Credit Risk Assessment using Machine Learning Techniques

Varsha Aithal, Roshan David Jathanna

Abstract*: *Analysis of credit scoring is an effective credit risk assessment technique, which is one of the major research fields in the banking sector. Machine learning has a variety of applications in the banking sector and it has been widely used for data analysis. Modern techniques such as machine learning have provided a self-regulating process to analyze the data using classification techniques. The classification method is a supervised learning process in which the computer learns from the input data provided and makes use of this information to classify the new dataset. This research paper presents a comparison of various machine learning techniques used to evaluate the credit risk. A credit transaction that needs to be accepted or rejected is trained and implemented on the dataset using different machine learning algorithms. The techniques are implemented on the German credit dataset taken from UCI repository which has 1000 instances and 21 attributes, depending on which the transactions are either accepted or rejected. This paper compares algorithms such as Support Vector Network, Neural Network, Logistic Regression, Naive Bayes, Random Forest, and Classification and Regression Trees (CART) algorithm and the results obtained show that Random Forest algorithm was able to predict credit risk with higher accuracy.

Keywords*: *Classification Algorithm, Credit Risk evaluation, Machine learning, supervised learning.

I. INTRODUCTION

Since the last decade, the field of banking risk management has bloomed, and the importance of credit risk evaluation has increased in many sectors. The addition of credit scoring and credit risk evaluation was a major advantage to the banking sector. Credit scoring is a statistical study carried out by the financial institutions and the lenders to predict the potential risk, corresponding to a transaction whereas, credit risk evaluation can be defined as identification of the risk levels associated with the credit transaction as to whether the party will meet the commitment towards the agreed terms.

Credit risk evaluation can be divided into two categories. In the first category, applicants are classified as "good" and "bad" credit risk. This is called application scoring where the data is categorized into groups based on financial data. In the second category, the payment pattern of the applicant along with payment history and other details are considered. This is called behavioral scoring [1]. This paper focuses on application scoring.

Over the last few years, the banking sector has developed several advanced systems in order to assess the credit risk related to a few features of their business. This has led to a better risk calculation and management, which in turn contributed to an effective business transaction. It can be used by a vast number of business associates when the risk is properly assessed.

The objective of this paper is to perform a comparative evaluation of different techniques such as Support Vector Network, Neural Network, Logistic Regression, Naive Bayes, Random Forest, and Classification and Regression Trees (CART) algorithm in order to find out the most accurate system for determining credit risk. With the purpose to evaluate the techniques, they are tested using a real dataset of German credit data, and the obtained results are used to analyze the better technique, that could be used to accurately evaluate the banking sector's credit risk.

The next section of the paper will be as mentioned below. Section II illustrates the literature review. Section III provides a brief introduction to the proposed methodology and techniques used. Section IV presents the implementation of the actual dataset and provides the preliminary results. Lastly, Section V explains the conclusion and future work.

II. LITERATURE REVIEW

In the past few years, different machine learning classification algorithms to evaluate credit risk have been proposed. This section provides a review of different techniques mentioned in the following papers which are used during the evaluation of credit risk.

Due to the rise in popularity of the Basel II necessity, the evaluation of credit risk became important. As a result of the Basel II necessity, there has been an increase in the number of customers applying for loans and competition among financial institutions whether to grant a loan. Li, Shiue and Huang [2] performed the evaluation of the customer loans using Support Vector Machine (SVM) but this was implemented on a data sample of small size.

A non-linear nonparametric model was constructed using machine learning algorithms to forecast the customer credit risk on a small percentage of the sample data. A Classification and Regression Trees (CART) algorithm was used to construct the model, which is a generally used analytical technique wherein an output variable is related to a set of input variables through a series of binary relations recursively were discussed by Khandani, Kim and Lo [3].

Revised Manuscript Received on November 7, 2019.

Varsha Aithal, Dept. of Computer science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. Email: varshaaithal30@gmail.com

Roshan David Jathanna, Dept. of Computer science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. Email: roshan.jathanna@manipal.edu

An effective method was needed to evaluate the credit risk of the customers which was discussed by Devasena [4], Gulsoy and Kulluk [5] and Huang, Liu and Ren [6]. In these papers, the author described the various supervised learning classification algorithm that was implemented on different data sets. Various metrics were used to compare different techniques. Devasena [4] discussed memory-based classifiers such as IBk classifier, Kstar classifier, and LWL classifier, which was implemented in the German credit data set. Gulsoy and Kulluk [5] performed analysis on different techniques such as Random Trees, simple CART, PART, J48, Fuzzy, and NBTrees. The metrics used to evaluate credit risk were precision, number of rules, Kappa statistics, recall, and accuracy. Huang, Liu and Ren [6] stated that the Probabilistic Neural Network (PNN) has minimum error rate and maximum AUC value, the analysis was carried out on Chinese enterprise dataset.

Khashman [7] examined the Emotional neural network (EmNN) model with 12 neurons that were applied to the Australian credit dataset in order to evaluate credit risk. EmNN is based on an emotional back propagation learning algorithm. Wang, Yu and Ji [8] compared the various classification techniques (Random Forest, Naive Bayes, XGBoost and RF-Bagging) using the ensemble model applied on the German credit dataset.

Zhong, Miao, Shen, and Feng [9] stated that one of the significant problems in a corporate credit rating is the management of credit risk, and to address this issue, scorecards are used widely. With the high dependence on user involvement, AI methods like ANN, SVM performs remarkably well in automatic credit rating. They also discussed the various techniques such as BP, ELM, SVM and SLNF that were applied to real financial data which was normalized and preprocessed.

Shukla and Nanda [10] stated that the verification of financial credibility is needed for credit evaluation. And this involves various factors to be assessed, for example social, economic background of a person, which have a mixed datatype and hence cannot be directly verified. As a result, they proposed an algorithm called "parallel social spider algorithm", which is experimented on Japanese, Australian and German datasets. This algorithm solves the problem of credit assessment effectively. The silhouette index that is obtained by this algorithm is greater than the other.

Claderia, Brandao, Campos, and Pereira [11] and Sapozhnikova et. al [12] presented different machine learning techniques like Logistic Regression, Neural networks, Bayesian Networks, and Random Forest, which were evaluated using dataset and gained better economic efficiency.

Soui, Gasmi, Smiti and Ghédira [13] mentioned about Multi-objective Evolutionary algorithm which was used to analyze rule-based credit risk models (SMOPSO, NSAG-II, MOEA / D and SPEA-2) which was evaluated based on 5 performance criteria – Comprehensibility, Fidelity, Accuracy, Scalability, and Generality which were applied on German and Australian data set. This research work analyses the various machine learning classifiers for successful credit risk evaluation.

III. PROPOSED METHODOLOGY

This section provides a brief description regarding the various techniques used to evaluate the model: Logistic Regression (Section III-A), Naïve Bayes (Section III-B), Neural Network (Section III-C), Support Vector Network (Section III-D), Random Forest (Section III-E) and CART (Section III-F) along with methodology described in (Section III-G).

A. Logistic Regression

Logistic regression can be classified as a statistical technique in which a binomial output with explanatory variables can be modeled. Data is made to fit into a linear regression model; a logistic function is used to predict the categorical dependent variable. A binary logistic regression algorithm can predict only two possible outcomes for a categorical response. A multinomial logistic regression algorithm can predict three or more categories without any order. Ordinal logistic regression function is used to predict three or more categories with ordering. In this technique, generalized linear models (GLM) are used, which is designed to execute the generalized linear model regression on the output of binary data [14].

B. Naïve Bayes

Naive Bayes classifiers are a group of classification algorithms built on Bayes Theorem. In general, Naïve Bayes is a type of graphical probabilistic model which could be used to create other models based on data or expert opinion. They are used in various fields for prediction, anomaly detection, etc. They consist of a graph with nodes and directed links between them, where each node represents a variable and the arcs represent the relationship among them [15]. The center of Bayesian learning uses the Bayes theorem which states that given a joint probability distribution over events A and B, then the conditional probability is given by

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

In fraud detection, information about the Naïve Bayes is not available, but the set of variables which are the cause of the frauds can be calculated using the same theorem.

C. Neural Network

Neural Networks (NN) is a part of Artificial Intelligence (AI), which is a learning model whose working is influenced by the functioning of a biological neuron [16]. The neural network comprises nodes, which operate on the input data fed to it and pass on the output to other nodes. The output of each node is defined as the activation or node value. The nodes are associated with weights which by alteration will help the network learn. These weights represent the magnitude to which an input might influence an output. The net linear calculation is implemented in a linear, ramp, move sigmoid, hyperbolic or Gaussian activation function. The Multilayer Perceptron Model is the preferred method for detecting fraud, as it is capable of distinguishing non-linear regions.

Retrieval Number: A4936119119/2019©BEIESP
DOI: 10.35940/ijitee.A4936.119119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

D. Support Vector Network

Support Vector Network (SVN) can be described as a supervised machine learning model that can be used to analyze data, for regression as well as classification analysis, using associated learning algorithms. It is commonly used for classification analysis. Each data element in this algorithm is sketched as a point in the m-dimensional area (m is the count of features) where the cost of each feature corresponds to a specific coordinate [9]. A hyper-plane is found, which best suits to classify the two classes appropriately. It is a selective classifier that is formally defined by an independent hyper-plane. Given supervised training data, an optimal hyper-plane is produced as an output that classifies the data.

E. Random Forest

Random Forest (RF) creates a forest and makes it random is a supervised machine learning algorithm. The bagging method is used to train the collection of Decision Trees known as a forest. Random Forest creates numerous decision trees, combines them to get a stable and accurate classification [8]. The major advantage of the Random Forest algorithm is that it can be applied to both classification and regression analysis.

F. Classification and Regression Trees (CART)

CART is an analytical technique wherein an output variable is related to a set of input variables through a series of binary relations recursively. The recursive relation partitions the independent variable into definite regions wherein the dependent variable is considered as a constant value (classification tree) or it is related to the independent variable (regression tree) linearly. CART is able to detect nonlinear associations among the input variables, which elevates the types of relations that can be captured [3].

G. Methodology

In order to effectively evaluate the credit risk using machine learning classification algorithms, the following architecture is proposed in this paper.
Step 1: The data set is split using feature extraction into training data and testing data.
Step 2: The various classification algorithms such as Support Vector Network, Neural Network, Logistic Regression, Naive Bayes, Random Forest, and Classification and Regression Trees (CART) are applied to the training data to build a training model.
Step 3: A predictive model is built using the test data.
Step 4: The predictive model's output is compared to the model built using trained data.

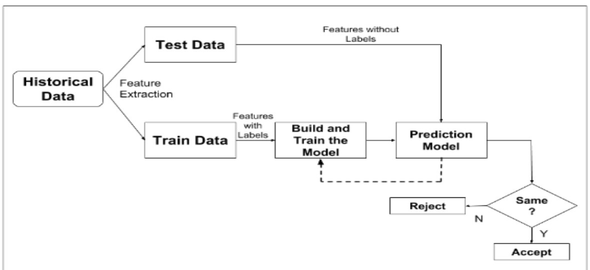

Fig. 1. Flowchart of the proposed methodology

IV. CASE STUDY

This section comprises of an observational study where different classification algorithms such as Random Forest, Naive Bayes, Logistic Regression, Support Vector Network, Neural networks, and CART were applied on the same dataset in order to evaluate the credit risk.

A. Dataset

The dataset used for this research purpose is German credit data [17] which categorizes customers based on the set of attributes, as a "good" or "bad" credit risk. This is an open-source dataset which is available on the UCI Machine Learning Repository [18]. The dataset comprises of 1000 instances of 21 attributes including a classification attribute for each instance. Previously this dataset has been used in credit scoring and credit risk evaluation and that is the reason to choose the same dataset for this analysis. The aim of this research is to predict the outcome of each instance as good or bad using the data set which is classified based on features or attributes, utilizing different machine learning classification algorithms, which are applied on the same data set to compare the accuracy of each of them. A total of 21 attributes are used for this analysis, each instance is characterized by the first 20 attributes and the last attribute is used to classify if a transaction is good or bad.

Table- I: Attributes and class of German credit dataset [17]

ATTRIBUTE NUMBER	DESCRIPTION	CLASS
1)	Creditability	Categorical
2)	Account Balance	Categorical
3)	Credit length (in months)	Numeric
4)	Status of payment	Categorical
5)	Purpose	Categorical
6)	Credit Amount	Numeric
7)	Savings in cost	Categorical
8)	Current employment period	Categorical
9)	Installment	Numeric
10)	Sex and Marital Status	Categorical
11)	Guarantors	Categorical
12)	Current address duration	Numerical
13)	Most precious resources	Categorical
14)	Lifespan	Numeric
15)	Simultaneous loans	Categorical

16)	Type of house	Categorical
17)	Amount of loans from this bank	Numeric
18)	Employment	Categorical
19)	Number of dependents	Categorical
20)	Telephone	Categorical
21)	Foreign Workers	Categorical

The table presents the different attributes and their classes which are either numeric or categorical in nature.

B. Data-Preprocessing

The dataset consists of a mixture of categorical and numeric variables. Categorical variable values are limited and based on a finite group whereas numeric variables can take any value from integer to decimal. The pre-processing of the dataset begins with a characterization of the dataset where the lower significance items were removed, and numerical variables were categorized. The most relevant attribute for credit risk evaluation was selected using Forward Stepwise Regression in the WEKA tool. The comparative gain of every variable was found using InfoGain, available in the WEKA toolbox [19].

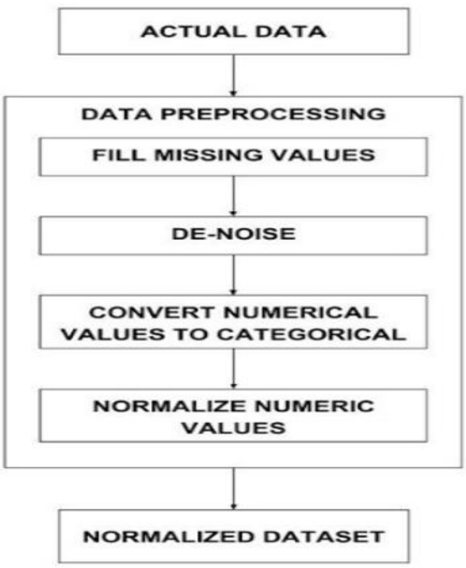

Fig. 2. Structural outline of the data pre-processing

C. Training and Testing dataset

The dataset is split for training and testing purposes. K-Fold-Cross-Validation technique is commonly used due to the higher accuracy. K-Fold-Cross-Validation technique has been used to split the data that operates on multiple percentage divisions, i.e. the data is layered into M folds, in which M-1 folds are used for learning and the M^{th} fold is used for testing of the data. This technique was used to verify the standard of the analysis which was performed by defining the sub-samples number to 10. The system is evaluated by comparing the results of the techniques using various parameters.

D. Analysis and Outcomes

The credit risk assessment techniques were evaluated using different environments and different parameters. These techniques were compared using common measures such as F1- score, specificity, accuracy, sensitivity, error-rate and precision. A brief description of each measure used for evaluating a technique is given below:

True Positive (TP) – classification of good credit risk as good

True Negative (TN) – classification of `bad credit risk as bad

False Positive (FP) – classification of bad credit risk as good

False Negative (FN) – classification of good credit risk as bad

Error rate - can be measured as the count of inaccurate predictions divided by the complete dataset.

$$ERR = \frac{FP + FN}{TP + TN + FN + FP} \quad (2)$$

Accuracy - can be measured as the number of accurate predictions divided by the complete dataset.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \quad (3)$$

Sensitivity/Recall - can be estimated as the count of accurate positive predictions divided by total positives.

$$REC = \frac{TP}{TP + FN} \quad (4)$$

Specificity - can be estimated as the count of accurate negative predictions divided by total negatives.

$$SP = \frac{TN}{TN + FN} \quad (5)$$

Precision - can be calculated as the amount of accurate positive predictions divided by total positive predictions.

$$PREC = \frac{TP}{TP + FP} \quad (6)$$

F1-Score - brings the steadiness between the specificity and recall.

$$F1 - S = \frac{2 * PREC * REC}{PREC + REC} \quad (7)$$

Table- II: Comparative outcome of different machine algorithm applied on German credit dataset

	ERR	**ACC**	**REC**	**SP**	**PREC**	**F1-S**
LR	0.25	0.75	0.91	0.34	0.76	0.83
NB	0.23	0.77	0.78	0.67	0.91	0.84
NN	0.26	0.74	0.80	0.59	0.82	0.81
SVN	0.29	0.76	0.81	0.50	0.80	0.83
RF	**0.22**	**0.78**	0.80	**0.67**	**0.91**	**0.85**
CART	0.23	0.77	**0.93**	0.39	0.78	0.84

Table II summarizes the outcomes of the various techniques used to evaluate the credit risk, which was described in Section III. Random Forest has a low error rate and high precision, specificity, precision, and F1-Score. CART has a good recall rate.

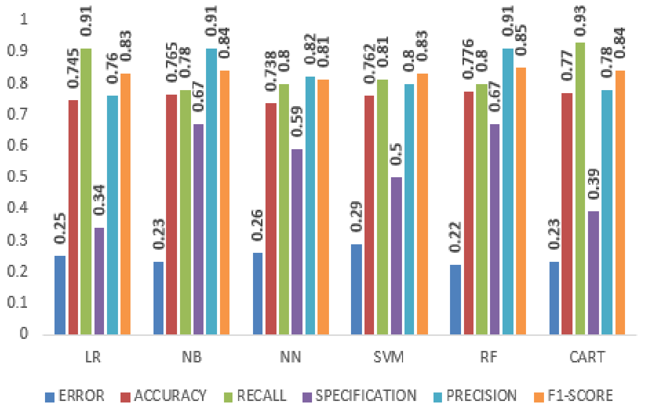

Fig. 3. Measures for different techniques

V. CONCLUSION AND FUTURE WORK

Global markets are full of risks and many attempts have been made to find quick and efficient ways to predict the future. The introduction of credit scores and credit risk evaluation was a major advantage for the banking sector. In this paper, different machine learning techniques were compared to evaluate the credit risk in the German credit dataset. These have been implemented and tested on various classification algorithms such as LR, BN, NN, SVN, RF, and CART. The techniques are tested by applying them on an existing dataset called German credit with thousand transactions per day. From the above analysis, using the Random Forest methodology provides higher accuracy of credit risk evaluation. As future work, various deep learning techniques can be evaluated to see if accuracy increases.

REFERENCES

1. A. Khashman, “Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes,” Expert Systems with Applications, vol. 37, no. 9, pp. 6233–6239, 2010.
2. S. T. Li, W. Shiue,M. H. Huang, "The evaluation of consumer loans using support vector machines." Expert Systems with Applications, vol. 30, no. 4, pp. 772–782, 2006.
3. A. E. Khandani, A. J. Kim, and A. W. Lo, “Consumer Credit Risk Models Via Machine-Learning Algorithms,” SSRN Electronic Journal, 2010.
4. C. L. Devasena, “Adeptness Evaluation of Memory-Based Classifiers for Credit Risk Analysis,” 2014 International Conference on Intelligent Computing Applications, 2014.
5. N. Gulsoy and S. Kulluk, “A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019.
6. X. Huang, X. Liu, and Y. Ren, “Enterprise credit risk evaluation based on neural network algorithm,” Cognitive Systems Research, vol. 52, pp. 317–324, 2018.
7. A. Khashman, “Credit risk evaluation using neural networks: Emotional versus conventional models,” Applied Soft Computing, vol. 11, no. 8, pp. 5477–5484, 2011.
8. N M. Wang, J. Yu, and Z. Ji, “Personal Credit Risk Assessment Based on Stacking Ensemble Model,” IFIP Advances in Information and Communication Technology Intelligent Information Processing IX, pp.328–333, 2018.
9. H. Zhong, C. Miao, Z. Shen, and Y. Feng, “Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings,” Neurocomputing, vol. 128, pp. 285–295, 2014.
10. U. P. Shukla and S. J. Nanda, “Designing of a Risk Assessment Model for Issuing Credit Card Using Parallel Social Spider Algorithm,” Applied Artificial Intelligence, vol. 33, no. 3, pp. 191–207, 2018.
11. E. Claderia, G. Brandao, H. Campos, and A. Pereira, "Characterizing and Evaluating Fraud in Electronic Transactions", Eighth Latin America Web Congress, vol. 12, pp. 115-122, 2018.
12. M. U. Sapozhnikova, A. V. Nikonov, A. M. Vulfin, M. M. Gayanova, K. V. Mironov, D. V. Kurennov, “Anti-fraud system on the basis of data mining technologies” International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 243-248, 2017.
13. M. Soui, I. Gasmi, S. Smiti, and K. Ghédira, “Rule-based credit risk assessment model using multi-objective evolutionary algorithms,” Expert Systems with Applications, vol. 126, pp. 144–157, 2019.
14. Y. Sahin and E. Duman, "Detecting Credit Card Fraud by ANN and Logistic Regression", IEEE, pp. 315-319, 2011.
15. A. Bahnsen, A. Stojanovic, D. Aouada and B. Ottersten, "Cost-Sensitive Credit Card Fraud Detection using Bayes Minimum Risk", 2013 12th International Conference on Machine Learning and Applications, pp. 333-338, 2013.
16. A. Srivastava, M. Yadav, S. Basu, S. Salunkhe and M. Shabad, "Credit Card Fraud Detection at Merchant Side using Neural Networks", 2016 International Conference on Computing for Sustainable Global Development (INDIACom), pp. 667-670, 2016.
17. Lichman, Moshe, “UCI Machine Learning Repository,” http://archive.ics.uci.edu/ml, 2013
18. ttps://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
19. Weka Machine Learning Project, https://www.cs.waikato.ac.nz/ml/weka

AUTHORS PROFILE

Ms. Varsha Aithal is currently pursuing an MTech degree in Computer science and Information Security in the CSE department at Manipal Institute of Technology, Manipal, Karnataka.

Roshan David Jathanna is currently working as Assistant Professor Selection Grade at the Department of Computer Science & Engineering, MIT, Manipal. Prior to his current assignment, he has worked in the industry for seven years as a developer. His areas of interests are machine learning, web services, and big data analytics.