International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.14, No.1 (Jul-23)

http://dx.doi.org/10.12785/ijcds/140113

Prediction of Bank Loan Status Using Machine Learning Algorithms

Yakobu Dasari [1], Katiki Rishitha [2] and Ongole Gandhi [3]

[1]*Department of CSE,VFSTR Deemed to be University, Guntur, India*
[2]*Department of CSE,VFSTR Deemed to be University, Guntur, India*
[3]*Department of CSE,VFSTR Deemed to be University, Guntur, India*

Received 23 Jun. 2022, Revised 6 May. 2023, Accepted 8 May. 2023, Published 1 Jul. 2023

Abstract: Major income of banks and any financial organization is generated by loans. Banks can issue loans only to specific authentic people or organizations due to restricted resources or credits. Those who actually can able to repay the taken loan amount along with interest are safe people to whom loan can be sanctioned, but finding eligible (safe) people is a monotonous process. The problem is addressed by various researchers in the literature, however, accuracy level of their models proposed is utmost of 80%. Hence in our work, we proposed a model in which various machine learning algorithms are aggregated with ensemble algorithms like bagging and voting classifiers. The pre-eminent objective of our work is to predict whether a particular person is eligible for the loan or not. Our proposed model reduces human efforts and processing time as well and produces more accurate results than existing models. Experimental results show that our model improves the performance of the existing model from 80% to 94%.

Keywords: Finance, Loan, Machine Learning, Bagging classifier, Voting classifier

1. INTRODUCTION

Sanctioning loans for people is essential for copious purposes like establishing a new organization, setting up businesses, etc. The investor will make a profit from the interest if they repay the loan otherwise, he will be in debt [1]. If the investor fails to predict whether the borrower can repay or not correctly, then he will lose the money. Precise loan prediction is an accepted real-life problem encountered by almost every finance company. Along with risks it also involves lots of human power and time for background verification and for finding safe (eligible) people [2].

The main business of practically all banks is the distribution of loans. The majority of a bank's assets are directly attributable to the profits made from the loans that the bank granted. The main goal in a banking setting is to invest one's money in a secure location. There is no assurance that the applicant chosen will result in the deserving appropriate applicant among all registrants, despite the fact that many banks and other financial giant firms now accept loans following a verification and approval process of regression. We can assess yet if the applicant is correct or not using this approach.

Loan forecasting is quite beneficial for both the applicant and the bank employee. The purpose of this paper is to offer the appropriate candidates a straightforward, quick, and uncomplicated method of selection. It might offer the bank particular advantages. The candidate may have a deadline to determine whether or not their loan will be approved. Jumping to a particular application allows it to be checked first thanks to the Loan Prediction System. No shareholders would indeed be able to change how the full prediction procedure is processed because it is done privately in this paper, which is intended solely for the planning authority of the lender and finance organisation. Reports can be provided to various bank departments in relation to a certain Loan Id such that they can respond appropriately to requests. Other formalities should be completed in all the other departments.

Hence there is a need for a system that automates the entire process with minimal errors. Many researchers proposed various machine learning based systems to address the problem. But from these existing systems, we have observed that almost every system has pros along with some drawbacks. It is found that a major demerit is the low performance i.e accuracy. This may be due to various reasons like ineffective pre-processing techniques or insufficient datasets or even inefficient techniques that might be used for converting categorical values to numerical values. Hence in our we have used effective pre-possessing techniques for filling missing values and dataset of sufficient size is taken and a label encoder is used in categorical to

E-mail address: dy_cse@vignan.ac.in, rishithakatiki@gmail.com, og_cse@vignan.ac.in

numerical conversion. Features of label encoder are unique, Larger number of categories, it is more efficient when order doesn't matter.

In our work, the loan approval process is automated utilizing cutting-edge machine learning technology, which saves time and effort while also speeding up service to borrowers [3][4]. If the borrowers are satisfied with our work, they may recommend us to others which indirectly boosts the profits of the bank. For automation and better performance, we have used a Logistic regression classifier, SVC (support vector classifier), Decision tree, Random forest algorithms along with bagging and voting classifiers [5]. Bagging and voting are ensemble algorithms. The ensemble model uses multiple algorithms which gives better results when compared to stand- alone algorithms. An ensemble algorithm is a supervised technique that finds suitable/appropriate data which gives better predictions [6][7]. An ensemble model is a constant model which gives better results, better forecasting and reduces errors as well [8][9].
.

A. *The following contributions are made in our work:*

i) Investigated the performance of various existing models to predict the loan approval/rejection and found that the utmost performance of the existing model is nearing 80%.

ii) Proposed a model that improves the performance of machine learning algorithms from 80% to 94%.

The subsequent sections of the paper are systematized in this way, section-2 is the literature survey in which contributions of various researchers on the same problem are ventilated in brief. The proposed system is explained in detail in section-3. The overview of the methodology and proposed system is also explicated conscientiously. Results of the proposed system and performance and capabilities of the proposed approach are collated with other systems in section-4 and it is wound up in the conclusion segment .

2. LITERATURE SURVEY

Lin zhu, DajiErgu, Kuiyi Liu, DafengQiu, and Cai Ying proposed that the RF(random forest) machine learning algorithm produced trumped results when analogized to other ML algorithms Decision tree, Logistic regression, and SVC (support vector machine) for predicting loan approval [10]. Random forest not only showed better performance but also strong generalization ability [11]. Their model works for categorical data and numerical data as well [12].

Duan Jing proposed an MLP (Multi-Layer Perceptron) that consists of three layers that are hidden in DNN, trained with the help of an algorithm, back-propagation. The one-hot encoder is employed to metamorphose categorical data to numerical data [13]. SMOTE (Synthetic Minority Over-Sampling Technique) is utilized to balance the imbalanced data as major data belongs to the safe loan class to enhance the prediction accuracy. This proposed model gave better results than the previous single hidden layer MLP [14][15].

For credit data, Arujothi G, and Seethamarai C together proposed a classifier-based machine learning model. Many machine learning algorithms are involved in credit scoring. They have used both K-Nearest Neighbours (KNN) classifier and Min-Max Normalization with R-tool software, which gives higher accuracy than the single ML algorithms [16][17].

Nathan G, Haengjiu L, Shi Zha, and Raj M proposed HMM (Hidden Markov Model) which is statistical for loan approval process automation with the help of historical/previous payments data details from borrowers the probability is predicted. Many HMMs training is done during the training stage. They showed that more accuracy is achieved by training default data separately by segmenting them, if the probability is higher than the threshold then, a signal is sent by the monitoring system [18][19].

Girija A, Radhika M Pai, and Manoharan Pai M used dimensional reduction which considers selecting features and an extraction algorithm to handle the tremendous amount of data(financial data). In their work, they tried to understand the feature extraction along with the transformation algorithm with the help of feature analysis of data. They studied reduced dimension effects on numerous classification algorithms on IBM cloud (Bluemix) with the help of spark notebook, implementation of parallel and distributed is executed. Finally, the proposed enhanced feature reduction accuracy and further, execution time improved the model [20][21].

In order to forecast loan acceptance, Ashlesha Vaidya employed logistic regression like a stochastic and predicting method. The author noted that since they are simpler to create and offer the most effective predictive analysis, artificial neural network and logistic regression were mostly widely utilized for loan prediction. One of the justifications for this is that other algorithms typically perform poorly when trying to forecast from non-normalized data. However, because there is no need that the explanatory variables upon which the forecast takes place have a normal distribution, logistic regression is able to handle the strong positive impact and dynamic factors with ease [22].

In their research article, Mohana Kavya and Tejaswini developed a loan prediction method that automatically determines the weight of each characteristic involved in loan approvals and processes the same features in relation to their associated weight on new test data. Six machine-learning classification strategies have been constructed using R to select the most worthy loan applicants. Decision Trees(DT), Random Forest(RF), Support Vector Machine(SVM), Linear Models, Neural Networks, and Adaboost are some of the models. The decision tree appears to be better on the loan forecasting system and has the highest accuracy of any model, according to the authors' findings [23].

A. *Brief on the existing system:*

Generally, if we see in the case of traditional loan approval, the applicant details have to be checked by the employees of the bank. If the applicant's details meet the criteria, then the loan will be granted otherwise it will be rejected after lots of manpower, and working hours and there is also a good chance of human errors. Many algorithms have been used to increase the performance of the model especially many machine learning algorithms like logistic regression classifier, random forest, etc were used in the existing system. Individually classifiers were implemented on the data and compared but they were unable to handle all the drawbacks and errors and hence don't get much accuracy/efficiency. The below figure1 describes the general flow of activities of the existing systems.

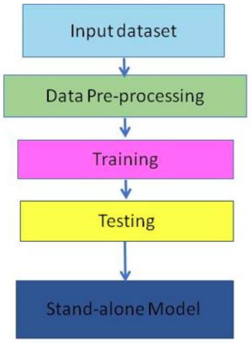


Figure 1. Over view of existing system

B. *Limitations of the existing system:*

i. Lots of manpower, working hours. ii. Chance of human errors. iii. Efficiency/Accuracy not up to the mark.

3. PROPOSED MODEL

To subdue the constraints of the existing system, we proposed a new model by combining various machine learning algorithms that are previously enhanced with an ensemble algorithm [24].

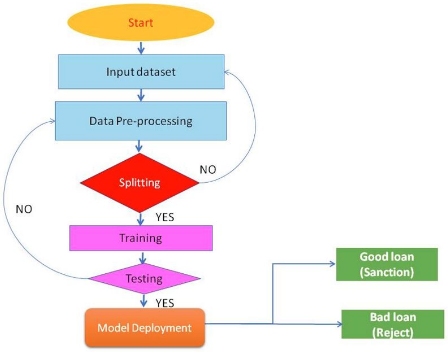


Figure 2. Over view of proposed model

These enhanced models are combined to form a high-performance model to predict the loan approval precisely. After collecting the data that has to be cleaned by pre-processing techniques, train the model with previously available data followed by testing against present data. During training and testing, we have to implement the basic classifiers which are enhanced using a bagging classifier which is an ensemble algorithm next, every enhanced classifier after bagging is given to the voting classifier [25]. The voting classifier is also an ensemble algorithm that takes outputs of multiple classifiers as input and forms the best model which gives the highest accuracy and lowest error rate. Figure 2 and figure 3 gives us the overview and detailed workflow of proposed model respectively.

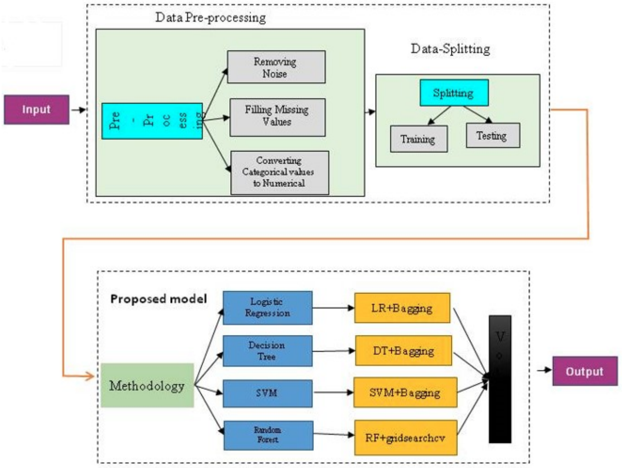


Figure 3. Block diagram of the proposed model

A. *Methodology:*

i. Data Collection: First we have to collect data from large repositories like Kaggle which consists of old or previous loan records [8]. We have collected nearly 1000 records and there are twelve attributes as shown in table I along with their description.

TABLE I. List of all attributes of the loan data set

S.No.	Name of the attribute	Description
1	Loan$_{I}d$	Unique loan Id
2	Gender	Male/Female
3	Marrital status	Applicant married(Y/N)
4	Dependents	No.of dependents
5	Deducation	Graduate/Under-graduate
6	Self$_{e}mployed$	Self employed(Y/N)
7	Applicant$_{i}ncome$	Applicant income
8	Coapplicant$_{i}ncome$	Coapplicant income
9	Loan$_{a}mount$	Amount in thousands
10	Loan$_{a}mount_{t}erm$	Term of loan in months
11	Credit$_{h}istory$	Credit history meets guidelines
12	Propert$_{a}rea$	Urban/Semi-urban/rural
13	Loan$_{s}tatus$	Loan approved(Y/N)

ii. Pre-Processing: The data which we have collected may contain missing values. We have to get rid of those missing values by filling those gaps otherwise it will cause inconsistency. For better results/performance, we have to treat outliers and for better accuracy, categorical values have to be converted into numerical values [8].

TABLE II. List of attributes that may affect the result

S.No.	Name of the attribute	Description
1	Education	Graduate/Under-graduate
2	Self_employed	Self employed(Y/N)
3	Applicant_income	Applicant income
4	Loan_amount	Amount in thousands
5	Loan$_{a}mount\_term$	Term of loan in months
6	Credit_history	Credit history meets guidelines
7	Loan_status	Loan approved(Y/N)

iii. Training and Testing: After collecting and pre-processing our data, we can be able to train and test the model but before that, we have to split the entire dataset into two parts:i. training and ii. testing. For instance,70% data is meant for training purpose and 30% data is meant for testing purpose [8]. The above table II shows the attributes which affect the result. Highly educated applicants have more changes to get their loan sanctioned. People with higher remuneration, a lesser amount of loan, lesser loan term, and who repaid the loan previously have high chances of getting the loan.

B. *Classifiers used:*

We have used Logistic Regression Classifier, SVC, Decision Tree, and Random Forest algorithms. All these algorithms are first bagged and then, the outputs of the bagging classifier are given to the voting classifier as input and finally get a better result than previous methods.

i. Bagging Classifier:
Bagging Classifier is a machine learning algorithm that forms random subsets from the original training dataset. During sample formation, some data may be repeated while other data may be left. Next, the classifier is implemented on every sample in parallel after the predictor (classifier) is trained, then it will predict the required data by aggregating all the predictions of the predictors. The reason behind choosing this bagging ensemble method is that it reduces the variance within the noisy data set. The working of the bagging classifier is as shown in the figure 4. As previously said, after collecting the data set it is making random subsets in the second stage of the diagram, and then each subset is given a classifier individually.

$$R\{(a_n; b_n), n = 1, ...N\}$$

Let "Rs" be the random samples with replacement, where 'b' can be a class or target response 'n' be the number of samples. If 'b' is numerical, we will take the average

$$\emptyset_i(a) = avg_i\emptyset(a, R_i)$$

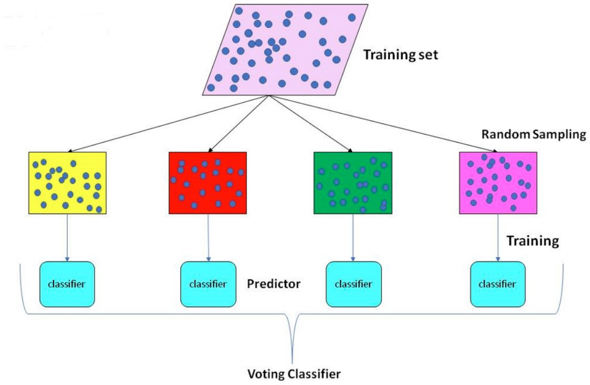


Figure 4. Bagging classifier

ii. Voting Classifier:
Usually, the voting classifier will achieve the highest accuracy than the other best classifiers. There are mainly two types of voting. They are: i. hard voting and ii. soft voting. We are using hard voting, besides hard voting is termed as majority voting. It aggregates each technique's prediction at first, and then it will predict the class by voting that is, which class gets the highest votes(majority votes) will be considered and given as the final result or output. It has non-biased nature and various models can be taken into deliberation. The working of the voting classifier is as shown in the figure 5.

$$b = mode\{r_1(a), r_2(a), ...r_m(a)\}$$

where 'b' is the class label 'r' is the classifier m be the multiple classifiers

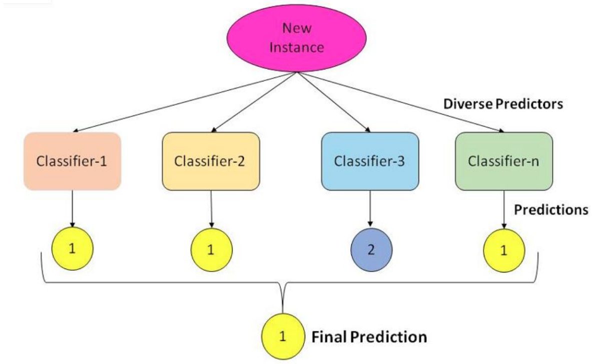


Figure 5. Voting classifier

For example, if three classifiers are classifying a training sample as shown below

C-1 (Logistic Regression + Bagging) L1

C-2 (Decision tree + Bagging) L0

C-3 (SVC + Bagging) L1 Where C= Classifier, L= class

Then class-1(L1) will be the final result based on the below formula as

$$b = mode\{1, 0, 1\} = 1$$

'b'is the majority class label

iii. Logistic Regression Classifier: Logistic Regression is a voguish classification technique that comes under supervised learning. It is employed to prognosticate the target variable's (dependent variable's) probability. The target variable will only have 2(two) classes, for example, yes/no, accept/reject.

$$s(b) = \frac{1}{1 + e^{-}(Q_0 + Q_1 x)}$$

let 's' be the logistic function, β_0 is the intercept, x is the constant

$$0 \leq (\beta_0 + \beta_1 x) \leq 1$$

iv. SVC (Support Vector Classifier): SVC is used for fitting the provided data. Support vector classifier gives the best-fit plane which categorizes our data. After that hyperplane, we have to give some attributes to our classifier to get the predicted class.

v. Decision Tree: A decision tree also a of the supervised strategy present in ML. In this, the whole data split continuously concerning the parameter. Its structure looks like a tree with branches and nodes. Nodes depicts the dataset features and branches depicts the rules. The leaf node delineates the final output.

$$Y(X) = \sum_{i=1}^{n} Y_i * l_i(X)$$

'Y' is the constant value, where 'Yi' is the chosen value from the Ri region, 'Ii' is the indicator function equal to 1.

vi. Random Forest: Random forest is also a supervised tactic that uses legion decision trees. It is used not only for classification but also for regression. A 'RF' is faster than a decision tree for training. One can work easily with many features as it works on the subset. It uses many decision trees for prediction. Random Forest is used along with GridSearchCV. It increases the model performance by finding the optimal hyper parameters. Bypassing all combinations of values into the dictionary and evaluating the using Cross-Validation.

$$thenormfi_s = \frac{fi_s}{\sum_r \in allfeaturesfi_r}$$

$$theRFfi_s = \frac{\sum_r \in alltreesnormfi_s r}{N}$$

'RFfis' is the 'I' assessed from everytree in this model(RF), 'normfisr' be the normalized characteristic (for s in r), and 'N' be the total no. of trees.

4. RESULTS

Our proposed model is implemented on jupyter notebook version 6.4.5 present in the anaconda navigator using windows 10 operating system, with AMD Ryzen 5 processor of 8 GB RAM. We implemented our method on over 1000 records collected from Kaggle, one of the largest repositories which consists of a wide range of datasets. The empirical outcomes manifest that our model yields finer results than the existing one. The following metrics measures the performance of the model.

A. *Performance evaluation metrics:*

i. Accuracy: It is one of the metrics used to evaluate which strategy is the most effective at discovering patterns and correlations among data samples utilizing input or even training data. Accuracy of proposed model is computed using the following formula.

$$Accuracy = \frac{TRP + TRN}{TRP + FLN + TRN + FLP}$$

Where TLP=true positives, TLN =true negatives, FLP=false positives, FLN =false negatives

ii. Precision: It is a metric that measures the consistency of the model's positive predictions, and it is computed using the following formula.

$$Precision = \frac{TRP}{TRP + FLP}$$

Where TRP=true positives, FLP=false positives

iii. Recall: The ability of a system to locate all similar instances in a dataset is referred to as recall and it is determined by using the following formula.

$$Recall = \frac{TRP}{TRP + FLN}$$

Where TRP=true positives, FLN =false negatives

iv. F-score: It is among the most significant ML evaluation measures. It concisely summarises a model's prediction performance by merging two previously opposing metrics, precision, and recall.F-score of proposed model is computed by using the following formula.

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The below table III figure 6 shows the precision values for all the four classifiers before and after bagging and after applying voting classifiers. It is found that the precision value is increased from 82% to 99%.

TABLE III. Precision of four algorithms before & after applying the model

S.No.	Algorithm	Precision (%) before	After bagging	After voting
1	Random Forest	82.58	82.58	99
2	Logistic Regression	96.96	96.96	99
3	Decision Tree	98.48	98.48	99
4	SVC	99.24	99.24	99.25

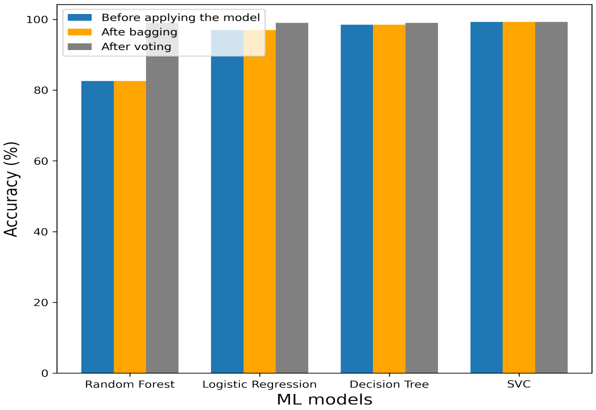


Figure 6. Precision of four algorithms before & after applying the model

The below table IV & figure 7 describes the recall values for all the four classifiers before and after bagging and after applying voting classifiers. It is observed that the recall value increased from 90% to 98%.

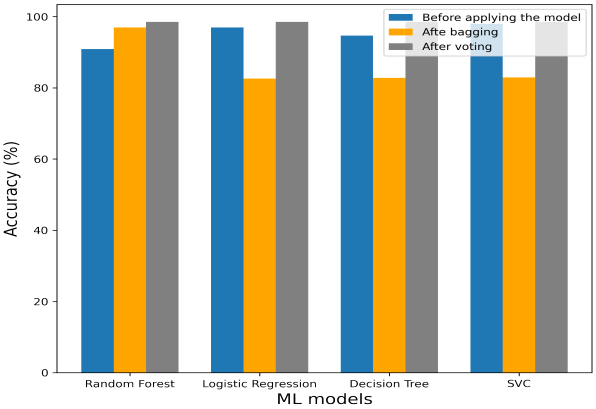


Figure 7. Recall of four algorithms before & after applying the model

TABLE IV. Recall of four algorithms before & after applying the model

S.No.	Algorithm	Recall (%) before	After bagging	After voting
1	Random Forest	90.9	96.96	98.5
2	Logistic Regression	96.96	82.58	98.5
3	Decision Tree	94.69	82.8	98.5
4	SVC	98	82.91	98.5

The table V figure 8 depicts the F1-score values for all the four algorithms before and after bagging and after applying voting classifiers. It is observed that the F1-score value is increased from 87% to 91%.

TABLE V. F1-score of four algorithms before & after applying the model

S.No.	Algorithm	F1-score (%) before	After bagging	After voting
1	Random Forest	87.24	87.24	91
2	Logistic Regression	89.19	89.19	91
3	Decision Tree	88.33	88.33	91
4	SVC	90.24	90.24	91

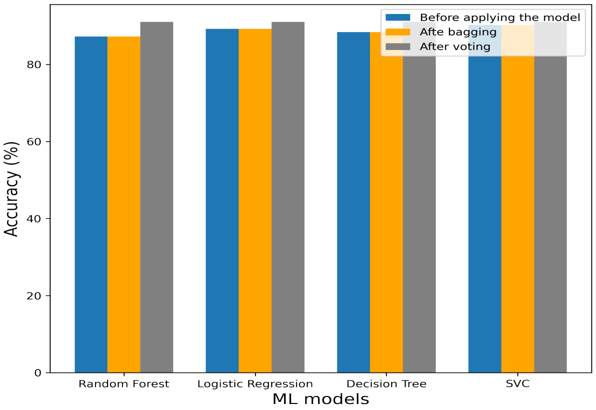


Figure 8. F1-score of four algorithms before & after applying the model

The table VI figure 9 shows the accuracy values for all the four algorithms before and after bagging and after applying voting classifiers. It is found that the accuracy value is increased from 81% to 94%.

TABLE VI. Accuracy of four algorithms before & after applying the model

S.No.	Algorithm	Accuracy (%) before	After bagging	After voting
1	Random Forest	81.08	83.24	94
2	Logistic Regression	83.24	83.78	94
3	Decision Tree	82.16	84.32	94
4	SVC	83.86	84.86	94

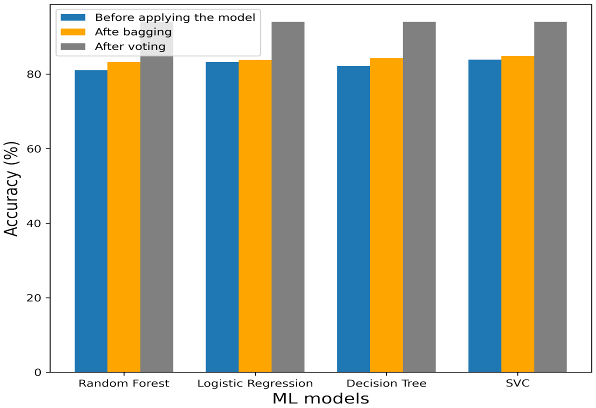


Figure 9. Accuracy of four algorithms before & after applying the model

B. *Results discussion:*

As previously said, after the bagging classifier is implemented on every basic classifier, all the bagged outputs are given as input to the voting classifier. From figure 3, we can see that Logistic Regression(LR), Support Vector Classifier(SVC), Decision Tree (DT), and Random Forest(RF) are given to the voting classifier after the bagging classifier is implemented on them to enhance the accuracy/performance and finally, it is giving 94% of accuracy. Initially, when Logistic Regression, SVC, Decision Tree, and Random Forest are implemented individually they are giving results with nearly 82% accuracy, when the four algorithms implemented with bagging classifier their accuracy is increased to nearly 85%, in order to increase the accuracy, we gave the bagged classifiers to voting classifier therefore, the accuracy has increased to 94%. Along with accuracy, prediction results also printed in a CSV file as ‘vote.csv’ with two columns, one with loanId and another with loan status. Hence, one can be able to easily see which loan is accepted and which is not.

5. CONCLUSION & FUTURE SCOPE

We tried to ameliorate the potential of the online loan approval system by using various algorithms Logistic regression classifiers, SVC (support vector classifier), Decision tree algorithm, and Random forest algorithm present in machine learning with a dataset of nearly a thousand records of people who applied for loan before and achieved 94% accuracy. On all those algorithms bagging classifier is implemented first and then the voting algorithm is applied to increase the performance by rectifying errors. As a future scope, the prediction accuracy can be enhanced by adding new features like more dependents, dependent vs. independent variables to find additional patterns, and interest rates, etc. Usage of neural network frameworks like PyTorch and Tensorflow may also produce better results.

LIST OF ABBREVIATIONS

Below table VII shows the list of abbreviations and their definitions.

TABLE VII. List of abbreviations and definitions

S.No.	Abbreviation	Definition
1	DT	Decision Tree
2	RF	Random Forest
3	SVC	Support Vector Classifier
4	LR	Logistic Regression
5	ML	Machine Learning
6	DNN	Deep Neural Network
7	MLP	Multi-Layer Perceptron
8	SMOTE	Synthetic Minority Over-Sampling Technique
9	KNN	K-Nearest Neighbours
10	RAM	Random Access Memory
11	GB	Giga Byte
12	SVM	Support Vector Machine

References

[1] H. S. Bhat and D. Zaelit, “Forecasting retained earnings of privately held companies with pca and l1 regression,” *Applied Stochastic Models in Business and Industry*, vol. 30, no. 3, pp. 271–293, 2014.

[2] V. B. Djeundje and J. Crook, “Identifying hidden patterns in credit risk survival data using generalised additive models,” *European Journal of Operational Research*, vol. 277, no. 1, pp. 366–376, 2019.

[3] M. A. Sheikh, A. K. Goel, and T. Kumar, “An approach for prediction of loan approval using machine learning algorithm,” in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020, pp. 490–494.

[4] T. Rahul, Y. S. Deepika, T. V. Vivek, and M. Shereesha, “Bank loan prediction using machine learning.”

[5] K. Gupta, B. Chakrabarti, A. A. Ansari, S. S. Rautaray, and M. Pandey, “Loanification-loan approval classification using machine learning algorithms,” in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, 2021.

[6] A. Shaik, K. S. Asritha, N. Lahre, B. Joshua, and V. S. Harsha, “Customer loan eligibility prediction using machine learning,” *JOURNAL OF ALGEBRAIC STATISTICS*, vol. 13, no. 3, pp. 2053–2062, 2022.

[7] A. Attig and P. Perner, “The problem of normalization and a normalized similarity measure by online data.” *Trans. Case Based Reason.*, vol. 4, no. 1, pp. 3–17, 2011.

[8] C. Gomathy, M. Charulatha, M. AAkash, and M. Sowjanya, “The loan prediction using machine learning,” 2021.

[9] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, “Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning,” *Electronic Commerce Research and Applications*, vol. 31, pp. 24–39, 2018.

[10] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, “A study on predicting loan default based on the random forest algorithm,” *Procedia Computer Science*, vol. 162, pp. 503–513, 2019.

[11] M. P. Bach, J. Zoroja, B. Jaković, and N. Šarlija, “Selection of variables for credit risk data mining models: preliminary research,” in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017, pp. 1367–1372.

[12] P. Cho, W. Chang, and J. W. Song, “Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision,” *IEEE Access*, vol. 7, pp. 16 925–16 939, 2019.

[13] M. Malekipirbazari and V. Aksakalli, “Risk assessment in social lending via random forests,” *Expert Systems with Applications*, vol. 42, no. 10, pp. 4621–4631, 2015.

[14] J. Duan, “Financial system modeling using deep neural networks (dnns) for effective risk assessment and prediction,” *Journal of the Franklin Institute*, vol. 356, no. 8, pp. 4716–4731, 2019.

[15] I. Brown and C. Mues, “An experimental comparison of classification algorithms for imbalanced credit scoring data sets,” *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446–3453, 2012.

[16] G. Arutjothi and C. . Senthamarai, “Prediction of loan status in commercial bank using machine learning classifier,” *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 416–419, 2017.

[17] K. Arun, G. Ishan, and K. Sanmeet, “Loan approval prediction based on machine learning approach,” *IOSR J. Comput. Eng*, vol. 18, no. 3, pp. 18–21, 2016.

[18] H. Lee, N. Gnanasambandam, R. Minhas, and S. Zhao, “Dynamic loan service monitoring using segmented hidden markov models,” in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 749–754.

[19] A. Mochón, D. Quintana, Y. Sáez, and P. Isasi, “Soft computing techniques applied to finance,” *Applied Intelligence*, vol. 29, no. 2, pp. 111–115, 2008.

[20] G. Attigeri, M. M. Pai, and R. M. Pai, “Analysis of feature selection and extraction algorithm for loan data: A big data approach,” in *2017 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2017, pp. 2147–2151.

[21] G. Wang, J. Ma, L. Huang, and K. Xu, “Two credit scoring models based on dual strategy ensemble trees,” *Knowledge-Based Systems*, vol. 26, pp. 61–68, 2012.

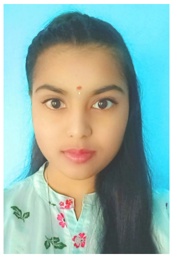
[22] A. Vaidya, “Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval,” *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, 2017.

[23] J. Tejaswini, T. M. Kavya, R. D. N. Ramya, P. S. Triveni, and V. R. Maddumala, “Accurate loan approval prediction based on machine learning approach,” *Journal of Engineering Science*, vol. 11, no. 4, pp. 523–532, 2020.

[24] A. Goyal and R. Kaur, “A survey on ensemble model for loan prediction,” *International Journal of Engineering Trends and Applications (IJETA)*, vol. 3, no. 1, pp. 32–37, 2016.

[25] A. Kim and S.-B. Cho, “An ensemble semi-supervised learning method for predicting defaults in social lending,” *Engineering Applications of Artificial Intelligence*, vol. 81, pp. 193–199, 2019.

Yakobu Dasari, working as Assistant Professor in Vignan’s Foundation for Science, Technology Research (Deemed to be University), Guntur Andhra Pradesh, India. He completed M. Tech at Dept. of CSE, Pondicherry Central University, Puducherry, Currently Pursuing Ph. D at VFSTR, Guntur. He is the member of various professional bodies. His research areas are Cloud Computing, Machine Learning, Deep Learning.



Katiki Rishitha, currently an under graduate student of Vignan’s Foundation for Science, Technology Research (Deemed to be University), Guntur Andhra Pradesh, India. She is the member of various professional bodies. Her research areas are Data Mining, Machine Learning, Deep Learning.

Ongole Gandhi, working as Assistant Professor in Vignan’s Foundation for Science, Technology Research (Deemed to be University), Guntur Andhra Pradesh, India. He completed M. Tech in JNTUV, Vizianagaram, Currently Pursuing Ph. D in JNTUK, Kakinada. He is the member of various professional bodies. His research areas are Data Mining, Machine Learning, Deep Learning.