




Document Information

| | |
|--------------------------|--|
| Analyzed document | 1637927935318_final minor report.docx (D120102691) |
| Submitted | 2021-11-27T06:25:00.0000000 |
| Submitted by | Praneet Saurabh |
| Submitter email | praneetsaurabh.set@modyuniversity.ac.in |
| Similarity | 11% |
| Analysis address | praneetsaurabh.set.modyun@analysis.orkund.com |

Sources included in the report

| | | | |
|----------|--|--|----------|
| W | URL: http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3669/P13389%20%252820%2525%2529.pdf?sequence=1&isAllowed=y Fetched: 2021-11-27T06:27:00.0000000 |  | 1 |
| W | URL: https://github.com/ameyanator/Credit-Card-Fraud-Detection Fetched: 2021-11-27T06:27:00.0000000 |  | 2 |
| W | URL: https://www.sciencedirect.com/science/article/pii/S1877050915007103 Fetched: 2021-11-27T06:27:00.0000000 |  | 3 |

Entire Document

A Minor Project Report on

"

CREDIT CARD FRAUD DETECTION"

69%

MATCHING BLOCK 1/6

W

<http://dspace.daffodilvarsity.edu.bd:8080/bits ...>

In partial fulfillment of requirements for the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering Submitted by

Drishti Narang (180041) Ishita Goyal (180048) Janhavi Bajaj(180049)

Under the Guidance of Dr. Praneet Saurabh

Department of Computer Science and Engineering SCHOOL OF ENGINEERING AND TECHNOLOGY Mody University and Science and Technology Lakshmangarh, Distt. Sikar-332311

December 2021

ACKNOWLEDGEMENT

I sincerely express my gratitude to my guide for her benevolent guidance in completing the report on Credit Card Fraud detection. Her kindness and help have been the source of encouragement for me. I am grateful to her for the guidance, inspiration and constructive suggestions that helped us in the preparation of this project.

Drishti Narang (180041) Ishita Goyal (180048) Janhavi Bajaj(180049)

CERTIFICATE

This is to certify that the minor project report entitled "-Credit Card Fraud Detection" submitted by Drishti Narang, Ishita Goyal, Janhavi Bajaj, as a partial fulfillment for the requirement of B. Tech. Computer Science(BDA) Semester examination of the School of Engineering and Technology, Mody University of Science and Technology, Lakshmangarh for the academic session 2021-2022 is an original project work carried out under the supervision and guidance of Dr. Praneet Saurabh has undergone the requisite duration as prescribed by the institution for the project work.

PROJECT GUIDE: HEAD OF DEPARTMENT Signature: Signature: Name: Name: Date: Date:

EXAMINER-I: EXAMINER-II Signature: Signature: Name: Name: Date: Date:

1. INTRODUCTION

With an ever-increase in the field of technology, the usage of e-commerce websites have increased manifold. Due to the increase in the number of people flocking on these sites the usage of credit cards have also increased. Gone are the days, when people were skeptical to use credit cards to carry out any online transactions. Pertaining to the current situation, 90% of the people these days prefer to use credit cards. The reason for choosing credit card over cash is because of its profitable returns. With an increase in the usage of credit cards, the frauds associated with cards are also increasing. Today, credit card frauds account upto 0.2% of the total transactions. With this number being low does not give us an assurance that the number will be this low in the future too. Each year we witness an increase in the number of frauds. Keeping in mind the above facts, there are two methods to prevent this- one being fraud detection and the other being fraud prevention. Fraud prevention as the name suggests are the number of steps which are carried out by the credit card companies to prevent any such frauds. These include double layer of security like OTP and security questions. These are the methods which are devised but these are not the full-proof to prevent any such fraud. The other method as the name suggests is Fraud Detection which is usually carried out post the occurrence of fraud. This is carried out by the credit card company once they are notified about the same. They carry out specific methods to undo the fraud. In order to do this, the companies have to regularly update their fraud detection softwares to keep pace with the new frauds. The main goal of our underlying project is to detect the transactions from a given dataset and to prevent the same. In the

underlying pages we have carried out a series of steps to detect and prevent the same. Machine Learning algorithms perfectly detect and suits for this scenario.

2. DATASET USED

79%

MATCHING BLOCK 2/6

W

<https://github.com/ameyanator/Credit-Card-Frau ...>

The dataset contains transactions made by credit cards in September 2013 by European cardholders. In this dataset we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class i.e. frauds account for 0.172% of all transactions. It only contains numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues,

our source of dataset could not

90%

MATCHING BLOCK 3/6

W

<https://github.com/ameyanator/Credit-Card-Frau ...>

provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes binary values, 1 in case of fraud and 0 otherwise.

SOFTWARE DETAILS

The software that we have used in our project is Collab Notebook. Collab Notebook is basically a IDE provided by Google for executing out various Machine Learning Problems. What attracted us towards it and made us to use the same is its interactive environment, its easy to comprehend features and the its feature of being the Cloud Platform. Though there are many IDE's available for carrying out machine learning problems like Jupyter Notebook, Pycharm etc, but Collab is the one which needs not to be downloaded as some external softwares on our machines. The work that we do in Collab gets saved as notebook with the extension of .ipynb and gets saved to Google Drive of the account that we are using. Another striking feature of the same that it enables the users to combine the code with HTML, images etc. It also enables to add comments to make the code user- friendly. The range of libraries that are available in Collab and can be easily installed makes it the best platform to perform and Data Science related Problem. It also enables the user to import any dataset be it numeric, image etc. which adds another feather to it's feature, Because of its Google's cloud server we can easily enable to use GPUs and TPUs .

CHALLENGES OF CREDIT CARD FRAUD DETECTION

While designing this model, we came across a number of challenges being: Unavailability of the dataset: Pertaining to the topic, credit card fraud is an extremely confidential data which makes it difficult to find any public datasets for training the model. Moreover, the dataset that is widely available on the Internet is the one which has been performed with PCA(Principal Component Analysis) which in turn changed the columns of the dataset. The main reason behind this is the privacy of the users and security concerns of the companies. The unavailability of the dataset also makes it difficult to train our model with utmost accuracy. Highly Skewed Data: This is another challenge that we faced while designing our model, the dataset available was highly skewed that is it majorly consisted of normal transactions and the fraud transactions forming a very small fraction of the dataset. This also does not detect the accuracy of the model in the real world perfectly. Dynamic Fraudulent Behaviour : As with an increase in the number of fraud transactions the type of frauds also increase each day. The model that we design today to detect the frauds might not be applicable to detect the frauds tomorrow because of the ever- changing nature of the frauds. Due to this the classification of the fraud and non- fraud transactions that we have done in this model might change the other hour. There is no fix pattern to this. Because of this, the problem becomes complex and also its detection.

REAL WORLD APPLICATIONS

Credit

card fraud is increasing day by day

57%

MATCHING BLOCK 4/6

W

[https://www.sciencedirect.com/science/article/ ...](https://www.sciencedirect.com/science/article/...)

with the development of modern technology and global communication systems. Credit card fraud costs consumers and the financial company millions and billions of bucks annually, and fraudsters constantly try to find new rules and tactics to commit illegal actions

and scams.

82%

MATCHING BLOCK 5/6

W

[https://www.sciencedirect.com/science/article/ ...](https://www.sciencedirect.com/science/article/...)

Thus, fraud detection systems have become a necessity for banks and financial institutions, to minimize their losses

and better customer service. Various machine learning techniques are used for fraud detection systems. Some common ones used

100%

MATCHING BLOCK 6/6

W

[https://www.sciencedirect.com/science/article/ ...](https://www.sciencedirect.com/science/article/...)

are Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbor algorithms (KNN). These techniques can be used alone or

can be combined with each other and used with the help of ensemble learning techniques. There are a lot of trials done using various approaches and the performance evaluation is done on the basis of their working on real life datasets.

Our Approach

4.1 DATA EXPLORATION Data Exploration as the name suggest is knowing our dataset better. Since, our Credit card Fraud Detection dataset is highly imbalanced with the fraud transaction forming an extremely low ratio. This issue can be imposed as a serious problem while fitting our data model. To remove this and to reduce our issues while fitting the data model, we perform the process of random under sampling which is used to remove the samples from major class. We perform under sampling in the ratio of 1:1.

Initial Attempts

4.2.1 Naïve Approach Using Unsupervised Learning

Before applying any algorithm we need to remove the outliers. To remove the outliers we first detect the features with positive and negative correlation. After this we remove the outliers that lie outside the range of 2.5 time of IQR to curb the effect of outliers in the data model. To find the correlation we use the `.corr()` function, to find the correlation between the 'Class' column and other columns. We pass the `under_sample` as data which is the result of concatenation of fraud and non- fraud indices.

After finding the correlation we detect the features with positive and negative correlation.

After we are done with the correlation part we apply isolation forest algorithm to fit our model and check for the accuracy.

4.2.2 APPLYING ISOLATION FOREST ALGORITHM

Initially, we started by using Isolation forest. Isolation forest isolates the observations by randomly choosing a feature and then it chooses a value which in turn is used to split the maximum and minimum values of the feature. It starts by isolating the outliers which we have also performed. This algorithm works on the principle of recursion.

After applying the Isolation forest algorithm, we fit the data into the model and tag the outliers and then calculate the various classification metrics.

On application of isolation forest algorithm which is usually considered as an anomaly detection algorithm, we find that the accuracy comes up to 99.76%. The isolation forest algorithm since the name suggests that it isolates the features and then decides whether a particular feature here belongs to fraud transaction or non- fraud transaction. Though keeping in

mind that it is one of the most commonly used anomaly detection algorithm, we also implemented random forest classification algorithm as to check our model on some other algorithm also.

Train-Test Split Evaluation

The train-test split procedure is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for predictive modeling problems. It is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. This procedure is not appropriate when the dataset available is small but in our case it seemed quite appropriate to us since our dataset was sufficiently large to apply it. The procedure involved taking a dataset and dividing it into two subsets. • Train Dataset: It is used to fit the machine learning model. • Test Dataset: It is used to evaluate the fit machine learning model.

Our objective was to estimate the performance of the machine learning model on new data: data not used to train the model. So, basically we divided our data into two parts with the ratio of 80:20 so as to achieve good accuracy and also avoid over fitting at the same time. Next we applied the machine learning algorithm Isolation Forest on the train and test dataset. By using this algorithm we got the accuracy of 99.76%. Since, the accuracy could have been better we then tried Random Forest Algorithm on the train and test dataset. By using this algorithm we got the accuracy of 99.95%. The precision score was 94.87% and recall 75.51%. The F1 score was 84.09%. The correlation coefficient was 84.61%. The evaluation parameters were quite better than the ones we got by applying Isolation Forest Algorithm so we took Random Forest Algorithm as our final approach.

4.3 Final Approach- APPLYING RANDOM FOREST ALGORITHM

Random Forest as the name suggest randomly selects a feature and then demarcate a feature into fraud or non- fraud transaction. Random Forest is an example of supervised learning algorithm. This algorithm randomly selects m features from T . For node 'D' it calculates best split point among ' m ' features. Post this, it splits the node into two daughter nodes using the best split. Repeat the above steps until ' n ' number of nodes. We build our forest by repeating these steps. Here T is number of features. D is number of trees to be constructed.

To start to apply random forest algorithm we firstly import the Random Forest Classifier from sklearn.ensemble package. After this we fit our Train data set into the data model and then predict the output for test data set.

Post applying the algorithm we calculate the evaluation parameters to check the correctness of our model. Namely we calculate accuracy, precision, recall and F1 score.

As it can be easily assessed that the accuracy of isolation forest comes up to 99.76% and the accuracy of random forest comes up to 99.95%. Judging on the basis of accuracy we see select random forest algorithm as our final model to make our project. Here, while calculating the evaluation parameters True Positive are the transaction that are non- fraud and are hence predicted as non- fraud. False Positives are the transactions that are fraud but predicted non- fraud transaction. After this, we calculate confusion matrix to check how our model is performing. In laymen language confusion matrix can be described as the report card of our model. We use the function confusion_matrix to calculate the print the same. Since it is a matrix with rows having actual fraud and non- fraud transactions and the columns having the predicted fraud and non- fraud transactions.

The Confusion Matrix came as follows:

As it is clearly visible from the above matrix that the actual fraud transaction are $24+74=98$ and the actual non- fraud transaction are $56860+4=56864$.

Evaluation Parameters

In our project we used accuracy, precision, recall, F1 score and correlation coefficient as evaluation parameters. We also made a confusion matrix for the model for a better visualization of the true positive, true negative, false positive and false negative values.

Precision is the fraction of relevant instances among the retrieved instances, higher precision indicates more relevance of the algorithm used. $\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}$

Recall is the fraction of relevant instances that were retrieved. $\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})}$ F1 score is used to find the balance between the precision and the recall. $\text{F1 score} = 2 * ((\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}))$

$\text{Recall}/(\text{Precision} + \text{Recall})$

Matthews Correlation coefficient is used in machine learning as a measure of the quality of binary classifications.

CONCLUSION

Detection of Fraud and Non- Fraud transactions was a tough one to do. There is no mistake in detecting a non- fraud as a fraud transaction but detecting a fraud as a non- fraud can be serious threat. To make this project a success, we had to prevent the misclassification of the transactions. One wrong detection of the type of transaction could disrupt the entire model. To achieve this we used a couple of ML algorithms like Isolation Forest and Random Forest. Post this project we learnt the nuances of fraud detection and how can we blend the classroom study into real world models in perfect composition. This gave an alarming sign of frauds happening every year and what exactly is the background study behind such life- saving softwares. Credit Card Frauds can land a person in financial crunch. These algorithms help to detect at the correct time and also helps to prevent. The accuracy and the various evaluation metrics tell us that which algorithm is better and can be used to detect and prevent and give better results. Through this model, we can easily now comprehend the problem statement and give a perfect solution to the same.

Hit and source - focused comparison, Side by Side

| | |
|-----------------------|--|
| Submitted text | As student entered the text in the submitted document. |
| Matching text | As the text appears in the source. |

| 1/6 | SUBMITTED TEXT | 117 WORDS | 69% MATCHING TEXT | 117 WORDS |
|-----|--|-----------|--|-----------|
| | In partial fulfillment of requirements for the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering Submitted by | | in Partial Fulfillment of the Requirements for the Degree of Bachelor of in Computer Science and Engineering Supervised By | |
| | W http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3669/P13389%20%252820%2525%2 ... | | | |

| 2/6 | SUBMITTED TEXT | 60 WORDS | 79% MATCHING TEXT | 60 WORDS |
|-----|---|----------|---|----------|
| | The dataset contains transactions made by credit cards in September 2013 by European cardholders. In this dataset we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class i.e. frauds account for 0.172% of all transactions. It only contains numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, | | The contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, | |
| | W https://github.com/ameyanator/Credit-Card-Fraud-Detection | | | |

| 3/6 | SUBMITTED TEXT | 82 WORDS | 90% MATCHING TEXT | 82 WORDS |
|-----|--|----------|--|----------|
| | provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes binary values, 1 in case of fraud and 0 otherwise. | | provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' the transaction Amount, this feature can be used for example-dependant cost-sensive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. | |
| | W https://github.com/ameyanator/Credit-Card-Fraud-Detection | | | |

| 4/6 | SUBMITTED TEXT | 40 WORDS | 57% MATCHING TEXT | 40 WORDS |
|-----|--|----------|--|----------|
| | with the development of modern technology and global communication systems. Credit card fraud costs consumers and the financial company millions and billions of bucks annually, and fraudsters constantly try to find new rules and tactics to commit illegal actions | | with the development of modern technology and the global superhighways communication. Credit card fraud costs consumers and the financial company billions of dollars annually, and fraudsters continuously try to find new rules and tactics to commit illegal actions. | |
| | W https://www.sciencedirect.com/science/article/pii/S1877050915007103 | | | |

| 5/6 | SUBMITTED TEXT | 18 WORDS | 82% MATCHING TEXT | 18 WORDS |
|-----|--|----------|--|----------|
| | Thus, fraud detection systems have become a necessity for banks and financial institutions, to minimize their losses | | Thus, fraud detection systems have become essential for banks and financial institution, to minimize their losses. | |
| | W https://www.sciencedirect.com/science/article/pii/S1877050915007103 | | | |

| 6/6 | SUBMITTED TEXT | 21 WORDS | 100% MATCHING TEXT | 21 WORDS |
|-----|--|----------|---|----------|
| | are Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbor algorithms (KNN). These techniques can be used alone or | | are Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbor algorithms (KNN). These techniques can be used alone or | |
| | W https://www.sciencedirect.com/science/article/pii/S1877050915007103 | | | |