Predicting the level of Income Qualification for Bank loan Approval

[1]Mrs.B.Gnana Prasuna , [2]Mrs.M.Swathi Sree, [3]Mrs.K.Srilatha

Assistant Professor of CSE
Stanley College of Engg. & Tech. for Women, Abids, Hyderabad.

***Abstract*- Every person relies on banks and the loans offered by national and domestic banking sectors due to the banking and financial sector's rapid growth. In India 67% people are rely on loans to meet their financial needs. Banks receive numerous loan applications daily from customers and other people, but not all of them are approved. Banks often handle loan applications after confirming and assessing the applicant's eligibility, which is a difficult and time-consuming process. The majority of lenders use their credit score and risk assessment algorithms when reviewing loan applications and deciding whether to approve loans. In spite of this, some applicants miss payments on their bills every year, costing financial institutions a sizable sum of money. In this work, machine learning (ML) algorithms are used to identify trends in a dataset of loans that have been granted and make predictions to find the deserving loan applicants. Customers' prior information, including age, income type, loan annuity, most recent credit bureau report, employer type, length of employment, and family history, will be used to conduct the study. In this paper, we primarily concentrate on determining the family's level of poverty as well as the applicant's credit score and risk assessment to determine whether they are eligible or not.**

***Keywords*- Random forest, cross validation, Machine Learning.**

I. INTRODUCTION

The distribution of loans is essentially every bank's core activity. The majority of a bank's assets consist of the profit generated from loans that the bank has disbursed. Placing their money in trustworthy hands is the main objective of the banking industry. Loans are now frequently granted by banks and other financial organizations following a protracted verification and validation process, but there is no assurance that the chosen applicant is the most deserving of all applicants. The banking industry, like many other businesses, is increasingly striving to take use of the opportunities provided by contemporary technologies to enhance their operations, boost productivity, and cut costs. The most popular machine learning function used today for applications in the banking industry is predictive analytics. The majority of lending platforms' capacity to assess credit risk determines whether they are successful or not. Because of the everyday growth of data brought on by the banking sector's digitization, people prefer to apply for loans online. The use of machine learning (ML) as a standard method for data analysis is growing in popularity. Calculations are being used by individuals from various industries to address issues depending on their industry expertise. It's challenging for banks to get loans approved. The likelihood of making a mistake is high since bank workers must oversee a large volume of applications every day. Your eligibility for a loan depends mostly on your income and your ability to repay it. Eligibility for loans is also based on a few other considerations. This essay takes the applicant's family's overall poverty level, age, and regular payments—like rent or any other EMIs—into account. This essay aims to present a quick, easy, and effective procedure for choosing competent candidates. It could offer the bank special advantages. Each characteristic involved in loan processing may be automatically evaluated by the Loan Prediction System, and the same features are processed in accordance with their corresponding weights on new test data. It is possible to set a deadline for the applicant to learn whether or not their loan will be accepted. You can quickly access a single application and prioritize its evaluation using the Loan Prediction System. This strategy enables you to focus on particular.

II. LITERATURE REVIEW

An assertion regarding what one anticipates will happen in the future is known as a prediction. Every day, predictions are made. While some were very serious and based on mathematics, others were only educated guesses. Predicting what will happen in the future, whether it be in a few months, a year, or ten years, can help us in a number of ways. A subset of advanced analytics known as predictive analytics uses a number of techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data and produce forecasts. In order to ensure the best possible predictive models, Ugochukwu.E. Orji et al. [1] created ML models that obtained high performance accuracy in predicting loan eligibility by applying ensemble ML methods (bagging and boosting). Client loan prediction utilizing supervised learning techniques for loan applicants as valid or fail-to-pay clients has been proposed by L. Udaya Bhanu et al. [2]. This article compares the ways in which different algorithms are implemented in order to forecast consumer loans. Logistic regression, random forest, KNN, SVM, and decision tree classification were used to achieve the best results. Random forest produces the most accurate outcomes out of all algorithms. In order to forecast deserving loan applicants, Miraz A Mamun et al. [3] employed machine learning (ML) algorithms to uncover patterns from a common dataset of loan approvals. The study will be conducted using historical customer data, including age, income type, loan annuity, most recent credit bureau report, employer type, and length of employment. The most relevant features—those that have the greatest influence on the prediction outcome—were found using ML techniques like Random Forest, XGBoost, Adaboost, LightGBM, Decision Tree, and K-Nearest Neighbor. Using common measures, these algorithms are

contrasted and evaluated against one another. The highest accuracy of these was 92%, attained through logistic regression. Additionally, it was chosen as the best model since it outperformed other machine learning techniques in terms of F1-Score, which is 96%, by a wide margin. Support Vector Machine (SVM) and Random Forest (RF) were utilized by Sachin Magar et al. [4] to forecast a borrower's loan eligibility. Rithulaa et al. [5] found that while evaluating several machine learning techniques to forecast client eligibility for mortgage loans, both the LR and SVM models accurately forecast eligibility with estimated RMSE values of 81.62%.

Using machine learning and the decision tree technique, Dr. C. K. Gomathy and colleagues [6] have developed a loan prediction system that automatically identifies qualified candidates and returns the data on qualified consumers to a CSV file.

According to predictions [7], predictive analytics will be the most often utilized machine learning feature for applications in the worldwide banking sector in 2020.On the other side, the boosting method has no effect on a special class of algorithms tasked with transforming poor learners into strong learners.

Z. Tian [8] introduced a more efficient method known as the Gradient Boosting Decision Tree. With proper data preprocessing and feature selection, models are compared based on their performance. Gradient Boosting Decision Tree, has been proven to be one of the best that obtain the highest accuracy (92.19%), f1 score (91.83%), and AUC value (0.97). The experiment proved that this model has the best ability of classification and generalization.

C. R. D. Devi and R. M. Chezian [9] says that the commercial credit company's bankruptcy and potential major butterfly impact are both plausible outcomes of the extreme credit risk. The credit rating method used by the commercial credit company for its clients is essential in relation to the issue raised above. The expert system, rating system, and credit scoring system are the three basic techniques for estimating credit risk.

In the age of big data, a significant quantity of data can be used to determine a person's creditworthiness. We may use this mathematical model to identify the credit quality of customers [10] by training it using previous data on the borrower's credit behavior and ultimate credit rating over a predetermined period.

III. METHODOLOGY

In the flow chart (Fig. 1), the proposed model's design is displayed. The main goal of this paper is to identify patterns in the datasets that are used in the loan sanctioning process and to build a model based on those patterns. Predict the accuracy using the ML classification algorithm.

Making sure the appropriate people receive enough assistance is a challenge for many social programs. When a program focuses on the poorest demographic of the population, it can be difficult. This population group is unable to submit the requisite income and expense records to demonstrate their eligibility.

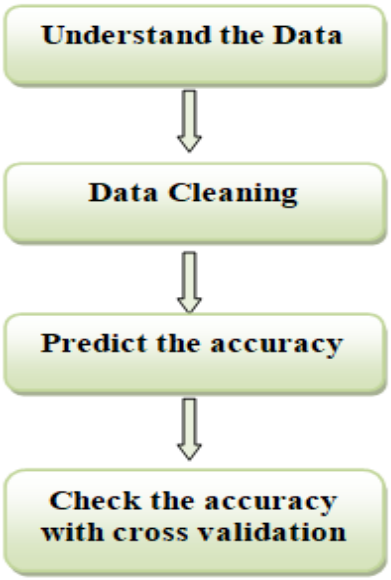

Fig 1:Work flow

This paper mainly focuses on what parameters are mainly impacting the loan sanction. To find the dependencies the methodology of the work is progress as bellow

- Understand the type of data.
- Examine your dataset to see if there are any biases.
- Determine the family's overall poverty level, including that of the head of the household.
- Use a random forest classifier to forecast accuracy. Use a random forest with cross-validation to test the accuracy.

In the first step of data cleaning remove null values and check for bias in the data. Clean all these then set the poverty level of the family which will find the bias in the data

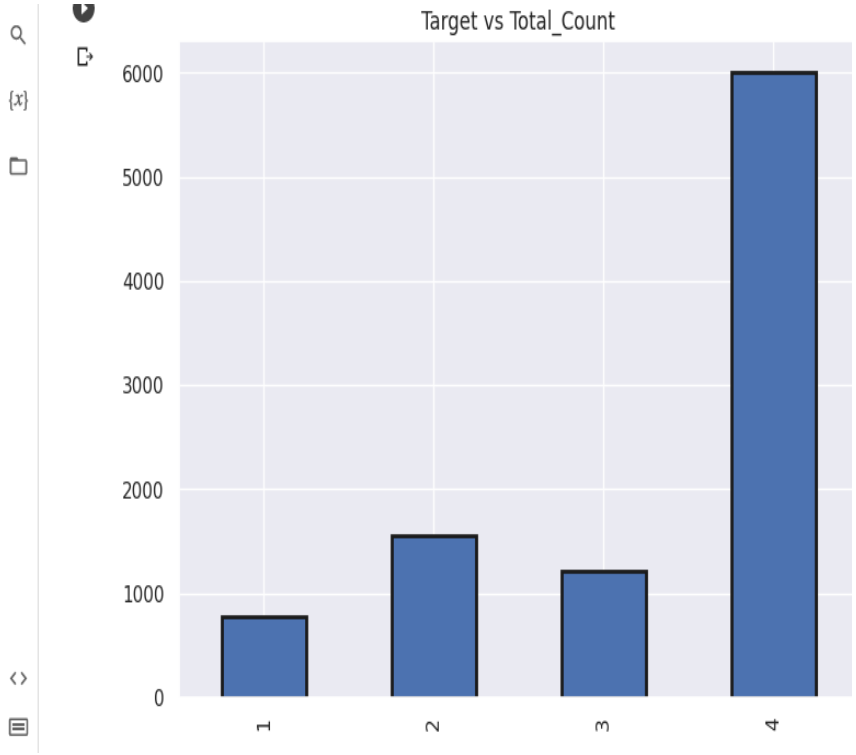

Fig 2

The above results clearly show that extreme poverty is the smallest count in the train data set. Which means the dataset is biased. Now verify the amount of poverty for each household member. Look to see if any homes lack a family head. Set the family's head of household and its members to the same degree of poverty. Set the family's head of household and its members to the same degree of poverty.

The accuracy of the model is precession and is 93%, and f1 score is 92%. By applying cross validation we will come to know which attributes are majorly impacting the model.

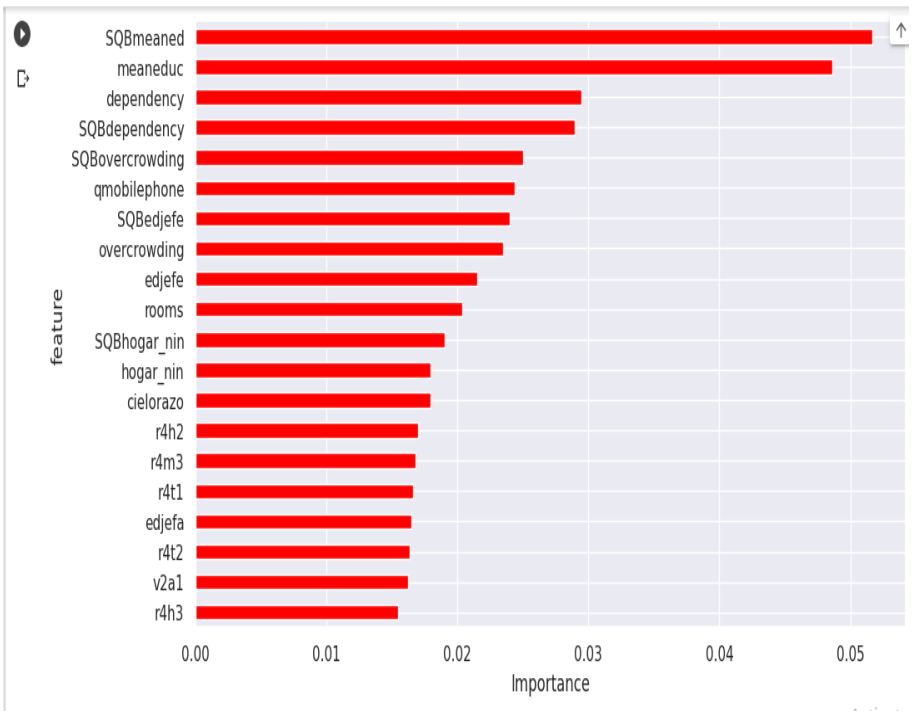

Fig:3

From the above we can conclude that meaneduc, dependency, overcrowding has significant influence on the model.

IV. RESULT AND CONCLUSION

The Random Forest classifier is used in this study to forecast accuracy and to cross-validate the results to determine which attributes have the most influence on the model. We can draw the conclusion that the Forest classifier, with 93% accuracy, produces better outcomes.

REFERENCES:

1. Ugochukwu .E. Orji, Chikodili .H. Ugwuishiwu, Joseph. C. N. Nguemaleu, Peace. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," 978-1-6654-7978-3/22/$31.00 ©2022 IEEE .
2. L. Udaya Bhanu , Dr. S. Narayana," Customer Loan Prediction Using Supervised Learning Technique", International Journal of Scientific and Research Publications, Volume 11, Issue 6, June 2021.
3. Miraz Al Mamun, Afia Farjana and Muntasir Mamun," Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis", Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management, Orlando, Florida, USA, June 12-14, 2022.
4. Sachin Magar , N.S.Nikam , Nilesh Taksale , Suprem Hajare, "Loan Eligibility Prediction using Machine Learning Algorithm", © 2022 JETIR August 2022, Volume 9, Issue 8 www.jetir.org (ISSN-2349-5162.
5. Rithulaa. A, Nidhhi. S, "LOAN ELIGIBILITY PREDICTION USING CLASSIFIERS", e-ISSN: 2582-5208 International Research Journal of Modernization in Engineering Technology and Science Volume:03/Issue:04/April-2021.
6. Dr.C K Gomathy, Ms.Charulatha,Mr.AAkash ,Ms.Sowjanya, "THE LOAN PREDICTION USING MACHINE LEARNING", © 2021, IRJET | Impact Factor value: 7.529 | ISO 9001:2008 Certified Journal | Page 1322.
7. "Most commonly used A.I. application in investment banking worldwide 2020, by types." Statista, 15-Sept-2021 [Online]. Available: https://www.statista.com/statistics/1246874/ai-used-in-investment-banking worldwide-2020/ [Accessed: 29-Jan-2022].
8. Z. Tian, J. Xiao, H. Feng, & Y. Wei. "Credit risk assessment based on gradient boosting decision tree." Procedia Computer Science. 2020, Vol.174, pp.150-160.
9. C. R. D. Devi and R. M. Chezian, "A relative evaluation of the performance of ensemble learning in credit scoring," 2016 IEEE International Conference on
10. B. Zhu, W. Yang, H. Wang and Y. Yuan, "A hybrid deep learning model for consumer credit scoring," 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, 2018, pp. 205-208