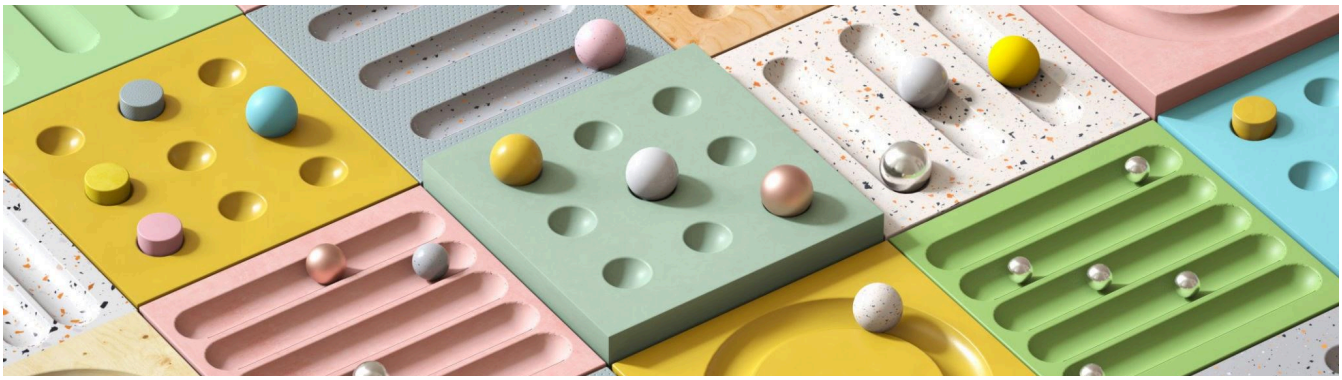

Fraudulent Insurance Claim Detection

Project Report

01. Overview

Global Insure, a leading insurance provider, faces significant financial losses due to fraudulent claims, currently identified through inefficient manual inspections. This project leverages historical claim data and advanced machine learning techniques to develop a predictive model for early fraud detection. Using a dataset of 1,000 claims with 40 features, we built and evaluated Logistic Regression and Random Forest models. The Random Forest model outperformed with an accuracy of 80.33% on validation data, identifying key predictive features such as claim amounts, incident severity, and customer tenure. This report outlines our methodology, findings, and recommendations to enhance Global Insure's fraud detection process, potentially reducing financial losses and improving operational efficiency.

INTRODUCTION

Global Insure processes thousands of claims annually, with a notable portion being fraudulent, leading to substantial financial losses.

The current manual fraud detection process is slow and often identifies fraud post-payout. This project aims to address the following

objectives:

1. Analyze historical claim data to detect patterns indicative of fraud.

2. Identify the most predictive features of fraudulent behavior.

3. Predict the likelihood of fraud for incoming claims using past data.

4. Provide insights to improve the fraud detection process.

We utilized a dataset containing 1,000 rows and 40 columns, including customer profiles, policy details, and incident specifics, to build data-driven solutions.

METHODOLOGY

1. DATA PREPARATION AND CLEANING

- Dataset Overview: Loaded a CSV file with 1,000 claims and 40 features (e.g., months_as_customer, total_claim_amount, fraud_reported).
- Cleaning: Handled missing values (e.g., replacing "?" with NaN), dropped irrelevant columns (e.g., _c39), and converted date types (e.g., dates to datetime).
- Outcome: A clean dataset ready for analysis.

2. EXPLORATORY DATA ANALYSIS (EDA)

- Training test split: Split the dataset into 70% training (700 rows) and 30% validation (300 rows).
- Fraudulent claims (~25%) showed higher claim amounts (vehicle_claim, property_claim).
- Severe incidents (incident_severity_Total Loss) were more associated with fraud.
- Shorter customer tenure (months_as_customer) correlated with fraudulent claims.

3. FEATURE ENGINEERING

- New Features: Created new columns injury_claim_ration, property_claim_ratio, vehicle_claim_ration, sum_sub_claims, claim_diff, customer_risk_score, and buckets for hobbies and auto makes.
- Encoding: Applied one-hot encoding to categorical variables (e.g., incident_severity, collision_type).
- Outcome: Enhanced dataset with 20+ predictive features.

4. MODEL BUILDING

- Logistic Regression:
    - Iteratively refined using statistical significance (p-values < 0.05).
    - The final model used 36 features, achieving 81% accuracy on training data.
- Random Forest:
    - Tuned via GridSearchCV with parameters (e.g., max_depth, n_estimators).
    - Selected top 20 features (e.g., vehicle_claim, incident_severity_Total Loss), achieving 95.53% accuracy on training data with balanced resampling.

5. PREDICTION AND EVALUATION

Validation Performance:

- Logistic Regression: 72.67% accuracy, sensitivity 72.97%, specificity 72.57%, F1-score 0.5684.
- Random Forest: 80.33% accuracy, sensitivity 64.86%, specificity 85.40%, F1-score 0.6194.

Key Metrics: Random Forest excelled in accuracy and specificity, making it more reliable for minimizing false positives(legitimate claims flagged as fraud)

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