Fraudulent Insurance Claim Detection

Problem Statement

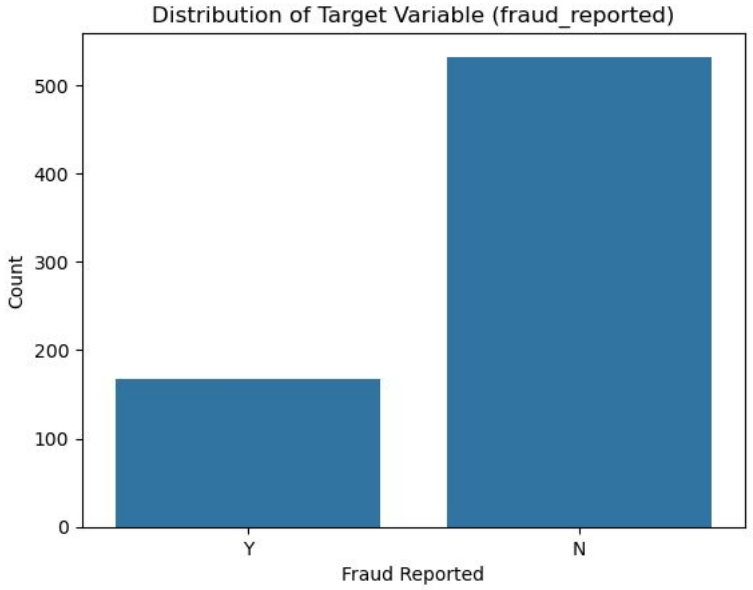
Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which 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.

Dataset & Assumptions

- 1000 records, multiple categorical & numerical features
- Cleaned invalid entries like `'?'`
- Dropped irrelevant columns (`policy_number`, `incident_location`, etc.)
- Assumption: `fraud_reported` is the target (binary classification)

Exploratory Data Analysis (EDA)

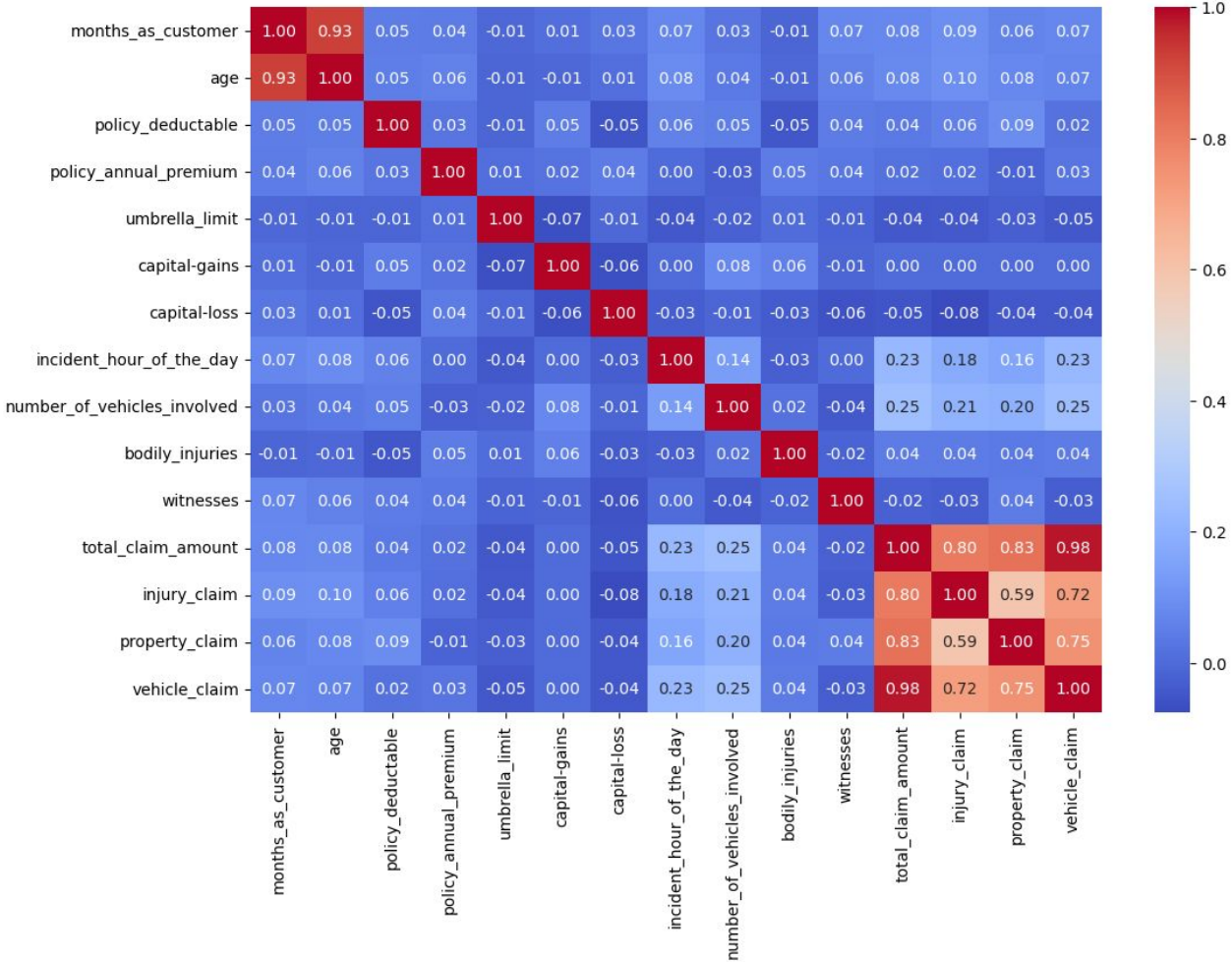
Observed Class imbalance in the target variable



Methodology

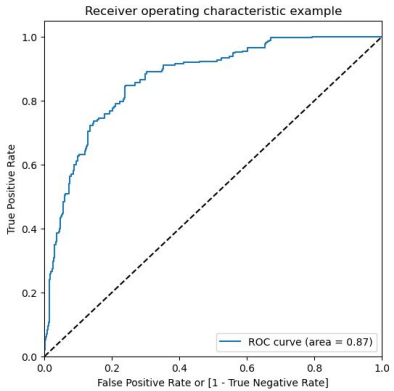
- Preprocessing: scaling, encoding, feature engineering
- Models: Logistic Regression, Random Forest
- Feature Selection:
- Logistic: RFECV
- Random Forest: Feature importance
- Tuning: GridSearchCV (Random Forest)

Correlation Map of Numerical Variables



Model Performance

Metric	Logistic Regression	Random Forest
Train Accuracy	78%	93%
Test Accuracy	73%	77%



Key Insights

- Fraud cases are associated with higher claim amounts
- `incident_type`, `insured_relationship`, and `authorities_contacted` are important predictors
- Random Forest handles non-linear relationships better than Logistic Regression

Recommendation

- Use Random Forest model in production
- Monitor performance over time (data drift, fraud strategy changes)
- Business impact: reduce false positives and detect more genuine frauds

Business Implications

- Reduced fraud losses = increased profits
- Faster, automated claim approval for genuine users
- Data-driven fraud detection strategy enhances credibility

Q&A

Q1: How can we analyse historical claim data to detect patterns that indicate fraudulent claims?

Answer:

- By performing Exploratory Data Analysis (EDA) on historical claims, we identified suspicious trends
- Visual tools like boxplots, likelihood analysis, and correlation heatmaps help in spotting anomalies.
- Using machine learning, we trained models on past data to learn patterns associated with fraud.

Q&A

Q2: Which features are most predictive of fraudulent behaviour?

Answer:

- Random Forest:
  - incident_severity_Minor Damage
  - property_claim, vehicle_claim, injury_claim
  - months_as_customer, umbrella_limit
  - incident_day, capital-gains, capital-loss
  - incident_week_of_year, witnesses

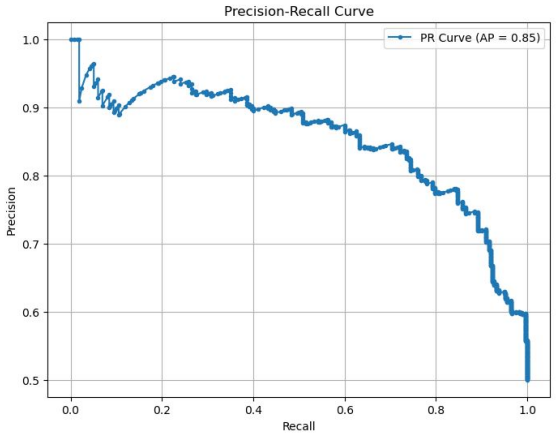
- Logistic Regression:
  - incident_month, incident_week_of_year
  - vehicle_claim, injury_claim
  - incident_severity_Trivial Damage, property_damage_Unknown/YES
  - incident_severity_Minor Damage, Total Loss
  - insured_education_level_High School, umbrella_limit

Q&A

Q3: Can we predict the likelihood of fraud for an incoming claim, based on past data?

Answer:

- Yes, Our trained models can predict the probability of fraud for a new claim
- The Random Forest model, tuned using GridSearchCV, achieved:
  - Train Accuracy: 93%
  - Test Accuracy: 77%
- Logistic Regression also showed decent predictive performance.



Q&A

Q4: What insights can be drawn from the model that can help in improving the fraud detection process?

Answer:

- Certain patterns like high claim amount + single-vehicle accident + police involvement → higher fraud likelihood
- Models can be integrated into the claim processing pipeline to flag suspicious claims in real time
- Helps prioritize manual review for only high-risk claims → improves operational efficiency