International Journal of Circuit, Computing and Networking

E-ISSN: 2707-5931
P-ISSN: 2707-5923
IJCCN 2022; 3(2): 44-48
Received: 11-05-2022
Accepted: 10-06-2022

Shruti Mishra
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamil Nadu, India

Shailki Sharma
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamil Nadu, India

Shreyansh Singh
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamil Nadu, India

Corresponding Author:
Shruti Mishra
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamil Nadu, India

Loan approval prediction

Shruti Mishra, Shailki Sharma and Shreyansh Singh

DOI: https://doi.org/10.33545/27075923.2022.v3.i2a.48

Abstract
A bank loan is an extension of credit by a bank to a customer or business. Lending is one of the primary financial products of any bank and the main profit comes directly from the loan's interest.
The loan corporations grant a loan after an intensive technique of verification and validation. However, they still don't know for sure if an applicant would be able to repay their loan. Banking processes usually use manual policies to determine whether or not an applicant is suitable for a loan from their bank.
The main purpose of this project was to determine whether an applicant is eligible for a loan by gathering data from multiple sources while using machine learning techniques to extract useful information. This would help lending organizations and banks to make the right decision for every loan approval.

Keywords: primary financial, learning techniques, help lending organizations and banks

Introduction
Machine Learning is a subcategory of Artificial Intelligence which focuses on teaching computers how to learn without using a given set of explicit rules. Machine learning systems use experience unlike a rule-based system which will perform any task with a given set of explicit rules. Performance can be improved by exposing the system to more data. It learns and adapts without following any given set of instructions. Our project focuses on using existing customers' details to analyse them further by applying a few machine learning techniques and predicting which future applicants can be approved for the loan.

Literature survey
We start with an all the more wide, methodical literature review that spotlights on the utilization of machine learning in the field of banking risk management in general. This section provides an outline of some of the previous work on developing ML models utilizing various algorithms to further develop the loan prediction process and assist banking authorities in selecting a qualified candidate with very low credit risk.
Risk management in banks has become progressively significant in guiding bank decision-making since the worldwide financial crisis. Accepting credit from prospective candidates is a significant part aspect of risk management. Given that loan prediction is a classification problem, we chose to use well-known classification techniques that have previously been used for a comparison problem.
Ashlesha Vaidya made use of logistic regression like a probabilistic and prescient way to deal with the process of loan prediction. He called attention to how logistic regression and artificial neural networks are generally utilized for loan prediction as they are more straightforward give the most reliable prescient results.
Suneel Sharma, Martin Leo, K. Maddulety's report researched as where the Ma- chine Learning is utilized in the fields of market risk, credit risks, operational risk and the liquidity risk just to conclude that numerous areas in bank risk management could benefit greatly from research into how machine learning can be applied to specific problems.
Regression Trees and Classification presented by Leo Breiman are mentioned to as CART. It best helps in both prescient and choice making issues. This binary tree method is a greedy method for choice of the best parting. Despite the fact that Decision trees provided us with a comparable level of accuracy. The benefits of Decision Trees in this case came from the fact that they gave equal weight to precision and prediction. This approach was successful in reducing the number of False Predictions and thereby lowering the risk factor. Another smart technique used by T. Sunitha and colleagues was to predict loan status using a Binary Tree and Logistic Regression.

A decision tree is an AI model that predicts the future. Anchal Goyal and Ranpreet Kaur investigate several ensemble algorithms. The ensemble method is a machine learning methodology that mixes numerous base models to create a single best-predictive model.

In their call for papers, Hong Wang and Lifeng Zhou used an improved random forests strategy to forecast credit default on imbalanced data sets. Predicting loan defaulters is an important function in the banking system since it directly affects benefit. Nonetheless, loan default data sets are incredibly unbalanced, resulting in poor algorithm performance. Finally, they came to the conclusion that the concept of coupled algorithms improves the model's precision.

Data mining is becoming increasingly popular in the field banking industry since it extracts information from massive amounts of data. Tarig Mohammed Ahmed and Aboobyda Jafar Hamid concentrated on implementing data mining methodologies for detecting credit hazard in the financial industry using three models: j48, bayes Net, and naïve Bayesdel. The authors used the Weka application to create and test models. They found a link between these algorithms in terms of correctly identifying data and accuracy in their research.

The information entered on the application form by each new applicant serves as a test data set. Based on the training data sets, the model will predict whether or not a loan will be approved.

Problem Statement

All types of home loans are handled by ABC Finance Company. Its consumers first apply for a house loan, after which the company verifies their loan eligibility. Because the process is time-consuming and laborious, the company opted to use machine learning to automate the loan approval process.

Now it was time to articulate the research work with ideas gathered in above steps by adopting any of below suitable approaches:

Machine Learning

Machine Learning is a study of computer science that uses experience to learn and improve itself, similar to humans. Many industries use machine learning algorithms to make predictions about their future growth and help them compete in their respective fields. An algorithm can be understood as a set of rules or instructions that is followed by a computer program.

Machine algorithms use algorithms to constantly improve itself from past experiences by using quality data sets. Advancements in this field are helping sectors like health-care, manufacturing, travel and financial services improve customer relationship management, reduce costs and can affect a company's bottom line.

Its importance can be understood by thinking of business companies using these algorithms to see patterns in business operations and give support to developing projects. Organizations can use machine learning tools to detect profitable possibilities and potential dangers more quickly. Leading companies like Uber, Meta and Google use machine learning to make data driven decisions which helps to keep up with the competition. One of the best known examples of machine learning in action is the recommendation system that enables Meta's news feeds.

The accessibility to enormous amount of data is directly proportional to the need of forming new predictive models with more accuracy. Human brains are the only ones on the planet capable of detecting patterns in data sets. Coming to the rescue to all this havoc is Machine Learning by designing algorithms that are efficient and swift and also forming data-driven models which are delivered in real time. Introducing intelligent alternatives to analyzing limitless quantity of available data, it can produce accurate findings and analysis. New approaches in the field are continually advancing, allowing for practically unlimited applications of machine learning. Machine learning and artificial intelligence are clearly here to stay, based on a large amount of data and facts.

Software and libraries used

Python: A high level programming language

Anaconda (jupyter Notebook) - A software used to run python programs

Libraries

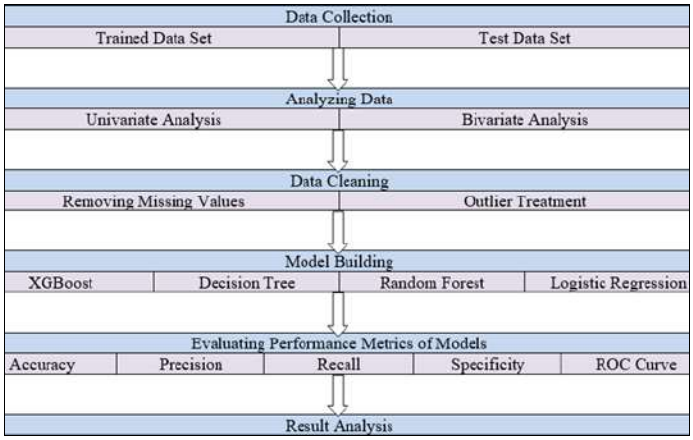
- **NumPy:** NumPy is a Python library that allows for more data storage while at the same time using less memory. It can be used as a multi-dimensional container for generic data.
- **SciKit Learn:** Scikit-Learn is one of Python's most popular machine learning li- braries. Because of its simple and easy-to-use API, Scikit-learn has gained widespread acceptance.
- **Pandas:** Pandas is a Python module that allows you to manipulate and operate on numerical tables and time series. Pandas make it simple to create, manipulate, and organize data. It is based on NumPy, which implies it requires NumPy to function
- **Matplotlib:** Matplotlib is an open source low level graph plotting library written in Python that fills in as a visualization utility.
- **Sea born:** The Python Sea born library is a generally well known data visualization library that is normally utilized for data science and machine learning assignments.
- **XG Boost:** Extreme Gradient Boosting is a well-known machine learning algorithm that finds the best tree model by using more accurate approximations. It is included in several popular machine learning libraries, such as scikit-learn for Python users.

Process

Methodology

Initiate Hypothesis

This is the first and foremost step that was performed even before looking at the data. It involved understanding the issue exhaustively by conceptualizing as many variables as we can.



Some of the things that we believed can influence the process of loan approval (the dependent variable in this loan prediction problem) are listed below:
Salary: Candidates with high income ought to have high chances of loan approval.
Previous history: Candidates who have paid off their previous obligations ought to have a better possibility of getting their loan approved.
Loan amount: The loan amount should also be considered in loan approval. Assuming that the loan amount is small, the possibilities of approval are probably going to be high.
Loan term: A loan for a more limited timeframe and for a lower sum ought to have better chances of approval.
EMI: If the sum is comparatively less that is to be paid monthly to reimburse the loan, greater is the likelihood of getting approval for the loan.

Reading the data
The train file was utilized to prepare the model so that it could learn from it. It incorporated every independent variable as well as the target variable.
All of the independent variables were present in the test document, however not the target variable. The model was utilized to foresee the target variable for the test data.

Understanding the data
- **Object:** Object format means variables are categorical. Loan ID, Self Employed, Dependents, Loan Status, Married, Education, Property Area, Gender are the categorical variables in our data set.
- **int64:** This is the format of Applicant Income. It is used to represent integer variables.
- **float64:** This format labels the variables that contain decimal values. They, too, are numerical variables. Loan Amount, Coapplicant Income, Loan Amount Term, Credit History are the numerical variables in our dataset.

Univariate Analysis
In this we examined each variable individually. Frequency tables were utilized to calculate the number of each category in a specific variable for categorical factors such as Loan ID, Gender, Married, Dependents, etc. Probability density plots were utilized to check the distribution of the variables. Probability density maps were also used to examine the distribution of a variable's numerical properties.

Bivariate Analysis
We looked at each variable separately in univariate analysis before looking at them in relation to the target variable.

For approved as well as unapproved loans, it was sensible to assume that the distribution of male and female applicants were almost equivalent.

Data Cleaning and Treatment
After investigating all the variables in our data, we could now credit the missing quantities and treat the outliers because outliers as well as missing data could have an adverse effect on the model execution.
- **Numerical variables:** Only numerical variables could be used to calculate the mean and median
- **Categorical variables:** mode based imputation

Loan Amount contained outliers so they were treated on the grounds that the existence of outliers influences the distribution of the data.
As a result of these outliers, while the right tail was longer, most of the data in the loan amount was on the left side. It is referred to as right skewness. The first step in removing skewness should most likely be log transformation. When log transformation is used, it greatly reduces the larger values but has little effect on the smaller values. As a result, a distribution that is similar to the normal distribution is obtained.

Feature Engineering
In view of subject area, new elements were thought of that could influence the target variable. The following three new highlights are created:
- Total Income - As discussed during bivariate analysis, Applicant Income and Compliant Income were joined. The possibility of loan getting approved could be high assuming that the total income is high.
- EMI - EMI is a regularly scheduled installment made monthly by the candidate for the repayment of the loan. This variable was created because individuals with high EMIs may struggle to repay the loan. EMI can be calculated by taking the ratio of loan amount to the loan amount term.
- Balance Income - When the EMI has been paid, this is the income left afterwards. This variable was created assuming this worth is high, the possibilities of an individual will repay the loan and hence increasing the chances of loan approval.

Model building
We proceeded with the model-building process in the wake of adding new elements. So we began with logistic regression and move gradually to more complex models like XG Boost and Random Forest.

Logistic Regression

It comes under the category Supervised Learning technique. Here, we use the well-known Logistic Regression to construct our model which gives an outcome of discrete value.

The output can be Yes or No, 0 or 1, true or false, and so on. It provides some probabilistic values that falls somewhere in the range of 0 and 1, rather than giving the exact values of 0 or 1. We got an accuracy score of 0.778 for this model, which is equivalent to 77.8%.

Decision Tree

A Decision Tree is a tree-like design that comprises of a root node, branches, and leaf nodes. The root node is the tree's highest node and each internal node exhibit a test on an attribute, each branch exhibits the result of a test, and each leaf node exhibits a class label.

To choose whether to divide a node into at least two sub nodes, decision trees employ numerous algorithms. The making of sub nodes extends the uniformity of the sub nodes that result.

We got an accuracy score of 0.681 for this model which is equivalent to 68.1%.

Random Forest

Random forest is a classifier that utilizes an average of a number of decision trees on various subsets of a given dataset to further develop classification accuracy.

A random sample of rows and a couple of randomly picked variables are used to construct a decision tree model for every individual learner. The final prediction can be an element of all of the individual predictions made by the learners. We got an accuracy score of 0.735 for this model, which is equivalent to 73.5%.

XG Boost

Extreme Gradient Boosting is defined a gradient boosting algorithm which is an open- source library that implements the in a proficient and effective way. We got an accuracy score of 0.767, which is equivalent to 76.7% and is close to Logistic Regression.

Conclusion

Out of all the classification algorithms used on the dataset, the Logistic Regression algorithm gave the best overall prediction accuracy.

Credit History, Balance Income, EMI, and Property Area were the most important factors for predicting the class of the loan applicant (whether the applicant would be 'approved' or 'not').

In the near future, this module of prediction could be integrated with the module of the automated processing system. The system is trained on an old training dataset in future software can be made such that new testing data can be used after a certain time.

We can train the XG Boost model using grid search to optimize its hyper parameters and improve its accuracy.

Future Scope

Time Series Analysis can be used to anticipate the approximate time when a client will default utilizing loan data from several years.

Future analysis can be performed by estimating the approximate interest rates that the loan applicant will be paid based on his profile if the loan is authorized. This is beneficial for loan applicants because some banks accept loans but charge customers exorbitant interest rates.

An app with a good user interface can be constructed that will accept numerous inputs from the user, such as name, address, loan amount, loan term, and so on, and forecast whether or not their loan application would be granted by the banks based on their inputs, as well as anticipated interest rates.

Acknowledgment

The authors acknowledge the support from SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu.

References

1. Vaidya A. Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In 2017 8th International Con- ference on Computing, Communication and Networking Technologies (ICCCNT) IEEE; c2017 Jul, p. 1-6.
2. Leo M, Sharma S, Maddulety K. Machine learning in banking risk management: A literature review. Risks. 2019;7(1):29.
3. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Routledge; c2017.
4. Sunitha T, Chandravallika M, Ranganayak M, Suma Sri G, Jagadeesh TVS, Tejaswi A. Predicting the Loan Status using Logistic Regression and Binary Tree, 2021 January 20. ICICNIS; c2020.
5. Goyal A, Kaur R. A survey on ensemble model for loan prediction. International Journal of Engineering Trends and Applications (IJETA). 2016;3(1):32-37.
6. Zhou L, Wang H. Loan default prediction on large imbalanced data using random forests. Telkomnika Indonesian Journal of Electrical Engineering. 2012; 10(6):1519-1525.
7. Hamid AJ, Ahmed TM. Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal. 2016;3(1):1-9.
8. Amin RK, Sibaroni Y. Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region). In 2015 3rd International Conference on Information and Communication Technology (ICoICT). IEEE; c2015 May p. 75-80.
9. Sheikh MA, Goel AK, Kumar T. An approach for prediction of loan approval using machine learning algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) IEEE; c2020 Jul p. 490-494.
10. Arora N, Kaur PD. A Bolas so based consistent feature selection enabled random forest classification algorithm: An application to credit risk assess- ment. Applied Soft Computing. 2020;86:105936.
11. Yang B, Li L, X, Ji, H, Xu J. An early warning system for loan risk assessment using artificial neural networks. Knowledge-Based Systems. 2001;14(5-6): 303-306.
12. Madaan M, Kumar A, Keshri C, Jain R, Nagrath P. Loan de- fault prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering. IOP Publishing. 2021;1022(1):012042.

13. Ruud M, Nilsen HB. A Comparative Study in Binary Classifica- tion for Loan Eligibility Prediction (Master's thesis, Handelshøyskolen BI); c2021.
14. Aditya Sai Srinivas T, Ramasubbareddy S, Govinda K. Loan Default Prediction Using Machine Learning Techniques. In Innovations in Computer Science and Engineering. Springer, Singapore; c2022. p. 529-535.
15. Yadav O, Soni C, Kandakatla S, Sawant S. Loan Prediction System Using Decision Tree.
16. Dosalwar S, Kinkar K, Sannat R, Pise DN. Analysis of Loan Availability using Machine Learning Techniques. International Journal of Advanced Research in Science, Communication and Technology, September; c2021. p. 15-20.
17. Manglani R, Bokhare A. Logistic Regression Model for Loan Prediction: A Machine Learning Approach. In 2021 Emerging Trends in Industry 4.0 (ETI 4.0) IEEE; c2021. May p. 1-6.