Loan Approval Prediction Using Machine Learning

Harjyot Singh Sandhu, Varun Sharma, Vishali Jassi

Department of Computer Science & Engineering, Lyallpur Khalsa College Technical Campus, Jalandhar

Corresponding author: Varun Sharma, Email: varunsharma@lkcengg.edu.in

As the banking sector improves, many take bank loans; It is difficult to choose the right applicant. When the process is done manually, a lot of misunderstandings can arise in choosing the right applicant. The System uses machine learning algorithms, so the system automatically selects right candidate. This is obtained by extracting Big Data from the previous data of persons those previously granted the loan, to comply that the machine is trained with the help of machine learning algorithms based on these experiences. Prior study from this period has exposed that there are lots of processes for exploration the setback of credit debt control. Although for accurate forecasts, profit maximization is essential; examining the nature of different ways and comparing them is necessary. To this writing, several Machine learning algorithms that have been used in the past are discussed, and their accuracy assessed.

Keywords: Logistic regression, Random forest, Decision tree, Support Vector Machine, Naive Bayes.

2023. In Vikram Dhiman & Pooja Dhand (eds.), *Emerging Trends in Engineering and Management*, 1–6. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-3-5-01

1. Introduction

Credits play a significant role in our daily life. People who are not financially strong can take loans from banks to start up their businesses or for other purposes. Banks get huge profits through loans. Banks have limited goods, so it is essential to choose the right applicant who repays the loan within the time limits. Many people apply for loans it is tough to choose the right one. Selecting the right candidate is the responsibility of the bank. When the process is done manually, many misunderstandings can arise in choosing the right applicant. Thus, we prepare a loan forecast. The system uses machine learning, automatically selecting the right candidate. This model supports both bank employees and applicants. The primary purpose of this model is to find the right ones and reduce the selection time for choosing the right ones. We used algorithms like RF Classifier-Random Forest, LR Classifier-Logistic Regression, Gaussian Naive Bayes, DT Classifier-Decision Tree, and SVM Classifier-Support Vector Machine to make a model. Apply that algorithm which gives more accuracy and gives the correct prediction.

2. Technology used

2.1 Machine Learning (ML): Machine learning is a sub- part of Artificial Intelligence that allows software applications to predict results accurately beyond traditionally programmed to do so. Machine learning algorithms use previous authentic information as input to predict new output. ML is the sector of study that permits computers to gather information without being traditionally programmed. ML is one of the many impressive methodologies that one would ever encounter. As the name suggests, it gives a computer that makes it like human being; the competence to learn. ML will be exercised today, perhaps in many areas than one might hopes. Some real-life problems cannot be solved using the Traditional programming paradigm; for such issues, we have to use machine learning. We must-have ML methods "To make data-driven decisions at scale."
The ML paradigm attempts to include data and expected outcomes. This program or model can then be used in the future to make the essential conclusion and provide likely results from another modern inputs.

3. Algorithms used

3.1 Logistic Regression: LR-classifier is one of the famous ML classifier that has a place with the classification of learning under supervision. It is used for forecasting the categorical subordinate variable using a given set of independent features. Logistic regression gets its name from the function used behind the method, the sigmoid function. LR classifier shapes the opportunity to create unique results with input. For example, you can get something from the most common two-result logistics return models. Two values, such as true/false, yes/no, etc. used to understand statistical programs by guess the link between a subordinate variable and one or more independent features

3.2 Random Forest: RF-classifier is one of the famous ML classifier that has a place with the classification of learning under supervision; it can be used for ML placement return issues. It depends on the perception of collective study. It is a operation of bringing together conglomerate classifiers to solve sophisticated problems and enhance effectiveness. As the name suggests, "random forest" is a classifier with several decision trees. The given dataset takes different subsets and an average value to boost the accuracy of this prediction. On behalf of relying on a single decision tree, a random forest gets a forecast for each tree. It is based on most predictions, predicts and then takes the result.

3.3 Decision Tree: DT-classifier is one of the most famous ML classifier that has a place with the classification of learning under supervision. The decision tree operates the tree portrayal to solve the complication in which each leaf node matches a class label and the attributes on the internal knot of the tree. Root node is the highest decision node in a tree that fits best predictor.

3.4 Support Vector Machine: SVM-classifier is one of the famous ML classifiers that has a place with the classification of learning under supervision. It is mainly used for classification. There we plot each given data as a point in the space of size n (where n is several attributes you have) for the value of each feature to a specific coordinate value. Then, we perform the classification, finding a hyper plane that distinguishes the two classes very well. The support vectors are purely individual observation coordinates. The SVM classifier is a boundary that best separates the two classes. The central idea of SVM-classifier is to find the extreme marginal hyper plane which best categorized collection of data into categories.

3.5 Naive Bayes: NB-classifier is one of the most straightforward and powerful classifier based on the classification of Baye's statement on the hypothesis of independence among forecasters. The NB model is easy and beneficial, especially for massive data sets. The NB classifier assumes that a class has (or does have) a particular feature unrelated to the presence (or absence) of any other attribute assigned to the class variable. This is naive because this is (almost) never true for real-world problems. That is, each feature is an equal contribution to the result.

4 Methodology

4.2 Import all the Essential Python Libraries and import the dataset.
4.3 Before building a model, the model must get through Data Preprocessing, dropping the unwanted columns, and identifying and handling missing values.
4.4 Performs the Exploratory Data Analysis (EDA), Splitting the dataset into training and testing dataset.
4.5 Build the prediction model.
4.6 Apply ML Algorithms to the testing dataset then evaluate the model's accuracy.

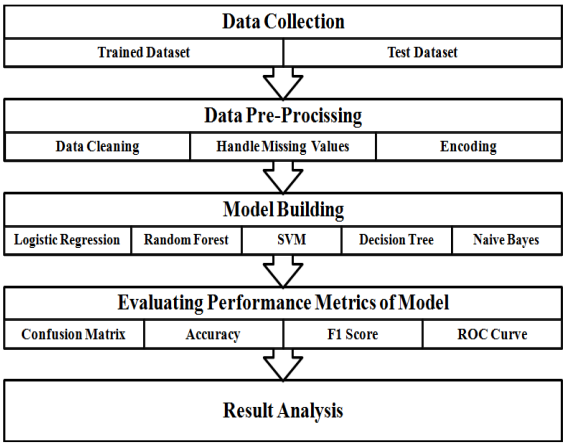

Fig. 1: ML Algorithm

5 About the Dataset

Features Name	Description
Loan_ID	Unique ID
Gender	Applicant Gender
Married	Applicant Marital Status
Dependents	Number of family members
Education	Applicant Qualification
Self_Employed	Applicant Employment Status
ApplicantIncome	Applicant 's monthly salary
CoapplicantIncome	Additional Applicant 's monthly salary
LoanAmount	Loan Amount
Loan-Amount_Term	Loan 's Repayment Period
Credit_History	Record of Previous Credit History
Property_Area	The Location of Property
Loan_Status	Status of Loan

Fig. 2: Dataset Chart

6 Result

	Model	Accuracy	F1_Score
1	Random Forest	99.072356	0.993598
3	SVM	83.302412	0.894614
0	Logistic Regression	82.189239	0.888631
2	Decision Tree	82.003711	0.887861
4	Gaussian NB	79.406308	0.868949

Fig.3 : Model Comparison**:** Comparison Table

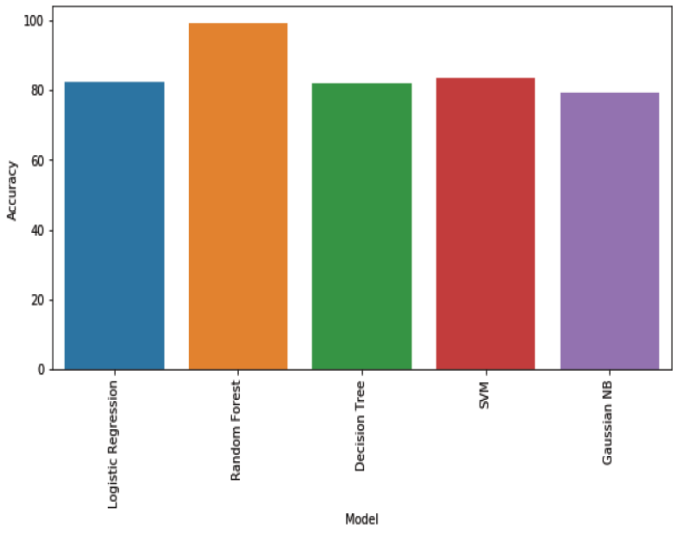

Fig. 4: Plot between Accuracy of Machine learning algorithms for Loan Approval Prediction System:

7 Conclusion

This project used different algorithms like RF Classifier, LR Classifier, Gaussian NB-classifier, DT Classifier, and SVM Classifier to train the model. After that, compare all of these algorithms. Random Forest classifier gives 99% accuracy, Support Vector Machine gives about 83 %, LR classifier gives nearly 82%, DT classifier also gives 82,% and Gaussian Naïve Bayes gives almost 79% accuracy. So from these algorithms, the Random Forest classifier provides more accuracy than all other algorithms. So Random Forest algorithm is best for such types of datasets.

References

[1] Short-term prediction of Mortgage default using ensemble machine learning models, Jesse C.Sealand on july 20, 2018.
[2] Ms. Neethu Baby, Mrs. Priyanka L.T., " Customer Classification And Prediction Based On Data Mining Technique" , International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 12, pp.December 2012.
[3] Alberto, Túlio C, Johannes V Lochter, and Tiago A Almeida. "Tubespam: comment spam filtering on YouTube." In Machine Learning and Applications (Icmla), Ieee 14th International Conference on, 138–43. IEEE. (2015).
[4] C. J. Burges. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2):121–167, 1998.
[5] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai, volume 14, pages 1137–1145, 1995
[6]Machine Learning Engineering" by Andriy Burkov, 2020

[7] Serving Machine Learning Models: A Guide to Architecture, Stream Processing Engines, and Frameworks" by Boris Lublinsky, O'Reilly Media, Inc. 2017
[8] PerfGuard: Deploying ML-for-Systems without Performance Regressions. H M Sajjad Hossain, Lucas Rosenblatt, Gilbert Antonius, Irene Shaffer, Remmelt Ammerlaan, Abhishek Roy, Markus Weimer, Hiren Patel, Marc Friedman, Shi Qiao, Peter Orenberg, Soundarajan Srinivasan and Alekh Jindal
[9]Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists.By Alice Zheng, Amanda Casari.
[10] Programming Fairness in Algorithms. Understanding and combating issues of fairness in supervised learning Towarsdatascience.
11] Do we need hundreds of classifiers to solve real world classification problems, by Amorim, D.G., Barro, S., Cernadas, E., & Delgado, M.F. (2014). Journal of Machine Learning Research.
[12] Trends in extreme learning machines: a review, by Huang, G., Huang, G., Song, S., & You, K. (2015), Neural Networks