Credit Card Fraud Detection

Problem statement :

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

In this assignment, delve into credit card fraud detection using a high-dimensional dataset. Address data imbalance by employing oversampling or under sampling techniques, followed by data normalization. To tackle high dimensionality, apply Principal Component Analysis (PCA) for dimension reduction, enhancing model efficiency.

Experiment with advanced anomaly detection algorithms, including Isolation Forest, Local Outlier Factor (LOF), and Autoencoders. Employ cross-validation for robust model evaluation, utilizing metrics such as Precision, Recall, F1-Score, and ROC-AUC to measure performance effectively.

Advance your skills by delving into hyperparameter tuning using grid search or Bayesian optimization, fine-tuning models to optimize detection accuracy. Visualization plays a critical role: plot data distribution in reduced dimensions and visualize anomaly patterns for insightful interpretation.

Craft a comprehensive report encapsulating your methodology, challenges faced, and insights gained. The report should detail the experimentation process, including code snippets, and explain the rationale behind your decisions. To complement the report, create a dynamic presentation that concisely communicates key findings, model performance, and visualization outcomes.

Introduction:

Data Exploration is done to analyse the data. Data is Normalized using Standard Scaler. Dimensionality Reduction is applied here using Principal Component Analysis (PCA), and experimented with a range of anomaly detection algorithms, including Isolation Forest, Local Outlier Factor (LOF), and Autoencoders. Visualization plot data distribution in reduced dimensions and visualize anomaly patterns for insightful interpretation.

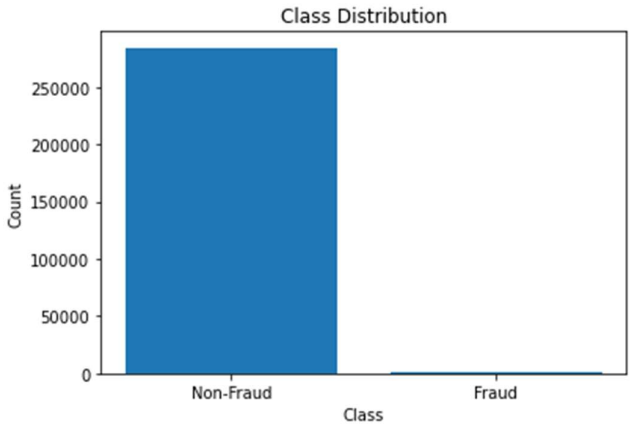
Data Source - *https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud*

Data Exploration and Preprocessing :

read_csv () is used to load the data into the notebook.

head() function is used to display the first five values in the dataset.

Value_counts() is applied on the target value to decide how many classes are present and the number of rows for each class.



Bar chart is used to display the distribution of data target variable in each class.

From data.info() we can conclude that the data doesn't contain null values and all the values are numerical.

data.describe() to display the statistics of the data.

Split the data to x and y as input and output variables respectively.

Splitting the data into training data and testing data to train the model and to test the model.

Over sampling with SMOTE

This technique adds the duplicate data to the class with less values to balance the data.

Under sampling with RandomUnderSampler

This technique removes the data from the class with more values to balance the data.

Both the techniques are used but the SMOTE gives good results.

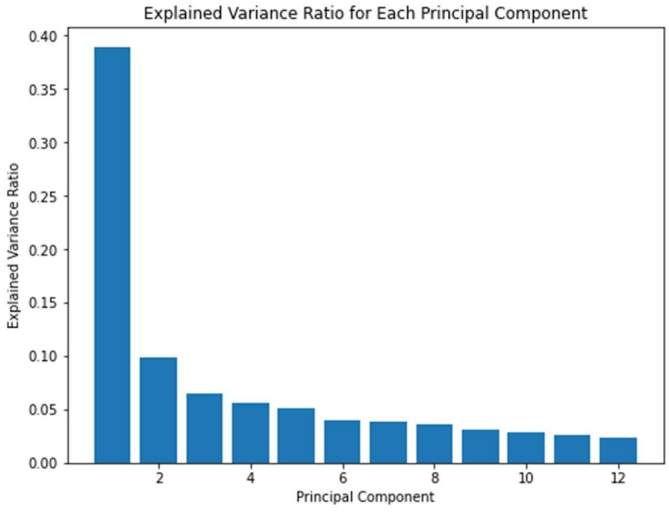
StandardScaler() is used to normalize the data. The range of the data is very high, so normalization is used to decrease the range.

Dimensionality Reduction with PCA :

Dimensionality reduction is a technique used in data analysis and machine learning to reduce the number of input variables or features in a dataset while preserving as much relevant information as possible.

PCA is used for Dimensionality Reduction and to take the principal components that capture the most variate data.

Hyper parameter is number of components that take as principal components. This can be



experimented using explained variance ratio for each component.

I took the number of components as 12.

Model Selection and Experimentation :

Isolation Forest – It is an anomaly detection algorithm. Isolation Forest is particularly effective for high-dimensional datasets.

Hyper Parameter – contamination, bootstrap

Results –

Precision: 0.010256410256410256

Recall: 0.10204081632653061

F1-score: 0.01863932898415657

ROC-AUC : 0.5425352505943288

Local Outlier Factor - The Local Outlier Factor (LOF) model is an unsupervised anomaly detection algorithm used to identify outliers or anomalies in a dataset.

Hyper Parameter – n_neighbors

Results –

Precision: 0.02247191011235955

Recall: 0.7959183673469388

F1-score: 0.04370972261137573

ROC-AUC : 0.8681248420865251

MLP Classifier - The MLP Classifier (Multi-Layer Perceptron Classifier) is a type of artificial neural network model used for classification tasks.

Hyper Parameter – Learning rate, alpha

Results –

Precision: 0.3018867924528302

Recall: 0.8163265306122449

F1-score: 0.44077134986225897

ROC-AUC : 0.9065365770675181

MLP Classifier the best-performing model because its recall value and ROC-AUC score is good compared with other models.

Cross validation is used to increase the accuracy of the model.

Hyperparameter Tuning :

Grid CV Search is used to find the best values as the hyper parameters.

Grid CV Search - GridSearchCV (Grid Search Cross-Validation) is a hyperparameter optimization technique commonly used in machine learning to systematically search for the best combination of hyperparameter values for a model.

Considered hyper parameters are,

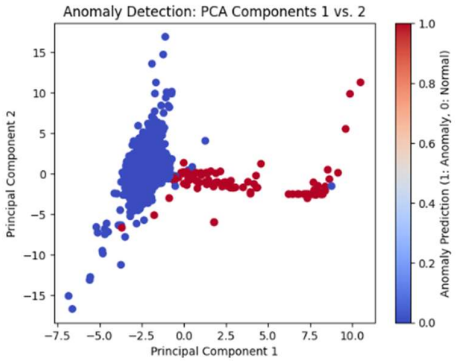
contamination : [0.05, 0.1], bootstrap : [True, False], n_neighbors = [20,30,50], learning_rate = ['constant', 'adaptive', 'invscaling'], alpha: [0.0001, 0.001, 0.01],

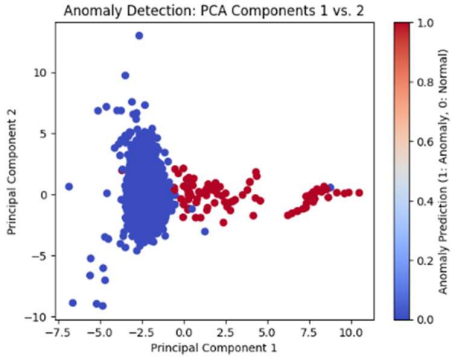
Best Hyper Parameters obtained are,

{'bootstrap': True, 'contamination': 0.05}, {'n_neighbors': 20}, {'learning_rate': 'invscaling'}, {'alpha': 0.001}

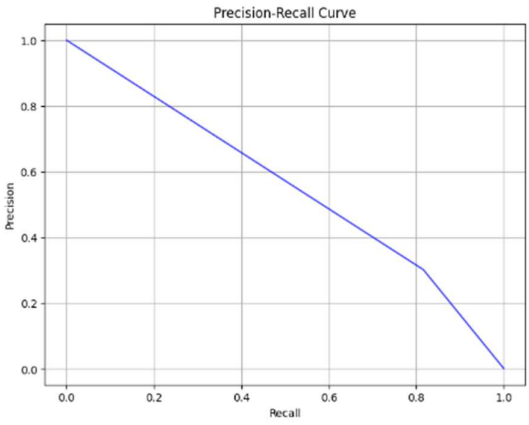
Visualization :

Visualization of PCA components –

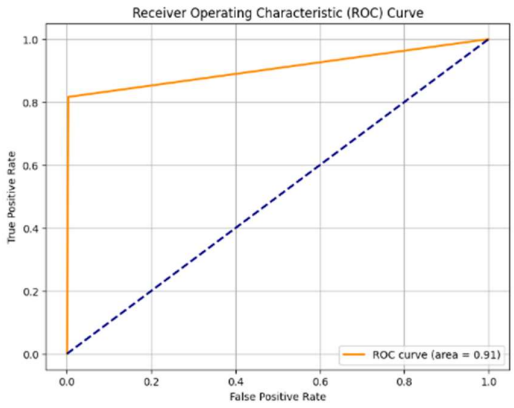




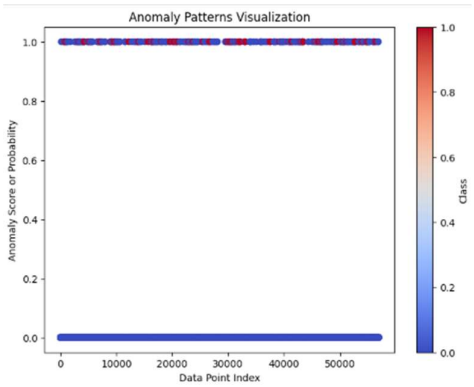
Component 1 vs Component 2 in classification of Anomaly and Normal.



Precision – Recall curve



ROC curve



Anomaly Patterns

Conclusion and Insights:

we addressed the class imbalance issue in the dataset by using the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class (fraudulent transactions). This helped balance the class distribution and improve model performance.

To tackle the high dimensionality of the data, we applied Principal Component Analysis (PCA) for dimension reduction.

we selected the best-performing model based on our chosen evaluation metrics. This model demonstrated the highest precision and recall for fraud detection.

Conclusion -

The credit card fraud detection project successfully addressed the class imbalance and high dimensionality challenges in the dataset.

The selected model, after thorough tuning and evaluation, exhibited strong performance in detecting fraudulent transactions.

Challenges Faced during the Project –

Required high training time to find the best Hyper Parameters using Grid Search CV.

Name :Teeguri Prasanna Kumar Reddy

Mail Id : 20131a4258@gvpce.ac.in

Mobile no: 6305919455