International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05837
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Using Policyholder and Incident Risk Factors in Predictive Modelling of Auto Insurance Claims

Dr. K. Satyam[1], Veera Leela Manikanta[2]

[1]Associate Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India.

[2] Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India.

ABSTRACT

Effective claim analysis is a major difficulty for insurance firms due to the substantial rise in auto insurance policies and claim requests brought about by the automobile industry's rapid growth. To reduce financial losses and increase operational effectiveness, accurate claim severity prediction and fraudulent claim detection are crucial. Using policyholder data, vehicle attributes, and incident-related variables, this study suggests a machine learning-based method for modelling auto insurance claims. Age, insurance type, premium amount, car information, accident severity, and claim costs are just a few of the many attributes included in the dataset.

To improve the quality of the data, preprocessing methods such as addressing missing values, encoding categorical variables, and feature normalisation were used. To examine trends in the dataset and forecast claim outcomes, including fraud detection, a number of classification algorithms were put into practice. The experimental findings show that, in comparison to conventional techniques, machine learning models may successfully detect high-risk claims and increase forecast accuracy. Insurance businesses can automate decision-making, identify fraudulent activity, and streamline claim management procedures with the help of the suggested system.

Keywords

Auto Insurance, Machine Learning, Claim Prediction, Fraud Detection, Risk Analysis, Classification Algorithms, Data Preprocessing, Predictive Modeling

I.INTRODUCTION

Due to the growing number of cars and policyholders globally, the auto insurance sector has grown quickly. Insurance firms now have a difficult time accurately and efficiently handling claims as a result of this expansion. The growing number of false claims and the challenge of determining claim severity based on several contributing elements, including policyholder behaviour, vehicle features, and accident conditions, are two of the fundamental issues in this field.

Conventional approaches to claim analysis mostly rely on manual research and rule-based systems, which are frequently laborious, prone to errors, and ineffective when handling massive amounts of data. These drawbacks emphasise the necessity of sophisticated, automated systems that can analyse intricate datasets and identify significant trends.

In several fields, including insurance, machine learning has become a potent tool for predictive analysis. Machine learning algorithms can find hidden links between factors and accurately forecast claim outcomes by utilising past data. These models can be applied to motor insurance to forecast the severity of claims, identify fraudulent activity, and support decision-making.

The goal of this project is to create a machine learning-based framework that uses policyholder demographics, vehicle data, and incident-related characteristics to model auto insurance claims. The suggested strategy seeks to increase overall claim management efficiency, decrease fraudulent claims, and improve prediction accuracy. The system offers a dependable and

scalable solution for contemporary insurance analytics by combining data pretreatment, feature selection, and classification algorithms.

II. PROBLEM STATEMENT

The intricate links between policyholder demographics, vehicle characteristics, and event conditions make it difficult for the auto insurance sector to effectively estimate claim severity and detect fraudulent claims. Conventional claim processing techniques frequently fall short of identifying hidden patterns in sizable and varied datasets because they rely on manual analysis and preset procedures. This results in erroneous claim evaluations, longer processing times, and monetary losses from fraud that goes unnoticed. As a result, a clever and data-driven strategy that can successfully evaluate several risk factors and produce precise forecasts for claim outcomes is required.

III. DATASET DESCRIPTION

Auto insurance records with comprehensive details on policyholders, cars, and incident reports make up the dataset used in this study. It is appropriate for machine learning-based analysis since it combines numerical and categorical characteristics.

Features like age, policy type, insurance premium, and umbrella limit are examples of policyholder-related characteristics. These factors aid in determining the customer's risk profile and insurance coverage. In order to examine behavioural patterns, demographic information such as gender, education level, occupation, interests, and relationship status are also included.

Features pertaining to incidents, including incident type, collision type, incident severity, authorities contacted, number of witnesses, and geographical details, are also included in the dataset. These characteristics are essential in identifying the type and severity of the mishap.

Total claim amount, injury claim, property damage claim, and vehicle damage claim are examples of financial-related characteristics that offer information about the financial impact of each incident. To evaluate vehicle-related risks, vehicle-specific characteristics such auto make, model, and manufacture year are also taken into account. This dataset's target variable, fraud_reported, reveals if a claim is fraudulent (Yes/No). Furthermore, classification jobs employ claim severity categories like minor damage, significant damage, and total loss. All things considered, the dataset offers a thorough understanding of insurance claims, facilitating efficient analysis and forecasting through machine learning methods.

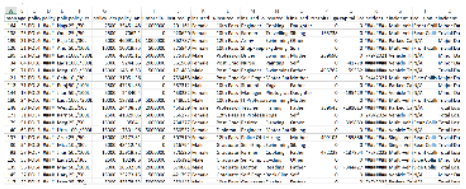

Fig: Dataset-1

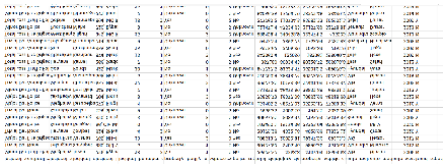

Fig: Dataset-2

IV. METHODOLOGY

The proposed system follows a structured machine learning pipeline to analyze auto insurance claims and predict fraudulent activities as well as claim severity. The methodology consists of multiple stages, including data preprocessing, feature selection, model development, and performance evaluation.

Data Preprocessing

In addition to some missing and inconsistent values, the dataset includes both numerical and category attributes. Preprocessing methods were used to guarantee the quality of the data. The proper imputation techniques, such as mean or mode replacement, were used to deal with missing variables. Categorical variables such as gender, policy type, occupation, and vehicle model were converted into numerical format using encoding techniques like Label Encoding or One-Hot Encoding. To ensure that every feature contributes equally during model training, feature scaling was also used to normalise numerical parameters like claim amount and premium values. To increase the accuracy and stability of the model, this phase is crucial.

Exploratory Data Analysis (EDA)

To comprehend the distribution of variables and find connections between features, exploratory data analysis was carried out. Patterns in claim amounts, incident severity, and fraud incidents were examined using statistical summaries and visualisations. The most important characteristics influencing claim prediction were found using correlation analysis. This stage directed the feature selection procedure and offered insightful information about the dataset.

Feature Selection

By eliminating unnecessary or redundant attributes, feature selection is essential for enhancing model performance. Based on their link with the target variable, significant characteristics such incident severity, total claim amount, number of witnesses, vehicle type, and policy premium were chosen. The model becomes more effective, lessens overfitting, and increases prediction accuracy by removing less important features.

Model Development

To forecast claim outcomes and identify fraudulent claims, a number of machine learning classification methods were put into practice. These consist of Support Vector Machine (SVM), Random Forest, Decision Tree, and Logistic Regression. The processed dataset was used to train each model, which was then optimised for improved performance. Because ensemble approaches like Random Forest can efficiently handle complicated relationships and heterogeneous data types, they outperformed other models.

Model Evaluation

Standard assessment criteria like accuracy, precision, recall, and F1-score were used to assess the models' performance. These metrics offer a thorough grasp of the model's efficacy in accurately categorising claims and identifying fraud. True positives, false positives, true negatives, and false negatives were also examined using a confusion matrix. The final model was chosen as the one with the best accuracy and balanced performance across all measures.

V. CONCLUSION

Using policyholder, vehicle, and incident-related risk indicators, this study offers a machine learning-based method for simulating auto insurance claims. In order to forecast claim outcomes and spot fraudulent activity, the suggested system efficiently examines intricate relationships within the information. Techniques for feature selection and data preprocessing greatly increased the quality of the input data and improved model performance.

When addressing heterogeneous data types and vast feature sets, ensemble approaches like Random Forest showed the best accuracy and dependability among the models that were put into practice. The findings show that machine learning can offer a scalable and effective way to automate claim analysis and minimise manual labour. All things considered, the suggested framework aids insurance firms in making better decisions, reducing monetary losses from fraud, and streamlining claim handling procedures. The system can be expanded in the future by including real-time data processing, deep learning techniques, and deployment as a web-based application for practical usage.

References

[1] J. Smith and A. Kumar, "Insurance Fraud Detection Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 12345–12356, 2020.

[2] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data Mining for Credit Card Fraud: A Comparative Study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] Kaggle, "Auto Insurance Fraud Detection Dataset," [Online]. Available: https://www.kaggle.com/. [Accessed: 2026].

[7] I. Witten, E. Frank, and M. Hall, "Data Mining: Practical Machine Learning Tools and Techniques," 3rd ed., Morgan Kaufmann, 2011.