FINANCIAL FRAUD DETECTION AND VALIDATION USING NEURAL NETWORKS

Dissertation proposal submitted in fulfillment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

TANYA SINGH

11309402

Supervisor

MS. HARLEEN KAUR

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

December 2017

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal entitled "FINANCIAL FRAUD DETECTION AND VALIDATION USING NEURAL NETWORKS" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms. Harleen Kaur. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Tanya Singh

11309402

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M. Tech dissertation proposal entitled "**FINANCIAL FRAUD DETECTION AND VALIDATION USING NEURAL NETWORKS"**, submitted by **Tanya Singh** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Harleen Kaur)

Date:

Counter Signed by:

1) **Concerned HOD:**

   HoD's Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   Internal Examiner

   Signature: _______________

   Name: __________________

   Date: ___________________

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the guidance to complete this dissertation proposal. I would like to express my special gratitude to my dissertation mentor, Ms. Harleen Kaur, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my dissertation, especially in writing this proposal.

Without the support from Lovely Professional University, it would not have been possible to complete this proposal. The faculties have always been very responsive in providing necessary information, and without their generous support, I would have lacked in accurate information on current developments.

I perceive this opportunity as a big milestone in my career development. I will strive to use the gained skills and knowledge in the best possible way and will continue to work on their improvement, to attain desired career objective.

Hope to continue cooperation with all of you in the future.

Tanya Singh

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Financial fraud is one of the wide-reaching glitches that have been observed by many academicians, auditors, and management marketing heads. The financial fraud problem has so many factors to consider, like, size of the dataset to be analysed, feature selection, the sensitivity of data, the accuracy of results, performance analysis of various techniques and many more. With many factors involved, the detection of fraud has become a cumbersome process to be done recurrently. Auditors have been exploited with the never-ending process of detection and prevention of fraud. The issue of financial fraud has been explored by many other researchers with the help of data mining. The use of data mining has recently proved the ease of detecting this financial fraud. Data mining is widely used to uncover the hidden patterns and forecasts future trends and behaviors in financial markets.

Though it has been proved through many researches that neural networks have the best performance among all other data mining techniques used for fraud detection, but this performance rate decreases after the implementation of validation theory. This gives birth to a new research question: What are the effects of implementation of different validation theories on neural networks?

This dissertation proposal is an attempt to study and report on the appropriateness of neural networks in the detection of financial fraud. The study is especially aimed at showing to the extent possible, relevance and reliability of neural network as a tool for controlling fraud.

CHAPTER 1
INTRODUCTION

According to the Oxford Dictionary [1], Fraud can be defined as “Unlawful or Felonious dishonesty practiced with the intention of financial or personal gain”. Well, in general, no set of defined words explain the definition of financial fraud. But, in simpler terms, it can be demarcated as “an unlawful act with intent to acquire unauthorized financial benefits”.

With the presence of many illegal performs, auditors have been overworked with the duty of uncovering fraud. The area of data mining has proficiencies to excerpt and uncover hidden patterns and truth behind the large chunks of data. It can also forecast many inclinations and behaviors in the financial market. Data mining can be divulged as “a procedure that uses multidisciplinary techniques like statistics and mathematical methods along with artificial intelligence and machine learning methods to abstract and extricate useful information and later gain knowledge from a large database” [13].

Data mining turns out to be an advanced tool for auditors that may assist them to detect fraud and for better decision-making processes. Data mining has been purposeful for many financial applications which includes loan assessment, bankruptcy prediction, fraud detection, real estate assessment, investment selection and much more.

1.1. Classification of Data Mining Techniques

Data mining is an enduring procedure within which the growth is spelled out by unearthing trends or patterns. It is the most beneficial technique in a tentative analysis scenario where there are no prearranged concepts about what will constitute a conclusion. The knowledge results obtained from data mining processes are used to assist human decision makers in efficient and effective decision making and finally to solve their complex problems.

Following listed are widely used Data Mining techniques.

1.1.1. Classification

Classification figures out a model to foresee the clear-cut labels of unidentified stuff to discriminate amongst objects of diverse classes. These definite labels are predefined, distinct and unordered [13]. This technique is used to assign objects to one of the pre-determined classes. It is a supervised learning method. The basic working of classification is described as:

- A collection of data which is identified as input data is used in the progression of the classification method. Each set of data contains its attribute set and a class label. The class label is a predetermined category. Input data is further divided into two sets – Train data and test data [6].
- Train data is randomly portioned data used to generate the classification model which is also acknowledged as a classifier. This is used to forecast the class of unidentified new record. Test data is the remaining portion of the dataset which is used to estimate the performance of the classification model.
- The various algorithms used under classification model are Naïve Bayes technique, decision tree, neural networks, Bayes Algorithm, Nearest neighbor and Support Vector Machine.

These algorithms are used in the exposure of credit card frauds, healthcare fraud, automobile insurance, corporate fraud and much more. Out of all applications of data mining, classification is the most learned models.

1.1.2. Clustering

Clustering is a data analyzing technique exercised to distribute the objects into some considerable bundle or clusters. These objects are identical to one another but are different from objects present in another cluster. This is the unsupervised approach of learning and is also famous as Partitioning or data segmentation method [4].

According to Yue et al. [16] , “Clustering can be defined as a method of decomposing or partitioning dataset into chunks so that the points in one group are similar to each other and are as alien as possible from the points in other groups.”.

Clustering techniques are commonly used to identify stable dependencies for many fields which include investment and risk management.

Common clustering algorithms are K-means algorithm, Hierarchical clustering and self-organizing maps [4]

1.1.3. Prediction

This data analysis methodology evaluates the numeric and well-ordered future values grounded on patterns of a dataset [17]. Prediction attributes are always continuous and not categorical.

Predictive analysis is a statistical method which takes data as input and extracts meaningful information from the dataset. The extracted information is then used to predict the behavioral trends and patterns. In simpler terms, Predictive analysis tends to learn from the dataset to predict the future trend for the sake of better decision making.

Neural Networks and regression are the generally used Prediction Algorithms. [4]

1.1.4. Regression

This method is used to disclose the relationship between independent variable and dependent variable. This technique aims to describe the relation between a dependent variable and one or more independent variable. It describes the change in the value of a dependent variable when any one of the independent variable is changed while others remain constant.

Regression technique follows basic mathematical statistical approach. Other mathematical methods used under regression are linear regression and logistic regression.

This is very beneficial for the exposure of credit card fraud, corporate frauds and many more [4]. The algorithm generally used is a logistic regression [5].

1.1.5. Visualization

Visualization is termed as the modest and easier way to epitomize data in an understandable data presentation. The input is multifaceted data and subsequent easy and clear uncovered patterns of complex data that can be effortlessly recited by the users through the progression of data mining.

This is closely related to the pattern detection capabilities of a human. This is done with the help of certain factors like size, vision and previous experiences.

The main aim of this technique is to communicate the meaningful information in a clear understandable format with the help of graphical means. This is termed as the best way to represent hidden trends and patterns from complex data [3] [4]

1.1.6. Outlier Detection

Outlier detection is the technique which aims to identify the objects which do not match the expected outcome or the predictable trend of the dataset. It is used to measure the dissimilarities between different or inconsistent data objects from the rest of the dataset. Different or inconsistent data that remain out of the data are called outliers. Their characteristics stand out from all the dataset available. Investigation of outliers is a basic function done through data mining.

This technique is also famous as anomaly detection and is applicable in many domains like fraud detection, intrusion detection etc. [18]

Commonly used algorithm under this approach is discounting learning algorithm.

1.2. Neural Networks

Artificial Neural Networks are nonlinear statistical computerized data demonstrating tools which are built to emulate the detection and pattern recognition capabilities of humans. Neural networks are the algorithms with the ability to recognize hidden patterns and to deduct the knowledge from those uncovered truths. The neural network is a method based on inductive learning and is inspired by the functionality of human brain which uses set of neurons for communication purposes.

Neural Networks consists of some fundamental elements that help in processing data. These elements are distributed in different layers. A multi-layered neural network consists of many neurons or units connected with each other. Different layers of neural networks consist of an input layer, an output layer and a hidden layer. There is also some weight associated with every connection.

The first layer of the neural network is the input layer and the last layer of the network is the output layer. As the name suggests, these layers serve as a source to intake input and exterminate an output. There may be more than one hidden layer between the input and output layer. Hidden layer is used as the model to process the input relations and present them as output. The number of layers depends on the type of neural network. For example, Back Propagation Neural Networks have one additional or more hidden layer while Self-Organizing Maps have only input and output layer.

Each neuron receives some signal from its connected neurons. Individual neurons take single or multiple inputs and produce an output which eventually becomes the non-linear transformation of input and corresponding connective weights. When an input signal reaches the hidden layer, the combined input signal is premeditated by multiplying the input signal with the connective weight.

The total input signal for neuron j is:

$$uj = \Sigma\ wij * xi, \qquad [1.1]$$

where

xi is the user specified input signal from neuron

i and wij are the connective weights between neuron i and neuron j. [14] [23]

This combined input strength is then compared with the threshold value. If the sum is greater than the threshold, then it is considered as output and is sent to the output layer. This output layer will further communicate this data with the help of some threshold function to another neuron as input. This is the process how neurons are fired. If the sum is not greater than the threshold values, then some changes or modifications can be made in the connective weights and threshold value using certain algorithms like backpropagation algorithm.

Neural networks are widely used as a technique for classification and clustering. It has many advantages over other data analyzing techniques. Neural networks are adaptive, and they generate robust models. One of the biggest advantages of using the Neural network is that one can modify the process of classification with the modification in the connective weights. They are frequently applied to find credit card fraud, corporate fraud, financial fraud and automobile insurance fraud

1.3. Performance of Neural Networks in comparison with other data mining techniques

Artificial Neural Networks can be best used as the armor against the financial fraud. Many researchers and academicians have explored the effectiveness of neural networks in the detection of financial fraud. Many comparisons have been made with other data mining algorithms like Bayesian Network, Bayes' Theorem, Decision tree and Logistic Regression. Many research papers have proved that the performance of Neural Network has always outperformed in comparison to others [5] [15] [24]

The performance of every technique includes some factors like Accuracy, Specificity, Sensitivity as shown in Fig. 1.1,1.2 and 1.3 [5] [24] respectively and many more. But according to researchers, accuracy attracts the major preference while selecting the best model for fraud detection. According to analysts, the technique should be chosen which has fewer misclassification and which consumes less time. Even if all the other factors are equal, less hectic and easy to implement technique should be given preference over others.

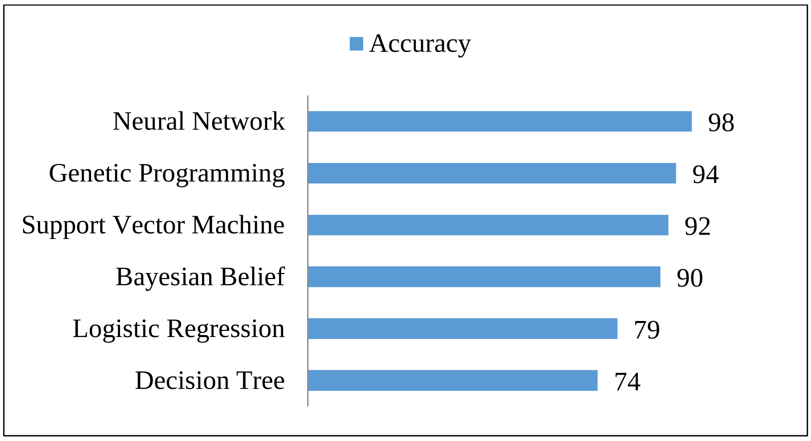


Figure 1.1 Accuracy Analysis of different data mining techniques

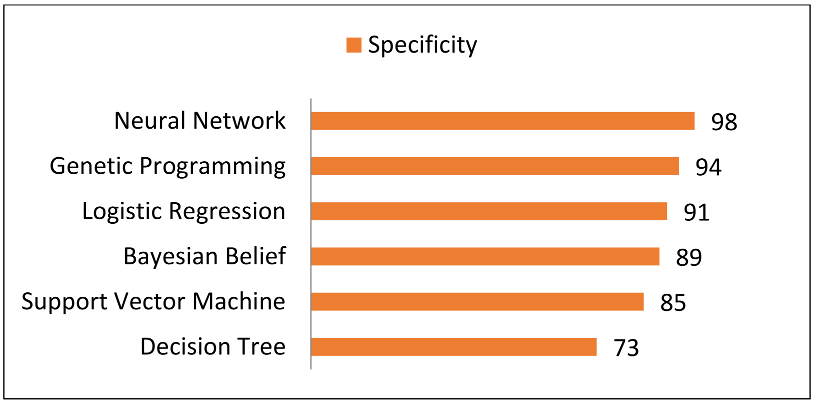


Figure 1.2 Specificity Analysis of different Data Mining Techniques

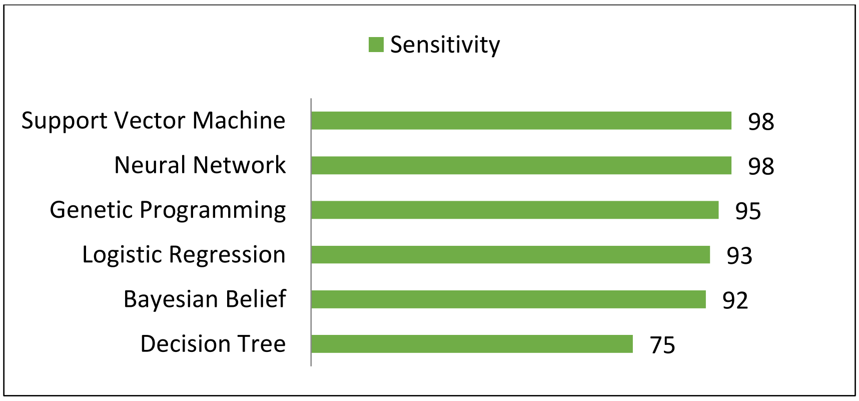


Figure 1.3 Sensitivity Analysis of different Data Mining Techniques

All different parameter analyzers of various data mining techniques have been compared and shown in Figure 1.4. This Figure clearly shows that Neural Network outshines in all the three factors. This is also one prove which says that neural network has outperformed in contrast with other techniques.

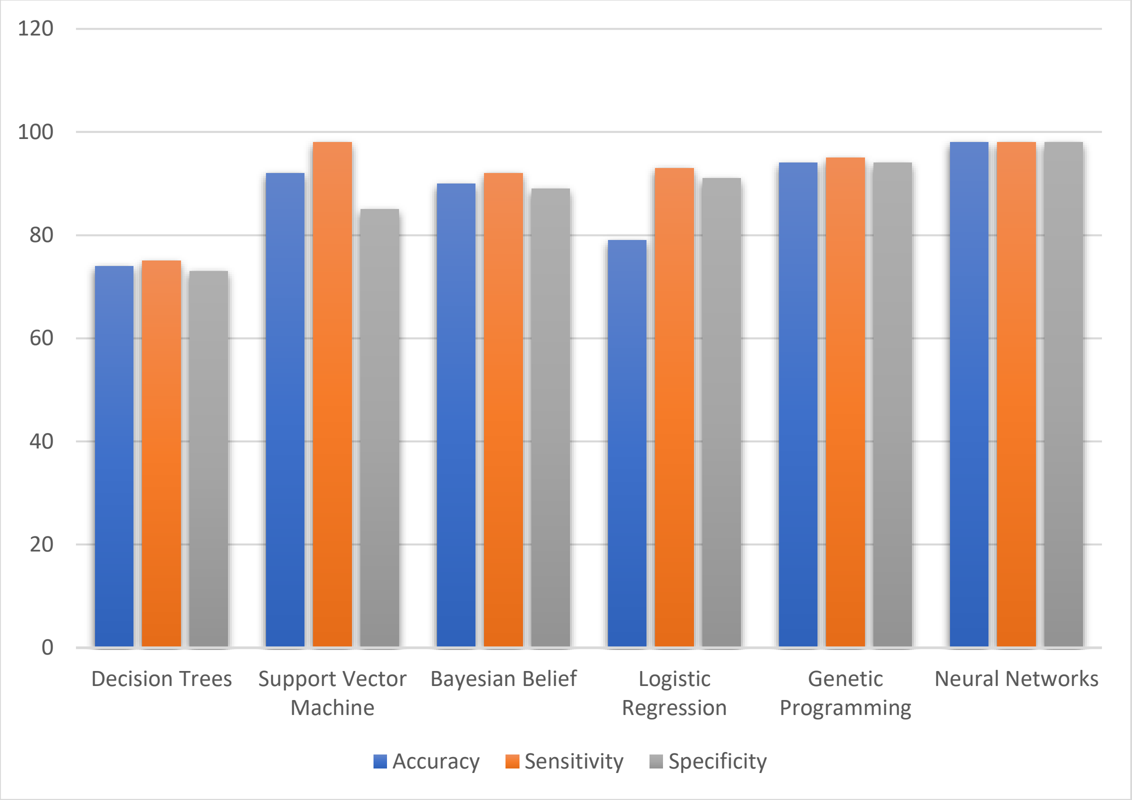


Figure 1.4 Comparison of parameter analyzers of different Data Mining Techniques

The outshining performance analysis of Neural networks bring it to the limelight. The implementation of the Neural network can be done through many software which are readily available in the market as open source. The technology market offers various open source libraries such as TensorFlow for the computation of mathematical numerical operations with the help of data flow graphs. Even neural networks have many algorithms and architectures like Generalized Adaptive Neural Network Algorithm [24] [25] which can be used for the detection of financial fraud. Every algorithm owns a different accuracy rate.

1.4. Validation of Neural Networks

The input dataset is divided into two sets of data namely, Train data and Test data. Training dataset is randomly allocated data from the input dataset which is used to predict the class of unknown new record. Test data is the remaining portion of the dataset which is used to evaluate the performance of the model. This performance rate may have some bias due to inevitable factors. To eliminate this bias, the performance of the model is compared with the previously found patterns. This is known as model validation.

Model validation is the process which estimates the correct and exact performance rate of any model without any bias. There are many approaches for model validation. One such technique for validation is cross-fold validation.

Cross fold validation is a method which is used to analyze how the test data will react to a generalized independent form of data. The goal of this technique is to test model while it is in training phase. There are two types of cross-validation which are stated below in brief:

1.4.1. Exhaustive Cross-Validation

This cross-validation method learns from the original sample and then it takes trials on all thinkable ways to split the sample into two sets i.e. training and a validation set. [19] [22].

They are further divided into the following:

1.4.1.1. Leave-p-out Cross-Validation

This method picks up ‘p’ observations from the sample as the validation set and the residual dataset is used as the training set. The process is repetitive on all the possible combinations.

In case of occurrence of an error, the error is averaged to all the combinations. This is how overall effectiveness is computed. LpO cross-validation involves training and validation of the model nCp times, where n is the number of observations in the original sample.

For moderately large p and for the large size of n, LpO cross-validation can become impracticable.

1.4.1.2. Leave-one-out Cross-Validation

It is a special case of Leave-p-out cross validation with p = 1. This method is highly preferable over Leave p-out cross-validation as it does not own the problematic unnecessary computation time because of $^nC_1 = n$.

1.4.2. Non-Exhaustive Cross-Validation

Non-exhaustive cross-validation methods are exactly opposite of exhaustive cross-validation methods. It does not figure out all thinkable ways to split the original sample.

They are further divided into following:

1.4.2.1. K-cross fold Cross-Validation

In this technique, sample dataset is first divided into k folds. Each fold is trained by the remaining folds of data. These folds are then tested by one holdout fold. This is how the average performance rate of the algorithm is evaluated. When k equals n i.e. the number of observations, this is exactly same as the leave-one-out cross-validation [3] [8] [9].

1.4.2.2. Holdout method

In this method, two sets d0 and d1 are randomly assigned with the data, namely the training set and the test set respectively. Some portion of training dataset is used to predict from the model which is previously trained on the rest of the data.

Error estimation becomes a base to judge the model. The size of each of the sets is unpredictable, but in general test set is smaller than the training set. After this division, training is performed on d0 and testing is done on d1. [19]

1.4.2.3. Repeated random sub-sampling validation

This method is widely famous as Monte Carlo Cross-validation method. In this technique, original dataset is casually split into two sets i.e. training and validation data. The data is then trained on the training dataset and then prognostic accuracy or mean square error is calculated using the validation data. After the calculation of prognostic accuracy, all the results are averaged over all the splits. This is not preferred method as the error is dependent on the division of the dataset.

Studies by many researchers have shown that there is a difference between accuracy rates in validation dataset and accuracy rates for training data. The implementation of validation theories brings the change in the performance of the data mining technique [3].

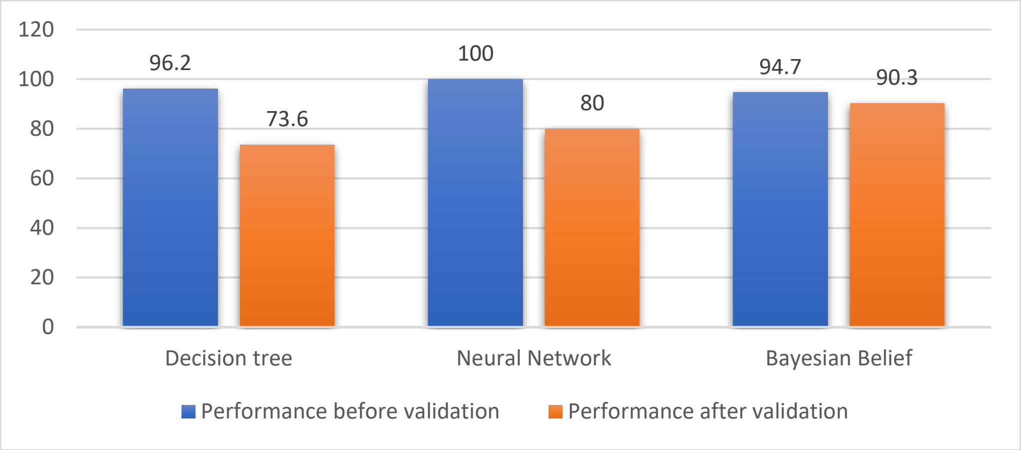


Figure 1.5 Performance Analysis of different techniques

The graph in Fig. 1.5 [3] clearly shows that the performance of data mining models normally degrades after the validation methods. Here 10 cross fold approach is applied to the data set.

CHAPTER 2

REVIEW OF LITERATURE

D.Zhang, L. Zhou, Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Man, and Cybernetics 34 (4) (2004) Nov.

This paper describes different data mining techniques in the context of financial application from both technical and application perspectives. In addition, a comparison is made of different data mining techniques and important data mining issues involved in specific financial applications. The paper also has its focus on existing data mining applications in the field of finance. The author describes various challenges associated with data mining techniques that reside in order to achieve effective financial management. With the help of this paper author has made clear distinctions of different types of frauds. The paper also defines the different data mining techniques. The author has also defined that Neural Networks outshines its performance among all other data mining techniques.

Kirkos, Efstathios & Spathis, Charalambos & Manolopoulos, Yannis. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32(4), 995-1003. Expert Systems with Applications. 32. 995-1003. 10.1016/j.eswa.2006.02.016.

The main intention of the paper is to explore the numerous data mining classification techniques in the detection of financial fraud. It also investigates the factors associated with financial fraud statement. The author compares the use of three different techniques namely, Decision Trees, Neural Networks and Bayesian Belief Networks for the identification of financial fraud detection. All the three models are then compared based on their performance. The author has also validated the model with the help of cross-validation. This paper also compared the performance of the model after the implementation of validation theories. This paper shows that the performance of the data mining models usually degrades after the implementation of validation theories.

Ngai, E.W.T., Hu, Y. H., Chen, Y., & Sun X. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support System (2010),

This paper presents a review of classification structure of data mining techniques for the detection of financial fraud. The author also provides a systematic academic literature review of different data mining techniques that are used for financial fraud detection. Articles published with reference to financial fraud detection in between 1997and 2008 are deeply studied by the author and beautifully summarized in the paper. This paper also discourses the gaps between FFD and the needs of the commerce to inspire additional research on neglected topics and concludes with several suggestions for further Financial Fraud Detection research.

Ravisankar P., Ravi V., Rao G.R. & Bose I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision Support System, 50, 491-500

This is a relative study of the performance of different analyzing algorithms of data mining used for the fraud detection like Multilayer Feed Forward Neural Network, Support Vector Machines, Group Handling, Genetic Programming and some others. The author has deeply explained the application of above-mentioned techniques to predict the financial fraud in companies. With the help of dataset of 202 Chinese companies, author experimented with all the techniques and hence has compared all the results. Author has also validated the results with the help of cross-validation and has described the results based on different parameters such as accuracy, sensitivity, specificity etc. The author also concludes that accuracy is the most attracted performance analyzer of any data mining technique and has recommended that technique with fewer misclassifications should be chosen for fraud detection.

C.-C. Lin, A.-A. Chiu, S. Y. Huang and D. C. Yen, Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments, Knowledge-Based Systems, 2015.

The objective of this paper is to examine all aspects of fraud triangle using the data mining techniques based on the previous academic literature available in the field. The author has also discussed the role of experts in defining the results of these techniques. In specific, the author has used both expert questionnaires and data mining techniques to sort out the different fraud factors and then rank the importance of them. The data mining methods used by the author in this paper are Logistic Regression, Decision Trees, and Artificial Neural Networks. In the end, author also explains the differences between different data mining tools and expert judgments.

Ubon Thongsatapornwatana, A Survey of Data Mining Techniques for Analyzing Crime Pattern. IEEE in Defence Technology (ACDT), 2016 Second Asian Conference

This paper reviews the literature on various data mining applications applied to solve the crimes. In addition to that, this paper provides insight about the data mining for finding the patterns and trends in crime to be used appropriately and to be a help for beginners in the research of crime data mining. The author has beautifully described various data mining techniques along with its pros and cons. The author describes the application areas associated with these data mining techniques. The paper also states the research gaps and challenges present in the data mining with respect to crime.

J. West, M. Bhattacharya, "Some Experimental Issues in Financial Fraud Mining", Proceedings of ICCS, 2016.

In this paper, the author has analyzed some of the relevant experimental issues of fraud detection with the focus on credit card fraud. According to the author, previous researchers have not made enough observations with respect to financial fraud detection. In general, many observations have been explored but not in context to financial fraud detection. Author has also investigated credit card fraud with the help of different performance metrics, controlled simulation, and concentration on detection algorithms.

West, Jarrod & Bhattacharya, Maumita & Islam, Md Rafiqul. (2015). Intelligent Financial Fraud Detection Practices: An Investigation. 10.1007/978-3-319-23802-9_16.

With the help of this paper, author has presented a comprehensive literature review on different types of fraud. Authors have clearly defined the difference between each fraud. The authors have also defined the data mining technologies used for each fraud detection. The paper defines that credit card fraud is one the most exploited fraud. This fraud is treacherous and can lead to disastrous results if not detected on time. The paper also distinguishes between different data mining techniques on the basis of their performance metrics. The author also states that neural network outshines other techniques in terms of accuracy and performance of model.

Mohammad Sultan Mahmud; Phayung Meesad ; Sunantha Sodsee. An evaluation of computational intelligence in credit card fraud detection. Computer Science and Engineering Conference (ICSEC), IEEE. 2016

The author of this paper has analyzed and compared various popular classifier algorithms which are widely used in the detection of credit card fraud. The author aims to rank the algorithms according to their performance. The author has a clear vision of providing an au fait evaluation of different tactics of classification, compare their performances on a wide series of thought-provoking credit transaction dataset, and attract assumptions on their applicability to credit card fraud exposure applications.

S. Kotsiantis, E. Koumanakos, D. Tzelepis & V. Tampakas, Forecasting Fraudulent Financial Statements using data mining, International Journal of Computational Intelligence 3(2) (2006) 104-110

With the help of this paper, the author has explored the effectiveness of machine learning algorithms and different techniques that can be used for detection of fraud. Author has also tried to identify factors associated with financial fraud. To sum up, the author indicates that the examination of financial evidence can be used in the documentation of FFS and emphasize the importance of financial ratios. In this paper different machine learning algorithms have been compared that can be used in the detection of fraud. Author has a clear emphasis on different financial variables and their effect in determining fraud detection.

Raghavendra Patidar, Lokesh Sharma. Credit Card Fraud Detection Using Neural Network. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011

In this paper, the author has tried to detect fraudulent transaction with the help of two famous data mining approaches- the neural network along with the genetic algorithm. According to author even though artificial neural network can work as a human brain when trained properly but still it is impossible for the artificial neural network to completely imitate the human brain. The author has made use of the Genetic algorithm for the better decision making about the features of a neural network like network topology, number of hidden layers, number of nodes etc. Author has used Supervised learning feed forward backpropagation algorithm to train the neural network. The author also describes the working of the neural network as well as problems faced during the training of neural networks. Author has made use of both the algorithms considering the fact that a talented man can do exclusively well when trained properly. With this thought in mind, the author has combined both the algorithms to find better results.

Anuj Sharma, Prabin Kumar Panigrahi. "A review of financial Accounting Fraud Detection based on Data Mining Techniques." International Journal of Computer Applications. 2012.

This paper presents a wide-ranging review of the literature on the application of data mining techniques for the recognition of financial accounting fraud and suggests a framework for data mining techniques based accounting fraud detection. This paper is a presentation of a comprehensive literature review of academic literature on the application of different data mining techniques which are widely used for the detection of financial fraud. This paper is an amalgamation of all the academic literature published between 1992 and 2011 for the detection of fraud. The paper is definitely a foundation stone to further researches in this area.

Kapardis, M. K., Christodoulou, C. & Agathocleous, M. (2010). Neural networks: the panacea in fraud detection? Managerial Auditing Journal, 25, 659-678

The main purpose that this paper is to examine the use of Artificial Neural Networks in the field of financial fraud detection. The author has developed a questionnaire which is filled by several auditors. This questionnaire was further used to develop Artificial Neural Network based on the questionnaire filled by the auditors. The author has in detail explanation about ANN in the paper. The results are validated with the help of leave-p-out cross-validation. Results are discussed after the implementation of validation theory. The author has obtained that ANN has high accuracy in the prediction of financial fraud. The author also states that if we use the same parameters used by auditors with the ANNs then ANN will be able to predict 95% accurate prediction.

Liou, F. M. (2008). Fraudulent financial reporting detection and business failure prediction models: a comparison. Managerial Auditing Journal Vol. 23 No. 7, pp. 650-662.

The purpose is to explore the differences and similarities between fraudulent financial reporting detection and business failure prediction (BFP) models, especially in terms of which explanatory variables and methodologies are most effective. In this paper, the author has used famous algorithms like Neural Networks, Logistic regression, and Classification trees to build the model for the prediction and detection of financial fraud. The author concludes that many variables are good at both detections of financial fraud and also a prediction of business failures. The paper also concludes that Logistic regression has outshined the field of prediction of business failures. The author also describes the importance of financial factors while predicting the business failures. During the comparison of all the three techniques, the author describes the working of all the methods along with its pros and cons. Author has also described the application areas of different methods.

Yue, X. Wu, Y. Wang, Y. Li, C. Chu, A review of data mining based financial fraud detection research, international conference on wireless communications Sep, Networking and Mobile Computing (2007), 55195522

This paper provides answers to basic questions about the occurrence the detection of financial fraud. It also has literature view on effective of algorithms in terms of fraud detection. In this paper, the author is concerned about the financial losses which happen due to Financial fraud statements. The author tends to answer few questions related to financial fraud statements. The author has answered questions like which algorithm should be used for the detection of fraud, the data features used to predict fraud and many questions like this. This paper includes a comprehensive academic literature review of techniques used during the detection and prevention of the financial fraud.

Ahmed, S.R. (2004). Applications of data mining in the retail business, International Conference on Information Technology: Coding and Computing 2 (2) (2004) 455 – 459

The author of this paper brings the light on the application of data mining techniques in the field of retail business. The paper also includes the detailed study of pros and cons of these techniques. This paper is the summary of all widely used data mining techniques.

Yamanishi, K., Takeuchi, J., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Mining and Knowledge Discovery 8 (3) (2004) 275–300

This paper briefly describes the effectiveness of outlier detection. The paper is focused on the Smart Sifter which is one of the effective outlier detection engines. The paper demonstrates the practical effectiveness of Smart Sifter using a real-world dataset. The paper has complete detail information about outlier detection and all the equations associated with the technique. The author has described Smart Sifter inside out with all the pros and cons. Author has in detail information about outlier detection. The paper concludes that use of SS can be used in various fields like fraud detection, trend detection etc

Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." Statistics survey 4 (2010): 40-79.

With the help of this paper, the author has an objective of presenting a survey on cross-validation. This paper describes cross-validation as a widely used technology because of its simplicity. The author has given guidelines to choose the best cross-validation technique according to the in-hand problem. The author has described the procedure of model selection and all the procedures used in model selection. This paper contains all the information about cross-validation and its types. The author has quoted cross-validation in detail along with the traces of history and some real-world examples. The paper has also indulged the different properties of cross-validation estimators. It also takes into account the statistical properties of cross-validation. This paper is a milestone for the beginners in the field of cross-validation.

Pandey, Yamini. "Credit Card Fraud Detection using Deep Learning." *International Journal of Advanced Research in Computer Science* 8, no. 5 (2017).

The author of the paper aims to describe deep learning algorithm and their applicability in the field of credit card fraud detection. Author has a deep learning package called H2O which is a framework used with large datasets and is widely used to evaluate deep leaning. Author has beautifully explained deep learning and its architecture along with its working. The author concludes the paper with a proof that deep learning algorithms show better accuracy when compared with normal detection algorithms. A conclusion can be made that deep learning model gives less error and hence high accuracy.

Lu, Yifei. "Deep neural networks and fraud detection." (2017).

The author has briefly explained about deep neural networks along with its architecture. The purpose of the author is to evaluate credit card fraud using deep neural networks and for the implementation, author has used two open source widely used libraries called TensorFlow and Scikit-learn. The author has also explained the implementation of deep neural networks with the help of TensorFlow. Author has also explained about TensorFlow and its features. The experiments carried out by author explained that we cannot improve the performance of the model by increasing hidden layers. Furthermore, the author concluded that the recursive sampling of imbalanced training set can upsurge the performance of the network on the test set.

Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. (1995)

This paper brings up the comparison between the most common estimation methods – Cross-validation and Bootstrap. In this paper, the author has made few changes in the parameter of cross-validation as well as bootstrap. For cross-validation, a change is made in the number of folds and whether the folds are stratified or not. A number of samples have been changed in bootstrap to find the effect of different parameters. The author summarizes his research work with the conclusion that stratified ten-fold cross-validation is recommended for model selection.

Sohl, J. E., & Venkatachalam, A. R. (1995). A neural network approach to forecasting model selection. Information and Management, 29(6), 297303.

In this paper, the author aims to test the feasibility of neural network for the process of model selection. Backpropagation neural network has been used to fulfill the objective. The paper has detailed information about the neural network and its working. The utility of neural network has been explored in the selection of forecasting models in great details. This paper is a proof that neural network can have a great level of accuracy even with small training dataset.

Fanning, K., Cogger, K. O. and Srivastava, R. (1995). Detection of management fraud: A neural network approach. International Journal of Intelligent Systems in Accounting, Finance, and Management 4 113–126

This paper includes the development of a fruitful discriminator of management fraud using two famous approaches -the generalized adaptive neural network architectures (GANNA) and the Adaptive Logic Network (ALN) approaches to designing neural networks. The discriminant functions can easily identify fraudulent and non-fraudulent companies with superior accuracy. This paper adds up one more effective technique of using Artificial Neural Network for the detection of management fraud. The new technique defined using ANN is less time consuming as compared to previous techniques.

Fanning, K., Cogger, K. O. and Srivastava, R. (1995). Neural Network Detection of management fraud using Published Financial Data. International Journal of Intelligent Systems in Accounting, Finance, and Management, 7(1), 21-24

This paper makes use of Artificial Neural Network as a model to detect fraud. The author uses ANN along with the standard statistical tools to investigate the usefulness of the new model. The paper has detailed knowledge about Artificial Neural Network, its parameters and working of ANN. At the end of the research, the author concludes that ANN is superior as compared to the standard statistical method when it comes to detection of financial fraud. The paper proves that addition of ANN to the analytical procedures is beneficial for the detection of financial fraud.

CHAPTER 3

PROBLEM DEFINITION

After the deep analysis of all the research work done in the field of fraud detection, credit card fraud comes out to be most perilous fraud. This type of fraud can lead to destructive results if not spotted on time. The traditional approach of detecting fraud is auditing. But Auditing has become more burdened process due to many financial fraud cases. The world is facing the problem and searching for the suitable options so that the fraud cases can be reduced.

Data Mining is a boon in this field of research. With its capabilities of detecting the uncovered hidden pattern from a large data set, it can also learn from the pattern and hence predict things beforehand. There are so many approaches and techniques of data mining which are beneficial for successful detection of frauds. The methods that can be used are Neural Network, Bayesian Belief, Support Vector Machine and many more algorithmic approaches.

Extensive academic literature review claims that application of Neural Network in the field of credit card fraud detection has blossomed the field with its high accuracy and performance rates. Though Neural networks have the best performance among all other data mining techniques used for fraud detection, this performance rate decreases after the implementation of validation theories.

This leads to one major research question: What is the effect of different validation theories on the result of Neural networks?

Efstathios Kirkos, C. Spathis and Yannis M. has explained that application of validation technique on neural network causes degradation in the accuracy rate of the model. This brings out another research gap: What will happen if bunch of validation theories are applied on same model simultaneously?

If one wants to club together different validation techniques and implement them on the same model, this can be done in three different ways. The first way is to implement different validation techniques like K cross-validation and holdout method sequentially and then average out the predicted accuracy by root mean square of results of each iteration.

Another method to implement the hybrid approach of validation techniques is to use one validation technique like k cross validation, recursively while resizing the dataset on each iteration. The accuracy can also be averaged out through root mean square method.

The above method can be implemented with some set of chosen validation approaches. This will be like heterogeneously implementation of above method.

This implementation of a combination of validation theories will let us know the behavior of Neural Networks towards the validation theories. This is also possible that these hybrid approaches of validation methods may have some positive effect on the result as compared to single validation method.

CHAPTER 4

SCOPE OF STUDY

The proposed research dives into the new world of exploration by providing a feasible, accurate and efficient method of detecting financial fraud. The use of data mining technologies makes the whole cumbersome process an uncomplicated effortless attempt.

The research carries the vision to utilize neural network and formulate a new validation theory that will contribute to the mine of knowledge in the field and further aid researchers in their explorative studies. The trained neural network works for prior detection of financial frauds, thereby, aiding in prevention of the same.

Though neural networks have outperformed the other technologies in fraud detection yet performance efficiency of neural networks is known to be degraded post application of validation theories. Hence, the research proposes a new validation theory that will have either no or negligible impact on the performance of the neural network. This will ensure authenticity and reliability of the results.

CHAPTER 5

OBJECTIVES OF STUDY

5.1 To train a neural network for the detection of financial fraud and implement all the validation approaches.

5.2 To validate the data using various validation techniques in a sequential order to obtain a more comprehensive measure of the accuracy.

5.3 To validate the data using various validation techniques in a recursive order homogeneously to obtain a more extensive measure of the accuracy.

5.4 To validate the data using various validation techniques in a recursive order heterogeneously to obtain a more comprehensive measure of the accuracy.

CHAPTER 6

PROPOSED RESEARCH METHODOLOGY

6.1 Sources of Data

The dataset for credit card fraud detection is the transactional data produced by European credit card holders in September 2013. The dataset is a set of transactions occurred in two days, where dataset contains 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. The dataset has been poised and investigated during a research alliance of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

6.2 Techniques used

For the training of Neural Network, I have made use of open source library called TensorFlow. TensorFlow is a symbolic mathematical library which is used for algorithmic approaches like Neural Network. It is an open source library which is widely used for data flow programming. TensorFlow allows us to perform a complex operation with great ease and efficiency. It was developed and maintained by Google for fast numerical computing. It was released under the Apache 2.0 open source license. The API is nominally for the Python programming language, although there is access to the underlying C++ API.

The initial comparison of all validation theories is done in R language with all the predefined libraries like Scikit-learn. The implementation of validation theory is being done in python. The new formulated technique will be written in python language.

6.3 Tools Used

All the python programming will be done through a tool called PyCharm 2017.2.3. PyCharm is an Integrated Development Environment which is used in computer programming. This tool is specifically used for the Python language. It is developed by the Czech company JetBrains.

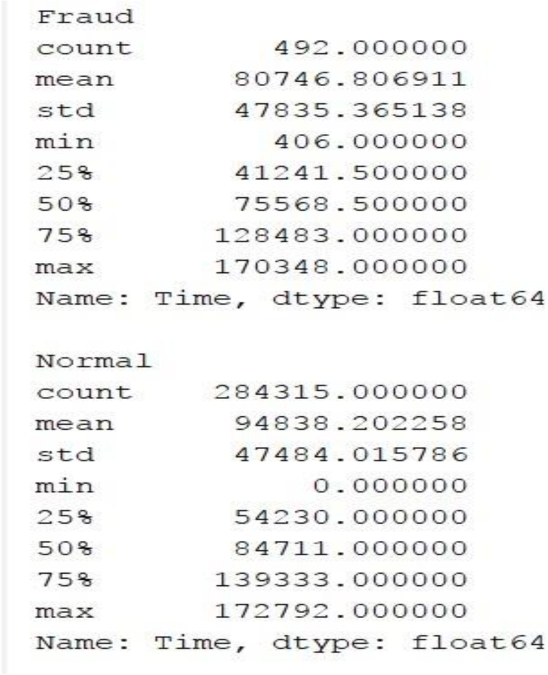
CHAPTER 7

EXPECTED OUTCOME

Objective 1: To train a neural network for the detection of financial fraud and implement all the validation approaches.

The above objective is successfully achieved. The Neural network is trained in python with the help of TensorFlow libraries and is written in a tool called PyCharm. All the visualization of analysis is done with the help of a library called t-SNE.

The actual dataset used contains 2,84,315 transactions. Out of all the transactions 492 transactions are fraud transactions as predicted by the model shown in figure 7.1.



```
Fraud
count         492.000000
mean        80746.806911
std         47835.365138
min           406.000000
25%         41241.500000
50%         75568.500000
75%        128483.000000
max        170348.000000
Name: Time, dtype: float64

Normal
count      284315.000000
mean        94838.202258
std         47484.015786
min             0.000000
25%         54230.000000
50%         84711.000000
75%        139333.000000
max        172792.000000
Name: Time, dtype: float64
```

Figure 7.1 Statistics of the dataset

The model here uses only 10,000 transactions because of limitations of hardware resources. With this portion of data, this predictive model has achieved accuracy of about 90%. The accuracy and cost analysis of trained neural network is shown in figure 7.2

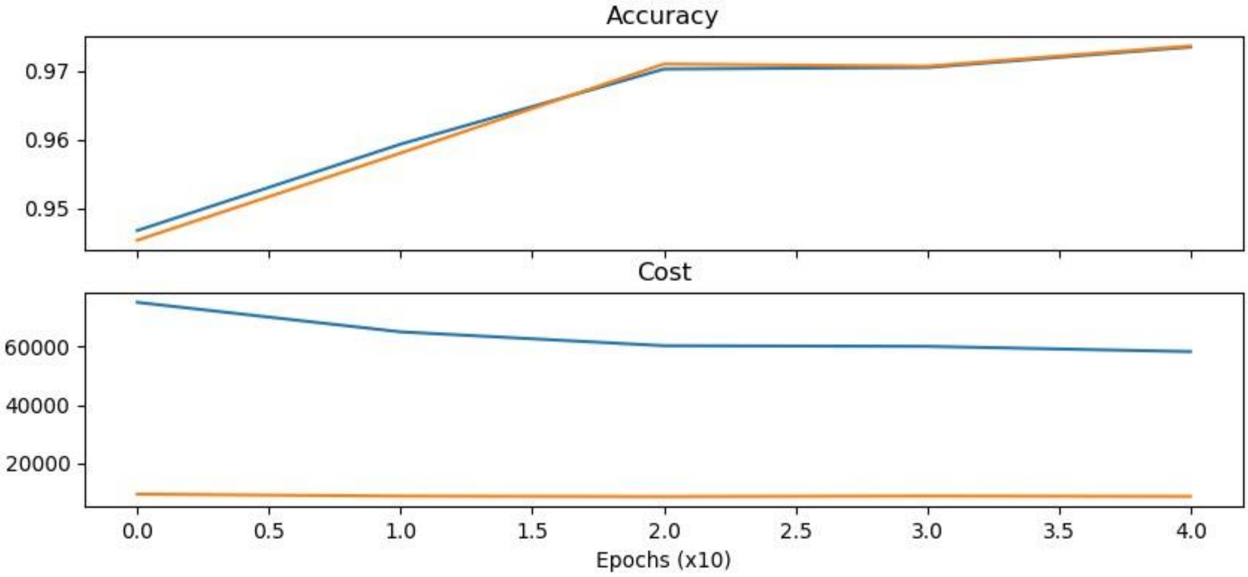


Figure 7.2 Accuracy and Cost Analysis of trained Neural Network

The result of fraud and non-fraud transactions are shown in figure 7.3. This visualization is crested with the help of t-SNE library where Blue color represents fraud transactions.

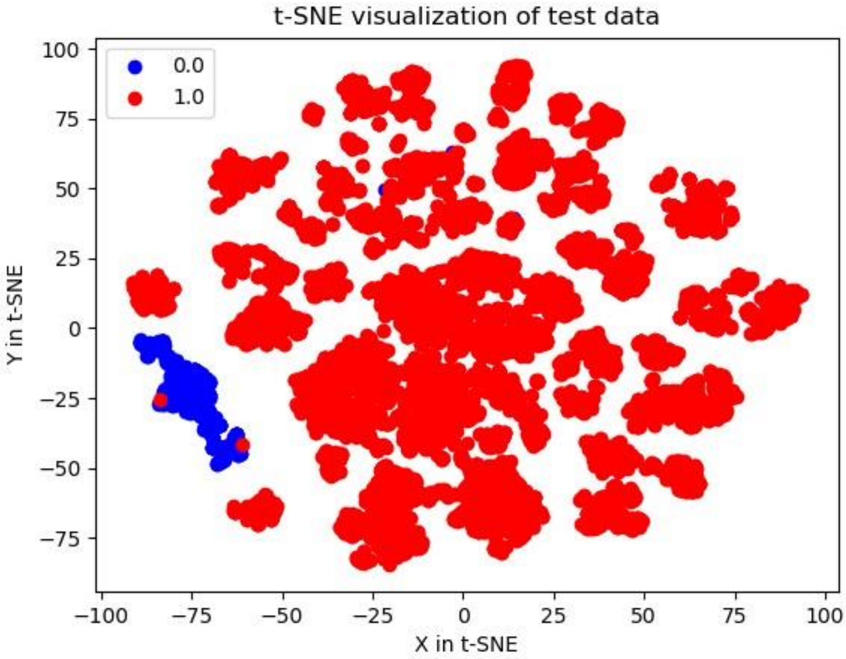


Figure 7.3 Visualization of Neural Network

Objective 2: To validate the data using various validation techniques in a recursive order homogeneously to obtain a more extensive measure of the accuracy.

The sequential order ensures that one technique doesn't influence the effect of the other. Each validation technique is applied one after the another and the obtained results are averaged out. As it is known through researches that validation is a statistical method. Accuracy is calculated with the help of average of iterations.

This is expected that this newly devised validation technology based on the serial application of different validation theories will affect the result of neural networks in positive essence and will have better accuracy than the previous validation theories.

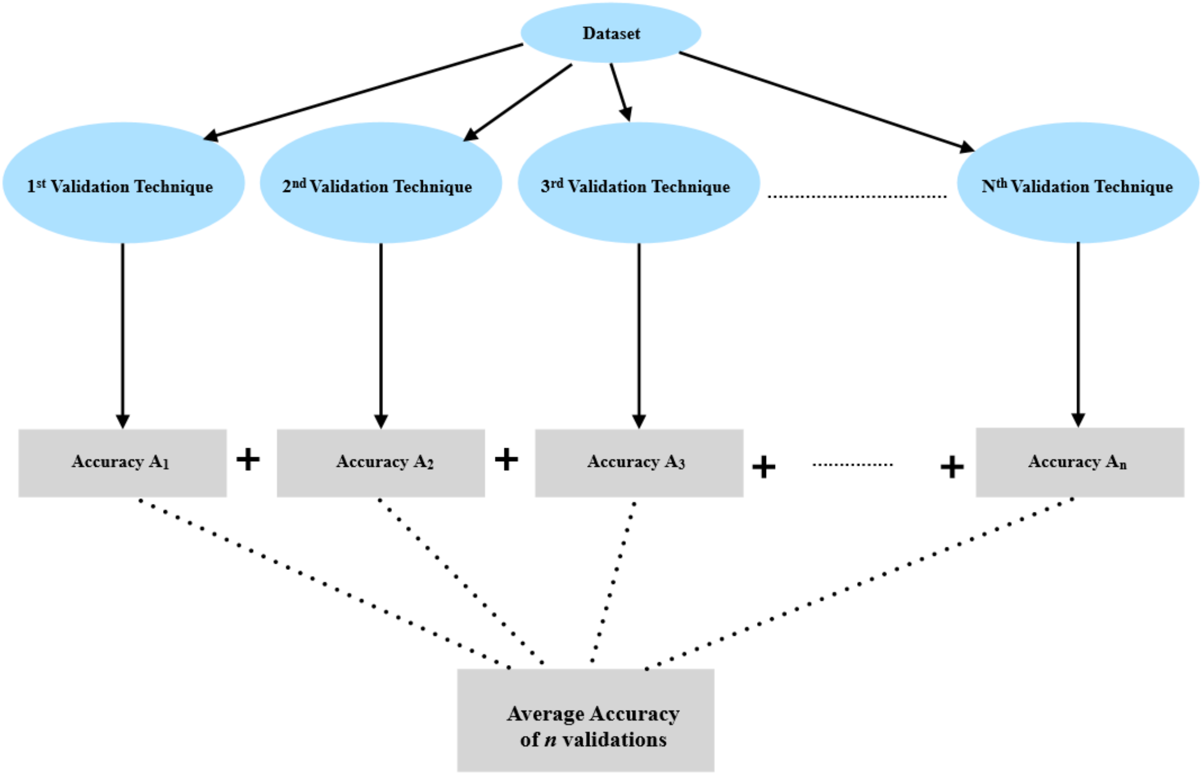


Figure 7.4 Draft technique for Objective 2

Objective 3: To validate the data using various validation techniques in a recursive order homogeneously to obtain a more extensive measure of the accuracy.

One validation technique is chosen, and it is uniformly applied on each individual iteration. Instead of applying validation on the data set as a whole, this validation technique branches out each iteration as a new set of data to be validated. In simpler terms, the output of the first step in an iteration becomes the input of the next step. So, each iteration is extensively validated, and the results are averaged out.

Figure 7.5 explains the rough plan to be executed in the fulfillment of the third objective. The new technique is based on the recursive application of validation theories.

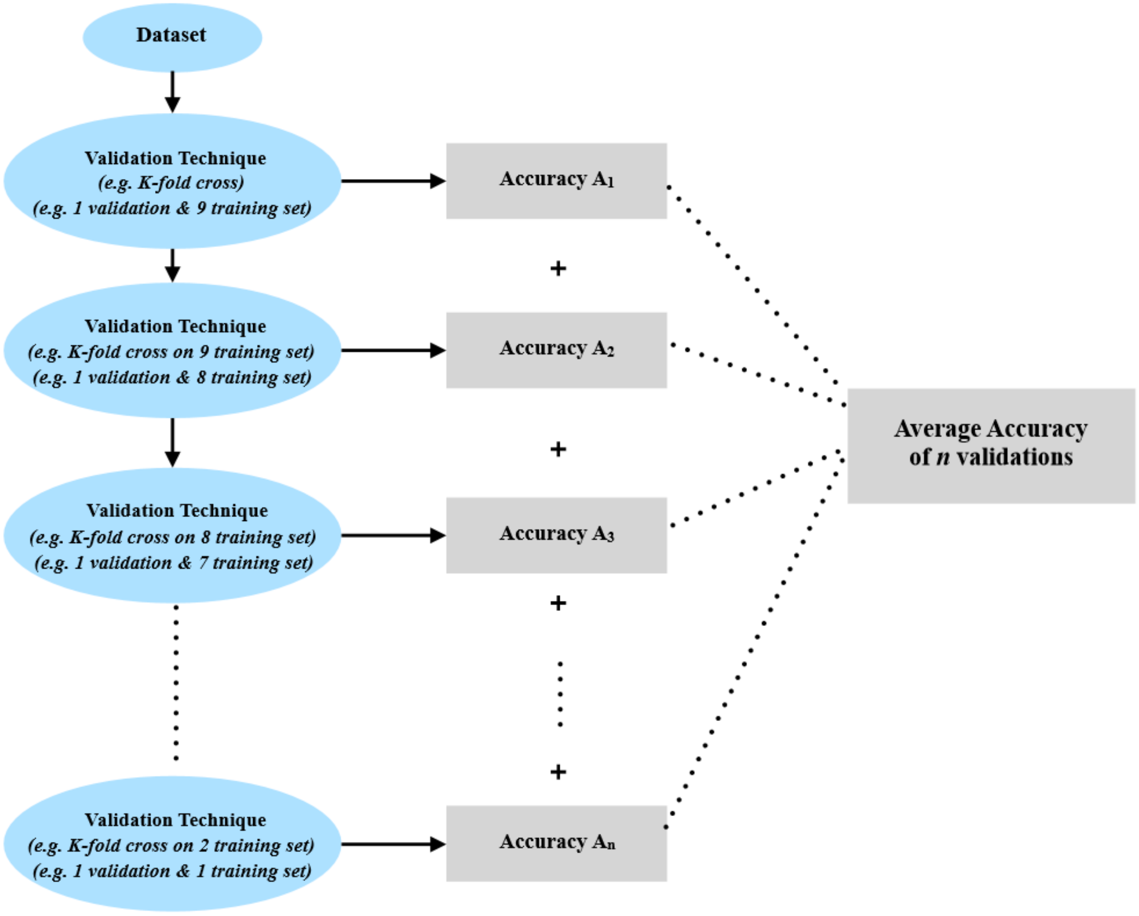

Dataset
Validation Technique
(e.g. K-fold cross)
(e.g. 1 validation & 9 training set)
Accuracy A1
+
Validation Technique
(e.g. K-fold cross on 9 training set)
(e.g. 1 validation & 8 training set)
Accuracy A2
+
Validation Technique
(e.g. K-fold cross on 8 training set)
(e.g. 1 validation & 7 training set)
Accuracy A3
+
+
Validation Technique
(e.g. K-fold cross on 2 training set)
(e.g. 1 validation & 1 training set)
Accuracy An
Average Accuracy of n validations

Figure 7.5 Draft technique for Objective 3

Objective 4: To validate the data using various validation techniques in a recursive order heterogeneously to obtain a more comprehensive measure of the accuracy.

The required validation techniques are chosen and are applied one after the other in a single iteration. Each iteration is branched out and the validation techniques are applied in an exhaustive manner heterogeneously. The output of the first step is obtained using a validation technique and then it becomes the input for the next step which would be validated using another technique. This technique aims to reduce the dependability on a single type of validation. The results obtained are averaged out over the cumulative number of steps in different iterations.

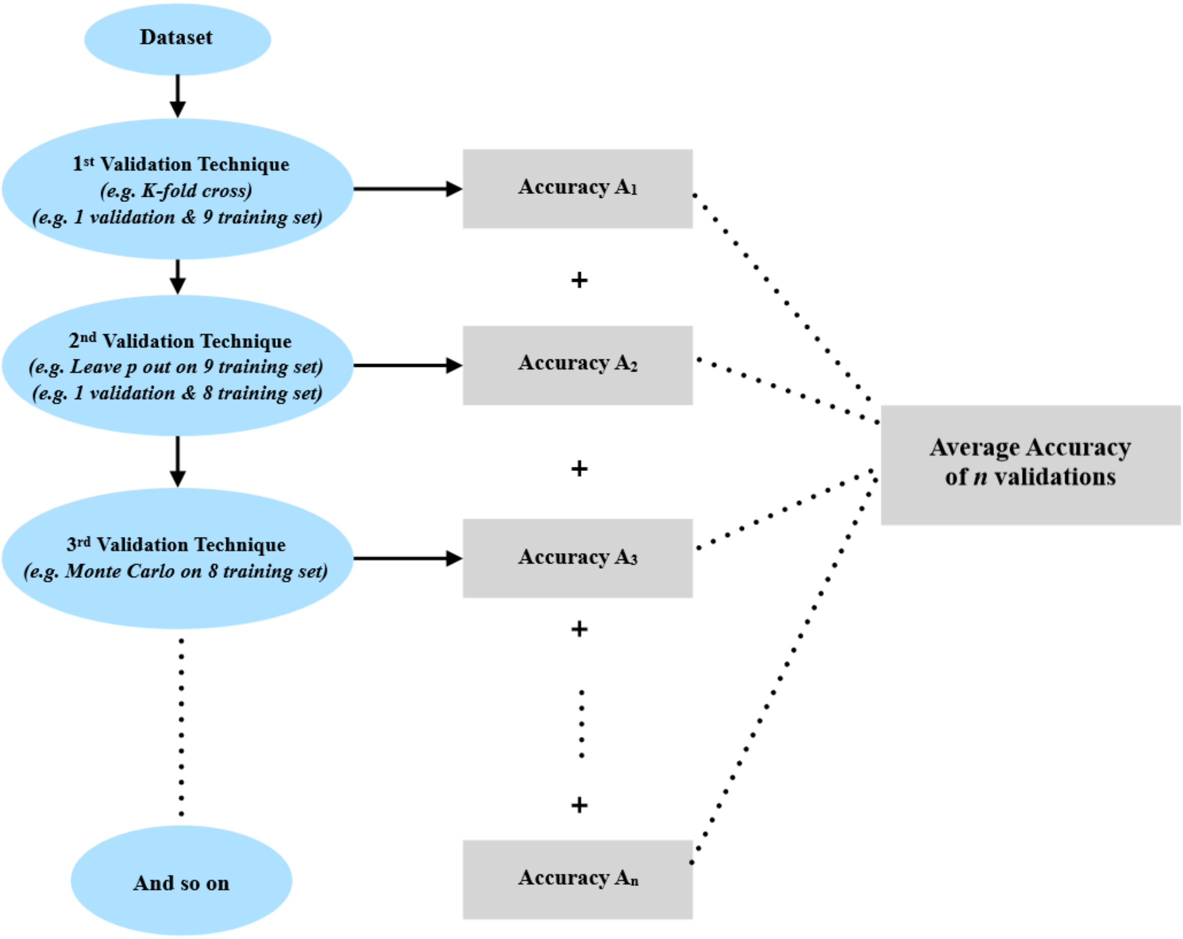

Dataset
1st Validation Technique
(e.g. K-fold cross)
(e.g. 1 validation & 9 training set)
2nd Validation Technique
(e.g. Leave p out on 9 training set)
(e.g. 1 validation & 8 training set)
3rd Validation Technique
(e.g. Monte Carlo on 8 training set)
And so on
Accuracy A1
+
Accuracy A2
+
Accuracy A3
+
+
Accuracy An
Average Accuracy of n validations

Figure 7.6 Draft technique for Objective 4

CHAPTER 8

SUMMARY AND CONCLUSIONS

The proposed research work aims at construction and training of a Neural network capable of detecting financial fraud. Financial frauds are of varied types and here, we have made credit card fraud as the focal point. By the end of the proposed research work, a Neural network is trained for credit card fraud detection which will be capable of distinguishing between a fraudulent and a non-fraudulent transaction. It will aid in preventing the unlawful and deceitful settlements in advance.

While undergoing a comprehensive study of the use of neural networks in the field of fraud detection, the performance efficiency of neural networks was compared with the all the available techniques in the arena. It was then revealed that after the application of the existing validation theories, the efficiency of the neural network is quite degraded. Therefore, the current manuscript proposes a new validation theory based upon a novel hybrid of the existing ones. This theory is expected to have a positive impact on the neural network while attempting to have minimal influence on its performance efficiency.

LIST OF REFERENCES

[1] Oxford Dictionaries | English. (2017). fraud | Definition of fraud in English by Oxford Dictionaries. [online] Available at: https://en.oxforddictionaries.com/definition/fraud

[2] D.Zhang, L. Zhou, Discovering Golden Nuggets: Data Mining in Financial Application, IEEE Transactions on Systems, Man, and Cybernetics 34 (4) (2004) Nov.

[3] Kirkos, Efstathios & Spathis, Charalambos & Manolopoulos, Yannis. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32(4), 995-1003. Expert Systems with Applications.

[4] Ngai, E.W.T., Hu, Y. H., Chen, Y., & Sun X. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support System (2010),

[5] Ravisankar P., Ravi V., Rao G.R. & Bose I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision Support System, 50, 491-500

[6] C.-C. Lin, A.-A. Chiu, S. Y. Huang and D. C. Yen, Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments, Knowledge-Based Systems, 2015.

[7] Ubon Thongsatapornwatana, A Survey of Data Mining Techniques for Analyzing Crime Pattern. IEEE in Defence Technology (ACDT), 2016 Second Asian Conference

[8] J. West, M. Bhattacharya, "Some Experimental Issues in Financial Fraud Mining", Proceedings of ICCS, 2016.

[9] West, Jarrod & Bhattacharya, Maumita & Islam, Md Rafiqul. (2015). Intelligent Financial Fraud Detection Practices: An Investigation.

[10] Mohammad Sultan Mahmud; Phayung Meesad ; Sunantha Sodsee. An evaluation of computational intelligence in credit card fraud detection. Computer Science and Engineering Conference (ICSEC), IEEE. 2016

[11] S. Kotsiantis, E. Koumanakos, D. Tzelepis & V. Tampakas, Forecasting Fraudulent Financial Statements using data mining, International Journal of Computational Intelligence 3(2) (2006) 104-110

[12] Raghavendra Patidar, Lokesh Sharma. Credit Card Fraud Detection Using Neural Network. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011

[13] Anuj Sharma, Prabin Kumar Panigrahi. “A review of financial Accounting Fraud Detection based on Data Mining Techniques.” International Journal of Computer Applications. 2012.

[14] Kapardis, M. K., Christodoulou, C. & Agathocleous, M. (2010). Neural networks: the panacea in fraud detection? Managerial Auditing Journal, 25, 659-678

[15] Liou, F. M. (2008). Fraudulent financial reporting detection and business failure prediction models: a comparison. Managerial Auditing Journal Vol. 23 No. 7.

[16] Yue, X. Wu, Y. Wang, Y. Li, C. Chu, A review of data mining based financial fraud detection research, international conference on wireless communications Sep, Networking and Mobile Computing (2007), 55195522

[17] Ahmed, S.R. (2004). Applications of data mining in the retail business, International Conference on Information Technology: Coding and Computing 2 (2) (2004) 455 – 459

[18] Yamanishi, K., Takeuchi, J., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Mining and Knowledge Discovery 8 (3) (2004) 275–300

[19] Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." Statistics survey 4 (2010): 40-79.

[20] Pandey, Yamini. "Credit Card Fraud Detection using Deep Learning." International Journal of Advanced Research in Computer Science 8, no. 5 (2017).

[21] Lu, Yifei. "Deep neural networks and fraud detection." (2017).

[22] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. (1995)

[23] Sohl, J. E., & Venkatachalam, A. R. (1995). A neural network approach to forecasting model selection. Information and Management, 29(6), 297303.

[24] Fanning, K., Cogger, K. O. and Srivastava, R. (1995). Detection of management fraud: A neural network approach. International Journal of Intelligent Systems in Accounting, Finance, and Management 4 113–126

[25] Fanning, K., Cogger, K. O. and Srivastava, R. (1995). Neural Network Detection of management fraud using Published Financial Data. International Journal of Intelligent Systems in Accounting, Finance, and Management, 7(1), 21-24