Fraudulent Claim Detection – Model Development Report

Prepared by: Pooja Singh, Tushar Rajput & Akanksha Garg

1. Problem Statement

Global Insure seeks to improve its ability to detect fraudulent insurance claims. Manual investigation is insufficient and delays genuine claims. The goal is to build predictive model that classifies incoming claims as fraudulent or legitimate using structured historical claim data.

2. Methodology Overview

- Cleaned and pre-processed structured insurance data.
- Performed detailed EDA to identify predictive patterns.
- Engineered features to enhance signal strength.
- Handled class imbalance using resampling techniques.
- Trained and evaluated Logistic regression and Random Forest models.
- Delivered business recommendations based on performance metrics.

3. Data Cleaning

- **Null values handled:** Imputed missing values in **'property_damage'** and **'police_report_available'**. Retained rows with valid **'fraud_reported'.**
- **Redundant columns removed**: Dropped ID-like and date columns (**policy_number, policy_bind_date,** etc.).
- **Low-frequency categories grouped:** Merged rare values in **'auto_model'**, '**auto_make'**, and **'Insured_hobbies'** into 'Other' groups.
- **Data types fixed**: Categorical features cast to category; numerics ensured proper dtype.

Outcome: Clean dataset with consistent types, minimal noise, and no major information loss.

4. Train-Validation Split

- Defined features and target (fraud_reported).

- Split into 70% training and 30% validation using stratification to maintain class balance.

5. Exploratory Data Analysis (EDA) – Training Data

- **Univariate analysis:** Analysed distributions using appropriate.
- **Correlation analysis:** Numerical features showed low multicollinearity.
- **Class imbalance confirmed:** Fraudulent cases ≈ 23.6%.
- **Bivariate analysis (Target Likelihood):**

  - **insured_hobbies:** High fraud rates in Chess and Cross-Fit.
  - **incident_type:** Single Vehicle Collisions showed higher fraud incidence.
  - 'incident_severity', 'collision_type', 'property_damage', and 'police_report_available' showed significant variation across fraud vs non-fraud.

Outcome: EDA highlighted patterns to guide feature selection and engineering.

6. Feature Engineering

- **Resampling (SMOTE):** Applied to training set only to balance minority class.
- **Feature creation:** Extracted 'incident_hour' from timestamp.
- **Redundant categories removed:** Simplified high-cardinality columns using frequency grouping.
- **One-hot encoding:** Applied to both train and validation sets using consistent columns.
- **Scaling:** Used MinMaxScaler selectively (only for Logistic Regression).

Outcome: Balanced, transformed, and meaningful feature matrix prepared.

7. Model Building

Logistic Regression

- Feature selection using **RFECV** and **VIF** analysis.
- Optimal cutoff chosen using Youden's J from ROC curve.
- Metrics:
  - o Accuracy: 0.73
  - o Precision: 0.47
  - o Recall: 0.82
  - o F1 Score: 0.60

Random Forest (Tuned)

- GridSearchCV tuned: n_estimators, max_depth, min_samples_split.
- **Metrics:**
  - o Accuracy: 0.77
  - o Precision: 0.57
  - o Recall: 0.12
  - o F1 Score: 0.20

Outcome: Two models built and validated. Logistic Regression offered better interpretability and slightly better recall. Random Forest had higher precision.

8. Model Evaluation

- Evaluation metrics used: Confusion Matrix, Accuracy, Precision, Recall, Specificity, F1 Score.
- Trade-offs identified:
  -High Accuracy due to class imbalance.
  -Logistic Regression captured more fraudulent cases.
  -Random Forest reduced false positives better.

Outcome: Both models assessed comprehensively with business - relevant metrics.

9. Insights & Recommendations

Key Insights

- High fraud rates in specific categories of 'insured_hobbies', 'incident_type', and collision_type.
- Low fraud correlation with police involvement and property damage reporting.

Model Suggestions

- Logistic Regression: Better recall; ideal when identifying more frauds is crucial.
- Random Forest: Better precision; suitable for reducing false positives.

Business Recommendations

- Use Logistic Regression for initial fraud screening.
- Deploy Random Forest as a second-stage filter.
- Investigate high-risk profiles (Cross-Fit hobbyists, single vehicle incidents) more closely.
- Collect more domain-specific variables (e.g., claim amount, past fraud record).