Financial Distress Prediction System Using Machine Learning Algorithm

Dr. P. Siva Kumar[1], Geetha Lakshmi[2], Kowsalya[3], Aishwarya[4]

[1]Professor and HOD, [2]Department of Information Technology, Manakula Vinayagar Institute of Technology, Puducherry, India.

Abstract - In today's world, taking loans from a financial institution has become a very common phenomenon. Every day a huge number of people create an application for loans, for a variety of purposes. But every these loan applications of applicants are not eligible for the loan. Every year, we read about a number of cases where people do not repay the bulk of the loan amount to the banks because of that they endure massive losses. The risk connected with creating a decision on loan approval is extremely large. Thus, the idea of our paper is to gather loan data from multiple data sources and use of Machine learning (ML) methods such as ACO for Feature selection and DNN for the classification process. It will classify the data as classes based on the eligibility criteria for the approval of credit to the customer or not. By using the confusion matrix we evaluating the obtained result from the classification process for the evaluation of the precision and accuracy for the loan prediction process.

Keywords — ACO, Classification, Confusion matrix, DNN, Feature selection, Machine Learning.

I. INTRODUCTION

ML is a function of artificial intelligence (AI) which gives the capability to the computer automatically learns without being explicitly programmed. The main goal of ML is to construct programs that are able to improve their performance using the past experiences collected during their operations and also build an analytical model depends on model data or training data in order to make forecast or decisions. Today many companies and financial sectors using ML techniques for making better decisions, upgrade productivity, predict disease, forecast weather and do many more things.

Financial distress prediction is highly important for financial sectors that aims to reduce the upcoming losses by evaluating feasible threat and avoids new loan applicant when the default loss is larger than a preset acceptance size. In the banking sector, the loan is one of the most important products of banking. Every bank is demanding to work out successful business strategies to promote customers to apply their loans. Still, there are some customer's acts negatively after their applications are approved. To prevent this situation, we have to find some methods to predict customers' behaviors and will focus on loan approval by using ML models. ML algorithms have an adequate performance on this purpose, which are widely used by banking. Permitting loans for both retail and business customers depends on credit scoring. By this method, we can estimate whether that specific applicant is safe or not and the entire procedure of validation of features is automated through ML method. It helps to determine whether to receive or decline loan application depends on 4 main input data:

Client Details: age, gender, marital status, job, incomes/salary, housing (rent, own, for free), geographical (urban/rural), residential status, existing client (Y/N), number of years as a client, total debt, account balance.

Credit Details: Net amount, need of credit, amount of the monthly payment, interest rate.

Credit history: Payment history and misconduct, Amount of current due, number of months in payment due, Span of credit history, time since last credit, Types of account in use.

Bank account conduct: mean monthly savings amount, maximum and minimum levels of stability, credit yield, change in payments, a change in balance, number of missed payments, times run over the credit limit, times changed home address.

II. NEED OF THE SYSTEM

Credit Prediction is most helpful for the workforce of banks in addition to the customer. The goal of this paper is to give a fast and simple method for selecting the deserving customer. This paper is absolutely for the administrating leverage of the finance sectors, the entire process of prediction is done exclusively. There are several prospects involved in bank loans, so as to decrease their capital loss; banks should perform the risk and rating the individual before sanctioning loans. In the absence of this process there are many chances that this credit may turn in to bad loan in the near future. Banks hold large volumes of customer behavior-related data from which they are unable

 DOI : 10.35291/2454-9150.2020.0554

to arrive at a decision point. This model can be used by the organizations in making the correct decision to approve or reject the credit request of the client.

III. Related Works

Projection of Loan Sanctioning Process

As enlarge in the number of clients applying for credit in the banks and non-banking financial sectors (NBFC), it is greatly difficulty for banks and non- banking financial sectors (NBFC) with less capital. In this paper, the aim is made to find the threat occupied in chosen a proper person who can repay the loan on time. In the paper, they used different classification techniques and compare their results to find the best classification techniques. This archived by the following sections (i) Collecting of Raw Data, (ii) Data Cleaning and (iii) Performance analysis. Experimental tests found that the Naïve Bayes model has better and good result than other models.

Loan Risk prediction Models

In this paper, they determine the validity of many concepts in the R language and rate it to begin the finest process to predict the finance level for an organization. They did the experiment 5 times on the identical data set in 11 techniques of classifications and find the test results that show the Tree Model for Genetic Algorithm (GA) is the optimal concepts to forecast the finance for clients.

Ensemble Technique prediction

Evaluating the possibility that an individual would default on their loan. They described a prototype in this paper which can be used by the organizations for making the correct or right decision to approve or reject the request for the loan of the customers. They use three different types of models (SVM Model, Random Forest Network and Tree Model for GA) and the Ensemble Model, which combines these three models and analyses the credit risk for optimum results.

Loan Approval using Machine Techniques

This paper applied ML in the prediction of loan approval. Three ML algorithms are used to guess the credit approval status of clients for the bank credits. The results showed that the prediction accuracy is 93.04% for linear regression, 95% for decision tree and 92.53% for random forest respectively. By the experimental results ends that the accuracy of the Decision Tree ML algorithm is better as compared to Logistic Regression and Random Forest ML approaches. Among three the accuracy of the DT algorithm is finest for the prediction of loans.

IV. Existing Techniques

Financial crisis prediction is a potentially good system that will predict the status by analyzing the details of the customer to approve their loan request using ant colony optimization (ACO).

In this method, they use only ACO-feature selection and ACO-classification algorithms based on five benchmark datasets, the submission of the ACO algorithm in the feature selection and classification process. Initially, the ACO algorithm undergoes the feature selection method and selects the optimal feature subset. Then, the selected feature subset is taken into the classification progression. In this way, the ACO-FCP model classifies the financial data and predicts whether a crisis occurs or not, But this system has some limitation is given below.

ACO algorithm has a deficiency of local optimization in the classifier (i.e.) local optimization finds the finest value within the nearby set of candidate solutions.

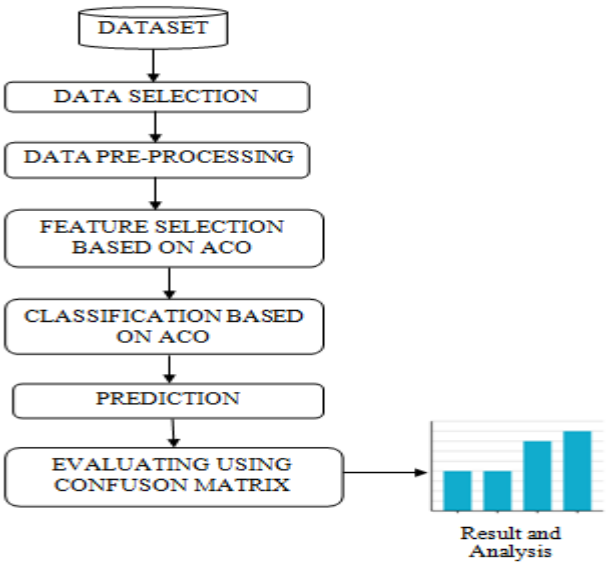

Figure 1: Existing model architecture diagram

V. Proposed Techniques

In financial distress prediction replica focuses on predict the status of customers for finance refunds by examining their conduct. The status of customer behavior is the input for our model. The result from the classification is used for the assessment on whether for supporting or refusing the customer application is developed. Utilizing several data analytics devices loan forecast and their severity is predicted. During this procedure, it can be needed for training the data utilizing several ML techniques and next related user data with trained data for forecasting the nature of the loan. Different functions and packages is utilized to practice the data and to construct the classification representation. In real-time customers, data sets may have frequent missing and impute data which need to be replacing with valid data generated by making use of the obtainable completed data. The dataset has countless attributes that describe the reliability of the clients in search of for several types of loans. The values of attributes in the dataset can have outliers that do not fit into the usual range of data. Therefore, it is required to remove the outliers before the dataset is used for added modeling.

 DOI : 10.35291/2454-9150.2020.0554

Figure 2: work flow diagram

Data selection

The data composed for processing may have lost values, noise. A data method with a high feature of data will generate proficient data mining outcomes. To enhance the eminence of data and accordingly the mining results, the composed data is to be pre-processed so as to enhance the outcome of the data processing. This leads to remove incompatible information from the mining method.

Data pre-processing

The dataset has many missing and noisy data which is replaced by several steps. Data preprocessing is one of the censorious steps in the data pre-processing which prepares and transforms from the raw facts set to the final data set. Data preprocessing is the greater time-ingest part of data processing. Data cleaning of the credit data pulls out many attributes that have no important about the conduct of a client. Data integration, data reduction, and data transformation are also to be applied for credit data selection. For effortless investigation of the credit records. Initially, the features which are censorious to making a loan reliability prediction are recognized and ranker as the search-method.

ACO based feature selection process

The ACO is a graph-based Meta heuristic we require to layout the complete graph. Each and every node of the graph determines a feature of the preliminary n feature set, and each boundary represents a selection from the one real characteristic to the other. Each ant will begin with an unfilled set of features and travels through the graph visiting the least number of nodes which is satisfy the traversal-stopping criterion, and finally, the output is subsided. As depicted in the figure, an ant started on node A executed a route up to node F after that stop containing create the subsequent subset {A, B, C, D, F} that convince the traversal stopping condition.

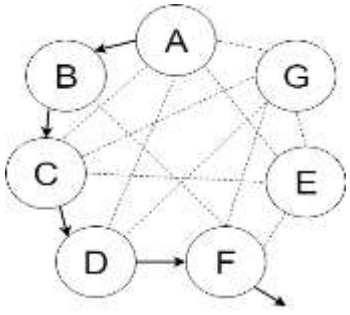

Figure 3: Graphical selection of the feature selection

DNN-Based Classification process

Deep Neural Network (DNN) architectures are intended to learn through multiple connections of layers where every single layer only receives a link from earlier and provides relations only to the next layer in the hidden part. The output layer neurons are equal to the number of classes for multi-classification. Firstly, the output from the ACO-feature selection is provided as the input of the DNN classification process. The DNN appears underneath a type of feed-forward network having the following pairs of Convolutional layer, max-pooling (MP) and varied fully connected (FC) layers. The unique strength of data presents in the input is given in the hierarchical feature extractor. A created feature vector undertakes classifier using the FC layers.

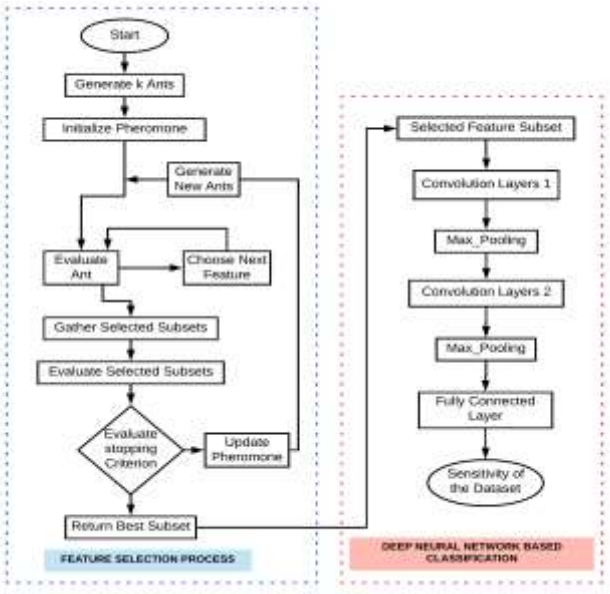

Figure 4: Architecture Diagram for proposed system

Each convolution level carry out a 2D conv of its input maps utilizing a rectangular filter generated through possibilities which all features goes to the needed class (in a simple classification example). The 2nd part is pooling (is known as down sampling) that will reduce the dimensionality of all features and continuing its most significant data. The pooling phase develops a "summary" of the most significant features. The DNN has several relations, weights, and nonlinearities. Almost the different pair of conv as well as MP layers, single FC is only combines by the conclusion of the feature vector. The resultant layer is usually an FC layer through the single neuron in separate classes endure verifying ensuring which all neuron's outcome establishment is reasonable as the probability of particular input goes to the class.

VI. Result And Discussion

As a result of this process is obtained by evaluating the classification outcome with the help of the confusion

 DOI : 10.35291/2454-9150.2020.0554

matrix. In order to assess the performance of the proposed ACO-FCP model accuracy measures are employed which are generated from a 2×2 confusion matrix. The efficiency of the feature selection method is validated in terms of obtained cost. To analyses the classification performance, various measures like False Positive value, False Negative Value, sensitivity, accuracy, True Positive value , True Negative value are used. For better classification results, the values of confusion matrix should be as low as possible whereas the values of sensitivity, specificity, accuracy values should be as high as possible. And, the cost of the feature selection technique should be low for effective performance.

Confusion matrix

A confusion matrix is the process of prediction results of a classification process. The numbers of positive and negative predictions are outlined with a count by accuracy, precision, and recall.

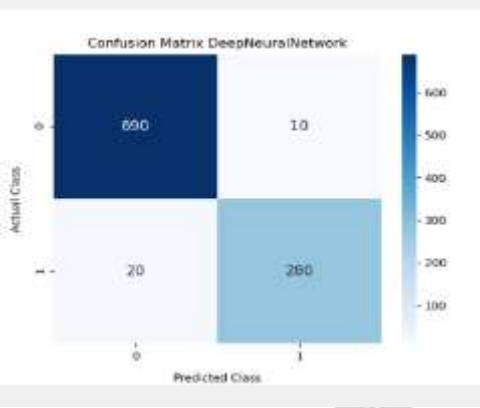

Figure 5: confusion matrix

Accuracy

The accuracy is defined as the full amount number of 2 correct forecasts (TP + TN) divided by the full amount number of a dataset (P + N).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP: True Positive: Predicted values correctly predicted as actual positive

FP: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive

FN: False Negative: Positive values predicted as negative

TN: True Negative: Predicted values correctly predicted as an actual negative

You can compute the accuracy test from the confusion matrix:

Techniques	Accuracy
ACO-FS and ACO-C	76.6
ACO-FS and DNN-C	97

Table 1: Result comparison

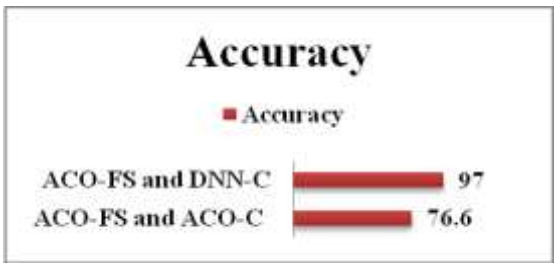

Figure 6: Accuracy graph

VII. CONCLUSION

This paper demonstrates the basic idea for predicting the occurrence of the financial crisis use the ML techniques such as ACO for Feature selection and DNN for classification algorithms. By using this application banks can minimize the number of imperfect loans and severe losses. Many python functions and packages are utilized for producing the loan facts and to make the classification representation. By using this process banks can identify the wanted information easily from a large amount of data and helps in finding successful loan prediction. This process is very useful for the banking sector for better risk management and marketing.

REFERENCES

[1] Uthayakumar J, Noura Metawa, K. Shankar, S.K. Lakshmanaprabu "Financial crisis prediction model using ant colony optimization" International Journal of Information Management

[2] Anchal Goyal, Ranpreet Kaur "Loan Prediction Using Ensemble Technique" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[3] Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta "Loan Sanctioning Prediction System" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-6 Issue-4, September 2016.

[4] Kumar Arun, Garg Ishan, Kaur Sanmeet "Loan Approval Prediction based on Machine Learning Approach" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 18-21.

[5] Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai "Machine Learning in Financial Crisis Prediction: A Survey" IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 4, July 2012.

[6] Martin, A., Uthayakumar, J., & M, N. (2014). Qualitative Bankruptcy data set [WWW document].

 DOI : 10.35291/2454-9150.2020.0554

Reposhttps://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy.

[7] Quinlan (1987). Australian credit dataset [WWW document]. URLUCI Mach. Learn. Reposhttp://archive.ics.uci.edu/ml/datasets/statlog + (Australian + credit + approval).

[8] Tomczak, S. (2018). Polish companies bankruptcy data data set [WWW document]. n.d. URLUCI Mach. Learn. Repos https: //archive. Ics.uci. edu/ml/datasets/Polish + companies + bankruptcy + data.

[9] Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Systems with Applications, 41, 2353–2361.https://doi.org/10.1016/j.eswa.2013.09.033.

[10] Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2012). Machine learning in financial crisis prediction: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 42, 421–436. https://doi.org/10.1109/TSMCC.2011.2170420.

[11] Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Systems with Applications, 41, 2353–2361.https://doi.org/10.1016/j.eswa.2013.09.033.

[12] Tsanas, A., Little, M. A., & McSharry, P. E. (2013). A simple filter benchmark for feature selection. Journal of Machine Learning Research, 1–24.

[13] Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. (2017). Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. Journal of King Saud University – Computer and Information Sciences.'

[14] Lin, Y., Guo, H., & Hu, J. (2013). An SVM-based approach for stock market trend prediction. The 2013 International Joint Conference on Neural Networks (IJCNN) 1–7.

 DOI : 10.35291/2454-9150.2020.0554