Machine Learning Based Model for Prediction of Loan Approval

2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) | 978-1-6654-6756-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIEM54221.2022.9853160

Bhanu Prakash Lohani
Assistant Professor
Amity School of Engineering and Technology
Amity University
Greater Noida, India
bplohani@gn.amity.edu

Mayank Trivedi
Amity School of Engineering and Technology
Amity University
Greater Noida, India
2219rashit@gmail.com

Ridhik Jeet Singh
Amity School of Engineering and Technology
Amity University
Greater Noida, India
ridhikjeetrjs@gmail.com

Vimal Bibhu
Assistant Professor
Amity School of Engineering and Technology
Amity University
Greater Noida, India
drvimal@gn.amity.edu

Shiv Ranjan
Assistant Professor
Amity Business School,
Amity University
Greater Noida, India
sranjan@gn.amity.edu

Pradeep Kumar Kushwaha
Assistant Professor
Amity School of Engineering and Technology
Amity University
Greater Noida, India
pkkushwaha@gn.amity.edu

***Abstract*— Banks are vital to financial management and controlling the economy of a country. Banks and financial institution distribute loans and these loans act as the core business part of almost every banks. The profits are earned from the loans distributed by the banks. The prime goal is to invest their assets in safe hands. The success of bank depends on the decision-making capability to evaluate risk of lending loan to the customer. Checking manually individual consumer's credibility for the loan approval is difficult, time consuming and risky. Thus, the banks aim to minimize the credit risks of defaulting. In this study we have applied logistic regression as a tool to predict whether an applicant is eligible for the loan or not. The data is collected from the Kaggle for studying and prediction.**

Keywords*— *Artificial Intelligence, Machine Learning, Automation, Data Mining, Predictive Analysis, Logistic Regression

I. INTRODUCTION

World is progress at rapid pace towards the automation, the automation of clearly defined and mundane operations. Artificial Intelligence is one of such field that excites scientists, technologists, and futurist regarding the development of automation. Huge amount of data is being generated with the digitization of banking sector, this data acts as the fuel for artificial intelligence and machine learning. Field of artificial intelligence deals with machines and computers to replicate human intelligence through programming [6]. A machine that can work at a human level intelligence can perform mundane tasks repeatedly using fewer resources. Scientists and engineers have obtained a thorough understanding of artificial intelligence over time, resulting in the creation of applications to solve real-world challenges and issues.

Regarding this paper, our main objective is to use machine learning, a subset of artificial intelligence, to predict loan safety in an efficient and non-biased manner. Machine learning is an AI application that allows you to develop analytical and predictive models without having to programme the system explicitly. Machine learning algorithms use data to find patterns and make inferences. There are various machine learning tools that can tailored for different purposes and requirements. Machine learning is currently being used by businesses to improve their operations, and the results have been quite favorable. The logistic regression approach is used to forecast the loan's safety. Logistic regression is a type of supervised machine learning algorithm which uses labelled dataset. The dataset includes features like- Education Gender, Income, Self-employed, Married, Loan Term, and more.

II. LITERATURE SURVEY

A prediction is a forecast and about what will occur or might happen in the future or in coming days. Predictions can scientific or just mere guesses. In predictive analytics we analyze data and make scientific predictions using machine learning, statistics, probability, data mining, and data engineering ideas. Specifically, using data mining for banking industry data helps us in enhancing the accuracy of predictions. A prominent machine learning approach for solving classification issues is logistic regression. The logistic regression method is a predictive analysis tool that explains the relationship between variables in a dataset. The dataset is collected from Kaggle. This evaluation of the literature aids us in carrying out our research and developing a trustworthy bank loan prediction model. Binary logistic regression can help banking industry assess credit risk. It is a time and resource consuming job to identify the characteristics of people who are likely to default on loans. Then we want to identify good and bad credit risks. For a bank with over many applicants for the loan it can be a tedious work [9][10][11].

978-1-6654-6756-8/22/$31.00 ©2022 IEEE

III. METHODOLOGY

As depicted in Fig. 1, the proposed methodology starts with the collection of data. Here, we have used a public dataset from Kaggle for exploration. Then, we clean and filter the data according to the requirement. In next phase, for the purposes of training and validation of the machine learning method, we divided the dataset into training and test datasets. In the final phase, we measure the accuracy of the model.

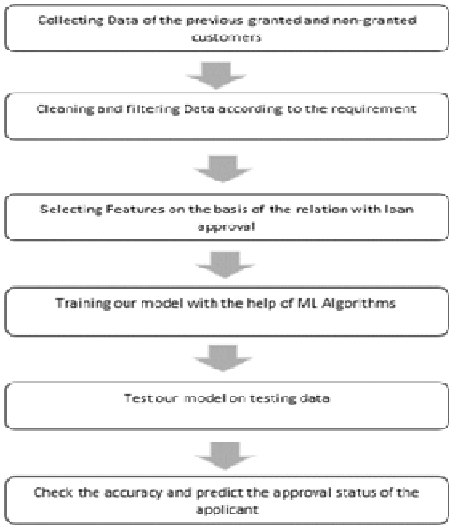

Figure 1 Proposed Methodology of Machine Learning for Prediction-

IV. DATASET DESCRIPTION AND PRE-PROCESSING

The dataset for the bank loan prediction model comes from Kaggle and includes applicants of various ages and genders. Education, marital status, income, and other characteristics are included in the dataset. We manage the missing value and normalize the data for future processing during pre-processing and feature engineering. The dataset is then split into two parts: training and testing. The training dataset is used to train the machine learning model, which is then tested and validated using the test dataset. The last step of pre-processing is to test the correlation between data attributes to find the most significant feature in prediction process.

A. Machine Learning

Predictive analytics is a type of data analysis that is used for the purpose of predicting about future events by analyzing the data. Predictive analytics includes concepts like machine learning, data mining and system modelling [4]. Machine learning is the science of getting computer systems to act and perform operations without being explicitly programmed for it. Machine learning helps to build such powerful models that can automatically learn from data which can predict future events and propose solutions without human intervention.
Machine learning is sub-divided into:

1. ***Supervised Learning (SL)***: Supervised learning is defined using the labelled dataset to train algorithm to classify data or predict outcome correctly. Both, input and output, datasets are labelled. Logistic regression is a supervised learning algorithm.

2. ***Unsupervised Learning (UL)***: In unsupervised learning, algorithms analyze and cluster the unlabeled dataset. The algorithm itself discover the hidden patterns or grouping without the need of any kind of human intervention.

3. ***Reinforcement Learning (RL)***: Reinforcement learning is a machine learning method that primarily focuses on agents and actions the agents ought to take in an environment. It is also concerned with the state of environment and agent to maximize the gain of reward. Reward is gained by the agent by performing the actions in environment.

B. Algorithms used for Prediction

For loan approval, the answer must in binary outputs such as “true” and “false” or “yes” and “no”. Hence, the logistic regression is selected for classification. It is a type of statistical analysis used for predictive modelling. Logistic regression can help us predict the likelihood of an event happening. It takes independent features and produces a categorical result. The chance of a category output occurring can also be found in the logistic curve shown in Fig. 2.

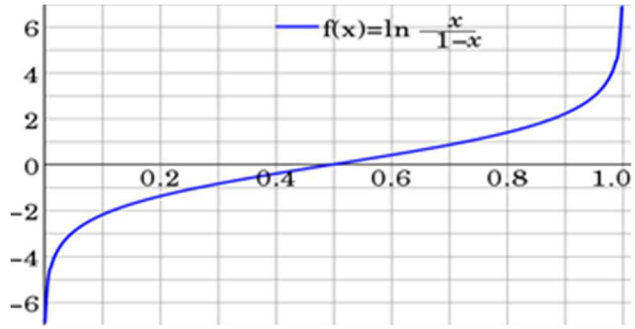

Figure 2 shows a general logit curve

Many different models can also be developed using Support vector machine, linear regression, or random forest to solve the defined problem. Logistic Regression model was selected because of its mathematical clarity and flexibility. Predictive models built using this approach can make positive difference in industry because the models help us understand relationships between attributes. For building the logistic regression model, we have used python as programming language. The main objective for using python over other programming languages is its readability, flexibility, good visualization options and having rich library ecosystem.

V. RELATED WORK

In the banking industry, the basic criterion is required for the approval of loan. Decision-making is a complex process including a number of factors that must be considered at all times. The model's output can take one of two forms: Accept or Reject. The deduced model is aimed to reduce the time and resources to make the decision of loan approval without compromising on the accuracy and maintain the standards of banks.

VI. PROBLEM STATEMENT

Dream Housing Finance is a corporation that deals with home financing. It can be found in both urban and rural regions. Buyer initially applies for a house loan with the company, which then verifies the buyer's credentials and determines whether or not the loan's eligibility criteria have been met. However, executing this work by hand takes a long time. As a result, automating the procedure (in real time) based on client information is advantageous. As a result, the final task is to simply identify the buyers who are suitable for a bank loan. The second question is how the business will benefit if we provide client segments. As a result, the more precisely and effectively the model predicts loan-eligible consumers, the more useful and effective it is for the Finance Company. The system of bank loan approval and its customer and staff interaction with the system is presented by use case diagram in fig.3. The decision making to define the eligibility of the customer is presented by activity diagram in fig. 4. The object sequence process is mentioned by sequence diagram in fig. 5.

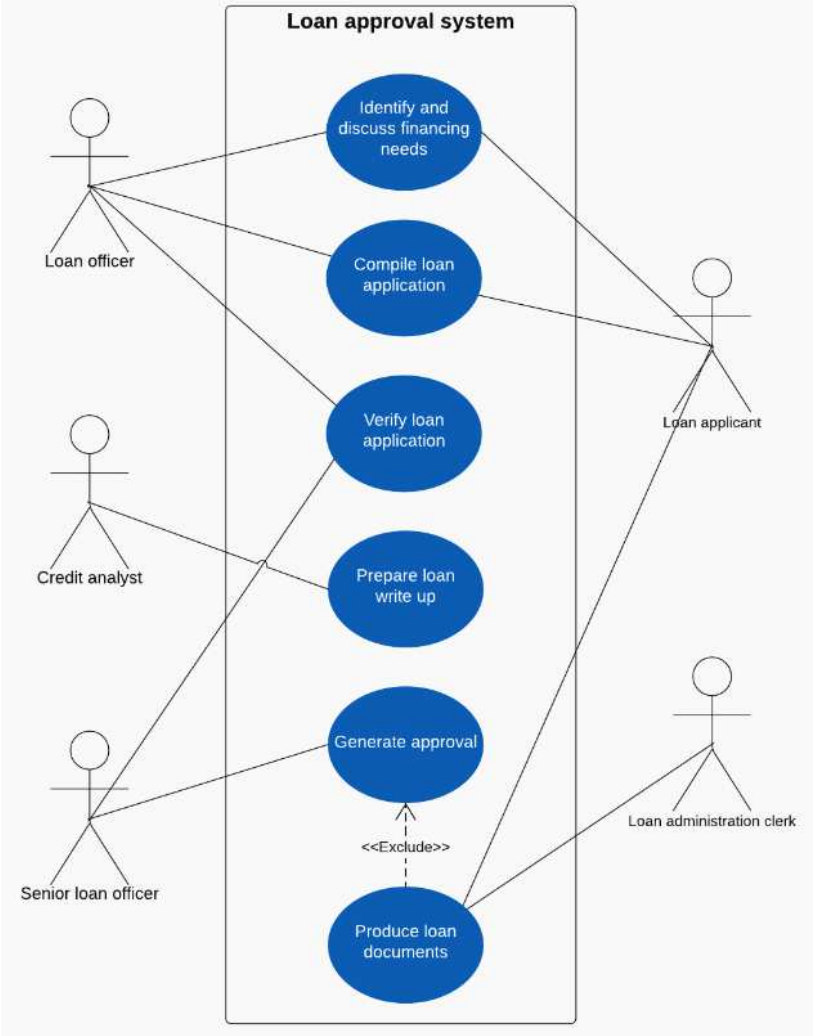

Figure 3. Use Case Diagram

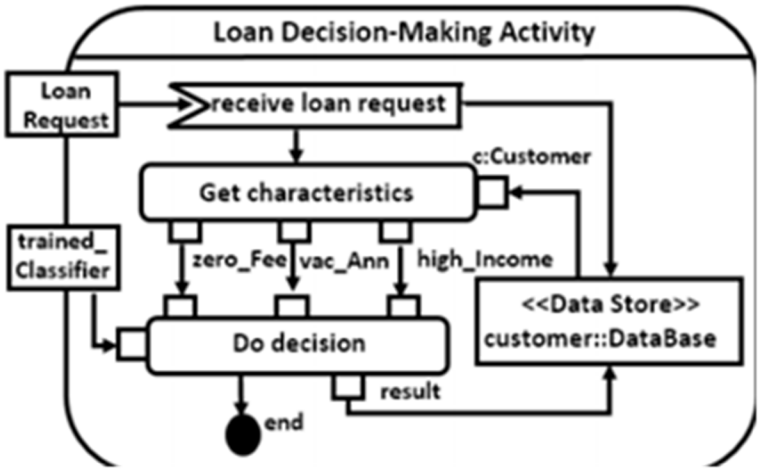

Figure 4 Activity Diagram

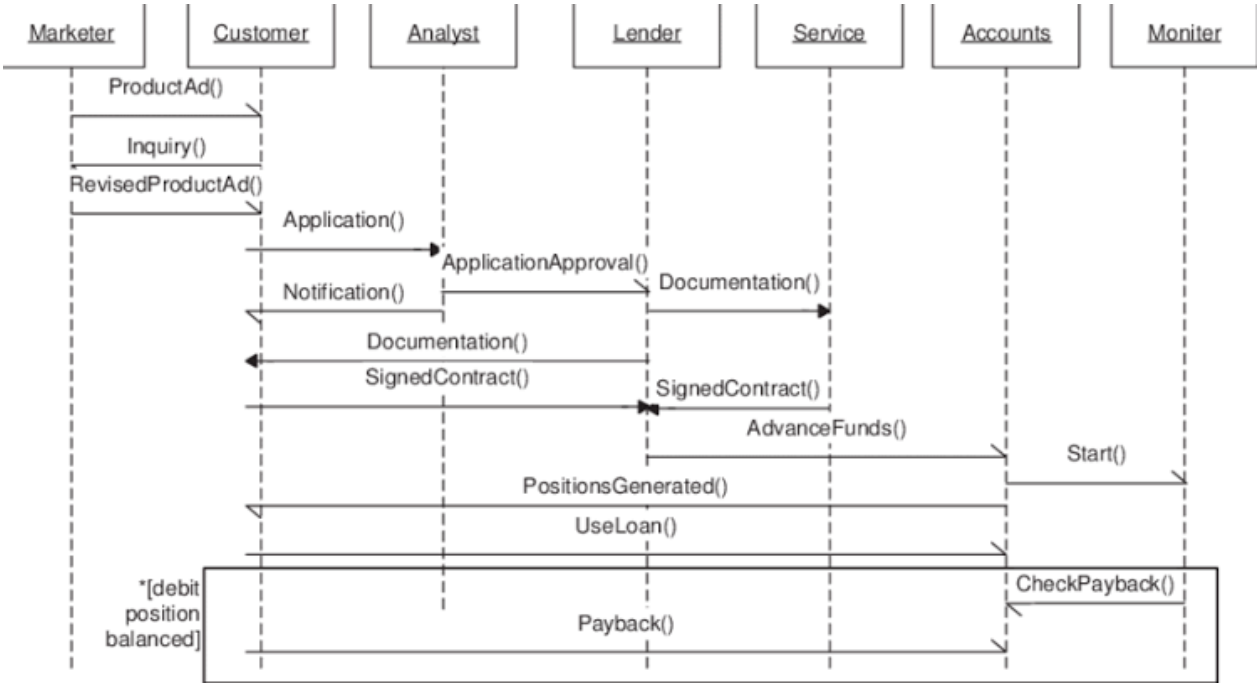

Figure 5 Sequence Diagram

VII. DATASET

The goal to be attained, as mentioned in the preceding section, is a classification problem, since it must categorize whether or not the loan should be provided after the required requirements are met. The dataset to apply the regression method through the machine learning algorithm is given in table 1.

A variety of classification algorithms can be used to solve it;

1. Logistic Regression (LR)
2. Decision Tree Algorithm (DTA)
3. Random Forest Technique (RFT)

TABLE I. DATASET DESCRIPTION

Table 1 Dataset introduction

Variable	Data-type	Description
Loan_ID	Categorical (non-numeric)	Unique loan ID
Gender	Categorical (non-numeric)	Male/female
Married	Categorical (numeric)	Yes/no
Dependents	Categorical (numeric)	No. of dependents
Education	Categorical (non-numeric)	Graduate/not-graduate
Self_employed	Categorical (non-numeric)	Yes/no
Applicant Income	Numeric feature	Applicant income
CoapplicantInc	Numeric feature	Co-applicant
ome		income
LoanAmount	Numeric feature	Loan amount in thousands
Loan_Amount _Term	Numeric feature	Term of the loan in months
Credit_History	Categorical (numeric)	0/1
Property_Area	Categorical (non-numeric)	Urban/semi-urban/rural
Loan_Status	Categorical (non-numeric)	Yes/no

VIII. RESULT ANALYSIS

The inferences we can draw from the experiment are:

1. When compared to non-married people, married people/couples have a higher percentage of house loan approval.
2. The percentage of applicants with 0 or 2 dependents who were approved for a house loan.
3. Graduates are given loan more frequently and easily as compared to non-graduates,
4. It is not a matter of concern whether the customer is self and
5. Model has significantly reduced the time and resources used by banks in decision making process employed or not as there is almost no correlation between Self-employed and Loan Status data points of applicants.

The result and probability of approval and outcome with the defined basis in given in fig. 6 and final regression value of the result is given in fig. 7/

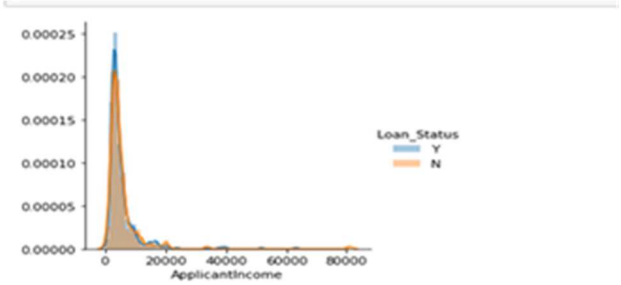

Figure 6. Probability of approval and outcome on different basis

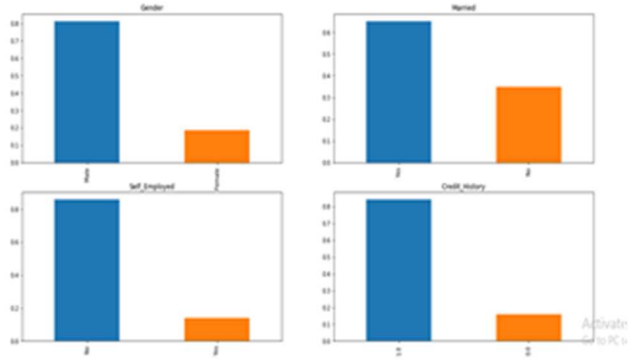

Figure 7. Final Regression Value Analysis

IX. CONCLUSION AND FUTURE SCOPE

From an analytical viewpoint and constraints on the component, it can be concluded that model is meeting all the requirements. After observing the results, it is clear that the prediction models can be integrated into the banking management system to reduce the credit risk and default rates for the loans. This component can be easily plugged into other systems too. The machine learning model is prepared on the trained and tested on data but in the future, it is possible to upgrade and further develop the predictive accuracy of model by using realistic big data from banking industry. The pre-processing phase starts with understanding of data, data cleaning, outlier detection, and removal and understanding problem statement. Logistic regression is used for its simplicity and had showed great performance.

Limitations and deductions:-

- For training, logistic regression necessitates a large sample size
- For predictive analysis, it imposes independent variables
- A reflection of bias in human decision making can be seen in model output

REFERENCES

[1] "Prediction Analysis of Cancer Cells Using ML Classification Algorithms", *Indian Journal of Public Health Research & Development*, 2021. Available: 10.37506/ijphrd.v12i2.14115.

[2] A. Jain, Y. Sharma and K. Kishor, "Prediction and Analysis of Financial Trends Using Ml Algorithm", *SSRN Electronic Journal*, 2021. Available: 10.2139/ssrn.3884458.

[3] N. Cholli, "Machine Learning Classification Models for Banking Domain", *SSRN Electronic Journal*, 2019. Available: 10.2139/ssrn.3363076.

[4] H. Meshref, "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms", *International Journal of Circuits, Systems and Signal Processing*, vol. 14, pp. 914-922, 2020. Available: 10.46300/9106.2020.14.117.

[5] M. J. Hamayel, M. A. Abu Mohsen and M. Moreb, "Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine," *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 33-37, doi: 10.1109/ICIT52682.2021.9491636.

[6] A. Shivanna and D. P. Agrawal, "Prediction of Defaulters using Machine Learning on Azure ML," *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2020, pp. 0320-0325, doi: 10.1109/IEMCON51383.2020.9284884.

[7] P. Kirubanantham, A. Saranya and D. S. Kumar, "Credit Sanction Forecasting," *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, 2021, pp. 155-159, doi: 10.1109/ICCCT53315.2021.9711790.

[8] R. Ahn, S. Supakkul, L. Zhao, K. Kolluri, T. Hill and L. Chung, "A Goal-Oriented Approach for Preparing a Machine-Learning Dataset to Support Business Problem Validation," *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2021, pp. 282-289, doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech52372.2021.00057.

[9] A. Chhillar, S. Thakur and A. Rana, "Survey of Plant Disease Detection Using Image Classification Techniques," in IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO' 2020), Noida, India, 2020, pp. 1339-1344, doi: 10.1109/ICRITO48877.2020.9197933.

[10] S. Ghosh, A. Rana, V. Kansal, "A Novel Model Based on Nonlinear Manifold Detection for Software Defect Prediction" in Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018, pp 140-145 (2019).

[11] S. Ghosh, A. Rana, V. Kansal, "A statistical comparison for evaluating the effectiveness of linear and nonlinear manifold detection techniques for software defect prediction" in International Journal of Advanced Intelligence Paradigms , Vol. 12, pp 370-391 (2019).