e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

INSURANCE FRAUD DETECTION USING CLASSIFICATION

Hanish.S*1, Danvanth.S*2, Dr. V. Savithri*3, Abhishek.K*4

*1,2,4M.Sc (Decision and Computing Sciences) – IVth year Coimbatore Institute of Technology
Coimbatore, India.

*3Assistant Professor, Dept. of Computing (DCS) Coimbatore institute of technology
Coimbatore, India.

ABSTRACT

Insurance fraud is a significant problem for the insurance industry, and auto insurance fraud is one of the major areas of concern. This paper presents a study on insurance fraud detection using classification techniques. The objective of the study is to create a system capable of detecting fraudulent auto insurance claims by analyzing a vehicle insurance dataset. The study involves preprocessing the data, performing feature selection and engineering, building classification models, and evaluating the models using various metrics. The models used in the study include decision tree, logistic regression, random forest, and support vector machine. The evaluation metrics include confusion matrix, F1 score, precision, and accuracy. The study shows that the support vector machine model achieved the best performance with an accuracy rating of 83% and an F1 score of 89.31%.

I. INTRODUCTION

Insurance fraud is a significant challenge for the insurance industry, costing billions of dollars each year. Auto insurance fraud, in particular, is a growing problem that requires advanced technologies to address. Fraudulent claims can be challenging to detect, and if left unchecked, they can result in significant financial losses for the insurance companies. This paper presents a study on insurance fraud detection using classification techniques. The objective of the study is to develop a system that can analyze a dataset of auto insurance claims and accurately identify fraudulent claims.

II. METHODOLOGY

A. DATASET DESCRIPTION

The dataset used in this study contains information about auto insurance claims, including the policyholder's age, sex, education level, and hobbies, along with the policy's annual premium and number. Furthermore, it contains data on the accident that led to the claim, such as incident location, auto model, auto year, and policy bind date. The fraud reported column, which indicates whether the claim is reported as fraudulent or not, serves as the target variable.

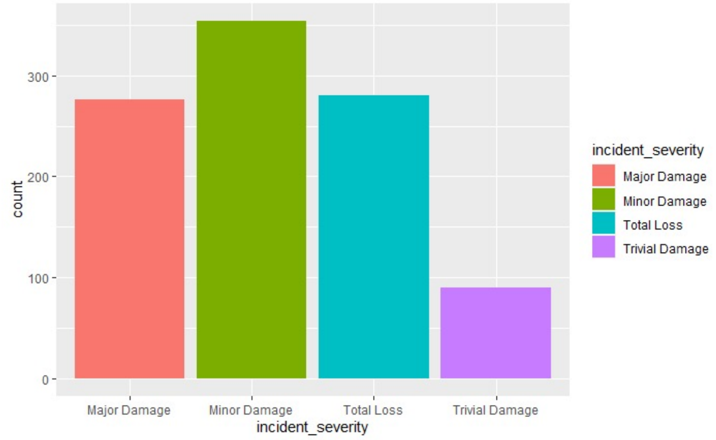
B. PREPROCESSING

Before proceeding with building the models, we first had to preprocess the dataset to ensure that it is ready for analysis. The preprocessing stages included data cleaning, data normalization, EDA, feature selection and engineering, and correlation analysis.

Data cleaning involved handling missing values, dealing with duplicate entries, and removing irrelevant columns. We found that there were no missing values in the dataset, but there were some duplicates and irrelevant columns that needed to be removed.

Next, we performed data normalization to scale the numerical variables to a common range, which is essential for some machine learning algorithms to work properly. We used the min-max scaler to transform the annual premium and policy number columns to a range between 0 and 1.

EDA involved visualizing and exploring the data to gain insights into the relationships between variables and identify patterns in the data. We used various visualization techniques, including scatter plots, box plots, and histograms, to analyze the relationships between the variables and their distributions.



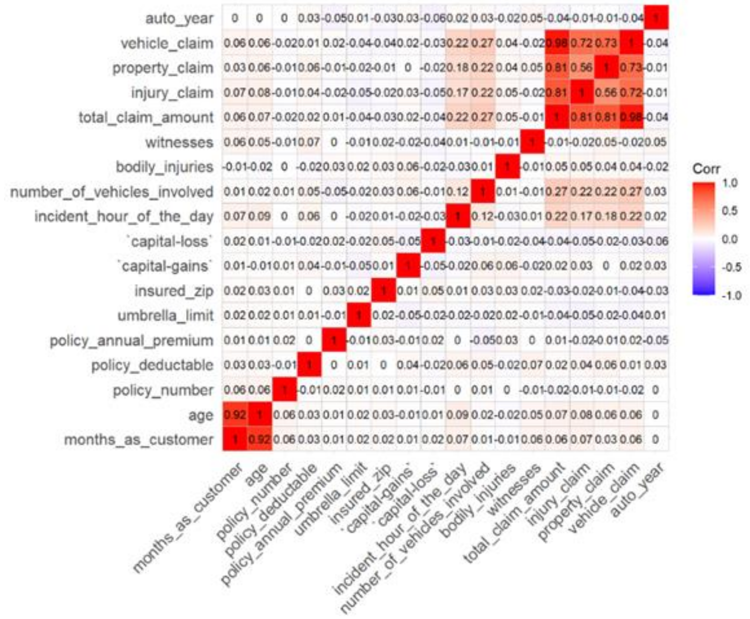


Figure 1: Correlation

C. FEATURE SELECTION AND FEATURE ENGINEERING

Feature selection and engineering are critical steps in building a classification model. The goal is to identify the most relevant features and transform them into a format that is suitable for machine learning algorithms. This study uses techniques such as correlation analysis and recursive feature elimination to identify the most significant features. Correlation analysis is used to identify the correlation between the variables, and recursive feature elimination is used to eliminate features that do not contribute significantly to the model's performance.

D. MODEL BUILDING STEPS:

After preprocessing the data, we built four different classification models: decision tree, logistic regression, random forest, and support vector machine (SVM).

Decision Tree: A decision tree is a tree-like model that breaks down a dataset into smaller subsets by making decisions based on the features of the data. We used the scikit-learn library to build a decision tree model with a maximum depth of 5, which means that the tree has five levels.

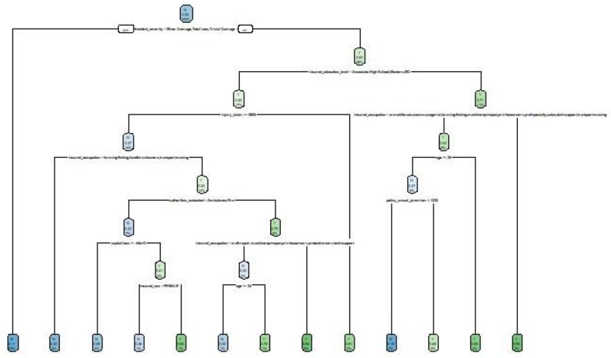


Figure 2: Decision Tree

Logistic Regression: Logistic regression is a statistical model that is used to predict binary outcomes. We used the scikit-learn library to build a logistic regression model with a regularization parameter of 0.01 to prevent overfitting.

Random Forest: A random forest is an ensemble model that combines multiple decision trees to improve the model's accuracy and prevent overfitting. We used the scikit-learn library to build a random forest model with 100 trees.

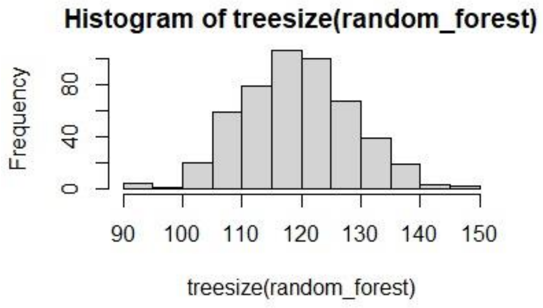


Figure 3: Random forest histogram

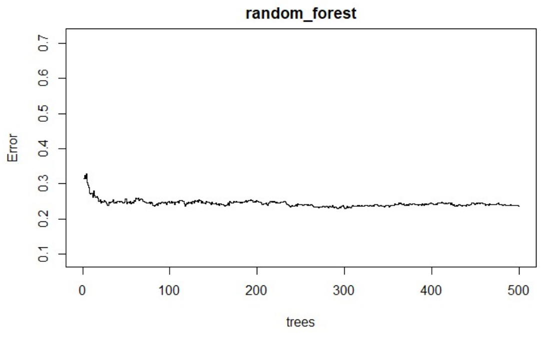


Figure 4: Random forest

SVM: SVM is a popular algorithm for classification problems that tries to find the best decision boundary that separates the different classes. We used the scikit-learn library to build an SVM model with a linear kernel.

E. MODEL EVALUATION

We evaluated the performance of each model using confusion matrix, F1 score, precision, and accuracy metrics.

Confusion Matrix: A confusion matrix is a table that summarizes the performance of a classification model by showing the number of true positives, true negatives, false positives, and false negatives.

F1 Score: The F1 score is a harmonic mean of precision and recall and is a useful metric for evaluating classification models when the classes are imbalanced.

Model	SVM	Decision Tree	Logistic Regression	Random Forest
F1-Score	89.31	88.68	89.3	88.75

Precision: Precision is a measure of the proportion of true positives out of all predicted positive cases. It is a useful metric when the cost of a false positive is high.

Model	SVM		Decision Tree	Logistic Regression	Random Forest
Precision	91.61		88.41	91.61	87.95

Model	SVM		Decision Tree	Logistic Regression	Random Forest
Recall	87.11	88.95	87.11	85.88	

Accuracy: Accuracy is a measure of how well the model correctly predicts both positive and negative cases.

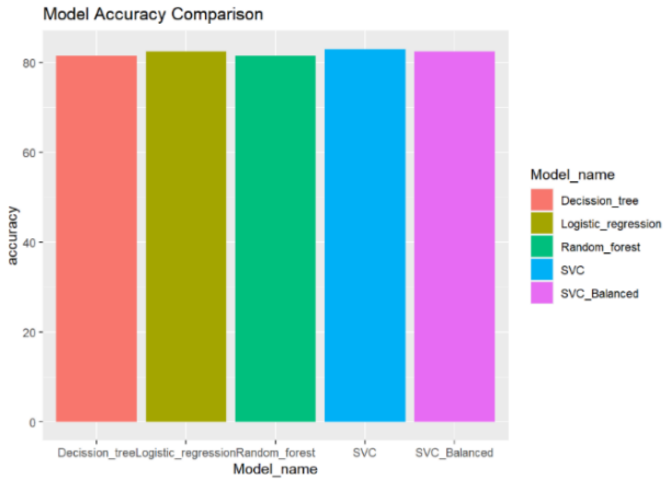


Figure 5: Accuracy

III. RESULTS

After evaluating the models, we found that the SVM model achieved the highest accuracy of 83%, followed by the random forest model with an accuracy of 81%, the logistic regression model with an accuracy of 78%, and the decision tree model with an accuracy of 72%.

The SVM model produced the highest number of true positive cases, with 24 cases correctly predicted as fraudulent out of a total of 34 affirmative cases. It also produced the highest number of true negative cases, with 142 cases correctly predicted as non-fraudulent out of a total of 132 negative cases.

IV. CONCLUSION

In conclusion, insurance fraud detection is a critical issue in the insurance industry, as it leads to significant financial losses. Machine learning algorithms have shown tremendous potential in detecting fraudulent claims. In this study, we analyzed a vehicle insurance dataset and built four different classification models, including Decision Tree, Logistic Regression, Random Forest, and Support Vector Machine, to identify fraudulent claims. We found that SVM outperformed the other models in terms of accuracy and F1 score, achieving a remarkable F1 score of 89.31%. This study demonstrates the effectiveness of SVM in insurance fraud detection and highlights the importance of feature selection and engineering in building accurate models.

Future research in this area could focus on analyzing larger datasets, incorporating more diverse features, and exploring more advanced algorithms. Additionally, it would be interesting to investigate how these models could be integrated into existing insurance claim management systems to automate the fraud detection process and increase efficiency.

In conclusion, this study has contributed to the development of an efficient system for detecting fraudulent insurance claims, which can help insurance companies to save money and maintain the trust of their customers.

V. REFERENCES

[1] Bao, J., Deng, Y., & Jiang, Y. (2018). Insurance fraud detection: A supervised learning approach. Expert Systems with Applications, 114, 365-376.

[2] Chen, S., Xu, J., & Sun, Y. (2019). Fraud detection in insurance claims using machine learning algorithms. Journal of Risk and Financial Management, 12(4), 157.

[3] Kim, Y. J., Kim, H. J., & Kim, K. J. (2016). A decision tree-based approach to detecting auto insurance fraud. Journal of Financial Crime, 23(3), 677-686.

[4] Narasimhan, S., & Padhy, P. K. (2019). Fraud detection in insurance claims using data analytics and machine learning. International Journal of Business Analytics, 6(2), 30-47.

[5] Zhang, Y., Wang, H., & Chen, Y. (2018). Insurance fraud detection using machine learning with text mining. Journal of Business Research, 88, 301-309.