

Loan Approval Prediction

Shubham Nalawade¹, Suraj Andhe¹, Siddhesh Parab¹, Prof. Amruta Sankhe²

¹ BE Student, Information Technology, Atharva College of Engineering, Mumbai

² Assistant Professor, Information Technology, Atharva College of Engineering, Mumbai

Abstract – Today a lot of people/companies are applying for bank loans. The core business part of every bank is the distribution of loans. The main objective of the banking sector is to give their assets in safe hands. But the banks or the financial companies take a very long time for the verification and validation process and even after going through such a regress process there is no surety that whether the applicant chosen is deserving or not. To solve this problem, we have developed a system in which we can predict whether the applicant chosen will be a deserving applicant for approving the loan or not. The system predicts on the basis of the model that has been trained using machine learning algorithms. We have even compared the accuracy of different machine learning algorithms. We got a percentage of accuracy ranging from 75-85% but the best accuracy we got was from Logistic Regression i.e., 88.70%. The system includes a user interface web application where the user can enter the details required for the model to predict. The drawback of this model is that it takes into consideration many attributes but in real life sometimes the loan application can also be approved on a single strong attribute, which will not be possible using this system.

Key Words: Machine Learning, Loan Approval Prediction, Web Application, Bank, Algorithms, Random Forest, Naïve Bayes, Logistic Regression, K Nearest Neighbor, Decision Tree.

1. INTRODUCTION

Despite the fact that our banking system has many products to sell, the main source of income for a bank is its credit line. So, they can earn from interest on the loans they credit [1]. Commercial loans have always been a big part of the banking industry, and lenders are always aiming to reduce their credit risk [5]. Nowadays in the market economy banks play a very crucial role. The profit or loss of a bank is largely influenced by loans, i.e., whether the customers repay the loans or default on them [1]. The banks need to decide whether he/she is a good(non-defaulter) or bad(defaulter) before giving the loans to the borrowers. Among the most important problems to be addressed in commercial loan lending is the borrowers' creditworthiness.

The credit risk is defined as the likelihood that borrowers will fail to meet their loan obligations [5]. To predict whether the borrower will be good or bad is a very difficult task for any bank or organization. The banking system uses a manual process for checking whether a

borrower is a defaulter or not. No doubt the manual process will be more accurate and effective, but this process cannot work when there are a large number of loan applications at the same time. If there occurs a time like this, then the decision-making process will take a very long time and also lots of manpower will be required. If we are able to do the loan prediction it will be very helpful for applicants and also for the employees of banks. So, the task is to classify the borrower as good or bad i.e., whether the borrower will be able to pay the debts back or not. This can be done with the help of machine learning algorithms.

2. LITERATURE SURVEY

In [1] they have used only one algorithm; there is no comparison of different algorithms. The algorithm used was Logistic Regression and the best accuracy they got was 81.11%. The final conclusion reached was only those who have a good credit score, high income and low loan amount requirement will get their loan approved. Comparison of two machine learning algorithms was made in [2]. The two algorithms used were two class decision jungle and two class decision and their accuracy were 77.00% and 81.00% respectively. Along with these they also calculated parameters such as Precision, recall, F1 score and AUC. The [3] shows a comparison of four algorithms. The algorithms used were Gradient Boosting, Logistic Regression, Random Forest and CatBoost Classifier. Logistic Regression gave a very low accuracy of 14.96%. Random forest gave a good accuracy of 83.51%. The best accuracy we got was from CatBoost Classifier of 84.04%. There was not much difference between Gradient Boosting and CatBoost Classifier in terms of accuracy. Accuracy of Gradient Boosting was 84.03%. Logistic Regression, Support Vector Machine, Random Forest and Extreme Gradient Boosting algorithms are used in [4]. The accuracy percentage didn't vary a lot between all the algorithms. But the support vector Machine gave the lowest variance.

The less the variance, the less is the fluctuation of scores and the model will be more precise and stable. Only the K Nearest Neighbor Classifier is used in [5]. The process of Min-Max Normalization is used. It is a process of decomposing the attributes values. The highest accuracy they got was 75.08% when the percentage of dataset split was 50-50% with k to be set as 30. In [6] Logistic Regression is the only algorithm used. They didn't calculate the accuracy of the algorithm.

3. DATASET

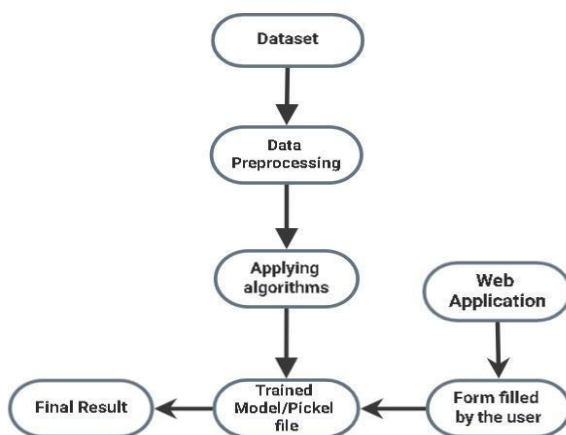
Table -1: Dataset Variables and their Description

Variable Name	Description
Loan_ID	Unique ID
Gender	Male/Female
Marital_Status	Applicant Married (Yes/No)
Dependents	Number of Dependents
Education_Qualification	Graduate/Undergraduate
Self_Employed	Self-Employed (Yes/No)
Applicant_Income	Applicant Income
Co_Applicant_Income	Co-applicant Income
Loan_Amount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Credit History meets guidelines
Property_Area	Urban/Semi urban/Rural
Loan_Status	Loan Approved (Yes/No)

The dataset includes 615 rows and a total 13 variables or columns. In which 12 are independent variables and 1 is dependent variable. There are also some null values in the dataset. These values are filled by using the mean and mode method and also the label encoder is used to convert the string values by 1 and 0. The Loan_Status attribute is the target variable.

4. METHODOLOGY

Fig-1: Flowchart



The proposed system includes a web application with a model trained by using machine learning algorithms deployed in it. There are a total 11 fields in the form which the user needs to fill. The dataset that we have used for training the model also includes 11 attributes. This dataset is pre-processed before using it for training the model. The pre-processing is done by replacing the null values in the dataset with mean and mode method and replacing the string values with 1 and 0 using label encoder.

Then the dataset was divided into two parts: train and test. 90% of the dataset is used for training purposes and 10% is used for testing the accuracy that the model will give for different algorithms. After splitting the dataset different algorithms were applied and each of them gave different accuracy. The best we got was from Logistic Regression i.e., 88%. Once the model is trained a pickle file is created of the model. When the client wants to predict his/her loan approval the client has to first fill a form by visiting our web application.

After filling the form, the user has to just click on the MAKE PREDICTION button and depending on the pickle file or the model that we have trained it will give the result as whether the loan of the customer will be approved or not. As we have also done the comparison of different machine learning algorithms in terms of their accuracy. The web application also includes a bar plot graph of the comparison of algorithms, insights of the dataset that we have used for training the model. This system will make it easier for the banks or organizations to do the job of loan approval prediction.

5. MACHINE LEARNING ALGORITHMS

5.1 Random Forest:

This algorithm is used for both classification as well as regression tasks. The flow of this algorithm is such that it begins by creating multiple decision trees. The final decision of the tree is taken based on the majority of the tree and they are chosen by the random forest. A decision tree is any try diagram which focuses on determining the course of action.

The branch of the tree resembles the possible decision, occurrence or reaction. When there are a lot of sub trees in the forest then this algorithm is used to avoid the overfitting of the model and also reduce the time required for training. It also helps in providing the highest possible accuracy. The advantage of this algorithm is that it can operate productively on large databases and provides highly accurate results by predicting the missing data.

5.2 Naïve Bayes:

This algorithm works best only on classification tasks. It is called a probabilistic classifier because it predicts the outcome based on the probability of the data. It calculates this probability with the help of bayes theorem which is why it is named as Naive Bayes algorithm. The assumption which is taken for the algorithm to work effectively is that all the possible features which will operate needs to show equal and independent contribution to the final outcome. This algorithm works in 3 basic steps:

- 1) Firstly, the data set which is used is converted into a frequency table based on the responses which we have to predict further.
- 2) Create a likelihood table by calculating the probabilities of all the classes in the data set.
- 3) Finally, we apply the Naive Bayesian formula to find the posterior probability for each class. The class which is having the highest posterior probability is the estimated outcome of the required prediction

5.3 Decision Tree:

This algorithm is also used for classification as well as regression problems. This algorithm consists of a tree diagram wherein the leaf nodes resemble the class label and internal nodes specify the attributes. The motive of using the decision tree is to train a model that will help to predict the class or value of the target variable by remembering simple decision tree rules which we get from the training data. Decision tree uses many specific algorithms to split a node into various sub nodes. For the purpose of predicting the class from the dataset, this algorithm begins from the root node of the tree and starts comparing the values of root features with the given dataset.

Based on the comparison result, it follows that branch and jumps onto the next node. It follows the same procedure again until it travels all the way to the leaf node of the tree.

5.4 Logistic Regression:

This algorithm is only used in classification problems. It gives probabilistic values as the outcome i.e., '0' or '1' and 'true' or 'false'. In this, instead of having a regression line as in linear regression, we fit an 'S' shaped logistic function called a Sigmoid Function. The values must be in the range of 0 to 1, which shouldn't go beyond this limit, so it forms a curve like the "S" shape.

For regression tasks, we use linear regression but for classification we use only logistic regression. Certain assumptions we need to consider while working with this

algorithm which says that the dependent variable based on which it predicts, should be categorical in nature. The independent variable (if any) should be independent of each other i.e., it should not have multicollinearity.

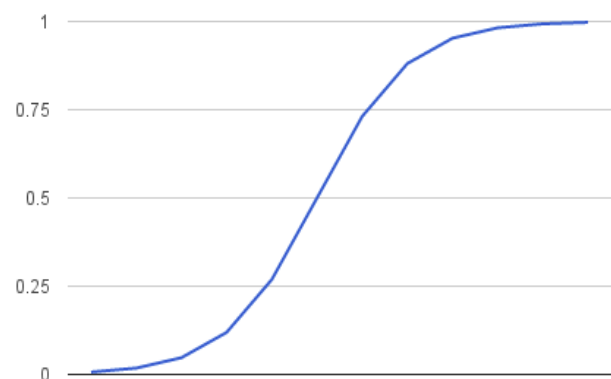
In its core, logistic regression uses a function called the logistic function. In math, the logistic function or sigmoid function is an S-shaped curve that can take any real number and map it into a value between 0 and 1, but never exactly between those limits.

$$1 / (1 + e^{-value})$$

Wherein:

1. 'e' will represent the base in the natural logarithm.
2. 'value' resembles the actual numerical value that we want to transform using the sigmoid function

Utilizing a logistic/sigmoid function, a range of numbers between 0 and 1 can be generated.



The Equation:

Unlike linear regression, logistic regression is represented by an equation, which is extremely similar to the equation for linear regression. For the equation to work, input values are combined linearly using weights or coefficient values. When compared to linear regression, the output value modeled is a binary value rather than a numeric value (0 or 1).

Here is the logistic regression equation:

$$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$$

Where:

- x is the input value

- y is the predicted output
- b_0 is the bias or intercept term
- Coefficient b_1 refers to the value of the single input (x).

A constant real value (the b coefficient) must be learned from each column in the input data to determine the coefficient for that column.

5.5 K Nearest Neighbor:

This algorithm is most commonly used in classification but can be used in regression as well. It depends on the existing labeled input and learns the function to provide the appropriate output when it is fed with new unlabeled data. It stores all the data available and classifies a new data point based on the existing ones in the training dataset.

This algorithm differs from the other algorithms in such a way that it doesn't require any assumption and so it is called a non-parametric algorithm. So, this becomes useful in dealing with nonlinear data. KNN follows 'feature similarity' and so the name is the nearest neighbor. 'k' is the value which is chosen to get the higher accuracy and selecting the right value is very important here. This process is called parameter tuning.

6. RESULT

6.1 Comparative Study of Different Algorithms

We have successfully compared different machine learning algorithms for the Property Loan dataset; they are Random Forest, Naive Bayes, Logistic Regression and K Nearest Neighbors. The Logistic Regression algorithm gave the best accuracy (88.70%).

Table -1: Comparison of Algorithms

Sr. No.	Algorithm	Accuracy
1.	Random Forest	79.03%
2.	Naive Bayes	85.48%
3.	Decision Tree	79.03%
4.	Logistic Regression	88.70%
5.	K Nearest Neighbor	80.64%

6.2 Implementation Output

First, we have our home page where we get information about our system, details of the developers of the system and also a button to go to the prediction page.

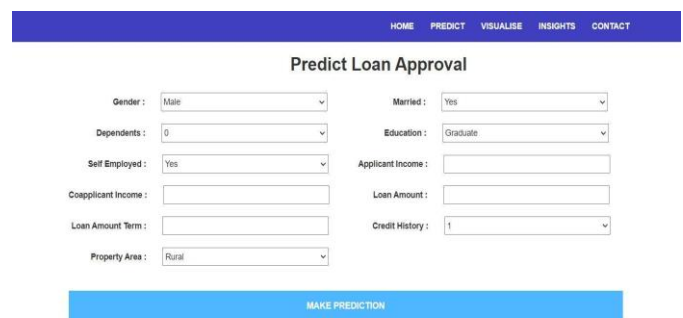


Nowadays in the market economy banks play a very crucial role. The banks need to decide whether he/she is a good(non-defaulter) or bad(defaulter) before giving the loans to the borrowers. To predict whether the borrower will be good or bad is a very difficult task for any bank or organization. The banking system use manual process for checking whether a borrower is a defaulter or not. No doubt manual process will be more accurate and effective, but these process cannot work when there are large number of loan applications at the same time. If there occurs a time like this, then the decision making process will take a very long time and also lots of man power will be required. If we are able to do the loan prediction it will be very helpful for applicant and also for the employees of banks. So the task is to classify the borrower as good or bad i.e. whether the borrower will be able to pay the debts back or not. This can be done using machine learning algorithms.

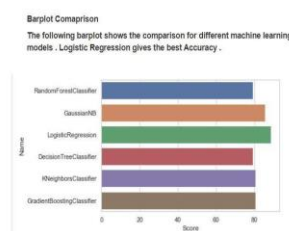
DEVELOPED BY :

SURAJ ANDHE Email Id : suraj.andhe91@gmail.com LinkedIn Github Kaggle	SHUBHAM NALAWADE Email Id : shubham.nalawade03122000@gmail.com LinkedIn Github Kaggle	SIDDHESH PARAB Email Id : parabisiddhesh95@gmail.com LinkedIn Github Kaggle
--	--	--

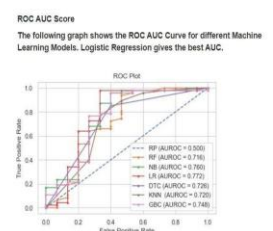
The next is the prediction page where the user can fill the form to check whether he/she is eligible for loan approval or not. It also includes comparison of different algorithms in terms of accuracy in graphical representation.



SNS BARPLOT



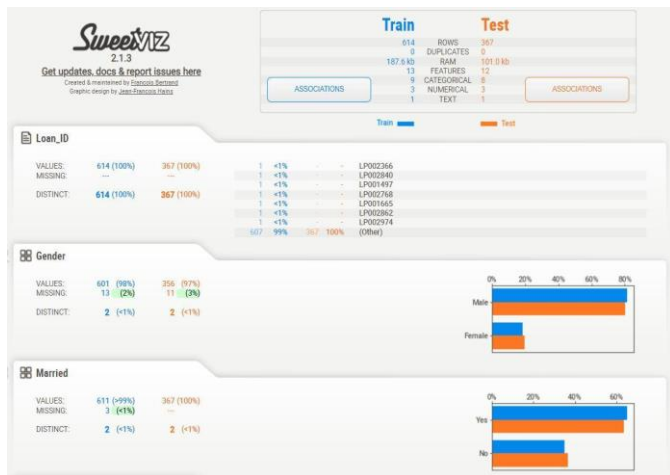
ROC AUC



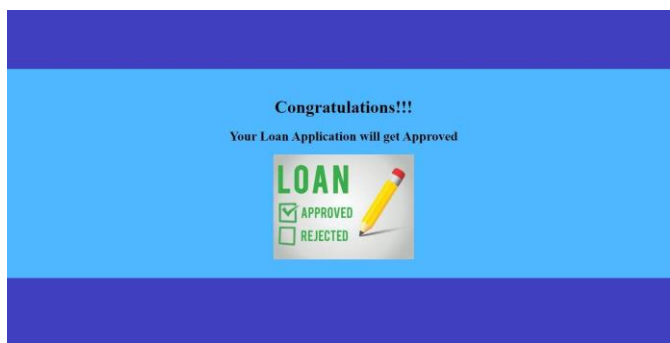
DEVELOPED BY :

SURAJ ANDHE Email Id : suraj.andhe91@gmail.com LinkedIn Github Kaggle	SHUBHAM NALAWADE Email Id : shubham.nalawade03122000@gmail.com LinkedIn Github Kaggle	SIDDHESH PARAB Email Id : parabisiddhesh95@gmail.com LinkedIn Github Kaggle
--	--	--

This page gives the report or analysis of the dataset that we have used to train the model.



The last is the result page where it shows the result of whether the loan application is approved or not.



7. CONCLUSION

For the purpose of predicting the loan approval status of the applied customer, we have chosen the machine learning approach to study the bank dataset. We have applied various machine learning algorithms to decide which one will be the best for applying on the dataset to get the result with the highest accuracy. Following this approach, we found that apart from the logistic regression, the rest of the algorithms performed satisfactory in terms of giving out the accuracy. The accuracy range of the rest

of the algorithms were from 75% to 85%. Whereas the logistic regression gave us the best possible accuracy (88.70%) after the comparative study of all the algorithms.

We also determined the most important features that influence the loan approval status. These most important features are then used on some selected algorithms and their performance accuracy is compared with the instance of using all the features. This model can help the banks in figuring out which factors are important for the loan approval procedure. The comparative study makes us clear about which algorithm will be the best and ignores the rest, based on their accuracy.

REFERENCES

- [1] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.
- [2] K. Alshouli, A. AlGhamdi and D. P. Agrawal, "AzureML Based Analysis and Prediction Loan Borrowers Creditworthy," 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 302-306, doi: 10.1109/ICICT50521.2020.00053.
- [3] B. Patel, H. Patil, J. Hembram and S. Jaswal, "Loan Default Forecasting using Data Mining," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154100.
- [4] S. Z. H. Shoumo, M. I. M. Dhruva, 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), 2019, pp. 2023-2028, doi: 10.1109/TENCON.2019.8929527.
- [5] G. Arutjothi, C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier" 2018 International Conference Sustainable Systems (ICISS)
- [6] Ashlesha Vaidya, "Predictive and Probabilistic approach using Logistic Regression" 2017 8th International Conference on Computing, Communication and Networking Technologies.