Building a Loan Approval Prediction Model Using ML Techniques

**1 Purushotam Naidu K\*, 2 K. Harika, 3K. Keerthi, 4M. Tejitha, 5D. Gayatri Sameera**

**1 Assoc.Prof., Dept. of Computer Science and Engineering (AI&ML), GVP College of Engineering for Women, Visakhapatnam, AP, INDIA.**

**2,3,4,5Student, Dept. of Computer Science and Engineering (AI& ML), GVP College of Engineering for Women, Visakhapatnam, AP, INDIA.**

# Abstract

Banks provide various services to the customers, viz. loans, insurances, multiple funds, but their significant profit comes from their loan or credit line. Banks take these loans that they credit to various customers, investors, global partners, and even among different banks.

The profit or loss of a bank largely depends on the loan that their customer takes and whether they can pay it back or not. By analysing defaulters of loans, banks can downgrade their Non-Performing Assets. The literature survey shows several methods for studying the loan default problems. Accurate guesses are very urgent for gain facilitation, so reviewing different ways and comparing them based on accuracy becomes urgent. This paper has made a practical approach for predictive analytics to predict loan defaulters. The further model has been trained to review different machine learning algorithms.Logistic Regression, Random Forest, Decision Tree, XGBoost (Extreme Gradient Boosting), Support Vector, Stochastic Gradient Descent, Gaussian NB, and K-NN. Results have been calculated on the dataset generated by Kaggle and predictive models to study and predict loan defaulters. Support Vector Clustering (SVC) model out-performs all the other tested algorithms with an accuracy of 88.7%, further raised to 91.2476% using Hyper-Parameter Tuning, along with an AUC validation score of 0.71

**Keywords:** Non-Performing Assets (NPA), Predictive Analytics, Random Forest (RF), XGBoost (XGB), Support Vector Classifier (SVC), AUC Validation Score

# 1.Intoduction

Loan approvals play an extremely important role in the financial sector. Loan approvals do not merely enable organizations to be at the helm of their credit risks; they ensure that the credit granted is unbiased and fair too. There are disadvantages involved with the traditional loan approach based on the manual computation of rules and standards which are predetermined. It is time-consuming and prone to mistakes because man is not perfect. It has been a revolution for the automation of predictions of loan approvals with increasing historical loan data.

Problem lies within the fact that the prediction of loan approval is actually a binary class where an output is going to either be "approved" or "rejected." All those aspects are being affected majorly, which are-applicant income, credit history, loan amount, employed status, and marital status. All these variables formed a part of the train dataset provided for training: train\_u6lujuX\_CVtuZ9i.

The conventional evaluation process also has some disadvantages, like decision-making bias, a non-efficient way of handling high volumes of applications, and it is not scalable. Machine learning eliminates all of these issues by finding a pattern in historical data; prediction is automated, allowing constant, data-driven decision-making.

This project deals with the development and application of ML models, particularly the Logistic Regression algorithm, for accurate prediction of loan approval decisions. With the careful pre-processing of a dataset and advanced modelling techniques, this project shows how

ML can be used to increase efficiency, minimize biases, and simplify the loan approval process.

# 2.Literature Survey

In [1] Logistic regression is developed as a learning tool for predictive and probabilistic decision-making in forecasting and loan approval prediction. Its emphasis is on the mathematical development, its application in deciding approval from applicant records, and even extends that scope further into other real-world applications of such an approach.

In [2] Financial institutions leverage data-driven models for target marketing and risk management. On Malaysian bank data a 98% accuracy could be achieved in loan responses with decision trees slightly above logistic.

In [3] model explores the application of logistic regression as a tool in machine learning for predicting decision making in loan approvals. It elaborates the mathematical representation of logistic regression to how it may be applied on determine whether or not a loan shall be approved through applicant information. Other application of this model that exits in the real world are also featured in the paper.

In [4] studies the performances of SVM and logistic regression models for loan defaults predictions are compared. The study makes use of the data obtained from Equity Bank, Kenya between the years 2006 and 2016.Variables used in the study include credit history, loan purpose and employment statues among the 1000 loan applicant. Accuracy was found to be higher when using the SVM (linear kernel) model compared to that of logistic regression (SVM: 88.29% train, 86.12% test; Logistic regression: 77.27% train, 73.33% test) and precision. The study recommends using SVM for better loan default prediction.

In [5] it compares the loan default prediction models through logistic regression and support vector machine (SVM) for Equity Bank in Kenya. The research was conducted on 1,000 loan applicants by taking into account factors like credit history and loan purpose. SVM model performed better than logistic regression in accuracy and precision with a high percentage of 88.29% vs. 77.27% train, 86.12% vs. 73.33% test. It is advised that SVM should be adopted in the financial institutions for precise loan default predictions.

In [6] it predicts loan defaults in banking can help in lowering Non-Performing Assets and sustaining loan income. This research compares different methods for predicting loan defaulters, focusing on the Logistic Regression model. It used data from Kaggle to compare performance in terms of sensitivity and specificity. The results of this research show that including personal attributes such as age, credit history, loan purpose, etc., improves the predictions, and therefore banks should consider more than wealth when granting loans to reduce default risk.

In [7] Advanced lending is the heart of all financial sectors but it involves high risk of credit exposure. To mitigate these, human entities are given FICO scores to assess reliability. AI algorithms such as Logistic Regression and Decision Trees are gaining acceptance in the field to forecast credit risk with historical data analysis. This paper is a survey of existing models using these AI techniques to analyse risk.

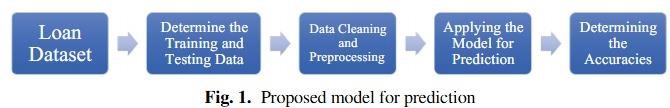
In [8] Loan eligibility is one important determinant in lending. However, such a process is too time-consuming with traditional methods. Machine learning algorithms, including Logistic Regression, can accelerate the decision to within minutes or seconds. This study applies Logistic Regression for loan approval predictions since it is also providing the capability for binary classification. Although evolution metrics were fair overall, the performance of the model, specifically in AUC, was very weak.

In [9] With the rise of online loan platforms such as Lending club, the assessment of loan default risks is important to financial security. In this study, a Logit model was applied to the data of Lending club, and it identified the factors that determine the default risks and risks and produced a predictive model. The accuracy metrics and visualizations from the results will provide insights into default causes and potential improvements to the model.

In [10] Logistic regression is used to describe a relationship between a binary response variable and explanatory variables, but multicollinearity among predictors may hamper the results. This study involves an analysis of survey data from 2,331 customers with highly correlated binary variables to predict housing loan approval. A categorical principal component analysis is applied in order to address multicollinearity and improve the reliability of the logistic regression model.

# 3.Methodology

There are two CSV files: the training dataset and the testing dataset. The training file has been used for the training model; predictive models will learn from this file. This file consists of independent variables and target variables. After that, python libraries Pandas, Seaborn, and SKlearn were uploaded. Loan Dataset determine the training and testing data. The architecture of the model can be seen in Fig.1 where data cleaning, preprocessing and applying the model for prediction determining the accuracies.



**Fig.1:** Architecture of the model

Proposed model for prediction. The testing file consists of features variables but does not include targeted variables. Dependent variables of testing data points have been analysed from the model. In this paper, progressive work on an imminent structure has been done to identify whether the customer shall be capable of repaying his debts or loans to the company or not.

The structure of the model is manifested in the flow diagram in Fig.1. This paper aims to assume specimens from data sets that have been used for the credit clearance process and build a structure based on the models considered in the preceding stage. For analytics purposes, important properties like credit score, duration of credit, purpose, etc., are gathered and used to search for the most appropriate attributes.

**3.1 About Data Preprocessing:**

In any machine learning process, preprocessing of data set is a state in which the data gets changed or coded to conduct it to a condition that the model can now parse it quickly. In other words, the independent variables or features of a dataset can now be smoothly elucidated by the algorithm.

Steps to be followed in Data Pre-Processing:

* Importing the Libraries
* Hypothesis Generation
* Comprehending the Data
* Exploratory Analysis of Data
* Univariate Analysis and Bivariate Analysis
* Treatment of Missing Value
* Encoding Categorical Data
* Feature Scaling
* Model Building Hyper-Parameter Tuning

**3.2 About Dataset:**

This paper there is a structure of a predictive model to guess if the customer would be capable of repaying. Design and Development of Loan Predictor Using Machine Learning 119 credit to the loan provider company or not. This model demands a dataset that includes all the factors of an individual regarding all his credit components, ultimately deciding whether he will be sanctioned a loan by any of the financial companies he is requesting. Therefore, we would have a large dataset for this problem that we will split into two parts: Training and testing sets.

* Dataset of training shall be used for training the model as the model shall learn and grasp from this set. The training dataset contains all the feature variables and targeted/test variables.
* Dataset of testing shall contain all the features variables but will be missing the target/dependent variables. This will be applied to the model to guess the dependent variables for testing data

The below Table:1 gives a detailed information about the different variables that are used in this model.

|  |  |
| --- | --- |
| **Variables** | **Their Descriptions** |
| ID of Loan | ID of Loan (Unique) |
| Gender | Male/Female |
| Age | Age of Customer |
| Marital Status | Applicant Marital Status (Y/N) |
| Area-Property | Rural/Urban/Semi-Urban |
| Self-Employed | Self-Employed Status (Y/N) |
| Status-Education | Customer education (Graduate/Non-Graduate) |
| Income-CoApplicant | Co-Applicant Income Status |
| Amount-Loan | Loan Amount in Thousands |
| Customers’ Dependents | Number of Dependents |
| Income-Applicant | Income of Applicant |
| Amount-Term Loan | Period of Loan/Credit in Months |
| History-Credit | Guidelines Meeting Credit History |
| Status-Loan | Approval of Loan Status (Y/N) |

**Table1:** Dataset variables description

From the above components that will make up the whole dataset, we can conclude that there are three formats of data types:

* Object
* Int64
* Float64

**Below steps shows the structure/pseudo steps for the loan prediction method:**

1. **Insert the Data.**
2. **Classify the testing and training data.**
3. **Data Preprocessing and Cleaning.** 
   1. Computation of Missing Values:
      1. For Categorical Variables: - use MODE.
      2. For Numerical Variables: - use MEAN and MEDIAN.
   2. Outlier Treatment.
4. **Begin the Model Building process:** 
   1. Removal of the load identifier
   2. Creation of the targeted/dependent variable: This approach will be focused on the dependent variable that is STATUS-LOAN.
   3. Dummy variable creation for categorical variable and validation will involve splitting of testing and training dataset.
   4. Application of different ML models: -
      1. DT model
      2. RF model
      3. LR model
      4. K-NN model
      5. SVC model

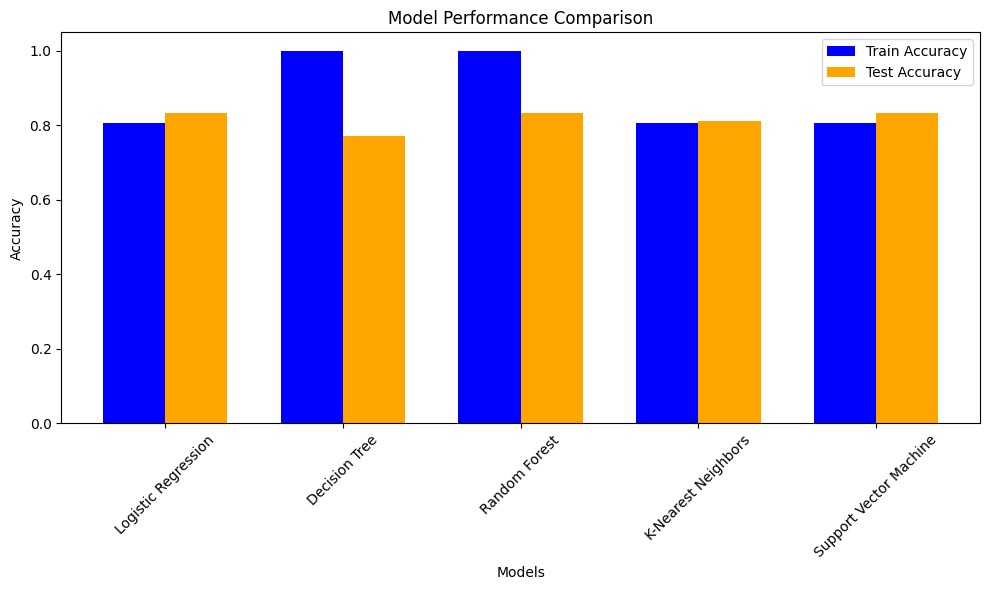
For these models Fig.2 gives the bar chart comparing the training and testing accuracy of different machine learning models.

Fig2: Model Performance Comparison

# Results

The dataset was pre -processed and cleaned. After that, EDA (Exploratory and Data Analysis) and FE (Feature Engineering) were also performed. Then, a model has been trained to identify if the applicant would be capable of repaying the credit or not. Whenever a bank sanction‘s

applicants’ credit, it spontaneously reveals several financial risks. Therefore, it becomes mandatory for banks to be acquainted with the clients registering for the credit. This complication stimulates an Exploratory Analysis of data on the given data set and, therefore, analyses the applicant’s behaviour. The data set that utilizes EDA goes through the filtering of essential columns, identification of dependent variables, derivation of new columns, normalization, and visualizing the data in the graphic and picturized format. Python language has been used for efficient and optimized

Design and Development of Loan Predictor Using Machine Learning and rendering of data. Further, Seaborn, Pandas, and Scikit learn libraries available in Python have been used to extract and process information from the given dataset. Finally, the rendered data has been transformed into suitable plots and graphs to better visualize the conclusions and perception. The matplotlib library of Python has been used to create the plots and graphs. Graphs and Plots show the visualized representation of different datasets. Fig. 3 gives the confusion matrix of loan prediction. The series of charts in the Fig.4 show the distribution of attributes such as gender, marital status, education, and income into categories. They may also detail the terms related to the loan, credit history, property area, and loan status.

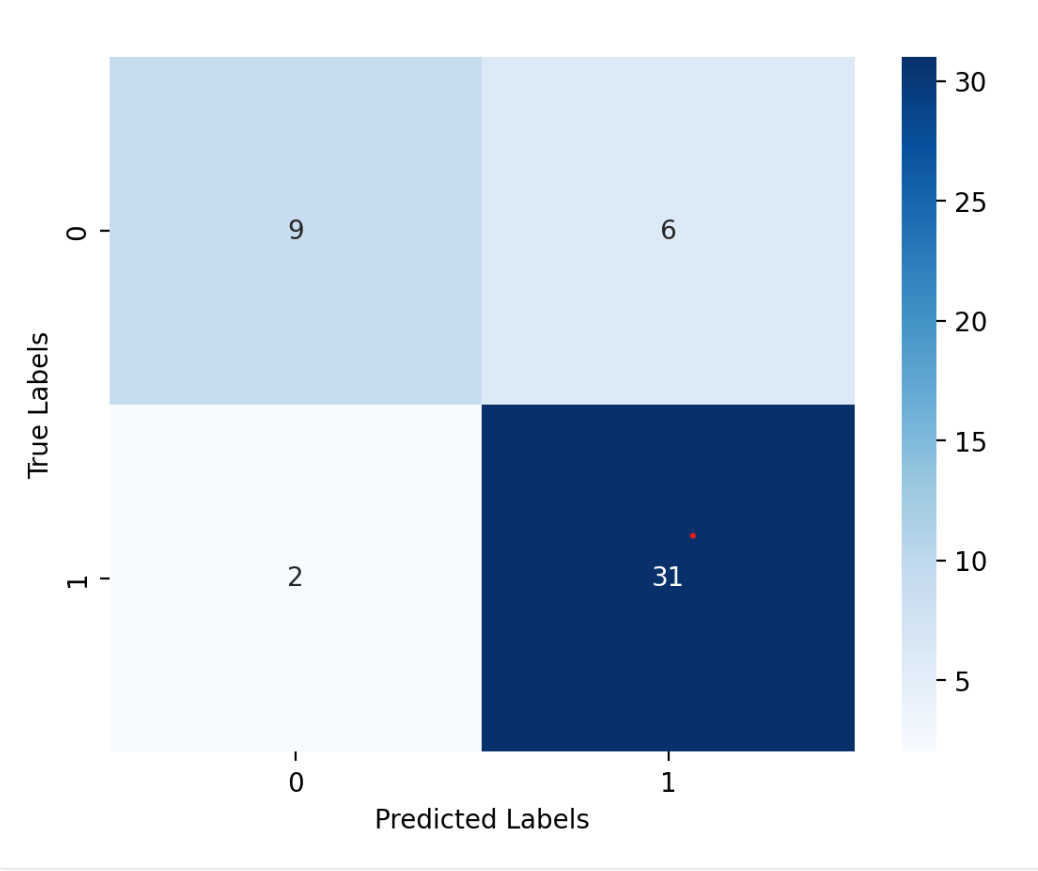
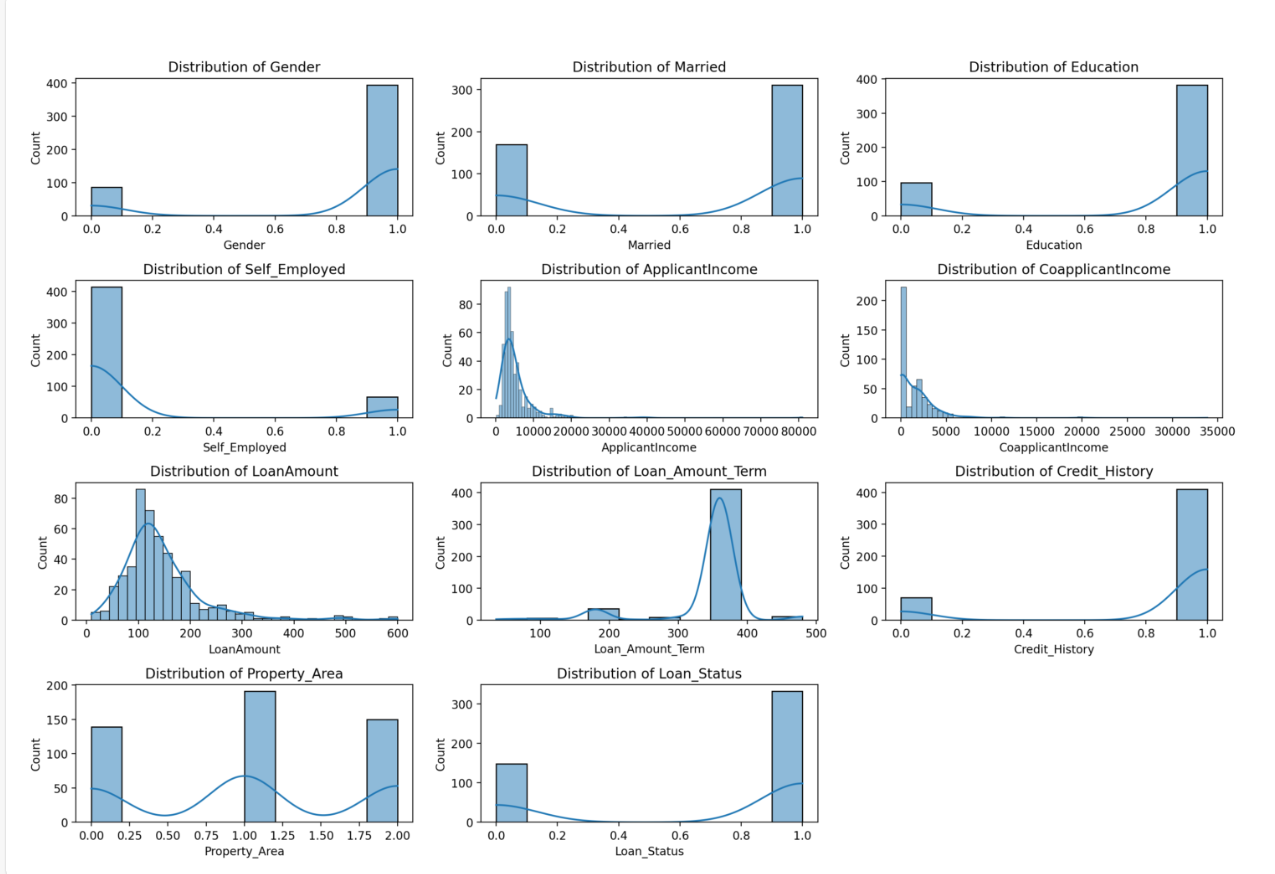


Fig 3: Confusion Matrix of loan prediction

Fig 4: Histogram of the dataset attributes

# CONCLUSION

Through detailed examination of the positive constraints, qualities, and attributes, it would clearly show that the SVC algorithm is very efficient and optimized. In comparison to other machine learning models, it shows a high level of accuracy in results delivery every time. It satisfies all requirements that a banking system requires, and it functions correctly and can easily work with others as well. The following factors are expected to influence predictions: system malfunctions, content errors, and improper weight assignments in computerized systems. Banking software will become more dynamic, reliable, and accurate with automatic processing units. This would be helpful in loan approval and rejection. The inclusion of other techniques, such as discriminative analytics and neural networks, will enhance accuracy.

The model offers a full pipeline to develop a machine learning system for predicting loan approval. It covers all important steps such as encoding data, replacing missing values, scaling features, exploratory data analysis, and model evaluation. Logistic Regression is a fitting choice for binary classification problems where the target variable loan approval is to be encoded as 1 and loan denial is encoded as 0. Some evaluation metrics, including accuracy and classification reports, are necessary for testing how well the model performs on unseen data.

# REFERENCES

1. Manglani, Richa, and Anuja Bokhare ."Logistic regression model for loan prediction: A machine learning approach." *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*. IEEE, 2021.
2. Vaidya, Ashlesha. "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval." *2017 8th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2017.
3. Dm, Obare, and Muraya Mm. "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 (2018): 266-271.
4. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. "An approach for prediction of loan approval using machine learning algorithm." *2020 international conference on electronics and sustainable communication systems (ICESC)*. IEEE, 2020.
5. Sobika, K. Aditya, et al. "An Empirical Study on Loan Prediction Using Logistic Regression and Decision Tree." *2021 International Conference on Advancements in*

*Electrical, Electronics, Communication, Computing and Automation (ICAECA)*. IEEE, 2021.

1. Agbemava, Edinam, et al. "Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana." *European Scientific Journal* 12.1 (2016).
2. Sarungu, Cornelius Mellino. "Loan Eligibility Prediction Using Logistics Regression Algorithm."
3. Adewusi, Amos Olaolu. "AN EVALUATION OF CLASSIFICATION PERFORMANCE OF ARTIFICIAL NEURAL NETWORK AND LOGISTIC REGRESSION AS LOAN SCORING MODELS." *JOURNAL OF CONTEMPORARY RESEARCH IN THE BUILT ENVIRONMENT (JOCREBE)*: 136.
4. Adewusi, Amos Olaolu. "AN EVALUATION OF CLASSIFICATION PERFORMANCE OF ARTIFICIAL NEURAL NETWORK AND LOGISTIC REGRESSION AS LOAN SCORING MODELS." *JOURNAL OF CONTEMPORARY RESEARCH IN THE BUILT ENVIRONMENT (JOCREBE)*: 136.
5. Sreesouthry, S., et al. "Loan prediction using logistic regression in machine learning." *Annals of the Romanian Society for Cell Biology* (2021): 2790-2794.
6. 11.Chan, V.C.T., et al.: Designing a credit approval system using web services, BPEL, and AJAX. In: 2009 IEEE International Conference on eBusiness Engineering, Macau, 2009, pp. 287–294(2009)
7. Hamid, E.N., Ahmad, N.: A new approach for labelling the class of bank credit customers via classification method in data mining. Int. J. Inf. Educ. Technol. 1(2), 150– 155 (2011)
8. Zurada,J.Zurada, M.: How secure are “Good Loans”: validating loan-granting decisions and predicting default rates on consumer loans. Rev. Bus. Inf. Syst. (RBIS) 6(3), 65–84 (2011)
9. Seema, P., Anjali, K.: Credit evaluation model of loan proposals for Indian Banks. In: World Congress on Information and Communication Technologies, pp. 868–873. IEEE (2011)