https://doi.org/10.1051/itmconf/20224403019

Loan Prediction System Using Machine Learning

Anant Shinde[1], *Yash* Patil[2], *Ishan* Kotian[3], *Abhinav* Shinde[4] *and Reshma* Gulwani[5]

[1,2,3,4]Department of Information Technology, RAIT, Nerul, India
[5]D.Y. Patil Deemed to be University, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India

Abstract. As the needs of people are increasing, the demand for loans in banks is also frequently getting higher every day. Banks typically process an applicant's loan after screening and verifying the applicant's eligibility, which is a difficult and time-consuming process. In some cases, some applicants default and banks lose capital. The machine learning approach is ideal for reducing human effort and effective decision making in the loan approval process by implementing machine learning tools that use classification algorithms to predict eligible loan applicants.

1 Introduction

Machine Literacy is a subset of artificial intelligence that allows computer programs to automatically learn from former tasks. It works by analysing the data, relating patterns, and incorporating minimum mortal intervention. Nearly any work that can be done using a data description pattern or set of rules can be done using a machine learning machine. This allows companies to modify processes that preliminarily only humans could make hypotheticals for client service calls, accounts, and reviews.

Loan distribution is the middle enterprise of virtually every bank. Loan distribution is the middle enterprise of nearly every bank. Utmost of a bank's means come directly from the gains it makes from the loans it makes. The main thing of the banking terrain is to put wealth in a safe place. Currently, many banks / financial companies approve loans after a verification and validation redemption process, but it is not yet clear if the selected applicant is correct among all applicants. Through this system, it is possible to predict whether a particular applicant is safe and the entire process of verifying the characteristics will be automated by machine learning technology. Credit forecasting is very useful for both bank employees and applicants.

The goal of this system is to provide a quick, immediate and easy way to select good applicants. It can offer banks special benefits. The credit forecasting system can automatically calculate the weights for each feature that participates in credit processing, and the new test data will process the same features for the assigned weights. The model can set a deadline to see if the applicant can approve the loan. Credit analysis allows to jump to specific applications and check according to priority. This system is exclusively for bank / financial company management authorities, the entire forecasting process is carried out privately and no stakeholders can change the process. The results of a particular credit ID can be sent to various departments of the bank so that they can take appropriate action on demand. This helps all other departments handle other paperwork.

2 Literature Survey

According to the authors, the forecasting process begins with data clean-up and processing, missing value substitution, data set experimental analysis, and modelling, and continues to model evaluation and test data testing. A logistic regression model has been executed. The highest accuracy obtained with the original dataset is 0.811. Models are compared based on performance measurements such as sensitivity and specificity. As a result of analysing, the following conclusions were drawn. However, other characteristics of customers that play a very important role in lending decisions and forecasting defaulters should also be evaluated. Some other traits, such as gender and marriage history, do not seem to be considered by the company [1]. A credit credibility soothsaying system that helps companies make the right opinions to authorize or reject the credit claims of guests. This helps the banking assiduity to open effective distribution channels. This means that if the customer has a minimum repayment capacity, their system can avoid future risks. Including other techniques (using the Weka tool) that are better than the general data mining model has been implemented and tested for domains [2]. The author suggests that, a credit status model for predicting loan applicants as valid or standard customers. The proposed model shows a score of 75.08 when classifying loan aspirants using R-Package. Lenders can use this interpretation to make mortgage choices for mortgage operations. In addition, comparative studies were conducted at different iterative levels. The replication position is a 30- grounded ANN model that offers a more advanced delicacy than other situations. This model can be used to avoid large losses in marketable banks [3]. Six machine learning classification models were used to predict Android applications. The model is available in open-source software R. This application works well and meets the requirements of all banks. The downside of this model is that it gives each element a different weight, but in reality, it may be possible to approve a loan only based on a single powerful element, which is not possible with this system. This component can be easily connected to many other systems. There are cases of computer failure, and the most important weights of content errors and features are fixed by the automatic prediction system, and soon, so-called software may be safer, more reliable and more [4]. Risk assessment and forecasting is an important task in the banking industry in determining whether a good and lazy loan applicant is applicable. To improve the accuracy of risk, risk assessments are conducted in primary and secondary education. Customer data is extracted and related attributes are selected using information gain theory. Rule forecasting is performed for each credit type based on predefined criteria. Approved and rejected applicants are considered "Applicable" and evaluated as "Not Applicable". Corresponding experimental results have shown that the method proposed predicts better accuracy and takes less time than existing methods [5]. The main purpose of this design is to prognosticate which customers will be repaid with a loan because the lender needs to anticipate the problem that the borrower won't be suitable to repay the threat. Studies of three models show that logistic regression with a rating is superior to other models, random forests, and decision trees. Poor credit seekers aren't accepted, presumably because they have the option of not paying. In utmost cases, high-value appliers may be eligible for a reduction that may repay the loan. Certain sexual orientations and marriage status appear to be out of the reach of the company [6].

3 Feature Engineering

Predicated on the field knowledge, this system can develop new features that can affect the target variables. Created three new functions:

3.1 Total Income – In Fig 1, as explained in the bivariate analysis, combined the income of the applicant with the income of the co-applicant. The higher the total income, the more likely there are to get loan approval.

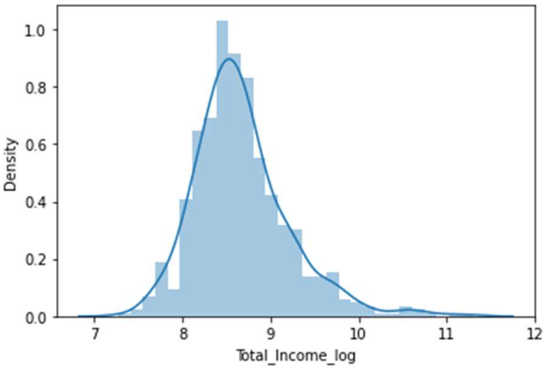

Fig. 1. Density vs Total Income Log

3.2 EMI – In Fig 2, EMI is the yearly volume that the seeker must pay to reimburse the loan. The model behind this variable is that people with lofty. EMI may possess challenges with the prepayment of their loans. EMI can exist figured by taking the rate of the loan volume to the majority the loan volume rate of the loan volume to the majority of the loan volume.

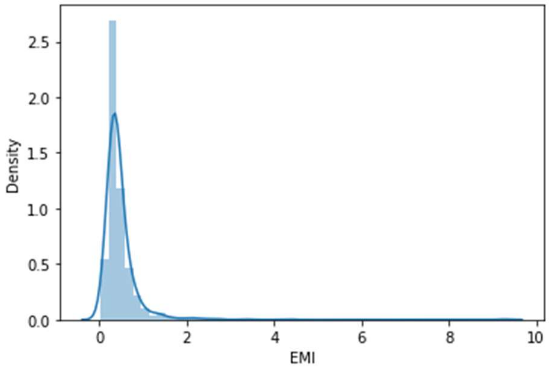

Fig. 2. Density vs EMI

3.3 Balance Income – In Fig 3, this is the return deserted over after compensating the EMI. The model behind creating this variable is that the advanced the valuation, additionally probable a person is to reimburse the loan and thus additionally probable it's to authorize the loan.

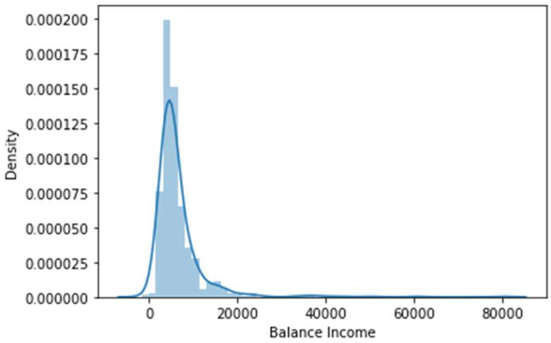

Fig. 3. Density vs Balance Income

4 Proposed Framework

4.1 Business understanding:

In the early stages, the base is on deriving the design from a custom outlook and rephrasing that lore into data mining challenge delineations and primary designs.

4.2 Data convention:

The data convention aspect focuses on the original data library, data command, relating data rate outcomes, and a subset of stake for undertaking retired data.

4.3 Data processing:

The data processing aspect includes all the conditioning to produce the concluding dataset.

4.4 Modelling:

The algorithm which will be used for data modelling is Logistic Regression using stratified k-folds cross-validation and Random Forest.

Logistic Regression using stratified k-folds cross-validation: This system uses validation to see how robust the model is against hidden data. This is an approach for booking distinct exemplifications of reports that don't train the model. Latterly, the system tests the model in this illustration and also finalize it. Some of the generally applied confirmation styles are the confirmation incubate path, k-fold cross-validation, Leave one outcross-validation (LOOCV), and stratified k-fold cross-validation. In Table 1, analyzed the mean validation and f1-score of Logistic Regression with the k-folds model.

Table 1. Accuracy for Logistic Regression model

The mean validation accuracy for this model turns out to be	0.7214
The mean validation f1 score for this model turns out to be	0.8279

In Fig 4, Visualized the roc curve:

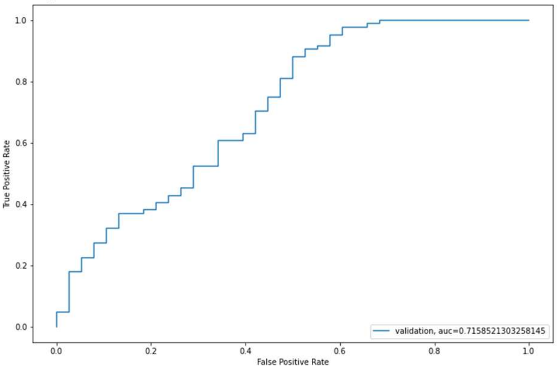

Fig. 4. AUC value of 0.5626

Random Forest: This system was tested to reduce the exactness by conforming to the hyperparameters of this model. The model used grid search to master optimized valuations for hyperparameters. Grid - search is a way to elect the stylish one from the family of hyperparameters parameterized by the parameter grid. adapted the max_depth and n_estimators' parameters. max_depth determines the maximum depth of the tree and n_estimators determine the number of trees used in the random forest model. In Tabel 2, generated the mean validation accuracy for the hyperparameters.

Table 2. Mean Validation Accuracy of Hyperparameters

Mean Validation Accuracy	0.7947

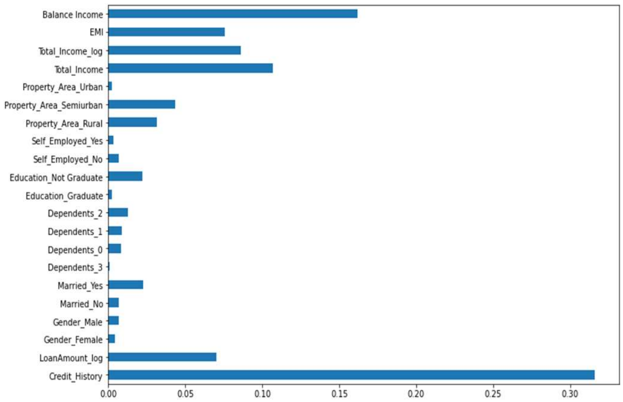

Fig 5. Feature importance predicting the target variable

In Fig 5, Credit_History is the most major point succeeded by Balance Income, Total Income, EMI. Accordingly, feature engineering assisted the model in forecasting the target variable.

5 Conclusion

Borrowers use a loan application to qualify for a mortgage. The above research employs a logistic regression algorithm-based prediction model. To create a logistic classification model that predicts loan status, over 600 sample data were collected and evaluated. The algorithm can obtain a maximum accuracy of about 82 percent and regression models are used to obtain such precision. The model can anticipate outcomes and is quickly adaptable to a wide range of inputs. Also, this strategy saves the banking industry and its staff a significant amount of time.

References

[1] M. Sheikh, A. Goel, T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," International Conference on Electronics and Sustainable Communication Systems (ICESC), (2020).

[2] S. M S, R. Sunny T, "Loan Credibility Prediction System Based on Decision Tree Algorithm," International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 09, (2015).

[3] A. Kumar, I. Garg and S. Kaur, "Loan Approval Prediction based on Machine Learning Approach," IOSR Journal of Computer Engineering, (2016).

[4] Dr K. Kavitha, "Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques," IJARCSSE - Volume 6, Issue 2, (2016).

[5] P. Dutta, "A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION", International Research

Journal of Modernization in Engineering Technology and Science, (2021).

[6] G. Arutjothi, Dr C. Senthamarai, “Prediction of Loan Status in Commercial Bank using Machine Learning Classifier,” Proceedings of the International Conference on Intelligent Sustainable Systems, (2017).

[7] P. Supriya, M. Pavani, N. Saisushma, N. Kumari and K. Vikas, “Loan Prediction by using Machine Learning Models,” International Journal of Engineering and Techniques, (2019).

[8] R. Salvi, R. Ghule, T. Sanadi, M. Bhajibhakare, “HOME LOAN DATA ANALYSIS AND VISUALIZATION,” International Journal of Creative Research Thoughts (IJCRT), (2021).

[9] B. Srinivasan, N. Gnanasambandam, S. Zhao, R. Minhas, “Domain-specific adaptation of a partial least squares regression model for loan defaults prediction,” 11th IEEE International Conference on Data Mining Workshops, (2011).

[10] M. V. Reddy, Dr B. Kavitha, “Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis,” International Conference on Signal Acquisition and Processing, (2010).

[11] G. Chornous, I. Nikolskyi, “Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring,” IEEE Second International Conference on Data Stream Mining & Processing August (2018).