- **Yassine Sfaihi**
- www.linkedin.com/in/yassinesfaihi
- www.github.com/yassinesfaihi
- www.kaggle.com/yassinesfaihi

[Author] : Yassine Sfaihi

Fraud Detection with XGBoost and LightGBM: Using SMOTE and Model Weight Tuning

Introduction

This notebook is an exploratory data analysis and machine learning pipeline for detecting fraud transactions using **XGBoost and LightGBM algorithms**. The main goal of this project is to create a model that can accurately detect fraud transactions. To do this, the data is first loaded into a Pandas DataFrame and basic information about the data is provided, such as the number of rows and columns, the data types of the columns, and a statistical summary of the dataset.

Next, various visualizations are plotted to better understand the distribution of the target variable (Class) and other columns. This includes bar plots to show the count of each class, histograms to show the distribution of the amount for each class, box plots to show the distribution of variables across classes, and histograms for all columns.

In addition to exploring the data, the code checks for missing values and removes outliers. It also checks for NaN or infinite values and removes them. The data is then split into features (X) and target (y) and processed through pipelines that include **StandardScaler, SMOTE** (for oversampling), and XGBoost or LightGBM.

XGBoost Model

The XGBoost model is a binary classification model used to predict whether a transaction is fraudulent or not. The pipeline includes StandardScaler, SMOTE, and XGBoost. Cross_val_predict is used to obtain predicted probabilities of the target variable and a confusion matrix is calculated. Precision, recall, and f1-score are also calculated and a precision-recall curve is plotted.

Hyperparameter tuning is performed using **GridSearchCV**.

The hyperparameter being tuned is **"scale_pos_weight"**, which balances the weight between positive and negative classes in the target variable "y".

The scoring function used to evaluate the models during tuning is the average precision score (PRC-AUC). The best hyperparameter value is reported and the mean PRC-AUC scores for all configurations are displayed.

- **Yassine Sfaihi**
- www.linkedin.com/in/yassinesfaihi
- www.github.com/yassinesfaihi
- www.kaggle.com/yassinesfaihi

The model is then trained with the best hyperparameter value and the precision-recall curve is plotted again, showing the trade-off between precision and recall for different thresholds of the predicted probabilities.

LightGBM Model

The LightGBM model is also a binary classification model used to predict whether a transaction is fraudulent or not.

The pipeline includes StandardScaler, SMOTE, and LightGBM. Cross_val_predict is used to obtain predicted probabilities of the target variable and a confusion matrix is calculated. Precision, recall, and f1-score are also calculated and a precision-recall curve is plotted.

A grid search is performed on the scale_pos_weight parameter of LightGBM to find the best value that maximizes the average precision score.

The model is then trained with the best hyperparameter value and the precision-recall curve is plotted again, showing the trade-off between precision and recall for different thresholds of the predicted probabilities.

Summary

This notebook provides a comprehensive solution for detecting fraud transactions using XGBoost and LightGBM algorithms.