Loan Prediction System Using Supervised Machine Learning

Binitdev Pandey
(University of Mumbai)

Adnan Shaikh
(University of Mumbai)

Kanchan Mengune
(University of Mumbai)

Zeeshan Ansari
(University of Mumbai)

ABSTRACT

In today's fast growing world every person gets to the point where they need credit for something, in Banking, Car or any other organization which gives their clients loan benefits according to their status, as technology is getting better day by day clients 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 estimate 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 favour loan to customers reliant 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 conversion different ML algorithms are applied, selecting patterns of the algorithm which have highest accuracy and finally applying patterns on Test Dataset to foresee whether to approve credit or not.

Keywords

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

1. INTRODUCTION

Providing loan and generating income in form of interest is revenue of finance and banking industry in order to provide loan one must know about customer financial history regarding customer value or history of paying dues.

To do these things we need large number or manpower back in days but now days we can perform this process by having certain information about customer so that we can feed it to algorithm that can evaluate customer potential credit limit and to ease the way of doing for finance sector. In order to do that we collectively use various algorithms to maximize percentage of accuracy and providing better infrastructure to banking sectors and their customers is need of changing environment and future.

2. LITREATURE REVIEW

Loan calculation is ongoing most important topic of finance sectors credit evaluation based on record is main method or way to calculate this. And increasing use of ML and AI is gaining popularity in this sector and predicting credit on finance history is required in finance sectors so in order to achieve it, many algorithms and method is being used right now.

In paper [1] author: Sara Zahia, Et Al. proposed to apply ML cataloguing algorithm to binary logistic regression which will aid them in decision-making tool for banking sector, which will help them to avoid losses in advance risk in turnover. [1] In a logistic model it is cleared that the modelled variable which it shows the probability of whether loan given will be paid or not by the end date, it describes the credit situations as well as the contributors by a quantity of explanatory parameters specification.[1] The significance of model's constraints were assessed tested. with the use of performance level metrics

In paper [2] uthor: Jasmin K Et Al. proposed network interruption recognition on tree-based algorithm. The dataset of NSL-KDD, is more improved than the KDDCUP'99 dataset, which is used for evaluation of detection algorithm. The arriving network packets are normal or an attack is categorized by detection algorithm, describing every design of organization traffic is based on its features.[2] Author conclude that tactic based on the sum rule structure can produce better results than specific classifiers when classifiers are merged

In paper [3] Authors: Eman A.T Et Al The result of cardiac markers probability of changing outcome predictions and the selection of good cut is evaluated with the help of receiver operating-characteristic curve method. [3] Decision tree study of medical features and merging cardiac markers with demographic is seen useful for mortality and sternness in patients with SARS predictions.

In paper [4] Authors: Carlos G Et Al Author proposed a novel sparseness attentive algorithm for theoretically acceptable weighted quantile sketch and handling sparse data for estimation learning.

In paper [5] Authors: Weiwei L Et Al The authors analysed insurance business data for disproportion distribution, In disproportion dataset of pre-processing algorithms it is concluded, in which given ensemble

random forest algorithm build on Apache-Spark which is used in the imbalanced cataloguing of the big scaled business data, the results showed that it is more appropriate in the insurance product approval or probable client study than already existing strong classifier like SVM and Logistic Regression with the help of the ensemble random forest algorithm. [5] The KNN united with planning of bootstrap under sampling algorithm might be helpful into pre-processing of imbalanced cataloguing algorithms. Combined with bootstrap sampling pre-processing and the ensemble learning algorithms could diminish the learning progression further, and it also have relation with other disproportion data mining algorithms.

In paper [6] Authors: Iftikar A Et Al discussed regarding privacy or security intervention prediction is must in information system and future networks, since everything are getting heavily reliant on these. To make sure that in future privacy invasion is avoided and respected.[6] many methods is used but ML is common in many areas

In paper [7] Author: Robert E.S author discussed regarding a method based on functional gradient origin, as an estimate of logistic regression of AdaBoost has been understood as in a repeated-game playing procedure. [7] Various areas, consist of concept of gaming and linear programming, support vector machines, Brownian motion, Bregman distances, logistic regression and maximum-entropy procedures like iterative grading is connected to AdaBoost revealed

3. METHODLOGY

3.1 Data Set

Attribute	Values	Type
ID	Unique ID (Nominal)	Original
Gender	Male/Female (Nominal)	Original
Marital status	Yes/No (Nominal)	Original
Dependents	Number family member depends on client (Numeric)	Original
Qualification	Graduate/Undergraduate (Ordinal)	Original
Owned Business	Yes/No (Nominal)	Original
Primary Income	Primary Income (Numeric)	Original
Secondary Income	Secondary Income (Numeric)	Original
Loan Amount	Loan Amount in thousands (Numeric)	Original
Loan_Payment_Month	Term of lean in months (Numeric)	Original
Cibil Score	CIBIL Score (Binary)	Original
Residence	Urban/Semi/Rural (Ordinal)	Original
Loan Approval	Loan Approval (Binary)	Original
Total Income	Primary Income + Secondary Income (Numeric)	Derived
EMI	Total Income - (Loan Amount/Loan Payment month) *1000 (Numeric)	Derived
Balanced Income	Total Income - EMI (Numeric)	Derived

(Table:1)

3.2 Algorithms

Following ML Algorithms are successful to be used in Loan prediction structure:

1) Logistic regression.

2) Decision tree.

3) Random Forest.

4) XGBOOST

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) = $\frac{1}{1+e^{-x}}$ -- (1)

III) Projecting extreme values i.e.,
f: $[-inf, +inf]$ → [0,1]. --(2)

IV) Logistic regression is much similar to Linear regression Input values are combined linearly using coefficient values to forecast an output value y=f(x).

V) Calculation of logistic regression:

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

--(3)

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

--(4)

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.

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

--(5)

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

--(6)

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

$f(x) = y = 0.0000453978687$

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

3.2.1 Decision Tree

Decision Tree is classification algorithm. The aim is to generate a model that forecasts the value of a class variable by learning simple decision rules incidental from the data features. A tree can be understood as a piecewise constant approximation.

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

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

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

3.2.1 Random Forest

It is ML algorithm of classification. contains of many decision trees. The forest creates with the help of this algorithm is trained by bagging and bootstrap sampliing. it estimates the result based on voting of decision trees. A random forest eliminates the restrictions of a decision tree algorithm. It decreases the overfitting of datasets and rises 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.

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

(Table:2)

5. CONCLUSION

We analysed each variable to check if data is cleaned and normally distributed. We transform the data and removed null values we also created hypothesis to verify relations between the independent variables and the class variable. This is constructed based on results and made assumption regarding association is there or not. 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%. 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 set. In future we can integrate these models with credit system app, we can increase model accuracy by introducing attributes and can also predict future losses.

REFRENCES

[1] Sara Zahia, Boujemâa Achchab "Modeling Car Loan Prepayment" International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI) April 6-9, 2020, Warsaw, Poland

[2] Jasmin Kevric, Samed Jukic, Abdulhamit Subasi "An effective combining classifier approach using tree algorithms for network intrusion detection" springer article

[3] 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 "Association of cardiac biomarkers and comorbidities with increased mortality, severity, and cardiac injury in COVID-19 patients: A meta-regression and decision tree analysis" in Journal of Medical Virology · June 2020

[4] Guestrin, Tianqi Chen "XGBoost: A Scalable Tree Boosting System" research article

[5] Weiwei lin, Ziming wu, Longxin Lin, Angzhan Wen, and Jin li "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis" Digital Object Identifier 10.1109/ACCESS.2017.2738069

[6] Iftikhar Ahmaed, Mohmmad Basheri, Muhmmad Javed Iqbal, and Aneel Rahim "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection" Digital Object Identifier 10.1109/ACCESS.2018.2841987

[7] Robert E. Schapire "The Boosting Approach to Machine Learning an Overview" AT&T Labs Research Shannon Laboratory 180 Park Avenue, Room A203 Florham Park, NJ 07932 USA www.research.att.com/ schapire December 19, 2001

[8] R. Bekkerman. The present and the future of the kdd cup competition: an outsider's perspective.

[9] R. Bekkerman, M. Bilenko, and J. Langford. Scaling Up Machine Learning: Parallel and Distributed Approaches. Cambridge University Press, New York, NY, USA, 2011.

[10] J. Bennett and S. Lanning. The netflix prize. In Proceedings of the KDD Cup Workshop 2007, pages 3–6, New York, Aug. 2007.

[11] L. Breiman. Random forests. Maching Learning, 45(1):5–32, Oct. 2001.