Fraudulent Claim Detection

SYEDA FARHANA YASMIN

Introduction

Fraudulent insurance claims continue to be a major concern for insurers, driving significant financial losses and operational inefficiencies. The increasing volume and complexity of claims have rendered traditional manual fraud detection methods insufficient. In response, data-driven solutions have emerged as essential tools for enhancing fraud detection and prevention efforts.

This project, undertaken for Global Insure, leverages machine learning to modernize and strengthen the company's fraud detection framework. By systematically analyzing historical claims data and customer profiles, we developed a robust predictive model capable of classifying incoming claims as either fraudulent or legitimate. This early detection capability enables proactive intervention, reduces financial exposure, and enhances overall operational effectiveness.

Problem Statement

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimise financial losses and optimise the overall claims handling process.

Business Objectives

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.
Based on this assignment, you have to answer the following questions:

- How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
- Which features are most predictive of fraudulent behaviour?
- Can we predict the likelihood of fraud for an incoming claim, based on past data?
- What insights can be drawn from the model that can help in improving the fraud detection process?

Approach

- Load Data
- Clean Data : Handle Missing Values, Visualize Data
  - Special focus on NULL and ‘?’ values.
  - Identify the categorical and numerical columns and replace ‘?’ with ‘UNK’
- EDA :
  - Univariate Analysis
  - Corelation Analysis
  - Target Likelihood Analysis
- Test-Train Split
  - Feature Engineering
  - Data resampling for class imbalance
  - Apply VIF for checking redundant columns
  - Dummy variable creation and feature Scaling
- Feature Selection (RFECV for Top Predictors)
- Model Building
  - Logistic Regression : Find Optimal Cut off with ROC
  - Random Forest
- Evaluation Metrics and Insights: Accuracy, Precision and Recall.

Summary

Summarizing the main features of the data:

- Majority of customer with peak concentration is with Global Insure for 100 to 300 months.
- Majority of customers are between 30 to 45 years of age.
- Most customers have annual premium between 800 and 1200.
- The vast majority (672 cases) involved only 1 vehicle, indicating mostly single-vehicle incidents or claims.
- Most claims are for claim amount of 50K to 65K.
- Out of the total claims – ~500 claims are genuine and ~180 are fraudulent claims.
- Using Likelihood analysis and p_value analysis, most of the categorical features do not significantly affect the target variable. Only significant features are ‘incident_state’ and ‘authorities_contacted’.

Questions and Answers

Q1: How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
Response: To detect fraudulent claims in historical claim data, we need to start by exploring and preprocessing of the data, handling missing values and outliers. Identify key features such as claimant history, claim type, geographical location, and temporal patterns that may indicate fraud. Use exploratory data analysis (EDA) and visualizations to uncover unusual patterns, then apply machine learning models—such as Logistic Regression, Random Forest, or anomaly detection algorithms like Isolation Forest or Autoencoders—if the data is labeled or unlabeled. Evaluate the models using metrics like precision, recall, and AUC, and continuously monitor and update the system to adapt to evolving fraud patterns.

Q2: Which features are most predictive of fraudulent behaviour?
Response: Based on the data available it seems that total_claim_amount, authorities_contacted, incident_state and policy_premium play a vital role.

Q3: Can we predict the likelihood of fraud for an incoming claim, based on past data?
Response: Yes, Random Forest model built here with 99% recall rate can be used to predict the likelihood of fraud for an incoming claim based on past data.

Q4: What insights can be drawn from the model that can help in improving the fraud detection process?
Response: The Random Forest produced good results and the model is good to use for predicting fraudulent claims.

Conclusion

In this analysis, we compared the performance of two machine learning models—Logistic Regression and Random Forest—using recall as the evaluation metric, with a cut-off threshold of 0.3.

Logistic Regression demonstrated a strong performance with a train recall of 84.63%, but it achieved perfect recall of 100% on the test set. This suggests that while the model performs well on unseen data, there might be some overfitting on the training set, as it has a noticeable gap in recall performance between train and test datasets.

Random Forest, on the other hand, showed exceptional recall performance with a train recall of 99.05% and a perfect 100% test recall. The Random Forest model is able to generalize well on the test data, demonstrating both high accuracy and excellent recall performance.

Overall, both models performed well with high recall, particularly on the test set. However, the Random Forest model displayed more robust performance, with both high train and test recall, making it a potentially more reliable choice for this classification task.

Thank You