诈骗样本	0.75	0.77	0.76	0.75	35
总体	0.88	0.84	0.85	0.88	139

从表中可以看出 LightGBM 模型对无诈骗样本的各个评估指标上都有高于 0.9 的评分，但模型在诈

骗样本的预测评估指标都低于无诈骗样本为 0.75 上下。整体模型预测正确率达到 0.88，说明整体上 LightGBM 模型在数据集上表现较好，具有较高的精确度，且没有明显的数据过拟合现象。绘制 LightGBM 模型的 ROC 曲线图，如图 9 所示。

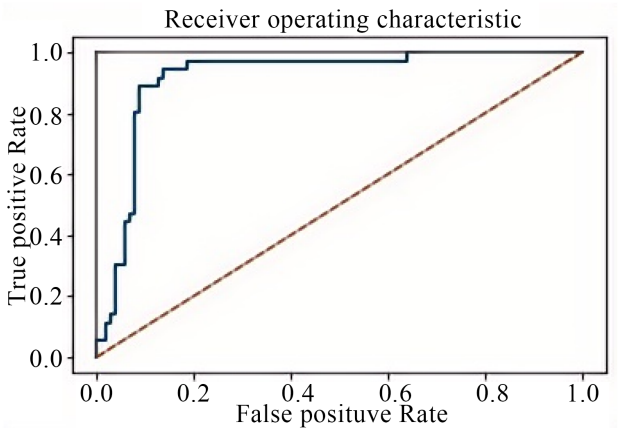

Figure 9. The ROC curve of the LightGBM model
图 9. LightGBM 模型 ROC 曲线

从图中可见 ROC 曲线接近于垂直纵轴，AUC 值为 0.918，说明模型整体性能处于一个较好的水平，模型具有一定的实际意义。最后再得到 LightGBM 模型特征重要性前十的特征，如下图 10 所示。

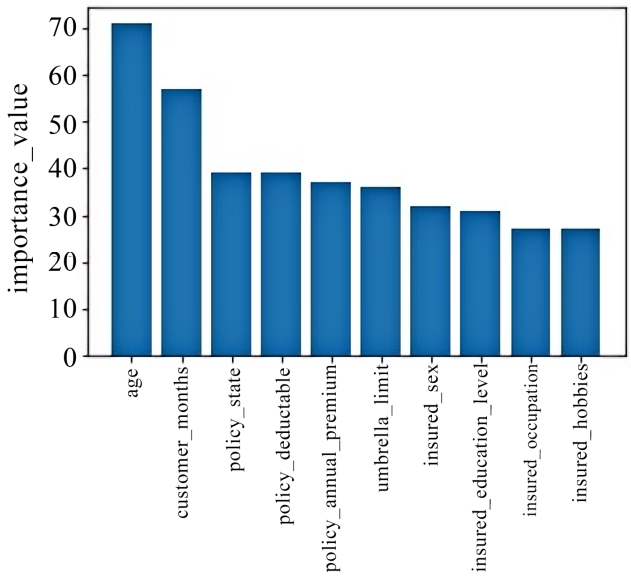

Figure 10. LightGBM model importance ranking
图 10. LightGBM 模型重要性排序

模型特征重要性前十的特征分别为：改为年龄、客户月数、投保所在地区、保险扣除额、每年的保费、保费责任上限、被保人性别、被保人教育水平、被保人职业、被保人兴趣爱好。

3.2. 模型对比分析

对上述各模型的测试集预测结果的评估指标进行对比分析，如表 11 所示。

Table 11. Comparison of model evaluation indicators
表 11. 模型评估指标对比

	正确率	精确率	召回率	F 值	支持度
LightGBM	0.88	0.84	0.85	0.88	139
随机森林	0.71	0.71	0.71	0.79	139
逻辑回归	0.69	0.98	0.69	0.81	139
KNN	0.65	0.73	0.68	0.7	139

从表 11 中可以看到 LightGBM 模型的预测准确率相较于其他模型来说较高，模型整体性能更优。但存在对非保险诈骗的预测精度高于保险诈骗预测的精度的情况。随机森林模型、knn 模型、logistics 回归不能很好地学习训练数据对测试数据类别进行划分，得到的测试集预测结果几乎都为不欺诈，模型性能较差。

4. 总结与讨论

在整个实践过程中，经过数据预处理，我们得到了适合建立模型的数据，接下来进行特征筛选，筛选掉 customer_months、vehicle_claim。再进行了数据可视化分析，了解到保险欺诈的事故大都为单车或三车的追尾事故，损失较大。被保人年龄大都在 30~40 之间，善于思考且具有强目标性，从事管理岗居多。保险欺诈多发生于白天，具有低买高保的特点。接着分别建立逻辑回归模型、KNN 模型、随机森林模型、LightGBM 模型，其中逻辑回归和 KNN 模型、随机森林预测效果不佳；LightGBM 模型预测效果最好。

对于保险公司而言，用模型预测是否会发生保险欺诈是方便高效的。但对于社会整体，我们需要从多个方面入手来防范保险欺诈。这是一项系统工程，需要有关方面共同努力，形成共识，多部门密切配合。在一般认知方面，公众认为基本医疗保险领域的欺诈行为最为多见；社会保险欺诈在待遇支付与领取阶段中更为集中；主要的欺诈实施主体为自然人和社会保险服务机构及其工作人员。在欺诈行为可责罚性认知方面，公众认为对于社会保险欺诈应当予以惩罚，且普遍不认同轻罚[9]。

对于以上现状，保险欺诈的应对策略应把握以下几个方面[10]：

(一) 完善相关法律法规，加大欺诈处罚力度：进一步明确界定保险欺诈行为，这样有利于法律法规的准确适用。

(二) 拓宽宣传教育范围，高度警惕欺诈行为：保险公司应加强在客户中以灵活、多样的形式开展普法教育，使消费者明白保险诈骗行为属于违法犯罪行为，情节严重的将受到法律的制裁。

(三) 强化公司内部管理，提高风险防范能力：重视风险防范，转变发展方式。完善产品管理，从源头防范风险。规范承保流程，强化核保管控。完善理赔制度，强化定损、核损、核赔等关键环节管控。

(四) 加强同业合作，建立反欺诈信息平台：由于保险行业同业间长期缺乏有效的沟通与联系，信息的不对称给骗赔者带来了可乘之机。面对这种情况各保险公司之间应建立信息交换网络，建立黑名单制度，将实施保险欺诈者在行业内甚至社会范围内公布，防止其继续进行欺诈，并在社会上形成一种反保险欺诈的威慑力。

因此，利用大数据手段进行防范预测与社会方面加强管控双双结合，离创建更好的商业保险环境就更进一步。除了本文所用的模型，还有其他有效手段可以进行有效防范，例如区块链技术、人工智能技术等等。

参考文献

[1] 王素芬. 社会保险反欺诈何以可能: 基于公众认知的策略选择[J]. 深圳大学学报(人文社会科学版), 2021, 38(2): 84-94.

[2] 熊珈. 新科技背景下我国商业健康保险反欺诈路径研究[J]. 保险职业学院学报, 2022, 36(3): 61-66.

[3] 谢廷廷. 融入文本信息的 P2P 网贷平台跑路预测模型研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2019.

[4] 翟晓风. 基于 XGBoost 算法的中国上市公司财务舞弊预测模型研究[D]: [硕士学位论文]. 北京: 中国财政科学研究院, 2022.

[5] 赵沛. 二分类 Logistic 回归模型对上市公司财务状况的预测效度研究[D]: [硕士学位论文]. 南宁: 广西大学, 2015

[6] 邵亚洁. 基于复合 CatBoost 模型的 P2P 网贷违约分类预测[D]: [硕士学位论文]. 兰州: 兰州大学, 2019.

[7] 陶能发. A 股上市公司财务造假预测模型研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2020.

[8] 钱苹, 罗玫. 中国上市公司财务造假预测模型[J]. 会计研究, 2015(7): 18-25, 96.

[9] 胡嘉麟. 基于 LightGBM 模型的车辆保险购买兴趣预测研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2021.

[10] 沙靖岚. 基于 LightGBM 与 XGBoost 算法的 P2P 网络借贷违约预测模型的比较研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2017.