Credit Approval Decision using Machine Learning Algorithms

Conference Paper · October 2022

DOI: 10.1109/ICRITO56286.2022.9964942

6 authors, including:

Krishna Mridha
Case Western Reserve University School of Medicine
42 PUBLICATIONS **658** CITATIONS

SEE PROFILE

Md Hasanul Kabir
Nanjing Normal University
2 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Ajay vikram Singh
Middle East College
80 PUBLICATIONS **730** CITATIONS

SEE PROFILE

Credit Approval Decision using Machine Learning Algorithms

Krishna Mridha
Computer *Engineering Marwadi University* Rajkot, India
krishna.mridha108735@marwadiuniversity.ac.in

Dipayan Barua
Informaation Technology
Marwadi Univeristy, Rajkot, India
Dipayan.barua108033@marwadiuniversity.ac.in

Meghla Monir Shorna
Computer science and Engineering
Daffodil International University, Dhaka, Bangladesh
meghla25896@gmail.com

Hasan Nouman Nouman
Information Technology
Marwadi University Rajkot, India
hasan.nouman108409@marwadiuniversity.ac.in

Md Hasanul Kabir
Computer Science & Technology
Henan Polytechnic University, China
hasanul.kabir09@gmail.com

Dinesh Kumar
Computer Engineering, Artificial Intelligence & BigData,
Marwadi University Rajkot, India
dinesh.kumar@marwadieducation.edu.in

***Abstract*— The credit approval process for customers seeking a loan has been an important and risky full decision to take for banking institutions. The standard procedure involves a host of information regarding the customer based on what the bankers decide whether to grant a loan which often is time-taking and exhaustive. Therefore, this paper proposes a solution based on machine learning algorithms that reduce risk factors associated with the customer's behaviors that could suggest selecting a trusted person to save the bank's effort and assets. The meth-od utilizes customer information such as monthly income, prior loan history, investment, education, gender, etc., up to twelve attributes that are further used to construct a machine learning model of the decision-making on credit approval. Five different machine learning algorithms named Gradient Booster (GB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (K-NN), and Gradient Booster Classification (GB) are built with a 90% training dataset. Among these, LR has reported the best accuracy as much as 80.43% for the available dataset. We applied these methods to a dataset collected from Kaggle, which contains 615 rows and 13 columns where rows represent customers and columns feature and decisions on credit in 'yes' or 'no'. Future work includes the attempt to improve prediction accuracy with the application deep learning algorithm.**

Keywords—Credit, Loan, Machine Learning, Decision making.

I. INTRODUCTION

Today, people depend on bank loans for a day to meet their needs. The rate of loan applicants has been rising at a very rapid pace in recent years. In the acceptance of loans, the risk is still involved. This paper took data from past customers of the various banks to which loans were made on several criteria that have been authorized. The machine learning technique is then informed on that. To get correct data, enroll. Our key target for this analysis is aimed at predict the protection of the loan [1,3]. The financial authorities are very aware of the clients' payments of the credit sum. The loan approval choices are not correct upon taking many precautionary measures and conducting research on the loan applicant. This mechanism has to be streamlined such that the issuance of loans is less costly and banks experience fewer losses.

Artificial Intelligence (AI) is an evolving technology every day now. The implementation of Ai overcomes many real-world issues. Machine Learning is an AI technique that is very beneficial in structures of prediction. From training data, it generates a model. The model generated by the learning algorithm (which is machine learning) is used when performing the prediction. Using a portion of the overall data, the machine learning model learned the method and checked the remaining data.

It is fascinating for scientists and engineers to make a machine think just like a person, and they aim to accomplish this objective by using artificial intelligence. Machine intelligence is the idea of simulating a machine, human intellect [6]. Artificial Intelligence dates before the formation of computing and has diversified into various areas since that time [6]. In this area, innovators have gained a great comprehension over the years that have contributed to the creation of developed techniques as well as further implementation of these solutions to complex problems.

Concerning this article, one of the principles is used and often extended to a real-world implementation in the area of machine learning. Machine Learning is an instrument that promotes the creation without specific programming of computational models [4]. Data mining algorithms are created to respond to the specifications of the problem. All the leading-edge companies are now using machine learning technologies to gain faster market growth and figure it has been shown that promising outcomes are being achieved. With companies producing such a large amount of data, it becomes impossible to directly exploit data, thus machine learning, as a solution, provides the potential for analytical modeling.

II. LITERATURE REVIEW

The author is familiar with a framework [2] to successfully understand a loan applicant's risk of default. The figures from the predictions indicate that the built model is highly reliable. The concept creation for the forecast is taken up in the history of regression models using the sigmoid activation function as the intended binary answer is either 0 or 1 [10, 11]. The univariate, modified forms, and multiple variance analysis reviews would include the Inner observable

and unobservable variable view [6, 8]. An efficient model is developed to predict the correct borrowers who will have applied for loans were introduced [5]. The Decision Tree is used to foresee the characteristics essential for reputation. The study in [6] research supports credit-scoring vector machine-based models generated using various default meanings. The study concluded that perhaps the comprehensive defined systems in their efficiency are greater than the limited concept variants. K. Gautam [7] introduced a model of cost-benefit analysis for the enforcement of a loan for clients using the Optimization technique. The pre-processing procedure comprises the following sub-processes to perfect the prediction performance: Outlier identification, ranking, exclusion, imputation reduction, and dataset balance through the proportional bifurcation of the method of data preprocessing. Besides, the method of feature selection increases the precision of prediction. The DT system results in 94.3 percent prediction precision when analyzed. Arun, Ishan, and Sangeet [9] recommended a load prediction process utilizing Ml techniques. The sub-processes usually involve the collection of data, feature engineering, preparation, running tests, and parameters for the Production of the outcomes suggested. For the observation and loan calculation process, data sets with 10 characteristics were used. LM, DT, RF, SVM, NN, and Adboost are different ML methods used in the present process. Besides, researchers reported a few essential parameters for different Ml algorithms that play an important role in the loan prediction phase, so that it lets bankers authorize loans to users based on their criteria.

III. Methodology

The suggested model is the estimation of the loan being issued to the clients by the institution. The goal for developing the proposed system is labeling and hence the use for the design of the model, logistic regression with sigmoid function is used. Data pre-processing is the main field of the application where it takes most resources, preceded by Function Technology and then Model Collection, and Data Exploration Review. Feed the algorithm with two different datasets, and then precede the model.

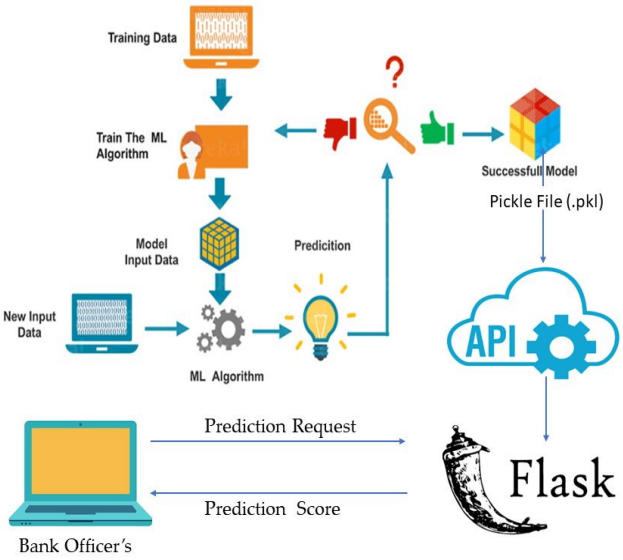


Figure 1: Proposed Architecture of Credit Approval Decision.

A. Data collection

Data sets are drawn from the financial industry. The data collection is in the ARFF format that Weka embraces. The ARFF (Attribute-Relational file format) file consists of tags containing names, attribute forms, values, and the details themselves. We use 13 features for this article, such as gender, marital status, qualification, salary, etc. The table below illustrates the data collection we used.

Table 1: Data set variables along with description and type

Feature Name	Description	Data Type
Loan Id	Unique Id	Integer
Gender	Male/Female	Categorical
Married	Yes/No	Categorical
Dependents	Number of Dependents	Numerical
Education	Graduate/Undergraduate	Categorical
Self-employed	Yes/No	Categorical
Applicant Income	Applicant Income	Numerical
Co-applicant Income	Co-applicant Income	Numerical
Loan Amount	Loan Amount(Thousand)	Numerical
Loan Amount Term	Term of Loan in months	Numerical
Credit History	Credit History meets guidelines	Numerical
Property Area	Urban/Semi Urban/ Rural	Categorical
Loan Status	Yes/No	Categorical

Currently, In the machine learning algorithm, we first introduce the classification models in which the model is developed in this data set of tested samples. The submissions of new candidates will serve as evaluation details to be filled out at the time the application is submitted.

B. Pre-processing:

The methodology of data mining is being used in preprocessing for reshaping raw data collected in an online

portal into raw information in helpful and effective formats. The gathered data includes missing values that may contribute to inconsistencies. To achieve optimal results, data must be pre-processed to boost the algorithm's usefulness. The outliers will have to be removed and then also variable transformation will have to be performed. Encoding processes can minimize the size of the data. Wavelet Converts and approaches for PCA (Principal Component Analysis) Efficient for reductions are.

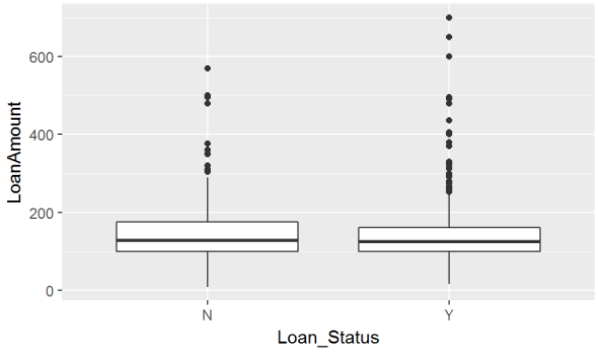
C. *Feature computation:*

An appropriate input dataset that is consistent with the per machine modeling method specifications is configured in function engineering. In our model, the library of Pandas and Numpy was imported to operate. So the reliability of the paradigm of machine learning increases.

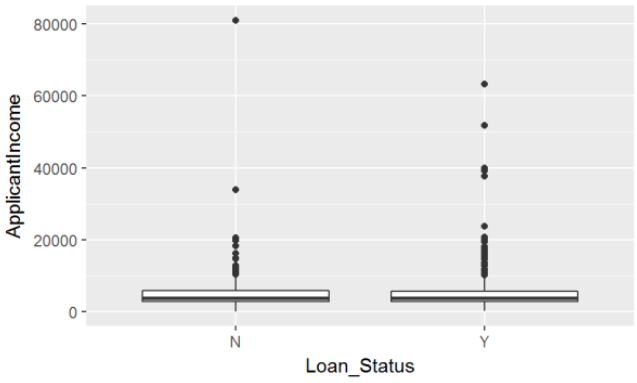
a) Missing Data handling: Missing information is characterized as attributes that are not accessible and that, if identified, will be significant. Missing information may be everything from the missed sequence, defective function, missed files, missing data, the mistake in data entry, etc. This identification approach aims to substitute missing details with factual evaluations of the missing characteristics. Mean, Median or Mode may be used as an approximation of ascription. The conceptual basis of the mean substitution is that the mean from a regular distribution is a rational gauge for an objectively chosen perception.

Now, if we check again the missing value, then we can see that all of the missing values are filled with mean and mode.

b) Handling Outlier: The information is seen visibly to identify the outliers and the outliers are treated later. When the visualized actions of the outliers are of high accuracy and exact. Another statistical tool for finding outliers is percentile rank. It assumed certain percentages of importance in this system, As an outlier, take it from the top or take it from the back. To set the quantitative measure again, the main point is here, and this depends on the nature of the variables as stated above.



(a)



(b)

Figure 2: Showing Outlier of (a) Loan Amount and (b) Applicant Income for Loan_Status

c) One Hot Encoding: For hot encoding, deep learning encoding processes are usually used. The values are divided into single and multiple columns of 0 and 1 values. An association between the encoded segment and the community is seen by all such values. In the machine learning algorithm, in this data set of tested samples, we first add the classification models where the model is created.

from sklearn.preprocessing import LabelEncoder
number=LabelEncoder()

IV. Result and Discussion

A. *Machine learning methods*

The next step is to build a classification algorithm to identify two collections of loan approvals based on a machine learning methodology. All machine learning methods typically rely on a feature matrix created from the data acquired from the method of acquisition. Therefore, many machine learning algorithms have played with the dataset [12, 13], because no rule of thumb dictates the preference of a machine learning algorithm for the available data to be categorized. Gradient booster, quick-to-test machine learning algorithm, vector support machine, logistic regression, K-nearest neighbors, Gaussian Naïve Bayes, and Random for-est, for example, as seen in Table 2.

a) Train the algorithm: Using a different collection of information, the separate category algorithms are trained. For training and the remaining percent for testing the algorithms, the dataset is further split into 90 percent. Also, to select an appropriate sample, each class in the complete dataset is represented in the training and testing datasets in approximately the correct proportion.

b) Testing the algorithm: The 5 algorithms are used on the research dataset to estimate the success of the algorithm. To estimate how accurately these algorithms work in operation on a different collection of data and their related confusion matrices and calculated classification accuracies, cross-validation was performed on the implementations. We follow the widely used metrics in the research when assessing the efficiency of the different classifiers. They require precision, consistency, memory, and specificity. Using Python scikit-learn method with input data as the uncertainty matrix items, these values are

determined. In this article, a "negative" example refers to no (meaning that there will be a default in the loan payment), while the "positive" example refers to yes (meaning there will be a default in the loan payment) (signifying there will not be a default in the payment of the loan).

B. Performance of the algorithms

Machine learning models are built with the necessary training set described above. 10-fold cross-validation is implemented for improved tuning of algorithms' hyperparameters and usability in the method. The results are shown in Table 1 in terms of classification reliability with one set of training and testing sets, the average accuracy of the 10-fold training set and test set, and the standard deviation of the 10-fold training and test set selection process.

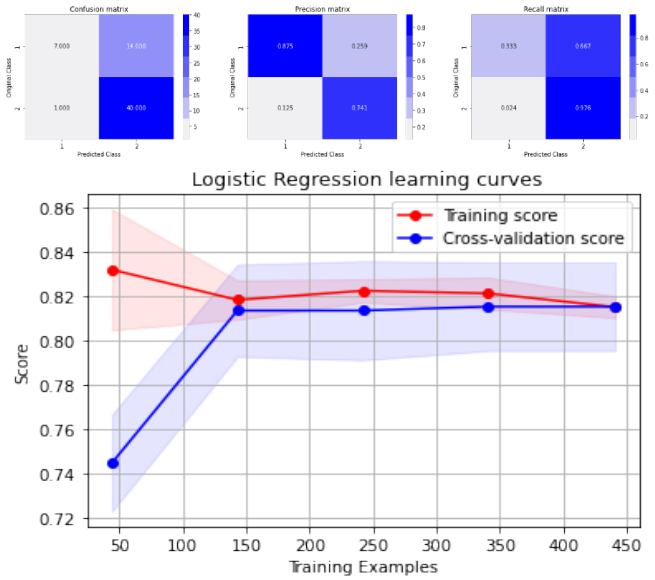


Figure 3: Confusion, precision, Recall, and Learning curve of the Logistic model

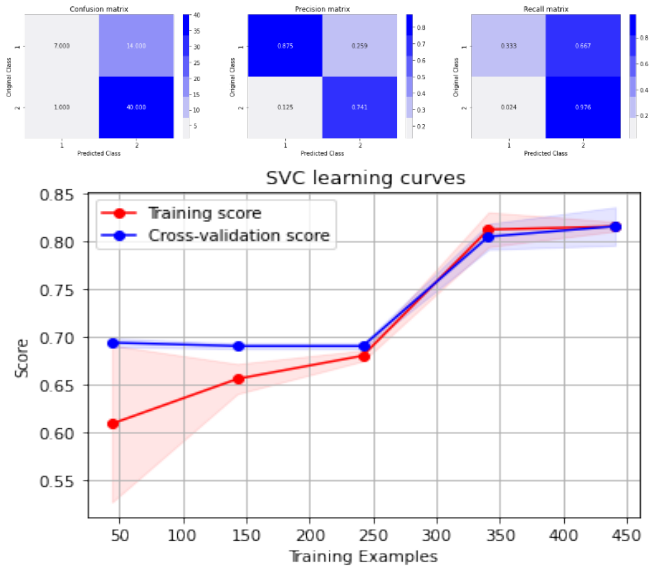
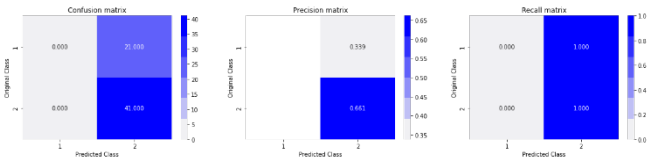


Figure 4: Confusion, precision, Recall, and Learning curve of the SVM model



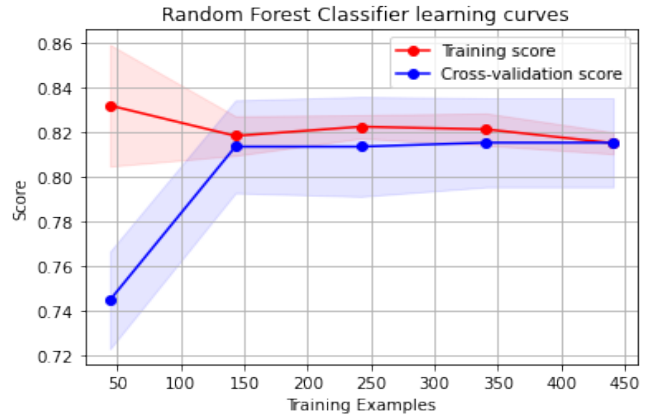


Figure 5: Confusion, precision, Recall, and Learning curve of Lo Random Forest Model

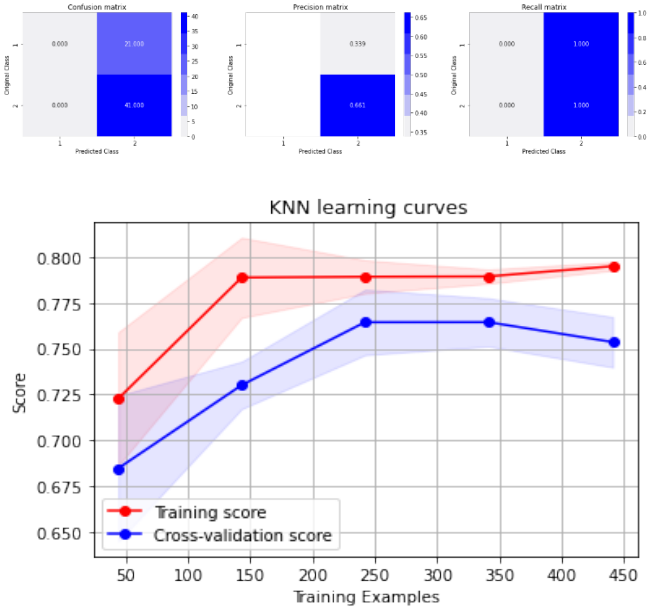


Figure 6: Confusion, precision, Recall, and Learning curve of the KNN model

Table 2.Comparison of between10-folder accuracy vs without cross-validation accuracy

Machine learning algorithms	K-fold Mean Ac (%)	Accuracy (%)	Std (%)
Gradient Booster	80.25	75.80	4.46
Support Vector Machine	81.52	75.80	4.98
Logistic Regression	81.52	80.43	4.98
K-Nearest Neighbors	79.16	74.12	5.06
Gaussian Naïve Bayes	81.15	75.80	5.01
Random Forest	75.88	66.19	7.35

According to the results shown in Table 2, it is evident that the Logistic Regression algorithm performs best with this

data not only 1-fold but also exhibits the least variance while performing k-fold training and test data selection and subsequent training.

Table 3. Performance measures like (a) Precision, (b) Recall, and (c) F1-score.

	Classes	Gradient Booster	Ran-dom Forest	Logistic	Naïve Bays	KNN	SVC
Precision	0	0.75	0.00	0.88	0.88	0.78	0.88
	1	0.76	0.66	0.94	0.74	0.74	0.74
Averaged		0.75	0.33	0.91	0.81	0.76	0.81

(a)

	Classes	Gradient Booster	Random Forest	Logistic	Naïve Bays	KNN	SVC
Recall	0	0.43	0.00	0.33	0.33	0.33	0.33
	1	0.93	1.00	0.99	0.98	0.95	0.98
Averaged		0.68	0.50	0.67	0.67	0.64	0.67

(b)

	Classes	Gradient Booster	Random Forest	Logistic	Naïve Bays	KNN	SVC
F1-Score	0	0.55	0.00	0.50	0.48	0.47	0.48
	1	0.84	0.80	0.95	0.84	0.83	0.84
Averaged		0.70	0.40	0.72	0.66	0.65	0.66

(c)

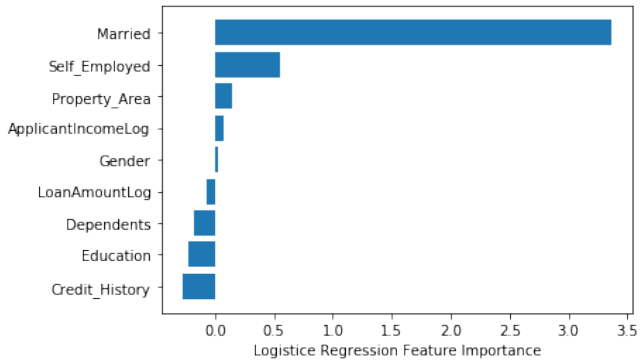
Table 3 shows other performance measures of a classifier like precision, recall, and F1 scores for all different algorithms applied to both classes. The Logistic Regression method shows superiority by yielding the best recall and precision. This suggests that the possibility of type-1 error (i.e., false positive, that shows class2 is wrongly detected as class1), and type-2 error (i.e., false negative, that shows class1 is wrongly detected as class1) are almost NIL. The F1-score, then, which is determined by considering the correlation coefficient of the accuracy and recall, is 0.99.

This is always in researchers' interest to know what features are contributing more and what is less in the process of decision-making of classification. The best classifiers, Logistic Regression, and the worst classifier Gradient booster use all four features mentioned before in their importance as shown in Figures 3 (a) and (b), respectively. The impact on the classification of the included features is estimated based on principal component analysis and feature selection methods that are translated into importance scores. From Figures 3 (a and b), it is evident that in the case of the best classifier married contributes by showing the highest importance score whereas credit history values show the least impact. On the other hand, in the case of the worst classifier Loan AmoutLog contributes by showing the highest importance score whereas property Area values show the least impact.

Here, showing the feature importance of the best and worst performance algorithms.

Logistic Regression feature name and its importance in percentage.

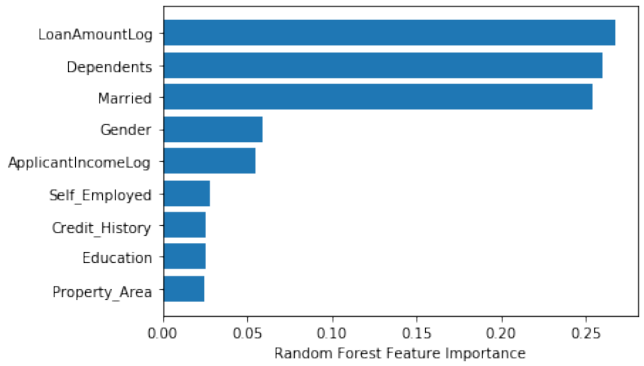
```
Credit_History     :       -0.27%
Education          :       -0.22%
Dependents         :       -0.18%
LoanAmountLog      :       -0.06%
Gender             :        0.02%
ApplicantIncomeLog:         0.08%
Property_Area      :        0.14%
Self_Employed:          0.56%
Married:        3.37%
```



(a)

Random Forest feature names and their importance in percentage.

```
Property_Area      :            0.02%
Education          :            0.03%
Credit_History     :            0.03%
Self_Employed:              0.03%
ApplicantIncomeLog:             0.06%
Gender             :            0.06%
Married            :            0.25%
Dependents         :            0.26%
LoanAmountLog      :            0.27%
```



(b)

Figure 7. Importance metric of the features in classification using: (a) Logistic Regression, (b) Random Forest.

C. User Interface:

By following this article, we developed a web application as discussed earlier in this paper. This particular application has been deployed on Heroku Server. This application flows the Logistic Regression classifier because the Logistic Regression classifier shows the best accuracy to predict the

credit approval decision. Figure 4 picture is the reference for the user interface page.

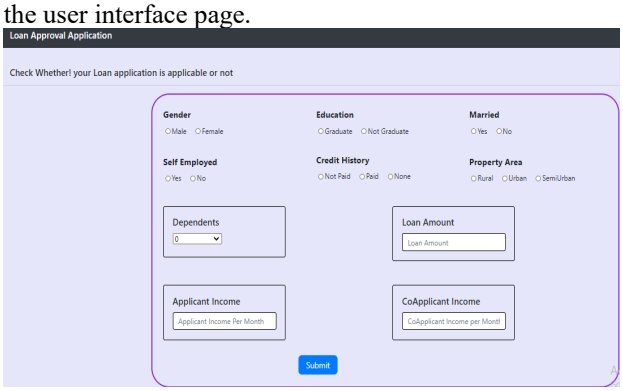


Figure 8: User Interface

When the institution authority clicks to open the application, after opening this application they can see this type of interface. On this page, the user should have to client's previous data which data is already collected from any sources. If we follow this application, there is so much input that we have to give. After giving the input value we have to press the submit button. The submit button takes the new page where we can see the predicting result. The result contains two types of values. One for "Yes" and another for "No".

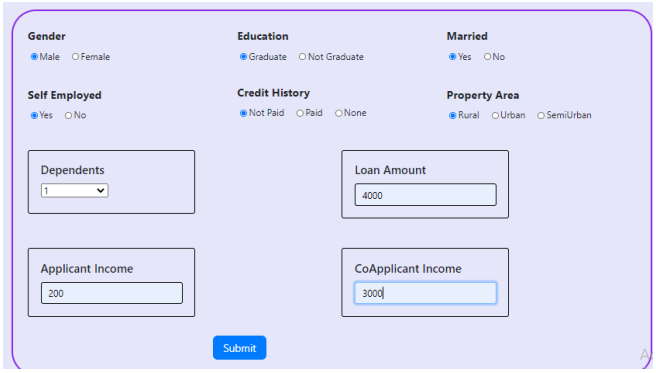


Figure 9: Giving Input

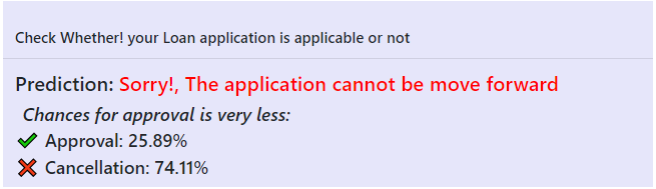


Figure 10: The Result Page

The application will automatically predict the probability of getting credit from particular institutions. This application gives the credit approval and cancellation of both probabilities. If the approval probability is more than 50%, then the machine can predict the positive probability otherwise it will be giving cancellation.

V. CONCLUSION AND FUTURE WORK

In this article, we used a machine learning technique to research the dataset of bank credit to predict the ability to repay customers. To decide which techniques are the best fit for testing bank credit datasets, we used various machine learning techniques on the dataset. The experiment shows that the majority of the algorithms work credibly well in terms of their precision and other algorithms, aside from the k nearest neighbors' quality measurement metrics. The average accuracy of between 74% and over 8% was obtained by both of these algorithms. We have also determined the most significant characteristics that impact customers' financial health. Compared to the case of using all the characteristics, these most significant aspects are then used on certain chosen algorithms and their output precision. No substantial variation in their statistical accuracy and other parameters was shown by the experimental findings. These effects have a lot of repercussions. The system can be used as a method to inform banks as to which variables are relevant in assessing customers' creditworthiness. Besides, the outcome showed that machine learning techniques are not sufficient for the analysis of bank credit datasets. To establish the danger framework of banks, we plan to create a hybrid machine learning model that will integrate the most critical features that assess the ability to repay clients. A detailed review of other supervised learning models other than these three can also be conducted in the future to analyze the strength of loan acceptance forecasting machine learning techniques.

ACKNOWLEDGMENT

This work is carried out under Leadingindia.ai a national initiative of skilling and research of the Government of India on artificial intelligence and its application in various domains of engineering and science.

REFERENCES

[1] P. Supriya1, M. Pavani 2, N. Saisushmam, N. Vimala Kumari4, K. Vikas, " Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques - Volume 5 Issue 2, Mar-Apr 2019, PP.144-147.

[2] R. Kumar, V. Jain, P. Sharma, S. Awasthi4, G. Jha, "Prediction of Loan Approval using Machine Learning", International Journal of Advanced Science and Technology Vol. 28, No. 7, (2019), pp. 455-460

[3] V.S Kumar, A. Rokade, S. MS, "BANK LOAN APPROVAL PREDICTION USING DATA MINING TECHNIQUE", International Research Journal of Modernization in Engineering Technology and Science, Vol. 02, Issue. 05, May-2020, pp. 965-970.

[4] M. Ahmad Sheikh, A. Kumar Goel, T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), pp. 490-494.

[5] M. E. Chandra Blessie, R. Rekha," Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 9 Issue. 1, November 2019, pp. 2714-2719.

[6] N. MADANE1, S. NANDA, "LOAN PREDICTION ANALYSIS USING DECISION TREE", Journal of The Gujarat Research Society, vol. 21, Issue. 14, December 2019, pp. 214-221.

[7] K. Gautam, A. Pratap S.ingh2, K. Tyagi, S. Kumar, "Loan Prediction using Decision Tree and Random Forest", International Research Journal of Engineering and Technology (IRJET), Vol. 07, Issue. 08, Aug 2020, pp. 853-856.

[08] Raj, J. S., & Ananthi, J. V., *"Recurrent neural networks and nonlinear prediction in support vector machine"* Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

[09] X.Frencis Jensy, V.P.Sumathi, Janani Shiva Shri, “An exploratory Data Analysis for Loan Prediction based on nature of client s”, International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-4S, November 2018.

[10] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, K. Vikash, “Loan Prediction by using Machine Learning Models”, International Journal of Engineering and Techniques.Volume 5, Issue 2, Mar-Apr 2019

[11] N. Madane, S. Nanda,” *Loan Prediction using Decision tree”,* Journal of the Gujrat Research History, Volume 21 Issue 14s, December 2019.

[12] G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, *"Machine learning and its Application"*. Advanced Lectures, Springer, 2001.

[13] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow" O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, March 2017