

Credit Card Fraud Detection Using ML Algorithms

R. Charan¹, K. Kaushanthi Devi², G. Sai Teja³, G. Sai Siddhartha⁴, D. Sitharam⁵

B. Tech Students, Department of Information Technology^{1,2,3,4,5}

Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: Bank credit card default rates are (as predicted by S&P dow Jones index) at all time high for the past few months. This possess a real challenge for banks and other lending financial institutions, as they are now more than ever are in need of a robust risk prediction model to generalize the economic behavior of their current and potential future clients. In real world massive stream of payment requests are coming in every second which is huge challenge to monitor. Machine Learning algorithms help in solving these huge challenges to an extent. These predictive models would not only benefit the lending institutions but also the customer as it would make them more aware of their potential defaulting rate. The goal of this project is to resolve this problem by building and comparing predictive risk models using supervised learning algorithms. We use python libraries to perform various Machine learning algorithms on the credit card dataset and data diagrams to have better visualization of data. Here the credit card fraud detection is done on the basis of some attributes in the dataset like the time factor and the amount factor. The accuracy of the algorithms is also being determined and the most efficient one is said with the highest accuracy obtained..

Keywords: Numpy, pandas, sklearn, logistic regression, knn, decision trees, random forest, XG boost, data visualization, pyplot

I. INTRODUCTION

In the present era due to digitalization everything is digitized and the use of online transactions has increased rapidly. This also led to the foundation of digital scams and the fraud of credit card. Fraud is the crime of gaining money or financial benefits. Most of the banking companies are trying to minimize the fraud activities and reduce the risk for the customers. The other ways in which it could be interpreted is that a person using some other credit card for a transaction without proper permission or in some other case the owner is unaware of the transaction. This enormous amount of credit card fraud can be prevented by the use of Machine learning algorithms and we can extract better solutions to detect fraudulent activities and reduce the number of activities being done.

This problem is challenging due to the continuous and unbalanced data. As the enormous transactions taking place within a second it's difficult for a human to look after every transaction. Hence a virtual machine is needed to detect the fraudility of the transaction. Therefore, Machine Learning algorithms help in solving these huge challenges to an extent.

II. PROPOSED SYSTEM

The goal of this project is to resolve this problem by building and comparing predictive risk models using supervised learning algorithms. The idea is to identify the credit card frauds using various machine learning algorithms and make a comparison among them to find the efficient algorithm that has the highest accuracy rate. In this we perform logistic regression, k nearest neighbor, decision tree, random forest and XG Boost. We also use matplotlib.pyplot , seaborn and t-distributed stochastic neighbor embedding for better visualization of the large data.

III. TECHNOLOGIES USED

3.1 Google Colab

Colab is used for machine learning purpose and it also provides a platform that allows anybody to learn and execute python code.

This project is done in a colab notebook to run the python code by uploading the dataset on which we perform certain required operations.

3.2 Python Libraries

There are several libraries available in python for data visualizations for better understanding of the data. The numpy is used for array operations and also matrices. Pandas is used to analyze data, it is widely used in machine learning tasks. Pyplot can be imported using import matplotlib. Pyplot is a sub-module of the matplotlib library for Python. It consists of various methods for plotting two dimensional graphs.

3.3 Data

The data has been taken from the kaggle website. It is in the form of a single csv table or an excel sheet. Then we scale the attributes required, split the data into train and test data and perform the algorithms and get the efficient algorithm with highest accuracy.

IV. FLOW DIAGRAM

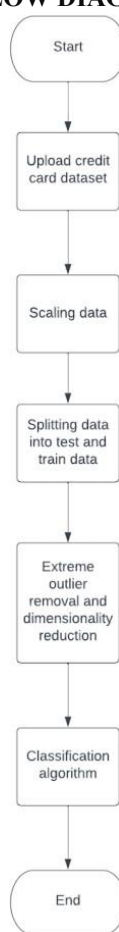


Figure 1: Flowchart Diagram for credit card fraud detection system

The flowchart is nothing but a type of diagram that represents an algorithm or process, showing the steps as boxes of various kinds. In our project the flowchart has started from the uploading of Credit card dataset. Scaling of the amount and time. Then as a next step we will split the data into test and train data and then we will perform the dimensionality reduction on it and use various classification algorithms like the knn, logistic reduction, naive bayes, decision tree, random forest and the score calculation for the accuracy of detection is measured and the best model with the highest accuracy is predicted.

V. IMPLEMENTATION

- First for the project we have to open a new notebook in the google colab.
- Before doing the project we must know the predefined functions and we should now have to user defined

functions.

- We need to import the dataset.
- Scaling Amount and Time.
- Splitting Data into Train and Test.
- Creating a subsample data set with balanced class distributions.
- Extreme Outlier Removal.
- Dimensionality Reduction.
- Classification Algorithms.

VI. RESULTS

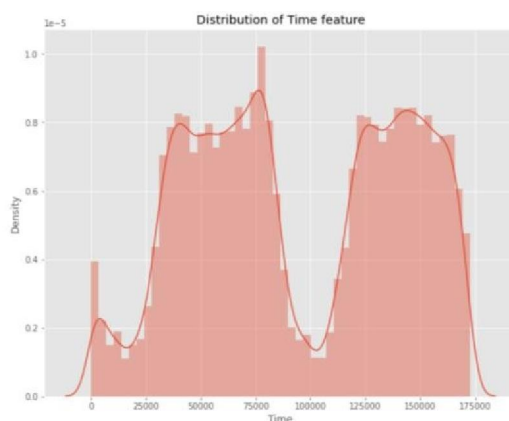


Figure 2: Distribution of Time Feature

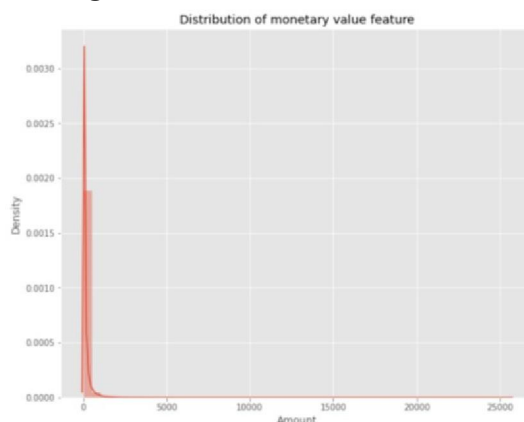


Figure 3: Distribution of Monetary value feature
count of fraudulent vs.non-fraudulent transactions

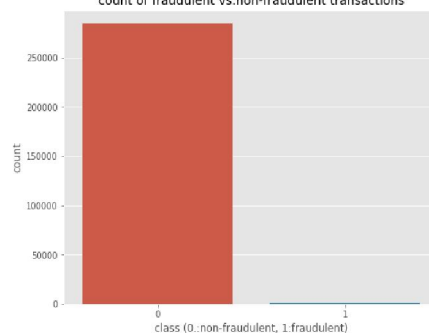


Figure 4: Count of transactions

Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. The number of non-fraudulent transactions are represented using 0 and the number of fraudulent transactions are represented using 1. This graph shows the times at which transactions were done within two days. Here we can observe that we have very few fraud transactions.

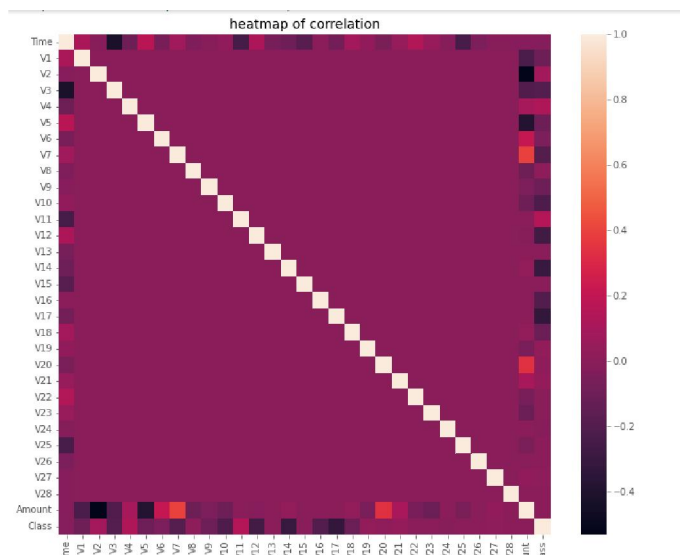


Figure 5: Heatmap of correlation

After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable. This shows us better visualizations of the data using various colors.

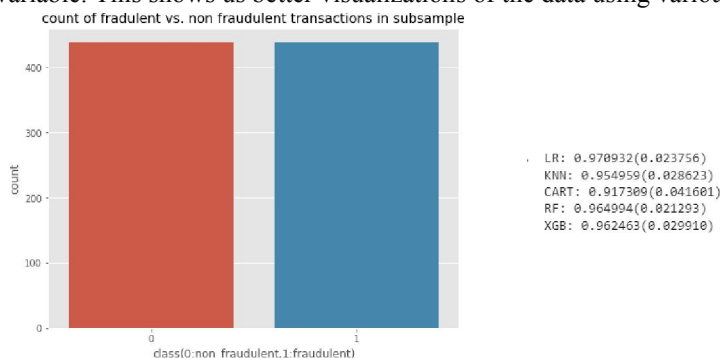


Figure 6: Accuracy of algorithms

We have got the highest accuracy with the logistic regression model followed by random forest, xgb, knn and then decision tree.

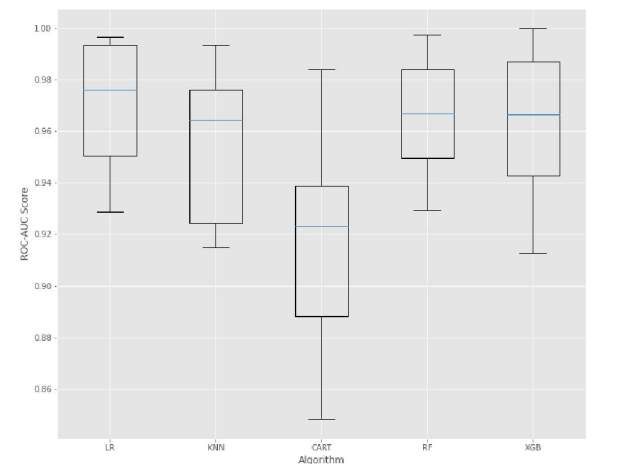


Figure 7: Classification of algorithms

Comparison of classification algorithm, logistic regression (LR), k nearest neighbour (KNN), Decision tree (CART), XG Boost (XGB). The accuracy is highest here for logistic regression.

VII. SCOPE FOR FUTURE WORK

We didn't end up with our goal of 100% accuracy, which means that there is future work to do to make it 100% accurate. Since the data science field is growing exponentially there will be an improvement in these algorithms to make it 100% accurate. We have analyzed the dataset with many algorithms to get a better result. Future scope of this project is to combine these algorithms and get a 100% accuracy of the system. More room for improvement can also be found in the dataset. These processes will make the system more accurate and reduce false positives.

VIII. CONCLUSION

This project focuses on detecting and minimizing credit card fraud activities with different machine learning algorithms. This paper also explained in detail how such machine learning models are trained to get a better result with implementation and experimental results. The best algorithm reached over 98% accuracy and 99% precision. Recall and F-score also are quite high. This is possible because of the effective cleaning of data. Pre-processing the data is a crucial step in the analysis of machine learning algorithms. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

REFERENCES

- [1]. Dataset: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>
- [2]. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE Transactions On Neural Networks And Learning Systems, Vol. 29, No. 8, August 2018
- [3]. "Credit Card Fraud Detection through Parenclitic Network Analysis- By Massimiliano Zanin, Miguel Romance, Regino Criado, and Santiago Moral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages
- [4]. "Credit Card Fraud Detection Based on Transaction Behaviour –by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017
- [5]. Clifton Phua¹, Vincent Lee, Kate Smith & Ross Gayler "A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia
- [6]. "Survey Paper on Credit Card Fraud Detection by Suman", Research Scholar, GJUS&T Hisar HCE, Sonapat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014
- [7]. <https://towardsdatascience.com/credit-card-fraud-detection-using-machine-learning-python-5b098d4a8edc>
- [8]. "Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence
- [9]. "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016