**Machine Learning based Method for Insurance Fraud Detection on Class Imbalance Datasets with Missing Values**

**ABSTRACT:**

Insurance fraud, particularly within the automobile insurance sector, is a significant challenge faced by insurers, leading to financial losses and influencing pricing strategies. Fraud detection models are often impacted by class imbalance, where fraudulent claims are much rarer than legitimate claims, and missing data further complicates the process. This research tackles these issues by utilizing two car insurance datasets—an Egyptian real-life dataset and a standard dataset. The proposed methodology includes addressing missing data and class imbalance, and it incorporates the AdaBoost Classifier to enhance the model’s accuracy and predictive power. The results demonstrate that addressing class imbalance plays a crucial role in improving model performance, while handling missing data also contributes to more reliable predictions. The AdaBoost Classifier significantly outperforms existing techniques, improving prediction accuracy and reducing overfitting, which is often a challenge in fraud detection models. This study presents valuable insights into how improving data quality and using advanced algorithms like AdaBoost can enhance fraud detection systems, ultimately leading to more effective identification of fraudulent claims. These enhancements can significantly aid insurance companies in reducing financial losses, improving decision-making, and refining pricing models.

**INTRODUCTION**

1

1.Introduction :

Insurance fraud, particularly in the automobile insurance sector, poses a significant challenge for insurance companies. Fraudulent claims not only result in substantial financial losses but also influence pricing strategies, ultimately leading to higher premiums for legitimate policyholders. One of the major hurdles in fraud detection is the class imbalance problem, where fraudulent claims are far less frequent than legitimate ones. This imbalance often leads to biased models that fail to correctly identify fraudulent claims. Furthermore, missing data exacerbates the issue by further complicating the training of effective predictive models. These challenges have prompted researchers to explore more sophisticated techniques to enhance fraud detection.

In response to these challenges, the proposed study utilizes two car insurance datasets—an Egyptian real-life dataset and a standard dataset—to develop a more robust fraud detection system. The methodology focuses on addressing both the class imbalance and missing data problems. Specifically, the research introduces the AdaBoost Classifier, a powerful machine learning algorithm that improves prediction accuracy by enhancing weak classifiers through ensemble learning. By applying AdaBoost, the model effectively handles both class imbalance and overfitting, common issues faced by traditional fraud detection models. The study demonstrates that addressing class imbalance significantly enhances the model’s performance, while the treatment of missing data ensures that the predictions remain reliable and accurate. The AdaBoost Classifier outperforms existing models, making it a promising approach for more effective fraud detection. This work provides valuable insights into how advanced machine learning algorithms, when combined with improved data quality handling techniques, can lead to more reliable and efficient fraud detection systems, ultimately helping insurance companies reduce financial losses and improve decision-making processes.

**SCOPE OF THE PROJECT**

The scope of this project focuses on addressing the challenges of insurance fraud detection within the automobile insurance sector. It aims to enhance the accuracy and efficiency of fraud detection models by utilizing advanced machine learning techniques, specifically addressing the class imbalance problem and missing data issues. The project involves working with real-life datasets, applying the AdaBoost Classifier, and evaluating the model's performance in comparison to existing systems. Additionally, the study explores how these enhancements can lead to better prediction accuracy, reduced overfitting, and more reliable fraud detection systems.

**OBJECTIVE**

The objective of this project is to develop an advanced and efficient insurance fraud detection system tailored for the automobile insurance industry. The project focuses on addressing key challenges such as class imbalance, where fraudulent claims are underrepresented compared to legitimate claims, and missing data, which often affects the model’s accuracy. By leveraging machine learning techniques, particularly the AdaBoost Classifier, the aim is to enhance prediction accuracy and reduce overfitting, ensuring the model generalizes better on unseen data. This project also seeks to provide a framework for improving the overall effectiveness of fraud detection systems, leading to more reliable identification of fraudulent claims and aiding in better decision-making and pricing strategies for insurance companies. Through this, the project aims to contribute to reducing financial losses for insurers and improving the overall integrity of the insurance system.

**EXISTING SYSTEM:**

Existing fraud detection models in the insurance industry typically rely on conventional machine learning algorithms like Logistic Regression, Decision Trees, and Random Forests. However, they face challenges when working with imbalanced datasets, where fraudulent claims are relatively rare compared to legitimate claims. This imbalance leads to poor model performance, as the classifiers are biased towards the majority class. Furthermore, missing data and inconsistencies in the dataset contribute to unreliable predictions, affecting the overall accuracy of these models. Despite their utility, traditional methods often suffer from overfitting, where models perform well on training data but fail to generalize to unseen data. Overfitting occurs because these models can become too complex when trying to fit noisy or incomplete data, resulting in reduced model robustness. While some approaches use sampling methods like undersampling or oversampling to address class imbalance, they do not always achieve optimal results, especially when the dataset contains many missing or noisy entries.

**EXISTINGSYSTEM DISADVANTAGES:**

* Class Imbalance
* Handling Missing Data
* Overfitting
* Limited Performance with Complex Fraud Patterns
* Inefficient Fraud Detection

**LITERATURE SURVEY**

**Title:** Encoding High-Cardinality String Categorical Variables

**Author:** Patricio Cerda, G. Varoquaux

**Year:** 2022

**Description:** Statistical models usually require vector representations of categorical variables, using for instance one-hot encoding. This strategy breaks down when the number of categories grows, as it creates high-dimensional feature vectors. Additionally, for string entries, one-hot encoding does not capture information in their representation.Here, we seek low-dimensional encoding of high-cardinality string categorical variables. Ideally, these should be: scalable to many categories; interpretable to end users; and facilitate statistical analysis. We introduce two encoding approaches for string categories: a Gamma-Poisson matrix factorization on substring counts, and the min-hash encoder, for fast approximation of string similarities. We show that min-hash turns set inclusions into inequality relations that are easier to learn. Both approaches are scalable and streamable. Experiments on real and simulated data show that these methods improve supervised learning with high-cardinality categorical variables. We recommend the following: if scalability is central, the min-hash encoder is the best option as it does not require any data fit; if interpretability is important, the Gamma-Poisson factorization is the best alternative, as it can be interpreted as one-hot encoding on inferred categories with informative feature names. Both models enable autoML on the original string entries as they remove the need for feature engineering or data cleaning.

**Title:** Predicting Insolvency of Insurance Companies in Egyptian Market Using Bagging and Boosting Ensemble Techniques

**Author:** Ahmed A. Khalil , Zaiming Liu1 , Ahmad Salah , Ahmed Fathalla , And Ahmed Ali

**Year:** 2022.

**Description**: Insolvency is a crucial problem for several insurance companies that suffer from it. This problem has direct or indirect effects on both the people working in the financial business and normal citizens. Thus, in insurance companies, the ability to predict insolvency is in great demand. There are several efforts proposed to predict insurance company insolvency using computer science methods (e.g., support vector machine and fuzzy systems). Each country has its own data patterns due to interior matters. Thus, insurance companies from different countries may have different data patterns. Consequently, the utilized predictive model should adapt to the dataset at hand. To our best knowledge, despite there are several efforts to build an insolvency predictive model, none of these efforts explored the Egyptian market. In addition, even the existing efforts did not utilize the ensemble learning methods in the insolvency prediction problem. In this context, we have two main contributions to this work. First, we proposed the first public access dataset of Egyptian insurance companies. The collected dataset was gathered from 11 Egyptian insurance companies during the years 1999 to 2019. The dataset consists of a set of 22 ratios (21 input features and one output feature), e.g., retention and investment yield alongside the solvency ration (i.e., the target feature). In the second contribution, we proposed exploring the performance of the ensemble learning methods to address the insolvency prediction problem. Thus, we proposed building several insolvency predictive models using ensemble learning and classic machine learning models. Next, the proposed models are evaluated on different accuracy metrics, e.g., Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The experimental results revealed that the ensemble learning-based models outperformed the classic machine learning-based models. Moreover, the correlation analysis between the utilized 22 financial ratios revealed that the most significant ratios, for the task of predicting the solvency ratio, are the technical provisions to shareholders’ funds, insurance companies’ debit balances to shareholders, and earnings after taxes to shareholders’ funds.

**Title:** Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach

**Author:** by Abdul Rehman Khalid ,Nsikak Owoh ,ORCID,Omair Uthmani 1,Moses Ashawa ,Jude Osamor, John Adejoh

**Year:** 2024.

**Description:** In the era of digital advancements, the escalation of credit card fraud necessitates the development of robust and efficient fraud detection systems. This paper delves into the application of machine learning models, specifically focusing on ensemble methods, to enhance credit card fraud detection. Through an extensive review of existing literature, we identified limitations in current fraud detection technologies, including issues like data imbalance, concept drift, false positives/negatives, limited generalisability, and challenges in real-time processing. To address some of these shortcomings, we propose a novel ensemble model that integrates a Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Bagging, and Boosting classifiers. This ensemble model tackles the dataset imbalance problem associated with most credit card datasets by implementing under-sampling and the Synthetic Over-sampling Technique (SMOTE) on some machine learning algorithms. The evaluation of the model utilises a dataset comprising transaction records from European credit card holders, providing a realistic scenario for assessment. The methodology of the proposed model encompasses data pre-processing, feature engineering, model selection, and evaluation, with Google Colab computational capabilities facilitating efficient model training and testing. Comparative analysis between the proposed ensemble model, traditional machine learning methods, and individual classifiers reveals the superior performance of the ensemble in mitigating challenges associated with credit card fraud detection. Across accuracy, precision, recall, and F1-score metrics, the ensemble outperforms existing models. This paper underscores the efficacy of ensemble methods as a valuable tool in the battle against fraudulent transactions. The findings presented lay the groundwork for future advancements in the development of more resilient and adaptive fraud detection systems, which will become crucial as credit card fraud techniques continue to evolve.

**Title:**  Machine Learning Method for Return Direction Forecast of Exchange Traded Funds (ETFs) Using Classification and Regression Models

**Author:**  Raphael Paulo Beal Piovezan, Pedro Paulo de Andrade Junior & Sérgio Luciano Ávila

**Year:** 2023

**Description**: — This article aims to propose and apply a machine learning method to analyze the direction of returns from exchange traded funds using the historical return data of its components, helping to make investment strategy decisions through a trading algorithm. In methodological terms, regression and classification models were applied, using standard data sets from five reference markets, in addition to algorithmic error metrics. In terms of research results, they were analyzed and compared to those of the Naïve forecast and the returns obtained by the buy & hold technique in the same period of time. In terms of risk and return, the models mostly performed better than the control metrics, with emphasis on the linear regression model and the classification models by logistic regression, support vector machine (using the LinearSVC model), Gaussian Naive Bayes and K-Nearest Neighbors, where in certain data sets the returns exceeded by two times and the Sharpe ratio by up to four times those of the buy & hold control model.

**Title:** Machine Learning-Based Approaches for Real-Time Fraud Detection and Prevention

**Author**: Vikas R. Shetty, Pooja R, [Rashmi Laxmikant Malghan](https://researcher.manipal.edu/en/persons/rashmi-laxmikant-malghan)

**Year:** 2023**.**

**Description:** The proliferation of internet services in various industries, especially the financial sector, has increased financial fraud. Fraud detection and prevention are critical to protecting both individuals and organizations from significant financial loss. However, the lack of publicly available datasets containing fraud is a major challenge. This study aims to address these issues using advanced machine learning techniques. Known for their ability to provide insight into data, decision trees are used for real-time fraud detection. In addition, deep learning techniques and artificial neural networks (ANN) are used to detect complex fraud patterns, while logistic regression is used to model the probability of fraudulent events. The accuracy of these methods, including decision trees, logistic regression, and ANN, is fully evaluated, with accuracies of 99.8%, 99.9%, and 99.94%, respectively. These findings provide valuable guidance for companies on choosing effective anti-fraud strategies and shed light on the adaptability of algorithms to real financial contexts, contributing to machine learning-based fraud detection.

**PROPOSED SYSTEM**

The proposed system enhances insurance fraud detection by incorporating the AdaBoost Classifier, an ensemble technique that improves predictive performance by combining multiple weak classifiers to create a strong, accurate model. AdaBoost's ability to reduce overfitting makes it especially effective for handling noisy or incomplete datasets, ensuring that the model generalizes well to new data. This classifier is also integrated with techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address the class imbalance problem, thereby improving the detection of fraudulent claims that are underrepresented in the dataset.

Furthermore, the proposed system employs more robust data preprocessing methods to handle missing data effectively. These preprocessing techniques ensure that the model works with cleaner, more complete datasets, improving overall prediction accuracy. The combination of AdaBoost with these advanced data handling methods makes the model more reliable, scalable, and efficient. By addressing both class imbalance and missing data, the proposed system outperforms traditional methods in terms of accuracy and robustness, providing a more effective solution for fraud detection in the insurance industry.

**PROPOSED SYSTEM ADVANTAGES:**

* Enhanced Fraud Detection
* Effective Missing Data Handling
* Optimized Model Performance
* Reduced Overfitting
* Increased Accuracy in Predictions

**System Requirements**

**HARDWARE REQUIREMENTS**

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It should what the system do and not how it should be implemented.

* PROCESSOR : DUAL CORE 2 DUOS.
* RAM : 4GB DD RAM
* HARD DISK : 500 GB

**SOFTWARE REQUIREMENTS**

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the teams and tracking the team’s progress throughout the development activity.

* Operating System : Windows 10
* Platform : Spyder3
* Programming Language : Python
* Front End : Spyder3

**SYSTEM ARCHITECTURE:**

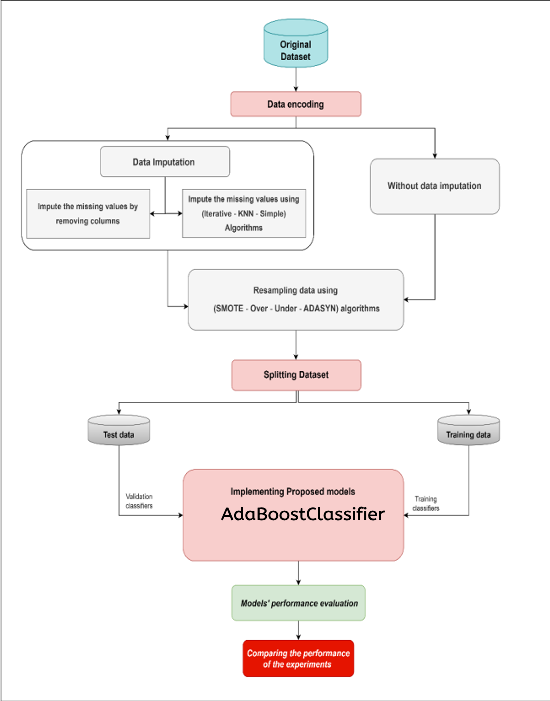


Fig 4.11: System Architecture

**REFERENCES**

[1] A. A. Khalil, Z. Liu, and A. A. Ali, “Using an adaptive network‐based fuzzy inference system

25, model to predict the loss ratio of petroleum insurance in Egypt,” Risk Management and Insurance Review, vol. no. 1, pp. 5–18, 2022, doi: 10.1111/rmir.12200.

[2] C. Bockel-Rickermann, T. Verdonck, and W. Verbeke, “Fraud analytics: A decade of research:

Organizing challenges and solutions in the field,”Expert Syst Appl, vol. 232, p. 120605, 2023, doi:

<https://doi.org/10.1016/j.eswa.2023.120605>.

[3] Y. Wang and W. Xu, “Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud,” Decis Support Syst, vol. 105, pp. 87–95, 2018, https://doi.org/10.1016/j.dss.2017.11.001.

[4] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, “Performance comparative study of machine learning algorithms for automobile insurance fraud detection,” in 2019 Third

International Conference on Intelligent Computing in Data Sciences (ICDS), 2019, pp. 1–4. 10.1109/ICDS47004.2019.8942277.

[5] R. P. B. Piovezan, P. P. de Andrade Junior, and S. L. Ávila, “Machine Learning Method for Return Direction Forecast of Exchange Traded Funds (ETFs) Using Classification and Regression Models,” Comput Econ, 2023, doi: 10.1007/s10614023-10385-4.

[6] A. A. Khalil, Z. Liu, A. Salah, A. Fathalla, and A. Ali, “Predicting Insolvency of Insurance Companies in Egyptian Market Using Bagging and Boosting Ensemble Techniques,” IEEE Access, vol. 10, pp. 117304–117314, 2022, 10.1109/ACCESS.2022.3210032.

[7] N. Boodhun and M. Jayabalan, “Risk prediction in life insurance industry using supervised learning algorithms,” Complex & Intelligent Systems, vol. 4, no. 2, pp. 145–154, 2018, doi: 10.1007/s40747-0180072-1.

[8] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, “A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques,” Artif Intell Rev, vol. 56, no. 11, pp. 13407–13461, 2023, doi: 10.1007/s10462-023-10472-w.

[9] M. Liao, S. Tian, Y. Zhang, G. Hua, W. Zou, and X. Li, “PDA: Progressive Domain Adaptation for Semantic Segmentation,” Knowl Based Syst, vol. 284, p. 111179, 2024, https://doi.org/10.1016/j.knosys.2023.111179.

[10] A. Khalil, Z. Liu, and A. Ali, “Precision in Insurance Forecasting: Enhancing Potential with Ensemble and Combination Models based on the Adaptive Neuro Fuzzy Inference System in the Egyptian Insurance Industry,” Applied Artificial Intelligence, vol. 38, no. 1, p. 2348413, Dec. 10.1080/08839514.2024.2348413. 2024,

[11] A. K. I. Hassan and A. Abraham, “Modeling insurance fraud detection using ensemble combining classification,” International Journal of Computer Information Systems and Industrial Management Applications, vol. 8, pp. 257–265, 2016.

[12] V. R. Shetty and R. L. Malghan, “Safeguarding against Cyber Threats: Machine Learning-Based Approaches for Real-Time Fraud Detection and Prevention,” Engineering Proceedings, vol. 59, no. 1, p. 111, 2023.

[13] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, “Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach,” Big Data and Cognitive Computing, vol. 8, no. 1, p. 6, 2024.

[14] A. A. Khalil, Z. Liu, and A. Ali, “Enhancing operational efficiency of insurance companies: a fuzzy time series approach to loss ratio forecasting in the Egyptian market,” Journal of Business

Analytics, pp. 1–19, 10.1080/2573234X.2024.2393609.

[15] M. Hanafy and R. Ming, “Improving imbalanced data classification in auto insurance by the data level approaches,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.

[16] B. Baesens, S. Höppner, I. Ortner, and T. Verdonck, “robROSE: A robust approach for dealing with imbalanced data in fraud detection,” Stat Methods Appt, vol. 30, no. 3, pp. 841–861, 2021, doi: 10.1007/s10260-021-00573-7.

[17] S. Subudhi and S. Panigrahi, “Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud,” in 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), 2018, pp. 528–531. 10.1109/ICDSBA.2018.00104.

[18] T. Olalekan Yusuf and A. Rasheed Babalola, “Control of insurance fraud in Nigeria: an exploratory study (case study),” J Financ Crime, vol. 16, no. 4, \pp. 418–435, Jan. 2009, doi: 10.1108/13590790910993744.

[19] R. Bhowmik, “Detecting auto insurance fraud by data mining techniques,” Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 4, pp. 156–162, 2011.

[20] K. Nian, H. Zhang, A. Tayal, T. Coleman, and Y. Li, “Auto insurance fraud detection using unsupervised spectral ranking for anomaly,” The Journal of Finance and Data Science, vol. 2, no. 1, pp. 58–75, 2016, doi: <https://doi.org/10.1016/j.jfds.2016.03.001>.

[21] G. Kowshalya and M. Nandhini, “Predicting Fraudulent Claims in Automobile Insurance,” in

2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1338–1343. 10.1109/ICICCT.2018.8473034.

[22] L. Goleiji and M. Tarokh, “Identification of influential features and fraud detection in the

Insurance Industry using the data mining techniques (Case study: automobile’s body insurance),” Majlesi J Multimed Process, vol. 4, pp. 1–5, 2015.

[23] S. Goundar, S. Prakash, P. Sadal, and A. Bhardwaj, “Health Insurance Claim Prediction Using Artificial Neural Networks,” International Journal of System Dynamics Applications (IJSDA), vol. 9, no. 3, pp. 40–57, 2020.

[24] J. Debener, V. Heinke, and J. Kriebel, “Detecting insurance fraud using supervised and unsupervised machine learning,” Journal of Risk and Insurance, vol. 90, no. 3, pp. 743–768, Sep. 2023, doi: https://doi.org/10.1111/jori.12427.

[25] A. Urunkar, A. Khot, R. Bhat, and N. Mudegol, “Fraud Detection and Analysis for Insurance Claim using Machine Learning,” in 2022 IEEE International Conference on Signal Processing,

Informatics, Communication and Energy Systems (SPICES), 2022, pp. 406–411. 10.1109/SPICES52834.2022.9774071.

[26] Y. Abakarim, M. Lahby, and A. Attioui, “A Bagged Ensemble Convolutional Neural Networks Approach to Recognize Insurance Claim Frauds,” Applied System Innovation, vol. 6, no. 1, 2023, doi: 10.3390/asi6010020.

[27] B. Xu, Y. Wang, X. Liao, and K. Wang, “Efficient fraud detection using deep boosting decision trees,” Decis Support Syst, vol. 175, p. 114037, 2023, doi: https://doi.org/10.1016/j.dss.2023.114037.

[28] S. Subudhi and S. Panigrahi, “Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection,” Journal of King Saud University - Computer and Information Sciences, vol. 32, no. 5, pp. 568–575, 2020, doi: <https://doi.org/10.1016/j.jksuci.2017.09.010>.

[29] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of Performance of Data Imputation Methods for Numeric Dataset,” Applied Artificial Intelligence, vol. 33, no. 10, pp. 913–933, Aug. \2019, doi: 10.1080/08839514.2019.1637138.

[30] G. G. Sundarkumar, V. Ravi, and V. Siddeshwar, “One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection,” in 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),

2015, pp. 1–7. doi: 10.1109/ICCIC.2015.7435726.