Loan approval prediction using Machine learning (ML) Techniques

Vamsi Sri Naga Manikanta Murukonda
20010620-T144
vamu21@student.bth.se

Lahari Gaddam
20000917-T083
laga21@student.bth.se

Avan Gogineni
20000821-T138
avgo21@student.bth.se

Sree Lakshmi Hiranmayee Kadali
20010920-T244
srkd21@student.bth.se

***Abstract*—The emphasis on the importance of making the right decisions in banks and financial institutions is not new. The complexity and quality of the information in investment sectors like banking and financial institutes are quite crucial. They require to take accurate, timely, and relevant decisions using this information. Not only it helps in increasing productivity but it also helps in boosting the service. The need for bank loans grows with the increase in people's demands. Because of the time-consuming and challenging process of loan approval, verifying and evaluating if a person is eligible for a loan or not with an automated process can save the lengthy process of approval and validation. This research aims to focus on this problem and help customer to check their credibility for loan approval thereby reducing the time spent for loan approval around financial vendors. A logistic Regression model is used in this expert system for predicting the result of loan approval. The dataset is of (11*401) size and contains different features. Four algorithms 'Random Forest', 'Support Vector Regression', 'Logistic Regression', and 'Ridge Regression' are used in identifying if the person is loan worthy or not after performing a literature review. Logistic regression is selected because of its highly accurate performance with 87% compared to other models.**

***Index Terms*—Loan Approval prediction, Machine learning, Python GUI**

I. Introduction

Credit scoring systems in financial institutions help assess the credit risk of borrowers to decide if the loan application can be approved or not. It gathers all the necessary information such as the customer's age, income type, loan annuity, last credit report, employment, period of employment, etc, and predicts the output. Selecting loans with high success and low risk is very critical when lending. Although human judgment and expertise are very crucial in determining whether to approve a loan or not, the DSS tool can help make effective decisions and accurate predictions with the use of the right machine learning models.

Generally, there are two types of credit scoring models, first which classifies applicants into classes like 'fund' or 'reject'/ 'good' or 'bad'. Secondly, the model discovers the patterns of customer repayment and updates the underwriting system for potential risks. In this study, we focused on an automated decision support system that shows if loan applicants are "worthy" or "not worthy".

For, this study we have developed a decision support system that uses the 'Logistic Regression' machine learning algorithm to accurately take appropriate decisions in selecting a loan application with low risk. In this procedure, the system determines if the person is worthy or not after he/she enters the required information like "Income", "No of cards he/she holds", "Previous Limit" and "Balance" as input. The input parameters are selected from the dataset after analyzing the principle features for loan approval. Furthermore, the system requires at least input two parameters, or else it doesn't display any result.

The developed system is a simple version of an automated credit risk decision support system that helps the loan application if he/she is loan worthy by himself. The system is limited in scope because we use very limited yet important features for predicting the result. Although human expertise is the ultimate decision maker in loan approval, this system can help the person or can be used in financial institutes for calculating credit scores quickly just by entering the values.

The dataset is made up of 11 rows and 401 columns with different features like age, income, balance, education, cards, rating so on. We have selected the four most important features in calculating the credit score. Four models are initially trained before model selection. When we anticipated the data using logistic regression, the model outperformed other models with an accuracy of 87%. As, a result a DSS model with the logistic regression approach is used. The proposed model-based DSS implementation is given in the following sections.

II. Division of Labour

We considered the contributions of team members in developing the decision support system for loan approval prediction into four parts: B: "Designer", C: "Programmer", D: Documenting, and E: "Reviewer(Proofreading)". The workload share for developing the system is divided based on his/her skills that perform best.

Name	Duty
Vamsi Sri Naga Manikanta Murukonda	B,C,D,E
Avan Gogineni	B,C,D
Lahari Gaddam	B, D, E
Sree Lakshmi Hiranmayee Kadali	B,D,E

Tasks completed for developing the system:

1. Researching and determining project problems and objectives.
2. Formulating the problem solution.
3. Work division between team members and scheduling meetings for implementation.
4. Choosing the appropriate dataset for credit risk analysis.
5. Performing literature review for algorithm selection.
6. Experimenting with different machine learning techniques after data preprocessing.
7. Choosing a single ML model which outperforms other models based on accuracy.
8. Creating the final ML model and code.
9. Streamlit with Python libraries is used for implementing the system's GUI and connecting it to the code.
10. Important features from the dataset are selected for taking input parameters from the user.
11. The feedback analyzer gives the output if the person is worthy or not.
12. Arranging meetings with teammates regularly after each task to keep the work progressing.
13. Documenting and reporting the project result.

III. Project Analysis

A. Background

[8] talks about how the logistic regression model helps in predictive analysis for studying loan defaulters. Researchers worked on an open-source Kaggle database to observe the performance of logistic regression and stated that it helps efficiently in predicting credit risk.

[9] researchers focus on analyzing different machine learning models with explainable AI to optimize investment decisions. For the study, various machine learning models like logistic regression, decision trees, AdaBoost, Random Forest, and sequential neural networks are selected. The results indicated that ensemble classifiers like Random Forest and Neural Networks outperformed.

[10] talks about how different algorithms help in prediction. Logistic regression and linear discriminant analysis are the most used techniques for providing scorecards whereas other techniques like ridge regression, support vector regression, random forest, and neural networks are not widely used because of their lack of transparency. Overall, the paper talks about the statistical aspects of credit score analysis.

The above research papers talk about how widely logistic regression, Random Forest, Ridge Regression, and support vector regression is used in predicting credit score. Therefore, we want to focus on the four algorithms for predicting loan approval.

B. Problem Definition

The project's main goal is to help financial institutions, banks, and loan applicants check if he/she is loan worthy or not. It helps loan vendors for making accurate and timely decisions compared to the time and resource-consuming process of loan approval. A Python GUI web application is developed for predicting the output after the user inputs the necessary parameters for prediction.

C. Problem Objectives

1. To find an appropriate dataset that has features appropriate for classifying the loan applicant as 'good(worthy)' or 'bad (not worthy)'.
2. Analysing the data set by principal component analysis and extracting main features which help in developing the most efficient machine learning model for loan worthiness.
3. Experimenting with different machine learning models, focusing mostly on models used quite extensively in loan approval prediction after the literature review and finalizing four models for our decision support system.
4. Implementing data preprocessing, and building models for performance comparison.
5. Selecting a highly reliable model which can analyze customer data and predict loan approval.
6. To visualize the loan approval prediction using Graphic User Interface for predicting the output.

D. Problem Solution

Predicting the parameter "rating" i.e. credit risk for loan approval and giving feedback if the person is worthy or not worthy based on the prediction. The prediction is only done if the user gives necessary information in input parameters like "Income", "Previous Bank Limit", "No of cards he/she holds" and "Balance".

A simple web application GUI is developed for the use of loan applicants, financial institutes, banks, and money lenders.

E. System Criteria

Some of the tools and approaches used for developing the decision support system for the project ”Loan Approval using Machine learning techniques” is given below:
Tools: Streamlit, Jupyter notebook
Programming Language: Machine learning using Python
Packages: Numpy, Pandas, sklearn

IV. Project Design

A. Use case Diagram

In simple words, a use case diagram can be described as a graphical representation of the user's interaction with the system and the actions performed. A use case diagram consists of 4 main elements following

- Actors, which resemble users.
- Systems, it is a boundary that consists of uses cases in it.
- Use cases are all the possible tasks that can be performed by the user.
- Relationships are an interaction between the user and the use cases that can be connected using different types of connections.

A proper use case gives a clear picture to understand the behavior of a system from the perspective of a user.

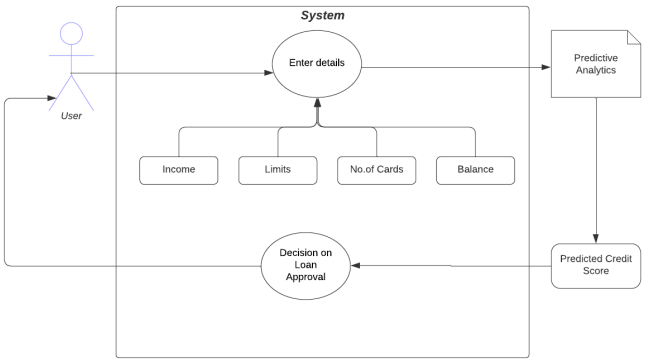

Fig. 1. Use Case Diagram

B. Activity Diagram

An activity diagram can be defined as a flow of actions in a procedure. It shows the step-by-step process of the actions in sequential order. By observing the below figure we can understand the workflow of the loan approval prediction system. The activity diagram starts with the ”start” node and continues with the internal steps and ends with the ”Decision For Loan Approval” node.

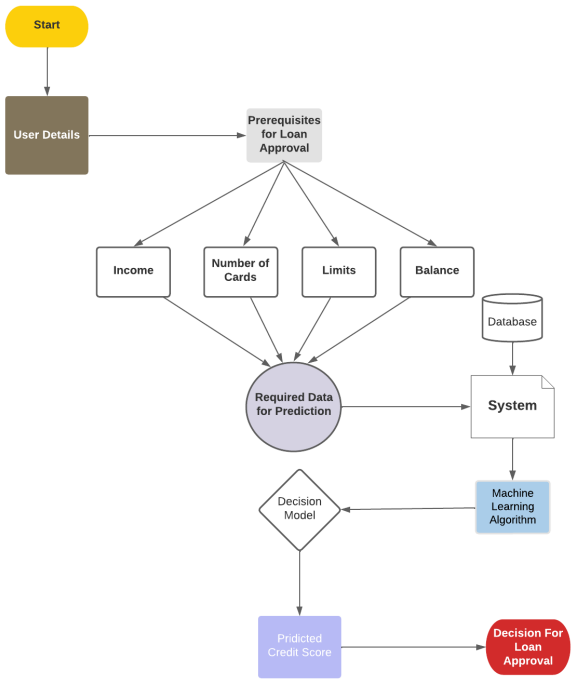

Fig. 2. Activity Diagram

C. Model Description

The user can use the graphical user interface (GUI) to give the inputs as asked. The user should enter the details such as Previous Limit, Balance, Income, and No. of cards. Upon clicking the ”Predict” button, the GUI sends the data to the machine-learning model which has been trained to predict the credit score of the corresponding user based on the analysis of the data set. Finally, the system decides whether the user can be approved for a loan if he/she has a decent or good credit score.

V. Project Implementation

The project has been done using various standard Python libraries. They are:

- **Streamlit:** It is an open-source Python framework widely used to create applications of machine learning and data science with ease. By using Streamlit, anyone without knowledge of front-end web development can easily build interactive web applications. It has all the basic user interaction elements such as input fields, check boxes, radio buttons, drop-downs, and buttons.

  In this web application, we have used text inputs, increment buttons, and simple buttons for the GUI.

- **Numpy:** It is one of the most widely used Python libraries which is used to perform large multi-dimensional array computations. It is a simple-to-use and robust library that can be used to manage data that is in the form of

arrays.

- **Pandas:** It is another widely Python library that is used for data manipulation and analysis. It has many applications such as data cleaning, data transformation, data exploration, and data manipulation. It offers the programmer various data structures that can be used to manage tabular data. Pandas library is used a number of times in this project to clean the data, transform the data, and many more.

- **Sklearn:** also called Sci-kit learn, is a Python library used to implement various machine learning models. Machine learning models such as Classification, Clustering, Regression, and other statistical modeling tools can be implemented using this library.
  In this project, we have used Sklearn to preprocess the data set and build various machine models to find the most accurate machine learning algorithm by validating them with the help of elements such as confusion matrices and accuracy.

A. Data Set

The 'credit.csv' data set that is utilized for the project is taken from GitHub provided by 'The UCL School of Public Policy'. The data set consists of 11 columns and 401 rows. The data set contains the data values of both numerical and variable values.

In this project, we have selected 4 columns among 11 columns. The columns are as follows

- Income
- Limit
- Cards
- Balance

By performing preprocessing of the data set the above columns are selected for final data.

B. GUI(Graphical User Interface)

In this project, we have used Streamlit to build a machine-learning web application that can predict the credit score of the user based on his input for a few fields. The GUI is made of a title, four input fields named Income, Limit, No. of Cards, and Balance, Increment button for each input field so that the user can press the buttons to increment or decrement button to change the value, and a submit button named 'Predict'.

Based on the inputs given by the user, the machine learning model computes the credit score of the user and displays a text message about whether the user is eligible for a loan.

C. Model

These are the machine learning models that are studied during the implementation of the project

- Logistic Regression
- Ridge Regression
- Support Vector Regression
- Random Forest

1) **Logistic Regression:**
   It is a machine-learning algorithm in which the probability of something happening is predicted. The algorithm takes a few factors or attributes and calculates the probability of the event using a logistic function. Logistic Regression is deemed to be most accurate in those cases where there are factors that might affect the probability of an event.

2) **Ridge Regression:**
   It is a variant of linear regression in which the algorithm find the best-fit line between independent variables and dependent variables. The concept of the "Ridge penalty" is what makes this variant special. The ridge penalty is where the algorithms add penalty coefficients to the regression equation to generalize the model. By doing this Noisy variables are reduced to make the model more accurate.

3) **Support Vector Regression:**
   It is a machine learning algorithm that is used to predict continuous variables that are numerical. It finds the best-fit line or curve that has the maximum distance between predicted values and actual data values. It is a very flexible and accurate algorithm that transforms linear data into a multi-dimensional space to find the relation between linear and non-linear variables with higher accuracy.

4) **Random Forest:**
   It is a machine learning algorithm that is simply a group of decision trees that are combined to predict more accurate values and classifications. Random Forest is very efficient as it can handle huge amounts of data, can recognize complex relationships between variables, and avoid overfitting of variables. It is widely used for its expected high accuracy and robustness.

In this project, we have trained and built all the models for the algorithms. However, we selected Logistic Regression as it got the highest accuracy of 87%.

VI. Conclusion

In this project, we successfully designed an interactive web application that is able to decide whether a user is eligible of getting approved for a loan. This application can be used to quickly check if a person can be approved for a loan based on his/her Income, Limits, No. of Cards, and Balance. Many machine learning algorithms such as classification, regression, and Clustering algorithms were trained and models were built. However, most of the regression algorithms got higher accuracy. Therefore, Logistic Regression which got

the highest accuracy is chosen for the web application. This machine learning model is linked to a web application with the help of an open-source Python framework called "Streamlit". A good-looking and simple web application is designed to make sure the user can have easy interaction.

This web application can be further improved by gathering more data, using other hybrid machine learning algorithms that have higher accuracy, and improving the graphical visualizations for the website by adding an interactive credit score meter to increase the readability for the user.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[9] Tyagi, Swati. "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions." arXiv preprint arXiv:2209.09362 (2022).

[10] Gilbert Saporta. Some Statistical Aspects of Credit Scoring. 3rd world Conf. on Computational Statistics Data Analysis, Oct 2005, Limassol, Cyprus. ⟨hal-01125140⟩