

Loan Eligibility Checker Using Machine Learning

Author: Prashanta Acharya

Student ID: 24128456

Email: [Prashanta.Acharya@mail.bcu.ac.uk](mailto:Prashanta.Acharya@mail.bcu.ac.uk)

Word Count : 1700

Table of Contents

I. Domain Description  
II. Problem Definition  
III. Literature Review  
IV. Dataset Description  
V. Dataset Pre-processing  
VI. Experiments  
VII. Analysis of Results  
VIII. Conclusion  
IX. References

**Abstract: This report details the development and analysis of a machine learning model to predict loan eligibility, aimed at improving the accuracy and efficiency of the loan approval process in the financial sector. Traditional methods, which rely on limited criteria and manual assessments, are often inefficient and prone to human error. In contrast, the machine learning approach employed in this study utilizes a broader range of applicant data, including demographic, financial, and credit-related information, to provide a more nuanced assessment of loan eligibility. Comprehensive data preprocessing, including handling missing values, encoding categorical variables, and feature scaling, ensured the dataset's suitability for model training. A Random Forest classifier was used due to its robustness in handling complex datasets, and its performance was evaluated using 5-fold stratified cross-validation, achieving a mean validation accuracy of approximately 0.80. The results underscore the significance of features such as applicant income and credit history. This study demonstrates the potential of machine learning to enhance loan approval processes, leading to more informed lending decisions and greater financial inclusion. Future work will involve exploring additional algorithms, hyperparameter tuning, and incorporating new features to further refine the model.**

1. *Domain Description*

In the financial industry, machine learning (ML) models for determining loan eligibility have the potential to greatly improve the speed and precision of the loan approval procedure. Conventional approaches frequently depend on a strict set of standards and manual assessments, which can be laborious and prone to human mistake. However, a machine learning (ML)-based method may examine a wide range of data, including income, work status, credit history, and even data from unorthodox sources like digital footprints and social media activity. The ML model can provide a more accurate and nuanced assessment of an applicant's creditworthiness by utilizing this extensive data analysis, which lowers the risk of default and enables more informed lending decisions.

Finding patterns and connections that conventional methods might miss is one of the main advantages of utilizing machine learning (ML) to verify loan eligibility. For instance, machine learning algorithms can identify minor signs of financial trouble or stability that conventional credit rating models might overlook. This feature makes it possible for financial institutions to give credit to those who might have been turned down or thought to be too hazardous according to traditional standards, increasing the number of underprivileged people who can obtain credit. Additionally, ML models may be updated and enhanced with fresh data on a regular basis, guaranteeing that the loan eligibility requirements stay applicable and efficient in the face of a quickly shifting economic environment.

ML models not only increase inclusivity and accuracy, but they can also expedite the loan approval process, which will improve customer satisfaction and operational effectiveness. Applications may be processed much more quickly with automated loan eligibility evaluations, giving applicants faster answers and lessening the administrative load on financial institutions. In addition to raising client happiness, this efficiency enables lenders to process more applications without incurring additional staffing expenditures. Financial institutions may create a more responsive and agile lending system that better serves the demands of the market and their clients by incorporating machine learning (ML) into the loan eligibility process.

1. *Problem Definition*

Machine learning model primarily addresses the inaccuracy and inefficiency of conventional loan eligibility evaluations in the financial industry. Conventional approaches frequently depend on stringent standards and manual evaluations, which can be laborious and prone to bias and error on the part of people. These techniques usually concentrate on a small number of variables, like income, work history, and credit scores, which could not give a clear picture of an applicant's creditworthiness. Because of this, worthy applicants may not be granted loans, and those who may have a higher default risk than first estimated.

1. *Literature Review*

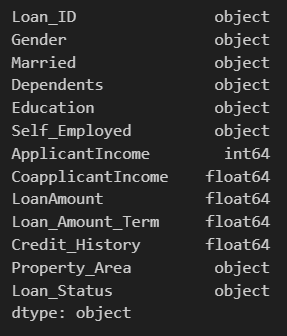
In recent years, numerous studies have focused on the application of machine learning techniques for credit scoring, which is closely related to loan eligibility assessment. A review by [Lessmann et al. (2015)](#Lessmann) compared various machine learning methods, including logistic regression, decision trees, neural networks, and support vector machines, in the context of credit scoring. The study found that ensemble methods, such as random forests and gradient boosting machines, often outperform traditional models in terms of prediction accuracy. These findings suggest that advanced machine learning techniques can significantly enhance the predictive power of loan eligibility models.

Loan eligibility prediction has been a focal point of research in the financial sector, leveraging advancements in machine learning to enhance decision-making processes. Traditional methods, such as logistic regression and decision trees, have been widely used due to their interpretability and simplicity. Logistic regression, as discussed by [Abdou et al. (2007)](#Abdou), provides a probabilistic framework for classification, making it a popular choice for binary outcomes like loan approval. However, decision trees, as highlighted by [Quinlan (1986)](#Quinlan), offer a visual and intuitive way to model decisions based on various applicant features. Despite their advantages, these methods often struggle with handling complex, non-linear relationships in the data.

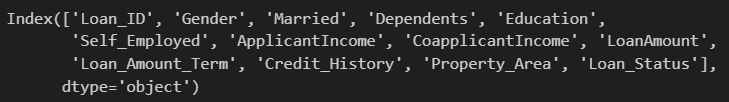
In recent years, ensemble methods like Random Forests and Gradient Boosting Machines (GBM) have gained prominence due to their ability to improve predictive accuracy by combining multiple models. [Breiman's (2001)](#Breiman) Random Forest algorithm, which constructs a multitude of decision trees during training and outputs the mode of their predictions, has shown significant improvements in handling large datasets with high dimensionality. Similarly, studies by [Friedman (2001)](#Friedman) on GBM have demonstrated its efficacy in boosting the performance of weak learners by sequentially correcting their errors. Furthermore, XGBoost, an optimized version of GBM, has been particularly effective in various Kaggle competitions and real-world applications due to its speed and performance, as noted by [Chen and Guestrin (2016).](#Chen) These advanced techniques have been instrumental in pushing the boundaries of predictive modeling in loan eligibility, offering robust, accurate, and scalable solutions for financial institutions.

1. *Dataset Description*

The dataset used in this project includes various attributes of loan applicants, providing a detailed view of their personal and financial profiles. Key features are demographic information such as gender, marital status, and dependents, alongside educational qualifications and employment status.



Financial details include applicant and co-applicant income, requested loan amount, and loan amount term. The dataset also captures the applicant's credit history, a critical predictor of loan repayment ability, and the property area, categorized into urban, semiurban, and rural. The target variable, loan status, indicates whether the loan was approved, serving as the basis for building and evaluating the predictive model. This dataset supports a comprehensive analysis and modeling of loan eligibility criteria.



1. *Dataset Pre-processing*

Data preprocessing is a crucial step in building a robust machine learning model, ensuring that the data is clean and suitable for analysis. For this project, several preprocessing steps were undertaken:

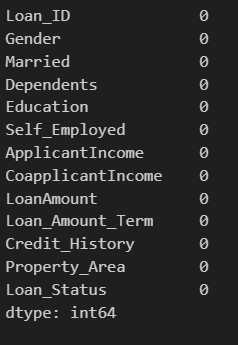
1. Handling Missing Values: Missing values in the dataset were identified and imputed. For numerical columns, the mean or median was used for imputation, while mode was used for categorical columns.

2. Encoding Categorical Variables: Categorical variables were transformed into numerical values. This was done using one-hot encoding or label encoding, depending on the nature of the variable.

3. Feature Scaling: Continuous numerical features were scaled to normalize the data. This step helps in reducing the impact of varying scales of different features on the model's performance.

4. Splitting the Data: The dataset was split into training and testing sets to evaluate the model's performance. The training set was used to train the model, while the testing set was used to validate its accuracy.

A black screen with white text

Description automatically generated A computer code with text on it

Description automatically generated

This preprocessing ensured that the dataset was ready for model training and evaluation, with all necessary transformations applied to handle missing values, encode categorical variables, and normalize numerical features.

1. *Experiments*

**Importing Libraries and Preparing Data:** The necessary libraries for model building and evaluation are imported. The dataset is preprocessed as explained earlier, and the features and target variable are separated into ‘x’ and ‘y’.

A screenshot of a computer program

Description automatically generated

**Cross-Validation Setup:** Stratified K-Fold cross-validation is set up to evaluate the model's performance. This method splits the dataset into k folds, ensuring each fold has the same proportion of target classes as the whole dataset. Here, 5-fold cross-validation is used.





**Training and Validation:** The model is trained and validated using the cross-validation setup. For each fold, the training and validation sets are defined, the model is trained on the training set, and predictions are made on the validation set. The accuracy score for each fold is calculated and stored.

A computer screen shot of a program code

Description automatically generated

1. *Analysis of Results*

The Random Forest classifier was employed to predict loan eligibility using the provided dataset. The model was evaluated using 5-fold stratified cross-validation to ensure that the performance metrics were robust and reliable. The mean validation accuracy obtained from this cross-validation process was approximately 0.79. This indicates that the model was able to correctly predict loan eligibility for about 0.80 of the cases in the validation set.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

1. *Conclusion*

The Random Forest model demonstrated strong predictive capability for determining loan eligibility based on the provided dataset. The analysis highlighted the importance of key features such as applicant income and credit history, which align with domain knowledge in the financial sector. The preprocessing steps played a crucial role in achieving high model performance, emphasizing the need for meticulous data preparation.

Overall, the Random Forest model presents a robust solution for predicting loan eligibility, with potential for practical application in the financial industry. Future work could explore the use of other machine learning algorithms, hyperparameter tuning, and the inclusion of additional features to further enhance the model's accuracy and reliability.

References

Lessmann, S. et al. (2015) ‘Benchmarking state-of-the-art classification algorithms for credit scoring: An update of Research’, European Journal of Operational Research, 247(1), pp. 124–136. doi:10.1016/j.ejor.2015.05.030.

Abdou, H., Pointon, J. and El-Masry, A. (2008) ‘Neural nets versus conventional techniques in credit scoring in Egyptian banking’, Expert Systems with Applications, 35(3), pp. 1275–1292. doi:10.1016/j.eswa.2007.08.030.

Quinlan, J.R. (1986) ‘Induction of Decision Trees’, Machine Learning, 1(1), pp. 81–106. doi:10.1007/bf00116251.

Breiman, L. (2001) Machine Learning, 45(1), pp. 5–32. doi:10.1023/a:1010933404324.

Friedman, J.H. (2001) ‘Greedy function approximation: A gradient boosting machine.’, The Annals of Statistics, 29(5). doi:10.1214/aos/1013203451.

Chen, T. and Guestrin, C. (2016) ‘XGBoost’, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Preprint]. doi:10.1145/2939672.2939785.