A Training Report

On

CREDIT CARD FRAUD DETECTION

Submitted in partial fulfilment of requirements for the award of the

Degree of

Bachelor of Technology

In

Computer Science & Engineering

Submitted By

AKASH GUPTA

(20211502717)

Under the guidance of

Dr. Rachna Jain

(Assoc. Professor)

Department of Computer Science & Engineering

Bharati Vidyapeeth's College ofEngineering

A-4, PaschimVihar,New Delhi-110063

July, 2019

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled "**CREDIT CARD FRAUD DETECTION**", in partial fulfilment of the requirement for the award of the degree **Bachelor of Technology** and submitted in **Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering**, **New Delhi** (**Affiliated to Guru Gobind Singh Indraprastha University**) is an authentic record of my own work carried out during the period from June – July 2019 under the guidance of **Dr. Rachna Jain, Asst. Professor.**

The work reported in this has not been submitted by me for award of any other degree of this or any other institute.

(Akash Gupta)
(20211502717)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. They are permitted to appear in the In- House Summer Training.

Dr. Rachna Jain
(Asst. Professor)

ABSTRACT

Fraud is any malicious activity that aims to cause financial loss to the other party. As the use of digital money or plastic money even in developing countries is on the rise so is the fraud associated with them. Frauds caused by Credit Cards have costs consumers and banks billions of dollars globally. Even after numerous mechanisms to stop fraud, fraudsters are continuously trying to find new ways and tricks to commit fraud. Thus, in order to stop these frauds, we need a powerful fraud detection system which not only detects the fraud butalso detects it before it takes place and in an accurate manner. We need to also make our systems learn from the past committed frauds and make them capable of adapting to future new methods of frauds.

In this report we have introduced the concept of frauds related to credit cards and their various types. We have explained various techniques available for a fraud detection system such as **Support Vector Machine (SVM), Bayesian Network, K- Nearest Neighbor (KNN), Random Forest and Decision Trees**. An extensive review is done on the existing model for credit card fraud detection and has done a comparative study on these techniques on the basis of accuracy, and by confusion matrix.

The results indicate the optimal accuracy for Support Vector Machine, Naïve Bayes, K- Nearest Neighbor, Random Forest Classification and Decision Tree Classification are 91.87%, 90.24%, 89.83%, 91.05% and 88.62% respectively.

ACKNOWLEDGEMENT

I express my deep gratitude to **Dr. Rachna Jain**, Asst. Professor, Department of Computer Science & Engineering, for her valuable guidance and suggestion throughout my training. We are thankful to **Ms. Shilpa Gupta** (CSE,3rd year,1st Shift) for their valuable guidance.

We would like to extend my sincere thanks to **Head of the Department, Dr. Kirti Gupta** for his time to time suggestions to complete my project work. I am also thankful to **Dr. Dharmender Saini, Principal** for providing me the facilities to carry out my project work.

Sign

(Akash Gupta)

20211502717

TABLE OF CONTENT

LIST OF IMAGES

LIST OF ABBREVIATIONS

KNN	K Nearest Neighbour
SVM	Support Vector Machine

LIST OF TABLES

INTRODUCTION

Year after year, the damages inflicted by the credit card fraud problem are growing rapidly. In the year 2014 alone, it was estimated that the total global monetary loss was in 16.31 billion dollars — accumulation of damages from card issuers, acquiring banks and merchants [1]. Although new technology is introduced to the public and being mandated by the government agency [2] to replace the old magnetic stripe to EMV (Europay-MasterCard- VISA) [3], some group of individuals are starting to challenge its design and implementation [4] [5]. The advancements in our technology and ease of availment opened more opportunities for fraudsters' ability to speed-up their execution plan and retain their anonymity at the same time. Surely, a layer of security will not be enough to protect the card holders, merchants, and issuing banks from a possible attack. There should be another layer that must be available to proactively detect these anomalies.

Evaluation of credit card related fraud cases in the past two (2) decades reveals that the top five (5) modus-operandi performed by the fraudsters are: (i) counterfeit credit cards, (ii) lost or stolen, (iii) no-card fraud (e.g., giving card information to non-legitimate telemarketer), (iv) stolen cards during mailing fraud, and lastly (v) identity-theft fraud [6]. These fraud cases occupy 81% of the known fraud types in the credit card industry. It may seem very common to the banks and merchants but to date, they are still being victimized by such attacks.

In retrospect, data-mining techniques are now well established. Nevertheless, researches pertaining to this area are very limited due to privacy issues. Bank customers are well protected by several laws [7]–[8] [9] that prohibit the disclosure of their personal information without proper consent. However, these papers were able to acquire data — their strategy is to combine information from customers, accounts, cards, and transaction datasets. This study will focus primarily on the transaction details made by the card holder.

The intention of this study is to fully explore the effectiveness of utilizing the credit card transaction logs to differentiate anomalous from legitimate transactions. With this, various learning algorithms available in Weka will be evaluated by measuring their effectiveness in predicting the correct classification of the input dataset.

LITERATURE SURVEY

Erkin et al. [[10], Springer 2013] proposed privacy preserving using distributed clustering. Here, K-means Clustering Algorithm is used. The goal was to stop the user from reading the private information while the performance should not be affected. The work was good but there are large numbers of problems as it uses squared Euclidean distance. Implementation may be done to solve this problem by using combination of technique which consists of Hidden Markov Model, Behavior Based Technique, and Genetic Algorithm. By filtering the obtained data fraud is detected to obtain better result.

Er. Teng et al. [[11], Springer 2012] proposed Customer Credit Scoring based on HMM/GMDH hybrid model. It uses the combination of Hidden Markov Model and Group Method of Data Handling. The Hybrid Model uses customer features as input and considers customer information in modeling process, and result show that it can improve the effect of Credit Scoring [11]. The work was good but more work can be done by improving accuracy.

Tejpal singh et al. [[12], Nuicone 2012] implemented Credit Card fraudulent Detection system using observation probabilistic in Hidden Markov Model. In this K-means Algorithm is used for clustering and Hidden Markov Model for fraudulent detection. Hidden Markov Model focuses on spending profile of cardholder and detects fraud, so sometimes it detects wrongly frauds. The work was good but more improvement may be done for detecting accurate frauds.

Alowais et al. [[13], IEEE 2012] proposed credit card fraud detection using personalized or aggregated model. As banking industry suffers lost in millions of dollars caused by credit card fraud [13]. Here, models are created for each card holder known as personalized model and compared with aggregated model. The work was good but accuracy of personalized models are found to be better than personalized model. This can be improved by using Credit Scoring Model.

Edwin Raj et al. [[14], IEEE 2011] had proposed Analysis on Credit Card Fraud Detection methods. In this various technique of Credit Card Fraud Detection are compared. Each technique has its own advantage and disadvantage. The work was good but combination of various methods will give better result.

METHODOLOGY & IMPLEMENTATION

3.1 About the dataset

The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where 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, we cannot 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' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

3.2 The Problem with Imbalanced Classes

Imbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio of observations in each class.

Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error.

Fig 1

3.3 The Problem with Accuracy

Here we can use the Dummy Classifier to always predict “not fraud” just to show how misleading accuracy can be.

We got an accuracy score of 99.8% — And without even training a model! Let’s compare this to logistic regression, an actual trained classifier.

Maybe not surprisingly, our accuracy score decreased as compared to the dummy classifier above. This tells us that either we did something wrong in our model, or that accuracy might not be our best option for measuring performance.

3.4 Resampling Techniques

3.4.1 Oversample Minority Class

Oversampling can be defined as adding more copies of the minority class. Oversampling can be a good choice when you don’t have a ton of data to workwith.
We will use the resampling module from Scikit-Learn to randomly replicate samples from the minority class.

3.4.2 Undersample Majority Class

Undersampling can be defined as removing some observations of the majority class. Undersampling can be a good choice when you have a ton of data -think millions of rows. But a drawback is that we are removing information that may be valuable. This could lead to underfitting and poor generalization to the test set.

3.5 Implementation

In this research, the initial stage is to do online shopping which includes Registration, Login, Banking details and so on. After feeding all these details transaction is processed. Then each and every transaction is tested with proposed model which uses Random Forest Algorithm, Decision Tree Algorithm, Naïve Bayes, KNN Algorithm, SVM Algorithm. So by testing with these methods values are obtained and the one with highest accuracy is taken out. As declared

within the previous section to enhance the performance of Credit Card Fraud Detection, this analysis work projected an economical method referred to as Credit Card Fraud Detection; that performs higher and will improve the performance of prevailing system. Then each and every transaction is tested individually with proposed model which uses Random Forest Algorithm, Decision Tree Algorithm, Naïve Bayes, KNN Algorithm, SVM Algorithm.

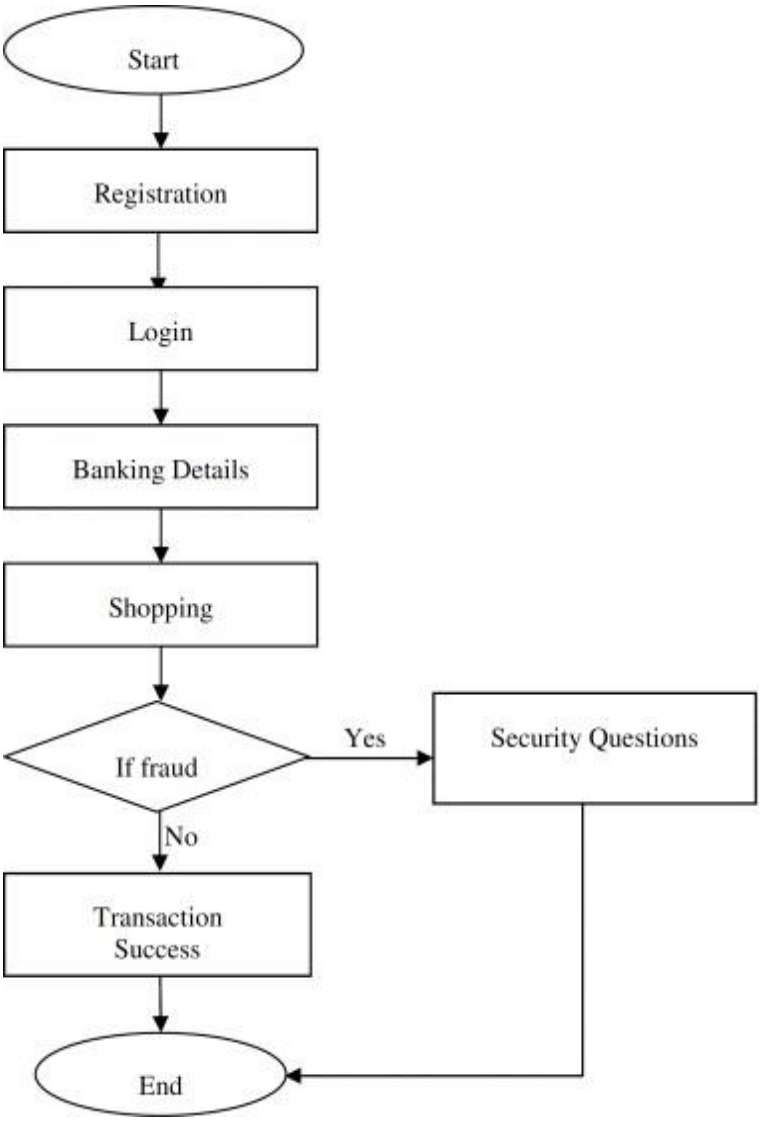

Fig 2

3.6 Results and Analysis

A. Random Forest Classifier

For the data tested on sets of the same ratios, the classifier experienced the smallest decrease in the F-1 score. While the precision of the Random forests model experiencing only a slight

drop, the recall experienced large decreases as the normal to fraudulent ratio increased. While applying on large dataset, the random forest algorithm reached the optimal normal-to-fraud ratio i.e., 30:1 ratio. As seen in Figure 11, the Random Forest algorithm was indicative of the pattern observed in Figure 10, as the F-1 score reaches its peak score, the precision and recall lines intersect each other at the same ratio. As a result, for the random forest algorithm, it would be necessary to find the optimal ratio to train the dataset on in order to use random subsampling to reduce bias. The Random Forest is the most efficient algorithm for testing on biased datasets at the optimal dataset because, unlike the SVM, the Random Forests algorithm does not require much processing power and time to train efficiently.

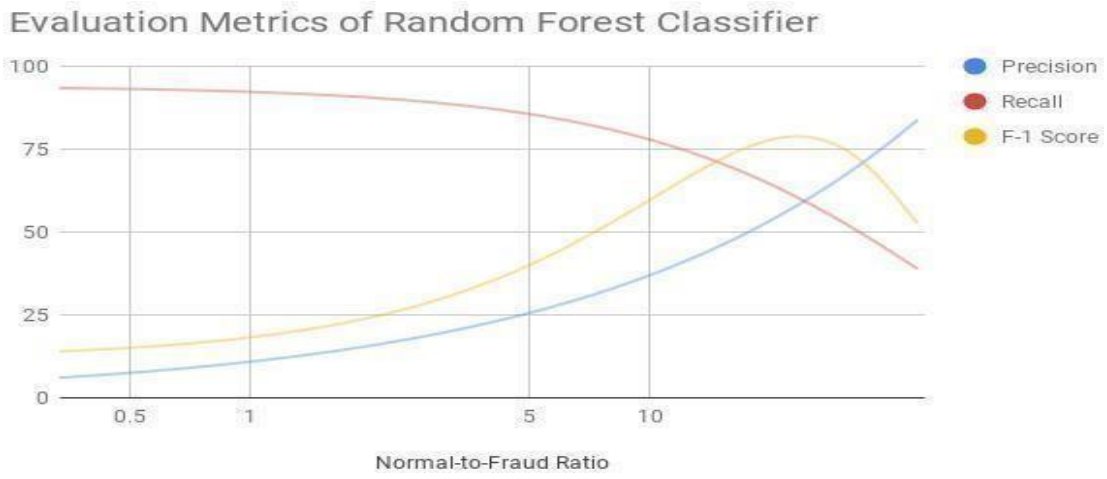

Fig 3

B. K Nearest Neighbor (KNN)

When tested with the 25:75,40:60, 50:50, 60:40, 75:25, and 90:10 non-fraudulent to fraudulent test split, all factors consistently showed that, with an increasing ratio, the F-1 Score decreased. For example, in the25:75 split, the F1 Score is about 81%.

However, at the90:10 split, the F1 Score dropped to about 19%. However, when training the algorithm with the 75:25, 60:40, 50:50, 40:60, 25:75, and 10:90 non- fraudulent to fraudulent test split and tested the whole dataset, the F-1 Score was higher when the ratio was higher.

This is possibly because the testing set is part of the training set. In other words, for many of the cases, the Euclidean distance calculated was zero. Thus, the testing set assumed to be fraudulent based on itself.

Thus, unlike with the Random Forest algorithm, the intersection of recall and precision do not correlate with the training dataset at which KNN is its most effective.

C. Support Vector Machine (SVM)

Results of the Support Vector Machine algorithm tested and trained with adjusted datasets of approximately equal fraud to non-fraud ratios indicated an average F-1 Score of 79.96%, regardless of the combination of factors. Unlike most models presented in this paper, the F-score did not decrease as the normal-to-fraud ratio decreased. It stayed at approximately 84%

Results of the Support Vector Machine algorithm varied when trained with an adjusted dataset but tested with the unadjusted and highly skewed data set. F-1 scores averaged 94.07% for the SVM with all combinations of factors, which is significantly higher compared to the data for the five other algorithms.

The average F-scores of the tests under different ratio combinations but same factor combination was averaged to determine the ideal combination of factors for analysis. The average F- 1 scores for the Support Vector Machine analyzing the "hour1", "field3", and "hour 1 and field 3" factor combinations of transactions was calculated to be 94.51%, 94.4%, and 93.27%, respectively.

In contrast to other algorithms, the SVM algorithm took much more time and computing power to complete the fitting of the model. Compared to the Random Forest Classifier, the model takes much more time to execute. As a result, when processing real-time data such as credit card transactions on datasets that would be much larger than the UCSD- FICO set, the SVM model would have to be made more efficient in order to process and classify fraud data in a reasonable amount of time after the transaction.

D. Naïve Bayes Classifier

When the Naive Bayes classifier was used to predict fraud with a training and testing data set of equal bias and length, there was a clear trend in the F-1 score.

Ratio s	Field1	Hour1	Field3	Hour1 + Field3	Avera ge
25-75	83.5	86.3	85.7	86.5	85.5
40-60	70	76	70	76.7	73.12
50-50	64.3	68	44	71	61.82
60-40	58.7	56	33	62	52.4
75-25	0	43	20	51	28.5
90-10	0	19	10	20	12.2

Table 1

As seen in Table 1, the Naive Bayes Classifier had the highest average f-score (85.5) when trained and tested with the dataset with the lowest normal- to- fraud ratio of 25:75. The F-1 - score dropped exponentially following a logistic curve as the normal-to-fraud ratio increased. When training and testing with the 90:10 ratio, the F-1 score reached its average lowest (12.2). This trend can be explained by the functionality of the Naive Bayes Classifier. As there is a greater normal-to-fraud ratio, the algorithm is provided with more evidence of trends pointing towards nor- mal transactions relative to evidence of fraudulent transactions. With a greater percentage of normal transactions within the dataset, there is also greater variability in the range of values within each field which indicate a normal transaction. Thus, there is overlap between the patterns indicating fraudulent and normal transactions. Coupled together, these two reasons make the algorithm overwhelmingly predict every transaction as fraudulent, decreasing the F-1 score.

When the Naive Bayes classifier was used to predict fraud with a training and testing data set of different bias and length, a different pattern emerged. When graphed, the average f-scores of the fields which were trained form a inverted-U curve. There is a clear peak in the F-1 score right after the ratio 75:25. This shape can be explained by the nature of training and testing data. By training the algorithm with data sets of lower ratios, a testing set of a similar ratio was expected, which thus decreased the precision of the classifier but increased the recall. As the normal-to-fraud ratio increased closer to the peak found, precision increased and recall decreased as the ratio in the training set grew closer to the testing set. However, once reaching the peak, denoting the optimal ratio for the Naïve Bayes Classifier to function, a similar effect indicated in Table 3 took place and decreased the efficacy of the classifier. The Naive Bayes classifier performed well for this project, however the simplicity of the classifier prevented it from observing patterns of great complexity within the data which ultimately reduced its efficacy in detecting credit card fraud.

CONCLUSION

The main concern of this research was to detect least and accurate false fraud detection. Credit card fraud is lack of security. Credit Card Fraud can be used for applications as paying with credit card make it easier to avoid loses from fraud. This project has examined the performance of five kinds of classification models namely Random Forest Algorithm, Decision Tree Algorithm, Naïve Bayes, KNN Algorithm, SVM Algorithm. A real-life dataset on credit card transactions is used in our experiment. Among all the methods, the SVM one has the maximum accuracy.

Although SVM obtains good results on small set data, there are still some problems such as imbalanced data. Our future work will focus on solving these problems. For example, the voting mechanism assumes that each of base classifiers has equal weight, but some of them may be more important than others. Therefore, we also try to make some improvement for this algorithm.

The results indicate the optimal accuracy for SVM, Naïve Bayes, KNN, Random Forest Classification and Decision Tree Classification are 91.87%, 90.24%, 89.83%, 91.05% and 88.62 % respectively.

FUTURE SCOPE

This research on detecting credit card fraud has great potential for future implications. If a dataset with unencrypted fields was released to the public, the true factors which can be traced for credit card fraud detection can known. Therefore, credit card companies can be informed about the most important factors to analyze when predicting credit card fraud and improve the efficiency of their notification systems.

Furthermore, the results ofthis project were limited by the small sample size of fraudulent cases provided by the data set. By using a larger dataset with a greater number of fraudulent cases, the algorithms can be trained to make predictions of greater precision. In order to pursue these goals, more computing power may be required. It may be important to consider using a Graphical Processing Unit like the Nvidia Jetson II to improve the productivity of training and testing each algorithm with a larger, more complex dataset.

Other methods for bias prevention, such asother resampling techniques, cost-sensitive learning methods, and ensemble learning methods could also be tested in future datasets to discover the best method of dealing with skewed data sets.

Ultimately, the results of this research project can provide insight on the best algorithm to be used in other cases of data analysis on skewed data sets, such as in natural disaster prediction.

REFERENCES

[1] Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K. Majumdar, "CreditCard Fraud Detection Using Hidden Markov Model" 2008

[2] V.Bhusari ,S.Patil ," Study of Hidden Markov Model in Credit Card Fraudulent Detection " 2008

[3] S Pratap Singh, Shiv Shankar P. Rakesh and Vipin Tyagi "Problem Reduction In Online Payment System Using Hybrid Model" 2011

[4] Wong, N., Ray, P., Stephens, G. And Lewis, L, "Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results" 2011

[5] Panigrahi, S. kundu, A. Sural, And Majumdar, A. "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning" 2009

[6] Wang, L., Zhang, S., LI, Y., WU, R. And YU, Y. "An Attribute-weighted Clustering Intrusion Detection Method" 2013

[7] Porkess, R. And Mason, S., "Looking at debit and credit card fraud" 2011

[8] Dorronsoro, J., Ginel, F., Sgnchez, C. And Cruz, C., "Neural fraud detection in credit card operations" 1997

[9] Abdul Razak, T. And Najeeb Ahmed, G., "Detecting Credit Card Fraud using Data Mining Techniques - Meta-Learning" 2015

[10] Erkin, Z., Toft, T. and Lagendijk, R. "Privacy-preserving distribution Clustering" 2013

[11] Ge-Er Teng, Chang-Zheng He, Jin Xiao, Xiao-Yi Jiang, “Customer credit scoringbased on HMM/GMDH hybrid model” 2012

[12] Ashphak Khan, Tejpal Singh, Amit Sinhal, “Implement Credit Card Fraudulent Detection System Using Observation Probabilistic in Hidden Markov Model” 2012

[13] M Alowais, L Soon, “Credit Card Fraud Detection: Personalized or Aggregated Model”2012

[14] S. Edwin Raj, A. Annie Portia “Analysis on Credit Card Fraud Detection Methods” 2011