LOAN ELIGIBILITY PREDCTION ANALYSIS USING VARIOUS MACHINE LEARNING ALGORITHMS

P. Nagraj   
Department of Computer Science and Engineering  
Kalasalingam Academy of Research and Education,Virudhnagar, Tamilnadu, India

K. V. S. Sai Ram Santosh Babu  
Department of Computer Science and Engineering  
Kalasalingam Academy of Research and Education,Virudhnagar, Tamilnadu, India  
9921004355@klu.ac.in

K. Nikhil   
Department of Computer Science and Engineering  
Kalasalingam Academy of Research and Education,Virudhnagar, Tamilnadu, India  
9921004374@klu.ac.in

D. Hari Tejaswar Reddy  
Department of Computer Science and Engineering  
Kalasalingam Academy of Research and Education,Virudhnagar, Tamilnadu, India  
99210041032@klu.ac.in

**Abstract— One of the most fundamental issues that banks and other financial institutions must deal with is loan prediction because it has a big impact on profits. Even though there are several established methods for extracting data from loan applications, most of them seem to be functioning poorly given the reported increase in the number of subprime loans. Gradient boosting, an Extreme Gradient Boosting approach, is used in this work to forecast loan default. Statistics on loans from the internet lending company Super Lender corroborate the prediction. We review demographic data and information from the loan application. The F1-Score, Accuracy, Recall, and Precision area of the investigation are among the critical evaluation metrics we list here. This study provides a strong foundation for loan credit approval by identifying risky clients from a large pool of loan applications using predictive modelling.**

# *keywords—introduction, literature survey, style, background, logistic regression, k nearest neighbor, random forest classifier, , extra tree classifier, dataset, data preprocessing,, data analysis, log transformation, label encoding and correlation matrix, training and testingconclusionreferences (key words)*

# INTORDUCTION

The banking industry is growing quickly as a result of an increase in loan demand for a variety of reasons. Additionally, as more people apply for credit, there are more bad debts accruing, which costs banks, financial institutions, and credit startups money.

The credit loan is money given by a bank or financial institution to a customer or debtor, who agrees to repay it with a set interest rate and by a certain date. These loans are allowed for a variety of purposes, including student loans, mortgages, auto loans, and personal loans.

In order to reduce the number of problematic borrowers, financial institutions and banks should develop robust models that accurately anticipate the loan approval status using the data already available.

A loan is the primary source of income for banks. The money made from loans accounts for the vast majority of bank profits. There is no guarantee that the chosen hopeful is the right hopeful, even if the bank approves the loan after a drawn-out verification and testimonial process. This process takes more time when carried out manually. There are a total of eleven different types of criteria that could affect a debtor loan, including marital status, gender, education level, dependents, loan size and term, self-employment, applicant and co-applicant income, credit history, and geographic location. These are all independent factors that could influence the dependent variable loan status, which the trained model predicts, or they might not. Numerous research has suggested that different elements or circumstances can affect good debt.

Strong models must be developed in order to find hidden patterns in the data that are provided that are vital to accuracy and also boost the model's effectiveness. Although there are numerous machine learning models that can forecast loan default, no method has been optimized to be 100% effective. Using the gradient boosting technique, the loan approval dataset is evaluated in this study, and some important factors that could have an impact on loan debts are selected. Additionally, we assess the results using the precision, accuracy, and confusion matrix.

The main benefit of gradient boosting is that it can be used to predict both continuous and categorical target variables. When used as a continuous target variable, mean square error is the cost function, while log loss is used when it is used as a categorical target variable.

The bias error, a sort of error in machine learning, will be the major focus of this gradient boosting because it can deal with missing data.

By utilizing various models with various algorithms, the model's accuracy has enhanced as compared to the prior loan prediction. Before applying one hot encoding, the model's testing set accuracy was roughly 83.4%. This method allowed it to be raised to 91.3%. K-fold cross validation was used to improve the model by determining whether the approach performed better on the given dataset.

# LITERATURE SURVEY

|  |  |  |  |
| --- | --- | --- | --- |
| Ref no | Title | Algorithms Used | Accuracy |
| 1 | Bank Loan Prediction System Using Machine Learning, Anshika Gupta, 2020 | Logistic Regression and Random Forest | 85 |
| 2 | An Approach for Prediction of Loan Approval using Machine Learning Algorithm, Mohammad Ahmad Sheikh, 2020 | KNN, SVM | 89 |
| 3 | Customer Loan Prediction using Supervised Learning Techniques, L Udaya Bhanu, 2021 | supervised learning techniques like Regression and SVM | 91 |
| 4 | Loan Eligibility Prediction Using Machine Learning Algorithm’s, Sachin Magar, 2022 | Decision Tree | 93 |
| 5 | A Study on Predicting Loan Default Based On the Random Forest Algorithm, Lin Zhu, 2019 | Random forest algorithm | 95 |

# BACKGROUND

We describe the model we used to forecast loan defaults in this section. We employed supervised learning when building a model, where input and output are provided to the model, which is denoted by (x, y), which is function y = f (x).

We utilized one hot encoding to convert all category data into numerical values, and the model then compared the output it predicted to the original output that was provided. There are many algorithms, including neural networks, support vector machines, decision trees, and random forests. The gradient boosting approach was selected for its improved accuracy and precision.

# Logistic Regression

A categorical dependent variable (also known as the outcome variable) and one or more independent variables (also known as predictor or explanatory variables) are analyzed using the statistical technique of logistic regression. Binary outcomes, such as yes/no, true/false, or 1/0, are modelled using this particular sort of regression analysis.

Finding the best-fitting model that connects the likelihood of a specific outcome to the values of the independent variables is the aim of logistic regression. It accomplishes this by converting a linear combination of the predictor variables into a probability value between 0 and 1 using the logistic function. The logistic function, commonly known as the sigmoid function, features an S-shaped curve that asymptotically moves closer to 0 or 1 depending on the input value.

The probabilities of the dependent variable being 1 or 0 can be predicted using the logistic regression model's estimated coefficients for each independent variable. The model can also be used to evaluate theories regarding how the dependent variable and predictor variables are related.

In order to forecast results and comprehend the variables that affect them, several disciplines, including medicine, biology, the social sciences, and marketing, frequently use logistic regression.

# K NEAREST NEIGHBOR

A supervised learning technique used for classification and regression analysis is the K-Nearest Neighbor (KNN) classifier. KNN is a non-parametric method that makes no assumptions about how the data are distributed in their underlying form.

The input-output pairs in the KNN algorithm's training data are examples of labelled input-output pairs with a vector of features as the input and a class label as the output, respectively. Based on their Euclidean distance (or other distance metrics) to the input vector, the algorithm looks for the K data points in the training set that are closest to the input vector when it is given a new input vector. Then, using the majority decision of the K-nearest neighbors, the class label of the input is predicted.

Since KNN is a lazy learning algorithm, it doesn't create a model during training. To make predictions, it only keeps the training data in memory. For small datasets, this makes the technique quite effective; however, for bigger datasets, it may be computationally expensive.

The ease with which KNN can handle multi-class classification issues is one of its benefits. Its shortcomings include its sensitivity to noise in the data and irrelevant features. The training data must be stored in a lot of memory, and large datasets might result in lengthy processing times.

# RANDOM FOREST CLASSIFIER

A well-liked supervised learning approach for classification tasks is Random Forest Classifier. It is a member of the family of ensemble learning algorithms, which mix several models to enhance performance and lessen overfitting.

In Random Forest, a number of decision trees are trained using a random selection of features at each split on various subsets of the training data. This produces a varied collection of decision trees that can effectively capture various features of the data. Each decision tree in the forest separately forecasts the input's class label during the prediction phase, and the final prediction is based on the trees' majority vote.

The key benefit of Random Forest is that it can handle high-dimensional data with intricate feature interactions. Estimates of feature relevance are also provided, which might be helpful for feature selection, and the risk of overfitting is decreased.

Random Forest is frequently used for classification tasks like picture identification, customer segmentation, and disease detection in a variety of industries, including finance, healthcare, and bioinformatics. However, it might not be appropriate for real-time applications with big datasets and can be computationally expensive.

# EXTRA TREES CLASSIFIER

A machine learning algorithm from the family of ensemble methods is called Extra Trees Classifier. For classification tasks, it is a particular kind of decision tree method.

Extra Trees Classifier constructs a number of decision trees and then combines their outputs to make predictions, similar to other ensemble approaches. In contrast to other decision tree algorithms, it randomly chooses feature subsets for each decision tree and divides nodes. This randomization aids in lowering the model's variance and guards against overfitting.

Another bagging strategy used by Extra Trees Classifier is to train each decision tree using a separate random subset of the training data. This improves the overall performance of the model by producing a broad set of decision trees that may capture various features of the data.

The fact that Extra Trees Classifier is computationally effective and capable of handling huge datasets with high-dimensional features is one of its benefits. Additionally, it has a little bias and performs well with noisy data. On some datasets, though, it might not perform as well as other algorithms, and it might need more trees to perform at its best.

In conclusion, Extra Trees Classifier is an effective machine learning technique for classification problems, especially when dealing with huge datasets and high-dimensional features.

# DATA SET

The data set gathered is separated into a training set and a testing set for the purpose of predicting loan failure clients. The 80/20 rule is typically used to separate the training set from the testing set. The decision tree-created data model is applied to the training set and hooked on the test take precision.

It is done to forecast test set. the following characteristics:

|  |  |
| --- | --- |
| **Variable** | **description** |
| Loan ID | Unique id |
| Gender | Male / Female |
| Married | Applicant married or not |
| Dependents | Number of dependents |
| Education | Applicant education (graduate or not) |
| Self Employed | Self-employed or not |
| Applicant Income | Applicant income |
| Co applicant Income | Co applicant income |
| Loan Amount | Loan amount in thousands |
| Loan Amount Term | Term of loan in months |
| Credit History | Credit history meets guidelines |
| Property Area | Urban or rural or semi urban |
| Loan Status | Loan approved or not |

Table 1. sample of dataset

Dataset Source: - loan prediction dataset

# DATA PREPROCESSING

Data processing is the process of turning unstructured data into usable information using a variety of tools like sorting, organizing, cleaning, and analysis. The first step in the process is gathering data from multiple sources, which might be in text, image, or video format. After that, the data is changed into a format that can be quickly processed and examined.

Data cleaning to eliminate any errors, duplication, or unnecessary information, data transformation to convert data into a standardized format, and data integration to combine data from diverse sources are some examples of the data processing procedures. Statistical techniques and data visualization tools can be used to Analyse the data after it has been processed in order to acquire insights and make data-driven decisions.

Spreadsheets, databases, programming languages like Python and R, data visualization programmers like Tableau and Power BI, and machine learning algorithms are some of the regularly used tools and methods for processing data. Organizations that effectively handle their data can find patterns, trends, and anomalies in their data that can help them gain a competitive edge and make wise decisions.



Removing null values from data

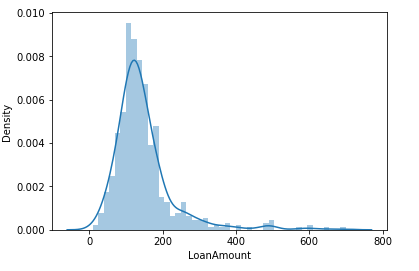
# DATA ANALYSIS

The practice of methodically reviewing and interpreting data in order to derive important insights and reach judgements is known as data analysis. To find patterns, trends, and relationships in the data, a variety of methods and instruments are used.

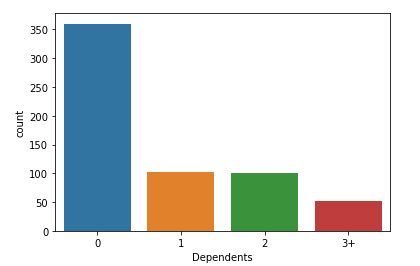
The purpose of data analysis is to turn unstructured data into knowledge that can be applied to decision-making or problem-solving. This can be done using a variety of methods, including machine learning, data visualization, inferential statistics, and descriptive statistics.

In descriptive statistics, the features of the data, such as its central tendency, variability, and distribution, are summarized and described. Based on a sample of data, inferential statistics involves drawing conclusions about a wider population. To help detect patterns and trends in the data, data visualization entails developing visual representations of the data. Making predictions and automatically identifying patterns in data are both aspects of machine learning.

Many different industries, including business, healthcare, education, and government, use data analysis. It is an essential part of data-driven decision-making and can give businesses a competitive edge by spotting areas for development and improvement.



Dist plot on loan amount



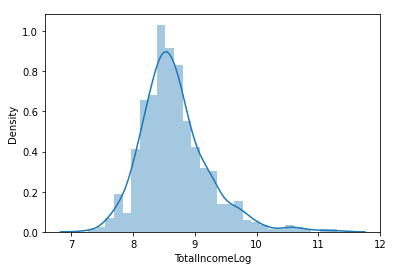
Count plot on Dependents

# LOG TRANSFORMATION

Taking the logarithm of a set of data values is a mathematical process known as log transformation. The opposite of the exponential function, the logarithm function is frequently used to modify data that is extremely skewed or changes widely.

The log transformation can be used for several purposes, such as:

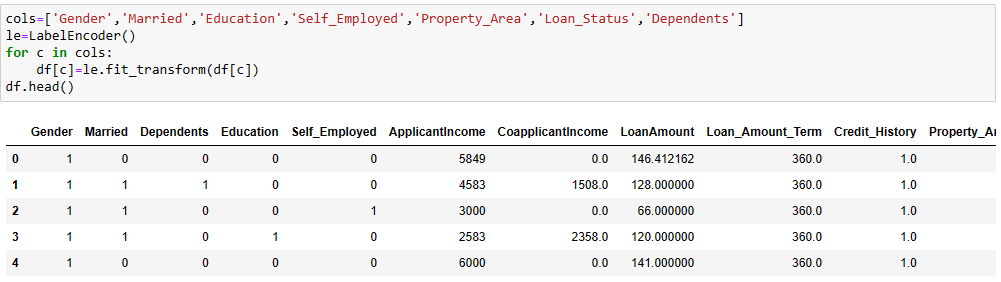
* Reducing skewness: The log transformation can be used to make a dataset's distribution more symmetrical if it is extremely skewed.
* Rescaling data: The log transformation can be used to rescale data to a smaller range when a dataset's value range is very wide.
* Linearizing relationships: Nonlinear relationships between variables can be linearized using the log transformation. For instance, if there is an exponential relationship between two variables, taking the log of both variables can make the relationship linear.
* Data normalization: By using the log transformation, data can be made more comparable between values on various scales.



Log transformation on total income

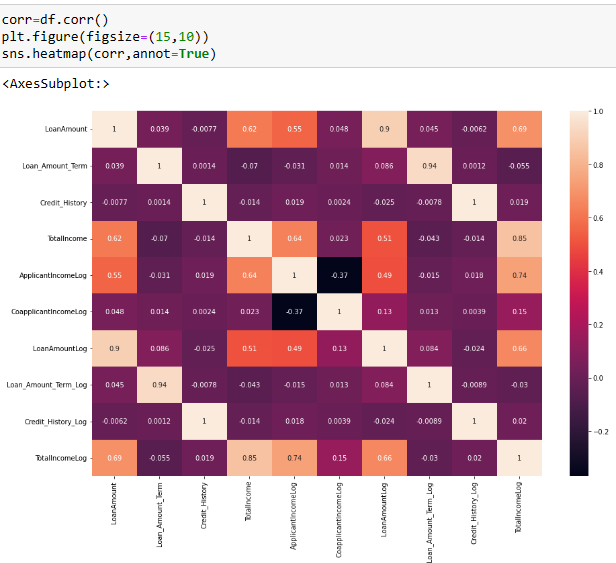
# LABEL ENCODING AND CORRELATION MATRIX

Label encoding is process of converting categorical data into numerical values, where unique category is assigned to a unique integer. This is often necessary in all machine learning models and data analysis



Label Encoding

A unique category is assigned to a unique number during the label encoding process, which transforms categorical data into numerical values. In any machine learning models and data analysis, this is frequently required.



Correlation Matrix

# Training and Testing

Prior to training and testing, the data is split 80/20 between train data and test data.

The process of teaching a machine learning model how to generate predictions in the data and how to classify the data using input data is referred to as training a dataset.

The model learns about the input labels that correspond to the output data with the aid of the training data. The model develops its grasp of how to make predictions on fresh data while being trained to identify patterns that are present in the input data. The aim of training is to build a model that can predict outputs with high accuracy for novel, unforeseen inputs.

Testing in the sense finding the best model that suits our data set by finding accuracy, recall, precision and f1 score.

|  |  |  |
| --- | --- | --- |
| Model | Train set | Test set |
| Logistic regression | Accuracy:80.19 %.  Recall: 0.983  Precision: 0.780  F1-Measure: 0.869 | Accuracy:82.16 %.  Recall: 0.978  Precision: 0.814  F1-Measure: 0.888 |
| KNN | Accuracy:73.19 %.  Recall: 0.920  Precision: 0.742  F1-Measure: 0.822 | Accuracy:77.3 %.  Recall: 0.978  Precision: 0.771  F1-Measure: 0.862 |
| Random Forest | Accuracy:82.52 %.  Recall: 0.983  Precision: 0.802  F1-Measure: 0.883 | Accuracy:87.03 %.  Recall: 0.993  Precision: 0.853  F1-Measure: 0.917 |
| Extra Trees | Accuracy:80.65 %.  Recall: 0.986  Precision: 0.782  F1-Measure: 0.873 | Accuracy:82.7 %.  Recall: 0.985  Precision: 0.815  F1-Measure: 0.892 |

Testing of models on train set and test set

# CONCLUSION

In this study, gradient boosting was used to successfully forecast bank loan failure. The aim was to forecast whether a loan application will forget to make a payment. Recall, accuracy, precision, and f1-score were among the performance metrics that were calculated as part of the evaluation, which was done using the Python programming language.

According to the investigation, two important variables that our system considers when determining whether or not a customer will default on a payment are their location and age. This study provides a strong foundation for loan credit approval by identifying risky clients from a large pool of loan applications using predictive modelling. Comparing the performance of the gradient boosting algorithm to that of SVM, naive bayes, and decision trees. The decision tree has higher training accuracy than other models, but testing accuracy is lower because, as shown by the results, the decision tree will produce greater variations in results when there is a small change in the data, which could cause instability in the tree.

This model receives the greatest accuracy values following the application of numerous pre-processing procedures to various models. The primary basis for the model prediction is the dataset preprocessing methods. The gradient boosting will iteratively improve based on each and every row of the data set's previous error. Up until the dataset's conclusion, this iteration is ongoing. A more accurate and time-effective real-time loan prediction model may be developed in the future. In the upcoming model, the complexity may be relatively reduced.

# REFERENCES

1. [1] Author Anshika Gupta “Bank Loan Prediction System Using Machine Learning” published year 2020, method Machine Learning (Logistic Regression and Random Forest).
2. [2] Author Mohammad Ahmad Sheikh “An Approach for Prediction of Loan Approval using Machine Learning Algorithm” published year 2020, method Machine Learning Algorithm (logistic regression, SVM, KNN).
3. [3] Author L Udaya Bhanu “Customer Loan Prediction using Supervised Learning Techniques.” Published year 2021, method Machine Learning Algorithm (logistic regression, SVM, KNN, Random Forest, Decision Tree).
4. [4] Author Sachin Magar “Loan Eligibility Prediction Using Machine Learning Algorithm’s” published year 2022, method Machine Learning Algorithm (SVM, Random Forest).
5. [5] Author Lin Zhu “A Study on Predicting Loan Default Based On the Random Forest Algorithm” published year 2019, method Machine Learning Algorithm (SVM,Logistic Regression and Decision Tree).
6. Kumar Arun, Garg Ishan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)
7. Haomin Wanga, Wei Chena, Fang Daa, Zhima Credit Score in Default Prediction for Personal Loans, The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021).
8. Trevor Fitzpatricka, Christophe Mues16 October 2021, how can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments, European Journal of Operational Research (volume 294)
9. Nithi Sopitpongstorna, Param Silvapulle, Jiti Gao, Jean- Pierre Fenech, May 2021, Local logit regression for loan recovery rate, Journal of Banking & Finance (volume 126)
10. Kathe Rutik Pramod, Nashik, Dapse Punam Laxman, Ghorpade Dinesh B, anhale Sakshi Dattatray, Avhad Pooja Prakash, 6 June 2021, An Approach for Prediction of Loan Approval Using Machine Learning Algorithm, IJCRT (volume 9).
11. Lakshmana Prabu, S.K. & Shankar, K. & .M, Ilayaraja & Nasir, Abdul Wahid & Varadarajan, Vijayakumar & Chilamkurti, Naveen. (2019). Random forest for big data classification in the internet of things using optimal features. International Journal of Machine Learning and Cybernetics. 10. 10.1007/s13042-018-00916-z.
12. Shankar, K. & Elhoseny, Mohamed & chelvi, E. & Lakshmanaprabu, S.K. & Wu, Wanqing. (2018). An Efficient Optimal Key Based Chaos Function for Medical Image Security. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2874026.
13. Bagherpour, A. Predicting mortgage loan default with machine learning methods. University of California/Riverside, 2017.
14. Buhlmann, P., Hothorn, T., et al. Boosting algorithms: Regularization, prediction and model fitting. Statistical Science, 22(4):477–505, 2007.
15. Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785– 794, 2016.
16. Chiang, R. C., Chow, Y.-F., and Liu, M. Residential mortgage lending and borrower risk: the relationship between mortgage spreads and individual characteristics. The Journal of Real Estate Finance and Economics, 25(1):5–32, 2002.
17. Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In icml, volume 96, pp. 148–156. Citeseer, 1996.