Credit Card Fraud Detection

by

Harsh Yadav

2020IMG-026

A report submitted for Summer Internship Project

Integrated Bachelor of Technology

in

IPG-MBA

ATAL BIHARI VAJPAYEE

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT

GWALIOR - 474015, MADHYA PRADESH, INDIA

Report Certificate

I hereby certify that the work being presented in the report, entitled Credit Card Fraud detection, Integrated Bachelor of Technology in IT + MBA and submitted to the institution is an authentic record of my/our own work carried out during the period *July 2022* to *August 2022*. I also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

_____

Dr. Vinal Patel
Dr Gaurav Kaushal
Dr Santosh Singh Rathore
Dr K.V Arya

Date_____

The final copy of this report has been examined by the signatories and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above-mentioned discipline.

Candidate's Declaration

I hereby certify that I have properly checked and verified all the items as prescribed in the checklist and ensured that my thesis is in the proper format as specified in the guideline for thesis preparation.
I declare that the work contained in this report is my own work. I understand that plagiarism is defined as anyone or combination of the following:

(1) To steal and pass off (the ideas or words of another) as one's own

(2) To use (another's production) without crediting the source

(3) To commit literary theft

(4) To present a new and original idea or product derived from an existing source.

I understand that plagiarism involves an intentional act by the plagiarist of using someone else's work/ideas completely/partially and claiming authorship/originality of the work/ideas. Verbatim copy as well as a close resemblance to someone else's work constitute plagiarism. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, and websites, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.
I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report/dissertation/thesis are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable. My faculty supervisor(s) will not be responsible for the same.

Signature:

Name: Harsh Yadav

Roll. No: 2020IMG-026

Date:

Abstract

This project represents the design, and implementation process for the Minor Project (Integrated Bachelor of Technology in Information Technology + MBA 2020-2025). The project discusses the ways to detect credit card transaction frauds. In this analysis, we discuss the various methods used to reduce the number of measurement variables for a simplified model. The project also throws a shadow on challenges dealt with during the process of implementation. A clear description of the experiments conducted is laid down in the report.

Keywords: Logistic Regression, Decision Tree, Random Forest

Dedication

The use of machine learning in fraud detection has been an interesting topic now days. A credit card fraud detection aims to identify the fraudulent transactions before they happen based on the similar historical data. This detection prevents the credit card companies from fraudsters.

Acknowledgments

I want to extend my heartful thanks to **Prof. Karm Veer Arya, Dr. Vinal Patel, Dr. Santosh Singh Rathore** and **Dr. Gaurav Kaushal** for constantly guiding me through the project. They helped me to develop an excellent practice of reading recent literature before pursuing the work. This practice increased my thirst for knowledge without any doubt. Their expertise in the field of Computer Vision helped me to think better and more innovative. I am indebted to all the professors for allowing us to develop an industry-grade project in these times of peril. They gave their valuable time to evaluating and giving much-needed insights about the project.

Thanks to my family for their constant support.

I would also like to thank Dr. Andrew NG for starting a MOOC specialization in deep learning, and *the Journal of Statistics Education* for conducting the study.

Harsh Yadav

Contents

Chapter

1. Introduction

This chapter presents an overview of the context as a part of the project developed in section 1.1. Section 1.2 introduces the objectives of the project. Next, section 1.3 presents the implementation workflow step by step. Finally, in section 1.4, the result of the research carried out is briefly introduced.

1.1 Context

This project is a part of B.Tech. Information technology curriculum for the second semester. The objective is to prevent those credit card transaction which are fraudulent. In this project, we used Logistic Regression to realize the above-stated objectives.

1.2 Implementation workflow

The workflow followed during the implementation is as follows:

Step 1: Importing the important libraries

Step 2: Data collection for the Machine learning model

Step 3: Pre-processing of the data set

Step 4: Train the models

Step 5: Record results

Step 6: Comparing the results of the models to get the best one.

1.3 Objectives

- To formalization of the fraud-detection problem that realistically describes the operating conditions of frauds that everyday analyse massive streams of credit card transactions.
- To design and assess a new technique that effectively addresses credit card frauds.
- To Timely identification of fraudulent transactions can prevent the fraudsters from further committing such illicit crimes.

1.4 Research results

There is a slight increase in training and test accuracy after modifications which were done considering the result of the various methods performed.

2. Methodology

This section includes a write-up on tools and methods used while implementation. It lays down a clear description of the process used during development and testing.

2.1 Tools

There are varieties of tools used during the development of the project. The major tools used have been listed below:

- Pandas: This library for machine learning is free to use based on the Torch library, utilised for applications involving computer vision and natural language processing, primarily developed by Facebook's AI Research lab. It is open-source software that is available for free under the Modified BSD licence.
- Sklearn: Sklearn is the most useful and robust library for machine learning in Python. It provides multiple efficient tools for machine learning like classification, regression, clustering, and dimensionality reduction by way of a consistence interface in Python. The below model have been used from sklearn:
    1. Logistic Regression: Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring.
    2. Decision Tree Classifier: A decision tree is a flowchart-like structure in which each internal node represents a test on a feature, each leaf node represents a class label and branches represent conjunctions of

features that lead to those class labels. The paths from root to leaf represent classification rules.

3. Random Forest Classifier: Random Forest is a supervised machine learning algorithm that is used widely in classification and regression problem. It makes decision trees on various samples and makes a decision based on the majority vote.

- Matplotlib: Matplotlib is a data visualization and graph plotting library for Python. It provides an opensource alternative for MATLAB.

2.2 Workflow

The workflow followed during the implementation is as follows:

Step 1: Importing the important libraries

Step 2: Data collection for the Machine learning model

Step 3: Pre-processing of the data set

Step 4: Train the models

Step 5: Record results

Step 6: Comparing the results of the models to get the best one.

2.3 Conclusion

This section presented the workflow followed during the implementation of the project.

3. Experiments and results

This section discusses the analysis conducted and the corresponding results obtained. Note: These results may vary from machine to machine.

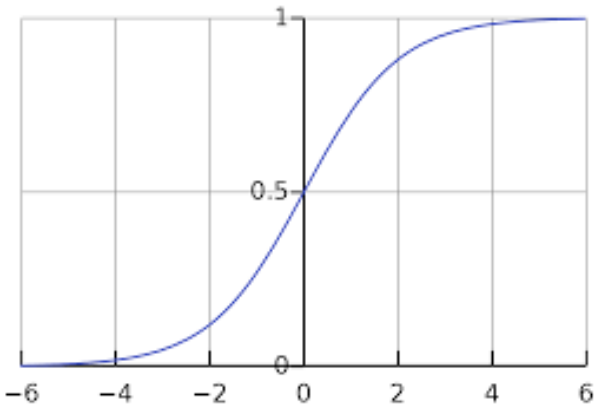
3.1 Experiment

The idea for the experiment has been inspired by the research paper Credit Card Fraud Detection using Machine Learning and Data Science (IJERT). The idea is to use various models and compare their accuracy to get the best or most accurate model. The various algorithms are:

I) Logistic regression:

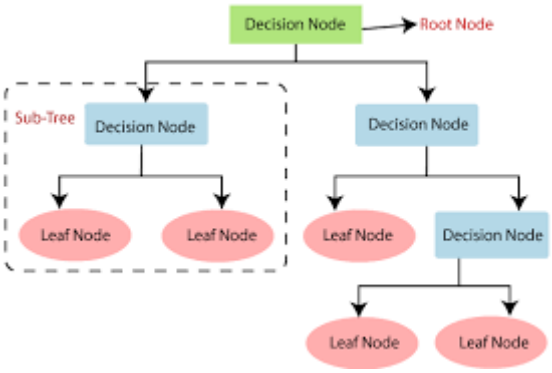
It is supervised classification algorithm for the discrete output values like true or false and yes or no. The Logistic Regression uses the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-(x)}}$$



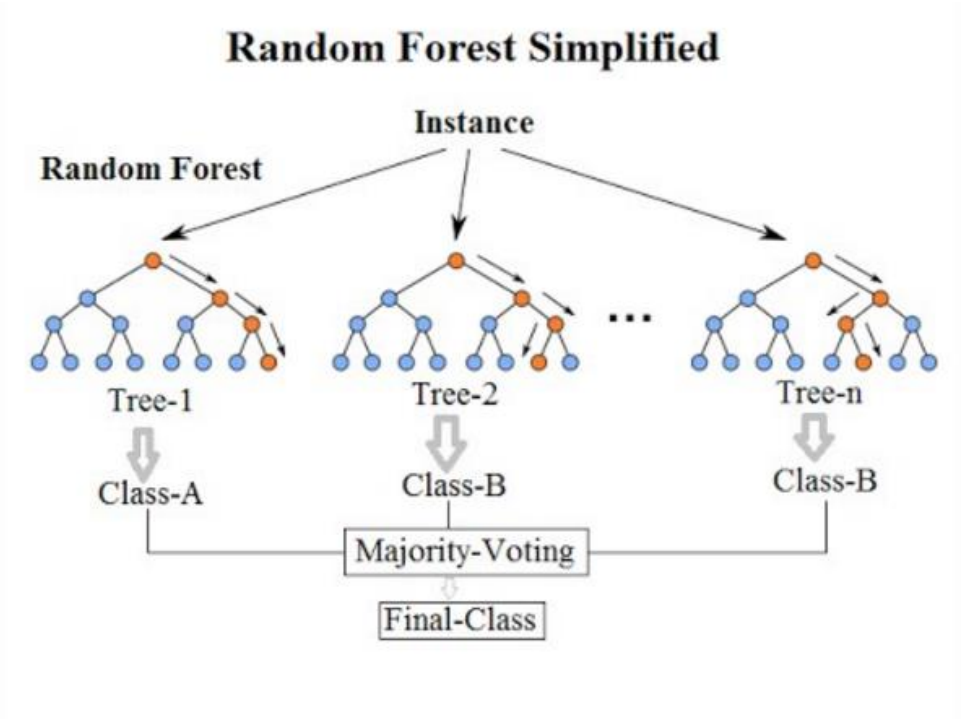
II) Decision Tree classification:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.



III) Random forest classification:

It is technique of supervised learning which can be used as classification and regression. It combines multiple decision tree to create a forest. In this algorithm each decision tree gives their classification and the random forest take the average of their classification which increases the accuracy significantly since there is chances that a tree classification has some error but the majority decision trees have not which lead the output in correct direction.



3.2 Experiment description

First, we import the important libraries and our dataset. Then we do pre-processing of the dataset. Then we will count the number of frauds happened in the past. Now we will take equal number of fraudulent and non-fraudulent data so that we can properly train our model. The non-fraudulent data is randomly taken from our dataset.

Now we will split the data into train and test data. And we will take 20% of test data and 80% of train data. Now we will create our logistic regression model. And now we will train our model using X_train and y_train in fit function. Then we will print the classification report, confusion matrix and accuracy. And similarily we do it for Decision tree classifier and Random Forest classifier.

3.3 Results and discussion

The results obtained in the above three models have been discussed below:

- Confusion matrix, accuracy, f1 score and mean squared error for the model using logistic regression is:

```
Confusion Matrix
[[86  6]
 [10 95]]
accuracy is -->  91.88
mean squared error is -->  0.08121827411167512
f1 score is -->  0.9223300970873787
```

- Confusion matrix, accuracy, f1 score and mean squared error for the model using Decision Tree Classifier is:

```
Confusion Matrix
[[92  0]
 [12 93]]
accuracy is -->  93.91
mean squared error is -->  0.06091370558375635
f1 score is -->  0.9393939393939393
```

- Confusion matrix, accuracy, f1 score and mean squared error for the model using Random Forest Classifier is:

```
Confusion Matrix
[[91  1]
 [ 7 98]]
accuracy is -->  95.94
mean squared error is -->  0.04060913705583756
f1 score is -->  0.9607843137254903
```

It can be seen that we get the best results with the third model that is using Random Forest Classifier.

3.4 Conclusion

We can conclude from the above results that the third model that is random forest classifier gives the best accuracy.

4. Discussions and conclusion

In this chapter, the work is concluded and future plan is presented. Next, the research contribution is presented. Finally, limitation of the work and possible future extensions is described respectively.

4.1 Contributions

The work presented in this paper is a step forward to a new way of thinking in Credit Card Fraud Detection. The approach is to compare different models and choose the one with best results and accuracy.

4.2 Limitations

The results presented in the paper does not have good accuracy. Therefore, there is a need for more accurate models to predict the fraud.

4.3 Future scope

There are some questions that need to be answered in the thesis. More research must be done on the idea presented in the paper using more accurate models. Work is in progress to explore the idea more and improve the model.

Bibliography

1. Credit Card Fraud Detection using Machine Learning Framework by Tejas Darekar, Kaustubh Kapadne, Nilesh Ambesange, Neeraj Joshi, Dr. S. U. Kadam.
2. Credit Card Fraud Detection using Machine Learning and Data Science by Maniraj S P SRM Institute of Science and Technology.
3. Dataset from Kaggle.com

Plagiarism Report