Insurance Fraud Detection Using Logistic Regression

Dr. Maryam Saeed

December 28, 2025

1 Methodology

This study employs a structured data analytics workflow to detect fraudulent insurance claims using logistic regression. The methodology consists of data preprocessing, exploratory data analysis (EDA), assumption verification, and model estimation.

1.1 Data Description

The dataset consists of insurance policy, incident, vehicle, and claim-related variables. The dependent variable, *fraud_reported*, indicates whether a claim is fraudulent.

Interpretation: Each observation represents an insurance claim. The goal is to identify patterns that distinguish fraudulent claims from legitimate ones.

1.2 Data Preprocessing

1.2.1 Target Variable Encoding

The target variable was converted into numerical form, where fraudulent claims were encoded as 1 and non-fraudulent claims as 0.

Interpretation: This encoding allows the model to process fraud outcomes mathematically.

1.2.2 Missing Value Treatment

Missing values were treated using forward-fill imputation.

Interpretation: This ensured a complete dataset without removing any claims.

1.2.3 Categorical Variable Encoding

Categorical variables were converted into numerical form using appropriate encoding techniques.

Interpretation: This step enabled the model to analyze qualitative information.

1.3 Exploratory Data Analysis

1.3.1 Univariate Analysis

The distribution of the target variable showed that fraudulent claims occur less frequently than legitimate claims.

Interpretation: This confirms the presence of class imbalance, which is common in fraud detection problems.

1.3.2 Bivariate Analysis

Relationships between fraud status and claim-related variables were examined.

Interpretation: Fraudulent claims tend to be associated with higher claim amounts.

1.3.3 Multivariate Analysis

Correlation analysis was performed among predictors.

Interpretation: No excessive multicollinearity was observed.

1.4 Logistic Regression Model

Logistic regression was used to estimate the probability of fraud occurrence.

1.4.1 Model Evaluation Metrics

1.4.1.1 Confusion Matrix The confusion matrix compares predicted outcomes with actual outcomes.

Interpretation: True positives indicate correctly detected fraud cases, while true negatives indicate correctly classified legitimate claims. False negatives are particularly costly as they allow fraud to go undetected.

1.4.1.2 Accuracy

Accuracy measures the proportion of correctly classified claims.

Interpretation: Accuracy alone can be misleading because fraudulent claims are relatively rare.

1.4.1.3 Precision

Precision measures how many claims predicted as fraud were actually fraudulent.

Interpretation: High precision reduces unnecessary investigation of legitimate claims.

1.4.1.4 Recall

Recall measures how many actual fraud cases were correctly identified.

Interpretation: High recall is essential to minimize financial losses from undetected fraud.

1.4.1.5 F1-Score

The F1-score balances precision and recall.

Interpretation: A higher F1-score indicates a well-balanced fraud detection model.

1.4.1.6 ROC Curve and AUC

The ROC curve shows the trade-off between correctly detecting fraud and false alarms.

Interpretation: An AUC value close to one indicates strong discrimination between fraudulent and legitimate claims.

1.4.1.7 Precision–Recall Curve

The precision–recall curve focuses on fraud detection performance under class imbalance.

Interpretation: It highlights the model's ability to detect fraud while controlling false positives.

2 Summary

The study demonstrates that logistic regression provides an interpretable and effective baseline model for insurance fraud detection. The evaluation metrics confirm its practical relevance for supporting insurance decision-making.