Loan Prediction using Machine Learning Algorithms

(Allaparthi Sree Roja Rani,11,cb.en.p2aid20011@cb.students.amrita.edu),
(Anjana M S,12,cb.en.p2aid20012@cb.students.amrita.edu),
(Medarametla Sravani,29,cb.en.p2aid20029@cb.students.amrita.edu)

Abstract: *Loans take up an extremely huge part in the present business world. Checking if a person is truly eligible for a loan is a very big task for all money lenders. Thus this project focusses on the computerization of the process based on various factors like gender, property, credit history, etc. Different machine learning algorithms such as logistic regression, k-nearest neighbors, support vector machine, clustering algorithms and boosting algorithms are applied to check the accuracy of the results in each algorithms, and to find out the algorithm with maximum accuracy. This document is divided into four sections (i) Introduction (ii) Algorithms Used (iii) Experimental Analysis (iv) Conclusion*

Keywords: *logistic regression, k-nearest neighbors, support vector machine, clustering, boosting, accuracy.*

I. INTRODUCTION

With the increased involvement of banking in various sectors, loans have become an irreplaceable part of any business cycle. This is profitable to both small scale businesses and the customers. However, the profit of the bank only depends on whether the customers repay their loans or not. If a bank grants a loan to a person who fails to repay it, then they would have to face several consequences. But if the loan process is too stringent, fewer customers will apply loans. Thus banks have to find a balance of how rigorous the process should be, and at the same time must ensure that the person repays the loan on time. This project helps to computerize and automate the process of checking if a person who has applied for a loan will be able to repay it.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight. A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

DATASET

The dataset used for this project is taken from Kaggle, which is a repository of various datasets for different machine learning problems. The following are the attributes present.

Attribute	Description
Loan_ID	Unique Loan ID
Gender	Male / Female
Married	Applicant married(Y/N)
Dependents	Number of dependents on the applicant(0,1,2,3, >3)
Education	Graduate/Under Graduate
Self_Employed	Self Employed(Y/N)
ApplicantIncome	Income of applicant
CoapplicantIncome	Income of coapplicant
LoanAmount	Amount for loan in thousands
Loan_Amount_Term	Term in months
Credit_History	Score based on previous records
Property_Area	Urban/ Semi urban/ Rural
Loan_Status	Loan approved(Y/N)

In the above attributes, Loan_Status is the target variable.

II. ALGORITHMS USED

The following machine learning algorithms are used for predicting if the loan is approved or not.

Logistic Regression: Logistic Regression uses a statistic model that uses a logistic function to predict a categorical dependent variable based on values of independent variable. Its parameters can be predicted by maximum likelihood estimation. The formula that is used for building a logistic regression model is $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

SVM: SVM is a supervised learning algorithm which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). SVM is based on the idea of finding a hyperplane that best separates the features into different domains. The points closest to the hyperplane are called as the support vector points.

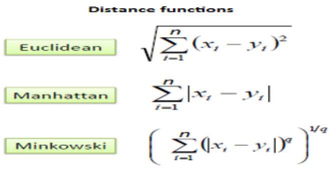
Decision Tree: A decision tree is a tree-shaped diagram that people use to determine a course of action. In a decision tree, each branch represents a possible decision, occurrence or reaction. Simply, Decision tree is a graph to represent choices and their results in form of a tree. For new set of predictor variable, we use this model to arrive at a decision on the category (Yes/No) of the data. The entropy is calculated as :

$$\sum_{i=1}^{C} -f_i \log(f_i)$$

Random Forest: The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. The information gain is calculated as:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right})$$

K Nearest Neighbors: It classifies a data point based on how its neighbors are classified. K is a parameter that refers to the number of nearest neighbors to include in the majority voting process. It is calculated using distance measures based on below functions:



Distance functions

Euclidean	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{n}\lvert x_i - y_i \rvert$
Minkowski	$\left(\sum_{i=1}^{n}(\lvert x_i - y_i \rvert)^q\right)^{1/q}$

Naïve Bayes: A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. It is based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The formula used for classification is $P(A/B) = \frac{P(B/A)*P(A)}{P(B)}$.

Hierarchical Clustering: It is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram. Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

K-Means: K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The *‘means’* in the K-means refers to averaging of the data; that is, finding the centroid.

AdaBoost: AdaBoost algorithm**,** short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instances.

XGBoost: XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models.

Gradient Boosting: It is machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

III. Experimental Analysis

The dataset was trained and tested by various machine learning models, both supervised and unsupervised. The correlation matrix is represented in the form of a heat map as shown below in figure [1]. The feature importance of each attribute present in the dataset is represented in the form of a graph. A hierarchical clustering with method type 'ward' is displayed below in figure [3]:

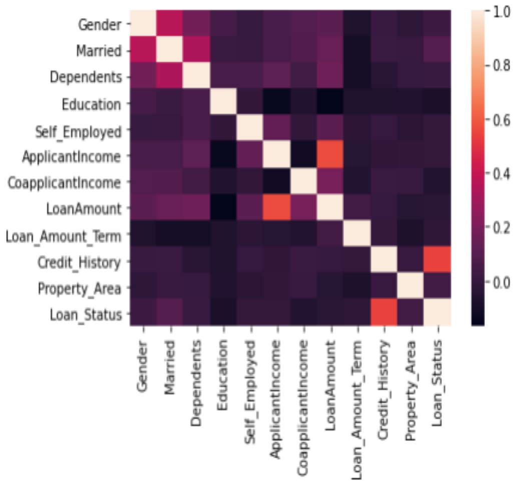


Figure [1]

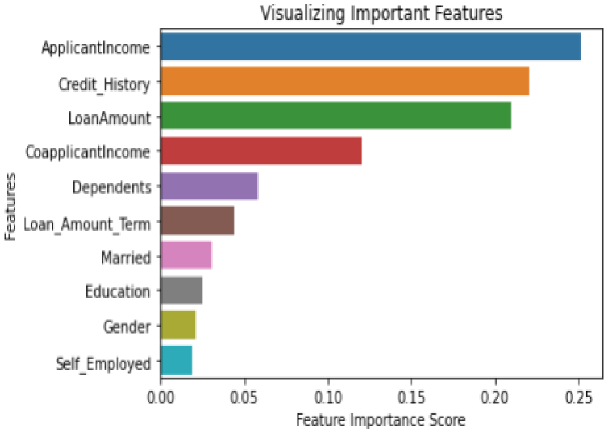


Figure [2]

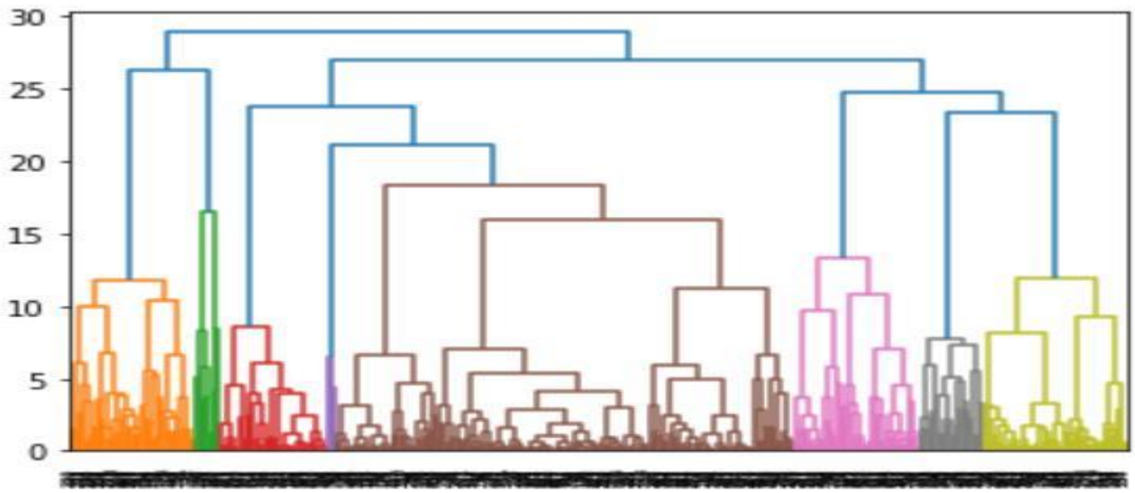
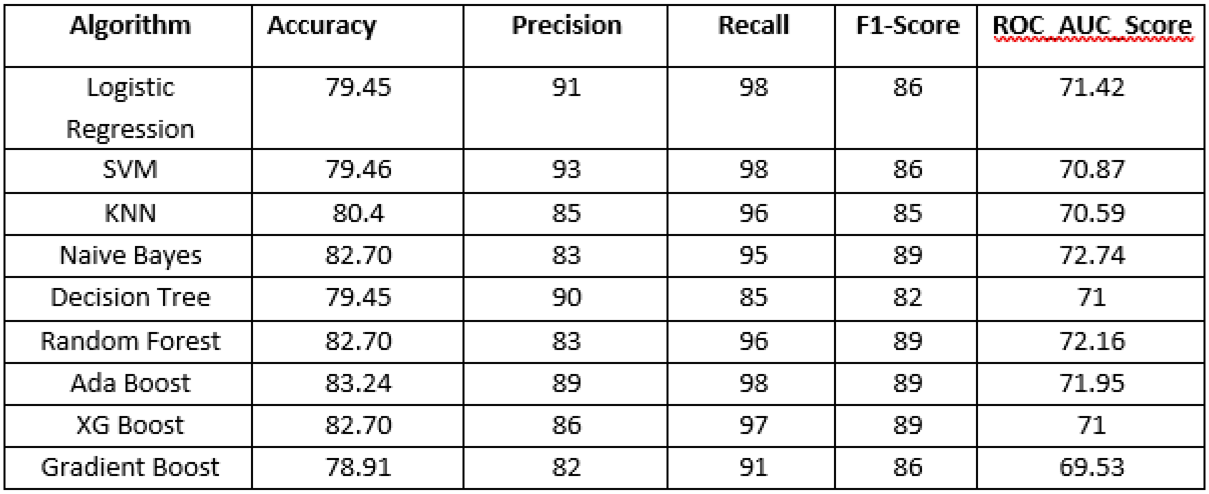


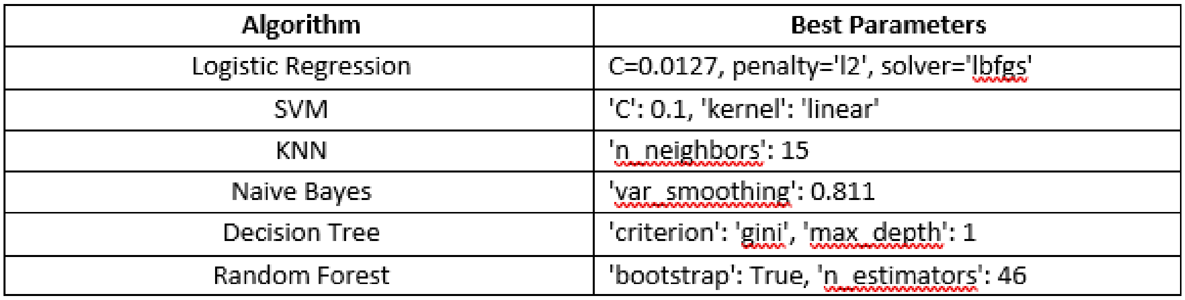
Figure [3]

Various machine learning models, both supervised and unsupervised ones were trained and tested. The accuracy, precision, recall, f-score were found out and are tabulated below.



Algorithm	Accuracy	Precision	Recall	F1-Score	ROC_AUC_Score
Logistic Regression	79.45	91	98	86	71.42
SVM	79.46	93	98	86	70.87
KNN	80.4	85	96	85	70.59
Naive Bayes	82.70	83	95	89	72.74
Decision Tree	79.45	90	85	82	71
Random Forest	82.70	83	96	89	72.16
Ada Boost	83.24	89	98	89	71.95
XG Boost	82.70	86	97	89	71
Gradient Boost	78.91	82	91	86	69.53

Parameters for Machine Learning Models:



Algorithm	Best Parameters
Logistic Regression	C=0.0127, penalty='l2', solver='lbfgs'
SVM	'C': 0.1, 'kernel': 'linear'
KNN	'n_neighbors': 15
Naive Bayes	'var_smoothing': 0.811
Decision Tree	'criterion': 'gini', 'max_depth': 1
Random Forest	'bootstrap': True, 'n_estimators': 46

IV. Conclusion

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems.

Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.

References

[1] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

[2] ‘Loan Approval Prediction based on Machine Learning Approach’ by Kumar Arun, Garg Ishan, Kaur Sanmeet.

[3] Loan Repayment Prediction Using Machine Learning Algorithms by Chang HanMaster of Applied Statistics in University of California, Los Angeles, 2019Professor Yingnian Wu, Chair.

[4] J.R. Quinlan. Induction of decision trees. Machine learning Springer, 1(1):81–106, 1086.

[5] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News(http://CRAN.R-project.org/doc/Rnews/), 2(3):9–22, 2002.

[6] S.S. Keerthi and E.G. Gilbert. Convergence of a generalize SMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002. .

[7] J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077.

[8] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[9] https://www.ijarcce.com/upload/2016/march-16/IJARCCE%20128.pdf