IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 3, May-June 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203115

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

Thatipally kavya , Udigiri Harshitha, Rajnish, K Raveendra Kumar

Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

ABSTRACT: Insurance fraud, particularly within the automobile insurance sector, is a significant challenge faced by insurers, leading to financial losses and influen cing 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 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 a challenge in fraud detection models.

.

I. 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.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203115

II. 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 system

III. 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.

IV. PROBLEM STATEMENT AND 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.

EXISTING SYSTEM DISADVANTAGES

1. Class Imbalance
2. Handling Missing Data
3. Overfitting
4. Limited Performance with Complex Fraud Patterns
5. Inefficient Fraud Detection

V. 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

ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203115

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.

VI. 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

1. Enhanced Fraud Detection
2. Effective Missing Data Handling
3. Optimized Model Performance
4. Reduced Overfitting
5. Increased Accuracy in Predictions

VII. APPLICATION

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.

VIII. FUTURE ENHANCEMENT

Future enhancements for vehicle insurance fraud detection using the proposed AdaBoost Classifier could significantly elevate the accuracy and efficiency of fraud detection systems. One key area for improvement is the integration of real-time analytics to detect fraudulent claims as they occur, leveraging the adaptive capabilities of the AdaBoost algorithm to refine predictive accuracy continuously. Additionally, expanding the data sources to include telematics data from vehicles, such as driving behavior, GPS logs, and real-time accident data, could provide deeper insights into fraudulent patterns, allowing the model to differentiate between genuine and suspicious claims more effectively.

IX. CONCLUSION

The conclusion of this project highlights the significant advancements achieved in vehicle insurance fraud detection by implementing the AdaBoost Classifier. The proposed model effectively addresses challenges associated with class imbalance and missing data, which are common in insurance datasets. By leveraging AdaBoost, the system enhances the detection of fraudulent claims with greater accuracy and reduced overfitting, ensuring a robust predictive model that adapts to various fraud patterns. This approach not only improves the precision of identifying fraudulent activities but also contributes to minimizing financial losses for insurance companies.

ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203115

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.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com