Fraudulent Claim Detection

April 7th , 2025

By

Arpan Das
Gagan Deep Madaan
Grishma G

Problem Statement

- Global Insure faces high losses due to fraudulent insurance claims
- Manual fraud detection is inefficient and resource heavy
- Objective: Build a data-driven model to identify fraud early using claim, customer, and incident data.

Data Overview & Preparation

- Dataset: 1000 rows, 40 columns.
- Key Steps:
- Dropped null columns (e.g., _c39)
- Converted date fields to datetime
- Identified numerical and categorical features.

Data Cleaning

- Dropped fully null columns
- Replaced placeholders like '?'
- Encoded binary variables Yes/No → 1/0
- Date formatting and data type fixes werer taken care

Train-Test Split

- 70:30 train-test split
- Stratified by target variable (fraud_reported)

Exploratory Data Analysis (EDA)

- Univariate: Histograms and boxplots revealed skewed distributions
- Bivariate: Higher fraud likelihood with:
- Craft repair occupation
- Hobbies like Chess, Crossfit
- Major damage & vehicle collisions
- 'Other relative' as claimant

Feature Engineering & Resampling

- Addressed class imbalance using RandomOverSampler
- Created dummies for categorical variables
- Scaled numeric features using MinMaxScaler

Model Building

- Model 1–3: Logistic Regression
- Evaluated using:
- Accuracy (≈ 88%)
- Sensitivity, Specificity, F1-score
- ROC Curve & Optimal Cutoff

Random Forest Model

- Achieved 100% on training data - Overfitting
- After tuning:
- Accuracy: ~92.7%
- Sensitivity: 96%
- F1-score: 0.93

Validation Results

- Logistic Regression:
- Accuracy: 85.3%, F1-score: 0.71
- Random Forest:
- Accuracy: 82.3%, F1-score: 0.66

Key Insights & Recommendations

EDA Insights:

- Fraud linked to occupation, hobby, damage type, and relationship to claimant. Craft repair, hobbies like chess or crossfit are fraud-prone
- Major damages and collisions show high fraud likelihood

Modeling:

- Logistic regression and RF models both are effective
- High recall and precision are crucial in fraud detection