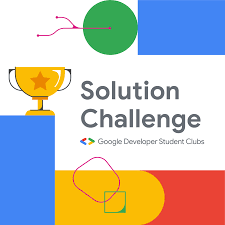


**PRESENTS**

****

**Project Name: Accounts Fraud Detection**

**Team Name: Akatsuki**

**Team Members:**

Abhinav Kumar (The NorthCap University | B.Tech ECE)

Piyush (The NorthCap University | B.Tech ECE)

Harsh Kumar Sharma (The NorthCap University | B.Tech ECE)

Sushant Panwar (The NorthCap University | B.Tech ECE)

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S NO** | **Title** | **Page No.** |
| 1. | Introduction | 1 |
| 2. | Inspiration of the project | 2 |
| 3. | What it does!!! | 3 |
| 4. | Technologies Used | 7 |
| 5. | Challenges Faced | 8 |
| 6. | What we learnt??? | 9 |
| 7. | Future Scope of Project | 11 |
| 8. | References | 12 |
|  |  |  |

**INTRODUCTION**

In this new era, the world has started being digital. It's really great success for all of us to be part of this wonderful experience. At the present time, most of us are using UPI, credit cards, debit cards online to do transactions.

That's great!

In colleges/Universities/campuses transactions are mainly done online and maintaining the data of many students is really a tough job. So, as there will be a large number of transactions, the problem which arises from this will be fraud cases of payment leading to face many big losses.

So there has to be way such that colleges and universities are able to detect/recognise fraudulent or unwanted transactions so that student or college-no one has to suffer.

Detecting the fraud transactions is of great importance for colleges & universities.

The main motive of this project falls under Campus Digital Solutions and is to classify whether the transactions done in campus are fraud or not.

Due to the theatrical increase of fraud which results in loss of currency worldwide each year, several modern techniques in detecting fraud are persistently evolved and applied to many businesses fields. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid undesirable behaviour. Undesirable behaviour is a broad term including delinquency, fraud, intrusion, and account defaulting.

So, in our project we would be analysing/building a fraudulent detection model using Machine Learning and Deep Learning and would inference and compare the results obtained by both to achieve maximum accuracy/meaningful results.

**1.**

**Inspiration for our Project**

While reading news over internet I get to know that a women paid for two students from Egypt to attend university of Wisconsin-Green Bay.

Staff have become suspicious after one of the college students wanted money back due to the fact his account turned into overpaid. The lady's credit card was used on-line in April to pay extra than $3,300 on the students' debts, and university personnel later decided the lady in no way gave all of us permission to apply her card.

After a monthslong probe, the two college students — ages 18 and 20 — had been cited in November for theft and a laptop facts offense. University police determined to trouble citations in preference to arrest them; however, such instances regularly can be prosecuted as felonies.

If the scholars are convicted of both noncriminal charges, they each could ought to pay nearly $750 in fines.

Colleges and universities are making efforts to prevent crimes in which someone's Social Security or credit card numbers are used to commit fraud. Fraud attempts targeting the college have occurred when criminals get the institution's bank routing and checking account numbers, which then are printed on fake checks and sent to unsuspecting parties as a payment.

That turned into the time we found out that college additionally faces trouble in these varieties of cases. Then we decided that if we were given any hazard to enhance our campus in a manner, we would work for bills branch wherein all of the transactions are completed so they may without difficulty stumble on fraud cases and save you any useless loss. Keeping the information in mind we made our project for detecting fraudulent cases among the many transactions which would help not only campus but it’s employees and in some cases the person making fraud transactions (as it might be done through his/her details without them realizing).

**2.**

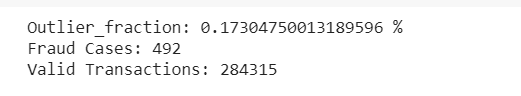
**WHAT IT DOES!!!**

Credit/Debit Card fraudulent detection, as the name suggests our project simply detects fraud cases amongst number of transactions in campus. As we proceed, we have a total of **284,807 transactions out of which 492 frauds** cases are there.

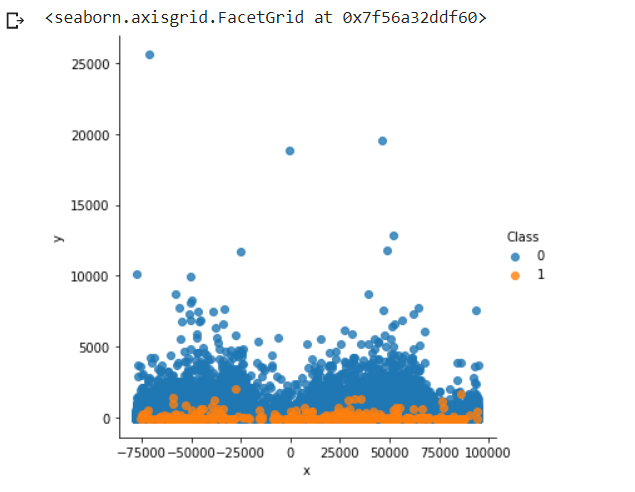
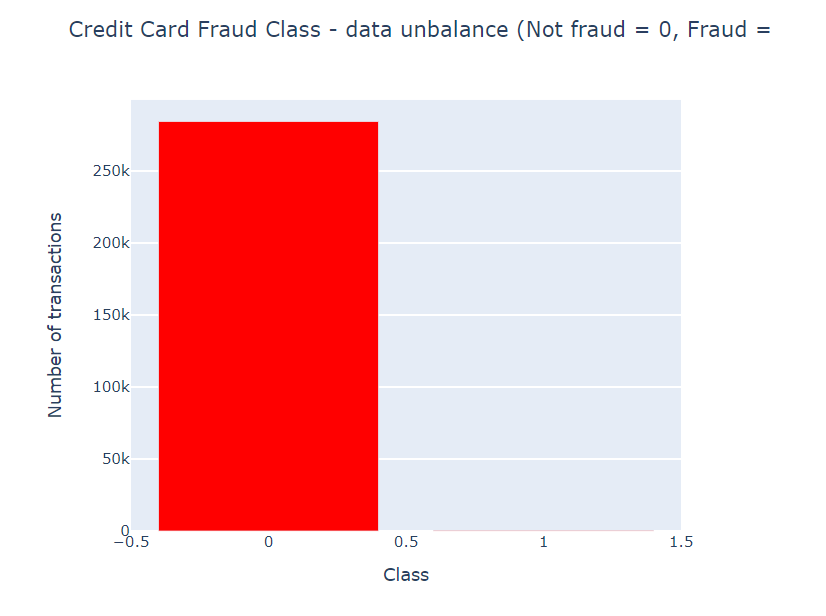
Firstly, we understood what our data is its different columns/features.

Then did the Data Pre-processing stuff that consists of:

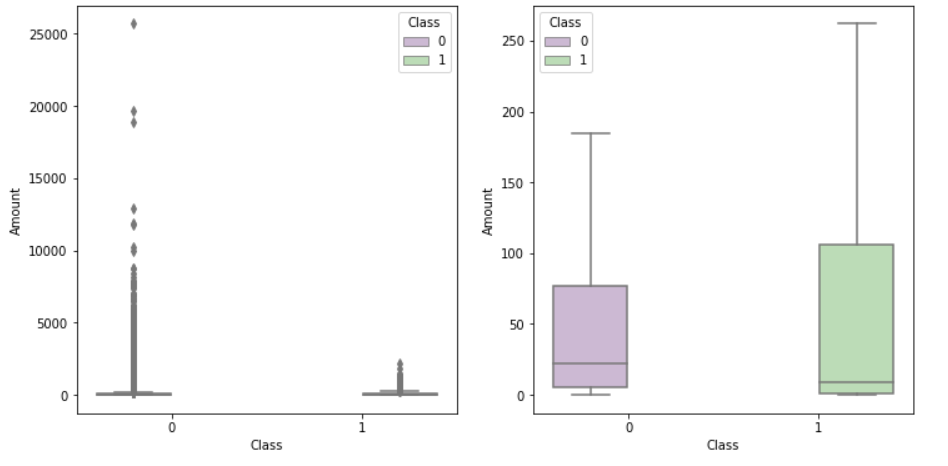
* Describing the dataset.
* Checking for null values in the dataset and removing if any
* And other basic pre-processing stuff.



Also, by plotting different histograms plot we observed that the transactions occur in a cyclic way. But the time feature does not provide any useful information as the time when the first transaction was initiated is not given. Thus, we'll drop this feature.



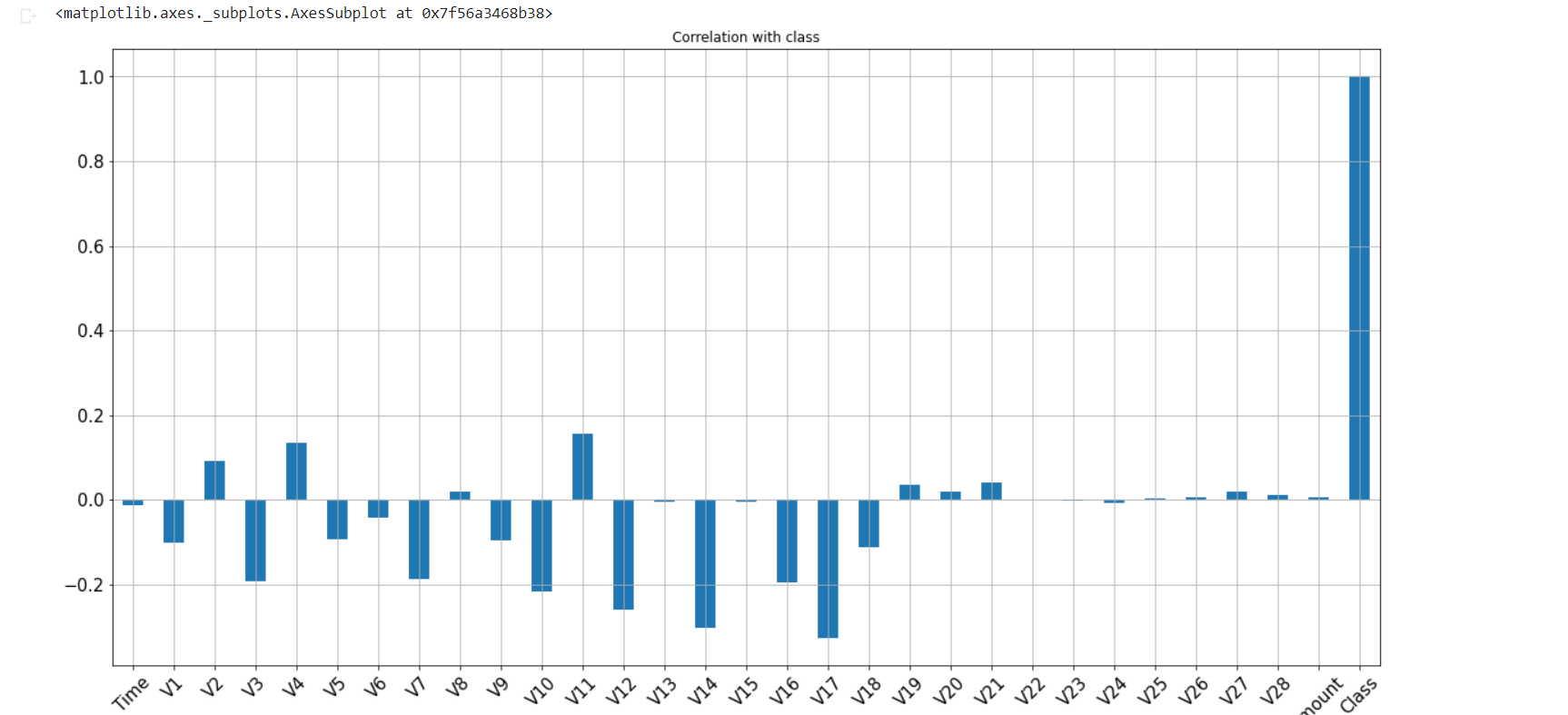
Then we plotted box plot and we observed that the Amount needs to be scaled. As if a feature has a variance order of magnitude larger than the rest of the features, then it might dominate and, make the estimator unable to learn from other features as expected.



**3.**

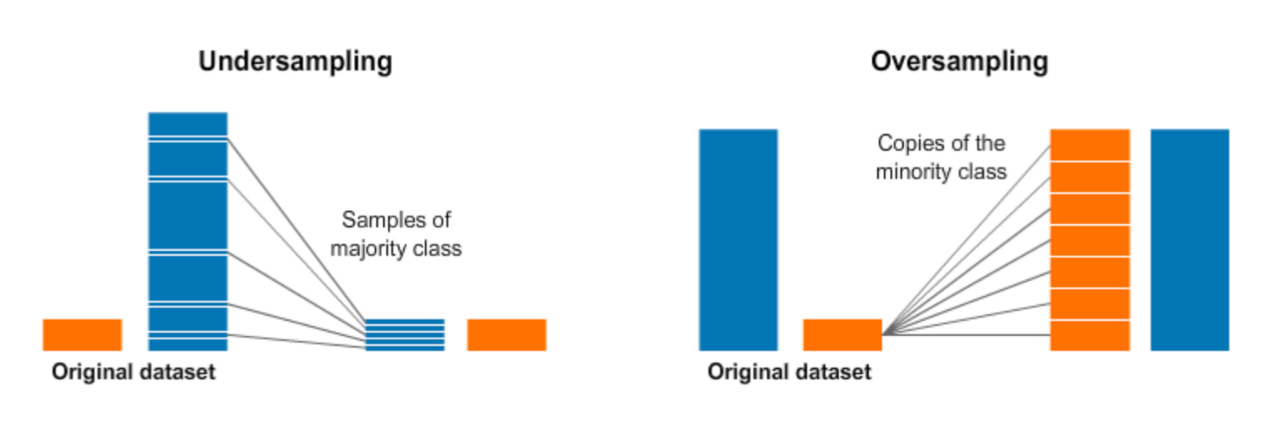
Further we had plotted the Features correlation to see dependence of the Class column on any feature.

As expected, there is no notable correlation between features V1-V28. There are certain correlations between some of these features and Time (inverse correlation with V3) and Amount (direct correlation with V7 and V20, inverse correlation with V1 and V5).



Further we had plotted the feature distribution plots.

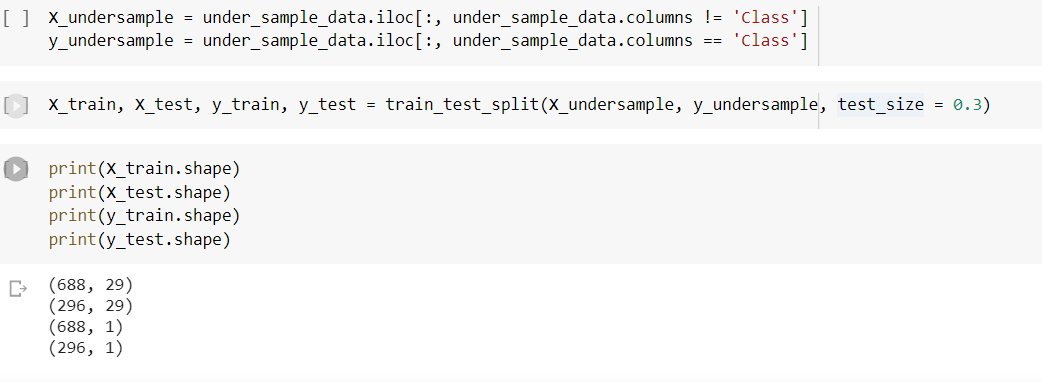
An issue about the dataset we have here is a class imbalance. Only 0.17% of 284,807 transactions are frauds. Sadly, the model is more sensitive to detect the majority class than the minority class. So, to tackle the problem of this imbalanced dataset, we have used In general, two techniques to tackle class imbalance, under-sampling, and over-sampling.



In **Undersampling** we randomly selected the same number of non-frauds as the fraud and created a new dataset. And then we trained our model on this dataset.

Undersampling is the technique of removing major class records from the sample. In this case, it is necessary to remove random records from the legitimate class (No fraud), in order to obtain a number of records close to the amount of the minority class (fraud) in order to train the model.

**4.**

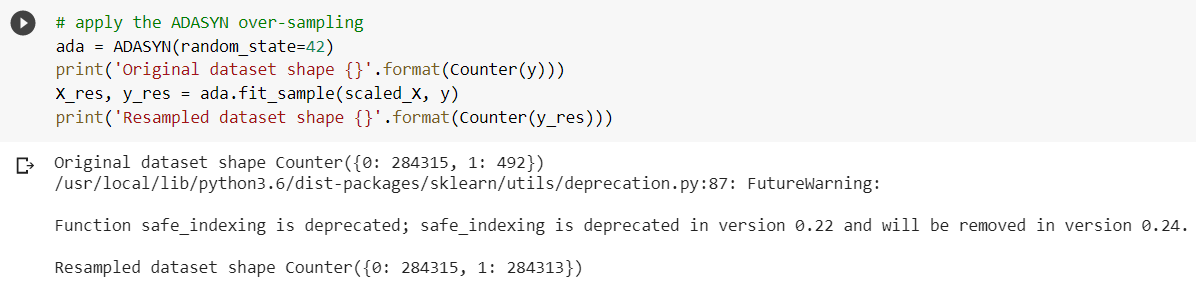


Here we had created a dataset with 984 transactions containing equal number of frauds and non-frauds.

The other method of handling this imbalanced data was **oversampling**.

As noted earlier, the dataset is unbalanced. Training unbalanced dataset with learning algorithms may lead to misclassification of minority class. Therefore, to compensate for the unbalancedness, I used ADASYN oversampling method as implemented in imbalanced-learn package to resample the dataset. ADASYN (ADAptive SYNthetic) is an oversampling technique that adaptively generates minority data samples according to their distributions using K nearest neighbour.

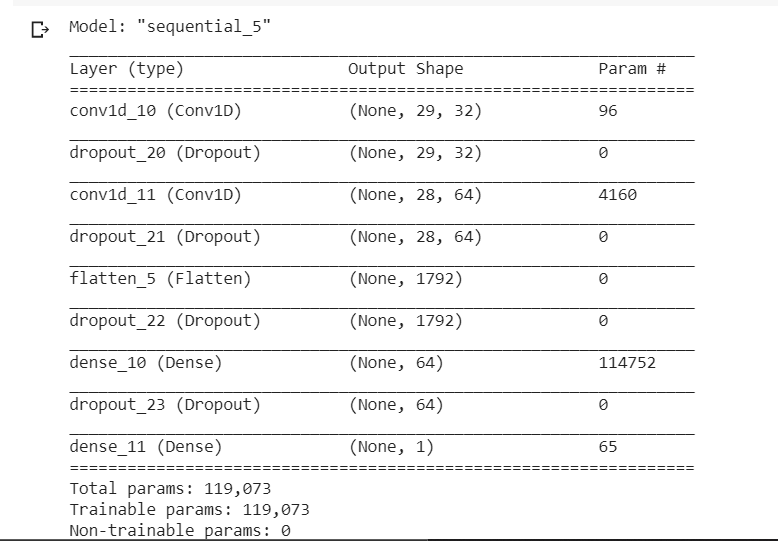
Oversampling is exactly the opposite: it means adding minority class records (fraud) to our training sample, thus increasing the overall proportion of fraud records.



So further we used Linear Regression, Logistic Regression, Random Forest, Decision Trees and Support Vector Machines, Machine Learning algorithm and inferred the results.

Also further built a simple CNN model and inferred its result.

**5.**



Further we applied Machine Learning algorithms and CNN model on the prepared data and inferred the results.

**6.**

**Technologies Used**

Our Project mainly focuses on machine learning and deep learning. We get to know a lot about how things work under them and how vast and interesting they are. They taught us how to build models and work on them accordingly, which would help us in many ways making things easy for us.

* **Machine Learning:** Machine learning is a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed. It is extensively defined as the functionality of a gadget to mimic sensible human conduct. Artificial intelligence structures are used to perform complex obligations in a manner that is much like how people resolve troubles. Machine getting to know starts with records — numbers, photographs, or textual content, like financial institution transactions, or sales reports. The information is accrued and organized for use as schooling facts, or the facts the gadget learning model will be trained on. The greater statistics, the better this system.
* **Deep Learning:** Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It’s achieving results that were not possible before.

**7.**

**Challenges Faced**

Credit/Debit Card Fraud Detection, our assignment deals with unwanted or fraud cases in our campus which allows in decreasing manpower and increase in efficiency, however working on it was never easy and performed in a single go. We confronted demanding situations which I’ll talk beneath, created issues at the start however helped us in studying more approximately getting to know and how to address it.

1. **Data Loading:** Our dataset became taken online from KAGGLE. Its overall length is round to 144MB so loading it on Python (Jupyter Notebook) was kind of hard and took time. So slow net customers would possibly face hassle whilst strolling this system because of its length.
2. **Runtime Error:** As all of us know Jupyter Notebook works simplest till a factor, maintaining it idle or running for lengthy would cease the session of the undertaking/work you’re doing. Our challenge included information in PCA layout, for this reason we had to expose its functions as a consequence so each person who tries to adapt to our code should easily apprehend. Also, we worked on go validation then we did oversampling and undersampling which made our task long enough for its session to give up.
3. **Connecting/Arranging our data:** The principal challenge looked while making our undertaking was the means by which to associate our information and what all to interface. As we referenced before our information was in PCA (hidden/summarized details) of Credit/Debit card. What all highlights would be suitable to show so that elements, utilizes and some way or another what they are would be obvious to the client. Then, at that point, after cross validation what all arrangement and relapse models to utilize so it makes our undertaking simple, clear and accommodating for its utilization. What's more subsequent to associating our information organizing it such that the client is effectively ready to get what's happening.

**8.**

**What we learnt???**

So, through this project we learned how to handle the data. Further we learned about Random Forest, Decision Tree and SVM Machine Learning algorithms and also convolutional neural networks. Also, we learned about different Undersampling and Oversampling Algorithms. This project helped me increase my knowledge in different fields like Machine Learning, Deep Learning.

From all the methods above used for this classification problem we observe that in case of:

1. **Unbalanced Dataset:**

Our Random Forest Model was giving best result with f-1 score as 90% and the model correctly predicted about 81% of fraud transactions.

Also, our unsupervised approach i.e., using islolation forest also gave good result and was able to identify 439 anomaly points out of 492 in the original dataset. Also, we have kept the threshold value a bit higher to so we will catch a lot of False positive as well.

As we know Accuracy in a highly unbalanced data set does not represent a correct value for the efficiency of a model. Initially, a method should be applied to balance the data before taking into account any performance evaluation metrics.

Also, the results from this unbalanced data are not meaningful as most of the classes were not fraud and our algorithm could not much understand about the behaviour of the Fraud transaction but still the model was able to perform really well i.e., unsupervised approach it performed well and it was able to identify 444 anomaly points out of 492 fraud transactions.

1. **Undersampled dataset:**

In case of our under-sampled data almost all the models performed really well, out of these Random Forest got f1-score as about 94% and was able to correctly predict fraud transactions as we can see from confusion matrix.

Also, our simple Neural Network model also gave good results with accuracy as 0.943 and from confusion matrix we can see that out of 159 fraud transactions in test data it correctly predicted about 147 transactions as fraud.

1. **Oversampled dataset:**

In case of over-sampled data, the size of dataset was very large, so we obtained good results on most of the models. But out of these our Random Forest model gave best results as we obtained accuracy 0.999877, precision of 0.999755, recall of 1.000000, and f1-score of 0.999877. So here we got good extrapolated results. Also, same can be seen if we observe the confusion matrix where we see that 0 transactions classified as normal but which were fraud, 93836 transactions classified as fraud and were actually fraud. So, from this we can say that it has correctly predicted all the fraudulent transactions in the test data set.

**9.**

One reason of getting these extrapolated results would be that our model performed really well and correctly learned the relationship between the features other could be that our dataset in oversampling was very large and it learned on this large dataset which resulted in extrapolation of these results.

**10.**

**FUTURE SCOPE OF THE PROJECT**

Our project deals with fraud transactions and there is no place where transactions are not done. Our project can handle number of organizations such as schools, banks etc having credit card transactions.

Banks can typically implement a single fraud detection and prevention system that tries to capture fraudulent transactions based on a model generalized to all their customers. This network model incorporates general fraud trends from different products across the bank.

In future, this technique will be standardized across all card associations and banks. Nonetheless, this approach is difficult because of customer’s privacy concerns for customer data.

Customers want their transactions approved in seconds, not minutes. To address this issue, better machine learning algorithms are needed to raise flags about fraudulent transaction in real-time.

**11.**

**REFERENCES**

[1] K. Fu, D. Cheng, Y. Tu, and L. Zhang, “Credit Card Fraud Detection Using Convolutional Neural Networks,” in Neural Information Processing, vol. 9949 of Lecture Notes in Computer Science, pp. 483–490, Springer International Publishing, 2018.

<https://medium.com/@palashshinde6/credit-card-fraud-detection-using-deep-learning-models-a8a918e29d4e>

[b] <https://www.pymnts.com/news/artificial-intelligence/2020/how-mastercard-uses-ai-to-fight-fraud-and-make-better-credit-decisions/#:~:text=By%20PYMNTS,Mastercard's%20Virtual%20Cyber%20%26%20Risk%20Summit>

[2]

TY - JOUR

AU - S P, Maniraj

AU - Saini, Aditya

AU - Ahmed, Shadab

AU - Sarkar, Swarna

PY - 2019/09/13

SP -

T1 - Credit Card Fraud Detection using Machine Learning and Data Science

VL - 08

DO - 10.17577/IJERTV8IS090031

JO - International Journal of Engineering Research and

**URL:** <https://www.researchgate.net/publication/336800562_Credit_Card_Fraud_Detection_using_Machine_Learning_and_Data_Science>

[3] <https://whatis.techtarget.com/definition/over-sampling-and-under-sampling>

[4] <https://towardsdatascience.com/adasyn-adaptive-synthetic-sampling-method-for-imbalanced-data-602a3673ba16>

[5] <https://www.datacamp.com/community/tutorials/convolutional-neural-networks-python>

[6] <https://www.geeksforgeeks.org/ml-linear-regression/>

[7] <https://towardsdatascience.com/machine-learning-simple-linear-regression-with-python-f04ecfdadc13>

[8] <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

[9] <https://machinelearningmastery.com/random-forest-ensemble-in-python/>

“Page 199, [Applied Predictive Modeling](https://amzn.to/2O3BuOa), 2013.”

[10] <https://scikit-learn.org/stable/modules/tree.html>

[11] <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>

**12.**