Loan Approval Prediction Using Logistic Regression Model in Machine Learning

Sayed Ebrahim Shiwa , S Prasad Babu Vagolu and Prof Vedavathi Katneni

Abstract*: Credit endorsement is a vital interaction for banking associations. The framework endorsed or dismisses the advance applications. Recuperation of advances is a significant contributing boundary in the budget summaries of a bank. It is exceptionally hard to anticipate the chance of installment of advance by the client. As of late numerous scientists dealt with advance endorsement forecast frameworks. AI (ML) methods are helpful in foreseeing results for huge measure of information. In this paper three AI calculations, Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF)are applied to foresee the advance endorsement of clients. The test results presume that the exactness of Decision Tree AI calculation is better when contrasted with Logistic Regression and Random Forest AI draws near.

***Keywords*: Random Forest Regression, Regression, Logistic Regression, Neural network, Machine Learning.**

I. INTRODUCTION

Advances are the center business of banks. The principle benefit comes straightforwardly from the advance's advantage. The advance organizations award an advance after a concentrated cycle of check and approval. Be that as it may, they actually don't have affirmation if the candidate can reimburse the advance without any challenges. Internet driving has acquired prominence in light of its higher proficiency in offering credit to customers and independent companies. It is one of the businesses that has been upset by innovation with the electronic loaning stage. Credit application choice is made consequently with electronic information driven calculations. Online banks have the adaptability to bring to the table little credits with transient developments. Borrowers that are rejected from customary financial frameworks can thus get the opportunity to get to credit. Web based loaning has developed with stages associating banks and borrowers to more enhanced plans of action like direct loaning, monetary institutional organizations.

II. PROCEDURE FOR PAPER SUBMISSION

Assemble a computerized framework for credit application handling (choice to support or reject the advance) in view of the different boundaries as chosen by the AI calculation. The reason for the archive is to clarify the High engineering that would be utilized for building up the mechanized credit acknowledgment framework. The design outline will give an outline of a whole framework, recognizing the fundamental parts that would be created for the item and their interfaces.

III. PRE-PROCESSING THE DATASET

Firstly, we need to prepare the dataset for applying various algorithms.

A. Exploratory Data Analysis (EDA)

EDA isn't indistinguishable from measurable illustrations albeit the two terms are utilized reciprocally. Measurable illustrations is an assortment of strategies - all graphically put together and all centering with respect to one information portrayal angle. EDA envelops a bigger setting; EDA is a way to deal with information examination that delays the typical suspicions about what sort of model the information follow with the more straightforward methodology of permitting the actual information to uncover its fundamental design and model. EDA is certainly not a simple assortment of procedures; EDA is a way of thinking with regards to how we take apart an informational collection; what we search for; what we look like; and how we decipher. The facts confirm that EDA intensely utilizes the assortment of strategies that we call "measurable illustrations", yet it isn't indistinguishable from factual designs essentially.

Steps in EDA

- Bivariate Analysis

  Bivariate analysis is one of the measurable investigation where two factors are noticed. One variable here is reliant while the other is free. These factors are generally indicated by X and Y. In this way, here we examine the progressions occured between the two factors and how much. Aside from bivariate, there are other two factual investigations, which are Univariate (for one factor) and Multivariate. In insights, we as a rule decipher the given arrangement of information and offer expressions and forecasts about it. During the exploration, an investigation endeavors to decide the effect and cause to finish up the given factors. Bivariate analysis is expressed to be an examination of any simultaneous connection between two factors or properties. This investigation investigates the relationship of two factors just as the profundity of this relationship to sort out if there are any errors between two factors and any reasons for this distinction. A portion of the models are rate table, dissipate plot, and so forth
- Preprocessing Data

In any Machine Learning measure, Data Preprocessing is that progression wherein the information gets changed, or Encoded, to carry it to such an express that now the machine can undoubtedly parse it. All in all, the highlights of the information would now be able to be effectively deciphered by the calculation. Data preprocessing is an information mining strategy that includes changing crude information into a reasonable arrangement. True information is frequently deficient, conflicting, and additionally ailing in specific practices or drifts, and is probably going to contain numerous mistakes. Information preprocessing is a demonstrated strategy for settling such issues.

IV. DATA VISUALIZATION

Data Visualization is the demonstration of making an understanding of information into a visual setting, similar to a guide or graph, to simplify data for the human frontal cortex to grasp and pull pieces of information from. The essential target of data discernment is to simplify it to recognize models, examples and exemptions in tremendous enlightening assortments. The term is habitually used proportionally with others, including information plans, information portrayal and genuine delineations.

Data visualization is one of the methods for the data science measure, which communicates that after data has been accumulated, taken care of and illustrated, it ought to be imagined for closures to be made. Data insight is moreover a segment of the more broad data show designing discipline, which means to recognize, discover, control, organize and pass on data in the most capable way possible.

A. Categorical Independent Variable vs Target Variable:

First of all, we will find the relation between the target variable and categorical independent variables. Let us look at the stacked bar plot now which will give us the proportion of approved and unapproved loans.

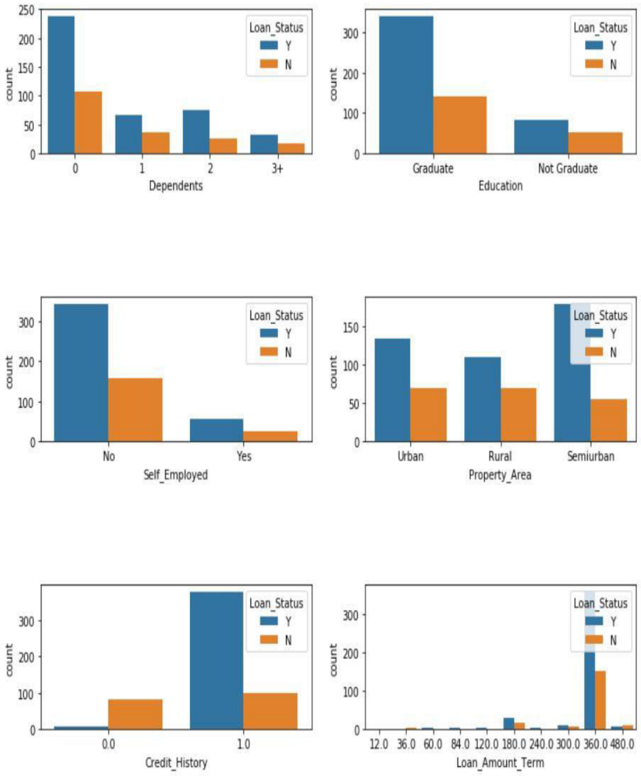


Figure 1

Observations

Plots above pass on after things about the dataset:

- Advance Approval Status: About 2/third of candidates have been allowed advance.
- Sex: There are a bigger number of Men than Women (approx. 3x)
- Martial Status: 2/third of the populace in the dataset is Marred; Married candidates are bound to be allowed advances.
- Wards: Majority of the populace have zero wards and are additionally prone to acknowledged for advance.
- Instruction: About 5/sixth of the populace is Graduate and graduates have higher proportion of advance endorsement
- Business: 5/sixth of populace isn't independently employed.
- Property Area: More candidates from Semi-metropolitan and furthermore prone to be allowed advances.
- Candidate with record of loan repayment are undeniably bound to be acknowledged.
- Advance Amount Term: Majority of the credits taken are for 360 Months (30 years).
- It appears individuals with a financial record as 1 are bound to get their advances endorsed.
- The extent of advances getting affirmed in the semi-metropolitan territory is higher when contrasted with that in country or metropolitan regions.

B. Visualize numerical independent variables with respect to the target variable :

Allow us to join the Applicant Income and Co-candidate Income and see the consolidated impact of Total Income on the Loan_Status.

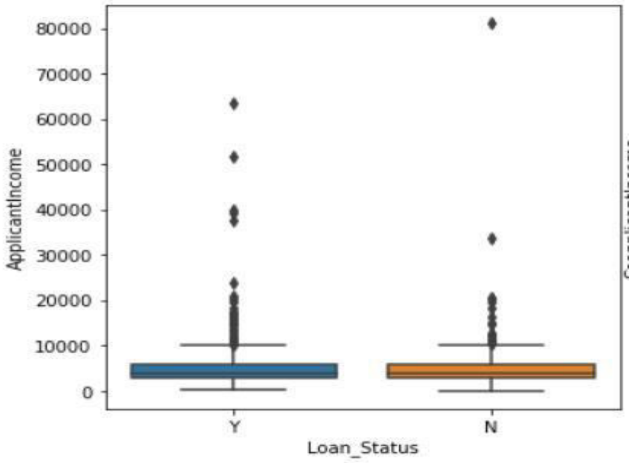


Figure 2

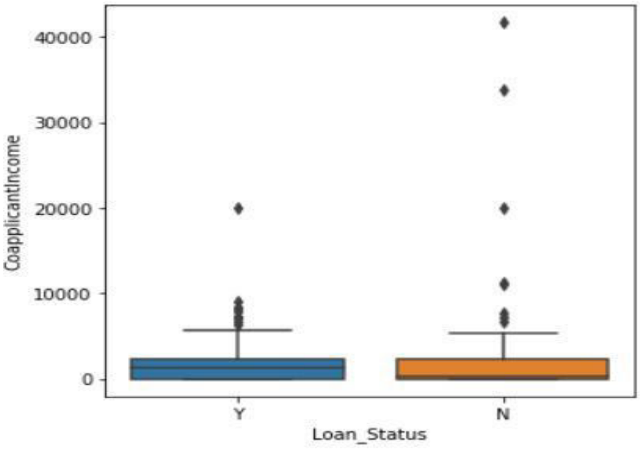


Figure 2

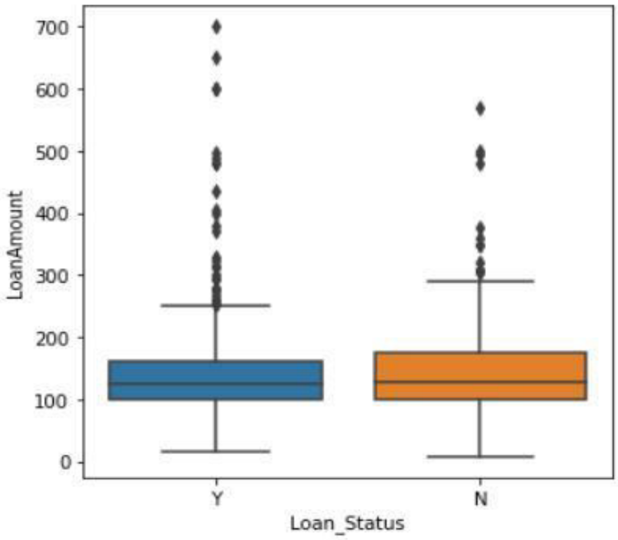


Figure 3

We can see that Proportion of advances getting affirmed for candidates having low Total Income is less contrasted with that of candidates with Average, High and Very High Income.

It shows that if co-candidates pay is less the odds of advance endorsement are high. In any case, this doesn't look right. The conceivable purpose for this might be that the majority of the candidates don't have any co-candidate so the co-candidate pay for such candidates is 0 and consequently the credit endorsement isn't subject to it. Along these lines, we can make another variable where we will join the candidate's and co-candidates pay to envision the consolidated impact of pay on advance endorsement.

V. MODEL TRAINING

A. *Decision Tree Classifier* -

Decision tree is a type of supervised learning algorithm(having a pre-defined target variable) that is mostly used in classification problems. In this technique, we split the population or sample into two or more homogeneous sets(or sub-populations) based on the most significant splitter/differentiator in input variables.
Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable..

B. *Random Forest Algorithm*-

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

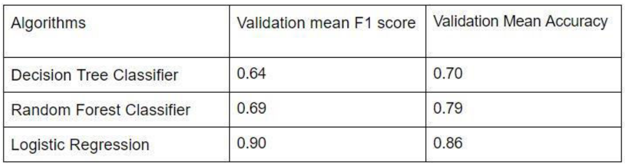
C. *Logistic Regression Algorithm-*

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

VI. RESULTS

At the point when every one of the outcomes are looked at Logistic Regression is performing admirably.

Table- II: Performance Comparison Chart for Algorithms-



Algorithms	Validation mean F1 score	Validation Mean Accuracy
Decision Tree Classifier	0.64	0.70
Random Forest Classifier	0.69	0.79
Logistic Regression	0.90	0.86

Calculated Logistic Regression: The validation mean F1 score is 0.90 while the validation mean accuracy is 0.86.

- The logistic regression model precision score is 0.86. In this way, the model does an excellent occupation in forecast.

Observations

- Logistic Regression Confusion matrix is very similar to Decision Tree and Random Forest Classifier. In this analysis, we did extensive analysis of input data and were able to achieve Test Accuracy of 86 % Logistic Regression Confusion matrix is very similar to Decision Tree and Random Forest Classifier. In this analysis, we did extensive analysis of input data and were able to achieve Test Accuracy of 86 %.

VII. CONCLUSION

This investigation gives fascinating data about the clients advance qualification. It relies upon different elements like Applicant Income, Co-applicant Income, Loan Amount and Credit History. From the exploration I comprehended that advance qualification relies upon the above factors. So by AI model we can channels clients who are qualified for advance dependent on above factors. It is apparent that Logistic Regression Confusion framework is basically the same as Decision Tree and Random Forest Classifier. In this examination, we did broad investigation of info information and had the option to accomplish Test Accuracy of 86 %.

REFERENCES

[1] H. Harb and L. Chen, Voice-based gender identification in multimedia applications, Journal of Intelligent Information Systems, 24(2), 179-198 (2005).

[2] Md. Sadek Ali1, Md. Shariful Isla1 and Md. AlamgirHossain, Gender recognition of speech signal, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(1), 1–9 (2012).

[3] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. In Proc. XII European Signal Processing Conf., volume 1, pages 341–344. Vienna, Austria,(2004).

[4] M. Sedaaghi, A Comparative Study of Gender and Age Classification in Speech Signals, Iranian Journal of Electrical & Electronic Engineering, 5(1), 1–12 (2009)

[5] S. Gaikwad, B. Gawali, and S.C. Mehrotra, Gender identification using SVM with combination of MFCC, Advances in Computational Research, 4(1), 69-73, 2012.

[6] Y.-M. Zeng, Z.-Y. Wu Falk, and W.-Y. Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in Machine Learning and Cybernetics, 2006 International Conference on. IEEE 2006

[7] C.S. Leung, M. Lee, and J.H. Chan (Eds.), "Gender Identification from Thai Speech Signal Using a Neural Network" ICONIP 2009, Part I, LNCS 5863, pp. 676–684, 2009

[8] Kumar R.,Dutta S.,Kumara shama,"Gender Recognition using speech processing technique using LABVIEW" IJAET May 2011

[9] Md. Rabiul Islam1, Md. Fayzur Rahman,"Improvement of Text Dependent Speaker Identification System Using Neuro-Genetic Hybrid Algorithm in Office Environmental Conditions ",IJCSI International Journal of Computer Science Issues, Vol. 1, 2009.

[10] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet," Front-End Factor Analysis for Speaker Verification" IEEE transaction on Audio, Speech, And Language Processing, Vol. 19, No. 4, May 2011. [11] Tomi Kinnunen, " Spectral Features for Automatic TextIndependent Speaker Recognition", Ph. Lic. Thesis, Department of Computer Science University of Joensuu , 2004.

[12] Milan Sigmund. "Gender Distinction Using Short Segments Of Speech Signal".

[13] Atal B., "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification", Journal of the acoustic Society of America 1974, pp. 55(6):1304-1312.

[14] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," Journal of Intelligent Information Systems, vol. 24,no. 2, pp. 179–198, 2005.

[15] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Transactions on acoustics, speech and signal processing, vol. 22, no. 2, pp. 135–141, 1974.

[16] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker verification over long distance telephone lines", IEEE Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, May 1989, pages 524--527.

AUTHORS PROFILE

Sayed Ebrahim Shiva Master of Computer Science, Department of Computer Science, GIS, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India. His area of interest is in python, Data Mining and Machine Learning.

S Prasad Babu Vagolu is an Assistant Professor in Department of Computer Science, GIS, GITAM. He has 8 years of Software Development Experience and 10 years of Teaching Experience. He is lifetime member of CSI, Indian Science Congress and passionate about research in Wireless Mesh Networks, Machine Learning and Artificial Intelligence.

Dr Katneni Vedavathi is currently working as HOD of Computer Science department, GITAM, Visakhapatnam. She completed her Ph.D. from Andhra University. She is member of CSI, Indian Science Congress and ISTE. She received multiple awards. Her areas of interest are Data

Mining, Cognitive Learning, Machine Learning and IOT.