Machine Learning Based Identification of Credit Card Fraud

Mohammad Nurul Abrar

ID: 2018–1-60-139

Md Maruf

ID: 2018–1-60-140

Md Solaiman

ID: 2018-1-60-128

Nayma Alam

ID: 2018-1-60-180

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering
East West University
Dhaka-1212, Bangladesh

May, 2022

Declaration

I, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by me under the supervision of name of your super visor , Professor, Department of Computer Science and engineering, East West University. I also declare that no part of this thesis/project has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

.....

Musharrat Khan

Supervisor

Signature

.....

Mohammad Nurul Abrar

2018-1-60-139

Signature

.....

Md Maruf

2018-1-60-140

Signature

.....

Md Solaiman

2018-1-60-128

Signature

.....

Nayma Alam

2018-1-60-180

Abstract

According to recent studies, credit card fraud has become a major concern for financial institutions, as well as for low-income people who struggle to make financial decisions. We presented a machine learning-based model for credit card fraud detection to resolve these concerns. We used several feature selection and dimension reduction techniques to get a better dataset, then applied various machine learning algorithms. We found that SVM and RF have the best accuracy after applying these algorithms. Univariate and XGBoost outperformed the other techniques in feature selection and dimension reduction.

Acknowledgments

For the better understanding, a sample Acknowledgment is given below.

As it is true for everyone, I/We have also arrived at this point of achieving a goal in my/our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, I/We would not like to make efforts to find best words to express my thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, I/We would like to express my deepest gratitude to the almighty for His blessings on me/us. Next, my/our special thanks go to my/our supervisor, "Name of Your Supervisor", who gave me/us this opportunity, initiated me/us into the field of "Name of your Thesis/Project Field", and without whom this work would not have been possible. His encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of my/our BS.c study were simply appreciating and essential. His ability to muddle me/us enough to finally answer my/our own question correctly is something valuable what I/We have learned and I/We would try to emulate, if ever I/We get the opportunity.

I/We would like to thank "Name of your Friend" for his excellent collaboration during performance evaluation studies; "Name of your Friend" for his overall support; "Name of Some one" for her helpful suggestions in solving tricky technical problems. Last but

not the least, I/We would like to thank my/our parents for their unending support, encouragement and prayers.

There are numerous other people too who have shown me their constant support and friendship in various ways, directly or indirectly related to my/our academic life. I/We will remember them in my/our heart and hope to find a more appropriate place to acknowledge them in the future.

Mohammad Nurul Abrar
May, 2022

Md Maruf
May, 2022

Md Solaiman
May, 2022

Nayma Alamr
May, 2022

Table of Contents

Chapter 1 Introduction

Data mining along with machine learning is one of the progressive research areas for detecting credit card fraud. Introducing credit cards in the banking industry has been known as one of the progressive steps toward the future. The credit card was released in 1950 [1]. But modern credit cards were invented in 1946 by John Biggins [2]. Although the concept was far older. A credit card is a portable thin plastic card that holds identification information and authorizes the person named on it to charge any type of purchases or services to his account. Nowadays, the information on the credit card is read by the ATM(Automated Teller Machines), bank and is also used in online banking system[]. There is a rapid growth of credit card transactions which has conducted to a factual rise in fraudulent activities. Credit card fraud is a widespread term for theft and fraud committed using credit card as a fraudulent source of funds in a given transactions. Usage of this plastic card with identification elements has massively increased for purchasing goods or paying bills. Involvement of this credit card in various fields also increases the vulnerability of the system or can be said as the miss-user of the system. In here missuser or defaulters are those who have not made the minimum payment for several months compared to traditional loans. Debt repayment in credit card loans is minimum as compared to the credit balance which will lead to a great risk on the loan lenders.

Generally, many data mining algorithms and the statistical methods are used to solved this fraud detection problem. Most of the fraud detection problems are based on meta learning, artificial intelligence and pattern matching. Here, so many importance are given to develop productive and secure electronic payment system to detect whether a transaction is fraudulent or not. As the users of this system are increasing enormously and risk is also for lenders that's why for good decision making machine learning and data mining various techniques will help. Predicting the defaulter or the fraud which is a must to control the risk for the lenders. In this paper, we will focus on credit card fraud and its detection measures. A credit card fraud occurs when one individual person uses other individuals card for their personal use without the knowledge of its owner.

Considering financial statement fraud fetch huge property damage to investors, a large number of researches have been conducted on the this area using various techniques and machine learning methods and make more accurate models. In the study of this paper[21], Six data mining techniques (Logistic Regression, Naïve Bayes, FLDA, J48, MLP, and IBK) are applied and the results of this research indicate that the neural network shows the highest accuracy. In this[20][22] research paper, Seven data mining techniques (KNN,ANN, SVM, decision tree, NBTree, RF, Rotation Forest) are applied and by utilizing SMOTE method, the predicting model produced by Random Forest has the best accuracy and performance with low error rate.

The goal of this paper is to increase the accuracy of the models and predict better by balancing the data and pre-processing along with that, find best algorithm with best dimension so that model works in optimized condition.

Chapter 2

Background

The worldwide customer uses credit cards around the clock and its user base is getting bigger and bigger. To detect fraud from that transaction history it needs to be classified and measured by certain rules. For that, there are some statistical methods and techniques available. By those techniques and methods, machines can learn previous fraud patterns and predict any possible fraud. In this paper, several dimension reduction methods, feature selection techniques and algorithms were used to find the best possible models and their accuracy for finding the fraud.

2.1 Feature Selection

I. **XGboost:** XGBoost stands for "Extreme Gradient Boosting". XGBoost is a distributed gradient boosting technique that is optimized for efficiency, flexibility, and portability. It uses the Gradient Boosting framework to create Machine Learning algorithms. It uses parallel tree boosting to tackle a wide range of data science issues quickly and accurately [3].

II. **Chi-square**: In statistics, the chi-square test is used to determine if two events are independent. We can get observed count and predicted count from the data of two variables. The Chi-Square test determines how far predicted count and observed count differ. The goal of feature selection is to choose characteristics that are strongly dependent on the response. Because the observed count is close to the expected count when two features are independent, the Chi-Square value is less. The high Chi-Square score suggests that the independence hypothesis is false. Simply put, the higher the Chi-Square value, the more dependent on the response the feature is, and it can be chosen for model training [4].

III. **GINI index**: The Gini Index, also known as Gini impurity, assesses the likelihood of a certain feature being incorrectly identified when randomly selected. It is said to be pure if all of the elements are related to a single class. Let us consider the Gini Index criterion. Like the qualities of entropy, the Gini index ranges from 0 to 1, with 0 expressing the purity of categorization, i.e. all items belong to one class or only one class exists. And 1 denotes a random

distribution of components among different classes. The Gini Index value of 0.5 indicates that items are distributed evenly across several classifications. The Gini index only performs binary splits on categorical target variables in terms of "success" or "failure." [5].

IV. **Univariate**: Univariate feature selection looks at each feature separately to see how strong of a relationship it has with the response variable. These strategies are straightforward to use and comprehend, and they are especially useful for better understanding data (but not necessarily for optimizing the feature set for better generalization). The best features are chosen using univariate statistical tests in univariate feature selection. Each feature is compared to the target variable to check if there is a statistically significant association between them. Analysis of variance is another name for it (ANOVA). We ignore the other features while analyzing the relationship between one feature and the target variable. That is why it is referred to as "univariate." Each feature has a test score associated with it. Finally, all of the test results are compared, and the features with the highest scores are chosen [6].

2.2 Dimension Reduction

I. **PCA:** Principal Component Analysis, or PCA, is a dimensionality-reduction approach for reducing the dimensionality of large data sets by transforming a large collection of variables into a smaller one that retains the majority of the information in the large set. PCA can help us enhance performance without sacrificing model accuracy. Other advantages of PCA include data noise reduction, feature selection (to a degree), and the capacity to generate independent, uncorrelated data features [7].

II. **Pearson correlation coefficient:** The population parameter is represented by the Greek letter rho (ρ), while the sample statistic is represented by the Greek letter r. The correlation coefficient is a single number that indicates the degree and direction of a linear relationship between two continuous variables. Which measures linear correlation between two variables.

III. **SVD**: Dimensionality reduction is the process of reducing the amount of input variables for a predictive model. A simpler predictive model with fewer input variables may perform better when making predictions based on new data. Singular Value Decomposition, or SVD, is the most widely used dimensionality reduction approach in machine learning [8]. With sparse data, it works better. Sparse data is defined as data with a large number of zero values. Sparse data is

created in numerous situations, for as in a product recommendation system. The number of columns in a truncated SVD factorized data matrix equals the truncation. It mathematically shortens the value of float digits by dropping the digits following the decimal place. If a matrix m×n given, the truncated SVD will produce matrices with the specified number of columns, whereas a normal SVD procedure will produce with m columns. It means that it will drop off all features or columns except the number of features provided to it.

IV. **Entropy**: Entropy-based feature selection is based on the criteria of IG. It chooses the features that provide the highest information gain. To compute the information gain of features, we must first calculate the probability and entropy of the classes in the data set.(Singh, K. & Raut, Abhinav. (2014). Feature Selection for Anomaly Based Intrusion Detection using Rough Set theory.) For each variable in the dataset, the information gain is determined. To separate the dataset, the variable with the most information gain is chosen. A higher gain usually signifies a lower entropy or less surprise [9].

2.3 Algorithm

I. **Naive Bayes**: Naive Bayes uses a simple probabilistic classifier to predict the probability of different classes based on various attributes. This classifier is one of the simplest algorithms. This algorithm calculates the set of probabilities by measuring the occurrence and group of values from the data set. Bayes's theorem and theory of probability are the basis for the Naive Bayesian classifier. Here, all attributes are considered independent variables. In real life scenario, independence assumption is not quite relatable. Although this algorithm performs quite well and learning is also fast in different supervised classification [10].

II. **SVM**: SVM is based on statistical learning theory which is a supervised model and can be used for regression along with classification tasks. . They can be used for learning and recognizing different patterns to predict future data or for classification. A constrained quadratic optimization problem is used to train. SVM uses a set of nonlinear basis functions to map inputs onto a high-dimensional space.The basic principle behind SVM is to discover an ideal hyperplane that can divide two classes of support vector instances [11].

III. **Random Forest**: Random forests are a collection of regression or classification which is a supervised machine learning algorithm for trees that are trained on boot samples of training data with random feature selection for tree construction. Following the tree construction process, each tree must vote for the most popular

class. When creating each individual tree, it employs bagging and feature randomization in order to generate an uncorrelated forest of trees whose committee prediction is more accurate than that of any one tree. This method of tree voting is called Random forest.

IV. **Decision Tree**: For credit scoring structures, this is the most commonly used algorithm. A decision tree is made up of several internal nodes that run several tests on input variables and attributes in order to divide the data into smaller chunks. It also has leaf nodes that assign a class to each of the observations we obtained after running various tests. This process continues until the required requirements are met. We can claim that the node's purity improves as the target variable increases.

V. **KNN**: It is a machine learning technique that uses a non-parametric classifier. This method is based on a basic analogy, in which a particular test case is compared to training cases that are similar to it. The K-nearest technique determines the pattern space for the k training observations that are comparable to the new case when a new sample is introduced. The k "nearest neighbors" of the new sample are these k training samples. The "degree of closeness" is calculated using a distance metric, such as Euclidean distance. If the new sample has k-nearest neighbors, it will be classed as the most common class among them using this classification technique. If K = I, the sample is simply classed as the nearest neighbor's class. It is usual practice to use a bigger value of k when there are more training examples. However, when dealing with huge datasets, the K-nearest approach demands a lot of time and work.

VI. **Logistic Regression**: Logistic Regression is a type of generalized linear model. The probability for categorization problems with two possible outcomes are modeled using logistic regression. It's a classification problem extension of the linear regression model. The logistic regression model uses the logistic function to squeeze the outcome of a linear equation between 0 and 1 instead of fitting a straight line or hyperplane [12].

VII. **Multilayer Perceptron (MLP)**: Hopfield network, self-arranging neural networks, mean-field theory machine, RB (radial basis) function, and multi-layer perceptron are some of the neural networks that have been developed and studied. MLP is a very significant technique for problems in a bigger area. Unlike other classification algorithms like Support Vectors or Naive Bayes Classifier, MLPClassifier does classification using an underlying Neural Network [13]. Multilayer perceptrons (MLPs) that are trained using the standard backpropagation algorithm are called feed forward neural networks. Because this approach is supervised learning, it requires the training of accurate targets. Because of their ability to learn how to convert input to the appropriate target, they are often utilized in pattern categorization. With one or two hidden layers, they can estimate practically any input-output map. Multi-layer perceptrons are

used in many neural networks. It is possibly the most popular network architecture. The units are arranged in a layered feed-forward topology. Each unit calculates a biased weighted sum of its inputs. The activation function is then used to generate their output. The model of a multilayer perceptron includes input, output, weights, and threshold values [14].

VIII. **CNN:** .A convolutional neural network (CNN or ConvNet) is a deep learning network architecture that learns from data without the requirement for manual feature extraction. CNNs are particularly useful for recognizing objects, faces, and scenes by looking for patterns in images. The primary idea behind the CNN algorithm is to reduce the number of training samples by removing data that exhibits similarity and not adding any more information [15].

Chapter 3

Literature Review

The dataset utilized in the study [16] was the financial fraud identification in the healthcare dataset. The sequential model and the other four models of a classifier based on Logistic Regression, KNN, Naive Bayes, and Support Vector Machine are developed. Accuracy, precision, specificity, sensitivity, and the Matthews correlation coefficient (MCC) with the rate of balance classification are applied for measuring the performance of all these classifier models. The performance of all these machine learning models is evaluated. The random forest model visualizes the better performance. The technique of the random forest model generates superior performance for the evaluation metrics applied. It generates the highest value for specificity and precision.

In the study of this paper [17], they have used two different datasets and propose a novel credit card fraud detection system based on Long Short Term Memory (LSTM) networks and attention mechanism. Attention mechanism approves a sequence based neural network to automatically focus on the data items that are the most important to the alignment task by a data-driven weighted average of local information taken in each term of the sequence which results in a high detection performance. The proposed model achieves better results in term of accuracy, precision and sensitivity (recall) than the compared classification GRU, SVM, LSTM, KNN and ANN methods, which demonstrate the effectiveness of proposed model in this paper on credit card fraud detection task.

In this research work [18], it has been proposed to select the important features required to detect fraud in financial statements. A multi-model approach is accepted in this work to select the most significant features. For the dataset selected, a total of 38 (M1-M38) models were trained and their performance was tested on testing data. Several kinds of models of regression, boosting, bagging, and tree ensembles are used in the training and predictions. By using the top metrics like accuracy, sensitivity, and precision; two models called, Parallel random forest and the Stochastic gradient boosting method were selected. The Parallel random forest yielded 100% accuracy, sensitivity, and negative precision on the training set of the undersampled dataset; and 65.48%, 72.72%and 98.87% of accuracy, sensitivity, and negative precision on the testing set respectively. Similarly, the Stochastic gradient boosting method yielded 91.73%% accuracy, 95.53% of sensitivity, and 95.16% of negative precision on the training set of oversampled dataset and 84.76%%, 39.39%and 98.11% of accuracy, sensitivity, and negative precision on the testing set respectively

The purpose of this study [19] is to propose a hybrid fraudulent financial statement detection model combining the PCA and XGboost to do feature selection, and then, RF, SVM, ANN and LR were applied to construct the fraud detection model, and the classification accuracy of each model was compared to determine the optimal model, where the study indicates that random forest gives the high performance compared to other methods.

In this study [20], in terms of risk management, they offer a data mining-based failure prevention system. This study focused on customer default payments in order Prediction of default payment of credit card clients using Data Mining Techniques to accurately estimate the likelihood of default. Because unbalanced datasets are a type of default dataset, this work uses the Synthetic Minority Over-Sampling Technique (SMOTE) to cope with them. Random Forest's predicting model provides the greatest accuracy and performance with a low error rate because of the SMOTE approach. The proposed Random Forest model classifies Default of Credit Card Clients in the test data by predicting the target variable 89.01% correctly. This effect also resulted in an improvement of ROC area (AUC=0.947, and F-measure (0.89) of Random Forest, which was higher than other classifiers.

According to this research [21], the neural network works best and has the highest accuracy in predicting credit card default. The data set is processed into six data mining algorithms (FLDA, Naive Bayes, J48, Logistic Regression, MLP, and IBK). In terms of risk management, accuracy can be predicted clearer and more concisely than just expressing binary result categories of "Credible" or "Not Credible". The results of this study demonstrate that the neural network works best and has the highest accuracy in predicting credit card default. With the use of data mining techniques, the study develops a model for predicting default cards. SMOTE and ADASYN algorithms are applied to balance the unbalanced data and improve the model's efficiency. Later, the two balancing approaches are compared to evaluate which one is more effective. To anticipate default credit cards, this balanced data is used as input to a machine learning system like SVM. The model's accuracy is calculated by evaluating it to other data models.

With the use of data mining techniques, the study [22] develops a model for predicting default cards. SMOTE and ADASYN algorithms are applied to balance the unbalanced data and improve the model's efficiency. Later, the two balancing approaches are compared to evaluate which one is more effective. To anticipate default credit cards, this balanced data is used as input to a machine learning system like SVM. The model's accuracy is calculated by evaluating it to other data models.

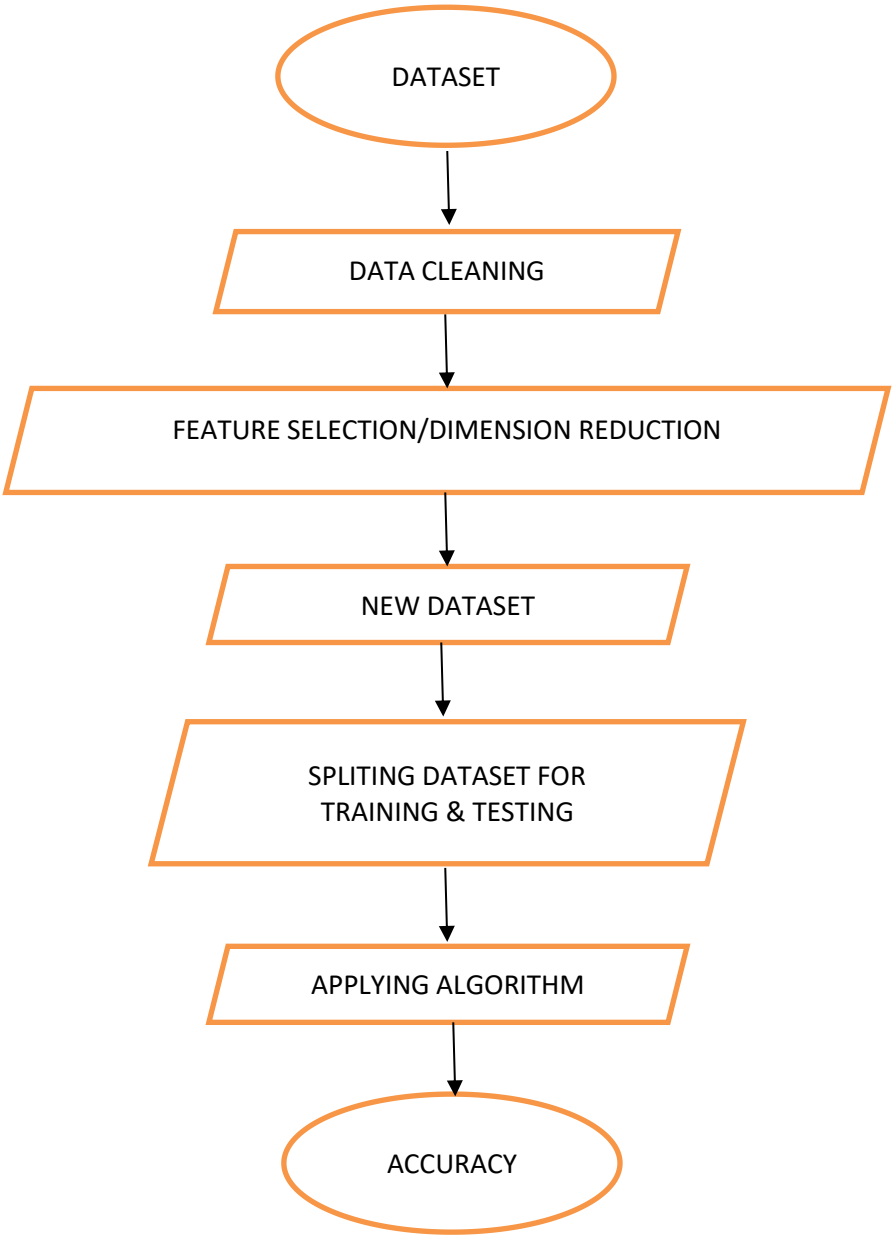
As fraudsters change fraud patterns quickly by assuming the consumers regular behavior. So, some fraudsters conduct frauds once using online channels and then transition to other ways, fraud detection systems must identify online transactions using unsupervised learning. This research aims to focus on fraud situations that cannot be identified using previous record or supervised learning, and develop a deep Auto-encoder and restricted Boltzmann machine (RBM) model that can recreate normal transactions and find anomalies from normal patterns. Backpropagation is used to set inputs equal to outputs by the proposed auto-encoder (AE) deep learning method which is an unsupervised learning algorithm [23].

This study [24], particularly focuses on machine learning and data mining methodologies, as well as the numerous datasets that have been studied for identifying financial fraud. To extract, synthesize, and narrate the results, they used the Kitchenham approach as a well-defined procedure. Major difficulties, gaps and limits in the field are discussed after choosing, compiling and evaluating 47 papers in this regard and also recommendations are given where further studies are needed. As supervised algorithms were used more frequently than unsupervised approaches like clustering, future research on fraud detection should focus on unsupervised, semi-supervised, bio-inspired, and evolutionary heuristic methods. Also, future research is expected to make use of textual and audio formed data.

In this paper [25], they proposed an improved Neural Arithmetic Logic Units which are a type of neural network architecture. These units are capable of implicitly modeling mathematical connections inherent in a neural network. They also create a synthetic benchmark dataset, which reflects the issue setting of automatically capturing such mathematical linkages within the data, inspired by a real-world credit payment application. For various network parameters, their new network design is assessed on two real-world and two synthetic financial fraud datasets. They evaluate their proposed model to a number of well-established classification methods and its results show that the proposed model is capable of improving the neural network's performance.

Chapter 4

Methodology



In our dataset we have 24 variables which is compromise of one dependent and 23 independents variables. At first, we start by cleaning up our dataset by removing or changing any data that is incorrect, incomplete, irrelevant, duplicated, or badly formatted. Then, to determine the most significant features, we used four feature selection techniques: XGBoost, Chi-Square, Gini Index, Univariate, and four dimension reduction techniques: PCA, Pearson Correlation Coefficient, SVD, and Entropy. As a result, it will assist us in obtaining a fresh and improved dataset. After using the above strategies, we construct a new dataset and divide it into two sections for training and testing. For training, we use 70% of the data and 30% for testing. After that, we used machine learning methods such as SVM (Support Vector Machine), RF (Random Forest), DT (Decision Tree), LR (Logistic Regression), NB (Nave Bayes), KNN (K-Nearest Neighbor), MLP (Multi-layer Perceptron), and CNN (Convolutional Neural Network).
We added the variables to the new dataset one by one in the order of importance from high to low for testing our algorithm accuracy in a variety of dimensions. As a result, we start with the most significant four variables, then add the fifth, and so on.

The results based on the relevance of variables provided by PCA, SVD, and other approaches utilized in this study show that as the number of variables increases, RF, LR, and SVM perform better and more consistently than the DT, KNN, NB, MLP, and CNN. We can see that the maximum accuracy is found on a different number of dimensions for different techniques. In the same way that PCA provides us the maximum accuracy for all techniques in the 16^{th} dimension. All approaches have the highest accuracy in the sixth and ninth dimensions. Finally, we put their results together to compare these methodologies and algorithms more intuitively, and we discover that RF outperforms them all.

Chapter 5

Experimental Results & Comparison

Table 1: Performance of algorithms with and without feature extractions

Algorithm	Accuracy (without feature selection)	PCA (16th Dimension)	XGBoost (11th Dimension)	Pearson correlation coefficient(20 Dimension)	Univariate (5th Dimension)	Chi-square (6th Dimension)	Entropy(9th Dimension)	Gini-Index(17th Dimension)	SVD(13th dimension)
SVM	81.84	77.71	81.91	78.23	**82.1**	78.22	81.2	81.95	78.22
Logistic Regression	80.87	77.67	**80.92**	78.23	80.85	78.22	80.12	80.83	78.23
Random Forest	81.37	81.61	80.92	81.41	**81.78**	74.55	81.41	81.27	81.41
Decision Tree	72.58	72.17	72.15	72.53	**81.96**	67.16	73.2	73.1	72.66
KNN	80.60	77.41	81.38	78.20	**81.96**	78.06	78.15	81.39	78.2
NB	66.84	76.67	**78.22**	76.80	**78.22**	**78.22**	77.85	68.82	76.8
MLP	81.2	77.51	81.67	77.97	**82.08**	78.10	78.15	81.63	77.96
CNN	80.10	80.31	80.43	80.42	**80.6**	71.65	80.37	79.72	80.48

This table shows the accuracy of different algorithms after applying dimension reduction and feature extraction methods. From the table it is quite clear that after applying feature selection and dimension reduction accuracy of the algorithms has increased. Best accuracy was shown after applying PCA in the 16th dimension. In Xgboost it was on the 11th dimension. In Pearson correlation coefficient it was on the 20th dimension. In univariate it showed best accuracy on the 5th dimension. In chi-square it showed best accuracy on the 6th dimension. In entropy it showed best accuracy on the 9th dimension.

In Gini-Index it showed best accuracy on the 17th dimension and SVD on the 13th dimension. Overall, after applying univariate feature selection all algorithms showed the best accuracy.

Table 2: Performance comparison of algorithms with different papers

Algorithm	**Our accuracy (Univariate)**	**SMOTE** [22]	**ADASYN** [22]	**WEKA** [21]	**SMOTE** [20]
SVM	**82.10%**	77.48%	77.33%	-	69.31%
KNN	**81.96%**	60.43%	57.98%	-	81.85%
Decision Tree	81.96%	68.10%	67.46%	80.3%	**86.53%**
Random forest	81.78%	76.98%	75.40%	-	**89.01%**
Logistic Regression	80.85%	-	-	**81%**	-
NB	**78.22%**	-	-	69.4%	-
MLP	**82.08%**	-	-	81.7%	-

The Univariate feature selection method is used in the comparison table-2 for seven machine learning algorithms (SVM, KNN, DT, RF, LR, NB, and MLP), with four algorithms (SVM, KNN, NB, and MLP) achieving the highest accuracy when compared to the other three related publications. In comparison to the p1 [21] and p2 [22] paper, our univariate feature selection technique showed good accuracy for two algorithms (DT and RF). However, from our univariate feature selection, p3 [20] has the highest accuracy and p2 has highest accuracy for LR algorithms from our univariate feature selection technique.

Chapter 6

Conclusion & Future Work

Credit card fraud has recently become a major concern for financial institutions all around the world. Various methods have been used in the past to detect fraudulent activities also to get at an accurate financial decision, a large number of researchers have researched into credit card fraud detection. For these purposes, data mining models have been widely used. However, the need to examine different dependable methods to detect fraudulent credit card transactions still continues. We used several feature selection and dimension reduction approaches in this study to determine the best dataset then we applied different machine learning algorithms from which we could get greater accuracy. In comparison to all prior study in this sector, SVM with Univariate in the fifth dimension attained the best accuracy of 82.1 %, which is one of the highest accuracy rate achieved. Though there have been certain studies that have been more accurate. For simplicity and better accuracy, we have used machine learning based approaches. We learned from this research that simple machine learning models combined with a dataset balancing strategy can produce excellent results without the use of more complex algorithms and methods. As a result, applying a simple strategy rather than a complex one can help people make more correct decisions in the Default of Credit Card Clients.

Due to lack of time, many different methods, testing, and experiments have been postponed. Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or simply curiosity. There are some ideas that I would have liked to try during the development of the Credit Card Fraud detection. This thesis has been mainly focused on the use of feature selection and dimension reduction techniques to find out the best result but there are also more techniques that can be used for acquiring highest accuracy. In the credit card domain, several techniques of feature selection and extraction should be examined to see how they affect prediction accuracy. Future research should focus on determining the most appropriate hybrid model among state-of-the-art machine learning algorithms to determine the most accurate hybridized model in the previously described area.

References

[1]The Editors of Encyclopedia Britannica, “Credit card,” *Encyclopædia Britannica*. Dec. 29, 2017. [Online]. Available: https://www.britannica.com/topic/credit-card

[2]“Who Invented Credit Cards?,” *Million Mile Secrets*, May 14, 2020. https://millionmilesecrets.com/guides/who-invented-credit-cards (accessed May 25, 2022).

[3]T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[4]Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, “A Chi-Square Statistics Based Feature Selection Method in Text Classification,” *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Nov. 2018, doi: 10.1109/icsess.2018.8663882.

[5]A. S. Manek, P. D. Shenoy, M. C. Mohan, and others, “Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier,” World wide web, vol. 20, no. 2, pp. 135–154, 2017.

[6]A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” in 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), 2015, pp. 1200–1205.

[7]I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016.

[8]P. Lin, J. Zhang, and R. An, “Data dimensionality reduction approach to improve feature selection performance using sparsified SVD,” in 2014 International Joint Conference on Neural Networks (IJCNN), 2014, pp. 1393–1400. doi: 10.1109/IJCNN.2014.6889366.

[9]A. S. Raut and K. R. Singh, “Feature Selection for Anomaly Based Intrusion Detection using Rough Set theory,” in Int. Conf. Industrial Automation and Computing (ICIAC), 2014, pp. 31–38.

[10]S. A. Mohammed and others, “Construction of a Prediction Model for Banking Loans Risk Using Data Mining Techniques,” phdthesis, Sudan University of Science, 2019.

[11]T. Evgeniou and M. Pontil, “Support Vector Machines: Theory and Applications,” *Machine Learning and Its Applications*, pp. 249–257, 2001, doi: 10.1007/3-540-44673-7_12.

[12]M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-6849-3.

[13]W. H. Delashmit and M. T. Manry, "Recent Developments in Multilayer Perceptron Neural Networks," 2005.

[14]G. Panchal, A. Ganatra, Y. Kosta, and D. Panchal, "Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers," International Journal of Computer Theory and Engineering, vol. 3, no. 2, pp. 332–337, 2011.

[15]L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," Journal of big Data, vol. 8, no. 1, pp. 1–74, 2021.

[16]A. Mehbodniya et al., "Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques," Security and Communication Networks, vol. 2021, 2021.

[17]I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," Journal of Big Data, vol. 8, no. 1, pp. 1–21, 2021.

[18]K. Maka, S. Pazhanirajan, and S. Mallapur, "Selection of most significant variables to detect fraud in financial statements," Materials Today: Proceedings, 2020.

[19]J. Yao, J. Zhang, and L. Wang, "A financial statement fraud detection model based on hybrid data mining methods," in 2018 international conference on artificial intelligence and big data (ICAIBD), 2018, pp. 57–61.

[20]A. Subasi and S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," in 2019 International Engineering Conference (IEC), 2019, pp. 115–120.

[21]M. Pasha, M. Fatima, A. M. Dogar, and F. Shahzad, "Performance comparison of data mining algorithms for the predictive accuracy of credit card defaulters," Int. J. Comput. Sci. Netw. Secur, vol. 17, no. 3, pp. 178–183, 2017.

[22]A. S. Shetty and R. Manoj, "Prediction of default credit card users using data mining techniques," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7, pp. 816–821, 2019.

[23]A. Pumsirirat and L. Yan, "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine," International Journal of advanced computer science and applications, vol. 9, no. 1, pp. 18–25, 2018

[24]M. N. Ashtiani and B. Raahemi, “Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review,” IEEE Access, 2021.

[25] D. Schlör, M. Ring, A. Krause, and A. Hotho, “Financial Fraud Detection with Improved Neural Arithmetic Logic Units,” in Workshop on Mining Data for Financial Applications, 2020, pp. 40–54.