ISSN 2278-3091
Volume 9, No.1.3, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5691.32020.pdf
https://doi.org/10.30534/ijatcse/2020/5691.32020

Optimized algorithm for Credit Scoring

Annie Chacko[1], Dr.A.Antonidoss[2],Aleena Sebastain[3]
[1]Research Scholar-PTE, Dept. of CSE, Hindustan Institute of Technology and Science, Chennai, India, annievargh@gmail.com
[2]Associate Professor, Dept. of CSE, Hindustan Institute of Technology and Science, Chennai, India, aro.antoni@gmail.com
[3]Second Year M-Tech Student, Dept. of CSE, MBCCET, Idukki, India, aleenasebastain07@gmail.com

ABSTRACT

The rapid expansion of credit scoring technologies is increased today. Credit scoring will be considered as the significant element in the financial industries. It plays an important role in modern affairs such as credit customer selection, risk measurement, post-loan and after-loan supervision, comprehensive performance evaluation etc. Credit scoring has been recognized as a binary classification technique distinguishing applicants into two classes: good credit and bad credit, based on characteristics such as gender, age, occupation, and salary. These determine the applicability of loans for applicants. There are two main stream classification techniques, statistical techniques and machine learning techniques. Linear discriminant analysis and logistic regression are the two most commonly used statistical techniques in credit scoring. Machine learning techniques include K- nearest neighbor, support vector machine, decision tree and neural network. Use different best algorithms for classify the credit scoring data sets. Here uses four algorithms for the classification of credit scoring data sets and then the accuracy of different algorithms on the data sets will be obtained.

Key words: Credit scoring, classification techniques, machine learning.

1. INTRODUCTION

Credit Scoring is an important tool in financial institutions. Credit Scoring Model helps to reduce credit risk in Microfinance Institutions. The advantages of using credit scoring model included that it reduces the cost of credit analysis, enabling faster credit decision, insuring credit collections and diminishing possible risk. Different types of credit scoring models are available today. Different models use different classifier models for classification.

In machine learning and statistics, classification is a supervised learning approach and which is basically classifies different set of data into classes. Different type classifiers are there. They include Perceptron, Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor, Artificial Neural Networks and Support Vector Machine etc. Different credit scoring model uses different classification techniques.

There are two types of automatic credit scoring techniques. They are statistical techniques and artificial intelligence techniques. Some statistical techniques widely used are Linear Discriminant Analysis, Logistic Regression Analysis, and Multivariate Adaptive Regression Splines etc. The AI techniques include Artificial Neural Networks, Decision Tress and Support Vector Machines etc. In contrast statistical techniques, AI methods can be used as an alternative method for credit scoring.AI techniques does not assume certain data distribution. It will be automatically extract knowledge from training samples. Based on study conducted it will be shows that Al techniques superior to statistical techniques in dealing with credit scoring problems, especially for nonlinear problems.

The overall best AI techniques can be used for building credit scoring will be depends on the details of the problem, The data structure, the characteristics used, the degree to which it is conceivable to isolate the classes by utilizing those qualities and the objective of the qualification. Depend on the classifier used in credit scoring model the performance of the model will be changed.

The paper is organized as follows. Section 2 describes the survey done on the paper "Different classification model for credit scoring." This section goes through more various documents giving an overview of different techniques. Section 3 discusses the overall design of the proposed system. Section 4 presents a conclusion.

2. LITERATURE SURVEY

The following are the papers surveyed for the need of doing the project to find out the different classifiers used for credit scoring.

Credit rating is an estimate of the ability of a person or organization to fulfill their financial commitments based on previous dealings. There are two basic types of credit rating and one is for specific debut issues or financial obligations and other one is for debt issuers. Here proposes a machine learning technique, support vector machines to the credit rating. Here back propagation Neural Network as a benchmark and BNN and SVM achieved better prediction accuracy. The introduced learning method based on statistical learning theory, SVM, together with frequently used high performance method and BNN. SVM achieved higher accuracy when comparing with BNN. The BNN can have three layers i.e., input layer, output layer and hidden layer. Here uses financial variables as the input node and rating outcome as the output layer nodes [1].

The unrepresentative samples reduce the usefulness of data classifiers. Here presents a hybrid mining approach in the design of an effective credit scoring model. It will be based on the clustering and neural network techniques. Clustering techniques are used to preprocess the input samples in a network. Neutral network techniques are used to construct

the credit scoring model. A class wise classification process is involved in classification stage. To determining the number of clusters a self-organizing clustering algorithm will be used and it will be automatically determining the clusters. The K-means clustering algorithm is used to generate clusters of sampling belongs to new classes and also eliminate the unrepresentative samples from each class. To design the credit scoring model samples with new class label is used in the neural network stage. The proposed hybrid scoring model has two processing stages. In the clustering stage, samples are divided in homogeneous clusters. In the neural network stage, create models for predicting consumer loans and for that uses neural network with new class labels[2].

Here uses ROC curve analysis to compare the model performance of Logit and Fuzzy Logic. A Receiver Operating Characteristic Curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Fuzzy logic is a form of many-valued logic in which the truth values of variables may be any real numbers between 0 and 1. Logit model or logistic model it is used to model the probability of certain class or event. In terms of overall classification accuracy and prediction accuracy FL exceeds than Logit. Logit will be preferable in situations where higher accuracy in classifying no default firms is used. Fuzzy logic is the process of grouping elements into fuzzy set. Logit is a logistic regression. It will be consider as the most accurate of the traditional statistical method for prediction and classification problems [3].

Credit scoring will be the most important task in credit industry. Here uses two data mining techniques such as, classification and regression tree and multivariate adaptive regression splines. In terms of credit scoring accuracy CART and MARS will be perform better than traditional discriminant analysis, logistic regression, neural networks, SVM approaches. CART is a statistical procedure. Primarily used as the classification tool and where the objective is to classify an object into two or more populations. It can be used to analyze either categorical or continuous data using the same technology and it will be treated as single procedure. MARS is a flexible procedure which models relationships that are nearly additive or involve interactions with fewer variables. CART and MARS scoring model have better classification capability in terms of the average classification rate. The CART and MARS do not need long training process therefore it will save lots of modeling time when the data set is large. These two will be considered as the effective tool for handling forecasting and classification problems [4].
Hidden Markov model is a statistical model that can be used to describe the evolution of observable events that depend on internal factors, which are not directly observable. The HMM have limited capacity to recognize complex patterns involving more than first- order dependencies in the observed sequence. The drawback involving HMM can be overcome by using GMDH. Group Method of Data Handling is a family of inductive algorithms for computer-based mathematical modeling of multi-parametric datasets that features fully automatic structural and parametric optimization of models. Here propose a hybrid model by combining the HMM and GMDH for credit scoring. There are three phases in this model. Firstly, by multiple observations HMM will be trained. Secondly, adding GMDH into HMM. Then optimize the hybrid model into loops. The proposed hybrid model is shows better performance than HMM/ANN in terms of average accuracy, Type I error, Type II error and AUC [5].

Here propose a new combination approach based on classifier consensus to combine multiple classifier system of different classification algorithms. Here five well known base classifiers will be used. They are Neural Networks, Support Vector Machines, Random forests, Decision tree, naïve Bayes. The proposed model show better prediction performance than the other models in terms of average accuracy, area under the curve, H-measure and Brier score. The MCS also known as a ensemble model. It will be consider as an alternative for single classifier and show better predictive performance than the single classifier [6].

Credit scoring will be considered as the significant element in the financial industries. It plays an important role in modern affairs such as credit customer selection, risk measurement, post-loan and after-loan supervision, comprehensive performance evaluation etc. Here generate a novel ensemble model for credit scoring to different imbalance ratios datasets are proposed and it will obtain superior performance and high robustness to different datasets.

Firstly, according to credit scoring data characteristics the proposed method extends the Balance Cascade approach to generate adjustable balanced subsets based on the imbalance ratios of training data. Balance Cascade approach is an educated undersampling strategy, which can successfully defeat the shortcoming of data lost by haphazardly evacuating redundant samples with irregular undersampling strategies. It removes correctly sampled information in each iteration. Secondly, the proposed model adopt two kinds of tree-based classifiers ie, random forest and extreme gradient boosting. These two classifiers used as the base classifier for the ensemble model. It also uses stacking to generate the predicted result. The parameters of the base classifiers are optimized using particle swarm optimization algorithm. Based on other comparative algorithms the average performance of the proposed model will be higher than the other. In credit scoring the credit applicants are good users and the remaining small parts is bad users ie, there is a data imbalance ratios. The cost of misclassifying good object as bad is much larger than the cost of misclassifying bad object as good. Also, imbalanced datasets accompany certain difficulties for the development of a classification model that do not occur for balanced data sets. Sampling is the one of the most commonly used approach to optimize the predictive performance in high imbalance ratio data. There are two types of sampling approaches. They are oversampling and undersampling. Here a novel ensemble model is constructed to obtain good predicted result than the other methods

adapting to different imbalance ratios.

Firstly extends the supervised undersampling approach. For that extends Balance Cascade approach. By estimating data imbalance ratios it is used to construct adjustable data subsets. Here random forest and extreme gradient boosting are used as the tree- based base classifiers and the constructed subset are used for training the base classifiers. Here uses stacking to create the predicted result of former layer as the new informative features of the latter layer. The PSO algorithm will be used to optimize the parameters of the base classifiers. Then the optimized classifier will be used to construct the optimized ensemble model. Finally the ensemble model is trained by subsets predict the testing set and then the obtained result are compiled as the final prediction result. The model will be applied to different datasets then the result shows that the proposed will be superior in performance [7,10].

Here presents a credit scoring model to predict default clients and reduce credit risk of financial institutions by applying data mining algorithm and there by support decision making process of credit approval. The proposed credit Scoring Model was build using Oracle Data Miner software package. It uses Generalized Linear Model for classification. The proposed credit scoring model has great predictive confidence and accuracy. Here uses Oracle Data Miner software package for selecting the most relevant attributes for the implementation of credit scoring model. ODM is the software package for data mining by oracle. Here uses transform tool of ODM to remove the redundant attributes. Here uses GLM algorithm for classification. GLM implements logistic regression for classification of binary targets. There will have a drawback such as it will not fill the unknown valued during the pre-processing phase [8].

Here propose a deep learning approach for credit scoring method based on attention mechanism LSTM. It is a novel application of deep learning algorithm. The proposed model show better predictive accuracy compared with the traditional artificial feature extraction method and the standard LSTM model. Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM has feedback connections. It can not only process single data points, but also entire sequence of data. Here does not consider the time interval of the events when it be occur in the case of considering the sequence of operation in the case of online user operation behavior data [9].

By considering the different classification techniques then made a new method for credit scoring to overcome the defects related to different models. The proposed model shows better performance than others.

3. SYSTEM DESIGN

There are different types of defects related to different models and here presents a new method for credit scoring.Figure 1 shows the overall design. Have a look at consists of three stages: During the first stage data processing is employed. In this stage standardize the original data into a standardized format and also generate more representative samples. In the second stage feature selection is employed. In the last stage provide the evaluation metrics.

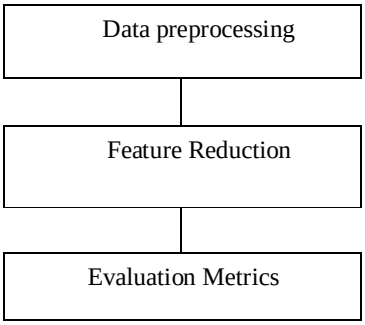

Figure 1: Sample design

The original data can have missing values and abnormal values that will be not favorable for the prediction of the proposed model. The data preprocessing will be done then standardize the original data. The original data will contain more representative features that will not be discovered. After the data preprocessing making the data more representative and standardized format.
After the data preprocessing feature reduction will be done. The principal component analysis will be done to reduce the dimensionality of the datasets, increasing interpretability but at the same time minimizing information loss. Then apply different algorithms for the classification of credit scoring. Here uses four algorithms. They are random forest, decision tree, logistic regression and support vector machine.

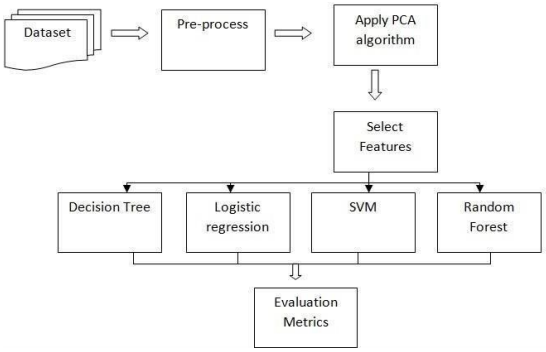

Figure 2: System design

By using Decision tree can split the dataset in different ways based on different conditions. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is binary in nature having data coded as either 1 or 0 and 1 stands for success and 0 stands for failure. Random forest is a supervised learning algorithm which is used for both classification as well as regression. It will create a decision trees on data samples and then gests the

prediction from each of them and finally selects the best solution by means of voting.

It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. Support vector machines are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. By using these algorithms classification will be done on the credit scoring data sets. Then the accuracy of the various algorithm on the credit scoring data sets will be analyzed.

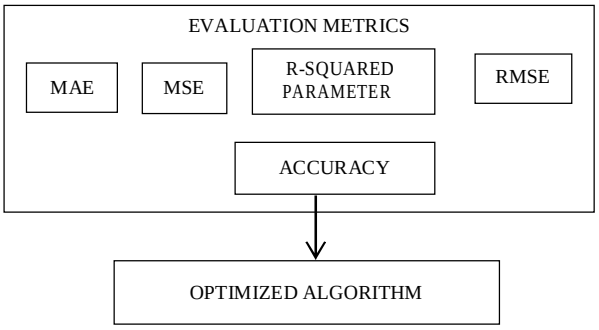

Figure 3: Final Description

The figure 3 shows the final design of the system. The different classification algorithm applied on different credit scoring data sets classification will be done and provide the results. It will include accuracy and error values like MSE value, MAE value, R- Squared parameter, RMSE. By comparing these algorithms SVM will be perform better than other algorithms.

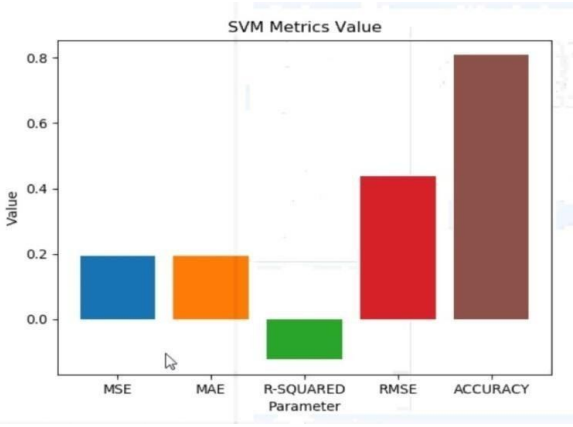

Figure 4: SVM Metrics Value

Each algorithm performs on the data sets and predicted values will be stored. MSE is the average of the squared error that is used as the loss function for leas squares regression. RMSE is the square root of the MSE. MAE is the Mean Absolute Error. It refers to the results of measuring the difference between two continuous variables. R-Squared parameter is the parameter of the model can be found by minimizing the squared error over all the data points. This is also known as the least square loss function.

Here four algorithms are used for classification and accuracy of each algorithm will be shows in evaluation metrics. The accuracy value of Random Forest will be 79%, for Decision Tree 71%, Logistic Regression 79% and SVM 80%. Figure 3.3 shows the Metrics Value of SVM and it will shows that by using SVM for classification get 80% accuracy value. The error values while using SVM will be MSE is 0.19, MAE is 0.19, R-Squared value is -0.123 and 0.438.So use SVM as the optimized algorithm for credit scoring.

4. CONCLUSION

There are two types of automatic credit scoring techniques. They are Statistical techniques and Artificial Intelligence techniques. In contrast statistical techniques, AI methods can be used as an alternative method for credit scoring.AI techniques does not assume certain data distribution. It will be automatically extract knowledge from training samples. Based on study conducted it will be shows that AI techniques superior to statistical techniques in dealing with credit scoring problems, especially for nonlinear problems. The overall best AI techniques can be used for building credit scoring will be depends on the details of the problem, the data structure, the characteristics used, the degree to which it is conceivable to isolate the classes by utilizing those qualities and the objective of the qualification. Here proposes four algorithms. It will be provide optimized algorithm for the classification of credit scoring data sets. The resultant evaluation metrics will shows the results with accuracy, MSE, MAE, R-Squared parameter and RMSE values. By comparing the algorithms SVM will be resulted as an optimized algorithm.

REFERENCES

1. Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu. **Credit rating analysis with support vector machines and neural networks: A market comparative study**, *Decis. Support Syst.*, vol. 37, pp. 543–558, Sep.2004 https://doi.org/10.1016/S0167-9236(03)00086-1
2. N.-C. Hsieh, **Hybrid mining approach in the design of credit scoring models**, May 2005. https://doi.org/10.1016/j.eswa.2004.12.022
3. T.-C. Tang and L.-C. Chi, **Predicting multilateral trade credit risks: Comparisons of Logit and Fuzzy Logic models using ROC curve analysis**, April 2005.
4. S. Lee, C.-C. Chiu, Y.-C. Chou, and C.-J. Lu, **Mining the customer credit using classification and regression tree and multivariate adaptive regression splines**, Feb 2006. https://doi.org/10.1016/j.csda.2004.11.006
5. G.-E. Teng, C.-Z. He, J. Xiao, and X.-Y. Jiang, **Customer credit scoring based on HMM/GMDH hybrid model**, *Knowl. Inf. Syst.*, vol. 36, pp. 731–747, Sep. 2013. https://doi.org/10.1007/s10115-012-0572-z
6. M. Ala'raj and M. F. Abbod, **Classifiers consensus system approach for credit scoring**, *Knowl.-Based Syst.*, vol. 104, pp. 89–105, Jul. 2016. https://doi.org/10.1016/j.knosys.2016.04.013
7. H. He, W. Zhang and S. Zhang, **A novel ensemble method for credit scoring: Adaption of different imbalance ratios**, May 2018.

8. Jasmina Nalić Info Studio d. o. o Sarajevo, Sarajevo, Bosnia i Herzegovina, **Using Data Mining Approaches to Build Credit Scoring Model**, 2018.
9. C. Wang, D. Han, Q. Liu, and S. Luo, **A deep learning approach for credit scoring of Peer-to-Peer lending using attention mechanism LSTM** , IEEE Access, vol. 7, pp.2161–2168,2018
10. Vinodh P Vijayan, Biju Paul **" Traffic scheduling for Green city through energy efficient Wireless sensor Networks "** International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.4, July – August 2019, ISSN 2278-3091 https://doi.org/10.30534/ijatcse/2019/ 81842019.