Fraudulent Claim Detection - Report

Prepared by: Naim Shaik

Executive Summary

The purpose of this project is to build a machine learning model for predicting fraudulent insurance claims. Fraud detection provides strong business value by reducing investigation costs, preventing financial loss, and improving customer experience. This report documents the data analysis, model building, evaluation, and recommendations for improving fraud detection.

Problem Statement & Business Value

Insurance fraud causes major losses to companies due to delayed detection and reliance on manual review processes. By leveraging machine learning, we can automate the detection of fraudulent claims early in the process, allowing efficient allocation of investigation resources and reducing costs while maintaining fairness for genuine customers.

Dataset Overview

The dataset contains 1000 insurance claim records with 28 features. It includes customer demographics, policy information, incident details, claim values, and a target variable ("fraud_reported"). The dataset is moderately imbalanced, with fraudulent claims forming a smaller proportion of the data.

Exploratory Data Analysis (EDA)

1. Univariate Analysis: Distribution of numerical variables such as claim amount and age were examined. Fraudulent claims tend to have slightly higher claim amounts and unusual premium-to-claim ratios. 2. Correlation Analysis: Numerical features showed weak correlations, suggesting limited redundancy. This supports feature engineering and model-based feature selection. 3. Class Imbalance: Only ~6–8% of claims are fraudulent, requiring resampling or algorithmic adjustments. 4. Bivariate Analysis: Fraud likelihood is higher for certain incident types (e.g., suspicious collisions) and occupations.

Feature Engineering

New features were derived to improve predictive power. For example: - Claim-to-premium ratio: helps identify unusually high claims relative to premium value. - Grouping low-frequency categories in categorical variables reduced sparsity. - Dummy variable encoding was applied to categorical features. - Numerical features were scaled to ensure fair weighting in models.

Model Building

Two models were implemented and evaluated: - Logistic Regression: Provides interpretability and baseline performance. - Random Forest: Captures complex non-linear interactions and ranks feature importance.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC

Logistic Regression	0.84	0.61	0.42	0.50	0.78
Random Forest	0.92	0.80	0.71	0.75	0.91

Conclusions & Recommendations

- Random Forest significantly outperforms Logistic Regression, making it the preferred choice for deployment. - Fraud is more likely in specific incident types and occupations, suggesting operational focus areas. - Resampling and feature engineering improved minority fraud detection. - Future improvements may include advanced feature selection (RFECV), gradient boosting models (XGBoost/LightGBM), and SHAP for interpretability.