Name: Hema N
Report: Credit card fraud Detection

Capstone Project: Prediction Of Credit Card Fraud

Problem Statement:

Credit card fraud detection is critical to ensure that customers are not charged for fraudulent transactions. The dataset provided contains transactions made by credit cards in September 2013 by European cardholders. It includes 284,807 transactions over two days, Whereas the data is highly unbalanced with 492 transactions identified as fraudulent.

Aim:

The main aim of the project is to build the machine learning model to detect fraudulent credit card transactions. This will help credit card companies to reduce the risk of fraud transaction.

Structure of dataset:

1. **Rows**: Each row represents a transaction.

2. **Columns**: There are 31 columns.
    - **Time**: The time elapsed between this transaction and the first transaction in the dataset (in seconds).
    - **V1 to V28**: The result of a PCA (Principal Component Analysis) transformation applied to the original features (for confidentiality reasons).
    - **Amount**: The transaction amount.
    - **Class**: The class label, where 1 indicates a fraudulent transaction and 0 indicates a legitimate transaction.

Exploratory Data Analysis (EDA)

Key Findings from the given dataset:

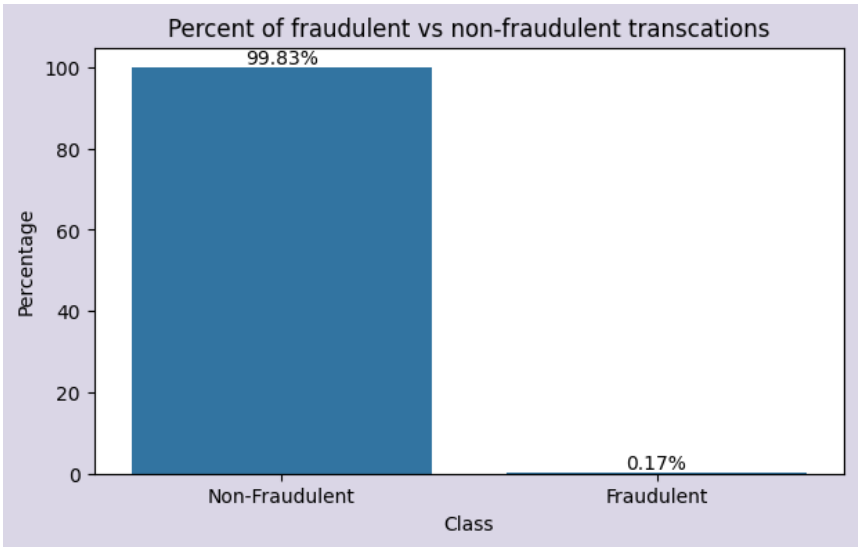
Distribution of Features:

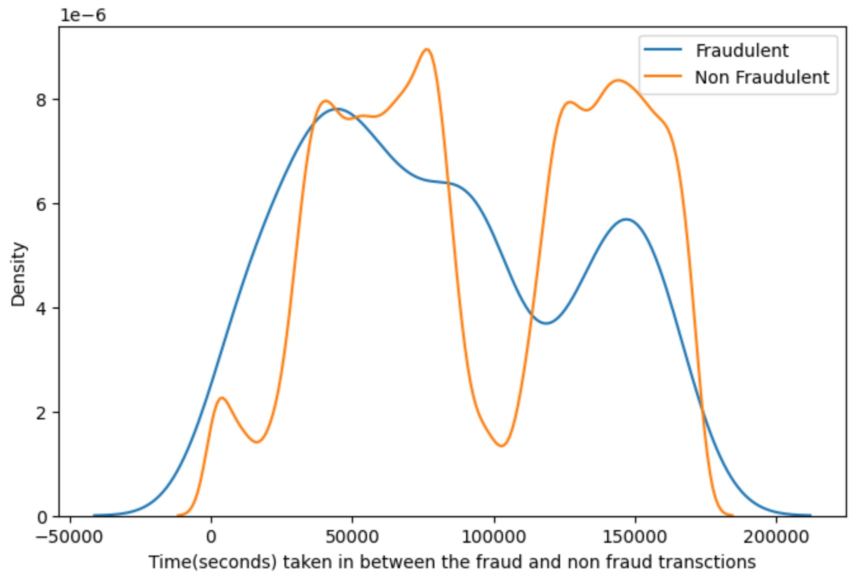
The dataset has been pre-processed with PCA, so features are named V1, V2, ..., V28. Class Imbalance: Visualizations confirmed the severe class imbalance. Amount Distribution: The transaction amounts vary significantly, with frauds often involving smaller amounts.

Distribution of Classes:

The distribution of the classes (fraudulent vs. non-fraudulent transactions) is examined to understand the class imbalance, which is crucial for model training and evaluation.

Name: Hema N
Report: Credit card fraud Detection





Data Exploration and Preprocessing

Data Loading and Initial Exploration:

• The dataset is loaded using pandas, and initial exploration involves checking the first few rows, summary statistics, and data types of the columns. Summary Statistics:

• Summary statistics provide insights into the distribution and scale of the data.

Checking for Missing Values:

• The dataset is checked for missing values, and it was found that there are no missing values in any of the columns.

This ensures that the dataset is complete and ready for analysis without the need for imputation.

Distribution of Classes:

• The distribution of the classes (fraudulent vs. non-fraudulent transactions) is examined to understand the class imbalance, which is crucial for model training and evaluation.

Handling Data Imbalance:

As we see that the data is heavily imbalanced, several approaches are employed to handle this imbalance

• **Under sampling**:

For balancing the class distribution, the number of non-fraudulent transactions is reduced to match the count of fraudulent transactions (492).

• **Oversampling:**

The number of fraudulent transactions is increased to match the count of non fraudulent transactions.

• **SMOTE (Synthetic Minority Over-sampling Technique):**

This oversampling technique uses the nearest neighbor algorithm to create synthetic data points thereby increasing the number of fraudulent transactions.

• **ADASYN (Adaptive Synthetic Sampling):**

Similar to SMOTE, ADASYN generates synthetic data, but it focuses on creating data points in regions with low density of minority class samples.

Model Building and Evaluation Models Used

We build three modules for Evaluation are listed below.

Logistic Regression (LR)

Decision Tree (DT)

XGBoost (XGB)

Name: Hema N
Report: Credit card fraud Detection

Model Performance Metrics:

Accuracy:

The ratio of correctly predicted transactions to the total transactions.

Precision (Positive Predictive Value):

The ratio of correctly predicted fraudulent transactions to all predicted fraudulent transactions.

Recall (Sensitivity):

The ratio of correctly predicted fraudulent transactions to all actual fraudulent transactions.

Sensitivity:

The ratio of correctly predicted non-fraudulent transactions to all actual non fraudulent transactions

Specificity:

The ratio of correctly predicted non-fraudulent transactions to all actual non fraudulent transactions

F1 Score:

The harmonic mean of precision and recall.

ROC-AUC:

The area under the receiver operating characteristic curve, which plots TPR vs. FPR.

Model Results Overview:

The performance of each model was evaluated using a confusion matrix and classification report, with key metrics summarized below:

Name	Accuracy	Precision	Recall	F1scorer	ROCAUC	Sensi tivity	specificity	FRP	PPV	NPV
Logistic Regression	0.910946	0.018083	0.952381	0.035492	0.982625	0.9524	0.9109	0.0891	0.0181	0.9999
Decision Tree	0.997811	0.422481	0.741497	0.538272	0.869875	0.7415	0.9983	0.0017	0.4225	0.9996
XGBoost	0.999134	0.721212	0.809524	0.762821	0.967321	0.8095	0.9995	0.0005	0.7212	0.9997

Name: Hema N
Report: Credit card fraud Detection

Hyperparameter Tuning:

Hyperparameters of the models are tuned using techniques like GridSearchCV or RandomizedSearchCV to find the optimal parameters that yield the best performance

Hyperparameter lr model,dt model,xgboost model

For each model, hyperparameter tuning was performed to identify the best set of parameters that maximize the model's performance. Below are the details for each model:

Logistic Regression:

Solver: The algorithm to use in the optimization problem. Best found: lib linear .

C: Inverse of regularization strength. Best found: 0.1

Decision Tree:

• criterion: The function to measure the quality of a split. Best found: gini

• Max Depth: The maximum depth of the tree. Best found: 10

• Min Samples Split: The minimum number of samples required to split an internal node. Best found: 2

• Min Samples Leaf: The minimum number of samples required to be at a leaf node. Best found: 1 Fitting 5 folds for each of 72 candidates, totalling 360 fits

XGBoost:

Fitting 5 folds for each of 162 candidates, totalling 810 fits

• Learning Rate (0.2) • Max Depth: Maximum depth of a tree. Best found: 5

• Subsample: The fraction of samples to be used for fitting the individual base learners. Best found: 0.7 • Colsample_bytree: The fraction of features to be used for fitting the individual base learners. Best found: 1.0

• Number of Estimators: The number of gradient boosted trees. Best found: 300 Hyperparameter tuning was performed using grid search and cross-validation to ensure robust selection of parameters.

Name: Hema N
Report: Credit card fraud Detection

Model Deployment:

The best-performing model is serialize using the pickle module for deployment We make the model deploy use to pickle file best_xgb model

First to Save the model to a file

with open('best_xgb_model.pkl', 'wb') as

file: pickle.dump(model, file) #load the model from the file

with open('best_xgb_model.pkl', 'rb') as file:

best_xgb = pickle.load(file) # Use the loaded model to make prediction

predictions = best_xgb.predict(X_test) and now print the model

Conclusion:

The project successfully demonstrated the application of various machine learning techniques to detect fraudulent transactions in a highly imbalanced dataset.

By employing strategies like **SMOTE** to address class imbalance and using a combination of different classifiers, we were able to achieve high performance metrics, particularly in precision and recall, which are crucial in fraud detection scenarios.

The high **ROC-AUC score** indicates that our model is effective at distinguishing between fraudulent and non-fraudulent transactions. However, there is still room for improvement, especially in increasing the recall rate to minimize false negatives, which are critical in fraud detection.