

Credit Card Fraud Detection Using Machine Learning & Data Science

Ishika Sharma¹ Shivjyoti Dalai², Venktesh Tiwari³, Ishwari Singh⁴, Seema Kharb⁵

^{1,2,3} Students, Computer Science Engineering, SRM University, Sonipat

⁴Asst. Professor, Dept. of Computer Science Engineering, SRM University, Haryana,

⁵Asst. Professor, Dept. of Computer Science Engineering, SRM University, Haryana, India

Abstract - A method for 'Credit Card Fraud Detection' is created in this study. As the number of scammers grows every day. Credit cards are used for fraudulent transactions, and there are several sorts of fraud. As a result, various techniques such as Logistic Regression, Random Forest, and Naive Bayes are utilized to tackle this problem. This transaction is evaluated individually, and whatever works best is carried out. The primary purpose is to detect fraud by filtering the aforementioned strategies in order to achieve a better outcome.

Key Words: Credit Card, Fraud Detection, Random Forest, Naïve Bayes, Logistic Regression.

1. INTRODUCTION

Credit card fraud is a broad word for theft and fraud perpetrated using or utilizing a credit card at the moment of payment. The goal may be to buy something without paying for it or withdraw money from an account without permission. Identity theft is often accompanied by credit card fraud. According to the Federal Trade Commission of the United States, the rate of identity theft remained steady during the mid-2000s, but it jumped by 21% in 2008. Even though credit card fraud, the crime most people connect with ID theft, fell to a fraction of total ID theft complaints in 2000, roughly 10 million transactions, or one out of every 1300, were fraudulent. In addition, 0.05 percent (5 out of 10,000) of all monthly active accounts were fake. Today, fraud detection systems keep track of a twelfth of one percent of all transactions performed, resulting in billions of dollars in losses. Credit card fraud is one of the most serious issues facing businesses today. However, to successfully detect fraud, it is necessary first to comprehend the processes of fraud execution. Fraudsters use a variety of methods to perpetrate credit card fraud. Credit Card Fraud is described as "when an individual uses another person's credit card for personal reasons while the card owner and the card issuer are unaware that the card is being used." Theft of the actual card or the critical data linked with the account, such as the card account number or other information that must be given to a merchant during a valid transaction, is where card fraud begins. Card numbers, usually the Primary Account Number (PAN), are often reproduced on the card, and the data is stored in machine-readable format on a magnetic stripe on the reverse.

2. METHODOLOGY

This part should provide the method and analysis used in your research project. Using keywords from your title in the first few phrases is a simple and effective method to follow.

A. Data Collection

The data-gathering phase is the first step in the project; this dataset comprises a collection of transactions, some of which are real and others are fraudulent. The data-gathering phase is the first step in the project; this dataset includes a collection of transactions, some of which are real and others that are fraudulent. The data-gathering phase is the first step in the project; this dataset comprises a collection of transactions, some of which are real and others are fraudulent.

B. Credit Card Dataset

A credit card transaction data set was gathered via Kaggle, and it comprises a total of 2,84,808 credit card transactions from a European bank. It divides transactions into "positive class" and "negative class." The data set is highly skewed, with roughly 0.172 percent of transactions being fraudulent and the remainder being legitimate; this indicates that just 492 of the 2,84,808 transactions are fraudulent, and the rest are genuine ones. So, we oversampled to balance the data set, resulting in 60% of fraud transactions and 40% genuine ones.

C. Preprocessing of Dataset

Selected data is formatted, cleaned, and sampled in this module. The following are some of the data pre-processing steps:

a) Formatting: The chosen data might not be in the correct format. We may prefer data in a file format over a relational database or vice versa.

b) Cleaning is the process of removing or correcting missing data. The dataset may contain records that are incomplete or have null values. Such records must be deleted.

c) Sampling: The class distribution in credit card transactions is uneven because the number of frauds in the dataset is fewer than the total number of transactions. As a

result, the sampling approach is utilized to tackle this problem.

D. Loading of Dataset

The dataset is loaded after it has been pre-processed. Various library functions can be used to load the dataset. In this case, we used the read CSV method of Python's Pandas library to load a dataset in CSV or Microsoft Excel format; in terms of python, it is called a DATAFRAME. `dataset = pd.read_csv('creditcard.csv')`

E. Splitting of Dataset

To compensate for the dataset's imbalance, we used the ADASYN oversampling technique, which oversamples both the number of fraudulent and genuine transactions to a specific number, resulting in a positive and negative range that is nearly equal. After the dataset is oversampled, the samples are split into Train and Test data. A suitable ratio is to be performed for the model (Usually, 70% for Train data and 30% for Test data are chosen, anyone can choose their ratio). The train dataset can be further split into train data and validation data.

F. Building Model

After the data has been split into train and test data which is 70% and 30%, respectively, the training data is now utilized for the model building. The dataset contains 31 features, out of which 30 features or columns are the independent features, and the last column called the CLASS column, is the dependent feature. So here, the dataset is split into four categories: xtrain, ytrain, xtest, and ytest, representing independent training features, dependent training features, independent test features, and dependent test features.

G. Algorithms

a) Logistic Regression -Regression is a regression model that analyses the relationship between multiple independent variables and has a categorical dependent variable. There are many different logistic regression models, including binary, multiple, and binomial logistic models. The Binary Logistic Regression model calculates the likelihood of a binary response based on one or more predictors.

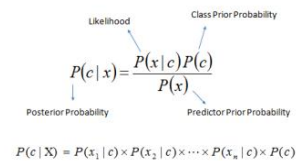
$$P = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Fig 1- Logistic Regression expression

The above equation represents the logistic regression in mathematical form.

b) Random Forest - Random Forest can be used to rank the importance of variables in a regression or classification problem in a natural way. Random forest is a tree-based

algorithm that creates several trees and combines the results to improve the model's generalization ability. An ensemble method is a technique for combining trees. Ensembling is nothing more than putting together a group of weak learners (individual trees) to create a strong learner. Random Forests can be used to solve problems involving regression and classification. The dependent variable in regression problems is continuous. The dependent variable in classification problems is categorical.



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig 2- Random Forest expression

c) Naïve Bayes - A Bayesian classifier is a statistical method that uses Bayes' theorem to calculate the probability that a feature belongs to a specific class. It is referred to as naive because it assumes that the possibilities of individual components are independent of one another, which is extremely unlikely to occur in the real world. The probability of an event occurring is calculated by considering the likelihood of another event occurring. It's possible to write it as:

$$Pr(c|X) = \frac{(Pr(X/c) \cdot Pr(c))}{(Pr(X))}$$

Fig 3- Naïve Bayes expression

Where the posterior probability of target class c $P(c|X)$ is calculated from $P(c)$, $P(X|c)$, and $P(X)$.

H. Training of Model

After building the model, the model is trained using Train data and validation data. The model is trained using the library function `fit()` function.

I. Evaluating the Model The model can be evaluated by using various metrics. These are

a) Interpreting Loss and Validation loss - Loss is the result of a bad prediction. A loss is a number indicating how bad the model's prediction was on a single example. Loss can be validation loss and training loss.

b) Interpreting Accuracy and Validation Accuracy - Validation accuracy and accuracy need to be converged in a good model.

c) Confusion Matrix - A confusion matrix summarizes classification problem prediction results. The correct and

incorrect predictions are totaled and broken down by class using count values. The confusion matrix's key is this. The confusion matrix depicts the various ways in which your classification model becomes perplexed when making predictions. It reveals the number of errors made by a classifier and the types of errors made.

	class 1 predicted	class 2 predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Here,

- Class 1: Positive
- Class 2: Negative Classification Rate/ Accuracy Classification Rate or Accuracy is given by the relation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig 4- Accuracy Expression

Recall - Recall can be defined as the ratio of the total number of correctly classified positive examples divided by the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN).

$$Recall = \frac{TP}{TP + FN}$$

Fig 5- Recall Expression

Precision - To get the precision value, we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High precision indicates that an example labeled as positive is positive (a small number of FP).

$$Precision = \frac{TP}{TP + FP}$$

Fig 6- Precision Expression

J. Saving the Model

After building the model, the model is saved to our device. The model can be saved in .pkl format or .h5 format. To save the model in .pkl format, python provides us a library named Pickle, and to save it in .h5 format, the Tensorflow library is

used. I have used the Pickle library and saved the models in .pkl format.

3. MODEL AND ANALYSIS

```

=== LogisticRegression ===
Model Accuracy: 91.2%

Confusion Matrix:
[[85582  8256]
 [    7  142]]
    
```

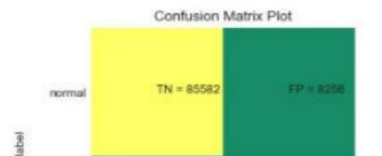


Fig 7- Logistic Regression

```

===== Model Test Results
=== RandomForest Classifier ===
Model Accuracy: 99.27%

Confusion Matrix:
[[93816  22]
 [   24  125]]
    
```

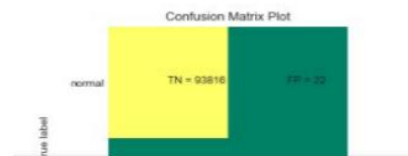


Fig 8- RandomForest Classifier

```

=== Naive Baiye Classifier ===
Model Accuracy: 89.4%

Confusion Matrix:
[[83840 9998]
 [   10  139]]
    
```

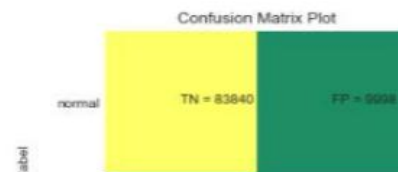


Fig 9: Naïve Bayes model

4. RESULTS AND DISCUSSION

The accuracy results we got from the three algorithms are shown in the following table below.

Methods	Frauds	Genuine	MCC
Logistic Regression	79.065	99.962	0.786
Random Forest	42.683	99.988	0.604
Naïve Bayes	83.130	93.730	0.219

Table- Accuracy Results

Training vs. Test Data in Dataset

```
In [106]: plt.figure(figsize=(10,8))
x=['Training data','Test data']
y=[len(x_train),len(y_test)]
print(x)
print(y)
sns.barplot(x,y,color='red')
plt.title('Training data vs Test data')
plt.xlabel('count')
plt.ylabel('count')
plt.xticks(['Non-Fraudulent', 'Fraudulent'])

Out[106]: Text(0, 0.5, 'count')
```

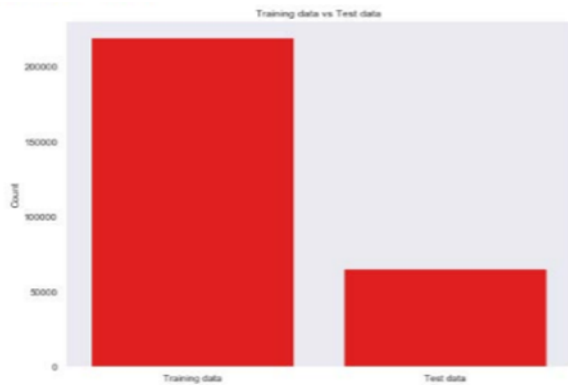


Fig 10- Training vs. Test Data

Normal vs. Fraud Transaction after Oversampling

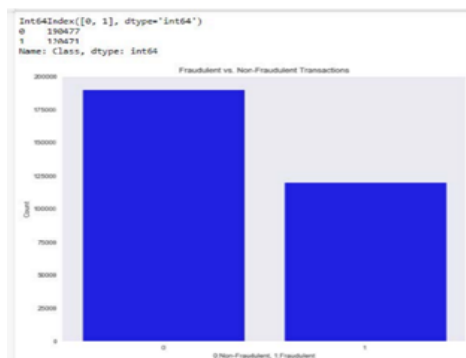


Fig 11- Fraudulent vs. non-Fraudulent

Correlation matrix

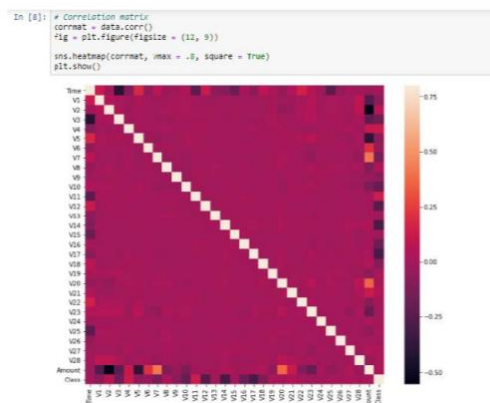


Fig12- Correlation Matrix

4. CONCLUSIONS

Various machine learning algorithms for detecting fraud in credit card transactions were reviewed in this paper. The accuracy, precision, and specificity metrics are used to evaluate the performance of this technique. To classify the transaction as fraudulent or authorized, I used three supervised learning techniques: Logistic Regression, Random Forest, and Naive Bayes. Using feedback and delayed supervised training, these classifiers were trained on a delayed supervised sample dataset of almost 284807 transaction records. Due to the massive imbalance, the dataset was subjected to an Oversampling technique, which resulted in the number of fraud and normal transactions being nearly equal. The training and test data were tested using the three Models, and the results were obtained. The accuracy of the Random Forest, Logistic Regression, and Naive Bayes was 99.27%, 91.20%, and 89.40%, respectively. From the Above project, it can be concluded that the Random Forest model is somewhat trustworthy, and its accuracy could be improved further with a larger and more balanced dataset. If some other algorithms can be combined with this one to form a Hybrid Algorithm, the results will be even better.

ACKNOWLEDGEMENT

The success and outcome of this project required a lot of guidance and assistance from many people, and we are highly privileged to have got this all along with the completion of our project. All that we have done is only due to such supervision and assistance, and we will not forget to thank them.

We are extremely grateful to **Dr. Paramjit S. Jaswal**, Vice-Chancellor, SRM University, and **Dr. Puneet Goswami**, Head of the Department, Department of Computer Science and Engineering, for providing all the required resources for the completion of my seminar.

Our heartfelt gratitude to our guide **Dr. Ishwari Singh**, for their valuable suggestion and guidance in preparing the research paper. Last but not least, we would express our obligation to all the people who have worked extensively on the topic and make the content available for free to all the aspiring people who want to grow in their community. I would say this report can be helpful to any aspiring student who wants to gain an overall idea about how high-performance computing works in practical life.

REFERENCES

[1] T. Mohana Priya, Dr. M. Punithavalli & Dr. R. Rajesh Kanna, Machine Learning Algorithm for Development of Enhanced Support Vector Machine Technique to Predict Stress, Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 20, Issue 2, No. 2020, pp 12-20

- [2] Ganesh Kumar and P.Vasanth Sena, "Novel Artificial Neural Networks and Logistic Approach for Detecting Credit Card Deceit," *International Journal of Computer Science and Network Security*, Vol. 15, issue 9, Sep. 2015, pp. 222-234
- [3] Gyusoo Kim and Seulgi Lee, "2014 Payment Research", *Bank of Korea*, Vol. 2015, No. 1, Jan. 2015.
- [4] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, Hao Fu, "Financial Fraud Detection Model: Based on Random Forest," *International Journal of Economics and Finance*, Vol. 7, Issue. 7, pp. 178-188, 2015.
- [5] Hitesh D. Bambhava, Prof. Jayeshkumar Pitroda, Prof. Jaydev J. Bhavsar (2013), "A Comparative Study on Bamboo Scaffolding And Metal Scaffolding in Construction Industry Using Statistical Methods," *International Journal of Engineering Trends and Technology (IJETT)* – Volume 4, Issue 6, June 2013, Pg.2330-2337.
- [6] P. Ganesh Prabhu, D. Ambika, "Study on Behaviour of Workers in Construction Industry to Improve Production Efficiency," *International Journal of Civil, Structural, Environmental and Infrastructure Engineering Research and Development (IJCSEIERD)*, Vol. 3, Issue 1, Mar 2013, 59-66
- [7] Manideep, A. P. S., and Seema Kharb. "A Comparative Analysis of Machine Learning Prediction Techniques for Crop Yield Prediction in India." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 13.2 (2022): 120-133.