Case Study Assignment

Fraudulent Claim Detection

Ketan Chincholikar, Milan Banura, Lakhan Varshney
5-4-2025

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 minimize financial losses and optimize the overall claims handling process

Business Objective:

- Global Insure aims to enhance its ability to detect fraudulent insurance claims by leveraging historical claim data.
- The company seeks to identify patterns and key indicators that differentiate fraudulent claims from genuine ones.
- By developing a predictive model, it intends to assess the likelihood of fraud in incoming claims, enabling proactive fraud detection and reducing financial losses.
- 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

High Level Approach:

To achieve the objectives, two classification models i.e. Logistic Regression and Random Tree were created and evaluated.

Data Preparation:

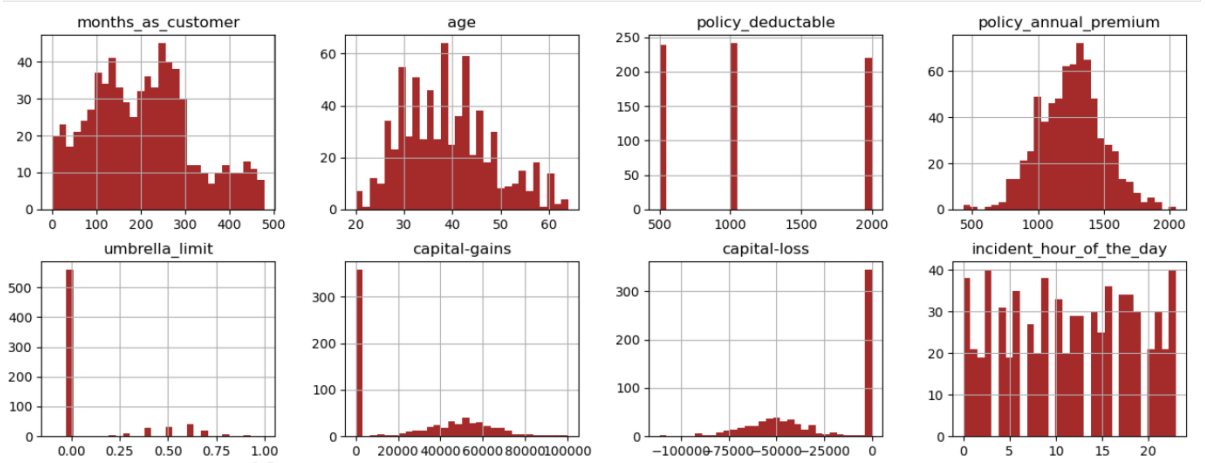
- Data was verified for null values. Column '_c39' has all null values and is dropped
- 36% values are missing for column 'property_damage' and 34.3% values are missing for column 'police_report_available'. We cannot impute the missing values using mode

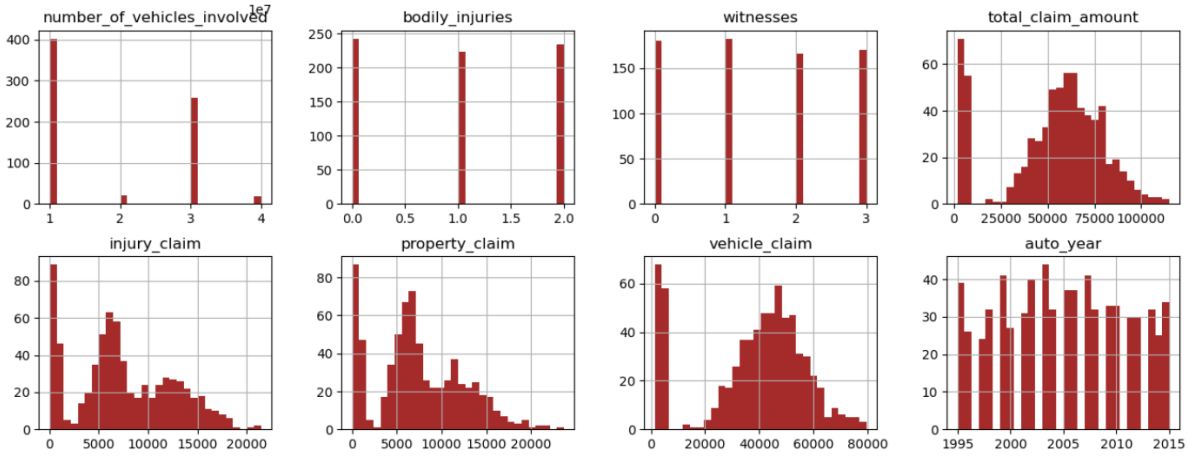
imputation as there is no dominant category values for these columns. Hence these columns are dropped

- There are 91 records with value "None" for 'authorities_contacted'. This is not a NaN. We substituted the values with ‘NotContacted’ instead of dropping the rows
- Columns with invalid values were handled next. 'collision_type' with value as "?" was handled
- Columns with negative values were identified and analyzed. There was no need to drop or impute any of the identified columns
- Columns with large proportion of the values are unique or near-unique values were identified and dropped. These columns are likely to be identifiers or have very limited predictive power. (policy_number, policy_bind_date, etc.)
- Data types were verified and fixed where necessary (incident_date)
- Created dummy variables for categorical columns using get_dummies().
- Rescaled numeric features and balanced the dataset using RandomOverSampler to handle class imbalance

Exploratory Data Analysis:

Univariate Analysis





The features that have an imbalanced or skewed distribution

- umbrella_limit: most customer have a limit of 0 (right skewed)
- capital_gains: values are right skewed around 0
- capital_loss: values are left skewed around 0
- Few mild right-skewed features like injury_claim, property_claim, vehicle_claim

Normally distributed features

- policy_annual_premium
- total_claim_amount

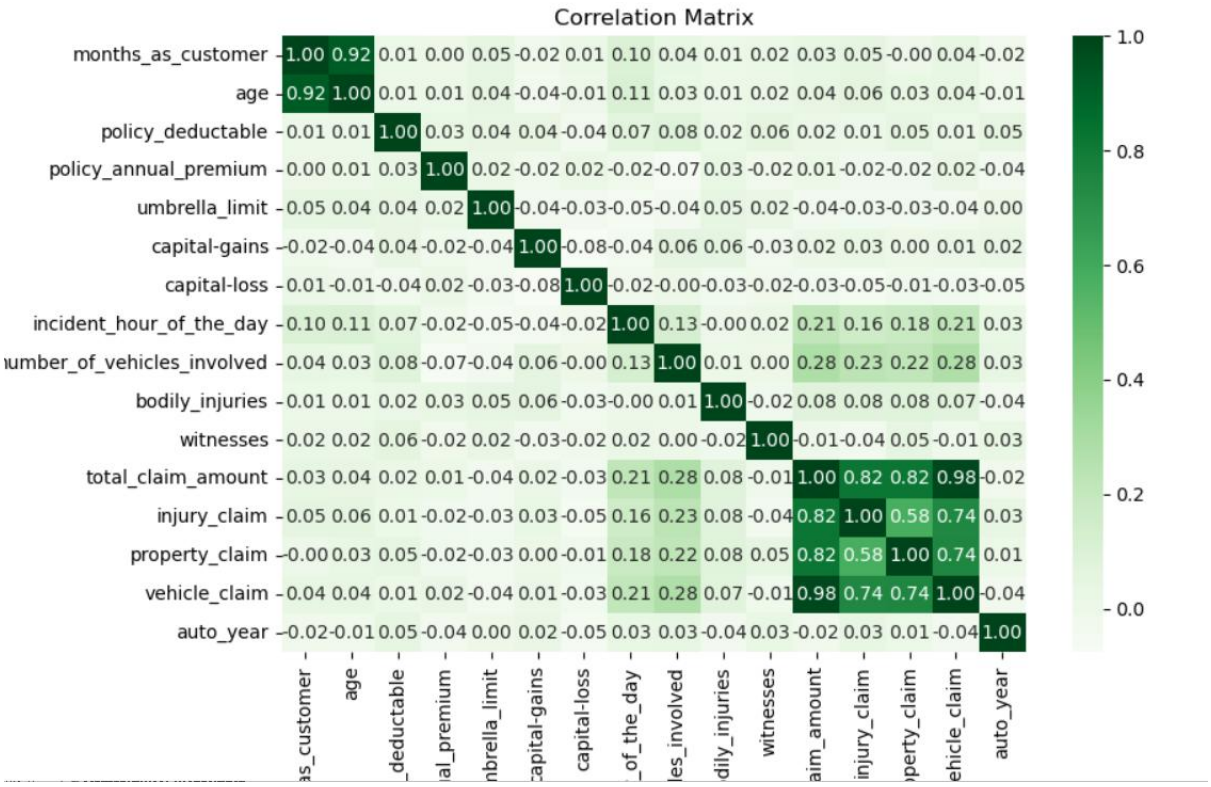
Features exhibiting category-like behavior

- bodily_injuries
- witnesses
- number_of_vehicles_involved
- policy_deductible
- auto_year

Correlation Analysis

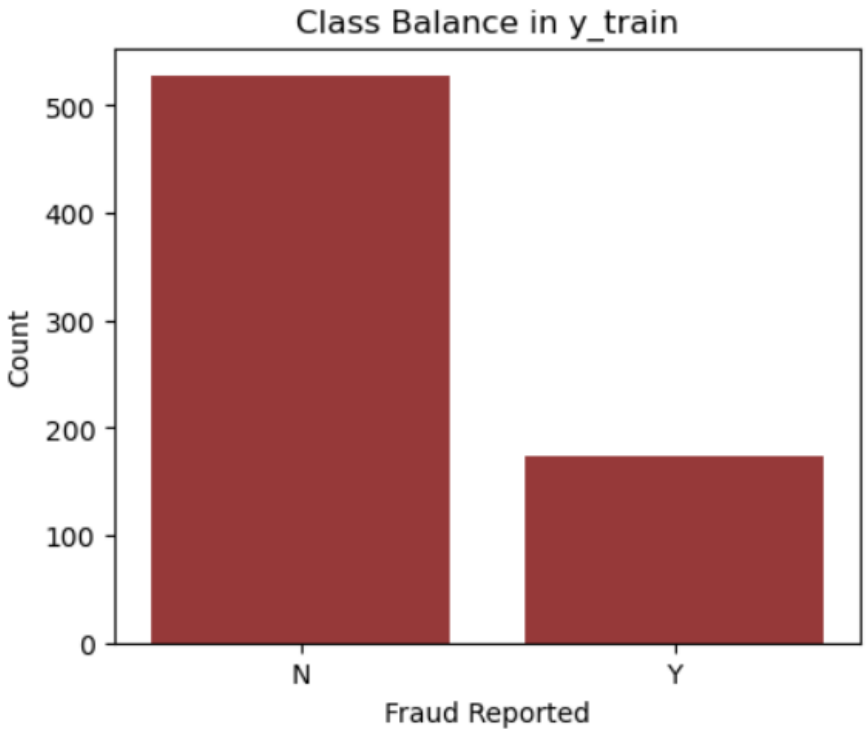
Strong correlation exists between following columns

- total_claim_amount & vehicle_claim - 0.98
- months_as_customer & age - 0.92
- total_claim_amount & injury_claim - 0.82
- total_claim_amount & property_claim - 0.82
- property_claim & injury_claim - 0.58



Class Imbalance

We observe a class imbalance with fraud reported in ~25% of cases



Bivariate Analysis

- policy_state: OH (26%) has higher fraud rates compared to IN (24%) and IL (24%)
- policy_csl: 100/300 has higher fraud rate of 28% compared to other values
- insured_sex: - Males and Females has same percentage of fraud rate (25%)
- insured_education_level: MD (28%), PhD and JD show slightly higher fraud rates. High school is the lowest (21%)
- insured_occupation: Fraud more likely in transport-moving (40%) and exec-managerial (33%) roles. They are lowest in handlers (15%) and adm (17%)
- insured_hobbies: Hobbies like chess (88%), cross-fit (83%) show high fraud risk
- insured_relationship: other-relative (29%) and unmarried (28%) have higher fraud rates
- incident_type: Single Vehicle Collision (29%) and Multi-vehicle Collision (27%) is more fraud-prone
- collision_type: Rear and Front Collisions (~30%) more fraudulent than Side or unknown
- incident_severity: Major Damage shows extremely high fraud likelihood (60%) vs. Total Loss (15%) and Minor Damage (9%)
- authorities_contacted: Ambulance and Fire lead to more fraud cases (30%, 30%)
- incident_state: OH (47%) has highest fraud likelihood
- auto_make: Audi,Mercedes, Ford (≥31%) lead in fraud; Accura is lowest (14%)

Feature Engineering:

- Handle class imbalance in the training data by applying RandomOverSampler resampling technique
- Created new features to capture the ratios - claim_to_premium_ratio, injury_claim_ratio, property_claim_ratio, vehicle_claim_ratio
- Identified new feature ‘is_weekend’ from date feature
- Dropped highly correlated features and date features as we have derived features
- Combined categories that occur infrequently or exhibit similar behavior to reduce sparsity and improve model generalisation

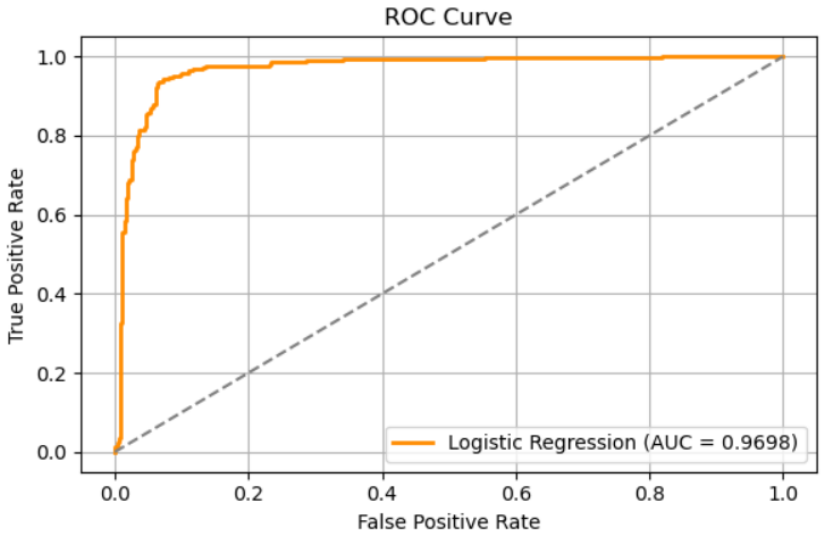
- Created dummy variables for categorical features
- Scaled the data using StandardScaler()

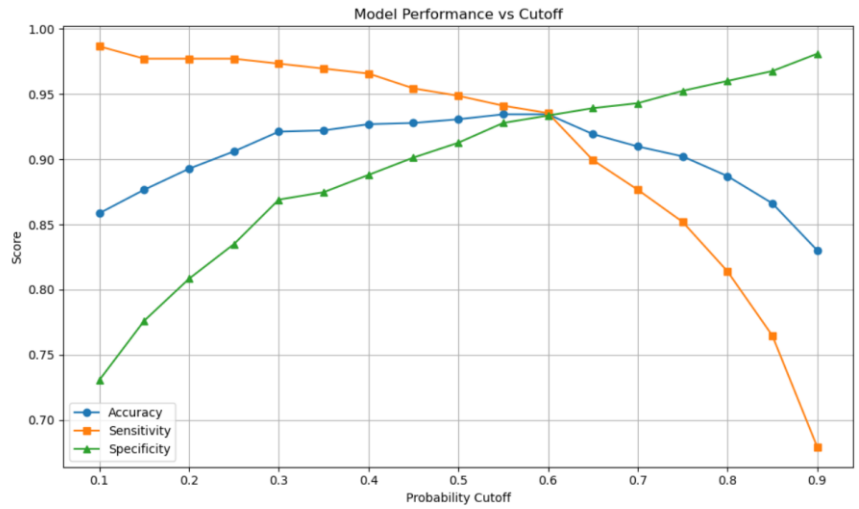
Model Building:

Logistic Regression Model

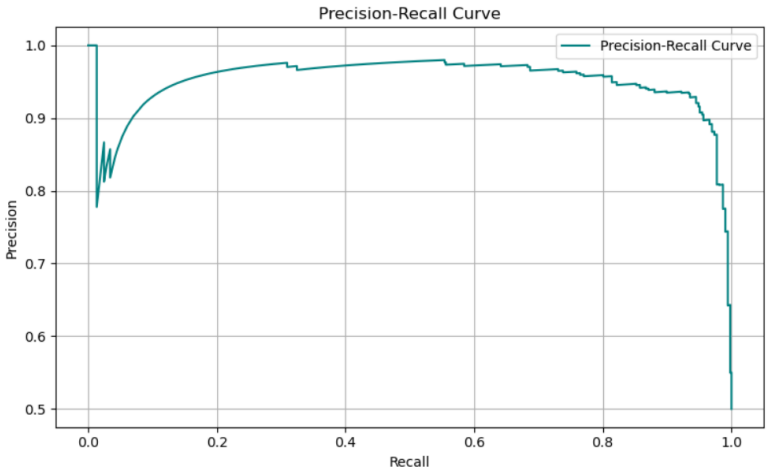
Cutoff	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
0.5	93%	0.95	0.91	0.92	0.95	0.93

AUC = 0.9698





Cutoff	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
0.6	93.5%	0.94	0.93	0.93	0.94	0.93



Random Forest Model

Hyperparameter Tuning	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
No	100%	1.00	1.00	1.00	1.00	1
Yes	94.87%	0.99	0.90	0.91	0.99	0.95

Model Evaluation:

Logistic Regression Model

Cutoff	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
0.6	79%	0.54	0.87	0.58	0.54	0.56

Random Forest Model

Hyperparameter Tuning	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
Yes	83.67%	0.77	0.86	0.64	0.77	0.70

Top 5 features as per RandomForest Model are

- incident_severity_Minor Damage
- incident_severity_Total Loss
- claim_to_premium_ratio
- insured_hobbies_chess

- policy_annual_premium

Top 5 features as per Logistic Regression Model

- insured_hobbies_chess
- insured_hobbies_cross-fit
- incident_severity_Minor Damage
- incident_severity_Total Loss
- incident_severity_Trivial Damage

Key Observations

- Random Forest offered better accuracy, precision, and F1 score, and handled complex patterns more effectively compared to Logistic Regression
- Feature importance from Random Forest added business insight into what contributes to fraud risk
- Overfitting concerns were mitigated through hyperparameter tuning and cross-validation, ensuring robust Random Forest performance.
- Variance recorded between training data and test data is high (F1 Score falls from 95% to 70% for Random Forest between training and validation). Model is not performing well on unseen data

Recommendation

- Use Random Forest as the primary fraud detection model, given its better accuracy and balanced performance across metrics
- Use Logistic Regression as a base model to identify the key variables impacting the fraud risk
- Use of PCA for dimensionality reduction and regularisation methods to reduce model complexity and have better fitment on the unseen data