e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

COMPETITIVE ANALYSIS FOR CREDIT RISK ASSESSMENT USING MACHINE LEARNING TECHNIQUES

Ankit Jain*1, Kunal Nandurkar*2, Sanket Vyawahare*3, Amit Nagarale*4

*1Student, Department of Electrical Engineering, MIT Academy of Engineering(Affiliated to SPPU), Pune, Maharashtra, India.

*2Student, Department of Electrical Engineering, MIT Academy of Engineering(Affiliated to SPPU), Pune, Maharashtra, India.

*3Student, Department of Electrical Engineering, MIT Academy of Engineering(Affiliated to SPPU), Pune, Maharashtra, India.

*4Professor, Department of Electrical Engineering, MIT Academy of Engineering(Affiliated to SPPU), Pune, Maharashtra, India.

ABSTRACT

An accurate machine learning tool can help financial institutions increase profits and reduce losses. This paper compares eight types of machine learning model accuracy. While some models are categorized and predict only whether a customer is a defaulter or a non-defaulter such as Extreme Gradient Boosting and some calculates probability of defaulting such as Logistic Regression. Among these eight algorithms Extreme Gradient Boosting provides a higher accuracy of 93% to identify defaulters. In credit risk dataset, classes are unequal and contain a small number of defaulters compared to non-defaulters. Univariate selections, feature importance and Heatmap are used to select the feature. Parameters such as accuracy, precision, recall, f1 scores are used to evaluate each model.

Keywords: Credit Risk, Loan defaulter , Extreme Gradient Boosting, Feature selection.

I. INTRODUCTION

Credit risk is an economic loss arising from a customer's failure to meet timely payment or interest or default on its transaction. Credit risk is an important issue in the financial institutions, including banks, and identification of effective variable involved in that risk, is the most critical element of this type of problem. [4] Therefore, it is important for assessing this risk for banks and other financial organizations in order to avoid economic losses that occur as a result of non-payment.

The machine learning model is used in various fields around the world to solve financial and banking problems. Machine learning gives computers the ability to predict using past data by finding a pattern in the data. Machine learning can be used to recognize the pattern of loan defaulters efficiently and precisely. Logistic Regression, K-Nearest Neighbor, Decision Tree, Naïve Bayes, Support Vector Machine, Random Forest, Multi-Layer Perceptron and Extreme Gradient Boosting will help in identifying a complex pattern in the dataset. By recognizing this pattern, it helps the financial institution to approve a good customer loan application for a profit increase and disapprove the loan application of a bad customer to reduce losses. Credit risk data contains several ineffective and unwanted features that cause data to overfit and reduce the accuracy of the classifier and give wrong results. To avoid this situation, feature selection is used techniques to minimize unwanted features and extract only the essential features. [3]

II. LITERATURE SURVEY

Credit risk has been a research problem in the interests of banks and financial institutions. The lending of credit is a large business venture and a major source of commercial bank's profits [6]. One way this problem can be solved with the help of probability, which is defined as the probability of loss due to the borrower's failure to make the required repayment on the loan [5], but it will not provide us with the key factors involved in the risk. If the risk involved is predicted, then it can greatly save the losses incurred for the banks [4]. The prediction of this risk factor can be made using machine learning algorithms. There are many IEEE papers available on this topic and most of them have used ML methods in their papers. But they do not work very well for some reason. One of the reasons is the improper dataset available for training ML model. This problem directly affects the accuracy of the model, so proper dataset is necessary to in order to find solutions. In referred papers, Neural

networks are used, but they need large datasets and involves more complexity that increases the calculation time [2].

III. DATASET FEATURES

To build our machine learning model, a dataset consists of details of different people with demography. The dataset contains 32581 numbers of rows out of which 25473 are non-defaulters and 7108 are defaulters. The dataset has 12 columns over which feature selection techniques are applied and selected 9 features in which 8 are input features and 1 is the target variable.

A. Dataset Attributes

Table 1 : Dataset Attributes and their values

Attributes	Values
Age	20 to 70
Income	4000 to 703800
Home Ownership	"RENT", "OWN", "MORTGAGE", "OTHER"
Loan Intent	"PERSONAL","EDUCATION,"MEDICAL","VENTURE","HOMEIMPROVEMENT", "DEBTCONSOLIDATION"
Loan Grade	"A", "B", "C", "D", "E", "F", "G"
Loan Amount	900 to 35000
Loan Interest Rate	5% to 23%
Loan Percent Income	0.01% to 0.78%
Loan Status	"0" : Non-Defaulter and "1" : Defaulter

B. Feature Selection

Feature selection is a process by which we select an optimal set of features from input features set using feature selection techniques. By removing unwanted features, the size of the data can be reduced and it can improve the complexity of time and space. Feature selection improves the model performance and saves time and space. [9]

IV. BLOCK DIAGRAM AND METHODOLOGY

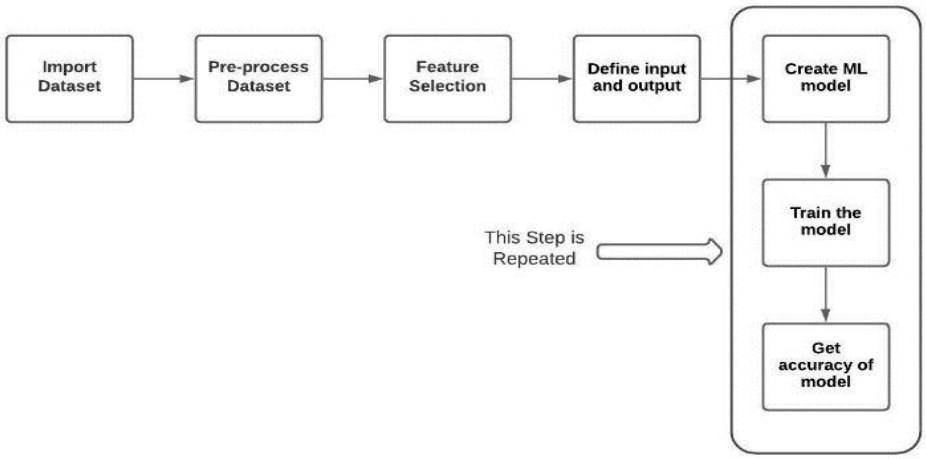

Figure 1: Block diagram

The first step is to import the dataset using the pandas library and then further preprocess the dataset by checking the null values and replace it with the mean or median values of the respective column. The categorical data in the columns is mapped into numerical values. After that the feature selection techniques are applied to the dataset and select only the optimal set of features. The set of input features (X) and the output (Y) are defined in the dataset. The input features are independent of each other and the output feature depends on the input features. Then the library is imported, and the ML Model is created, and the train-split-test method is used to separate the data into training data and testing data and then train the ML Model with training data and predict

using test data. Then the accuracy of the model is calculated by simply taking the ratio of the predicted testing data and the actual testing data. This method is repeated with each ML Algorithm and the accuracy of each algorithm is calculated. Finally, the accuracy of the algorithms is compared and then which of the algorithms is the best for this dataset is determined.

V. ALGORITHMS

A. Logistic Regression

Logistic Regression is a type of supervised machine learning algorithm used to predict binary or multivariate results (dependent variables) using a given set of independent variables. Dependent variations should be binary or multivariate such as "0" or "1", "True" or "False", "High" or "Medium" or "Low", etc. [7]. The Logistic regression does not give the exact value as "0" or "1" but rather the values between "0" and "1" and we get a "S" shaped curve logistic function. We map those values in the range of "0" and "1" using the Sigmoid function. If the value is greater than 0.5, then it is mapped as "1" and if the value is less than 0.5, then it is mapped as "0". [11]

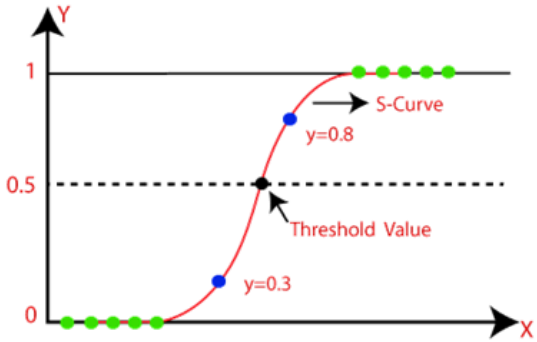

Figure 2 : Graph of "S" shaped curve

equation of the Sigmoid function is

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1$$

Here p is the probability of the event being a success and b_0 and b_1 are parameters.

B. K Nearest Neighbours Classifier

K-Nearest Neighbor is a kind of supervised machine learning algorithm which groups the data based on similarity therefore it is also known as non-parametric classifier [1]. The algorithm checks the similarity between new data and available data and then classifies this new data into the same category as available categories. There are three factors which affect the prediction are, Neighborhood number (k), distance calculation method and division rules. [11]. The KNN Algorithm is used for classification and regression but is mainly used for classification problems. There is a drawback to this algorithm, due to uneven dataset, the major class dominates and thus, the test points mostly falls under the major class category which leads to an error in prediction. This drawback can be solved by multiplying weights to each K Nearest Neighbor to the inverse of the distance from the point to the test point.

Euclidean distance is popular distance measure formula, the formula is given below

$$Distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Where x is new data and y is available data

C. Decision Tree

A supervised machine learning algorithm, which can be used to solve the problems based on regression and classification. It is mostly preferred for classification-based problems. It uses a tree-shaped classifier in which there are nodes present which are connected through the branches. The internal nodes represent the elements of the dataset and the branches represent the decision rule. The leaf node is the output of the decision node.

D. Naïve Bayes

Naïve Bayes is a type of supervised machine learning algorithm based on Bayes theorem that can be used to solve classification-based problems. It is useful for a large dataset due to its fast speed compared to other classification techniques. [1] It uses the Bayes theory of conditional probability which means finding the probability of event occurs given that another event has already occurred.

The Bayes theorem formula is given as

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

Where, A and B are events

P(A|B) = Probability of event A occurring given event B is already occurred

P(B|A) = Probability of event B occurring given event A is already occurred

P(A) = Probability of event A occurring

P(B) = Probability of event B occurring

E. Support Vector Machine

Support Vector Machine is a type of supervised machine algorithm. It is based on the statistical learning concept . It is used for both regression and classification problems. It attempts to define a hyperplane that can classify data in such a way that there is a wide margin between the hyperplane and the observations. By using SVM we can effectively divide data into two categories [10]

F. Random Forest

Random Forest is a type of ensemble learning approach used for both regression and classification problems. Leo Breiman developed this algorithm. [11]. A random forest is usually trained with bagging method. In simple terms random Forest uses many decision trees and combines them together and takes average to obtain accurate results and stable predictions. Random Forest searches for good features between a random subset of features and result as a better model. The random forest takes less training time compared to other algorithms and also maintains accuracy even if a large amount of data is not available.

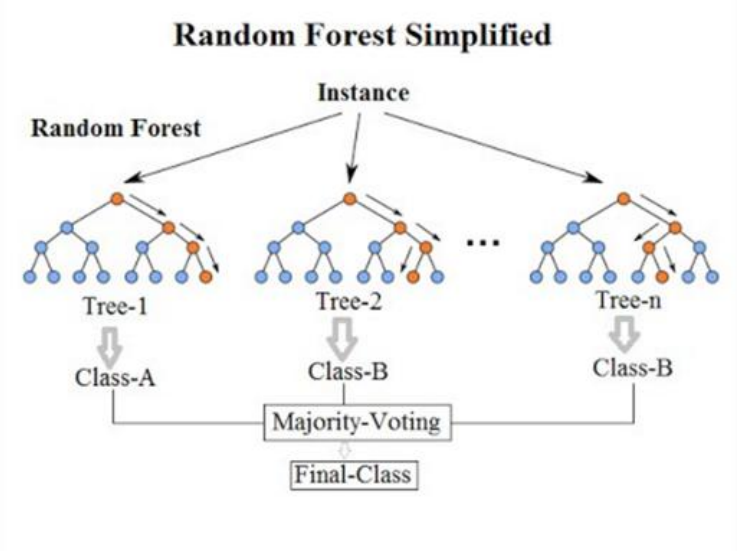

Figure 3 : Explanation of random forest

G. Multi-Layer Perceptron Classifier

The MLP classifier is supervised machine learning algorithm. MLP is a type of feed forward Artificial Neural Network (ANN). The MLP classifier consists of at least three layers the input layer, one or more hidden layers and one output layer. [1] We can change the number of hidden layers to increase complexity depending on our application. The MLP classifier is fully connected which means that each node in one layer is connected to all nodes in the next layer with some weights. In MLP the learning occurs in the perceptron by changing the weights based on the output error compared to the expected results.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

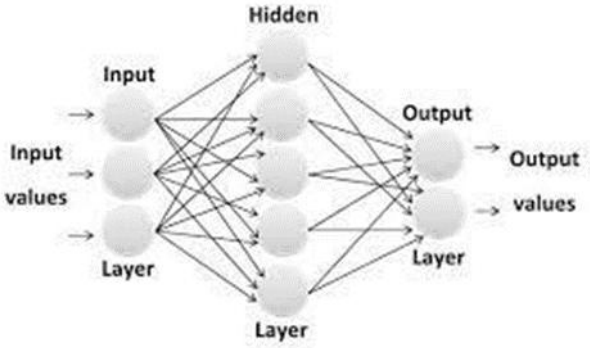

Figure 4 : Neural Network Diagram

H. Extreme Gradient Boosting

Extreme Gradient Boosting (XGB) is a decision tree-based ensemble machine learning algorithm . It uses a gradient boosting network. XGB is one of the gradient boosting techniques. XGB is much faster and has a better model compared to other algorithms. [7]. XGB falls under the boosting techniques and ensemble learning. Ensemble learning uses multiple models to increase accuracy. In boosting techniques error enhancement strategies are attempted to reduce the effects of previous models by adding additional components to the model.

VI. EVALUATION PARAMETER

A. Confusion Matrix

Confusion matrix is an evaluation parameter that describe performance of classifier and provide us error made by classifier and various types of error. There are four types of evaluation indicator which are explain below.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 5 : Confusion Matrix in Machine Learning

True Positive :- If observation is True and our classifier predict True.

False Negative :- If observation is True and our classifier predict False.

True Negative :- If observation is False and our classifier predict False.

False Positive :- If observation is False and our classifier predict True.

B. Classification Report

As discussed in [8] , Classification report gives us Accuracy, precision, recall and F1 score of model.

Accuracy = (TP + TN)/(TP + TN + FN + FP) : Accuracy is calculated as correct defaulters and non-defaulters are predicted divided by total number of defaulters and non-defaulters.

Precision = TP/(TP + FP) :Precision is calculated as correct number of defaulters predicted divided by total numbers of defaulters predicted.

Recall = TP/(TP + FN) : Recall is calculated as correct number of defaulters predicted divided by total number of defaulters.

F1 Score = (2 * Recall * Precision)/(Recall + Precision) : F1 score is weighted harmonic mean of precision and recall. F1 score is used for comparing models not accuracy.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

VII. RESULTS AND DISCUSSION

Based on various confusion matrix and classification report we calculate the model accuracy, precision, recall and F1 Score. We use another comparative analysis using K-fold Cross validation technique. We split the data train-test in ratio of 80:20 .

Logistic regression and KNN algorithm give us same accuracy of 0.837 . Decision tree gives us accuracy of 0.888 while Naive Bayes has accuracy of 0.823 and SVM have accuracy 0.811 . Random forest gives us second best accuracy 0.922 while MLP Classifier has accuracy 0.827 and XGB gives us best accuracy 0.926 .

Table 2: Comparison of accuracy for different Machine Learning Algorithms based on Classification Report

Algorithms	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.804	0.81	0.98	0.89
KNN Algorithm	0.836	0.90	0.97	0.93
Decision Tree	0.890	0.93	0.93	0.93
Naive Bayes	0.823	0.84	0.95	0.89
SVM Algorithm	0.811	0.81	0.99	0.89
Random Forest	0.920	0.92	0.98	0.95
MLP Classifier	0.833	0.86	0.95	0.90
XGB Algorithm	0.930	0.93	0.98	0.95

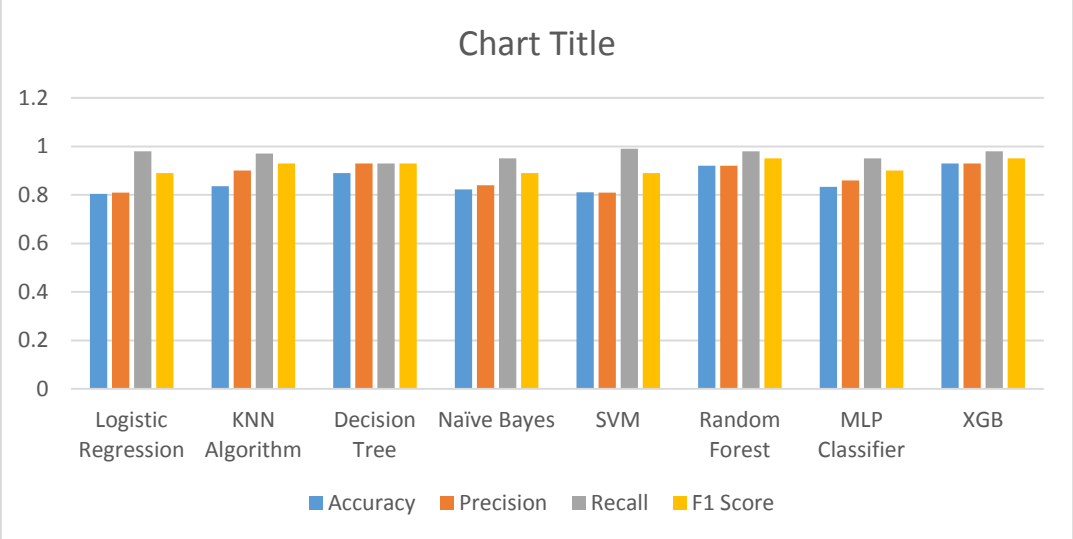

Figure 6 : Comparison of different machine learning algorithms-based classification report parameters accuracy, precision, recall, F1 score

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:05/May-2021 Impact Factor- 5.354 www.irjmets.com

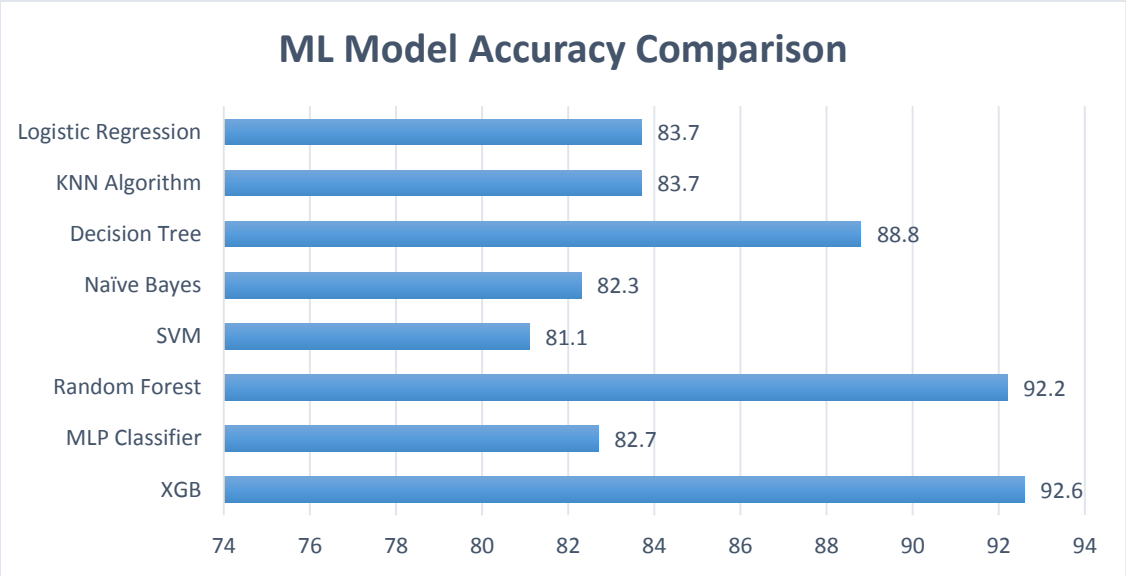

Figure 7 : Model comparison using K Fold Cross Validation Accuracy

VIII. CONCLUSION

We have compared 8 different machine learning algorithms on the credit risk dataset in this paper and XGBoost algorithm gives us best accuracy 0.93. We applied feature selection techniques to select optimal set of inputs to get better accuracy. Apart from using machine learning algorithms. we have also used these methods to find significant factor which impact the status of the loan. As customer don't understand the scientific methods but they can easily get the results from graphical representation.

ACKNOWLEDGEMENTS

We would like to thank Dr Rushikesh Borse, professor, MIT Academy of Engineering, Alandi for motivating and guiding us for this paper presentation.

IX. REFERENCES

[1] A. Bindal and S. Chaurasia, "Predictive Risk Analysis For Loan Repayment of Credit Card Clients," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 2508-2513, doi: 10.1109/RTEICT42901.2018.9012366.

[2] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 2023-2028, doi: 10.1109/TENCON.2019.8929527.

[3] Y. Li, X. Lin, X. Wang, F. Shen and Z. Gong, "Credit Risk Assessment Algorithm Using Deep Neural Networks with Clustering and Merging," 2017 13th International Conference on Computational Intelligence and Security (CIS), Hong Kong, 2017, pp. 173-176, doi: 10.1109/CIS.2017.00045.

[4] A. Mittal, A. Shrivastava, A. Saxena and M. Manoria, "A Study on Credit Risk Assessment in Banking Sector using Data Mining Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-5, doi:

[5] X. Mei and Y. Jiang, "Association rule-based feature selection for credit risk assessment," 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, 2016, pp. 301-305, doi: 10.1109/ICOACS.2016.7563102.

[6] C. Shi and K. Zhang, "Study on Commercial Bank Credit Risk Based on Information Asymmetry," 2009 International Conference on Business Intelligence and Financial Engineering, Beijing, 2009, pp. 758-761, doi: 10.1109/BIFE.2009.175.

[7] Y. Li, "Credit Risk Prediction Based on Machine Learning Methods," *2019 14th International Conference on Computer Science & Education (ICCSE)*, Toronto, ON, Canada, 2019, pp. 1011-1013, doi: 10.1109/ICCSE.2019.8845444.

[8] S. J. Shiv, S. Murthy and K. Challuru, "Credit Risk Analysis Using Machine Learning Techniques," 2018 Fourteenth International Conference on Information Processing (ICINPRO), Bangalore, India, 2018, pp. 1-5, doi: 10.1109/ICINPRO43533.2018.9096854.

[9] N. P. Singh and D. Singh, "Impact of Feature Selection Methods on the Perfromance of Credit Risk Classification Algorithms," 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 2019, pp. 1-6, doi: 10.1109/AICT47866.2019.8981771.

[10] W. Sun, C. Yang and J. Qi, "Credit Risk Assessment in Commercial Banks Based on Support Vector Machines," 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, 2006, pp. 2430-2433, doi: 10.1109/ICMLC.2006.258774.

[11] M. Yan, "Personal Credit Rating System Based on the Logistic Regression Method," 2019 International Conference on Economic Management and Model Engineering (ICEMME), Malacca, Malaysia, 2019, pp. 156-163, doi: 10.1109/ICEMME49371.2019.00040.