Prediction of Home Loan Status Eligibility using Machine Learning

[1]Milind Udbhav, [2]Robin Kumar, [3]Nitin Kumar, [4]Rohit Kumar, [5]Meenu Vijarania, [6]Swati Gupta

[1,2,3,4] Student, Department of Computer Science, School of Engineering and Technology, K R Mangalam University, Gurugram, India
[5,6]Associate Professor, Centre of Excellence, Department of Computer Science, School of Engineering and Technology, K R Mangalam University, Gurugram, India

[1]*udbhavpat@gmail.com*,[2] *robinattri0@gmail.com*,[3]*nitinkumarnn33@gmail.com*,[4]*rohit24kumar68@gmail.com*,[5]*meenuhans.83@gmail.com*,
[6]*swattiguptta@gmail.com*

Abstract

Getting a home loan and checking the eligibility is not so straightforward process, it is time-consuming for both the housing finance company and the customer who takes the home loan. A lot of time is wasted on the completion of the entire process. Even using an online application does not easy the process as many details are required such as Gender, Marital Status, Education, Dependents, Self Employed or not, Loan Amount, Loan Amount Term, Credit History, Property Area, and Loan Status. The loan status is the target data and others are its parameter through which some output and pattern can be obtained. The company also collects information about the area urban, semi-urban, rural wants customer wants its property. Overcome this issue machine learning is extremely helpful to automate the process to check the eligibility of customers but the fundamental issues that are used over several years do not care about the outliers and machine learning models, which results in less accuracy and affects the overall performance of the model. To reduce this issue the project uses a data cleaning technique for removing null, missing, and repeated values before applying the bivariate and multivariate analysis which helps to categorize the type of data whether it is numerical or categorical for understanding some unique relations and patterns which will help to increase performance and increase the accuracy as well as precision of the model. The machine learning algorithm contains many techniques like Random Forest, decision tree, and many others, but in the project, the two best classification models that are Logistic Regression, and Gradient boosting are used. Logistic Regression is a type of supervised learning which helps for better classification and predicts a discrete value and gradient boosting removes the error and mistakes of the previous model and helps improve overall performance. For automation of the process for eligibility of the customer, a dataset is collected by the housing finance company, then hidden trends and patterns are found which helps build a robust machine learning model. Evaluate the performance of the model performance metrics such as accuracy, precision, and f1score are used. The use of evaluation methods helps to produce the best model which is best to check the eligibility and make this process easier, hassle-free, and convenient. The project uses multiple techniques and multiple methods which makes it a differentiating factor when compared with other models. The use of modern technology not only saves time but also helps in changing the traditional methods and bringing the revolution to improve without compromising the time. Automation also reduces the data cleaner without outliers giving the perfect model which the company wants and can also be further used to change dependents factor for the easy-going process of checking the background details of customers to avoid the chance of getting fraud.

Keywords:

Linear Regression, Logistic Regression, Data Cleaning, Univariate Analysis, Bivariate Analysis, Evaluation Metrics.

1. Introduction

The projects aim to make a predictive modeling system that helps housing finance companies such as Dream Housing Company to check the eligibility criteria of the person which helps in determining whether the person is eligible for a home loan or not. The finance company has its presence in rural, semi-urban, and urban areas. The customer applies for the home loan and the company checks the eligibility of the loan. The company wants to automate the process of the loan eligibility process based on customer details provided by the customer while filling out an online form. Some of the details are taken such as Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and many others. For automating the process, the company has provided a dataset to identify the customer segments which are eligible for loan amounts. The automation process helps to find the hidden pattern and build a robust model. Trends and Patterns are used for Predicting the Loan Status of the candidates. It lasts the evaluation of the model is done with the help of evaluation metrics such as accuracy, Recall, and F1 Score. The project uses univariate and bivariate analysis for categorizing the type of data. The data is categorized into two types of categorical data such as male/female and numerical data such as co-applicant income, credit-card history, etc. The project uses two machine learning for getting a good precision score and accuracy. Two machine learning classifying models are used such as logistic regression which helps for better classification and predicts a discrete value [1]. The second model is the gradient boosting model which removes the error and flaws of logistic regression and helps to improve accuracy, performance, and function as improving elements of the model. Using the new automation technology process can be shortened and make the process easier for both customers as well as a finance company. The use of modern technologies such as artificial intelligence, machine learning algorithms reduces the workload and provides a better understanding for everyone, and provides reliable results as compared to manual and traditional approaches [2].

2. Machine learning Technique

A Machine Learning software algorithm was proposed to build a robust and efficient software algorithm that classifies individuals based on characteristics such as gender, education, number of dependents, marital status, employment, credit score, loan amount, and other factors to determine whether they are eligible for a loan or not. While it is the first line of command, it will surely reduce the workload of all other bank workers because the process for identifying client segments and those who are qualified for a loan amount will be automated, allowing them to target those clients individually [4]. And if the loan applicant fits the eligibility conditions for loan approval based on those aspects, this will be indicated. Metrics including accuracy, precision, and f1 score are used to assess the model's performance [5]. The application of evaluation methodologies aids in the development of the optimal model for determining eligibility and making the procedure more simple, quick, and efficient. The machine learning method includes a variety of techniques such as Random Forest, decision tree, and others, but in this project, the two best classification models used are Logistic Regression and Gradient Boosting [6]. As it is easier to compare develop and deliver the most accurate predictive analysis, logistic regression is commonly used for loan prediction. Other algorithms are often lousy at forecasting non-normalized data, which is one of the reasons for this.

However, because the independent variables on which the prediction is made do not have to be regularly distributed, the nonlinear impact and power factors are simply managed by Logistic regression. It possesses significant limits, and parameter estimation requires a large sample of data. The variables must also be independent of one another in logistic regression; otherwise, the model will overestimate the significance of the dependent variables. As Logistic regression [7] is a supervised learning technique that aids in improved classification and predicts discrete values. Gradient boosting helps to increase overall performance by removing the preceding model's errors and mistakes. Gradient boosting is a machine learning technique that may be applied to a wide range of problems, including regression and classification. It comes with a prediction model in the form of a graph a collection of shaky prediction models [8]. As prediction models, decision trees are frequently utilized. It is founded on the assumption that the best possible next decision will be made and will be taken into consideration. When used in conjunction with the previous models, this model creates a powerful combination. Decreases the overall prediction error. The key idea is to Identify the targeted outcomes for the next model. Lessen the likelihood of committing a mistake [9]. In this situation, the gradient refers to the loss function. Gradient, which is the value that each new tree should achieve forecast.

3. Literature Survey

M. A. Sheikh, A. K. Goel, and T. Kumar used information from past bank clients who had loans granted based on a set of criteria. The machine learning model is trained on the record to generate reliable results. The major goal of the research is to predict the loan's safety. To forecast loan safety, the logistic regression approach is applied. The data is cleaned initially to avoid missing values in the data set. Vaidya discusses logistic regression and how to numerically describe it. In his research, he uses logistic regression as a machine learning methodology to combine predictive and probabilistic methodologies to solve a specific problem of loan approval prediction. This study uses logistic regression to determine if a loan for a set of an applicant's data will be approved. It also goes through some of the Machine Learning mode's other applications in the actual world.

One of the most essential problems for the long-term viability and profitability of the highly competitive industry is assessing the risk connected with a loan application. These banks receive numerous loan applications from their clients and others regularly. Not all of them have been approved. Many banks review loan requests and make credit approval decisions using credit scoring and risk analysis tools. Nonetheless, there are several cases each year in which borrowers fail to repay the loans and default, causing financial institutions to suffer significant losses. To address the problem, the project employs a data cleaning technique that removes null, missing, and repeated values before applying bivariate and multivariate analysis which aids in categorizing the type of data, whether numerical or categorical to understand some unique relations and patterns that will aid in improving model performance and precision.

The project also employs two types of machine learning to achieve high precision and accuracy. There are two machine learning classification models used, one of which is logistic regression, which aids in categorization and predicts a discrete value. The second model is the gradient boosting model, which helps to increase accuracy, performance, and function as an improving aspect of the model by removing the error and defects of logistic regression. The introduction of modern automation technology can shorten the procedure and make it easier for both the customer and financial company. When compared to manual and traditional procedures, the introduction of modem technologies such as artificial intelligence and machine learning algorithms saves burden and provides a better understanding for everyone while also providing good outcomes.

4. Methodology

4.1 Univariate data Analysis

The most basic type of data analysis is univariate analysis. Uni signifies one, thus there is just one variable in the data. Univariate data necessitates a separate examination of each variable. The objective of gathering data is to answer a query, or more specifically, a research question. It is one of the most basic types of statistical analysis, and it is used to determine whether two sets of values have a relationship. The variables X and Y are involved. The examination of one ("uni") variable is known as univariate analysis.

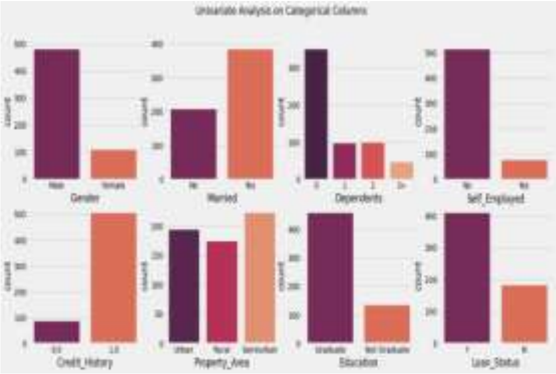


Fig.-1 Univariate Analysis on Categorical Columns.

In the given Figure it shows that More Loans are approved Vs Rejected, the Count of Male applications is more than females, the Count of Married applications is more than non-married, the count of the graduates is more than the non-Graduate, count of self-Employed is less than of non-Self-employed.

4.2 Bivariate Analysis

One of the most basic types of quantitative (statistical) analysis is bivariate analysis. It entails the examination of two variables (commonly labeled as X, and Y) to determine their empirical relationship. Simple hypotheses of connection can be evaluated using bivariate analysis.

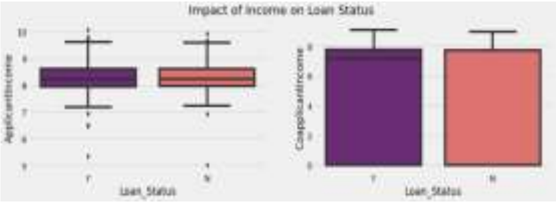


Fig-2 Impact of Income on Loan Status for Applicant Income and Co-applicant Income.

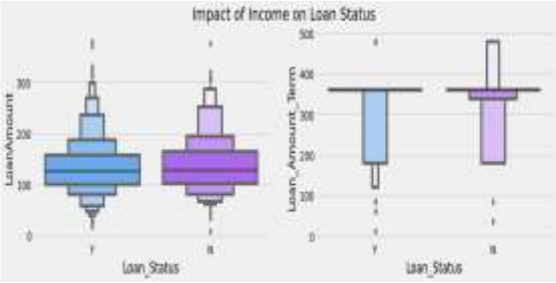


Fig-3 Impact of Income on Loan Status for loan amount on loan Amount and loan Amount status.

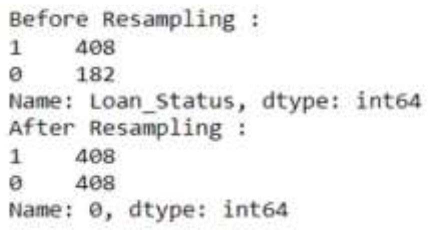
Here, In the Box Plot, (Fig-2) cannot show a clear value of loan time. The second figure (Fig3) shows that the number of loans as the duration increase becomes to pay the loan amount. When we know the value of one variable, bivariate analysis can help us figure out how much easier it is to know and predict the value of the other variable (potentially the independent variable) (see also correlation and simple linear regression).

4.3 *Resampling techniques*

Resampling techniques are a set of methods to either repeat sampling from a given sample or population, or a way to estimate the precision of a static. Although the method sounds daunting, the math involved is simple and only requires a high school level understanding of algebra.

The shape of the X and Y Train and Test

Before Resampling and After Resampling Results



```
Before Resampling :
1    408
0    182
Name: Loan_Status, dtype: int64
After Resampling :
1    408
0    408
Name: 0, dtype: int64
```

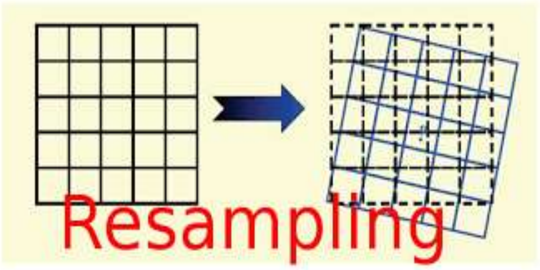
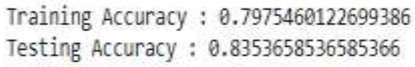


Fig-4 Change of pattern Before and after Resampling.

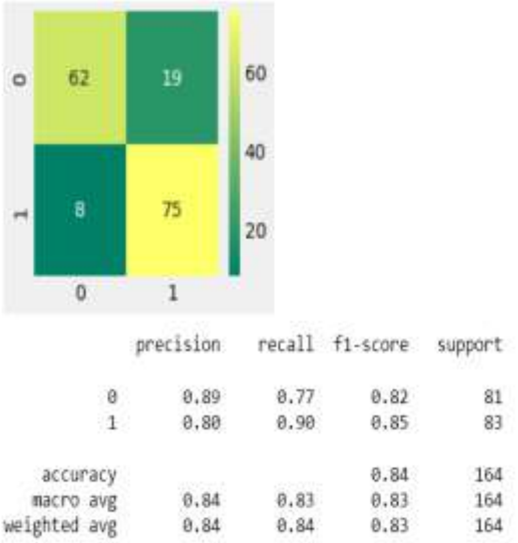
4.4 *Logistic Regression*

Despite the term "regression" in its name, Logistic Regression is a type of parametric classification model in the Machine Learning industry. This implies that logistic regression models contain a set number of parameters that are dependent on the number of input characteristics and provide categorical predictions, such as whether a plant belongs to a specific species or not. It performs complex calculations around probability into a straightforward arithmetic problem.

Accuracy



```
Training Accuracy : 0.7975460122699386
Testing Accuracy : 0.8353658536585366
```

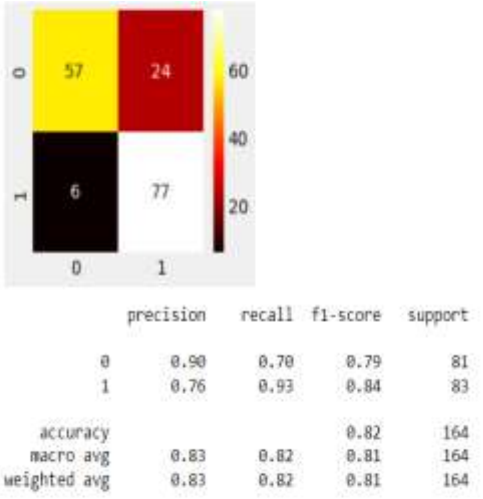


```
              precision    recall  f1-score   support

           0       0.89      0.77      0.82        81
           1       0.80      0.90      0.85        83

    accuracy                           0.84       164
   macro avg       0.84      0.83      0.83       164
weighted avg       0.84      0.84      0.83       164
```

Fig- 5 Confusion Matrix for Logistic Regression

4.5 *Gradient boosting*

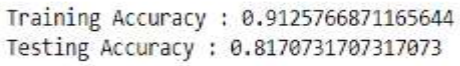
Gradient boosting is a machine learning technique that is used in a variety of tasks, including regression and classification. It provides a prediction model in the form of a group of weak prediction models. Decision trees are commonly used as prediction models. It is based on the belief that the best possible next decision will be taken. When paired with previous ones, this model reduces total prediction error The main concept is to determine the desired results for the following model. to reduce the chance of making a mistake. The gradient in this case refers to the loss function gradient, which is the target value for each new tree to forecast.



```
              precision    recall  f1-score   support

           0       0.90      0.70      0.79        81
           1       0.76      0.93      0.84        83

    accuracy                           0.82       164
   macro avg       0.83      0.82      0.81       164
weighted avg       0.83      0.82      0.81       164
```

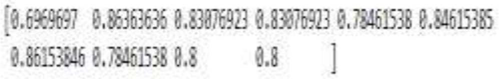
Fig-6 Confusion matrix for Gradient Boosting

Accuracy for Gradient Boosting



```
Training Accuracy : 0.9125766871165644
Testing Accuracy : 0.8170731707317073
```

Cross - validation Score



```
[0.6969697  0.86363636 0.83076923 0.83076923 0.78461538 0.84615385
 0.86153846 0.78461538 0.8        0.8       ]
```

For both models in real-life scenarios, the cross-validation score does not have much difference. It is just that the score for gradient boosting is a little higher as compared to that of logistic regression. Gradient boosting removes the error and flaws of logistic regression and helps to improve accuracy, performance, and function as improving elements of the model. Using the new automation technology process can be shortened and make the process easier for both customers as well as a finance company.

Conclusion

The project helped to understand the importance of resampling a technique that helped in the balancing of data and transformation techniques which helped to make data balanced. The two types of analysis that were performed are univariate and bivariate analysis to categorize the numerical and categorize data to get proper insights. Two machine learning model were used in the project which was logistic regression and gradient boosting. Among both the two gradients boosting provide better accuracy and precision as compared to logistic regression. Gradient boosting improves from the previous experience and improves the overall performance as well accuracy of the model. The project helped the company to automate the process and fastened the process for checking the person’s eligibility for loan criteria. Using the manual method, the process was insufficient and time taking, so the use of a machine learning algorithm was used to automate the process as well as make the company know the knowledge of genuine and fraudulent customers. It helped the company to understand the crucial factors that play a significant role in the loan status process. With the help of this project, the process became efficient as well as fast for both customers as well as company leading to reduce the efforts for both.

REFERENCES

[1.] Aditi Kacheria, Nidhi Sivakumar, Shreya Sawkar, Archana Gupta, ―Loan Sanctioning Prediction System‖, International Journal of Soft Computing and Engineering (IJSCE), vol. 6, no. 4, pp. 50-53, 2016.

[2.] Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal “Loan default forecasting using data mining” Department of Information Technology, St. Francis Institute of Technology, Mumbai, India (2020).

[3]. Yu-Jin and Yudan Zhu, "A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending," School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.

[4.] Masmoudi K, Abid L, Masmoudi A. Credit risk modeling using a Bayesian network with a latent variable[J]. Expert Systems with Applications, 2019, 127:157-166.

[5.]DM, O. and Muraya, M.M., 2018. Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. American Journal of Applied Mathematics and Statistics, 6(6), pp.266-271.

[6.] Dutta, P., A Study on Machine Learning Algorithm for Enhancement of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science,2021.

[7.] Vaidya A. Predictive and probabilistic approach using logistic regression: application to prediction of loan approval. In2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2017 Jul 3 (pp.1-6).

[8.] Ruzgar, B., and Ruzgar, N.S., 2008. Rough sets and logistic regression analysis for loan payment. International journal of mathematical models and methods in applied sciences, 2(1), pp.65-73.

[9]. Bagher pour, A. (2017), Predicting Mortgage Loan Default with Machine Learning Methods, University of California, Riverside.

[10.] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, K. Vikas, ―Loan Prediction by using Machine Learning.

[11.] X. Francis Jency, V.P. Sumathi, Janani Shiva Sri, ―An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients‖, International Journal of Recent Technology and Engineering (IJRTE), Vol. 7, No. 48, pp. 176-179, 2018.

[12.] Briceno Ortega, Ana Cecilia, and Frances Bell. "Online social lending: borrower generated content [C]." AMCIS 2008 Proceedings, 2008. 380.
[13] Dosalwar, S., Kinkar, K., Sannat, R., & Pise, D. N. (2021). Analysis of Loan Availability using Machine Learning Techniques. International Journal of Advanced Research in Science, Communication and Technology, September, 15-20.

[14] Dutta, P. (2021). A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science, 3.

[15] Vaidya, A. (2017, July). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[16] Hamayel, M. J., Mohsen, M. A. A., & Moreb, M. (2021, July). Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine. In 2021 International Conference on Information Technology (ICIT) (pp. 33-37). IEEE.

[17] Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), 18-21.

[18] Tejaswini, J., Kavya, T. M., Ramya, R. D. N., Triveni, P. S., & Maddumala, V. R. (2020). Accurate loan approval prediction based on machine learning approach. Journal of Engineering Science, 11(4), 523-532.

[19] Meenu Vijarania, Ashima Gambhir, Deepthi Sehrawat, Swati Gupta, Prediction of Movie Success Using Sentimental Analysis and Data Mining, Applications of Computational Science in Artificial Intelligence, Pages 174-189,2022.