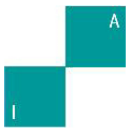
Inteligencia Artificial, 28(75) (2025), 140-169
doi: 10.4114/intartif.vol28iss75pp140-169



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

A Machine Learning and Explainable Artificial Intelligence Approach for Insurance Fraud Classification

Zaid Khan, Diana Olivia
Department of Information and Communication Technology,
Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, INDIA
Email: kzaidnba@gmail.com, diana.olivia@manipal.edu

Sucharita Shetty
Department of Computer Science,
Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, INDIA
Email: sucha.shetty@manipal.edu

Abstract This study tackles the pressing issue of fraud in the vehicle insurance market by introducing a comprehensive framework that integrates advanced detection models with Explainable Artificial Intelligence (XAI) and heterogeneous classifiers. The inclusion of XAI is particularly significant, as it enhances the interpretability and transparency of machine learning algorithms, which is crucial for maintaining the integrity and trustworthiness of insurance operations. Our methodology employs three distinct XAI techniques: Shapeley Additive Values (SHAP), Explain Like I'm 5 (ELI5), QLattice and Local Interpretable Model-Agnostic Explanations (LIME) to elucidate the decision-making process of machine learning models. This approach not only ensures model interpretability but also identifies key factors influencing fraud detection, including vehicle age, base policy, fault, deductible, and policyholder age. The standout contribution of our research is the development and validation of a multi-stack machine learning model that achieves an exceptional accuracy rate of 96 %, significantly outperforming traditional classifiers. This high level of accuracy, combined with the interpretability provided by XAI, underscores the potential of our framework to revolutionize fraud detection practices in the vehicle insurance sector. By offering a robust, accurate, and interpretable solution to fraud detection, this study makes a meaningful contribution to the field. It provides valuable insights and tools for insurance providers aiming to enhance their fraud detection capabilities, setting a new benchmark for the development of advanced, reliable systems in the industry.

Keywords: Insurance, Financial decision making, Predictive models, Fraud Detection, Machine Learning, Explainable Artificial Intelligence, Eli5, LIME, SHAP, QLattice.

1. Introduction

Insurance fraud is a pervasive and growing problem that significantly impacts both policyholders and insurance companies in all segments of the industry. The detection of fraud has garnered increasing attention in recent years, driven by the evolving nature of fraudulent activities, which continue to adapt to

ISSN: 1137-3601 (print), 1988-3064 (on-line)

new technologies and economic conditions. Although it is challenging to quantify the total financial losses caused by insurance fraud, the consequences are substantial, often leading to the exploitation of a profit organizationâs system without necessarily resulting in direct legal consequences. Although there is no universally accepted definition of financial fraud, it is generally described in the literature as a deliberate act that is contrary to law, rule, or policy with the intention of obtaining unauthorized financial benefit.

Given the severity of the economic implications, Insurance Fraud Detection (IFD) is critical to mitigate these impacts. IFD involves distinguishing between fraudulent and legitimate claims, thereby enabling decision-makers to develop appropriate policies to reduce the prevalence of fraud. Despite the potential of data mining to address some of these issues by leveraging large client databases, the fraud detection process faces significant technical challenges. One of the most pressing challenges is the issue of imbalanced datasetsâwhere fraudulent cases are vastly outnumbered by legitimate onesâresulting in models that perform well on legitimate cases but poorly on fraudulent ones. To address this imbalance, we employed the Adaptive Synthetic (ADASYN) method to synthetically balance the dataset.

This paper applies the stacking ensemble technique to the domain of credit risk assessment, aiming to enhance the accuracy and interpretability of predictive models in financial institutions. We aim to develop a robust and accurate pipeline for automating fraud detection. Using an open-source Kaggle dataset containing vehicle insurance data from US companies, we utilized ensemble models and implemented feature selection techniques such as the Sine Cosine Algorithm, Cuckoo Search, and Grey Wolf Optimizer prior to model training. We constructed several machine learning pipelines, incorporating classifiers and data-balancing techniques, to identify the top-performing architecture for fraud detection. Additionally, we employed a layer of explainable AI (XAI) tools to interpret the best-performing pipeline.
In this study, we leverage the concept of Heterogeneous Artificial Intelligence (Heterogeneous AI), which refers to the use of diverse types of AI models and algorithms working together to solve complex problems. Heterogeneous AI involves combining models that differ in architecture, learning mechanisms, and feature extraction methods, thereby capturing a wide range of patterns and relationships in the data. This approach contrasts with homogeneous AI systems, which rely on a single type of model or algorithm.
The use of Heterogeneous AI in our solution is crucial because it allows us to harness the strengths of different models, such as decision trees, support vector machines, and ensemble models, to improve the overall performance and robustness of the system. By integrating these diverse models into a stacked ensemble framework, we enhance the model's ability to detect complex patterns of fraud in insurance claims, leading to more accurate and reliable predictions. Moreover, the inclusion of Explainable AI (XAI) techniques ensures that the outputs of this heterogeneous system are not only accurate but also interpretable, allowing stakeholders to understand and trust the decision-making process.

The contributions of this study are as follows:

- Evaluation of Machine Learning Models: Conducted a comprehensive evaluation of various machine learning models specifically for insurance fraud detection, highlighting their effectiveness in real-world applications.

- Comparison of Data-Balancing Techniques: Performed a detailed comparison of advanced data-balancing methods, including ADASYN and Synthetic Minority Oversampling Technique (SMOTE), to address class imbalance in the dataset.

- Feature Selection Analysis: Analyzed feature selection techniques such as Cuckoo Search Optimization, Grey Wolf Optimization, and Sine Cosine Optimization, identifying the most effective methods for enhancing model performance.

- Assessment of Classifiers: Examined the performance of nine distinct classifiers, including advanced models like LSTM (Long Short-Term Memory), Stacked Ensemble, Extreme Boosting, Light Gradient Boosting, Adaptive Boosting, Categorical Boosting, as well as traditional classifiers like Decision Tree, Random Forest, K-Nearest Neighbors, and Logistic Regression.

- Development of a Novel Ensemble Model: Implemented a unique stacked multi-level ensemble model designed to predict the veracity of insurance claims, rigorously evaluated using K-Fold Cross-Validation to ensure robustness and reliability.

- Incorporation of Anomaly Detection: Integrated anomaly detection techniques to focus the model on identifying outlier instances that are highly likely to be fraudulent, thereby improving overall accuracy.
- Application of XAI Tools: Applied Explainable AI (XAI) toolsâsuch as Explain Like I am 5, QLattice, Shapley Additive Explanations, Local Interpretable Model-agnostic Explanations, and Feature Importanceâto validate and interpret the model's attributes. This dataset is underexplored, with no prior research utilizing XAI techniques, making this application particularly novel and significant.

The structure of this article is as follows: Section two provides a review of the relevant literature. Section three details the materials and methods used in this study. Section four presents the results and discussion, and Section five concludes the article.

2. Literature Review

Nalluri et al. [1] conducted an experiment on healthcare insurance fraud using Multi-Layer Perceptron, which outperformed DT (Decision Tree), RF(Random Forest), and SVM (Support Vector Machines) machine learning models with the F1 score of 81.44 %. The relevant 19 features were extracted using DT. Other personal information, such as personal facial traits, voice, and other biological information, as well as social network speeches and online transaction records, are also available. Consideration of these personal data, together with the use of AI techniques or deep learning methods to discover the best classifier, could improve the model's accuracy. Integrating personal data into fraud detection is of great importance. On the contrary, our contribution is that our approach uses advanced feature engineering to effectively incorporate a broader range of data types, enhancing model accuracy through the use of the Cuckoo Search algorithm, Grey Wolf optimization, and Sine Cosine algorithm. Debene et al. [2] considered the use of isolation forests and XGBoost for fraud detection. Neural Network (NN) and clustering-based detection were also compared. Their findings indicate that unsupervised algorithms outperform supervised algorithms. Our use of SHAP XAI analysis reveals that for identifying claim fraud, supervised and unsupervised learning techniques stress different aspects. Insurance fraud can be successfully detected via unsupervised learning, particularly in isolated forests. Despite the fact that there are few labeled fraud incidents, supervised learning performs well. Surprisingly, both unsupervised and supervised learning recognize new fraudulent claims based on varying input data. Conversely, Our model uses a similar XAI approach to ensure that both supervised and unsupervised models not only complement each other but also provide actionable insights, thus improving operational efficiency in fraud detection.
Healthcare insurance fraud has been detected by Mohanta et al. [3] using the SMOTE oversampling strategy and the Sequential Forward Selection (SFS) method. They employed a Gradient Boosting Machine (GBM), Linear Discriminant Analysis (LDA), Bagging classifier, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and stacking meta-estimator for classification. The best accuracy on the real-world healthcare insurance fraud dataset is obtained by applying feature selection to the Stacking classifier on a balanced dataset, yielding a 97.19 % accuracy rate. Work can be improved by using optimization techniques such as Cuckoo Search optimization and Sine Cosine Optimization to increase performance metrics by applying them to select relevant aspects of the dataset and also by minimizing the search space, which we used in our paper. In order to detect fraud behavior in vehicle insurance cases using real-world data, Jiaxi et al. [4] developed a vehicle Insurance Multi-modal Learning (AIML) framework and utilized computer vision and natural language processing techniques for knowledge-based algorithms. AIML comprises a visual data processing framework and an algorithm for processing car insurance data called the Semi-vehicle Feature Engineer (SAFE), which is self-designed. As opposed to models that solely use structural data, the results show that AIML greatly improves model performance in detecting fraud activity. There is a great need for the integration of multimodal data sources into a unified framework to ensure comprehensive fraud detection across various data types. on the other hand, our algorithm not only incorporates multimodal data but also uses machine learning to seamlessly integrate these diverse data streams, significantly enhancing detection capabilities Aarati et al. [6] have developed Machine learning algorithms like SVM, Logistic regression, Decision tree, and KNN for insurance fraud detection in health insurance datasets. These standard models are not always scalable. Conversely, our approach uses advanced ensemble techniques and stacked models to address these complexities, ensuring

higher scalability and better performance.
John et al. [7] have used Ensemble algorithms such as XGBoost, LightGBM, Extremely Randomized Trees (ET), Random Forest, and CatBoost, as well as the linear approach Logistic Regression, for Medicare fraud detection. They employed five ensemble approaches, as well as a Decision Tree, to choose features. The paper's novelty involves the use of an ensemble for feature extraction, which extracted 82 features from the highly imbalanced Big Data datasets. Their result shows that CatBoost and XGBoost performed better. Alternatively, our ensemble approach not only extracts features but also adapts to new and emerging fraud patterns through continuous learning and model updates through dynamic weight adjustment, anomaly detection, and periodic retraining.
With the use of temporal Medicare claims data, the Functional Principal Component Analysis (FPCA) method for examining the trajectory of the temporal covariates, and the distributional FPCA method for deriving features from the empirical probability density curve of the covariates, Shi et al. [8] identified fraud. They incorporated machine learning algorithms, such as Logistic Regression (LR), gradient boosting machine (GBM), Neural Network (NN), and Random Forest (RF), with Cost-Sensitive (CS) and traditional Non-Cost-Sensitive (NCS) methodologies. The best AUC for prediction ability is offered by the GBM model.
Lu et al. [9] offer a model for detecting health insurance fraud based on a layered attention mechanism. The effect of interactions between items in a healthcare context on fraud detection was investigated. In a model termed MHAMFD, these interactions were captured by an Attributed Heterogeneous Information Network (AHIN). To identify acceptable neighbors, several levels of behavioral relationships are used, which take into account the composite semantic information from the interweaving of different relationships and increase the quality of neighbor nodes. Extensive testing with two real datasets revealed that MHAMFD outperformed existing graph representation learning algorithms for fraud detection. Pavitha et al. [15, 16] research work introduces an explainable multistage ensemble model that integrates 1D Convolutional Neural Networks (CNNs) for credit decision-making. The proposed model aims to enhance the trustworthiness of credit evaluations by providing transparent and interpretable results. By leveraging a multistage approach, the model combines multiple 1D CNNs to capture complex patterns in credit data, resulting in more reliable and explainable predictions. Xiaoming et al. [17] have provided a comprehensive review of contemporary methods for assessing consumer credit risk, focusing on the latest advancements in classification algorithms, data characteristics, and learning techniques. Wirot et al. [18] introduce a novel approach for credit scoring that leverages a cost-sensitive neural network ensemble. The method addresses the challenges of imbalanced credit datasets by incorporating cost-sensitive learning into neural network ensembles. The proposed method [20–22] involves cost-sensitive neural network ensembles, which can be computationally intensive and complex to implement. This might limit its practical adoption, especially in resource-constrained environments. et al. Shounak [19] presents an innovative approach to credit decision-making using an explainable multistage ensemble model that incorporates 1D Convolutional Neural Networks (CNNs). The model is designed to enhance the interpretability and reliability of credit assessments by combining multiple 1D CNNs in a multistage framework. The multistage ensemble approach involving multiple 1D CNNs may lead to high computational demands, which could be a barrier to practical implementation
The remaining articles are tabulated in Table 1. Alternatively, our model incorporates similar temporal analyses and interaction data within a more comprehensive machine learning framework, optimizing detection accuracy and efficiency. This is better handled using LSTM which we have applied.

3. Materials and methods

An entire flowchart of our strategy is shown in Figure 1, operating with a specially created pipeline. The experimental steps of this proposed pipeline will be covered one by one in the following subsections.

3.1. Benchmark dataset description

The decision analytics division of an American business (EXL Service) supplied the vehicle insurance fraud dataset in 2020. The collection, which is open source and accessible to the public, comprises auto

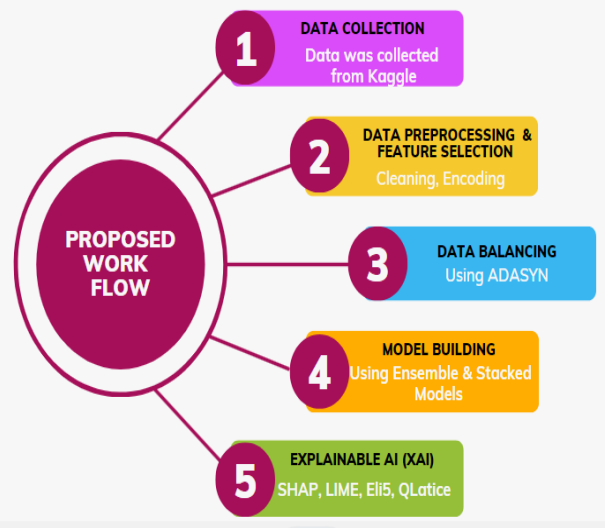


Figura 1: Methodology for Proposed Fraud Detection model

insurance data from US firms. The binary target variable (yes/no), which indicates if fraud has been reported or found, is one of the dataset's 33 total variables.

- **Age of Vehicle**: This describes the numerical difference between the model year and the current year of a vehicle.
- **Accident Area**: Location of the mishap.
- **Marriage Status**: The individual is either widowed, divorced, single, or married.
- **Fault**: The accident's causative fault's name.
- **Policy Type**: It is a document with the organization's goals, plans, and board member guidelines.
- **Sex**: Either a man or a woman.
- **Car Category**: Vehicles carrying passengers are denoted by M, vehicles carrying freight by N, vehicles with two or three wheels by L, and agrarian tractors with trailers by T.
- **Car Price**: The price of the car is displayed.
- **Driver Rating**: Customer experience informs driver ratings.
- **Days Policy Accident**: Duration of the incident in days. The age of the policyholder is shown in
- **Age of Policy Holder**.
- **Police Report Filed**: Reporting an accident to the authorities or not. The number of days that a policy can be claimed is indicated in
- **Days Policy Claim**.
- **Witness Present**: A witness's presence at the scene of the accident.
- **Agent Type**: 'Internal' or 'external' is how the agent type is described. This item **Number of Supplements** lists the additional reparations that were not included in the initial estimate.
- **Base Policy**: This displays the base policy for the insurance.
- **Claim Size**: The cost of claims' liability is mentioned.

- **Month**: The month the mishap happened.
- **Week of Month**: The week that the mishap happened.
- **Day of Week**: The accident's day of the week.
- **Make**: The brand name of the automobile.
- **Day of week Claimed**: The day of the insurance claim during the specified week. The month of the insurance claim is indicated in
- **Month Claimed**. The week of the insurance claim is indicated in
- **Week of Month Claimed**.
- **Rep Number**: The number assigned to the representative.
- **Deductible**: The sum subtracted from the compensation for the claim.
- **Policy Number**: Issued number.
- **Previous Claims**: Historical insurance claim figures.
- **Age**: The individual's age. The claimant's updated address is shown in
- **Address change Claim**.
- **No of Cars**: The claimant's car count.
- **Fraud Found (Target class)**: The dataset indicates whether or not fraud was discovered.

3.2. Data Pre-processing

Removing Irrelevant Dataset Features

The features such as Month, DayOfWeek, DayOfWeekClaimed, WeekOfMonthClaimed, WeekOfMonth, Make, MonthClaimed, PolicyNumber, RepNumber, AddressChange_Claim, and Age were dropped from the car insurance dataset for the following reasons:

Irrelevance Columns like Month, DayOfWeek, MonthClaimed, WeekOfMonthClaimed, and WeekOfMonth were not directly relevant to the specific analysis or modeling goals. The analysis is not focused on time series patterns or the day of the week, so these columns can be safely dropped.

Redundancy 'Age' and 'policyholder age' likely contained similar information, which led to redundancy. In such cases, it's common to retain one column to avoid duplicating information unnecessarily.

Identifiers Columns like 'PolicyNumber' and 'RepNumber' are unique identifiers or reference numbers, which are often not useful for modeling and are safely removed.

Data Quality and Relevance Columns like 'Make' do not provide substantial information for the analysis, so they have been dropped due to data quality issues.

Missing Values

The work checks for missing values in the dataset, which is essential for ensuring data quality and integrity.

Data Encoding

Categorical variables are encoded using one-hot encoding, resulting in a modification of the dataset where each category of a feature becomes a new feature itself.

Algorithm 1 Algorithm for Preprocessing, Balancing, Feature Selection, and Model Training

Data Preprocessing
1. Load the dataset.
2. Drop irrelevant features: Month, DayOfWeek, DayOfWeekClaimed, WeekOfMonthClaimed, WeekOfMonth, Make, MonthClaimed, PolicyNumber, RepNumber, AddressChange Claim, Age.
3. Handle missing values (e.g., imputation or removal).
4. Encode categorical variables using one-hot encoding.
Apply ADASYN for Data Balancing
5. Initialize ADASYN parameters and calculate synthetic samples needed.
6.
for each minority class sample **do**
    Compute and generate synthetic samples based on classification difficulty.
end for
Feature Selection Using Metaheuristic Algorithms
7. Apply metaheuristic algorithms: Sine Cosine Algorithm, Cuckoo Search, and Grey Wolf Optimizer to explore the feature space.
8.
for each algorithm: **do**
Run the algorithm for 100 iterations to explore the feature space thoroughly and identify relevant features.
Run the algorithm multiple times (10 times) to ensure consistent feature selection across different runs.
Track feature selection frequency.
Select consistent features.
end for
9. Combine selected features from all algorithms
Model Training and Evaluation
10. Define models: Random Forest, K Neighbors, Decision Tree, Logistic Regression, AdaBoost, CatBoost, LGBM, XGBoost, LSTM, and Stacked Ensemble.
11.
for each model: **do**
    Train on the data, apply k-fold cross-validation, and evaluate using metrics.
end for
Combining Predictions Using Logistic Regression
12. Train stacking models: (1) Random Forest, K Neighbors, Decision Tree; (2) AdaBoost, CatBoost, LGBM, XGBoost.
13. Generate predictions from both models.
14. Train a logistic regression model using these predictions as features.
15. For new data, generate ensemble predictions and use logistic regression for the final prediction.
16. Use k-fold cross-validation to ensure robustness.

Cuadro 1: Related research which detects Insurance Frauds using machine learning.

Ref.	Sources	Data Size	Total Features	Model used	Accuracy	Sensitivity	Specificity	AUC	Novelty/Findings	Limitations
[10]	decision analytics department of an American firm (EXL Service)	2434	33	logistic regression, support vector machine, and naÃ¯ve Bayes	0.9400	0.0070	0.997		Uncovers the most influential feature through the Boruta algorithm	Graph learning-based fraud detection methods neglect substantial structural homogeneity and fail to aggregate the features of two nodes that are structurally alike but far away from each other
[11]	244 companies in the Dhaka stock exchange report		11	ANN classifier, Averaged neural network Support vector machines, Support vector machines with linear kernel Naive Bayes classifier , Naive Bayes K-nearest neighbor classifier, K-nearest neighbors Ensemble classifier, Bagged CART	0.88	0.88	0.94	0.96		Applied default parameters , Not applying parameter optimization
[12]	Transactions made by credit cards by European cardholders: 284,807 transactions and Car holders from Brazilian bank: 374,823 transaction			LSTM, GRU, and ensemble model		0.74		0.8702	Developed an innovative system for voting based on artificial artificial neural networks and an ensemble framework based on sequential data modeling.	Accuracy and AUC score of this technique were greater, but only linearly separable patterns were learned.
[13]	Largest commercial banks in China.	153,685 transaction records	18		0.9811	0.583			presented a novel Presented a novel feature extraction system with an architecture for deep learning for fraud with credit cards detection.	Model has high calculating cost
[14]	Australian and England credit datasets	710	14	a novel dual-weighted fuzzy proximal support vector machine, model hybridizing fuzzy set theory, and proximal support vector machine, is proposed for credit risk analysis. linear regression, logistics regression, back-propagation neural network, standard support vector machine, proximal SVM, and dual-weighted fuzzy PSVM				0.9272	proposed FPSVM outperforms other SVM models	
[23]	loan listings from Renrendai	8650 samples	29	RF, SVM, LR and k-NN	0.9565			0.9662	ensemble-based machine learning methods, such as the RF and XGB,	Neglected the importance of feature engineering
[24]	Eikon database for small and medium-sized French firms		36	XGBOOST with most important features	90.62	0.935	0.868	0.964		Furthermore, it would be fascinating for future study to use interpretable ML models to anticipate failures in other industries, such as banks or financial institutions. Furthermore, future research should look into using interpretable ML models to predict

3.3. Data Balancing

Insurance fraud detection is a challenging task, compounded by the issue of class imbalance in the dataset. Typically, instances of fraud are considerably fewer compared to non-fraudulent cases. This imbalance leads to models that are biased towards predicting non-fraudulent cases, often overlooking the minority class, which, in this scenario, is the fraudulent cases. Hence, achieving a balanced dataset is crucial to enhance the model's ability to discern between fraudulent and non-fraudulent cases effectively. In our approach, ADASYN (Adaptive Synthetic Sampling) [25] is primarily utilized to address the imbalance issue. ADASYN is adept at generating synthetic instances of the minority class, thereby promoting a more balanced dataset conducive to effective model training and accurate fraud detection. We used ADASYN over SMOTE (Synthetic Minority Over-sampling Technique) because ADASYN is particularly effective when dealing with datasets where the imbalance is severe. The method focuses on generating synthetic samples next to the original samples, which are harder to classify, rather than generating samples uniformly like SMOTE.

While SMOTE simply replicates the minority class features based on linear interpolations between existing minority class samples, ADASYN takes an extra step by considering the density distribution. This approach can potentially lead to better generalization on unseen data, as the synthetic samples are tailored more specifically to the problem areas of the feature space.

ADASYN (Adaptive Synthetic Sampling): Focused on the minority class, ADASYN adaptively generates synthetic data points based on the difficulty of classification. It aims to equip the model better by creating more examples around the harder-to-classify instances. This adaptive synthetic sampling approach essentially steers the learning process, enabling the model to be more attuned to complex patterns indicative of fraudulent activity. Our methodology encompasses applying ADASYN to the training data, ensuring that the synthetic minority class samples are used for model training, thereby mitigating the overfitting risk associated with synthetic data points in the validation or test sets.

The application of ADASYN in our insurance fraud detection model aims to bolster the predictive modelsâ resilience and accuracy, enhancing their ability to unearth intricate patterns and anomalies associated with fraudulent cases, amidst the prevailing class imbalance. This thoughtful application of data balancing techniques aims at fostering models that are not only accurate but also robust in identifying and predicting fraudulent activities in insurance claims.

However, we recognize the potential risk of overfitting associated with ADASYN, as the model might become too specialized to the synthetic samples generated in the minority class regions. To address this concern, we have implemented several strategies to ensure the robustness and generalizability of our models. Specifically, we use rigorous validation techniques, including cross-validation, to evaluate model performance on unseen data and to prevent overfitting. By comparing the results of models trained with ADASYN-generated samples against those trained with SMOTE-generated samples, we can assess whether the precision improvements come at the cost of reduced generalizability. Additionally, we employ regularization techniques and monitor performance metrics to ensure that the model remains well-calibrated and generalizes effectively to new data.

ADASYN is used to balance the dataset by generating synthetic examples of rare events, like fraud claims, enhancing model sensitivity and accuracy in detecting these critical but underrepresented occurrences.

3.4. Feature Selection

Various metaheuristic algorithms are utilized to select the most crucial features for the model. These algorithms include:

- Sine Cosine Algorithm
- Cuckoo Search
- Grey Wolf Optimizer

Each algorithm selects a subset of features, and the union of these subsets is used for model training and evaluation. the selected features by each of these three algorithms (Grey Wolf Optimizer, Cuckoo

Search Algorithm, Sine Cosine Algorithm) are shown in Figure 2 Unlike PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis), which are fundamentally linear methods and might not capture non-linear interactions between features effectively, metaheuristic algorithms like Sine Cosine Algorithm, Cuckoo Search, and Grey Wolf Optimizer do not assume any specific data distribution or linearity. Metaheuristic algorithms are highly adaptable to different kinds of data and problems. They can be easily adjusted for specific needs, such as handling very large feature spaces typical in insurance data, which might be computationally intensive for methods like PCA or RFE (Recursive Feature Elimination) when the feature set is extremely large.

These algorithms were applied using PyMetaheuristic v0.2. The Sine Cosine Algorithm was configured with a population size of 50, a linear component (a) of 2, and 100 iterations to search for the optimal set of features. The Cuckoo Search Algorithm was similarly configured with 50 birds, a discovery rate of 0.25, and 100 iterations. Additionally, an alpha value of 0.01 and lambda value of 1.5 were used to control the algorithm's balance between exploration and exploitation. The Grey Wolf Optimizer was configured with a pack size of 50 and 100 iterations. Features were selected based on their correlation with the target variable (FraudFound) and the highest relevance score. The objective function used was to maximize the sum of absolute correlations of the selected features. These selected features were then used to train the classifiers, and their contribution to the overall model accuracy was evaluated through cross-validation.

From the chart, it's evident that certain features, such as "Deductible,DriverRating,MaritalStatus, PolicyType", and some more features as shown in Figure are consistently selected by all three algorithms, indicated by the prominent green, blue, and red segments.

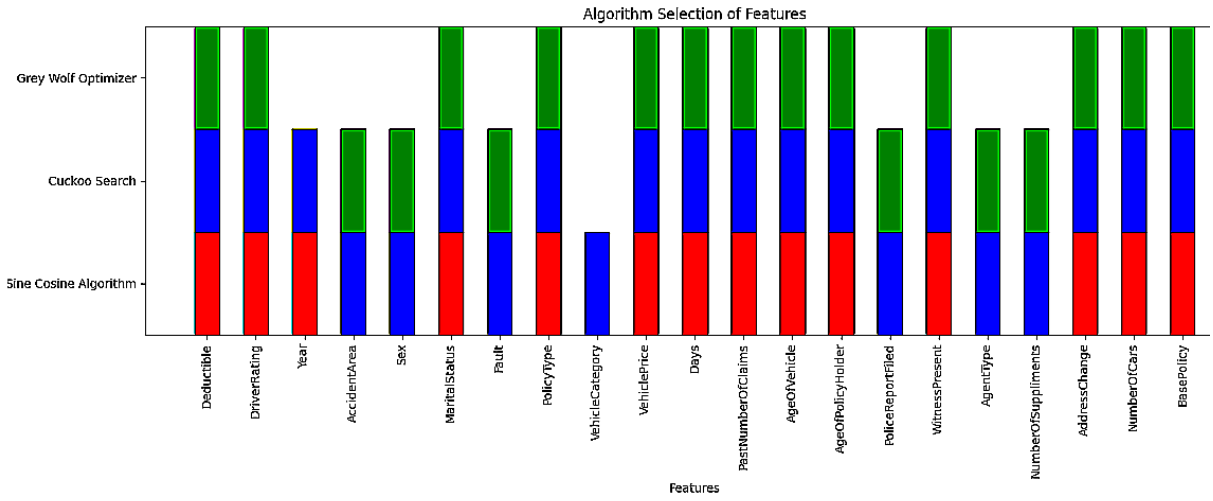


Figura 2: Features selected using different feature selection algorithms

This suggests that these features are highly influential in predicting outcomes within the ensemble models applied to this dataset. On the other hand, some features, like "VehicleCategory.[a]nd "WitnessPresent,"show variability in selection across different algorithms, indicating that their importance may be more context-dependent or that different algorithms prioritize them differently based on their internal mechanics.

The visual representation provides valuable insights into how different optimization algorithms contribute to the feature selection process, ultimately influencing the performance and interpretability of the ensemble models used in credit card insurance risk assessment.

3.5. Anomaly Detection

Anomaly detection is crucial in identifying fraudulent activities in insurance datasets.

Isolation Forest: By first choosing a feature at random and then choosing a split value at random between the feature's maximum and minimum values, the Isolation Forest algorithm separates obser-

vations. The reasoning is that since there are not many requirements, it is simpler to isolate anomaly observations.

In this approach, two methods were applied using the Isolation Forest algorithm. The first method involved giving higher weights to anomalies before fitting the model, resulting in better accuracy. In contrast, the second method involved filtering out anomalies, which led to lesser accuracy. Due to the improved performance, the first method, which involved giving higher weights to anomalies, was adopted. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and LOF (Local Outlier Factor)were also used but not considered as Isolation Forest gave the best accuracy. Isolation Forest is effective in handling high-dimensional data because it uses a random subspace method to build trees with a subset of features, reducing the dimensionality burden and avoiding the curse of dimensionality. Additionally, it relies on path lengths within these trees to determine anomalies, which is computationally efficient and less sensitive to the increased dimensions compared to distance-based metrics.

Visualization: To see the data in two dimensions, a scatter plot is created, with the anomalies shown in red and the normal points shown in green. The scatter plot uses the two principal components that were derived from PCA; the first principal component is represented by the x-axis, and the second principal component is represented by the y-axis. A considered dataset's Forest Scatter plot is displayed in Figure 3.

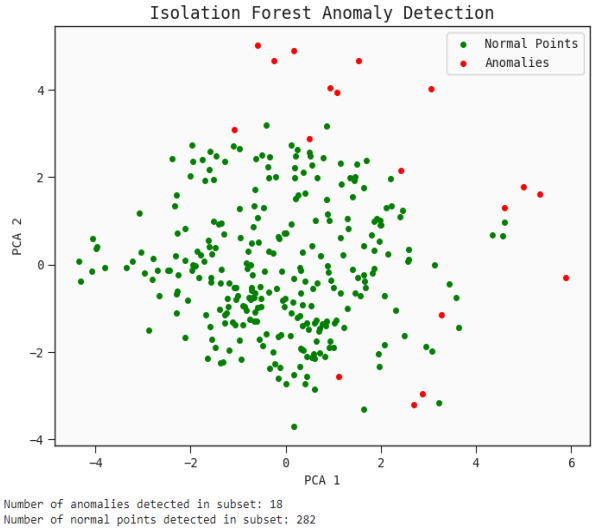


Figura 3: Isolation Forest Scatter Plot on a subset of data

In the realm of insurance fraud detection, leveraging algorithms such as Isolation Forest provides a powerful mechanism to identify and isolate fraudulent transactions effectively, thereby safeguarding the integrity of the insurance process. Through our methods and strategies, we found that assigning higher weights to anomalies before fitting the models yielded better accuracy in detecting potential fraudulent activities.

3.6. Model Training and Evaluation

Multiple classifiers are used for model training:

- Random Forest Classifier
- K Neighbors Classifier
- Decision Tree Classifier
- Logistic Regression
- AdaBoost Classifier

- CatBoost Classifier
- LGBM(Light Gradient Boosting Method) Classifier
- XGBoost(EXtreme Gradient Boosting) Classifier
- LSTM(Long Short Term) Classifier
- Stacked Ensemble Model

Each model is trained using the preprocessed and feature-selected dataset, taking into account the weights assigned during the anomaly detection phase.

3.7. Ensemble Method

Two separate ensemble models (stacking classifiers) are built using different base classifiers. From one side, stacking of Random Forest Classifier, K Neighbors Classifier, and Decision Tree Classifier is used, and for the other side, AdaBoost Classifier, CatBoost Classifier, LGBM Classifier, and XGBoost Classifier are stacked. The predictions from these two stacked ensemble models are then combined using logistic regression to make the final prediction.To combine the predictions from two different stacking ensemble models using logistic regression, first train each ensemble model separately on the dataset, with one model consisting of Random Forest, K Neighbors, and Decision Tree classifiers, and the other model consisting of AdaBoost, CatBoost, LGBM, and XGBoost classifiers. For each data point, generate predictions from both ensemble models. These predictions are then used as features to train a logistic regression model, where the target variable is the actual class label. The logistic regression model learns to weigh the predictions from each ensemble model to optimize the final prediction.The logistic regression model combines the outputs of the two stacking models by assigning weights to each model's prediction, based on their contribution to the final classification outcome. Specifically, the logistic regression model calculates the probability that a given input belongs to a particular class by taking a weighted sum of the predictions from both ensemble models. The model is trained to find the optimal weights that minimize prediction error,using a loss function (log-loss). During inference, the logistic regression model applies a threshold of 0.5 by default, classifying data points as positive if the combined probability is equal to or greater than 0.5.For new data, obtain predictions from both ensemble models, feed these into the trained logistic regression model, and make the final prediction based on the learned combination. K-fold cross-validation was also used to ensure that the logistic regression model generalizes well to unseen data. Our stacked ensemble model leverages the complementary strengths of multiple classifiers, offering improved predictive accuracy and robustness compared to traditional models.In the prior literature, the XGBoost Classifier was the best individual model in the prediction so we trained and tested that individually.

3.8. LSTM

In the pursuit of enhancing predictive accuracy and harnessing the potential of deep learning models, a Long Short-Term Memory (LSTM) neural network architecture was employed for the task at hand. The dataset was first preprocessed, and the training and testing sets were transformed into a suitable format for LSTM. LSTM was also carried out because LSTM networks are essential in insurance fraud detection for their prowess in handling sequential data and identifying anomalies like irregular claims. Their adaptive learning and real-time monitoring capabilities are crucial for staying ahead of evolving fraud tactics, ultimately reducing false positives and saving resources. Each sample is now presented in a three-dimensional shape (number of samples, time steps, features). This reshaping process ensures that the temporal dynamics of the data are effectively captured. The LSTM model, known for its ability to model sequential and time series data, was constructed. It comprised two LSTM layers, each with 50 units, with dropout and recurrent dropout applied to mitigate overfitting. The return of sequences in the first LSTM layer allows it to feed its output as input to the subsequent layer, enabling the model to learn complex temporal patterns in the data. A final dense layer with a sigmoid activation function was used to produce binary predictions. To optimize the model's performance, the Adam optimizer was chosen, and the binary cross-entropy loss function was employed to measure the model's predictive accuracy.

The model was trained for 100 epochs with a batch size of 32, and a portion of the training data was reserved for validation to monitor training progress and detect potential overfitting. The model's accuracy was assessed on the testing data, and the resulting accuracy metric, LSTM Model Accuracy: 0.939, was reported as a measure of the model's performance. LSTM (Long Short-Term Memory) LSTM was crucial for understanding the time-series data present in the dataset, such as the history of claims and policy updates, to effectively predict future claim events and detect fraudulent patterns by exploiting temporal relationships. There were many reasons for choosing an LSTM approach, too, which was directly applied to the original data features, as keeping time-dependent features and sequential features when using LSTM was important.

1. Temporal Features:-Age of Vehicle: Indicates how long a vehicle has been in use. Older vehicles might have different risk profiles for fraud compared to newer ones. LSTM can recognize if claims patterns correlate with vehicle age. -Days Policy Accident and Days Policy Claim: These features relate to the timing of incidents and claims post-accident. LSTM analyzes the sequence of days to detect unusual gaps or timing that might suggest fraudulent activities. - Month, Week of Month, Day of Week: Temporal data points like these helped the model learn seasonal trends or specific times when fraud is more likely. By analyzing claims across different times, LSTM can identify if claims are more likely to be fraudulent based on historical data during similar periods.

2. Sequential User Behavior: - Previous Claims: Reviewing the sequence and frequency of past claims by a policyholder allows LSTM to understand patterns that may indicate fraud, such as frequent claims within a short period. - Policyholder Age and Marriage Status: Changes in personal circumstances can influence claim patterns. LSTM can track these changes over time to identify anomalies in claiming behavior that might not match the expected profile.

3. Inter-related Features: - Police Report Filed and Witness Present: These binary features gain context through their sequence in the data. LSTM can discern patterns, such as frequently missing police reports in situations where claims are usually high, suggesting potential fraud. - Accident Area and Fault: By analyzing the location and cause of accidents sequentially, LSTM can identify if certain areas or types of accidents are frequently associated with fraudulent claims.

Feeding these and other features into the LSTM model effectively learns from the sequence and timing of events, which is critical in insurance contexts where the order and interval of actions can significantly indicate fraudulent activities. This capability to model long dependencies in sequences makes LSTM particularly suited for detecting fraud in insurance claims, where understanding the temporal sequence of events and their inter-relationships can highlight inconsistencies and anomalies indicative of fraudulent activities. Ability to Handle Long-Term Dependencies:

LSTMs were used over traditional RNNs(Recurrent Neural Networks) as it is designed to address the vanishing gradient problem that can occur, where the RNN struggles to learn correlations between events that occur at widely separated times. LSTM can remember temporal data over a longe r period of time. In insurance fraud detection, LSTMs outperformed ARIMA (Autoregressive Integrated Moving Average)/SARIMA(Seasonal Autoregressive Integrated Moving Average) and HMMs(Hidden Markov Models) due to LSTMs ability to handle long-term dependencies and complex, non-linear relationships between diverse data types. Unlike ARIMA/SARIMA, which were limited to linear, univariate forecasts, LSTMs excel in multivariate settings where interactions across various features are critical. HMMs, constrained by their reliance on short-term state dependencies, cannot capture the extensive temporal sequences as effectively as LSTMs, which use gates to manage information flow and mitigate noise. Consequently, LSTMs provide a more flexible, robust solution for detecting patterns and anomalies in the intricate and noisy datasets typical of insurance claims. So through LSTM age of the vehicle, , fault and the policyholderâs age were important features of its success

4. Results and discussions

To enhance the clarity of our results, the presentation has been reorganized into three distinct sections: (1) the overall performance of individual ensemble models, (2) a comparison of these models with the stacked ensemble approach, and (3) the application of Explainable Artificial Intelligence (XAI) techniques to interpret and validate the predictions made by these models. The inclusion of XAI not only ensures that our models are highly accurate but also transparent and interpretable, enabling stakeholders to

Algorithm 2 LSTM Training

Input: Sequential Training Data
Output: LSTM Model

1. Start
2. Reshape data for LSTM input (number of samples, time steps, features)
3. Construct LSTM network:
- Two LSTM layers with 50 units each, dropout and recurrent dropout
- Final dense layer with sigmoid activation for binary prediction
4. Compile the model using Adam optimizer and binary cross-entropy loss
5. Train the model for 100 epochs with a batch size of 32
6. Validate using a portion of training data
7. Evaluate the model on testing data
8. Report model accuracy
9. End

Cuadro 2: Comparison of data-balancing classification for Fraud detection

Models	ADASYN					Unbalanced dataset				
	Auc.	**AUC**	**Precision**	**Recall**	**F1-Score**	**Auc.**	**AUC**	**Precision**	**Recall**	**F1-Scor**
Random Forest	0.939	0.50	1.0	0.01	0.01	0.939	0.50	1.00	0.01	0.01
Decision Tree	0.91	0.61	0.26	0.27	0.26	0.923	0.66	0.37	0.37	0.37
Logistic Regression	0.939	0.51	0.67	0.01	0.03	0.938	0.500	0.33	0.01	0.01
KNN	0.92	0.51	0.10	0.04	0.05	0.935	0.51	0.21	0.02	0.04
CatBoost	0.94	0.57	0.68	0.13	0.22	0.946	0.59	0.72	0.18	0.29
LightGBM	0.94	0.54	0.67	0.09	0.15	0.945	0.56	0.86	0.13	0.22
XGBoost	0.95	0.66	0.78	0.32	0.45	0.953	0.66	0.78	0.33	0.47
Adaboost	0.94	0.51	0.25	0.03	0.05	0.936	0.51	0.27	0.03	0.05
Stacked Ensemble Model	**0.957**	0.71	0.74	0.43	0.55	**0.96**	0.75	0.77	0.51	0.61
LSTM Model	0.901	0.51	0.57	0.03	0.05	0.939	0.58	0.18	0.22	0.20

understand the decision-making process behind fraud detection and enhancing trust in the system's outputs.

4.1. Evaluation of the models

Models are evaluated using various metrics such as Accuracy, AUC(Area Under the Curve)-ROC(Receiver Operating Characteristic), Precision, Recall, and F1 Score.

The stacked models which are side one model and side two model as shown in Figure 4 are combined using logistic regression for the final stacked model using Stacking Classifier. Side one model has a random forest, k nearest neighbors, logistic regression, and a decision tree. Side one Model accuracy is 94.44, and the standard deviation is 0.26 percent. The two side models have Adaboost, Catboost, LGBM, and Xgboost. Side two Model is 95.60 percent and standard deviation of 0.42 percent.

This study explores a range of pipelines utilizing classifiers such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost Classifier, KNN, CatBoost Classifier, LightGBM, Adaboost, Stacked model, and LSTM, each paired with two distinct data balancing techniques, SMOTE and ADASYN, with ADASYN showing much better accuracy. A unique combination of algorithms and mathematical models for classification is employed in each pipeline, as detailed in Table 2.

The models were implemented using Scikit-learn v0.24, XGBoost v1.3, CatBoost v0.24, and LightGBM v3.1. For Random Forest, we used 100 trees with a maximum depth of 10 and the criterion set to Gini impurity. The Logistic Regression model was configured with L2 regularization and C=1.0 to control the strength of regularization.The Stacked Ensemble Model was constructed by combining Random Fo-

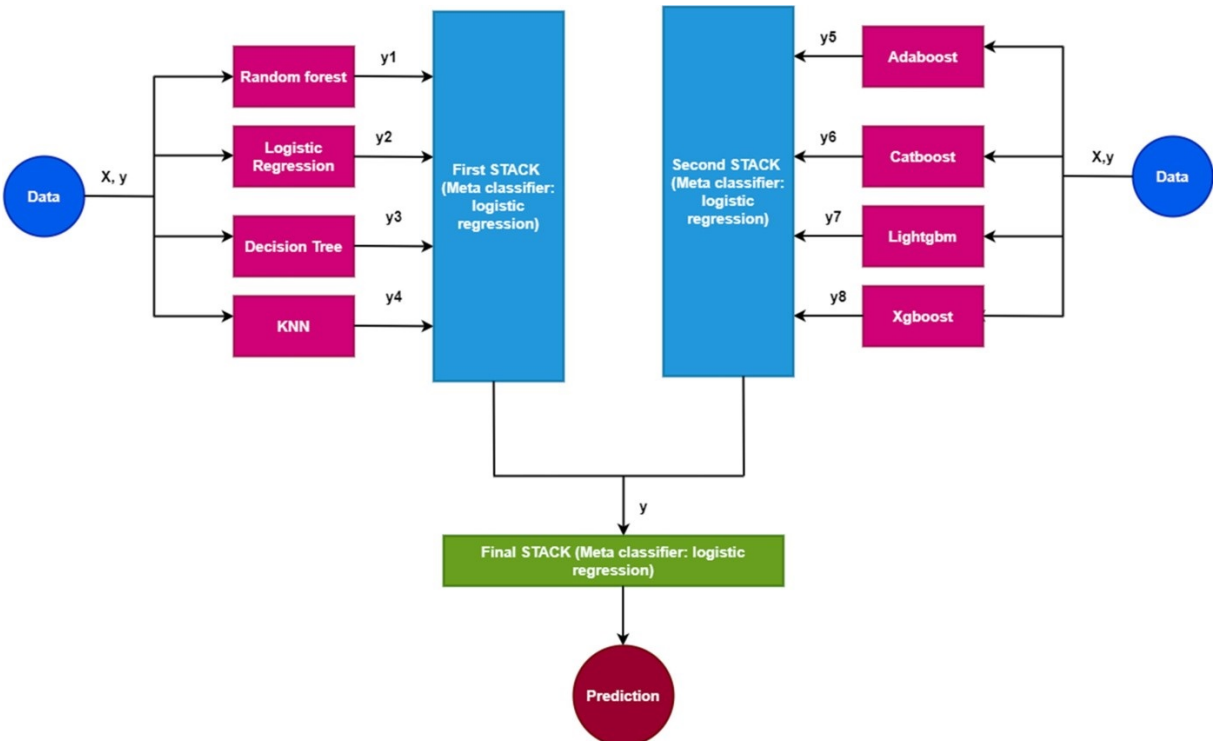


Figura 4: Custom stacking architecture to detect Insurance Fraud

rest, K-Nearest Neighbors (KNN), and Decision Tree as base models, with Logistic Regression acting as the meta-classifier to aggregate predictions. A second stacked ensemble included AdaBoost, CatBoost, LightGBM, and XGBoost, also using Logistic Regression as the meta-classifier. The stacking process was done using cross-validation predictions from the base models, which were then used as input features for the meta-classifier.All hyperparameters were optimized using Grid Search for the Random Forest, KNN, and Decision Tree models, while Random Search was applied for the XGBoost, CatBoost, and LightGBM models due to their larger hyperparameter spaces. The final models were evaluated using k-fold cross-validation to ensure robustness and reduce overfitting

Among these, XGboost demonstrated notable performance with 95 % accuracy, surpassing other models trained with the ADASYN approach. For the ADASYN dataset, the Stacked Ensemble Model demonstrates superior performance with the highest accuracy of 95.7 % and an Area Under the Curve (AUC) of 0.71. It also shows notable precision (0.74), recall (0.43), and F1-score (0.55). The XGBoost model shows impressive results with an AUC of 0.66, and balanced precision, recall, and F1-score values. The Adaboost model, Catboost model, and Lightgbm model, while slightly less effective, still achieve a commendable 94 % accuracy each. In the context of the Unbalanced dataset, the Stacked Ensemble Model again leads with an accuracy of 96 % and an AUC of 0.75, along with high precision, recall, and F1-score values. The XGBoost model closely follows with similar performance metrics, showcasing its robustness across different dataset types. The LSTM model shows reasonable performance, especially in the Unbalanced dataset with an accuracy of 93.9 %, a slightly higher AUC of 0.58 compared to the ADASYN dataset, and a balanced precision-recall profile. Importantly, the log loss of the final stacked ensemble model is notably low at 0.1404, indicating its high predictive accuracy and reliability in insurance fraud detection. Overall, the study highlights the effectiveness of these models in this field, especially when using sophisticated techniques like the Stacked Ensemble Model and XGBoost.

Our research demonstrates a significant improvement over the existing model presented in Table 1. While the existing approach [14] achieved high performance with an accuracy of 0.9811, it was noted for its high computational cost due to the complexity of the deep learning architecture employed. In contrast, our stacked ensemble model demonstrated high reliability, as evidenced by consistent performance metrics across different validation tests. The model not only provides a more balanced performance across

multiple metricsâincluding a precision of 0.957, recall of 0.74, F1 score of 0.77, and AUC of 0.96âbut also does so with a significantly lower computational overhead, indicating strong reliability in fraud detection. This makes our approach more efficient and practical for real-world applications where computational resources and time are critical. Our model achieves strong results without compromising efficiency, making it a more viable solution for credit card fraud detection.

4.2. Explainable Artificial Intelligence (XAI)

Our goal in this work is to construct pipelines for categorization and provide context for our predictions [34]. To explain various classifiers, we employed various XAI tools, such as ELI5, Quantum Lattice, Shapley Additive Explanations, and Local Interpretable Model-Agonistic Explanations. These XAI techniques are subsequently validated using insurance fraud literature and tree-based feature significance graphs.

4.2.1. Shapley Additive Explanations (SHAP):

A popular game theoretic classifier explanation tool is SHAP (Shapley Additive Explanations). Lundberg et al. [26, 27] proposed this mathematical model, which assesses each feature's Shapley values according to how well they contribute to a prediction. About an insurance dataset, the SHAP Beeswarm plot becomes a powerful tool for understanding the impact of various features on the modelâs predictions. The SHAP Beeswarm plot obtained for the Ensemble classifier employed in the insurance industry is shown in Figure 5. The features, such as fault, base policy, age of the vehicle, etc., are arranged along the y-axis in increasing order of significance or contribution to the prediction. Each dot on the plot represents one data point or prediction. The x-axis displays the SHAP values, quantifying the impact of each feature on the prediction. Moving along the y-axis from bottom to top increases the importance of the features in the prediction. Consider features like fault, base policy, and vehicle age. These features are prominent as they appear to have significant SHAP values. For instance, the 'FAULT' feature represents whether the policyholder was at fault in a claim. A higher SHAP value in 'FAULT' indicates a stronger effect on the prediction, possibly making the prediction more prone to being classified as a higher-risk policy. The plot's color gradient helps understand the feature values; redder dots indicate higher feature values. Other features like 'Policy Type,' 'Past no. of Claims', and 'Driver rating' also played roles in determining the model's prediction. For example, a policyholder with more past claims and a lower driver rating might be associated with higher SHAP values, indicating a higher propensity for risk or claims. The beeswarm plot offers a thorough global interpretive overview of the decision-making process used by the XgBoost classifier. It also sheds light on the features with the greatest predictive power in the insurance dataset. Thus helping to better understand and interpret the modelâs predictions in the context of insurance claims and risk assessment.

4.2.2. Local Interpretable Model-Agonistic Explanations (LIME):

LIME (Local Interpretable Model-agnostic Explanations) is a powerful tool for interpreting machine learning classifiers. Ribeiro et al. [28, 29] proposer of this methodology, elucidating how individual predictions could be unpacked for closer inspection. It operates by tweaking the inputs subtly and monitoring the prediction variance, enabling a deeper understanding of how each feature impacts the classifier's decision-making process. In our work, we applied LIME to interpret the predictions of a model tasked with insurance fraud detection. The application of LIME in our study facilitated insights into the modelâs decision-making patterns, especially in recognizing potential insurance frauds. Refer to Figure 6 for a vivid representation of our findings. In Figure 6, the LIME output reveals that the baseline probability for a claim being categorized as non-fraudulent, absent any feature influence, is 25.91 %. However, when introducing individual case features, the model predicts a 10.75 % probability of fraud, for instance, at hand, which contrasts with the actual model prediction probability of 7.25 %. Critical insights emerge from the weighted contributions of each feature. Notably, the 'Fault' feature, with a weight of -0.61, indicates a strong tendency for claims where the claimant is at fault to be deemed non-fraudulent. This suggests an underlying assumption that fraudulent activities are less likely when the claimant is responsible for the incident. Conversely,' Base Policy' and 'Witness Present' contribute to fraudulent cases. 'DriverRating,'

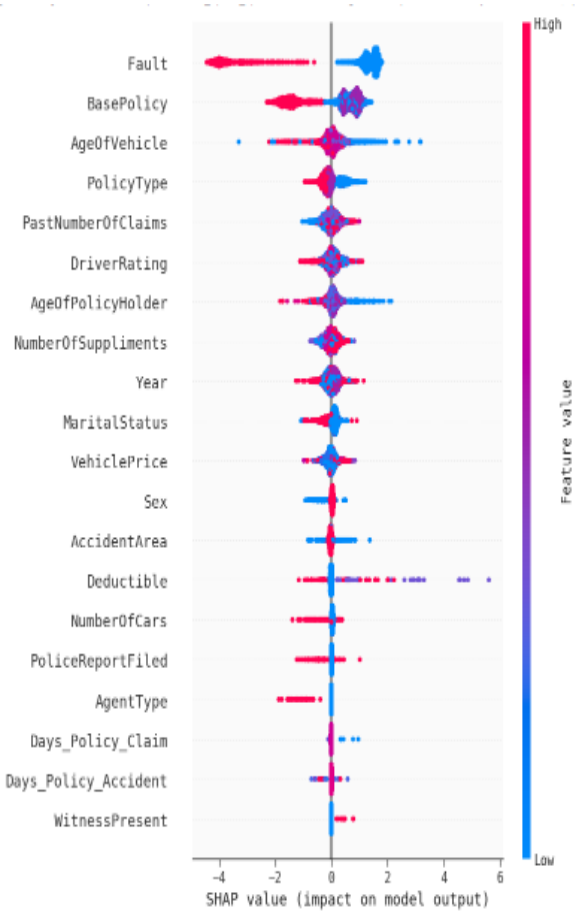


Figura 5: Plotting of SHAP Beeswarm acquired for Ensemble classifier

with a substantial positive weight of 1.36, suggests an unexpected correlation where higher driver ratings increase the likelihood of a claim being flagged as non-fraudulent. The 'AgeOfVehicle' feature, with a weight of +0.19, indicates that older vehicles are more likely to be associated with non-fraudulent claims according to the model's predictions. This could reflect a pattern in the training data where claims involving older vehicles are less often fraudulent, or it could be an artifact of how the model has learned to weigh certain features in its decision process.

In Figure 8, the Intercept was 0.3722, which acts as the baseline probability for the 'No Fraud' class without any feature influences. It indicates the model's inherent bias towards classifying instances as non-fraudulent. The Prediction_local, which was 0.0282, is the probability that the local surrogate model (the interpretable model generated by LIME) assigns to the instance being fraudulent. In this case, it predicts that there is a 2.82 % chance that the claim is fraudulent. The Right value was 0.0485, representing the original complex model's estimated probability that the instance is fraudulent. This is the output of the black-box model before being interpreted by LIME. 'Fault,' whose weight was -0.61, has a strong negative impact, pushing the prediction towards 'No Fraud." BasePolicy' whose weight was -0.05, contributed towards the likelihood of fraud. 'PoliceReportFiled' who weight was -0.17. If a police report is filed, it has a small negative impact on the probability of fraud, contributing towards 'No Fraud." DriverRating' with value of 1.36. A high driver's rating strongly increases the likelihood of no fraud. 'WitnessPresent' was -0.07. The presence of a witness slightly decreases the probability of fraud. 'Deductible', whose value was -0.19. A higher deductible contributes negatively to fraud probability. 'PolicyType' had a value of -0.22. This type of policy also contributes negatively to fraud probability, pushing the prediction towards 'No Fraud.' 'PastNumberOfClaims' had a value of 1.34. A higher number of past claims significantly increases the likelihood of no fraud. From the prediction probabilities box, we can see that the local

surrogate model is very confident in its 'No Fraud' prediction, with a 95 % probability assigned to 'No Fraud' and only a 5 % probability to 'Fraud'. This aligns with the 'Right' value of 4.85 %, indicating that the original model also considers the instance as likely not fraudulent, though with a slightly higher probability of fraud than the local model.

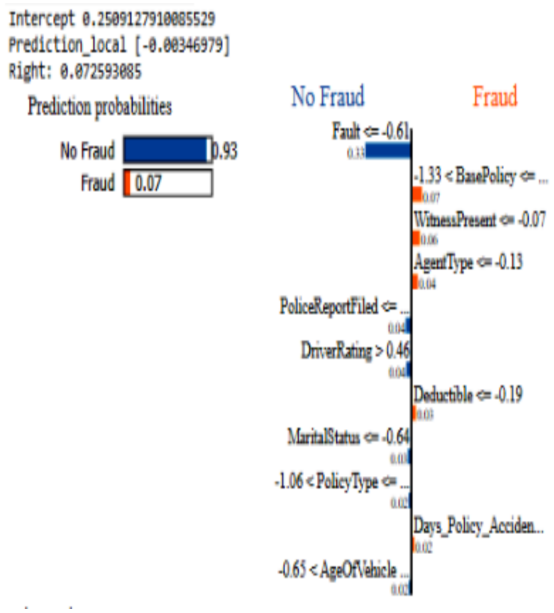


Figura 6: LIME Instance 0-Explanation 1

Feature	Value
Fault	-0.61
BasePolicy	-0.05
WitnessPresent	-0.07
AgentType	-0.13
PoliceReportFiled	-0.17
DriverRating	1.36
Deductible	-0.19
MaritalStatus	-0.64
PolicyType	-0.22
Days_Policy_Accident	0.05
AgeOfVehicle	0.19

Figura 7: LIME Instance 0-Explanation 2

In Figure 10, the model has predicted a specific instance as 'No Fraud' with a high probability of 0.97 versus 'Fraud' with a probability of 0.03. This suggests that the model is confident that the claim is not fraudulent. Features like 'PolicyType', with a weight of 4.83, have a significant positive impact, implying that something about the policy type is strongly associated with fraudulent claims. Also, 'BasePolicy' is directed towards fraudulent claims. Features like 'Fault' and 'Witness Present' have indications which

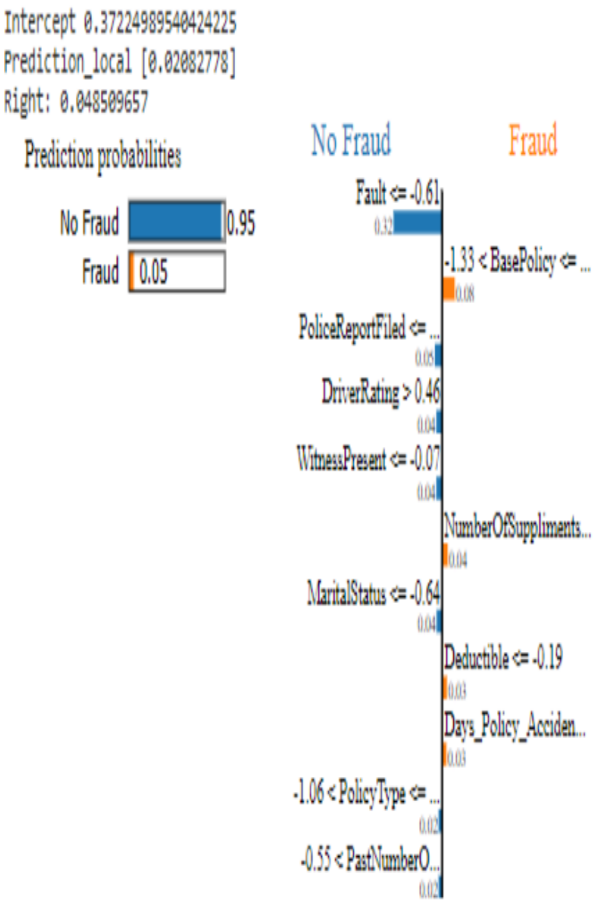


Figura 8: LIME Instance 1-Explanation 1

Feature	Value
Fault	-0.61
BasePolicy	-0.05
PoliceReportFiled	-0.17
DriverRating	1.36
WitnessPresent	-0.07
NumberOfSuppliments	-0.95
MaritalStatus	-0.64
Deductible	-0.19
Days_Policy_Accident	0.05
PolicyType	-0.22
PastNumberOfClaims	1.34

Figura 9: LIME Instance 1-Explanation 2

push the prediction towards 'No Fraud'.

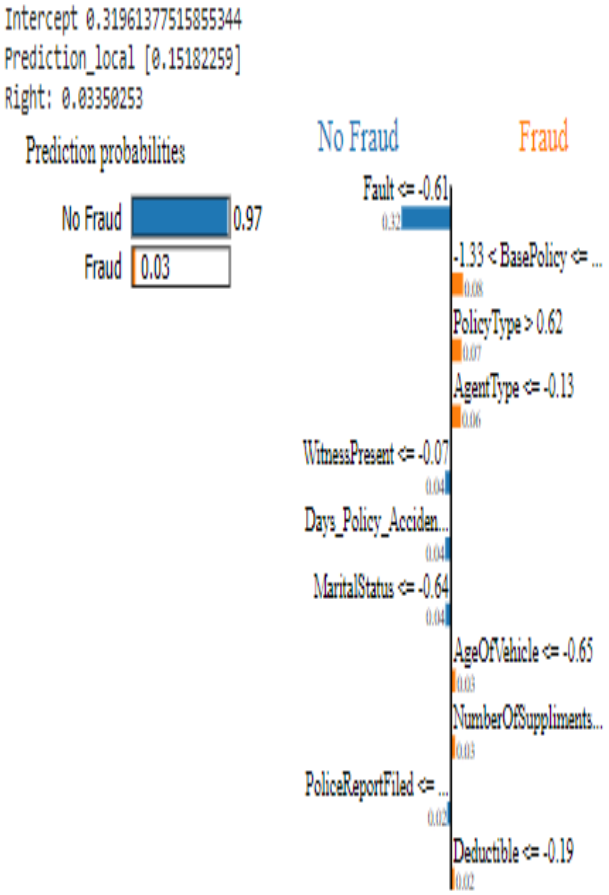


Figura 10: LIME Instance 2-Explanation 1

4.2.3. Explain Like Iâm 5 (ELI5 tool):

This Python module is instrumental for debugging and interpreting predictions made by machine learning classifiers, specifically those based on tree structures, within the realm of fraud detection. This [30] facilitates both regional and global explanations of the model's decisions. Explain Like I'm 5, or ELI5, is about breaking down complex concepts into simple, easy-to-understand explanations. Referencing Figure 12, a tabular interpretation of Xgboost's predictions about fraud detection is provided. From Figure 13 to Figure 18, features are methodically listed based on their contribution to the prediction, with the most influential feature occupying the foremost position. For instance, in our study, the 'Deductible,' 'Fault,' and 'AgeOfPolicyHolder' emerged as the most influential feature, receiving substantial weight during the modelâs construction and subsequently being selected as the root node. The sequential ordering from 'Deductible' to 'PastNumberOfClaims' illustrates a descending order of importance or influence each feature exerts on the prediction. This determination, elucidated through the modelâs global explanation, enables identifying and validating the most contributory features influencing such a decision. By leveraging ELI5 for global explanations, the integrity and reliability of the modelâs predictions regarding fraud detection are augmented, permitting a systematic validation against established research and benchmarks within the domain of fraud analytics.

Feature	Value
Fault	-0.61
BasePolicy	-0.05
PolicyType	4.83
AgentType	-0.13
WitnessPresent	-0.07
Days_Policy_Accident	0.05
MaritalStatus	-0.64
AgeOfVehicle	-0.65
NumberOfSuppliments	-1.60
PoliceReportFiled	-0.17
Deductible	-0.19

Figura 11: LIME Instance 2-Explanation 2

Global Explanation

Weight	Feature
0.0053 ± 0.0023	Deductible
0.0047 ± 0.0017	Fault
0.0043 ± 0.0024	AgeOfPolicyHolder
0.0026 ± 0.0037	AgeOfVehicle
0.0016 ± 0.0017	NumberOfSuppliments
0.0009 ± 0.0015	DriverRating
0.0007 ± 0.0016	AccidentArea
0.0006 ± 0.0014	BasePolicy
0.0004 ± 0.0011	MaritalStatus
0.0003 ± 0.0034	PolicyType
0.0002 ± 0.0009	WitnessPresent
0.0001 ± 0.0025	VehiclePrice
0.0001 ± 0.0011	PoliceReportFiled
0 ± 0.0000	VehicleCategory
-0.0001 ± 0.0003	Days_Policy_Accident
-0.0002 ± 0.0004	Days_Policy_Claim
-0.0003 ± 0.0014	PastNumberOfClaims
-0.0004 ± 0.0005	NumberOfCars
-0.0005 ± 0.0003	AgentType
-0.0010 ± 0.0009	Sex
… 1 more …	

Figura 12: ELI5 explanation plots: Global explanation

4.2.4. Feature Importance:

In the insurance fraud detection model, the feature importance technique was instrumental in identifying the most relevant attributes influencing the prediction. Different classifiers provided a spectrum of results, highlighting the featuresâ significance in various models.RandomForest Classifier: This model underscores the prominence of 'AgeOfPolicyHolder' and 'AgeOfVehicle.' The RandomForest's emphasis on these features suggests a correlation between the policyholder's age, the vehicle's age, and the likelihood of fraudulent claims. These attributes, therefore, emerge as critical indicators in the RandomForest's fraud detection paradigm. DecisionTree Classifier: Similar to RandomForest, this classifier also identifies 'AgeOfPolicyHolder' as a pivotal feature. However, it uniquely highlights 'Fault' as another significant

factor. The inclusion of 'Fault' indicates that the circumstances of the accident or claim play a crucial role in predicting fraud within the Decision Tree framework. AdaBoost Classifier: AdaBoost expands the feature landscape by bringing 'Deductible' and 'PolicyType' into the spotlight, alongside 'AgeOfVehicle'. This diversity implies that AdaBoost integrates policy-specific details and vehicle age, providing a broader perspective on fraud detection analysis. Light Gradient Boosting Machine (LGBM): LGBM classifier presents a varied set of important features, including 'AgeOfVehicle', 'AgeOfPolicyHolder', and 'NumberOfSuppliments'. This model seems to balance vehicle-related factors with policyholder demographics and the nuances of the claim (as indicated by 'NumberOfSuppliments'), offering a comprehensive view of potential fraud indicators. CatBoost Classifier: This model emphasizes 'Fault', 'VehicleCategory', and 'Year'. The inclusion of 'VehicleCategory' and 'Year' suggests an analytical focus on the type and model year of the vehicle, potentially correlating these with fraud propensity. This approach indicates CatBoost's inclination toward vehicle-specific attributes alongside the nature of the incident (Fault). XGBoost Classifier: XGBoost also emphasizes 'Fault' and 'AgeOfPolicyHolder,' similar to other models, but adds 'VehicleCategory' into its mix of critical features. This convergence on 'Fault' and 'AgeOfPolicyHolder' across multiple models reaffirms their significance, while the consideration of 'VehicleCategory' echoes a common theme observed in CatBoost, underscoring the relevance of the vehicle's characteristics. In conclusion, 'AgeOfPolicyHolder', 'Fault', and vehicle-related features (either 'AgeOfVehicle' or 'VehicleCategory') constitute the top three features in predicting insurance fraud, as indicated by the analysis of various machine learning classifiers.

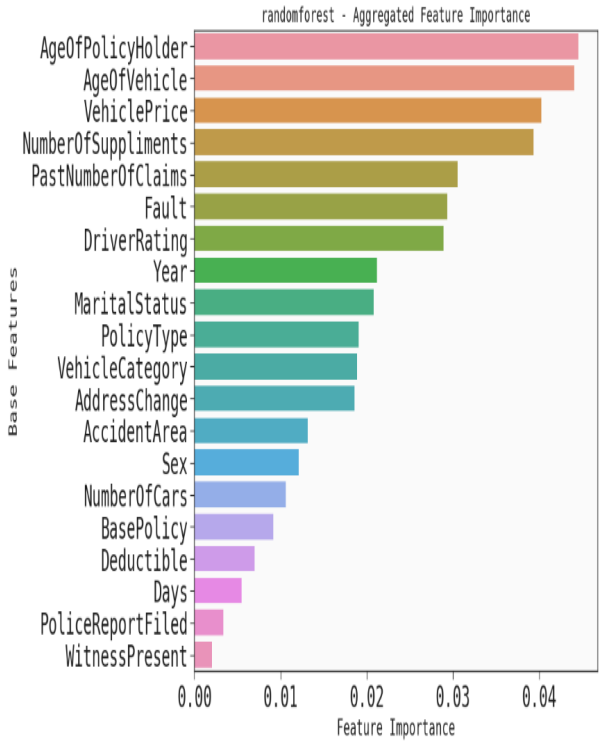


Figura 13: Random Forest

4.2.5. QLattice Interactive Plot: A Close Look at Insurance Fraud Detection

In our study, we utilized the QLattice [31, 32] interactive plot as shown in Figure 19 to meticulously dissect the decision-making process of our fraud detection model. We used Qlattice as it can handle both model complexity and interpretability efficiently. The plot vividly illustrates how each feature (clue) contributes to making a prediction - identifying whether a claim is fraudulent. In the presented model for insurance fraud detection, the interactive plot offers an intuitive schematic representation of the

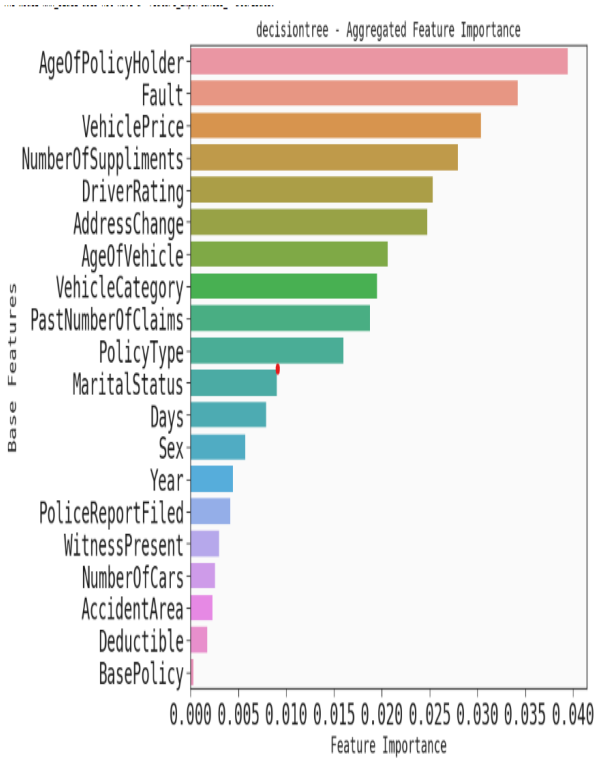


Figura 14: Decision Tree

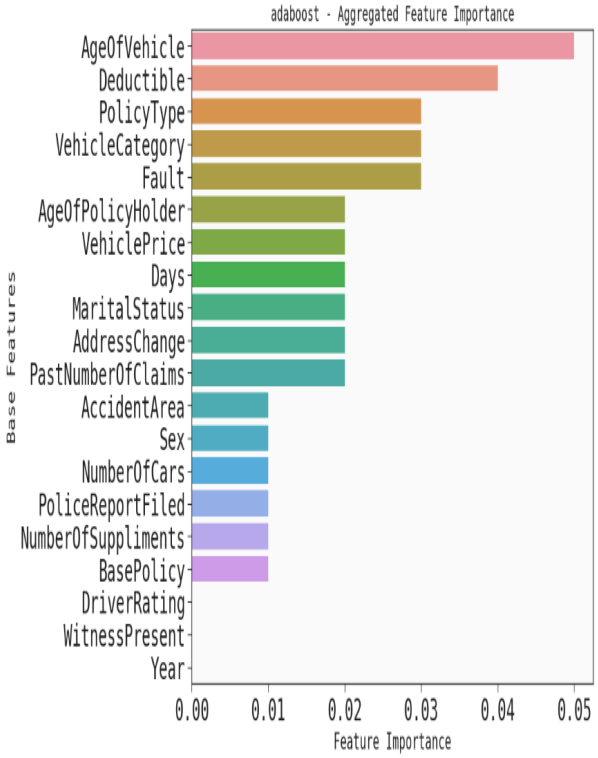


Figura 15: Adaboost

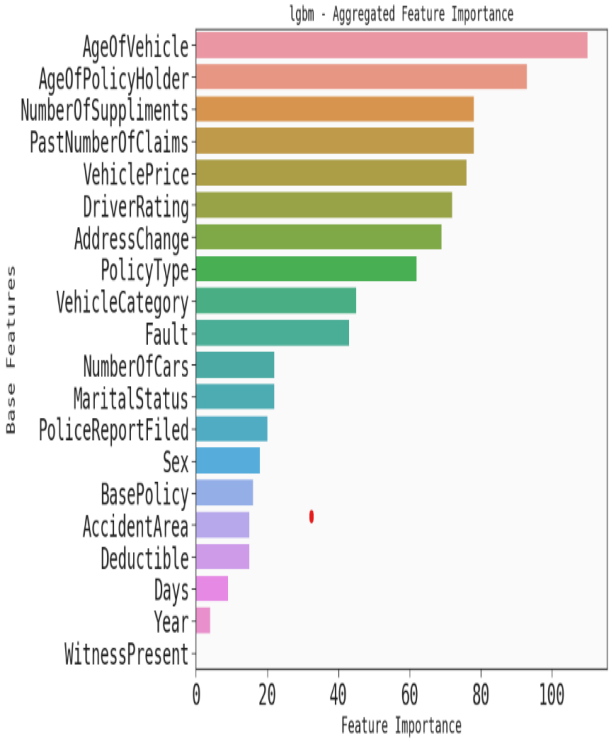


Figura 16: LightGBM

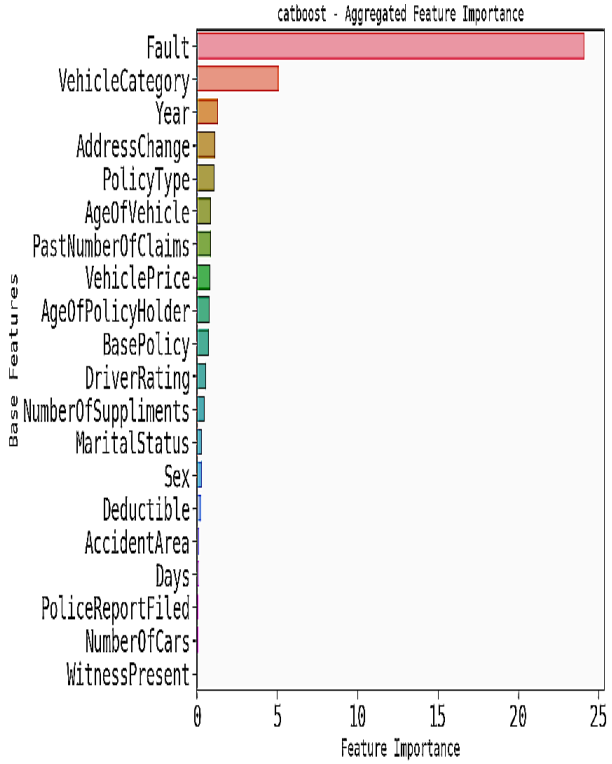


Figura 17: CatBoost

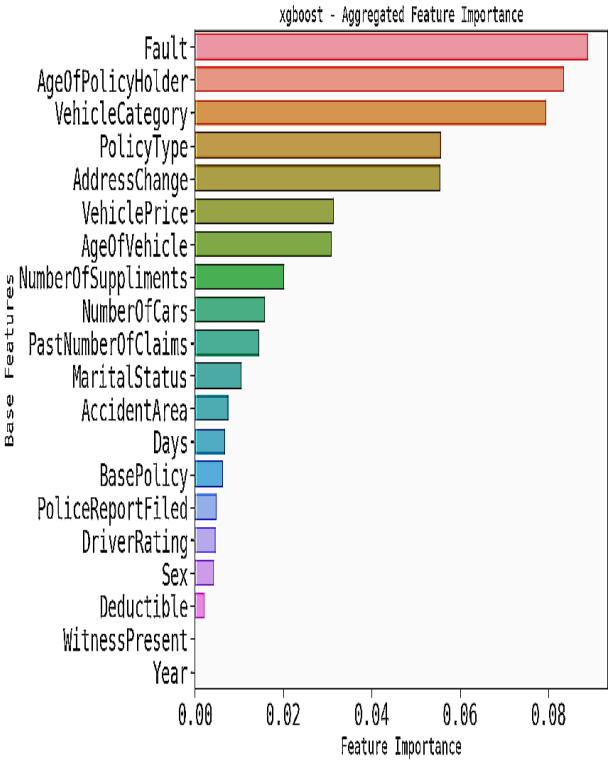


Figura 18: XGBoost

underlying computational process. The nodes, such as Fault, BasePolicy, and Deductible, indicate the critical input features or parameters harnessed by the model. Operations, signified by nodes like multiply and Gaussian, imply specific mathematical transformations applied to these input features. Conclusively, the node labeled FraudFound suggests the model's ultimate prediction output, highlighting whether a particular claim is fraudulent or genuine. Crucially, the âDisplaying activation of individual nodes"section sheds light on the activation strength of different nodes. Activation strength, in this context, may signify each node's relative importance or contribution to the decision-making process. This visualization offers an insightful perspective on feature importance, aiding in the interpretability of the model. Understanding which features play a more significant role in predicting fraudulent activity can be instrumental in refining and improving fraud detection mechanisms. Complementing the interactive plot, the ROC curve in Figure 20 provides an empirical evaluation of the model's classification performance. The trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various judgment thresholds is illustrated by the curve. Ideally, a competent model aspires to ascend rapidly towards a TPR of 1 with minimal FPR, straying as far from the diagonal line of no discrimination as possible. In this study, the model's ROC curve exhibits this behavior, showcasing its efficacy in distinguishing between fraudulent and legitimate claims. The reported Area Under the Curve (AUC) of 0.80 further underscores the model's robustness. AUC, capturing the overall performance across all thresholds, suggests that the presented model demonstrates good predictive capabilities. An AUC score of 0.80 is considerably higher than a random classifier (AUC=0.5), highlighting the model's effectiveness.

5. Discussion and Conclusion

Modern financial institutions must prioritize fraud detection, especially in critical and sensitive technical areas. Financial fraud has become more prevalent, especially in the vehicle insurance sector. In the past, several research and surveys on financial fraud detection have been created to address these problems through expert inspection and auditing. However, it is now impractical to detect fraud using such con-

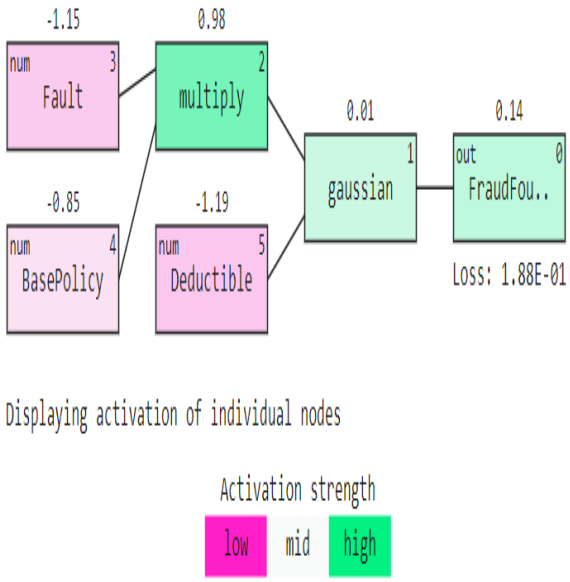


Figura 19: Qlattice Plot

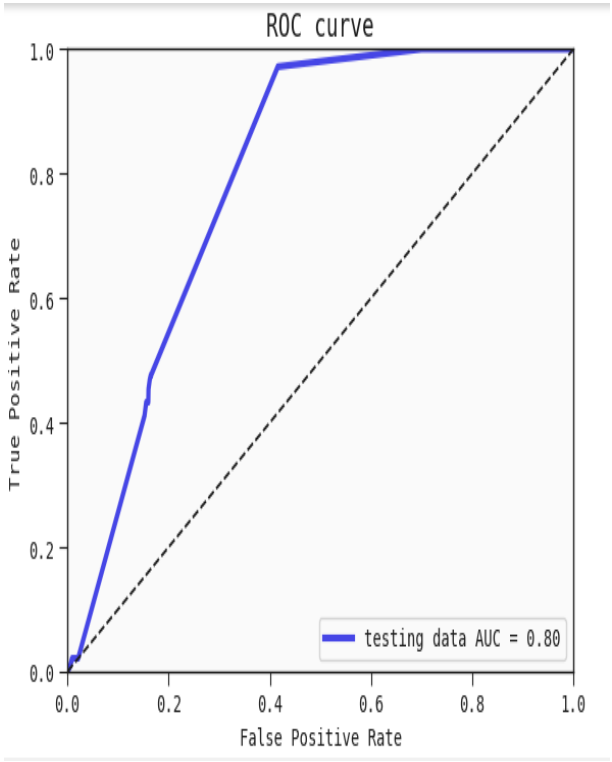


Figura 20: AUC curve for the best Qlattice Plot

ventional methods due to significant technological advancements. To investigate vehicle insurance fraud, we used nine prediction models: Random Forest, Decision Tree, Logistic Regression, KNN, CatBoost, LightGBM, XGBoost, Adaboost, and Stacked Ensemble Model and also LSTM classifier. Applying the stacking ensemble technique in this study not only advances the field of credit risk assessment but also sets a new benchmark for accuracy and interpretability in predictive modeling. For this study, the 2020

dataset on vehicle insurance fraud from US businesses is considered. It was gathered by the decision analytics division of an American company (EXL Service). To find the more pertinent characteristics in the data, we initially used the Sine Cosine, Cuckoo Search, and Grey Wolf Optimizer algorithms for feature selection. Age of Vehicle, Vehicle Category, Age of Policy Holder, Sex, Marital Status, Fault, Policy Type, Deductible, and Base Policy were the notable features that were chosen. It was found that the most important factors are the age of the vehicle, the base policy, fault, deductible, and the policyholderâs age.

An in-depth understanding of performance-improving aspects was provided by the application of nine models and balancing strategies. We first examined the parameters and their importance. We then used the ML models to effectively identify the fraud. Using explainable AI technologies like SHAP, LIME, ELI5, and Qlattice, we gained important insights into the reasoning behind the top-performing models. The XAI tools were verified using prior research proposals and feature importance. We came to the conclusion that attributes like BasePolicy, PolicyType, AgeofPolicyHolder, Deductible, AgeofVehicle, MaritalStatus, Fault, and PolicyCategory, Deductible all favorably affect fraud detection. Processes that aid with financing could be incorporated with these automated processes. The development of XAI can help bridge the gap between the fields of artificial intelligence and finance.

Compared to the aforementioned XAI-based literature papers [15–17], which employ an explainable multistage ensemble 1D Convolutional Neural Network for credit decision-making, our paper presents a more advanced approach by utilizing a stacked ensemble model that not only incorporates Explainable AI (XAI) but also achieves superior accuracy. While the previous work focuses on a specific type of neural network to enhance transparency, our stacked ensemble method leverages diverse base classifiers, leading to a more balanced and robust model. This approach not only improves interpretability but also significantly enhances predictive performance. Combining XAI techniques with a stacked ensemble model in our study results in better accuracy and more reliable credit scoring, making our approach a more effective and comprehensive solution for real-world credit decision-making scenarios. The results confirm the reliability of our stacked ensemble model, as evidenced by its consistent performance across various metrics and validation strategies. Explainable AI (XAI) tools further enhance reliability by providing transparency into the modelâs decision-making process.

Long-term growth for vehicle insurance companies would be made possible by the anticipated advantages of applying machine learning. The overall objective of this study is to facilitate the work of human investigators in the auto insurance industry, leading to more efficient identification and assessment of fraud. Because of this, our suggested framework may help insurance managers and businesses select the models and features for advanced machine learning and artificial intelligence-based fraud detection techniques.

Our research, however, is constrained because it only looks at US auto insurance data. which may not generalize to other regions such as Canada, Europe, or Asia. Future research should explore fraud detection strategies in these regions to determine if different approaches are needed. For more impactful feature selections, future studies should consider more feature engineering techniques. This includes integrating diverse data modalities in the future to gain a more thorough understanding of fraudulent behavior (text, images, voice, etc.). The need to examine how financial institutions and insurers may work together to share fraud-related information without compromising data privacy is important, enabling a more coordinated approach to preventing fraud. Further research could investigate the use of behavioral biometrics, such as speech recognition and mouse movement patterns, to improve user authentication and detect account takeover fraud. Examining the use of Deep Reinforcement Learning (DRL) in dynamic fraud detection and adaptive fraud prevention, where models can adapt and change their strategies in response to evolving fraud approaches, should also be studied. Finally, Examining how blockchain technology can be utilized to create immutable transaction records may provide additional security and a reliable audit trail, potentially reducing fraud.

In conclusion, our study provides a robust framework for improving fraud detection in the vehicle insurance sector using advanced machine learning and XAI techniques. By addressing the identified limitations and pursuing the suggested research directions, future studies can build upon these findings to develop even more effective fraud detection solutions.

Funding

This research received no external funding.

Data Availability

This review is based on publicly accessible databases; for more information, appropriate references could be looked up.

Declarations

The authors declare that none of the work reported in this study could have been influenced by any known competing financial interests or personal relationships.

Author Contributions

All authors contributed equally to this research. Sucharitha Shetty and Diana Olivia jointly conceptualized the study, developed the methodology, and Zaid Khan carried out the experiments. Data analysis and curation were performed collaboratively by all authors. All authors participated in writing the original draft and revising the manuscript for intellectual content.

Referencias

[1] Nalluri, Venkateswarlu and Chang, Jing-Rong and Chen, Long-Sheng and Chen, Jia-Chuan, 'Building prediction models and discovering important factors of health insurance fraud using machine learning methods,' in *Journal of Ambient Intelligence and Humanized Computing,* vol. 1, 2023, pp. 15–64.

[2] Jorn Debener and Volker Heinke and Johannes Kriebel, Detecting insurance fraud using supervised and unsupervised machine learning, *Journal of Risk & Insurance*, vol. 90, No. 3, 2023, pp. 743–768.

[3] Mohanta, Anuradha and Panigrahi, Suvasini, Health Insurance Fraud Detection Using Feature Selection and Ensemble Machine Learning Techniques, *Advances in Distributed Computing and Machine Learning*, vol. 1, No. 1, pp. 197–207, 2023.

[4] Yang, Jiaxi and Chen, Kui and Ding, Kai and Na, Chongqing and Wang, Meng, Auto Insurance Fraud Detection with Multimodal Learning, *Data Intelligence*, vol. 5, No. 2 pp. 388–412, Dec. 2022.

[5] E. H. Miller, "A note on reflector arrays," *IEEE Trans. Antennas Propagat.*, to be published.

[6] K. Aarati, A. Shirisha, K. Bindhu, Ch. Vandhana, Ch. Hasika, 'Identifying Health Insurance Claim Frauds using Machine Learning concept,' *Journal of Science and Technology*, No. 7, pp. 45-57, 2023.

[7] Hancock, John and Bauder, Richard and Wang, Huanjing, and Khoshgoftaar, Taghi, 'Explainable Machine Learning Models for Medicare Fraud Detection,' *Journal of Big Data*, vol. 10, No. 1, 2023

[8] H. Shi and M. A. Tayebi and J. Pei and J. Cao, Cost-Sensitive Learning for Medical Insurance Fraud Detection with Temporal Information, *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, No. 01, pp. 1-14, 2023

[9] Lu, J., Lin, K., Chen, R., Health insurance fraud detection by using an attributed heterogeneous information network (AHIN) with a hierarchical attention mechanism, *BMC Med Inform Decis Mak*, vol 23, No. 62, 2023

[10] Faheem Aslam and Ahmed Imran Hunjra and Zied Ftiti and Wael Louhichi and Tahira Shams, 'Insurance fraud detection: Evidence from artificial intelligence and machine learning', *Research in International Business and Finance*, vol. 62, 2022

[11] Abdullah, Mohammad, 'The implication of machine learning for financial solvency prediction: an empirical analysis on public listed companies of Bangladesh,' *Journal of Asian Business and Economic Studies*, Vol. 28, No. 4, 2021

[12] Forough, Javad and Momtazi, Saeedeh, 'Ensemble of deep sequential models for credit card fraud detection', *Applied Soft Computing*, Vol. 99, No. 11, 2021,

[13] Xinwei Zhang and Yaoci Han and Wei Xu and Qili Wang, 'HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture,' *Information Sciences*, Vol. 557, pp. 302-316, 2021

[14] aLean Yu and Xiao Yao and Xiaoming Zhang and Hang Yin and Jia Liu, 'A novel dual-weighted fuzzy proximal support vector machine with application to credit risk analysis,' *International Review of Financial Analysis*, Vol. 71, 2020

[15] N. Pavitha and Shounak Sugave. *Explainable Multistage Ensemble 1D Convolutional Neural Network for Trust Worthy Credit Decision*, *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, pp. 351â358, 2024

[16] N. Pavitha, P. Ratnaparkhi, A. Uzair, A. More, S. Raj, and P. Yadav. *Explainable AI for Sentiment Analysis*. In: J. Choudrie, P. Mahalle, T. Perumal, and A. Joshi, editors, *ICT with Intelligent Applications*, Smart Innovation, Systems and Technologies, vol. 311. Springer, Singapore, 2023

[17] X. Zhang and L. Yu. *Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. Expert Systems with Applications*, vol. XX, 2023

[18] W. Yotsawat, P. Wattuya, and A. Srivihok. *A Novel Method for Credit Scoring Based on Cost-Sensitive Neural Network Ensemble. IEEE Access*, 2021

[19] P. N and S. Sugave. *Explainable Multistage Ensemble 1D Convolutional Neural Network for Trust Worthy Credit Decision. International Journal of Advanced Computer Science and Applications*, 15(2), 2024

[20] W. Zhang, D. Yang, S. Zhang, "A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring," *Expert Systems with Applications*, vol. 174, no. 1, p. 114744, Jul. 2021. doi: 10.1016/j.eswa.2021.114744.

[21] X. Dastile and T. Celik, "Making Deep Learning-Based Predictions for Credit Scoring Explainable,"*IEEE Access*, vol. 9, pp. 50426-50440, 2021, doi: 10.1109/ACCESS.2021.3068854.

[22] P. Nooji and S. Sugave, .[Ex]plainable Multistage Ensemble 1D Convolutional Neural Network for Trust Worthy Credit Decision,"*International Journal of Advanced Computer Science and Applications*, vol. 15, 2024, doi: 10.14569/IJACSA.2024.0150237.

[23] Yi Liu and Menglong Yang and Yudong Wang and Yongshan Li and Tiancheng Xiong and Anzhe Li, 'Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China,' *International Review of Financial Analysis*, Vol. 79, 2022,

[24] Pedro Carmona and Aladdin Dwekat and Zeena Mardawi, 'No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure', *Research in International Business and Finance*, Vol. 61, 2022

[25] H. He, Y. Bai, E. A. Garcia and S. Li, 'ADASYN: Adaptive synthetic sampling approach for imbalanced learning', Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.), pp. 1322-1328, Jun. 2008.

[26] Baptista, M.L.; Goebel, K.; Henriques, E.M. 'Relation between prognostics predictor evaluation metrics and local interpretability SHAP values', *Artif. Intell.*, 2022

[27] Lundberg S.M., Lee S.I., 'A unified approach to interpreting model predictions,' *Adv. Neural Inf. Process. Syst.*, Vol.30, 2017

[28] Dieber, J.; Kirrane, S. 'Why model why? Assessing the strengths and limitations of LIME', *arXiv 2020*

[29] Ribeiro M.T., Singh S., Guestrin C., 'Model-agnostic interpretability of machine learning,' *arXiv preprint arXiv:1606.05386*, 2016

[30] Kuzlu, M., Cali, U., Sharma, V., GÃ¼ler, O., 'Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools,' *IEEE Access*, Vol. 8, 2020

[31] Bharadi V., 'QLattice Environment and Feyn QGraph ModelsâA New Perspective Toward Deep Learning', *In Emerging Technologies for Healthcare: Internet of Things and Deep Learning Models*, pp. 69â92, 2021

[32] Brolos K.R., Machado M.V., Cave C., Kasak J., Stentoft-Hansen V., Batanero V.G., Wilstrup C., 'An approach to symbolic regression using feyn,' 2021