Loan Prediction System Using Supervised Machine Learning

Adnan Shaikh **Kanchan Mengune**

Binitdev Pandey **Zeeshan Ansari**

ABSTRACT

In today's fast growing world everyone gets to the point where they need loan for something, in Banking, Car or any other system which gives their customers loan benefits according to their status, as technology is getting better day by day customers can apply for the loan through online system (Website or App) because of this large data (10,000+) can be piled up it will be difficult for any employee to evaluate all customers information (as well as giving additional salary just for passing loan) and depending on that approving their loan, it is necessary to have something at hand which can check whether to approve loan or not without interference of human. So, keeping this in mind we created ML (Machine Learning) Loan Prediction Project which can approve loan to customers depending on the necessary information they provide. This project goes through many steps from pre-processing data in proper format to transforming it to suitable format for applying ML algorithms, after transformation different ML algorithms are applied, selecting patterns of the algorithm which have highest accuracy and finally applying patterns on Test Dataset to predict whether to approve loan or not.

Keywords

Machine Learning, Kfold Stratified Sampling, Bootstrap, Logistic Regression, Random Forest, XG Boost and ADA Boost

1. INTRODUCTION

Credits are the core business of banks. The main revenue comes directly from the mortgage's interest. The loan firms grant a loan after an intensive process of confirmation and authentication. However, they still don't have guarantee if the applicant is able to repay the loan with no complications.

We'll build a analytical model to predict if an applicant is able to repay the loaning firm or not. We will prepare the data using Jupyter Notebook and use various models to predict the target variable. Housing Finance firm deals in all home mortgages. They have presence across all town, semi urban and countryside areas. Client first apply for home loan after that firm validates the client eligibility for mortgage.

The Firm wants to systematize the credit eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Edification, Number of Dependents, Revenue, Credit Amount, Credit History and others. To systematize this process, they have given a problem to recognize the clients sections, those are qualified for credit amount so that they can precisely target these clients.

2. LITREATURE REVIEW

3. MATERIALS AND METHODLOGY

3.1 Data Set

Attribute	Values	Type
Loan_ID	Unique ID (Nominal)	Original
Gender	Male/Female (Nominal)	Original
Married	Yes/No (Nominal)	Original
Dependents	Number family member depends on client (Numeric)	Original
Education	Graduate/Undergraduate (Ordinal)	Original
Self_Employed	Yes/No (Nominal)	Original
ApplicantIncome	Primary Income (Numeric)	Original
CoapplicantIncome	Secondary Income (Numeric)	Original
LoanAmount	Loan Amount in thousands (Numeric)	Original
Loan_Amount_Term	Term of lean in months (Numeric)	Original
Credit_History	CIBIL Score (Binary)	Original
Property Area	Urban/Semi/Rural (Ordinal)	Original
Loan_Status	Loan Approval (Binary)	Original
Total Income	Applicant Income + Co-applicant Income (Numeric)	Derived
EMI	Total Income - (Loan Amount/Loan Amount Term)*1000 (Numeric)	Derived

Balanced Income	Total Income - EMI (Numeric)	Derived

3.2 Algorithms

Following ML Algorithms are going to be used in Loan prediction system:

1) Logistic regression.

2) Decision tree.

3) Random Forest.

4) XGBOOST

5) ADABOOST

3.2.1 Logistic Regression

I) Logistic regression is a classification algorithm which uses logistic function or sigmoid function which takes value in the range [0, 1].

II) Sigmoid function: f(x) = $\frac{1}{1+e^{-x}}$

III) Projecting extreme values i.e
f: $[-inf, +inf]$ → [0,1].

IV) Logistic regression work much like Linear regression taking Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y=f(x)).

V) Equation of logistic regression:

$$y = \frac{e^{b0 + b1*x}}{1 + e^{b0 + b1*x}} = \frac{1}{1 + e^{-(b0 + b1*x)}}$$

VI) Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in input data has an associated b coefficient (a constant real value).

VII) The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation, which uses maxima-minima technique to find value of parameters such that error rates are as low as possible.

e.g.: Consider we want to create a model which predict loan to be pass or not on basis of Income.

We will find probability of Loan pass (1) and rejected (0) depending on Income.

i.e.$p(x) = p(L = 1|Income = value)\frac{e^{b0 + b1*x}}{(1 + e^{b0 + b1*x})}$

(loan pass when certain value limit in Income identified).

After simplifying this equation we get

$\ln\left(\frac{p(x)}{1-p(x)}\right) = b0 + b1 * x$. After applying maximum likelihood on given function

(Since, $p(x; b0, b1) = L(b0, b1; x)$) using our train set we will find value of b0 and b1 and we will apply these values with value of x to find corresponding value of y in test set. We will consider Loan pass if f(x) >= 0.5 and rejected if f(x) < 0.5.

$$y = \frac{e^{b0 + b1*X}}{(1 + e^{b0 + b1*X})}$$

$$y = \frac{e^{-100 + 0.6*150}}{(1 + e^{-100 + 0.6*X})}$$

$(consider, b0 = -100\ and\ b1 = 0.6)$

$f(x) = y = 0.0000453978687$

$since, f(x) < 0.5\ Loan\ is\ Rejected.$

3.2.1 Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

This algorithm split the node into n-nodes depending on best cost of split (lowest cost). Different type of cost function can be used, for classification decision tree we used Gini and Entropy, Gini index is given by:

$$\boldsymbol{G = 1 - \sum(pk(1 - pk))}$$

Here, pk is proportion of same class inputs present in a particular group. A perfect class purity occurs when a group contains all inputs from the same class, in which case pk is either 1 or 0 and G = 0, where as a node having a 50–50 split of classes in a group has the worst purity, so for a binary classification it will have pk = 0.5 and G = 0.5.

Entropy is given by: $\boldsymbol{E = -(\sum pklog(pk))}$

Here, pk means same as in Gini index but the value lies from [0, 1]. A perfect class purity occurs when a group contains all inputs from the same class, in which case pk is either 1 or 0 and E = 0, where as a node having a 50–50 split of classes in a group has the worst purity, so for a binary classification it will have pk = 0.5 and E = 1.

3.2.1 Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging and bootstrap (0.632) sampling. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. We will be using Greed Search to tune the model i.e. finding the best values for hyper parameters and RepeatedStratifiedKFold sampling which is an iterative sampling method combining both Stratified and KFold Cross Validation method.

3.2.1 XGBOOST

This algorithm only works with the quantitative variable. It is a gradient enhancing algorithm which forms solid rules for the model by boosting weak beginners to a strong learner. It is a fast and well-organized algorithm which newly conquered machine learning because of its high performance and speed.

3.2.1 ADA BOOST

AdaBoost algorithm, short for Adaptive Boosting, is a Enhancing method used as an Collective Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to imperfectly classified instances. Boosting is used to lessen bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. But for the first, each subsequent beginner is grown from previously grown beginners. In simple words, weak beginners are transformed into strong ones

4. RESULT

SR No.	Algorithms	Accuracy	Running Time	AUC Score
1.	Logistic Regression	81.11%	133ms ± 3.05ms	0.83
2.	Decision Tree	78%	34ms ± 915 µs	0.71
3.	Random Forest	81.40%	4.27s ± 83.2ms	0.69
4.	XGBOOST	76.54%	211ms ± 10ms	0.70
5.	ADABOOST	81.11%	228ms ± 5.83ms	0.74

Random Forest, Logistic Regression and ADABOOST gave the highest accuracies, Random Forest estimation time is highest and lowest AUC score because of modelling of large number of decision trees but these factors help the Random Forest to achieve highest accuracy and Logistic Regression estimation time is lowest out of these three classifiers which can be useful in evaluation of large data se

5. CONCLUSION

We did exploratory data Analysis on the features of this dataset and observe how each feature is distributed. We did bivariate and univariate analysis to see impact of one another on their features using charts. We analysed each variable to check if data is cleaned and normally distributed. We cleaned the data and removed NA values we also generated hypothesis to prove an association among the independent variables and the Target variable. And based on the results, we assumed whether or not there is an association. We calculated correlation between independent variables and found that applicant income and loan amount have significant relation. We created dummy variables for constructing the model we constructed models taking different variables into account and found through odds ratio that credit history is creating the most impact on loan giving decision finally, we got a model with co-applicant income and credit history as independent variable with highest accuracy. We tested the data and got the accuracy of 81 %

REFRENCES

[1] Modeling Car Loan Prepayment

Author: Sara Zahia , Boujemâa Achchab

[2] An effective combining classifier approach using tree algorithms for network intrusion detection

Authors: Jasmin Kevric, Samed Jukic, Abdulhamit Subasi

[3] Association of cardiac biomarkers and comorbidities with increased mortality, severity, and cardiac injury in

COVID-19 patients: A meta-regression and decision tree analysis

Authors: Eman A. Toraih, Rami M. Elshazli, Mohammad H. Hussein, Abdelaziz Elgam, Mohamed Amin, Mohammed El-Mowafy, Mohamed El-Mesery, Assem Ellythy, Juan Duchesne, Mary T. Killackey, Keith C. Ferdinand, Emad Kandil, Manal S. Fawzy

[4] XGBoost: A Scalable Tree Boosting System

Authors: Carlos Guestrin,Tianqi Chen

[5] An Ensemble Random Forest Algorithm for Insurance Big Data Analysis

Authors: WEIWEI LIN, ZIMING WU , LONGXIN LIN , ANGZHAN WEN , AND JIN LI

[6] Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection

Authors: IFTIKHAR AHMAD , MOHAMMAD BASHERI, MUHAMMAD JAVED IQBAL , AND ANEEL RAHIM

[7] The Boosting Approach to Machine Learning An Overview

Author: Robert E. Schapire