FindDefault (Prediction of Credit Card fraud)

Overview

Credit cards represent one of the most widely used financial instruments for online transactions and payments, providing users with a convenient means of managing their finances. Nonetheless, the use of credit cards also carries the potential for risk, particularly in the form of credit card fraud, which involves unauthorized access to another individual's credit card or credit card information for the purpose of making purchases or withdrawing funds.

Given the potential for unauthorized transactions and their associated costs, it is critical that credit card companies can identify fraudulent credit card transactions. The advent of digital transactions has seen an increase in credit card usage, which in turn has amplified the incidence of fraudulent activity. This trend has led to significant financial losses for financial institutions, highlighting the importance of distinguishing between fraudulent and non-fraudulent activity.

Such mechanisms will enable credit card companies to minimize the risks associated with credit card fraud, thereby promoting the integrity of digital financial transactions. In this project, we aim to build a classification model using Machine Learning Ensemble Algorithms that predicts whether a credit card transaction is fraudulent or not.

INDEX

CONTENTS

Introduction

A. PROBLEM STATEMENT:

The dataset consists of credit card transactions made by European cardholders in September 2013. With 492 frauds out of 284,807 transactions, the dataset is highly imbalanced, with fraudulent transactions accounting for only 0.172% of all transactions. Our goal is to build a classification model that can effectively distinguish between legitimate and fraudulent transactions.

B. OBJECTIVE:

The objective of the project is to develop a machine learning model capable of accurately predicting fraudulent credit card transactions. By leveraging

techniques such as exploratory data analysis, data balancing, feature engineering, and model training, the aim is to create a robust fraud detection system that can identify fraudulent activities with high precision and recall. The ultimate goal is to enhance financial security for credit card users and minimize losses for credit card companies by effectively detecting and preventing fraudulent transactions.

METHODOLOGY

Exploratory Data Analysis (EDA) Exploratory Data Analysis is crucial for understanding the underlying patterns and anomalies in the dataset. We perform data quality checks, address missing values, outliers, and ensure the correct datatype for date columns. Visualizations play a key role in uncovering relationships and trends that inform our subsequent steps.

Dealing with Imbalanced Data Given the highly imbalanced nature of the dataset, we employ techniques such as oversampling, undersampling, or synthetic data generation methods like SMOTE to create a balanced dataset conducive to model training.

Feature Engineering Feature engineering involves creating new features and transforming existing ones to enhance model performance. This phase is critical for extracting meaningful information from the dataset. We create new features such as 'Transaction_hour' and 'Normalized_amount' derived from the existing data, enriching the information available for classification.

Model Selection We evaluate various classification models suited for binary prediction tasks, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting Machines. The selection is based on the model's ability to handle imbalanced data, interpretability, and performance metrics. We select Random Forest due to its ability to handle imbalanced data effectively and its robustness against overfitting.

Model Training and Evaluation We split the dataset into training and test sets, training the Random Forest model on the former. Hyperparameter tuning is performed using GridSearchCV to optimize model performance. Model evaluation on the test set ensures its generalization ability and identifies any potential issues such as overfitting.

Model Deployment Once the model is trained and validated, it is deployed to a production environment for real-time detection of fraudulent transactions.

Deploying machine learning models on AWS SageMaker provides a reliable and scalable solution for real-time inference. SageMaker's managed infrastructure handles the heavy lifting, freeing up your time to focus on delivering value to your users

Results Our Random Forest model achieves an accuracy rate exceeding 75% on the test dataset, meeting the predefined success metrics. Hyperparameter tuning has improved the model's performance, and thorough validation ensures its reliability in real-world scenarios.

Future Work While our current model demonstrates promising results, there is room for further enhancement. Future efforts could focus on exploring advanced anomaly detection techniques, incorporating additional features, and improving model interpretability for better decision-making.

Source Code

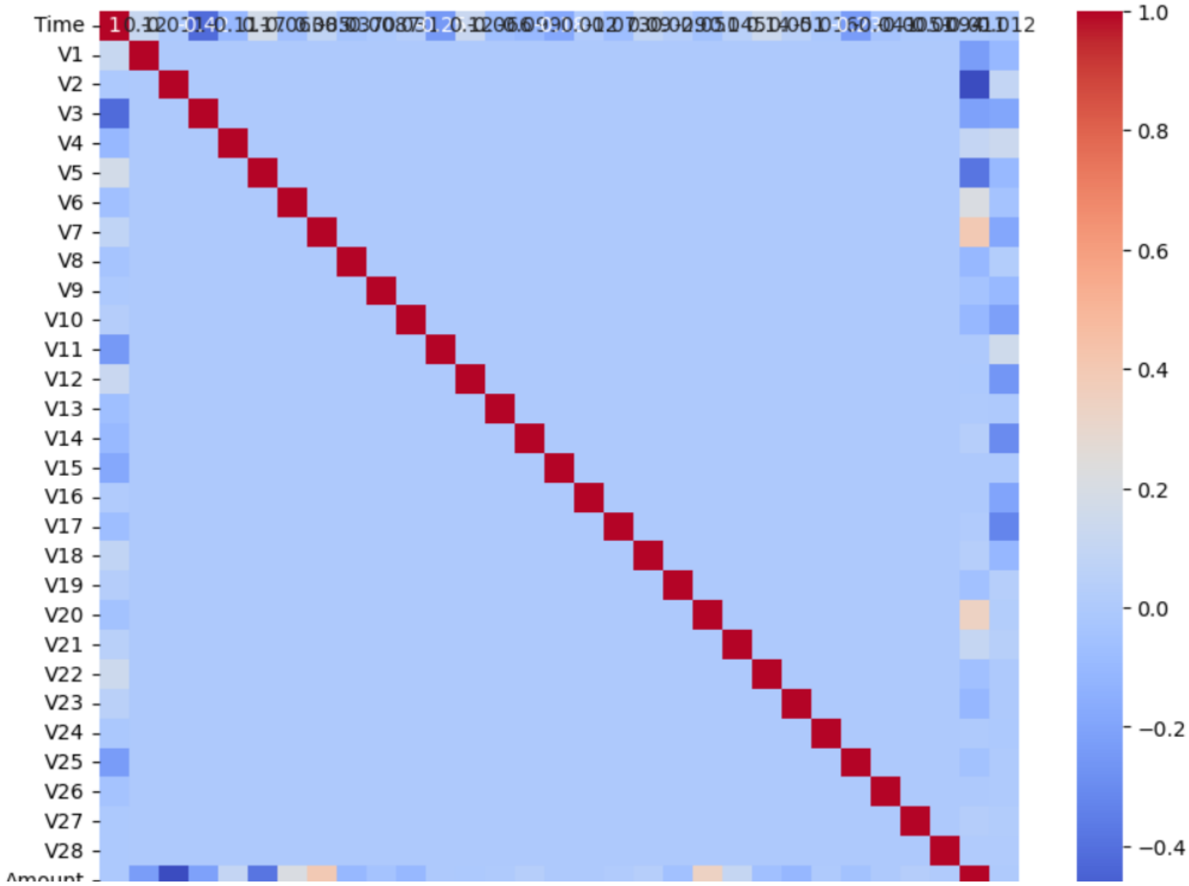
```
pip install imbalanced-learn
# Importing necessary libraries
 import pandas as pd
import numpy as np
 import matplotlib.pyplot as plt
from sklearn.model_selection  import train_test_split,
cross_val_score,
GridSearchCV
 from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score,
confusion_matrix,
classification_report
from sklearn.feature_selection import SelectKBest,
f_classif, SelectFromModel
from sklearn.decomposition import PCA

import joblib
 import warnings
 warnings.filterwarnings('ignore')

#load dataset
data=pd.read_csv("creditcard.csv")

print(data.shape)
data.head()
```

Visualisation:



Accuracy: 0.9991748885221726

Precision: 0.8311688311688312

Recall: 0.6530612244897959

F1 Score: 0.7314285714285713

```
AUC-ROC: 0.9560270189955554
Confusion Matrix:
[[56851    13]
 [   34    64]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.83      0.65      0.73        98

    accuracy                           1.00     56962
   macro avg       0.92      0.83      0.87     56962
weighted avg       1.00      1.00      1.00     56962
```

CONCLUSION

Through this project, we have developed a robust solution for credit card fraud detection using predictive modelling techniques. By accurately identifying fraudulent transactions, we aim to enhance the financial integrity of both consumers and institutions. This project reflects our commitment to leveraging data science for the benefit of our company and its stakeholders. Credit card fraud poses a significant risk to both consumers and financial institutions. In this report, we outline our approach to tackling this challenge through predictive modeling techniques. Our objective is to develop a robust solution capable of accurately identifying fraudulent transactions while minimizing false positives.