Report: Fraudulent Claim Detection in Insurance Claims

1. Problem Statement

Fraudulent insurance claims have become a significant problem for insurance companies, leading to considerable financial losses every year. Detecting fraudulent claims early in the approval process is crucial for minimizing these losses and ensuring smooth business operations. Traditionally, insurance companies rely on manual inspection to identify fraudulent claims, which is time-consuming, inefficient, and prone to human error.

Global Insure, a leading insurance company, seeks to leverage data-driven insights to automate fraud detection. The company wants to build a machine learning model that can classify claims as either fraudulent or legitimate based on historical data, reducing financial losses and improving the efficiency of the claims approval process.

Business Objective:

The goal is to predict which claims are likely to be fraudulent before approval, using features like claim amounts, customer profiles, and claim types, so the company can flag suspicious claims early.

---

2. Methodology

To approach the problem of fraud detection, we follow a structured methodology:

2.1 Data Collection and Preprocessing

The dataset used for this project contains historical insurance claim details, including information about the claims, customer profiles, and whether the claim was fraudulent or legitimate. The features include:

- Claim Amount: The total amount of the claim.

- Customer Profile: Demographic information like age and occupation.
- Claim Type: The type of claim (e.g., injury, damage, etc.).
- Claim Severity: How severe the claim is.
- Policy Duration: How long the customer has been with the insurance company.

Data preprocessing steps were performed to clean the data, handle missing values, and encode categorical variables to make them suitable for model building.

2.2 Exploratory Data Analysis (EDA)

To better understand the data, we performed Exploratory Data Analysis (EDA), which involved:

- Visualizing the distribution of key features (e.g., claim amounts, customer ages, etc.).
- Identifying correlations between features and the target variable (fraudulent vs. legitimate).
- Looking for patterns that could indicate fraudulent behavior.

2.3 Model Selection and Training

We selected two machine learning models for this project:

1. Logistic Regression: A simple, interpretable model suitable for binary classification tasks. It helps in understanding the relationship between the independent variables (features) and the dependent variable (fraudulent or legitimate).
2. Random Forest: A more complex model that uses an ensemble of decision trees to make predictions. It handles non-linear relationships better and can provide higher accuracy, though it is less interpretable than logistic regression.

Both models were trained on the dataset, and their performance was evaluated using various metrics, including accuracy, precision, recall (sensitivity), specificity, and F1 score.

---

3. Techniques Used

3.1 Data Preprocessing

- Missing Value Imputation: We handled missing data by either imputing or removing rows with missing values.
- Feature Encoding: Categorical variables were encoded into numerical values using techniques like one-hot encoding for variables like claim type and customer occupation.
- Scaling: Numerical features like claim amount and customer age were scaled to ensure they were on a similar scale, making the model more effective.

3.2 Model Building and Evaluation

We used two primary models to predict fraud:

- Logistic Regression: A statistical model that estimates the probability of an outcome. It is useful for predicting binary outcomes, such as fraud (1) or legitimate (0).
- Random Forest: An ensemble learning method that builds multiple decision trees and combines their results to improve prediction accuracy. Random Forest is particularly effective when the data has complex, non-linear relationships between features.

3.3 Performance Evaluation

Model performance was evaluated using the following metrics:

- Accuracy: The percentage of correct predictions out of all predictions.
- Sensitivity (Recall): The ability of the model to correctly identify fraudulent claims.
- Precision: The percentage of predicted fraudulent claims that were actually fraudulent.
- F1 Score: A balance between precision and recall.
- Specificity: The ability of the model to correctly identify legitimate claims.

---

4. Key Insights

4.1 Feature Importance

Through the evaluation of feature importance from the Random Forest model, we identified several key features that contributed significantly to detecting fraudulent claims:

- Claim Amount: Larger claims were more likely to be fraudulent.
- Incident Severity: Higher severity claims were associated with fraudulent activity.
- Customer Age and Profile: Certain age groups and customer profiles showed a higher likelihood of being associated with fraudulent claims.
- Claim Type: Some types of claims (e.g., injury claims) were more likely to be fraudulent than others.

4.2 Model Comparison

- Logistic Regression: This model provided an interpretable view of the relationships between features and the likelihood of fraud. However, its recall (61.22%) was lower compared to Random Forest, indicating it missed more fraudulent claims.
- Random Forest: The Random Forest model showed better performance in identifying fraudulent claims, with a recall of 75.51%, indicating it was better at capturing fraudulent claims. However, its interpretability was lower compared to Logistic Regression.

4.3 Business Implications

- Early Fraud Detection: Both models can be used to flag potentially fraudulent claims early in the approval process. This would reduce the chances of paying out fraudulent claims and help the company minimize financial losses.
- Operational Efficiency: Automating the fraud detection process would save time and resources, as manual inspections would only need to focus on flagged claims, rather than inspecting every claim manually.

---

5. Conclusion

The project demonstrated that machine learning models, particularly Logistic Regression and Random Forest, can effectively detect fraudulent insurance claims. While Random Forest showed better recall and overall performance, Logistic Regression offered more interpretability. This makes Logistic Regression a useful tool

for understanding feature importance, while Random Forest is better suited for production use where higher recall is needed.

Recommendations:

1. Deploy the Random Forest model for fraud detection in the production environment to maximize recall and minimize the risk of fraudulent payouts.
2. Monitor the model's performance continuously, adjusting thresholds and retraining periodically to adapt to evolving fraud patterns.
3. Enhance feature engineering by exploring additional features such as transaction history and behavioral data to improve the models further.

By implementing these strategies, Global Insure can significantly improve its fraud detection process, reduce financial losses, and optimize operational efficiency.

This concludes the overall approach and methodology for the fraud detection assignment.