Fraudulent Claim Detection

Rupam Mallick
Muthumariappan B

Problem Statement

- Global Insure faces significant financial losses due to fraudulent claims.
- Current fraud detection relies on manual inspections, leading to:
  - Time-consuming processes
  - Late detection (after payouts)
  - Inefficiency

Objective

Build a predictive model to classify insurance claims as fraudulent or legitimate using:

- Historical claim data (claim amounts, types)
- Customer profiles
- Other relevant features

Key Questions to Address:

1. Pattern Analysis – How can we detect fraud indicators in historical claims?
2. Predictive Features – Which factors best predict fraud?
3. Fraud Likelihood – Can we score new claims for fraud risk before approval?
4. Actionable Insights – How can the model improve fraud detection?

Pattern Analysis

How can we detect fraud indicators in historical claims?

We can analyze historical claim data to detect patterns by performing Exploratory Data Analysis (EDA). This involves:

- Univariate Analysis: Examining the distribution of individual features (both numerical and categorical) to understand their characteristics and identify any unusual patterns.
- Correlation Analysis: Investigating the relationships between numerical features to identify potential dependencies.
- Bivariate Analysis: Exploring the relationships between features and the target variable (fraud_reported) to understand how different feature values influence the likelihood of a claim being fraudulent. This is done for both numerical and categorical features.

Predictive Features

Based on the Random Forest models feature importance scores, the most predictive features of fraudulent behavior are:

- incident_severity_Minor Damage
- insured_hobbies_chess
- incident_severity_Total Loss
- witnesses
- incident_severity_Trivial Damage
- umbrella_limit
- property_damage_NO
- incident_state_WV
- policy_state_OH
- collision_type_Front Collision
- policy_csl_250/500
- policy_csl_500/1000
- collision_type_Rear Collision
- incident_state_NY
- authorities_contacted_Other
- insured_relationship_other-relative
- insured_relationship_own-child
- insured_education_level_MD
- insured_relationship_unmarried
- insured_education_level_JD
- insured_relationship_not-in-family
- auto_make_Audi

- insured_education_level_PhD
- insured_occupation_exec-managerial
- auto_make_Dodge
- insured_occupation_transport-moving
- insured_occupation_armed-forces
- insured_occupation_prof-specialty
- incident_city_Northbrook
- incident_state_VA
- insured_occupation_tech-support
- auto_make_Nissan
- insured_occupation_craft-repair
- insured_occupation_machine-op-inspct
- insured_occupation_priv-house-serv
- auto_make_BMW
- insured_occupation_sales
- insured_occupation_adm-clerical
- auto_make_Other
- insured_occupation_farming-fishing
- insured_hobbies_skydiving
- insured_hobbies_bungie-jumping
- insured_hobbies_yachting

Fraud Likelihood

- We can predict the likelihood of fraud for an incoming claim based on past data.
- By training machine learning models (like Logistic Regression and Random Forest) on the historical claim data, we can use these models to predict the probability of an incoming claim being fraudulent.
- The models learn patterns and relationships from the historical data that are indicative of fraudulent behavior.

Actionable Insights

- Focus on key features: The most important features identified by the models (e.g., incident severity, insured hobbies, number of witnesses, umbrella limit, etc.) should be prioritized in the manual review process. Claims with suspicious values or combinations of these features should be flagged for closer inspection.
- Automated flagging: The trained model can be used to automatically flag claims with a high predicted probability of fraud. This can help streamline the initial screening process and reduce the workload on manual reviewers.
- Identify unusual patterns: The analysis of feature importance and the relationships between features and the target variable can help identify unusual patterns or anomalies that may indicate fraudulent activity. These patterns can be used to refine fraud detection rules or develop new fraud indicators.
- Continuous monitoring and retraining: The model should be continuously monitored for performance and retrained periodically with new data to adapt to evolving fraud patterns.

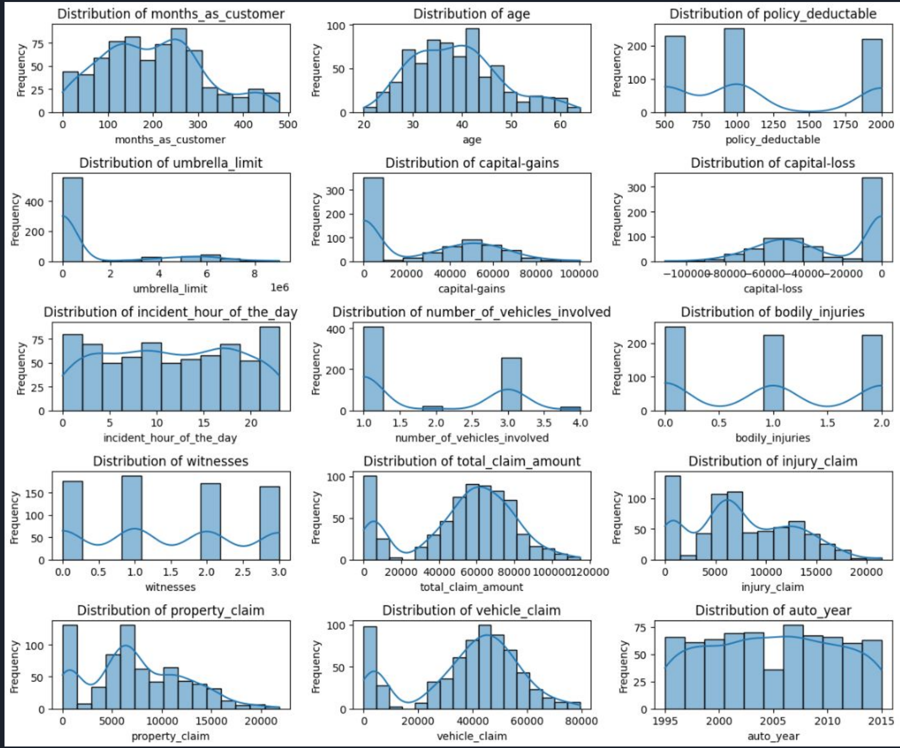
EDA on Training Data

Objective: Explore and visualize the training data to understand feature distributions, relationships, and identify patterns related to fraudulent claims.

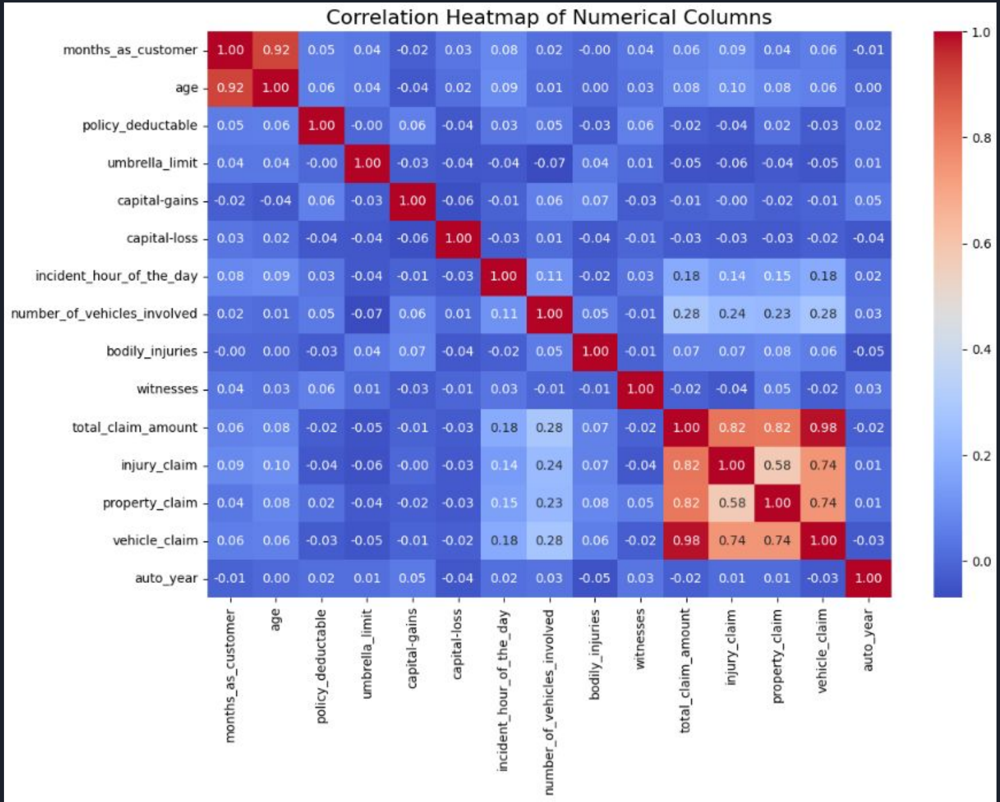
Actions:

- Univariate Analysis
- Correlation Analysis
- Check Class Balance
- Bivariate Analysis

EDA - Univariate Analysis

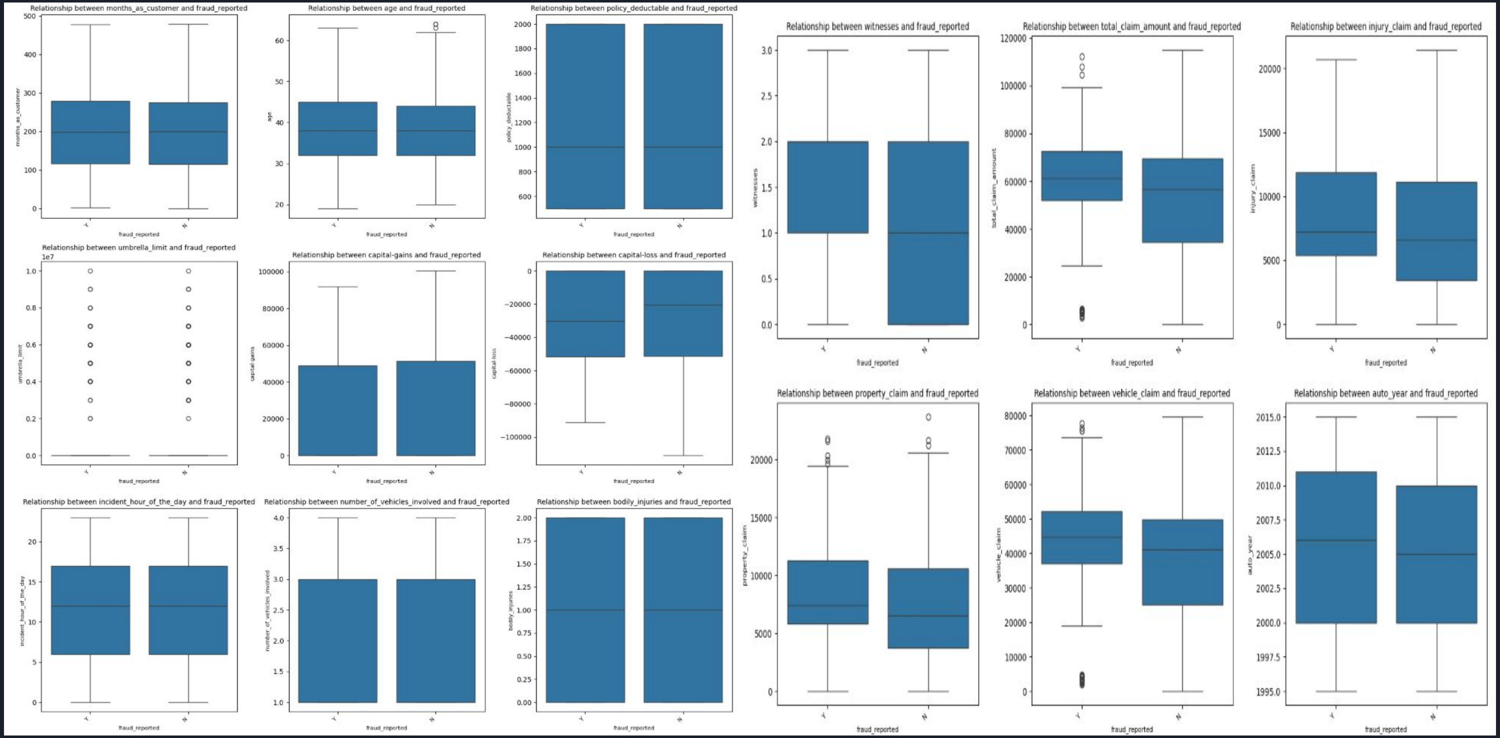


EDA - Correlation Analysis



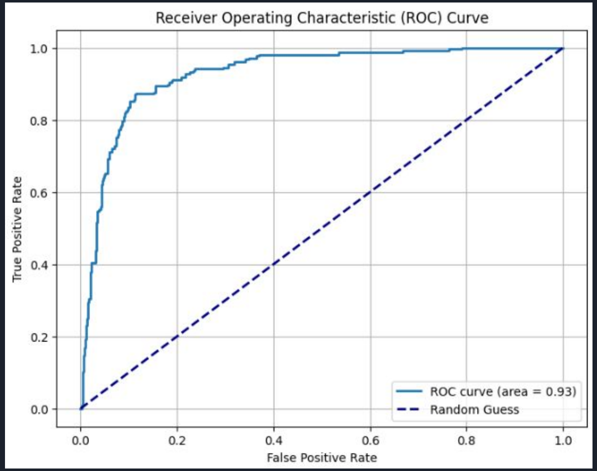
EDA - Class Distribution

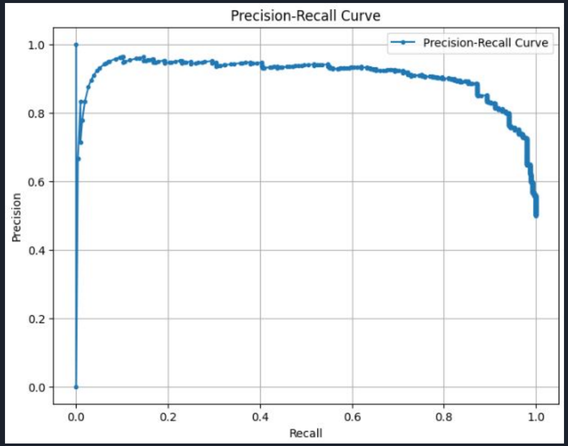
EDA - Bivariate Analysis Results



Logistic Regression: Optimal Cutoff

- Determined threshold by comparing:
  - Sensitivity
  - Specificity
  - Precision
  - Recall
- Used ROC and Precision-Recall curves





Conclusion - Final

In conclusion, the Random Forest model performed better in this analysis and can be used to predict the likelihood of fraud for incoming claims. The insights gained from the model, particularly the important features, can significantly help Global Insure improve their fraud detection process by enabling more efficient and data-driven screening of claims.