A REPORT ON

“**Preventing Credit Card Fraud : An ML Approach**”

Computer Science, College of Engineering and Applied Sciences

Advisor: Prof. Rohit J Kate

Submitted By:

Radhika Patali Atchakolu

Email: - atchako2@uwm.edu

Abstract

The amount of credit card fraud has grown significantly over the years. Due to the growing volume of daily digital trades that make credit cards more vulnerable to fraud, they are the most commonly used digital instalment method. Credit card organisations are looking for the best innovations and circumstances to identify and reduce exchange extortion at the credit score card. The objective is to identify specific instances of false extortion.

Neural groups, genetic algorithms, and k-manner bunching are some of the structures that can be used to identify credit card falsification. Over $4 trillion has been generated globally via persistent misrepresentation within the economy. The solution is to rely on project-wide data storage capabilities and cutting-edge field research techniques that support the use of automatic reasoning (AI) and AI (ML) approaches to deal with live one stride in the front of lawbreakers. ML competencies, misrepresentation and consistence devices can make investments their strength handling extra-complicated extortion issues.

We presented a machine learning-based transaction fraud detection model using feature engineering. The algorithm can learn from its mistakes and thus improve. You can improve stability and performance by processing as much data as possible. These algorithms can be used in the detection of online fraud transactions. A dataset of specific online transactions is used in these. Then, with the help of machine learning algorithms, we can identify unique or unusual data patterns that can be used to detect fraudulent transactions. For the best results, the XGBoost algorithm, which is a cluster of decision trees, will be used. This algorithm has recently taken over the world of machine learning. This ML algorithm outperforms others in terms of accuracy and speed.

Keywords: Fraud detection, Machine learning, Xgboost algorithm, classification, Data pre-processing, Prediction.

ACKNOWLEDGEMENT

I am deeply grateful to Professor Rohit J Kate for his valuable guidance and support during my project on "Preventing Credit Card Fraud: An ML Approach." His expertise in machine learning and patient teaching have been instrumental in my growth and understanding of this field. I am especially thankful to Professor Rohit J Kate for introducing me to the subject of Machine Learning through his course that I developed a deep interest in the field, leading me to aspire to work under his guidance. I am grateful for the opportunity to learn from him and for the inspiration he has provided throughout this journey.I am honored to have been his student and confident that the knowledge gained under his mentorship will prove invaluable in my future endeavors.

Introduction

Credit card fraud poses a significant threat to modern business relationships, and with technological advancements, new opportunities for fraudulent activities have emerged. This exposes organisations to substantial financial, operational, and psychological risks. Despite the advantages of electronic payments, credit card companies are experiencing an increase in fraudulent incidents due to evolving technologies.

In 2018, estimated losses from Visa fraud alone reached $24.26 million, and globally, fraud losses amounted to $27 billion in 2019. To address this issue, preventive measures and the study of fraudulent behaviour are crucial for minimising and preventing future occurrences.

Machine learning and data science communities play a vital role in finding automated solutions. However, the problem is challenging due to factors such as class imbalance and changing transaction patterns. Implementing a real-world fraud detection system involves automated scanning of payment requests, machine learning analysis of authorised transactions, and reporting of suspicious activities for investigation.

Feedback from professionals is used to train and update the algorithm, continually improving fraud detection performance. This project aims to leverage machine learning to develop an effective fraud detection system.

Motivation

With the rapid expansion of businesses worldwide, the need to provide excellent services to customers has grown significantly. To achieve this, companies handle vast amounts of data on a daily basis, including personal and financial information of their clients. Therefore, data security is of utmost importance to safeguard against exploitation or theft by other entities. In particular, the rise of online shopping has led to an increase in fraudulent activities such as Trojan attacks and spoofing, where criminals steal cardholder information.

Detecting credit card fraud has become a crucial area of interest for machine learning and computational intelligence, leading to the development of various automated solutions. As businesses accumulate vast volumes of data from diverse sources, including social media interactions and user purchasing behaviours, analysing and visualising this data allows for the identification of hidden patterns.

Overview

A. Recommendation systems

1. Transaction Monitoring: Develop a recommendation system that monitors and analyses transaction patterns in real-time. This system can identify potentially fraudulent activities based on anomalies or suspicious patterns, such as large or unusual transactions, frequent changes in purchase locations, or unusual time intervals between transactions.

2. User Behavior Analysis: Build a recommendation system that examines user behaviour and identifies deviations from normal patterns. By analysing factors such as spending habits, transaction frequency, and purchase categories, the system can detect any abnormal activities that may indicate fraudulent behaviour.

3. Risk Scoring: Implement a recommendation system that assigns risk scores to individual transactions or users. By considering various features, such as transaction amount, merchant reputation, and user history, the system can assess the likelihood of fraud. Based on these risk scores, appropriate actions can be taken, such as flagging high-risk transactions for further verification or declining suspicious transactions.

Main Contributions & Objectives

The rise of online transactions, particularly credit card usage, has brought about an increased risk of financial loss due to physical card theft or unauthorized access to credit card information. Hackers actively engage in fraudulent activities, posing a significant threat to individuals. Recognizing the need to detect fraudulent transactions and protect online credit card transactions, we developed a novel system named the "Online Credit Card Fraud Detection and Prevention System" that leverages machine learning techniques.

Our objective was to analyze this problem comprehensively and devise effective measures to combat credit card theft. In this system, we experimented with various algorithms to assess its accuracy. Notably, the system employs the XGBoost algorithm to determine the legitimacy of each transaction, effectively preventing credit card fraud.

Our primary objective is to accurately identify and differentiate fraudulent credit card transactions from legitimate ones. Various techniques, including neural networks, genetic algorithms, and k-means clustering, can be employed to achieve this goal. Given the staggering global cost of fraud, exceeding $4 trillion, it is crucial to utilize advanced investigative methods and robust data storage capabilities.

By harnessing the power of artificial intelligence (AI) and machine learning (ML), we aim to proactively combat fraudulent activities and stay ahead of criminals. Implementing algorithms such as Logistic Regression, Support Vector Machine (SVM), Decision Tree (DT), and XgBoost will enhance our fraud detection capabilities. This approach enables fraud and compliance teams to focus their efforts on addressing more sophisticated fraudulent concerns.

Credit card fraud leads to significant financial losses, often perpetrated through techniques like Trojan attacks and phishing. Detecting fraud is crucial, and this study utilizes algorithms to train behavior features of normal and fraudulent transactions. Data analytics uncovers hidden patterns for informed decision-making, despite challenges in available datasets. Multiple machine learning algorithms are used to identify fraudulent transactions, with oversampling techniques applied. Logistic Regression, XGBoost, DT, and SVM algorithms exhibit high accuracy in detecting credit card fraud.

Literature Survey

"A Comparison of XGBoost" - Gonzalo Martinez:

1. Evaluates XGBoost's performance in terms of training speed, generalization, and parameter setup.
2. Compares XGBoost with random forests and gradient boosting.
3. Considers carefully tuned models and default settings.
4. Highlights advantages of XGBoost over other algorithms.
5.

"Detection of Credit Card Fraud in Data Mining Using the XGBoost Classifier" - Rahul Goyal, Amit Kumar Manjhvar, and Vikas Sejwar:

1. Uses XGBoost classifier to detect credit card fraud.
2. Combines SMOTE technique and XGBoost for balancing the dataset.
3. Validates performance using publicly available datasets.
4. Demonstrates effectiveness of the system in fraud detection.

XGBoost

The XGBoost algorithm is an ideal choice for preventing credit card fraud through machine learning. Here's a concise explanation of why XGBoost is suitable for this application:

1. Handling Imbalanced Data: Credit card fraud detection involves imbalanced datasets. XGBoost effectively handles imbalanced data by incorporating techniques like SMOTE, which balances the dataset and improves fraud detection accuracy.
2. Gradient Boosting: XGBoost's gradient boosting technique focuses on difficult-to-classify instances, such as fraudulent transactions, enhancing the model's ability to detect fraud effectively.

3. Feature Importance: XGBoost provides insights into feature importance, helping identify crucial factors associated with fraudulent activities, guiding feature selection, and aiding in understanding fraud patterns.

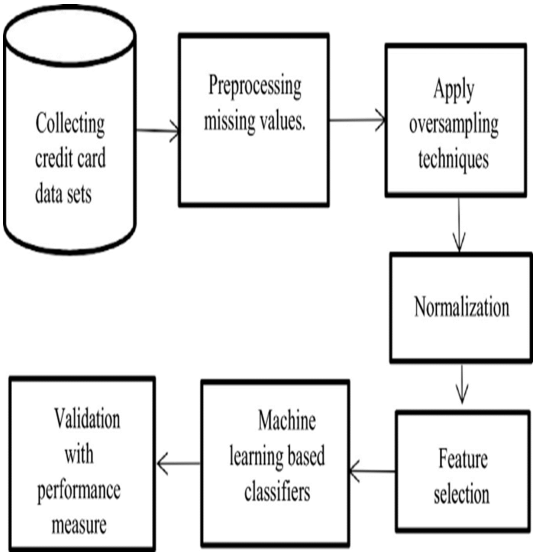
4. Robustness and Generalization: XGBoost's regularization techniques prevent overfitting and improve generalization, ensuring accurate fraud detection in unseen data.

5. High Performance and Scalability: XGBoost efficiently handles large datasets, enabling fast training and real-time or near-real-time fraud detection.

6. Tunability and Model Interpretability: XGBoost's customizable hyperparameters allow optimization for specific requirements, while feature importance analysis aids in model interpretability.

Overall, XGBoost's ability to handle imbalanced data, gradient boosting, feature importance analysis, robustness, scalability, and tunability make it a powerful choice for preventing credit card fraud using machine learning.

Architecture



IMPLEMENTATION

1. *Data Set:* The code starts by importing the necessary libraries and loading the train and test datasets. The dataset is obtained from Kaggle and has "fraudTrain.csv" and "fraudTest.csv" files. Test data has 555720 records, and the training data has 1296676 records.Both files have columns such as 'trans_date_trans_time', 'cc_num', 'merchant', 'category', 'amt', 'first', 'last', 'gender', 'street', 'city', 'state', 'zip', 'lat', 'long', 'city_pop', 'job', 'dob', 'trans_num', 'unix_time', 'merch_lat', 'merch_long', 'is_fraud', Unnamed: 0.

2. *Data Analysis and Visualization:* The code performs initial analysis and visualisation of the data. It includes generating a heatmap to visualise the correlation between different features in the dataset. It also creates bar plots to show the number of frauds by gender and category.

3. *Data Preprocessing:* The code performs several preprocessing steps on the data. It first downsampled the majority class (is_fraud=0) to balance the dataset. Then, it transforms the data by converting categorical variables (category, gender, merchant, job) into numerical representations using one-hot encoding and label encoding. It also calculates the age of the customers based on their date of birth.
4. *Data Split:* The code splits the preprocessed data into training and test sets. It selects specific features (amt, gender, age, category) as input variables (X) and the "is_fraud" column as the target variable (y).

5. *Data Scaling:* The code applies min-max scaling to the input variables in both the training and test sets using the MinMaxScaler from the scikit-learn library. This step scales the features to a specific range (usually between 0 and 1).

6. *Model Building:* The code builds several machine learning models for fraud detection, including Support Vector Machine (SVM), Decision Tree Classifier, Logistic Regression, XGBoost, and XGBoost with Hyperparameter Tuning. The models are trained on the training data (X_train, y_train).

7. *Model Evaluation:* For each model, the code makes predictions on the test data (X_test) and evaluates the performance using various metrics such as accuracy, confusion matrix, and classification report. It also generates a heatmap of the confusion matrix to visualize the model's performance.

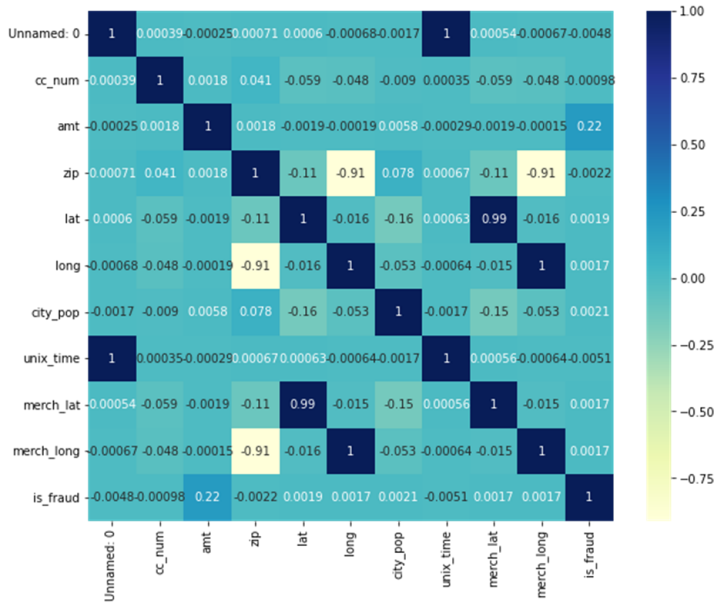
8. *ROC Curve:* The code plots the Receiver Operating Characteristic (ROC) curve for the Logistic Regression model. The ROC curve is used to assess the trade-off between the true positive rate and the false positive rate.

9. *Hyperparameter Tuning:* The code uses the hyperopt library to perform hyperparameter tuning for the XGBoost model. It defines a search space for different hyperparameters and uses a combination of random search and Bayesian optimization to find the best set of hyperparameters that maximize the model's performance.

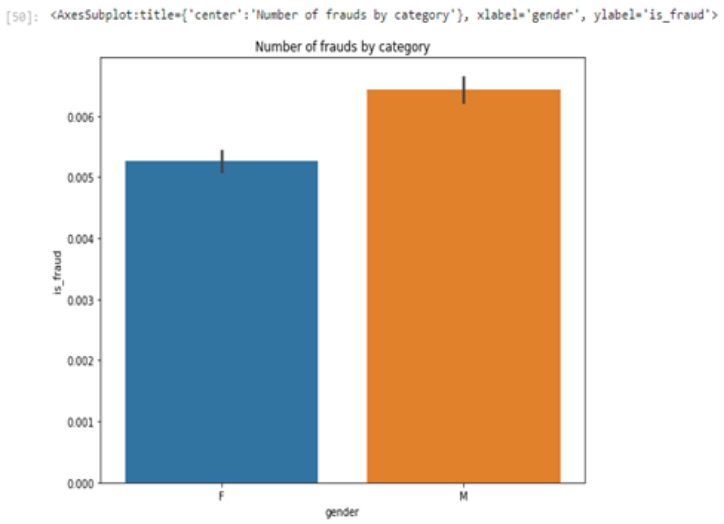
10. *Best Hyperparameters:* After performing hyperparameter tuning, the code prints the best set of hyperparameters found for the XGBoost model.

Results

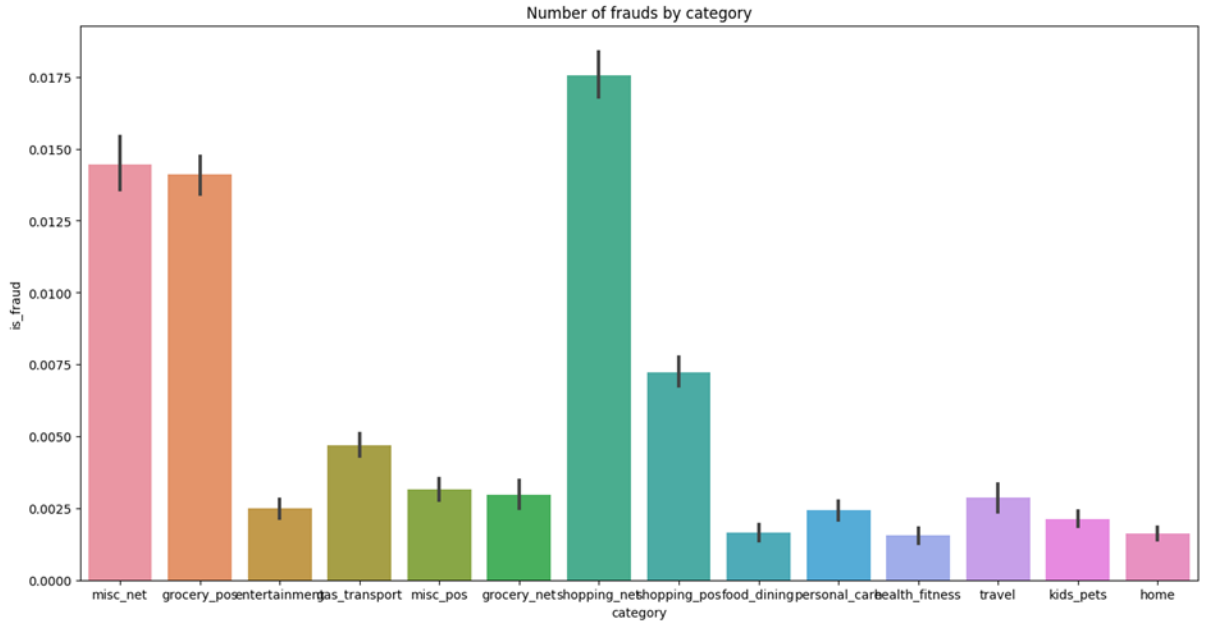
Correlation matrix



Box Plot



Bar Graph

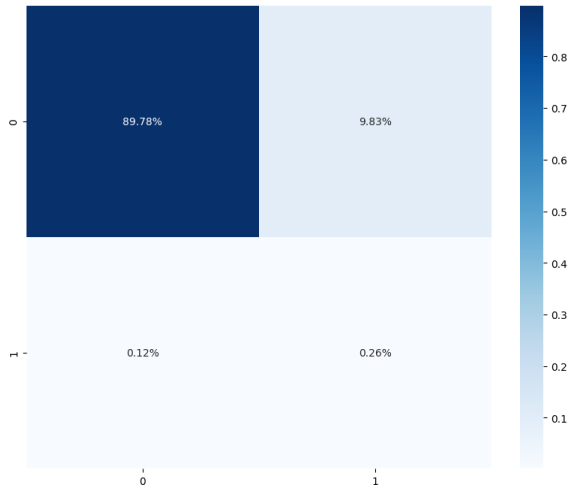


Data Columns

```
Int64Index: 15012 entries, 123118 to 1295733
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             15012 non-null  int64
 1   trans_date_trans_time  15012 non-null  object
 2   cc_num                 15012 non-null  int64
 3   merchant               15012 non-null  object
 4   category               15012 non-null  object
 5   amt                    15012 non-null  float64
 6   first                  15012 non-null  object
 7   last                   15012 non-null  object
 8   gender                 15012 non-null  object
 9   street                 15012 non-null  object
 10  city                   15012 non-null  object
 11  state                  15012 non-null  object
 12  zip                    15012 non-null  int64
 13  lat                    15012 non-null  float64
 14  long                   15012 non-null  float64
 15  city_pop               15012 non-null  int64
 16  job                    15012 non-null  object
 17  dob                    15012 non-null  object
 18  trans_num              15012 non-null  object
 19  unix_time              15012 non-null  int64
...
 21  merch_long             15012 non-null  float64
 22  is_fraud               15012 non-null  int64
dtypes: float64(5), int64(6), object(12)
memory usage: 2.7+ MB
```

SVM

Confusion Matrix



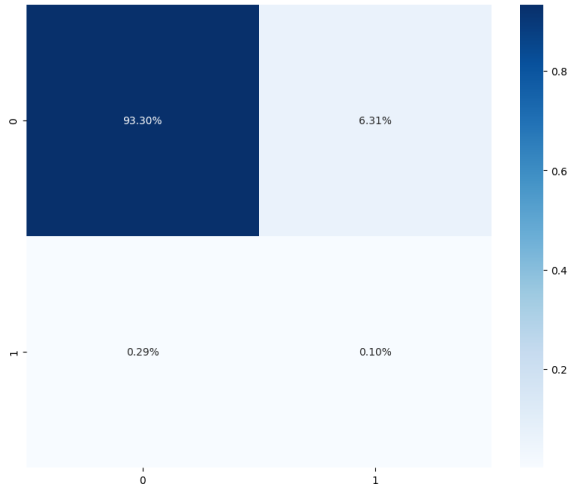
Classification Report

Classification report

	precision	recall	f1-score	support
0	1.00	0.90	0.95	553574
1	0.03	0.69	0.05	2145
accuracy			0.90	555719
macro avg	0.51	0.79	0.50	555719
weighted avg	0.99	0.90	0.94	555719

Decision Tree Classifier

Confusion Matrix



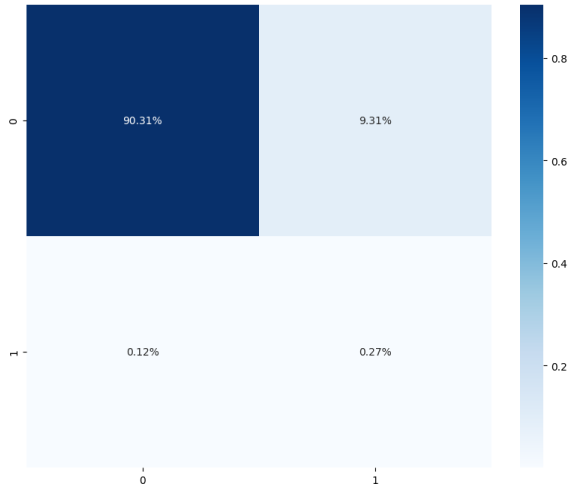
Classification Report

Classification report

	precision	recall	f1-score	support
0	1.00	0.94	0.97	553574
1	0.02	0.25	0.03	2145
accuracy			0.93	555719
macro avg	0.51	0.59	0.50	555719
weighted avg	0.99	0.93	0.96	555719

Logistic Regression

Confusion Matrix



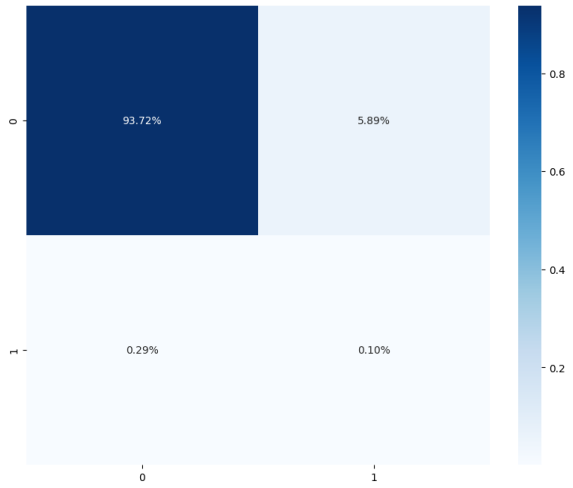
Classification Report

Classification report

	precision	recall	f1-score	support
0	1.00	0.91	0.95	553574
1	0.03	0.69	0.05	2145
accuracy			0.91	555719
macro avg	0.51	0.80	0.50	555719
weighted avg	0.99	0.91	0.95	555719

XGBoost

Confusion Matrix

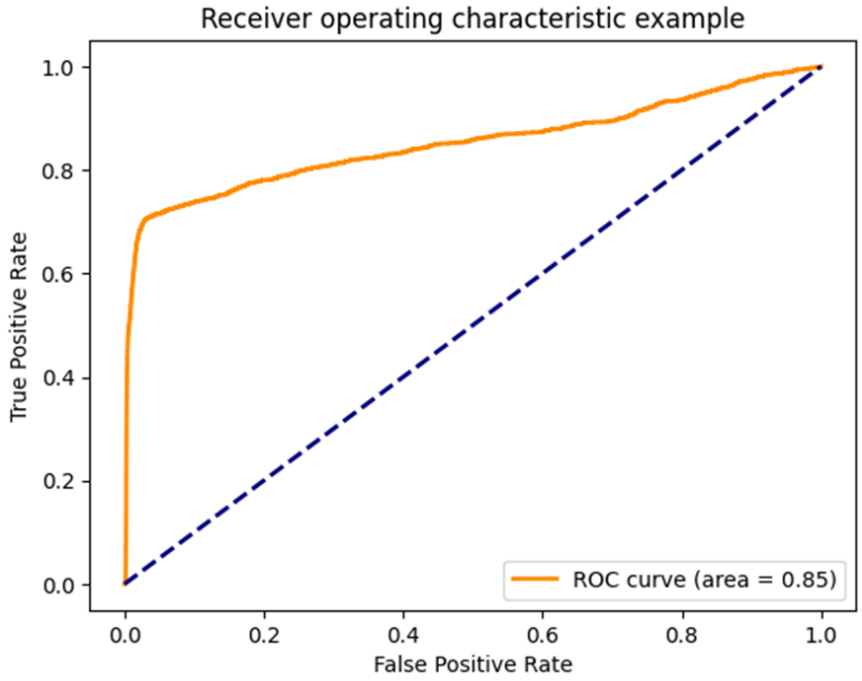


Classification Report

Classification report

	precision	recall	f1-score	support
0	1.00	0.94	0.97	553574
1	0.02	0.25	0.03	2145
accuracy			0.94	555719
macro avg	0.51	0.60	0.50	555719
weighted avg	0.99	0.94	0.96	555719

ROC_Curve



XGBoost with Hyper Parameter

Best Parameters

{

'colsample_bytree': 0.8356284452522001,

'gamma': 3.5024545402447966,

'max_depth': 8.0,

'min_child_weight': 3.0,

'reg_alpha': 151.0,

'reg_lambda': 0.5715472987377122

}

Conclusion

This project involved creating a machine learning model for credit card fraud detection, which went through several steps including data analysis, pre-processing, and model building. The model's accuracy was evaluated using cross-validation techniques and various metrics, and it achieved good results on the test dataset.

The approach used in this project could be extended to other areas, such as loan approvals and insurance claims.

In this project, I have used various techniques to improve the accuracy and efficiency of the model, such as hyperparameter tuning.

I have got similar accuracy as compared to other codes which are in Kaggle. the main difference between my project and those are they have been done with less no.of.parameters.

References

[1] "A Comparison of XGBoost"- Gonzalo, Martinez

[2] "Detection of Credit Card Fraud in Data Mining Using the XGBoost Classifier" - Rahul Goyal, Amit Kumar Manjhvar, and Vikas Sejwar

[3] S. Bachmayer, "Artificial Immune Systems," pp. 119-131 in Artificial Immune Systems, vol. 5132, 2008.M. Krivko, "A Hybrid Model for Plastic Card Fraud, "Expert Systems with Applications, vol. 37, no. 8, pp. 6070-6076, August 2010.

[4] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, Feb. 2011, pp. 602-613.

[5] "Plastic card fraud detection via peer group analysis," Advances in Data Analysis and Classification, vol. 2, no. 1, pp. 45-62, Mar. 2008.

[6] O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, "Credit Card Fraud Detection Using Machine Learning As A Tool.