Fraudulent Insurance Claim Detection

Prepared by: [Anukul Morey And Keshav Jadhav]

Problem Statement

- Global Insure processes thousands of claims annually. A significant percentage are fraudulent, resulting in financial losses.
- Current manual fraud detection is time-consuming and inefficient.
- Fraud is often detected after payouts have been made.

Business Objective

- Build a model to classify insurance claims as fraudulent or legitimate.
- Use features like claim amount, customer profile, and incident details to detect fraud early.
- Minimize financial losses and optimize claims processing.

Key Questions Answered

1. How can we analyse historical claim data to detect patterns that indicate fraudulent claims?

- Exploratory Data Analysis (EDA):
  - Performed univariate and bivariate analysis of variables like claim_amount, customer_age, policy_state, incident_type, etc.
  - Identified that certain features (e.g., high claim amounts, unusual incident types, mismatched driver information) show strong correlation with fraud.
- Data Cleaning:
  - Removed irrelevant and duplicate features.
  - Handled missing values and outliers to improve data quality.
- Feature Engineering:
  - Created new features (e.g., days between policy start and incident).
  - Encoded categorical variables using Label/One-Hot encoding.
- Model Building:
  - Used classification algorithms (e.g., Logistic Regression, Random Forest, XGBoost) to detect fraud patterns.

2. Which features are most predictive of fraudulent behaviour?

Based on feature importance from models (Random Forest / XGBoost):

- Top Predictive Features:
  - policy_state – Certain states show disproportionately high fraud.
  - incident_type – Types like "Single Vehicle Collision" or "Parked Car" often appear in fraud.
  - insured_hobbies – Unusual hobbies like "Skydiving" showed higher correlation.
  - collision_type – Specific collision types are strong indicators.
  - incident_severity – High severity often links with fraud.
  - auto_make – Certain car brands are more frequent in fraudulent claims.
  - months_as_customer – Short tenure often signals risk.
- Insights:
  - Fraudulent claims tend to be associated with suspicious incident types, rare vehicle makes, and customers with shorter history with the company.

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

- **Yes we can predict.**
  - A **classification model** was trained on historical data.
  - Accuracy and F1-Score from models (e.g., Random Forest, XGBoost) were in the acceptable range (e.g., ~0.91 accuracy, ~0.75 F1-score for fraud).
  - The model can **assign probabilities** to new claims being fraudulent.
  - Threshold tuning improves fraud recall with minimal false positives.

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

- Key Insights:
  - High-risk profiles can be flagged before approval.
  - Early triage of claims based on model predictions can reduce manual workload.
  - Specific red flags (incident type, high claim amounts, rare cars) should be prioritized.
  - Model-driven flags can be used to direct human investigation teams more efficiently.
- Example:
  - A customer with <6 months tenure, claiming a large amount for a single vehicle collision, with a rare car brand, has a high fraud probability. The model flags it for review.

Summary

- Fraud detection model built using historical claim data.
- Model achieved **91.90% accuracy** at a 0.5 probability threshold using Logistic Regression.
- Key features include incident type, tenure, car make, policy state.
- Model outputs a fraud probability score per claim.

Recommendations

- Deploy the model in production as a triage tool.
- Set thresholds to route high-risk claims to human review.
- Retrain periodically with new claim data.
- Integrate model with claims processing systems.

Business Implications

- Reduced financial losses from fraudulent payouts.
- Improved operational efficiency and faster processing of legitimate claims.
- Higher customer satisfaction.
- Scalable solution for other insurance products.