Insurance Claims Fraud Detection

The Syntax Writer

Insurance Claims Fraud Detection

Introduction

This project aims to detect fraudulent insurance claims using a dataset with 1000 rows and 40 columns. The target variable for prediction is 'fraud_reported'. Various data preprocessing, exploratory data analysis (EDA), and machine learning techniques are applied to build a robust model for fraud detection.

Insurance Claims Fraud Detection

Data Information

The dataset contains 19 numerical and 21 object columns. Initial data exploration reveals no null values except in the '_c39' column, which is dropped. '?' values are replaced with 'No Info'. The dataset is then analyzed for statistical properties, revealing skewness and outliers in several columns.

Insurance Claims Fraud Detection

Data Preprocessing

Data preprocessing includes handling missing values, converting categorical data to numerical using LabelEncoder, and addressing skewness and outliers. Numerical columns such as 'umbrella_limit', 'total_claim_amount', and 'vehicle_claim' are transformed using the Yeo-Johnson method. Outliers are handled using the Z-score method.

Insurance Claims Fraud Detection

Exploratory Data Analysis

Exploratory Data Analysis involves visualizing the data using various plots. Key findings include:

- High correlation between 'total_claim_amount', 'injury_claim', 'property_claim', and 'vehicle_claim'.
- Imbalance in the target variable 'fraud_reported'.
- Patterns of fraud in relation to various features like 'number_of_vehicles_involved', 'incident_city', 'incident_state', etc.

Insurance Claims Fraud Detection

Model Building and Evaluation

Several machine learning models are built and evaluated, including Logistic Regression, Decision Tree Classifier, KNeighbors Classifier, Random Forest Classifier, and ensemble techniques like AdaBoost, Bagging, and Gradient Boosting. Gradient Boosting Classifier is chosen for hyperparameter tuning due to its balanced performance between training and test data.

Insurance Claims Fraud Detection

Hyperparameter Tuning

Hyperparameter tuning is performed on Gradient Boosting Classifier using GridSearchCV. The best parameters are found to be:

- criterion: 'mse'
- n_estimators: 200
- learning_rate: 0.1
- random_state: 5

The model's accuracy improves after hyperparameter tuning.

Insurance Claims Fraud Detection

Final Model and Results

The final model, Gradient Boosting Classifier with the best parameters, achieves a high accuracy score. The model is saved using joblib for future use. The final classification report and ROC-AUC curve indicate the model's effectiveness in predicting fraudulent claims.