Insurance Claim Fraud Detection

Introduction:

Developing a predictive model for insurance fraud detection offers significant business value. Fraud investigations currently rely on manual processes such as reviewing claims, calling claimants and conducting background checks, which are time-consuming and inefficient. These delays allow fraud to go undetected whereas legitimate claims face unnecessary scrutiny. Predictive modelling enables early identification of high-risk claims, streamlining fraud investigations, reducing financial losses and improving operational efficiency. It also enhances customer experience by expediting legitimate claims. Ultimately, an effective fraud detection model leads to better decision-making, optimised resource allocation and increased profitability.

Objective:

The objective is to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features such as claim amounts, customer profiles, claim types and approval times, the company aims to predict the claims that are likely to be fraudulent before they are approved.

Data source is insurance_claims.csv

Target Variable identified is fraud_reported (Y/N).

Assumptions:

- The historical dataset accurately represents the characteristics of fraudulent claims.
- All missing values, inconsistencies, and outliers have been addressed correctly during preprocessing.
- The categorical encoding and feature transformation applied to the training data are consistent across validation/testing data.
- The model's performance on historical data generalizes well to future unseen claims.
- The timestamp fields like incident_date and policy_bind_date are assumed to be in a format that allows correct extraction of useful time-based features.

Data Preprocessing:

Redundant data, whether in the form of entire columns or specific values within columns,can negatively impact the performance, interpretability and efficiency of your machine learning models.

- Removed missing or empty data, duplicate or highly correlated features, hadled low variance data and features with low predictive power.

- Converted categorical variables using One-Hot Encoding.

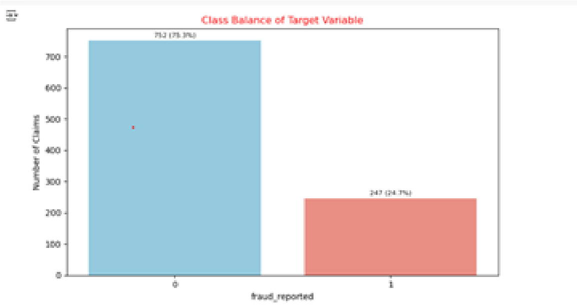
- Transformed date columns (policy_bind_date, incident_date) into numerical features such as year and month.

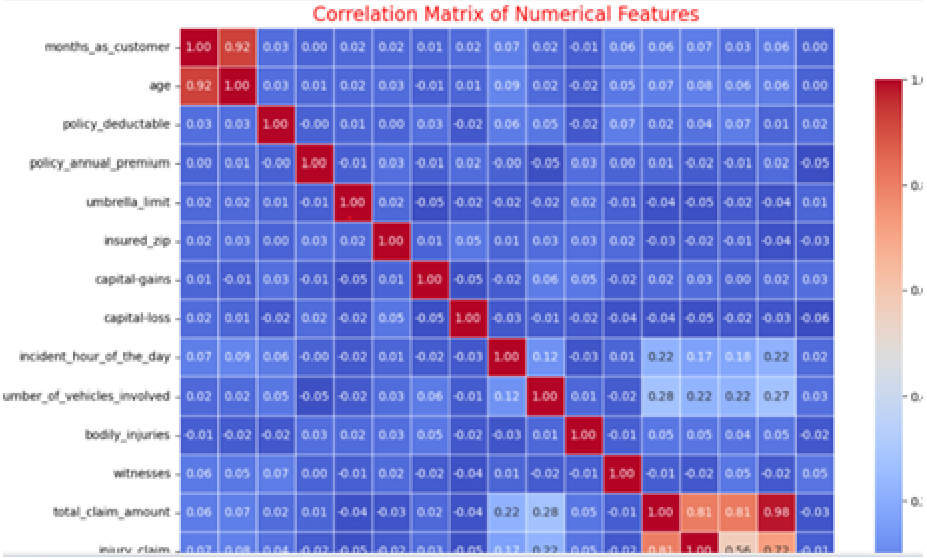
- Feature scaling was not required due to the nature of algorithms used.

- Dataset split into training, testing, and validation sets.

EDA:

Visualized data using bar plots, heatmaps and correlation matrices





- Class imbalance observed: ~13% of claims are fraudulent.

- High-value or exotic vehicles

- Certain cities and customer demographics

- Prior claims and high incident severity

Model Building:

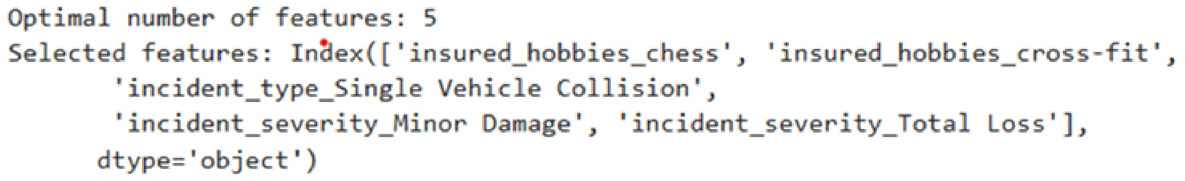
Feature Engineering and Selection:

Creating new features enhances the information available to the model, helping it capture complex relationships and patterns that may not be obvious from the original data. Well-designed features can introduce domain knowledge and improve the model's ability to distinguish between classes, leading to better performance.RFECV is an extension of Recursive Feature Elimination (RFE) that automates feature selection by using cross-validation to determine the optimal number of features. It iteratively removes the least important features while refitting a model, ensuring that only the most relevant features

Used Recursive Feature Elimination with Cross Validation (RFECV) for selection.

Evaluated multicollinearity using VIF and p-values.

Engineered new features like time difference between binding and incident dates.



```
Optimal number of features: 5
Selected features: Index(['insured_hobbies_chess', 'insured_hobbies_cross-fit',
       'incident_type_Single Vehicle Collision',
       'incident_severity_Minor Damage', 'incident_severity_Total Loss'],
      dtype='object')
```

The number of features selected is tuned automatically by fitting an **RFE** selector on the different cross-validation splits (provided by the cv parameter). The performance of the **RFE** selector are evaluated using scorer for different number of selected features and aggregated together. Finally, the scores are averaged across folds and the number of features selected is set to the number of features that maximize the cross-validation score.

Logistic Regression:

Optimal cutoff identified using Youden's J statistic- Youden's J statistic, also known as Youden's Index, is a simple and effective metric used to evaluate the performance of a binary classification model. It helps in finding the optimal cutoff threshold for classifying predictions, especially in imbalanced datasets like fraud detection.

Accuracy = 0.82

Sensitivity=0.62

Specificity=0.88

Random Forest Classifier:

Random Forest is one of the most popular and powerful machine learning algorithms used for both classification and regression tasks. It works by building multiple decision trees and combining their outputs to improve accuracy and control overfitting. While Random Forest is already a robust model fine-tuning its hyperparameters such as the number of trees, maximum depth and feature selection can improve its prediction.

GridSearchCV is used for Hyperparameter tuning. Hyperparameters are

n_estimators=200, max_depth=20, etc.

Accuracy: 0.91,Sensitivity: 0.80, Specificity: 0.93

Used the following evaluation metrics:

Confusion Matrix:

This matrix enabled us to visualize misclassification rates and understand the model's strengths and weaknesses in distinguishing fraud.

True Positives (TP): Correctly identified fraudulent claims.

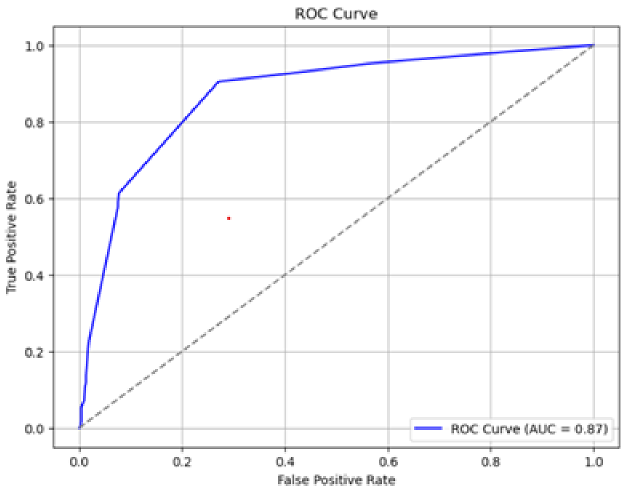
True Negatives (TN): Correctly identified genuine claims.

False Positives (FP): Genuine claims incorrectly flagged as fraudulent.

False Negatives (FN): Fraudulent claims missed by the model.

ROC Curve and AUC Score:

The ROC curve plotted the true positive rate(recall) against the false positive rate.



The AUC score quantified the model’s ability to distinguish between classes. AUC values close to 1.0 indicated best performance.

Random forest achieved the highest AUC, indicating robust discriminative power.

Precision measures how many flagged claims were actually fraudulent (minimizing false positives).

Recall measures how many actual frauds were detected (minimizing false negatives).

F1 Score is the harmonic mean of precision and recall, offering a balance between them.

Random Forest showed:

High recall, crucial in fraud detection to avoid missed frauds.

Balanced precision, reducing unnecessary investigation of genuine claims.

Strong F1 score, making it the most reliable model overall.

Predictions:

The Random Forest model was applied to the validation dataset.

Predictions were made using an optimal probability cutoff derived from Youden’s J statistic, improving classification performance over the default 0.5 threshold

Key Insights:

- Fraudulent claims often exhibit distinct patterns in terms of claim timing, incident severity, and policy details.
- Behavioural features (e.g., insured relationship, hobbies, claim amount) are used to identify suspicious activity.
- The most predictive features are incident_severity, insured_relationship, property_damage, incident_city and auto_make
- Robust against overfitting and capable of handling both numerical and categorical data.
- Feature importance from the model provided interpretable insights into fraud indicators.
- Handling missing values, encoding categorical data, and transforming timestamps significantly boosted model performance.
- Multicollinearity checks and RFECV helped refine the model to include only the most informative features.

Conclusion:

Random forest model is effective model for fraud detection. Careful preprocessing and feature selection improved model performance. This solution enables early detection of fraud, reducing financial loss and improving operational efficiency.

By,

Jyothsna Devi Duggasani

Bikash Sarkar