(IJTBM) 2019, Vol. No.9, Issue No. II, Apr-Jun e-**ISSN: 2231-6868, p-ISSN: 2454-468X**

AN ANALYSIS OF THE SUPERVISED AND UNSUPERVISED MACHINE LEARNING IN ENHANCING THE EFFICACY OF FINANCIAL ANALYSIS

Himanshu Dahiya

Bachelor of Technology (B.Tech), IT Manipal University, Jaipur-303007(Rajasthan), India

ABSTRACT

Data mining is the process of discovering patterns, corresponding to valuable information from the large data sets, involving methods at the intersection of machine learning, statistics, and database systems. Evolving from the fields of pattern recognition and artificial intelligence, machine learning explores the study and construction of algorithms that can learn from sample inputs. Financial data analysis is used in many financial institutes for accurate analysis of consumer data to find defaulters, to reduce the manual errors involved, for fast and saving time processing, to reduce the misjudgments, to classify the customers directly, and to reduce the loss of the financial institutions. We have analyzed a lot of machine learning techniques for financial analysis, namely models of supervised classification (Artificial Neural Networks, SupportVector Machine, Decision Trees), those of prediction (Cox survival model, CART Decision Trees), and also models of clustering(K-means clustering).

INTRODUCTION

Since 1980, a lot of researchers had the aim of their work to find solutions to search for a pattern in a large amount of data. There are used statistical, machine learning, and computational intelligence techniques grouped under the umbrella called data mining. Data mining is a technique for discovering interesting patterns from a large amount of data stored in the databases, data mart, data warehouse, or other information repositories. Datamart and data warehouses are tools that help in the management of business information. In the era of using the data warehouse through the development of the data mart, even though it will be limited to the use of any of the departments, but the information stored in the data warehouse is more important for a specific organization. A data warehouse is a set of data mart that provides information from the different operations in the company. It can contain information about the daily operations of the various factions of the company. Datamart, as part of a data warehouse, can provide transaction reports and analysis on some departments or operations in the company. Each company can keep the information of each department in its database, such as a database of the financial department, database of the sales department, database of production and that of marketing department Data warehouses have a basic architecture that can create applications from data mining. Data mining can be considered as a result of the natural evolution of information technology from multiple disciplines as database and data warehouse technology, statistics, high-performance computing, machine learning, computational intelligence(implying neural networks, fuzzy systems,

evolutionary computing, swarm intelligence and so on), pattern recognition, data visualization, information retrieval, image processing, and spatial or temporal data analysis. Data mining and Knowledge Discovery from the Database (KDD) are recent developments in the field of data management technologies. KDD is a kind of data mining designed to extract knowledge from a large amount of data. The standard procedure in performing data mining based on the Cross-Industry Standard Process of DataMining (CRISP-DM) involves six phases. See Figure 1.

These are the following:

1. Business understanding phase, consisting of choosing the objectives, understanding the business goal, learning situation assessment, and developing a project plan.

2. Data understanding phase, which consists of considering the data requirements and initial data collection, exploration, and quality assessment.

3. The data preparation phase, consisting of the selection of required data, data integration and formatting, data transformation, and data cleaning.

4. Modeling phase, consisting of the selection of appropriate modeling techniques, development, and examination of alternative modeling algorithms and parameter settings and finding the tuning of model setting according to an initial assessment of the model's performance.

5. Evaluation phase, corresponding to the evaluation of the model experiment results.

6. Deployment phase, representing an implementation step, where a model report is performed.

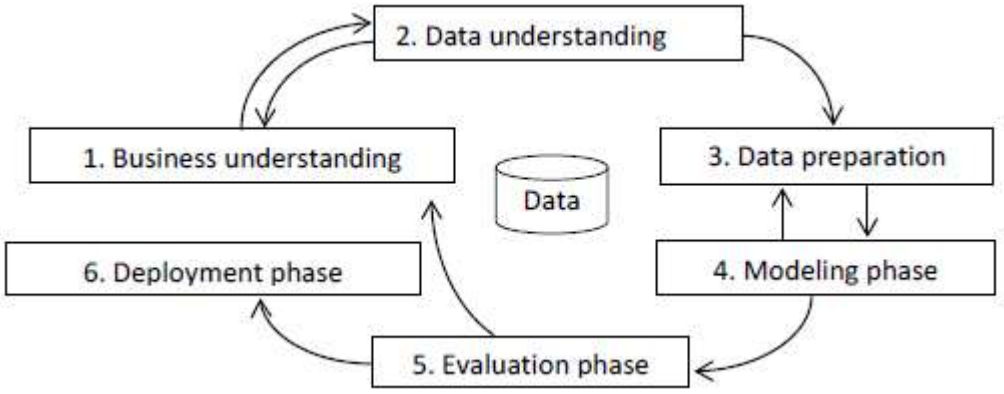

Figure 1. Cross-Industry Standard Process of Data Mining(CRISP-DM).

Data mining and machine learning are very necessary as a way of managing big data, enterprise efficiency, and business intelligence. Data mining provides enormous value in finance and banking. Banks need to find the hidden patterns in the large sets of data and thus, they can monitor the data in their database. Such data can be personal data that describe the financial status and the financial behavior before and by the time the client receives a credit. Most banks

and financial institutions have different services for customers, such as that of monitoring data for opening a savings account for each client's business. The schedule gives credit to customers in transactions such as mortgage business, car loan services, investment services, insurance services, and stock investment services. Other financial applications of data mining and machine learning are the prediction of financial events that will happen in the future, such as stock markets, foreign exchange rates, bankruptcy, credit rating of the bank's customer information, predictive financial and investment analysis, trading futures, understanding and managing financial risk in banks. As technology develops, it starts with bringing artificial Intelligence (AI) technology to be used in fund management, asset management, and other more financial institutions. Machine learning algorithms are used to isolate and analyze data from a large database.

Using this tool, one can find some patterns and can predict the outcome. However, there were many kinds of research using machine learning in banking to forecast future events that can help in decision-making processes. Nowadays, financial institutions and most banks are investing in information technology to bring data mining and machine intelligence techniques to manage the group of datasets in order to successfully operate in the presence of a competitive business. The banks are aware of the various risks. That may occur and adversely affect the business of the bank. Banks analyze the risk factors that are important. The quality of risk analysis may affect the financial performance of the business. There are risks of all institutions and organizations that may bring numerous direct and indirect losses. There are three major risks in banks corresponding to credit risk, operation risk, and market risk. Financial institutions should monitor credit risk management as appropriate. Banks are required to manage the credit risk compared to the risk of credit management individually. Credit risk management efficiency is very important and essential to the long-term success of banks. A popular tool used to evaluate the credit risk of individuals is Credit scoring. Credit scoring uses a report to evaluate some external components. The external reports process information on the status of credit risk data from credit bureaus and reliability party credit attributed together with the financial history and the current financial state of borrowers individually. Financial institutions have to remove unwanted features to distinguish between "good" and "bad" policies to manage the credit risk of each entity.

RELATED WORK

The bank is an organization that has a noteworthy job in the advancement of the economy of the nation. The obscure future practices of the clients are essential to Customer Relationship Management (CRM). It turns out to be progressively significant for the bank to anticipate their client's future choices to take reasonable activities in time. There are different territories where information mining and AI can be utilized in money related divisions like client division and productivity, credit examination, foreseeing installment default, advertising, deceitful exchanges, positioning speculations, streamlining stock portfolios, money the executives and determining activities, high hazard advance candidates, generally beneficial credit card customer and cross-selling. In 2010, M. C. Lee and C. To described the application of novel data mining

techniques for evaluation of the enterprise financial distress and credit prediction; there are improved the performance of algorithms by using Support Vector Machine (SVM) with3-folds cross-validation and Back Propagation Neural Network (BPN) by the four measured attributes. The data for this study have been collected from the database of a security firm in Taiwan. In this research, there are used 20 experimental samples for training data and 25 samples for testing data. By comparing the results, there has been shown that SVM gives higher precision of about 100% prediction accuracy and classification accuracy, implying low error rates, while BPN has led to 96% of prediction accuracy and 95% of classification accuracy. Many kinds of research about customer credit policy analysis were performed in 2012. K. Chopde et al. have studied the data mining techniques for credit risk analysis- in particular, the decision tree techniques. This research used data mining for credit risk analysis enabling the bank to reduce manual errors. This decision-making process is fast, it saves time processing, and it helps the bank to reduce the misjudgments. The research result found by the Meta Decision Tree (MDTs) used a base level classifier and the Random Forest (RF) classifier, leading to a more accurate classification score than the CART decision tree. Overall, the decision tree has proved to be a technique that can classify the customers directly with a good score, and thus, it can reduce the loss for the financial institutions in the best way.

I. G. Ngurah et al. used to suggest a decision tree model for credit assessment. This paper aims to identify factors that are necessary for a rural bank in Bali to assess credit applications. Current decision criteria in the credit-risk assessment are evaluated. The credit-risk assessment model has been applied to PT BPR X and it has used C5.0methodology; this model has used 84% of 1028 data as evaluation data to suggest the new criteria in analyzing the loan application. The result showed that PT BPR X could reduce nonperforming loans to less than 5%, and the bank can be classified or not as a well-performing one by applying data mining technology. In the same year, W. Chen etal29 proposed a hybrid data mining technique to build an accurate credit scoring model to evaluate credit risk based on the credit data set provided by a local bank in China. This research has proposed two processing stages: the first(clustering stage), meaning that the samples of accepted and new applicants are grouped into a homogeneous cluster by using K-means clustering. The second processing stage is the classification with Support Vector Machines (SVM). By comparison with other credit scoring models, here the samples. The previous model uses three or four classes rather than two (good and bad credit). In 2015, A. Byanjankaret al. described the application of Artificial Neural Networks (ANNs) for building scoring models in Peer to Peer Lending (P2P) to gain market share in the financial industry. This research used the neural network credit scoring model. The data have been divided in the following manner: 70% of the observations have been used for training, and 30% of observations have been used for testing. The neural network credit scoring model has shown a promising result in classifying credit applications to allow the lenders taking a smart decision in selecting a loan application and predicting the credit risk. In 2015, A. Gepp and K. Kumar proposed a semi-parametric survival analysis model consisting of Cox, Discriminate Analysis (DA), LogisticRegression (LR), and a non-parametric CART decision tree; the above models have been applied and compared to financial distress prediction.

Regarding classification accuracy, the CART model had led to the lowest error of classification, and regarding performance analysis of prediction accuracy, the Cox model had the lowest weighted error in 40% of the cases, while DA and the CART model had the lowest error in about 60% of the cases. The overall result provided empirical evidence that supports the use of survival analysis and decision tree techniques for financial distress. In 2016, R. G. Lopes et al. developed three predictive models: Generalized Linear Modelling (GLM), GradientBoosting Method (GBM), and Distributed Random Forest (DRF) by using the R language, to predict the recovery of credit operations in a Brazilian bank. All models have been built in 10-folds cross-validation, and there has been obtained a high evaluation result of the ROC curve. The GBM model has shown a better performance; this model has been used to help identify customers having the best potential with an 85.5% accuracy rate. This research has also identified the delinquent clients that had the highest probability of short-term recovery, to support the activities of account managers.

MACHINE LEARNING MODELS PROPOSED FOR FINANCIAL ANALYSIS

Classification techniques support Vector Machine (SVM) is a tool to find the hyperplane that can be used for classification; it is based on kernel functions. The Gaussian kernel is the most versatile kernel. By the width parameter of the Gaussian kernel function, one can control the flexibility of SVM classifier results. The Gaussian function can be used not only as a kernel for SVM but also for some exciting neuro-fuzzy classifiers. Decision trees are classifiers expressed as a recursive part of the instance space. Classification and Regression Trees (CART) model is an adaptable technique to depict how the variable Y disperses in the wake of relegating the figure vector X of the measurement. The CART model uses a doubletree to separate the estimated space into specific subsets on which Y dispersion is expected continuously. Counterfeit Neural Networks (ANNs) constitute a nonlinear measurement model dependent on the capacity of the human brain. ANNs give useful assets of information digging systems for information expert relationship displaying. ANNs can perceive the perplexing examples in info information, and they can foresee the result of the new autonomous info information precisely. ANNs have the amazing capacity to get importance from confounded information or uncertain information. It very well may be utilized to concentrate examples and recognize patterns utilizing explicit techniques. ANNs are entirely appropriate for distinguishing examples, and they are additionally very appropriate for expectation or gauging data. One of the most well-known ANNs is the Multi-Layer Perceptron (MLP), also named the Back-Propagation Neural Network (BPN); its algorithm is based on the computation of the errors of each output neuron after processing an input data. It is a general technique called automatic differentiation. BPN is characterized by backward propagation of output errors; namely, these errors are computed at the output layer, and the training is distributed back to the weights of the previous layers to reduce the output errors. Survival Analysis method is a new technique of credit-scoring model. A common way that banks can differentiate customer information when they apply for a loan from the bank. Banks can separate the good information from the bad information regarding the loan

application. The system can calculate the profitability of customers, and also it can evaluate the profit scoring from the customers. Worthwhile Survival analysis can predict the duration of the event will take place in advance and forecast the probability of occurrence of an event to occur. The H2O team discovered these famous data mining techniques to analyze the group datasets. These techniques are Generalized Linear Models (GLM), Gradient BoostingMethod (GBM), and Distributed Random Forest (DRF).GLM is similar to the linear regression model. Data mining techniques are used for regression analysis and data classification. GLM model is very popular because it is easy to be interpreted, and it is also a high-speed processing stage when used for large datasets. GBM model is a tool for prediction using regression or classification. It is an ensemble of the tree models and provides considerably accurate results. GBM model applies weak classification algorithms to incrementally change data, to create a series of decision trees. Finally, the DRF is an ensemble of tree models, where each tree is related to other trees. DRF is the most powerful technique for classification and regression. DRF can produce a major wood of arrangement or relapse trees as opposed to a solitary grouping or relapse tree. What's more, DRF assembles half the same number of trees for binomial issues with a solitary tree to gauge class 0 by probability (p0) at that point registers the likelihood of another class 1 as (p1). For multiclass issues, DRF is utilized to assess the likelihood of each class separately.

Clustering technique

Cluster analysis groups are the data mining techniques used to classify as variable or split into small groups of two or more. The objects within a group are similar to one another and different from the objects in other groups.K-means clustering is a method of clustering the observations into a specific number of disjoint clusters. In principle, K-means clustering aims to partition a dataset as {X1, X2,…, XN} into K subsets to minimize the distortion measure defined by the function where binary indicator rnk=1, only if data point Xn is assigned to the kth cluster(for the other cases, rnk =0)and μk denotes the mean of the kth cluster. In Table 1, there are given a lot of papers for financial (banking) applications with the corresponding machine learning techniques and research results.

Table 1. List of research papers with corresponding data mining tasks, machine learning techniques, and research results

Reference	Data mining tasks	Machine learning techniques	Research results
15	Clustering	K-means clustering	There can befound the panel data structure. There can reflect analysis for different periods.
29	Classification Clustering	SVM K-means clustering	There can be applied to panel data to find knowledge which is different from the regression knowledge discovered by the traditional linear regression.
34	Classification	SVM BPN	There are compared the results obtained by SVM and BPN for financial distress. The research results had shown that SVM leads to a lower error rate than BPN.
23	Classification	Decision trees – The CART model – MDT – RF	Decision trees techniques can reduce the manual errors, to obtainfaster and saving time processing,they can reduce the misjudgments, canclassify the customers directly and can reduce loss for the financial institutions.
24	Classification	Decision trees – PT BPR X, C.50	There is reduced the number of non-performing loans.
30	Classification	ANNS	There are classified the credit applications in order to allow the lenders taking a smart decision to select a loan application and to predictthe credit risk.
35	Classification Prediction	Decision trees – The CART model A survival analysis – Cox model – DA model – LR model	The presented results provide empirical evidence to support decision trees anda survival analysis in banks for financial distressto comparethe performance analysis. The CART model had obtained the best classificationaccuracy. In addition, the Cox, CART and DA modelshad led also to good prediction accuracy.
25	Classification	GLM Model GBM Model DRF Model	GBM model has shown a better performance. GBM had the highest probability of short- term recovery to support the activities of account managers and increase the efficiency of their approach with customers.

CONCLUSION

Data mining based on machine learning techniques is a technology that can be used to analyze existing data, applications and customer needs in order to build and maintain long-term customer relationships. It can build confidence for clients making customer satisfaction and business the longest. Using machine learning techniques for classification and clustering tasks is popular in the loan payment prediction and the customer credit policy analysis of the banking system. In this paper, we proposed data mining techniques that contain two mains processing stages. The classification stage consists of several models, including SVM, ANNs, Decision Trees, and BPN. We found that the SVM model and Decision Tree model are promising techniques for classification with financial applications. The techniques mentioned above can reduce manual errors; they can lead to faster and saving time processing; they reduce them are judgments for classifying the customers directly, and thus they can reduce the loss of the financial institutions. In the clustering stage, K-means clustering is the best performing model for customer credit management of the credit scoring model. The scoring methods are used to estimate the creditworthiness applicant. When credit loans and finances have the risk of being defaulted, credit managers have to develop and apply data mining methods to handle and analyze credit data in order to save time and reduce errors. Data mining (implemented mainly using techniques of machine learning) will be a challenge for future research in banking and financial areas.