Statistical Concepts in Fraud Detection

Univariate & Bivariate Analysis

Univariate analysis examines a single feature using mean, median, standard deviation, histograms, and boxplots.

Bivariate analysis looks at the relationship between two features. In this project, we used boxplots to compare numerical features across fraud labels to detect concentration of fraud behavior.

Skewness and Log Transformations

Features like transaction amount and balance differences are highly right-skewed. Most values are small, but a few are very large.

To normalize this and reduce variance, we use log1p (log(x+1)). This helps machine learning models interpret data more effectively.

Handling Class Imbalance

Fraud cases are rare. This imbalance can make accuracy misleading. For example, always predicting 'not fraud' gives high accuracy but 0% recall for fraud.

We focus on precision, recall, and F1-score instead. Techniques include class weighting, stratified sampling, and SMOTE.

Feature Engineering

We engineered features like orig_diff (oldbalanceOrg - newbalanceOrig) and dest_diff (newbalanceDest - oldbalanceDest).

These help track how much money is moved and are often very telling of fraud patterns.

Statistical Concepts in Fraud Detection

Feature Scaling

For algorithms like logistic regression or SVM, large value ranges can skew the model. Scaling helps stabilize training.

StandardScaler is used when features are normally distributed. MinMaxScaler is preferred after log transformations or to fit values into 0 to 1 range.

Encoding Categorical Features

Transaction type is categorical. Since there is no order, we use one-hot encoding to turn it into separate binary columns.

Alternatively, target encoding can be used for tree models by replacing each category with its average fraud rate.