Fraudulent Claim Detection

Submitted by:
Fresnel Fabian

PROBLEM STATEMENT

Outline

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 Use Case

GlobalInsure wants to build amodel 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

1. Data Preparation
2. Data Cleaning
3. Train-Validation Split
4. EDA on Training Data
5. Feature Engineering
6. Model Building
7. Prediction and Model Evaluation
8. Insights

1. Data Preparation

- Load and inspect data

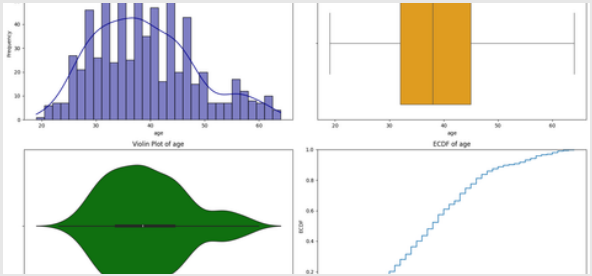
2. Data Cleaning

- Replace ‘?’ in some columns with nan value
- Null values in collision_type was filled with the mode.
- Null values in authorities_contacted, property_damage, police_report_available was filled with ‘No’
- No duplicate values found. Dropped column ‘_c39’ (completely empty
- Dropped umbrella_limit and capital-loss due to negative values
- Dropped policy_number, incident_location, insured_zip, policy_annual_premium, policy_bind_date
- Converted incident_date to datetime

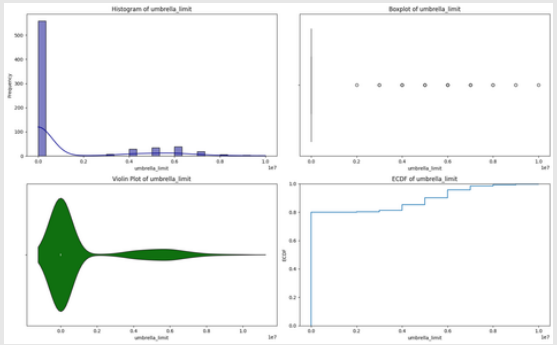
3. Train Test Validation Split

- Split the data into 70% train and 30% test by stratifying target variable due to imbalance

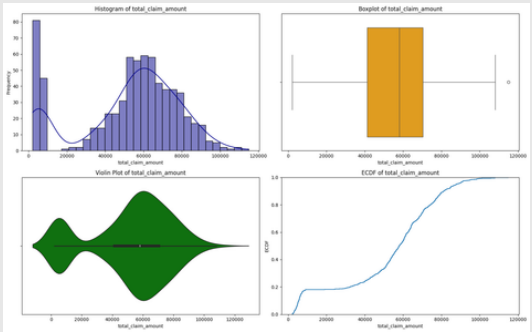
4. EDA on Training Data



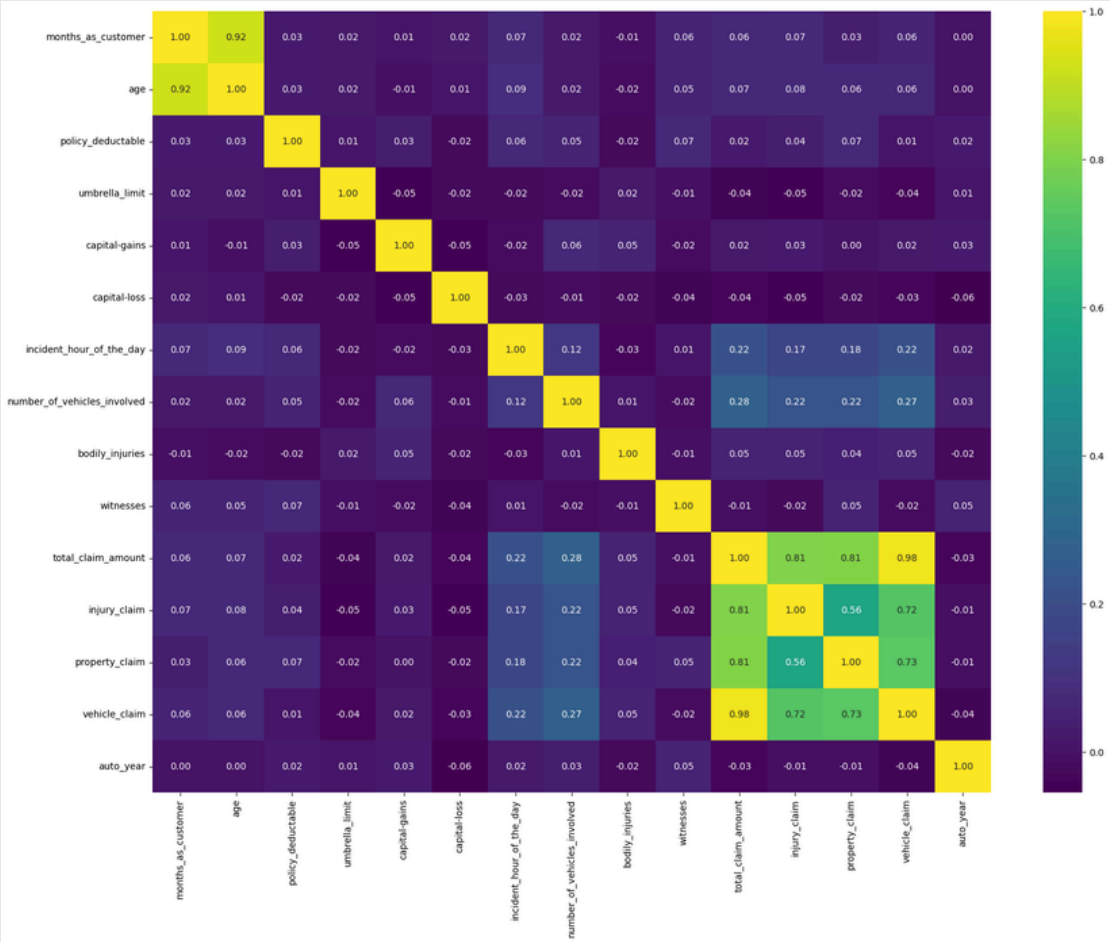
80% of the customers are between the age of 30 and 50 with the peak at 35 to 40



90% of the customers didn't take extra umbrella _limit(additional layer of liability coverage

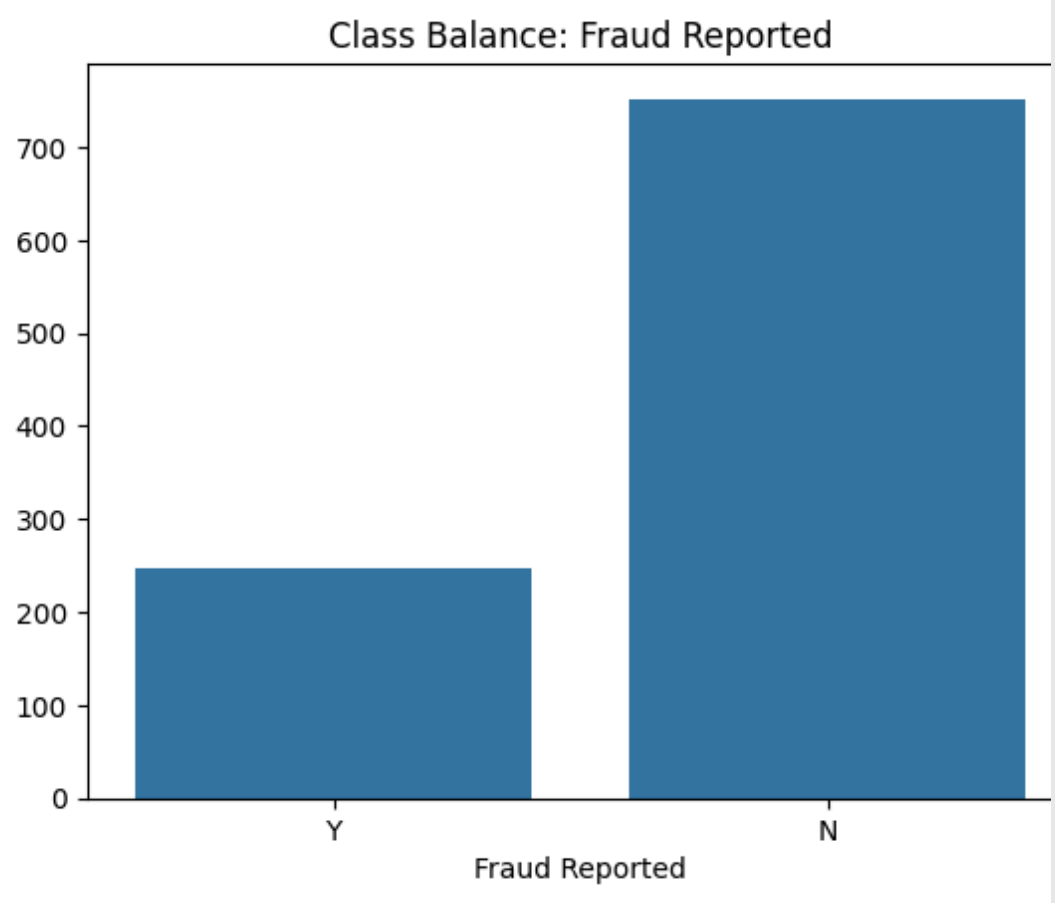


80% claims are between 40,000 and 80,000



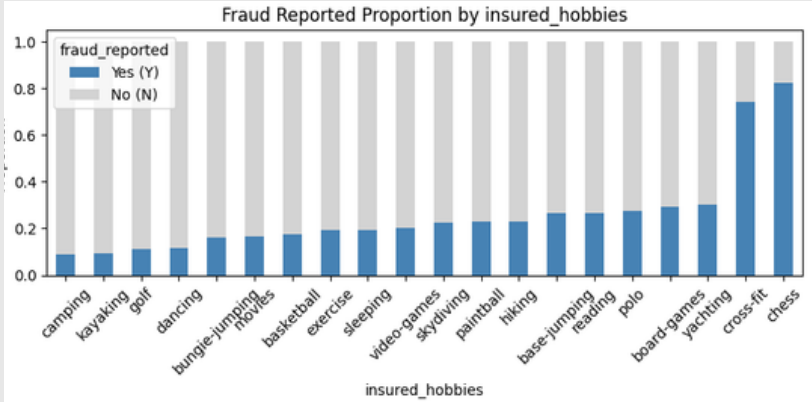
Insights from correlation matrix

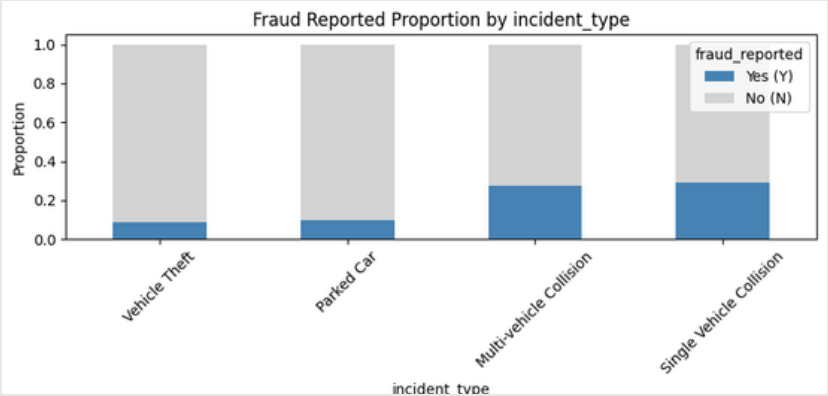
- Multi-collinearity between total_claim_amount and (injury_claim, property_claim and vehicle_claim
- High correlation between Age and months_as_customer



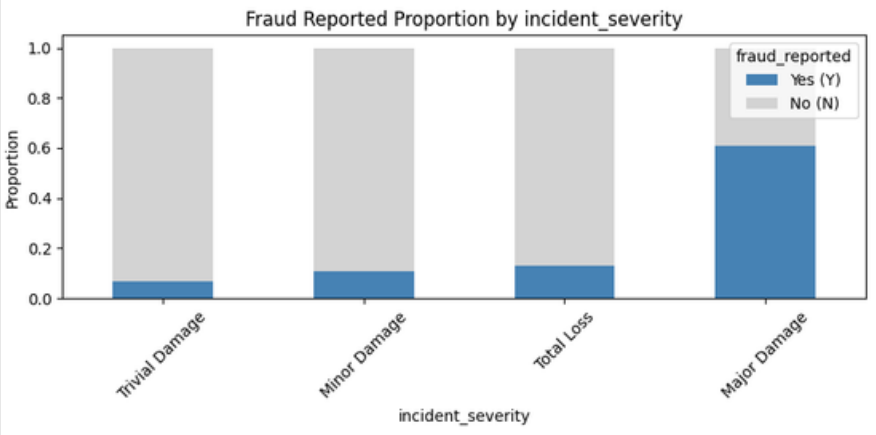
Class Balance bar chart

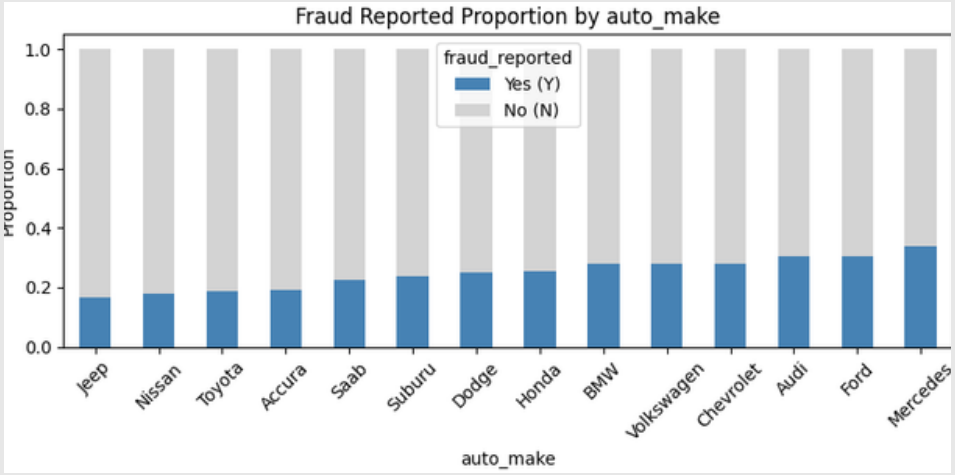
- Only a small fraction of the insurance claims are fraudulent 24.7%. The rest 75.3% are non-fraudulent claims.
- There is high imbalance in the data



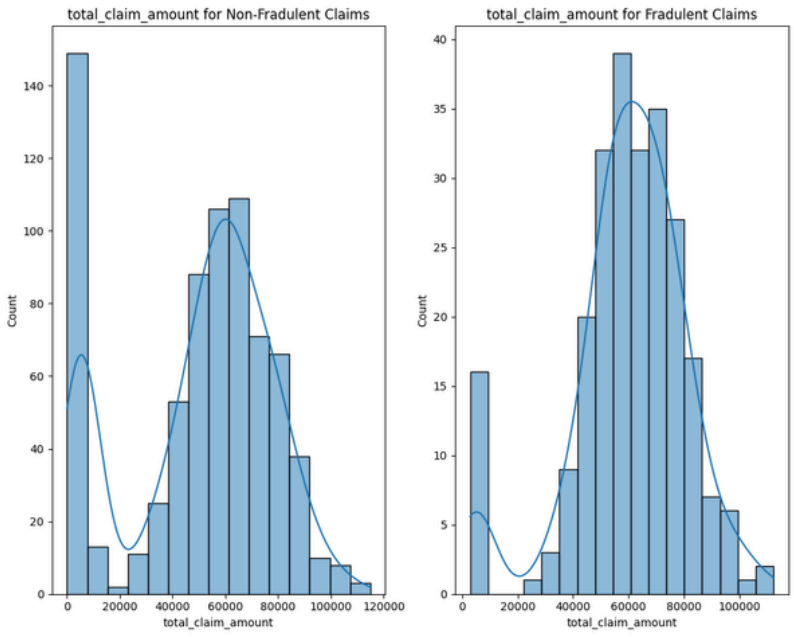


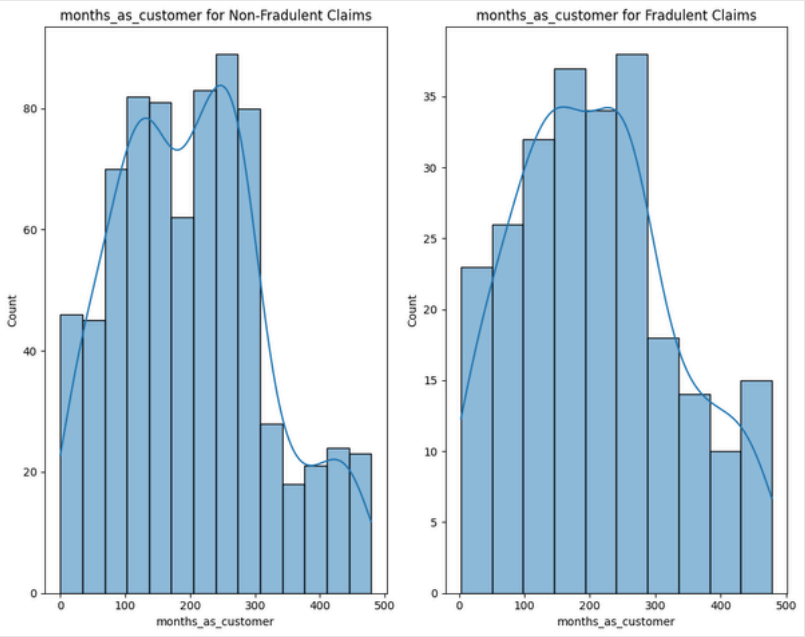
- Customers with certain hobbies show high fraudulent claims
- Fraudulent claims are more in single-vehicle and multi-vehicle collision





- High fraudulent claims for high severity incidents
- High fraudulent claims for certain automakers



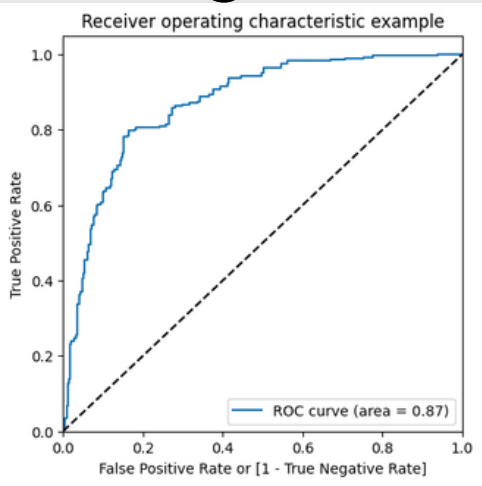


- Fraudulent claims are high for claims with high claim amount.
- Fraudulent claims are high for customer who have short tenure with the company.

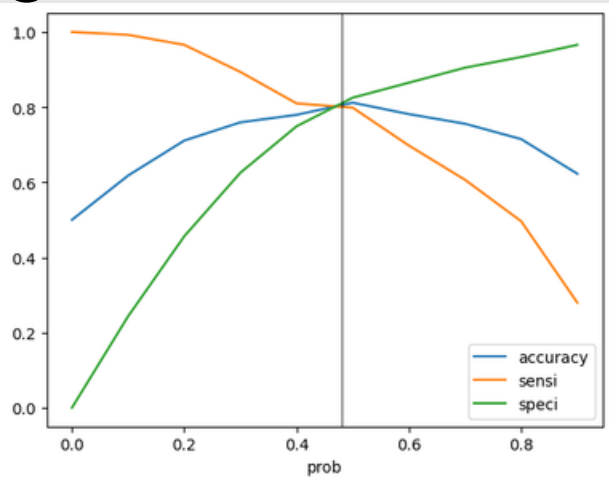
4. Feature Engineering

- Used oversampling to handle class imbalance
- Created new columns injury_claim_ration, property_claim_ratio, vehicle_claim_ration, sum_sub_claims, claim_diff, customer_risk_score
- Handled redundant columns total_claim_amount, auto_year, incident_hour_of_the_day, umbrella_limit, auto_model, incident_date, age as these variables contribute minimal information towards the prediction
- Combined Hobbies and car brands into categories
- Dummy variables
- Feature scaling using StandardScaler

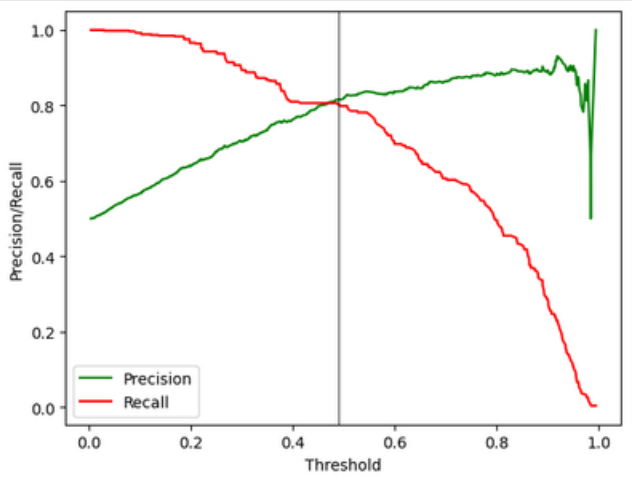
5.1 Logistic Regression Model Build



ROC Curve



Accuracy, Sensitivity & Specificity



Precision & Recall

- The roc curve with 0.87 score indicate strong predictive power
- The optimal threshold is close to 0.5. If we want to focus on catching fraud by minimizing false negatives we can favor sensitivity over specificity
- In the precision vs recall graph we can clearly see that when precision increases the recall drops. We can chose a lower threshold for high recall as missing fraud is more risky

5.1 LR Model Evaluation

True Negative	:	434
True Positive	:	420
False Negative	:	106
False Positve	:	92
Accuracy	:	0.8118
Sensitivity	:	0.7985
Specificity	:	0.8251
Precision	:	0.8203
Recall	:	0.7985
True Positive Rate (TPR)	:	0.7985
False Positive Rate (FPR)	:	0.1749
F1 Score	:	0.8092

Training prediction scores

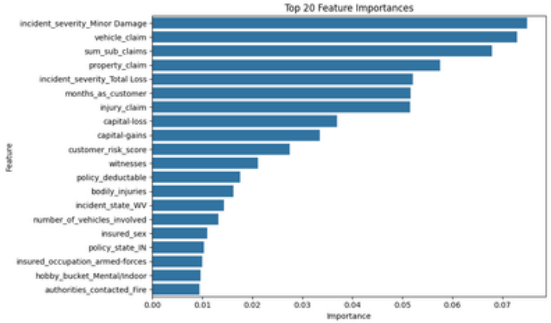
True Negative	:	427
True Positive	:	424
False Negative	:	102
False Positve	:	99
Accuracy	:	0.8089
Sensitivity	:	0.8061
Specificity	:	0.8118
Precision	:	0.8107
Recall	:	0.8061
True Positive Rate (TPR)	:	0.8061
False Positive Rate (FPR)	:	0.1882
F1 Score	:	0.8084

Test prediction scores

5.2 Random Forest Model Build

```
RandomForestClassifier
RandomForestClassifier(max_depth=15, max_features=11, min_samples_leaf=5,
                       n_estimators=50, n_jobs=-1, random_state=30)
```

- The best random forest model after hyper-parameter tuning using GridSearchCV has a max_depth of 15 and max_feature of 11 and 50 estimators.



Top 20 Features by importance

5.1 Random Forest Model Evaluation

True Negative	:	492
True Positive	:	513
False Negative	:	13
False Positve	:	34
Accuracy	:	0.9553
Sensitivity	:	0.9753
Specificity	:	0.9354
Precision	:	0.9378
Recall	:	0.9753
True Positive Rate (TPR)	:	0.9753
False Positive Rate (FPR)	:	0.0646
F1 Score	:	0.9562

Training prediction scores

True Negative	:	166
True Positive	:	56
False Negative	:	18
False Positve	:	60
Accuracy	:	0.74
Sensitivity	:	0.7568
Specificity	:	0.7345
Precision	:	0.4828
Recall	:	0.7568
True Positive Rate (TPR)	:	0.7568
False Positive Rate (FPR)	:	0.2655
F1 Score	:	0.5895

Test prediction scores

Evaluation & Conclusion

- How can we analyze historical claim data to detect patterns that indicate fraudulent claims?
  a. Exploratory Data Analysis (EDA) : Univariate and bivariate analysis revealed that fraudulent claims are more common among customers with shorter tenure, higher claim amounts, and certain incident types/severities. Correlation analysis helped identify multicollinearity and redundant features.
  b. Target Likelihood Analysis : By comparing fraud rates across categorical features (e.g., incident_severity, insured_hobbies), we identified which categories are associated with higher fraud likelihood.
  c. Feature Engineering : Creating ratios (e.g., injury_claim_ratio) and buckets (e.g., hobby_bucket) exposed hidden patterns and improved model interpretability.

- Which features are the most predictive of fraudulent behaviour?
  a. Feature Importance (Random Forest) : Top predictors include insured_occupation, insured_hobbies, incident_type, incident_severity, authorities_contacted, incident_state, months_as_customer, umbrella_limit, injury_claim, property_claim, and vehicle_claim.
  b. Logistic Regression Coefficients : Features with significant coefficients and low p-values (e.g., incident_severity_Major_Damage, hobby_bucket_Mental/Indoor, injury_claim_ratio) strongly influence fraud prediction.
  c. Categorical Buckets : Certain hobbies (chess, cross-fit) and luxury auto makes are linked to higher fraud rates.

- Based on past data, can we predict the likelihood of fraud for an incoming claim?
  a. Model Performance : Both Logistic Regression and Random Forest models achieved high accuracy, recall, and F1 scores on validation data, indicating strong predictive capability.
  b. Resampling : Addressing class imbalance with RandomOverSampler improved the model's ability to detect minority (fraudulent) cases.
  c. Cutoff Optimization : ROC and precision-recall analysis enabled selection of optimal probability thresholds, balancing sensitivity and specificity.

- What insights can be drawn to improve fraud detection processes?
    a. Process Recommendations : Focus manual review on claims with high-risk features (e.g., high claim ratios, major incident severity, certain hobbies/occupations).
    b. Continuous Model Updating : Regularly retrain models with new data to adapt to evolving fraud patterns.
    c. Feature Engineering : Use engineered features (ratios, buckets) to enhance detection and reduce false positives.
    d. Automation : Integrate predictive models into claims workflow for early flagging and prioritization of suspicious claims.
- Summary

A combination of thorough EDA, feature engineering, and robust modeling enables effective detection of fraudulent claims. Key features and engineered variables drive model performance, and continuous monitoring and updating of models are essential for maintaining high fraud detection


Thank
You