# 

E-commerce fraud detection using Machine Learning

Sahariar Ahmed Tonmoy and Abdul Wajed Khan

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)



Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Fall2020

E-commerce fraud detection using Machine Learning

Sahariar Ahmed Tonmoy and Abdul Wajed Khan

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Utpal Kanti Das

Chairman

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Hasibur Rashid Chayon

Co-supervisor, Coordinator and Associate Professor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Toyeer-E-Ferdous

Supervisor and Senior Lecturer

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Fall 2020

## **Letter of Transmittal**

30 September 2020

The Chair

Thesis Defense Committee

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

4 Embankment Drive Road, Sector 10, Uttara Model Town

Dhaka 1230, Bangladesh

**Subject:** Letter of Transmittal.

Dear Sir,

With due respect, we would like to approach you this is a great opportunity as well as immense pleasure for us to submit the thesis report entitled as “E-commerce fraud detection using Machine Learning”. This report is prepared as the partial fulfillment of the requirement of Bachelor of Computer Science and Engineering (BCSE) program, Department of Computer Science and Engineering, IUBAT – International University of Business Agriculture and Technology.

It was a great pleasure to do the thesis. Now, I am looking forward for your kind appraisal regarding this thesis report and I will be grateful if you kindly go through this report and evaluate our performance.

Yours sincerely,

\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_

Sahariar Ahmed Tonmoy Abdul Wajed Khan

18103263 18103038

## **Student’s Declaration**

We are Sahariar Ahmed Tonmoy and Abdul Wajed Khan are the student in the Department of Computer Science and Engineering, IUBAT – International University of Business Agriculture and Technology. The thesis report entitled as “E-commerce fraud detection” has been prepared for the partial fulfillment of thesis of Bachelor of Computer Science and Engineering (BCSE) program. We are declaring that this thesis report has prepared by us and it is not submitted anywhere else.

\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_

Sahariar Ahmed Tonmoy Abdul Wajed Khan

18103263 18103038

## **Supervisor’s Certification**

The supervisor’s certification of your thesis report is going to be added here. Keep the format as it is.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Toyeer-E-Ferdous

Supervisor and Senior Lecturer

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

## **Abstract**

Nowadays we use credit cards in our everyday lives, especially in online payment in e-commerce websites as almost all the e-commerce sites support online transactions. As the number of transactions is increasing with it the number of risks for fraud activities is increasing too. For that, researchers started using different machine learning algorithms to detect and analyse frauds in online transactions. In e-commerce online transaction data, the main problem is an imbalanced dataset as the majority of the data is not fraud this allows the models trends to be biased towards the majority samples for this most of the time the non-fraudulent activity are predicted as fraudulent activity. Our aim is to overcome this problem by using under sampling, oversampling methods using SMOTE technique and libraries like Sklearn to have the highest accuracy. In our research, we have used different feature engineering like the average amount for fraudulent and non-fraudulent cases and then compare each column with the amount to find out how much difference is in the amount. We used Recall Score, Precision Score, F1 Score, and Accuracy to evaluate the performance of the models and Logistic Regression, Decision Tree, Random Forest, SMV, Naive Bayes, Neural Network are models and Sklearn is the library.

*Keywords*—*fraud, credit card, e-commerce, feature engineering, machine learning*

## **Acknowledgments**

In the name of Almighty who is the most merciful and the most graceful.

I would like to give thanks to late Professor Dr. Md. Alimullah Miyan, Vice-Chancellor of IUBAT –International University of Business Agriculture and Technology to give permission to study in this university which is the most beautiful and renowned non-government university in this country. I would like to thank and convey the respect to our present honorable vice chancellor Professor Dr. Abdur Rab, IUBAT – International University of Business Agriculture and Technology. I would like to give gratitude to Dr. Utpal Kanti Das, Chairmen and Professor, Department of Computer Science and Engineering, IUBAT – International University of Business Agriculture and Technology to give the permission to study in the Department of Computer Science and Engineering and allow me to see the bright future in technological field of new era. I am very appreciative to Dr. Hasibur Rashid Chayon, Respected Coordinator and Associate Professor, Department of Computer Science and Engineering, IUBAT for his better direction and sustain throughout the semester.

I am extremely indebted to the Managing Director of Bit Net Solution for cooperating with me. I am really pleased and proud to express my feeling of gratefulness and profound respect to our respected faculties, especially Dr. Abhijit Saha, Professor, Department of Computer Science and Engineering, IUBAT for his scholastic guidance, helpful and unwiring efforts to execute this report.

Finally, I wish to thank my parents and my teachers who have been a great source of inspiration to me. Without their support, I would not have reached where I am today.

**Table of Contents**

**Letter of Transmittal** iii

**Student’s Declaration** iv

**Supervisor’s Certification** v

**Abstract** vi

**Acknowledgments** vii

**List of Figures** x

**List of Tables** xi

**Chapter I. Introduction** 1

1.1 Headings 4

**Chapter II. Literature Review** 5

2.1 Heading 3, Section Title 7

2.2 B-Head Style 7

2.3 C-Head Style and Accompanying Paragraph 8

2.3.1 Body Text with C-Head 8

2.4 Numbered List Style 8

2.5 Bulleted List Style 9

**Chapter III. Research Methodology** 10

3.1 Figures 12

3.1.1 Figure Style 13

3.1.2 Figure Description Style. 13

3.2 Tables 13

3.2.1 Table Title Style. 13

3.2.2 Table Cell Style. 14

3.2.3 Table Description Style. 14

**Chapter IV. Result and Discussion** 15

4.1 Title Style 18

4.2 Author Style 18

4.3 Centered Text Style 18

**Chapter V. Conclusion** 19

**References** 21

## **List of Figures**

Figure 3.1 Figure Title. 13

Figure 3.2 Figure Title2. 14

## **List of Tables**

Table 1.1 Evaluation of energy consumption of DSDV and DSR in MANET 2

Table 2.1 An example to understand how to write literature review 6

Table 3.1 Table Title. 15

## **Chapter I. Introduction**

The fraud in commerce is getting higher day by day. The attack of Covid-19 makes us stay more at home. The situation makes it harder for the people to go out and buy daily necessities, but in this time of technological advancement, the online marketplace becomes the alternative place of getting daily necessities. But there are always some criminals who always want to take advantage of such a situation by doing harm to the people. So, it is very important to make each customer’s experience as satisfying as possible, especially when it comes to the security of their accounts and money spent online. When thinking about how to decrease fraudulent, the first thing a banker, merchant, or other E-commerce participants should develop a new technique or process to tackle this. This technique includes find previous patterns and behavior in the data which cause fraudulent activity. There also so many types of frauds like mobile fraud, identity theft, friendly fraud, interception fraud, merchant app fraud and so on. This fraud detection and prevention plays an important role in the eCommerce site. Various types of frauds are now in this field to seize goods or money. A very huge number of transactions in e-commerce raises new kind of problems regularly. The number of e-commerce fraud is increasing each year since 1993. From a report we see in every 100$, 5.65 cents lost due to frauds. Detecting this activity by human observation is impossible, so we need a technique that can detect this activity in real-time. This technique shall be developed using machine learning because machine learning can learn by a pattern of data without the help of humans.

So, designing an efficient model or an algorithm is the key to reduce this problem. So, to solve this problem machine learning systems could help us to solve the tasks. As there are already so many solutions of previous problems that occur, we will try to get better efficiency and better accuracy than before by using algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine. Here, among these, we will get various kinds of accuracy. So, which's accuracy will be the closest to 100% that will be the perfect model for our work, but in our research data it is very difficult to arrange because data of eCommerce contain information about the customer and also contain information about credit cards so the data is very sensitive and impossible to work with real data.

The dataset we used there is 284807 rows and 31 columns. There is like 492+ frauds so the percentage of fraud is 0.17% and non-fraudulent activities are like 99.83%. The dataset used result are more in the classification of the majority of non-fraud than fraud. So, we will handle the imbalance data using the SMOTE (Synthetic Minority Oversampling Technique).

Our aim is to find out the best accuracy by our model with less training time. Whenever a new pattern is found by fraud, we need to train our model but this training process takes too much time and memory so we will find an efficient way to solve this problem So that our model can find the accuracy with more efficiency so the customer doesn’t go through the hassle of False Negative of the confusion matrix. But the time of our research is limited within this time it’s not possible to design the best model to solve the problem another limitation is the lack of balanced data or data.

## **Chapter II. Literature Review**

01: (Zhang et al., 2020) A new method proposed for solving the problem of" low-frequency users with small transaction amount", the current approaches cannot exactly illustrate transaction behavior for such users. they draw a new behavior for low-frequency users. To detect the user’s current transaction, they produce an approach based on" user behavior and Naive Bayes".

for low-frequency users with small transaction volumes, the existing methods cannot accurately describe their transaction behaviors for each user, or lead to a high misjudgment rate. So, they propose a new method for individual behavior construction, which can make the behavior of low-frequency users more accurate by migrating the current transaction group behavior and transaction status.

02. Chau et al. (2006) proposed a methodology called 2-Level Fraud Spotting (2LFS) to model the techniques that fraudsters often use to carry out fraudulent activities and to detect offenders. This methodology is used to characterize the auction users online as honest, dishonest, and accomplices. Methodologies that characterize fraud are essential for the first phase of the process since they are the starting point to create a model of the problem and define the best technique for its solution.

03. (Alghofaili, Albattah and Rassam, 2020). They use the "Long Short-Term Memory (LSTM) technique" to improve the traditional financial fraud detection systems. The model identifies unknown and sophisticated patterns of fraud quickly and with high accuracy (99.95% in less than a minute) on a real dataset. The unit is called a long short-term memory block because the program is using a structure founded on short-term memory processes to create longer-term memory. In general, it is an accepted and common concept in guiding recurrent neural networks.

04. (Wang et al., 2018) The design "a deep auto-encoder" by using the similarity of vector representations. To recreate unique bipartite chart taxonomy and estimated that experimental comparability about user conduct. An autoencoder is a type of artificial neural network used to learn efficient data coding in an unsupervised manner. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”.

05. Kang Fu, Dawei Cheng, Yi Tu, and Liqing Zhang at proposed a convolutional based neural network approach (CNN) to find fraudulent transactions. Convolutional neural network is a part of deep learning and is a type of advanced neural network composed of more than one hidden layer. In that document, finding more complex models of fraud and improving classification accuracy, a new feature of commercial entropy is proposed. For the first time, CNN is used to detect fraud used to detect frauds.

06. (Lucas et al., 2019). In this paper they benefit from the temporal data to produce " a strategy to quantify the covariate shift ", this strategy classifies the transactions in each day versus every other (that is there is a covariate shift between days when the classification is good and similar if the classification is not efficient. It is used as a distance matrix describing the covariate shift between days. Then applying the agglomerative clustering algorithm on it. Using a Random Forest classifier, they show that coordinating the information of the sort of the day previously recognized increments the Recall Precision- AUC by (2.5%).

07. (Yee, Sagadevan and Malim, 2018), this paper using Bayesian network classifiers namely Tree Augmented Naïve Bayes (TAN), K2, logistics, J48, and Naïve Bayes classifiers which is all supervised based. All the classifiers give more than 95.0% accuracy when preprocessing the dataset using Principal Component Analysis and normalization, in contrast to scores obtained before preprocessing the dataset. In general, all the Bayesian classifiers give notably better scores after being fed with filtered data.

08. (Author-Dinesh L. Talekar, K. P. Adhiya) Credit Card Fraud Detection System-A Survey: The credit card has become the most popular mode of payment for both online as well as regular purchase, in cases of fraud associated with it are also rising. Credit card frauds are increasing day by day regardless of the various techniques developed for its detection. Fraudsters are so expert that they generate new ways for committing fraudulent transactions each day which demands constant innovation for its detection techniques. Most of the techniques based on Artificial Intelligence, Fuzzy logic, neural network, logistic regression, naïve Bayesian, Machine learning, Sequence Alignment, decision tree, Bayesian network, meta learning, Genetic Programming etc., these are evolved in detecting various credit card fraudulent transactions. This paper presents a survey of various techniques used in credit card fraud detection mechanisms.

09. (Lucas et al., 2020).in this paper a feature engineering strategy has been Adopted by HMM, which permits us to integrate consecutive knowledge in the transactions in the form of HMMbased features. These features which based on HMM- have the ability to make a Random Forest classifier (non-sequential classifier) to use sequential information for the classification process. The strategy of the automated feature engineering based on HMM has "multiple perspective property" which enables us to benefit from incorporating a wide range of sequential information. Actually, the study differentiates the genuine from fraudulent behaviors of the card-holders & the merchants with respect to timing and amount features of the transactions.

10. (Porwal and Mukund, 2019). They produce an ensemble of clustering methods in large data sets to detect fraudulent patterns by using a consistency score for each data point. Their method approved their power in outlier detection in large datasets and changing patterns.

11. (Devi, Biswas and Purkayastha, 2019)). In this work, a random forest algorithm has been developed with cost-sensitive weights to improve the effectiveness of credit card fraud detection. In the training phase and for each tree, a cost-function has been defined, more weight has been assigned for the minority instances during training. The trees are arranged as stated by their ability prediction of the minority class instances. The model has not been checked for high-dimensional datasets. The performance of the proposed method is (F-measure [76.815], Gmean [82.298], and AUC [0.778].

12. (Makki et al., 2019). In this paper, they make a comparison among eight machine learning algorithms performance, and the class imbalance problem has been solved by some solutions. They studied the weakness and the performance of approaches applied to credit card fraud detection. They found that the LR, C5.0 decision tree algorithm, SVM, and ANN are the best methods according to the 3 considered performance measures (Accuracy, Sensitivity, and AUPRC).

13. (Li et al., 2021). They improve the support vector machine parameters by using a cuckoo search algorithm to raise the ability to detect fraud of credit cards. The results demonstrate that CSSVM is superior to SVM in Accuracy, Precision, Recall, F1-score, AUC, and superior to LR, RF, DT, and NB, whose accuracy is 98%.

14. (Priscilla and Prabha, 2020) Their research is a review on Credit Card Fraud Detection (CCFD), the problem of class imbalance in datasets and Machine Learning approaches. The conclusion of the review is Supervised Learning is highly used than the unsupervised for CCFD, the fraud detection techniques which are most commonly used are SVM, Bayesian classifiers, LR, and DT, and Random Forest is the most usage frequency Ensemble learning technique with a good performance. finally, Hybrid methods are better than using a single classifier.

15. (Trivedi et al., 2020). A feedback system mechanism based on machine learning methodology has been introduced as efficient fraud detection for a credit card. This method improves the rate of classifier detection aside from cost-effectiveness. The experiments show that random forest techniques show an accuracy percentage of 95.988 %, although SVM 93.228 %, LR 92.89 %, NB 91.2 %, Decision trees 90.9 %.

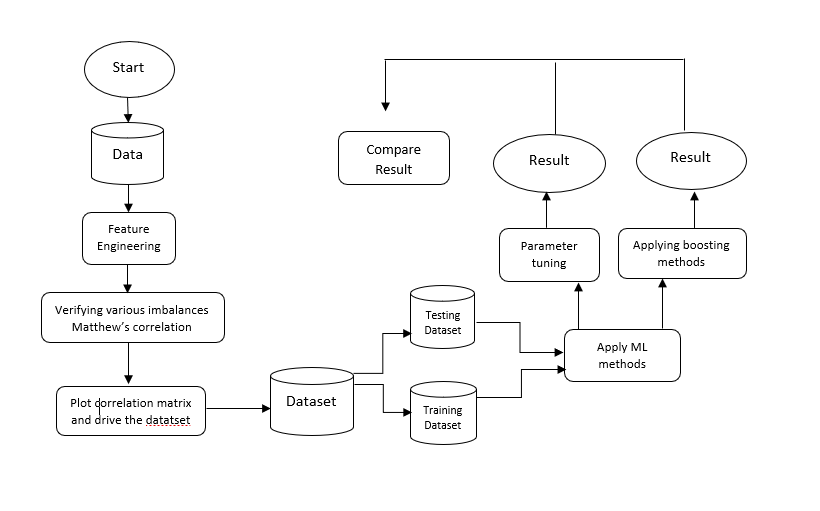
16. (B, Safont and Vergara, 2020). They present a novel approach to increase fraud detection performance by using "the differences in temporal dependence (sequential patterns)" between valid and non-legitimate credit card operations. It is assumed that in feature spaces with low dimension, the successive patterns are more notable than the features with high-dimension space of all of the card operation records. Linear and quadratic discriminant analyses, classification tree, and naive Bayes are the single classifiers used in this model. With these single classifiers and to get better results, Alpha integration has been applied

17. (S et al., 2020) They build the model using some supervised machine learning algorithms such as logistic regression, decision tree, support vector machine. In this project they used two features (time and amount) fundamentally for predicting if the transaction is fraud or not, the number of parameters of feature has been reduced by" time series analysis", then they use either average method, moving average, or window, the naive method and sessional naïve method for achieving the model but all these methods have some advantages and disadvantages.

18. (Mehana and Nuci, 2020)"An incremental learning approach "has been introduced as a new method for real-time fraud detection in online banking transactions. The model AFDM is analyzing transactions based on the transaction instead of accounts which gives more detailed information and represents a good assistant next to detection actions. In addition, the knowledge of the classification algorithm like NB can increment its Updatable transaction by transaction. This novel approach makes the detection and response possibility in real-time, and it gives accuracy (97,2). It also allowed learning new concepts of behavior changes immediately.

## **Chapter III. Research Methodology**

Here the proposed solution is described for eCommerce Transaction Fraud detection. The solution is proposed based on the fraudulent activity issues found from the core data analysis. To get better accuracy the data needed to be balanced and match the mismatch feature values in the dataset. So, the solution was designed in two parts, feature engineering and testing dataset with different algorithms.

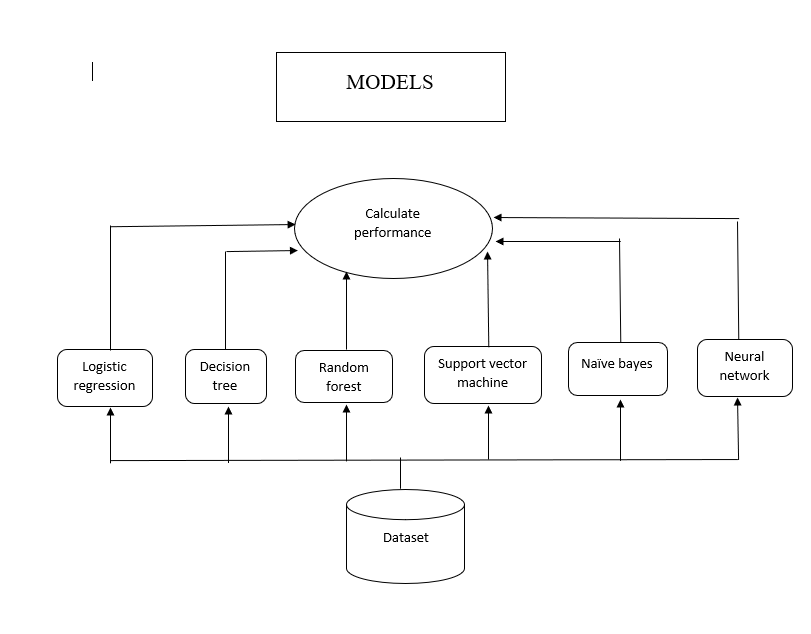


In our first part we have done feature engineering on the data we collected for Kaggle as a secondary data as it's impossible to collect the primary data that require to do reach because the data contain information about customers valuable information like cards number address etc. In our data columns from V1 to V28 are generated from PCA dimensionality and information about them is hidden. But rest of the columns have name and information, column names are given below:

|  |  |
| --- | --- |
| Columns Names | Columns Details |
| Time | Time columns contains the information about Number of seconds passed after the first transaction occurs |
| Amount | Amount contains information about the amount of money sent by card. |
| Class | This column contains 0 and 1 as to whether the transaction was done by fraud or not. 0 stands for not fraud and 1 stands for fraud. |

As the columns from V1 to V28 are generated from PCA dimensionality so we had to scale the rest of the data to work with. We used Sklearn’s RobustScaler method to scale the data. We also created an extra column that gives information about the difference between average amount of transactions incurred by the fraud and non-fraud. After that we compare the two average amounts with every transaction and figure out which average is closer to the amount then we subtracted the amount with the near close difference. Which will help out algorithms to give better accuracy. We also compared the result with the newly added column and without the column in the result section. In most of the research papers researchers used K-fold cross validation which is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. But the K-fold method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. As we are trying to avoid time and high computational power as most of the startup organizations don’t use highly configured devices. So, we didn’t use the method with it until our data cleaning and organizing is completed but our main data is imbalance and accuracy from it can’t take as genuine accuracy as the data is imbalanced, so we used balancing methods like under sampling and oversampling to balance.

In the second part we applied different algorithms to see which algorithm worked better with analyst data.



Logistic Regression**:** Logistic regression, also known as logit regression or logit model, is a mathematical model used in statistics to estimate the probability of an event occurring having been given some previous data. Logistic regression works with binary data, where either the event happens or the event does not happen. At the beginning we’ll use this model to get the result as it is simple and easy to understand. And logistic regression gives the probability of result being either 0 or 1 which is perfect to decide where its fraud or not.

SVM: A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. SVM finds patterns in data which is why we used SVM in our research. As SVM is best for two group classification and our data has two group fraud or nonfraud. Also, SVM uses pattern recognition which is the best way to find fraud in data.

Naïve Bayes:is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. We need fast results so we decided to use an algorithm that will give results fast.

Random Forests: or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.

Neural Network**:** is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, Neural networks refer to systems of neurons, either organic or artificial in nature.

## **Chapter IV. Result and Discussion**

Using Down sampling Method: Here we used Down sampling methods on data that were not modified using the Average difference. Here neural gives the highest accuracy which is almost 100% (99.7%).

Down sampling: In our research the easiest thing to do would be to learn the trivial classifier that says none of the emails are spam. This classifier would give us 99% accuracy, but it wouldn't learn anything interesting, and would have a 100% rate of false negatives.

To solve this problem, we had to do a "down sample", or learn from a subset of the data where 50% of the examples are frauds and 50% are not frauds.

**Class Precision Recall F1 Accuracy**

**------------------------------------------------------------------------**

**Logistic regression** 0.98 0.93 0.96 0.95

**Decision Tree**  0.98 0.93 0.96 0.95

**Random Forest** 1.00 0.88 0.94 0.94

**SVM** 1.00 0.92 0.96 0.96

**Naïve Bayes** 0.98 0.87 0.98 0.92

**Neural network** 1.00 1.00 1.00 1.00

The logistic regression performed very well in classifying the legitimate samples with resampling. Here the precision score is 0.98. This precision actually refers to the percentage of positive results, which are relevant. Then comes the recall score. In this model recall score is 0.93. It actually refers to the percentage of positive cases that are correctly classified. Last one is the f1 score and in this model the score is 0.96. It provides a single score that balances both the concerns of precision and recall in one number and it has a rule also for finding that score. Lastly, the accuracy score of the logistic regression model is 0.95. This was expected because we’re dealing with a balanced class.

The decision tree classifier performed well too. Here the precision score is 0.98 which is good, recall is 0.93 which is less than the previous model and f1 score is 0.96 which is similar to the previous one. And, the accuracy score of the decision tree classifier is 0.95 and we can say both models are close to another.

The random forest's score is good enough. Precision, recall, f1 scores are respectively 1.00, 0.88, 0.94. And, it’s accuracy is 0.94. That means the recall didn’t perform that much well and combinedly it is close enough according to above models.

The support vector machine performs so well. Precision score is 1.00, recall score is 0.92 and f1 score is 0.96 and the result of this model's accuracy is 0.96. We can see clearly that this model works well above all the models.

The naive bayes model's accuracy is 0.92 it works well. Here, the scores are 0.98, 0.87, 0.92 are respectively precision's, recall's and f1's. From all the models its accuracy is less.

The neural network's precision score is 1.00, recall score is 1.00, f1 score is 1.00. The accuracy of this model is 1.00. Actually, we have under-sampled our data. That’s why it cannot perform totally. It is the heart of deep learning. So, within a huge dataset it can perform well there. So, after under sampling our data it shows everything 100%.

Using SMOTE Oversampling Methods:

Here we used SMOTE Oversampling Methods on data that were not modified using the Average difference. Like in down sampling here Neural Network gives better result but not as much better as Down sampled.

**Class Precision Recall F1 Accuracy**

**----------------------------------------------------------------------**

**Logistic Regression** 0.99 0.94 0.96 0.96

**Decision Tree** 0.99 0.94 0.96 0.96

**Random Forest** 0.99 0.88 0.93 0.93

**SVM**  1.00 0.99 1.00 1.00

**Naïve Bayes** 0.99 0.87 0.93 0.93

**Neural Network** 0.98 0.98 0.98 0.98

The logistic regression model’s accuracy is 0.96. It’s precision, recall, f1 scores are respectively 0.99, 0.94, 0.96. So, we can say clearly this model performs well.

The decision tree classifier performed well too. Here the precision score is 0.99 which is good, recall is 0.94 which is same as above model and f1 score is 0.96 which is also similar to the previous one. And, the accuracy score of the decision tree classifier is 0.96 and we can say both models' scores are totally same.

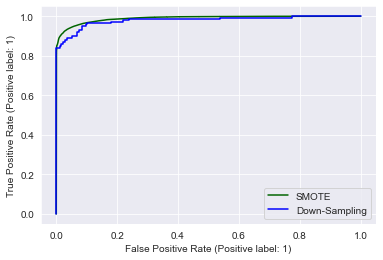
The random forest's score is good enough. Precision, recall, f1 scores are respectively 0.99, 0.88, 0.93. And, it’s accuracy is 0.93. That means the recall didn’t perform that much well and combinedly it is not that much close enough according to above models.

The support vector machine performs so well. Precision score is 1.00, recall score is 0.99 and f1 score is 1.00 and the result of this model's accuracy is 1.00. We can see clearly that this model works good above all the models.

The naive bayes model's accuracy is 0.93 it works well. Here, the scores are 0.99, 0.87, 0.93 are respectively precision's, recall's and f1's. From all the models its accuracy is similar to random forest and the score is less.

The neural network's precision score is 0.98, recall score is 0.98, f1 score is 0.98. The accuracy of this model is 0.98. As I mentioned previously, for the same reason another problem we are also facing is that all scores are the same. So, this model is accurate for our work.

Down sampling Vs Oversampling



## **Chapter V. Conclusion**

In our thesis paper we used a publicly collectable dataset that is provided by Université Libre de Bruxelles for learning purposes which is available in Kaggle. This dataset contains highly imbalanced data because it has only 0.172% fraud of 100% data. We applied different models to minimize false alerts and yet give the best result in fraudulent cases. In our research we found that adding different features doesn't help the models give better results but engineering extended features are more efficient. As we used different models, we found that neural networks perform better in terms of detecting the case whether it is fraudulent or not. But sometimes neural networks predict less fraudulent transactions on oversampled data as we didn’t detect outliers in our dataset. In future applying outlier detection, hybrid resampling can reduce fake alerts and further research on application neural network can bring great changes in fraud detection.

## **References**

‘Fraud detection with machine learning for e commerce and retailers’, *electronic health reporter,* Jan 10, 2020

Olena Kovalenko, ‘ The in depth guide to e-Commerce fraud detection ’, *Project Manager @ SPD-Group, leading AI/ML team and writing articles for our blog*, April 9, 2021

Denis baranov, ‘eCommerce payments fraud prevention: machine learning battle against cybercriminals’, *head of retail & distribution practice* Oct 11, 2021.

‘‘The Enhancement of Credit Card Fraud Detection Systems using Machine Leaming Methodology’’, Master of Applied Science, 2000 Soheila Ehramikar Department of Chernical Engineering and Applied Chernistry University of Toronto

(Zhang et al., 2020) A new method proposed for solving the problem of, ‘low-frequency users with small transaction amount’

IMPROVING ACCURACY IN CREDIT CARD FRAUD DETECTION USING EFFICIENT DATA CLASSIFICATION. School of Computer Science and Engineering Lovely Professional University Phagwara, December 2017 by TAMANNA CHOUHAN

Roy, Abhimanyu, et al. “Deep learning detecting fraud in credit card transactions.” 2018, Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2018

‘Credit card security and e payment’, enquiry into credit card fraud and e-payment by Jithendra Dara & Laxman Gundemoni, 2006:23 ISSN:1653-0187-ISRN: LTU-PB-EX--06/23—SE

Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar (2008). "Credit Card Fraud Detection using Hidden Markov Model". IEEE Transactions on dependable and secure computing,Volume 5; (37-48).

Yvan Lucas. ‘Credit card fraud detection using machine learning with integration of contextual knowledge’ Artificial Intelligence [cs.AI]. Université de Lyon; Universität Passau (Deutscheland), 2019. English.

‘Fraud Detection using Machine Learning in e-Commerce’. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, 2019

‘A Research on Credit Card Fraudulent Detection System’ By Devika S P, Nisarga K S, Gagana P Rao, Chandini S B, Rajkumar N. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019

PRESCRIPTION FRAUD DETECTION VIA DATA MINING: A METHODOLOGY PROPOSAL Karca Duru Aral M.S. in Industrial Engineering Advisors: Prof. İhsan Sabuncuoğlu, Prof. Halil Altay Güvenir July, 2009

December 2018 ‘Application of Machine Learning Techniques in Credit Card Fraud Detection’ Ronish Shakya

‘Credit Card Fraud Detection’, by Munira Ansari, Hashim Malik. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org NREST - 2021 Conference Proceedings.