Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766128

LOAN APPROVAL PREDICTION USING MACHINE LEARNING

Ch Seshadri Rao[1*], P Naveen[2], P Akhil Niyogi[3], S Bhuvana Satya[4]

[*1] Associate Professor, Department of CSE, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

seshadrirao.chinta@raghuenggcollege.in

[2,3.4] Students, Department of CSE, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

19981a05d0@raghuenggcollege.in, 19981a05d6@raghuenggcollege.in , 18981a05d6@raghuenggcollege.in

ABSTRACT:

Loan approval prediction is a crucial task in the banking sector that involves assessing the creditworthiness of a candidate based on historical data. The process of evaluating loan applications through manual methods has been known to be prone to errors and can be time-consuming. This highlights the need for an automated system that utilizes machine learning algorithms to provide efficient and accurate loan approval predictions. The objective of this study is to develop a machine learning model using different classification algorithms to predict loan approval based on historical data of loan applicants. The study identifies relevant features for the prediction task by collecting, pre-processing, and analyzing the data. Three different classification algorithms are used to build the predictive models, and their performance is evaluated using cross-validation techniques. The study concludes that machine learning algorithms can significantly improve the loan approval prediction process in the banking sector.The Random Forest with Grid Search CV algorithm outperforms XGBoost and random forest algorithms used in this study in predicting loan approval, and the developed model can expedite the loan approval process and reduce errors.

Keywords: Machine learning, Loan approval prediction, XGBoost, Random Forest, Grid Search CV, Cross-validation.

1.INTRODUCTION:

The banking sector is facing a challenge in managing the increasing number of loan applications due to digitization and online loan applications. Artificial intelligence (AI) has become a popular method for information analysis in various industries, including banking. AI algorithms can help banks in selecting deserving customers and minimize the chances of mistakes in the loan approval process. As loan approval is a significant source of profit for

banks, choosing the right customers is essential to avoid losses[1]. Automating the loan approval process can improve the efficiency of lending operations and lead to increased customer satisfaction and cost savings. To achieve these benefits, a robust loan prediction model is essential for accurately determining which loan applications to approve and which to reject, thus reducing the risk of loan default[2].The motivation behind this study is to develop a machine-learning model that can predict loan approval accurately and efficiently. Several studies have been conducted on loan approval prediction using machine learning algorithms. These studies have shown that machine learning algorithms can significantly improve the loan approval process's efficiency and accuracy. Various classification algorithms such as Decision Trees, Random Forest, and Logistic Regression have been used in these studies. This paper aims to explore various machine learning approaches that can accurately identify suitable loan candidates and assist banks in identifying potential loan defaulters, resulting in significantly reduced credit risk[3].

Problem Statement:

Automating loan approval using machine learning to predict candidate safety and expedite the process for the benefit of both banks and loan candidates[4].
The paper is structured into several sections, each serving a distinct purpose in presenting the research in a clear and organized manner. The literature survey provides an overview of previous studies on loan approval prediction using machine learning algorithms and identifies gaps that the current study aims to fill. The method section outlines the data collection process, feature selection, and classification algorithms used for prediction. The results section presents the findings of the research, including accuracy and performance of different algorithms used. The discussions section analyzes the results and discusses their implications, highlighting the strengths and limitations.The conclusion summarizes the main findings and contributions, emphasizing the significance of the study for improving loan approval processes and future research.Finally, the references section lists all sources cited in the paper, grounding the research in existing literature and research on the topic.

2.LITERATURE SURVEY:

The evaluation of loan applicants' creditworthiness based on their historical data is a critical aspect of the banking industry. As the number of loan applications grows, manual methods of evaluating loan applications become ineffective and prone to errors. Therefore, there is a growing need to develop automated systems that use machine learning algorithms to provide accurate and efficient loan approval predictions. This literature review seeks to examine previous research on the use of machine learning algorithms for bank loan approval prediction. Several studies have been conducted on bank loan approval prediction using machine learning algorithms.
L. Udaya Bhanu et al.[5], suggested automating the loan approval process based on customer details to save time for banks. They propose using a machine learning model to classify loan candidates and reduce credit risk. The goal is to predict loan approval using supervised learning

techniques, with random forest classification showing the best accuracy. The proposed application is effective and meets all requirements, with potential for integration into automated processing systems in the future.

According to Vishal Singh et al. [6], technology has significantly improved our quality of life, and this is particularly evident in the banking sector. Machine learning algorithms such as logistic regression, random forest classifier, and support vector machine classifier can be used to predict loan approval based on historical data. Given that loan recovery is crucial for a bank's profit and loss, various factors such as loan duration, amount, age, income, and credit history are taken into account when determining loan approval.

Zoran Erieiz [7] conducted a study that utilized Decision Trees, Neural Networks, and Logistic Regression models to predict loan approval based on various factors such as income, employment history, credit score, and debt-to-income ratio.

Ramachandra H V et al. [8], investigated the use of machine learning algorithms for loan approval prediction in their study. Specifically, they utilized Decision Trees and Logistic Regression models to predict loan approval based on factors such as age, income, loan amount, credit score, and loan duration.

Bhoomi Patel et al.[9], conducted a study where they applied data mining methodology to predict the likelihood of default from a dataset that included information on home loan applications. The aim of their study was to help banks make better decisions in the future regarding loan approvals.

Xin Li et al.[10], focused on the application of the LSTM-SVM model in user loan risk prediction. They discussed the current economic background and traditional risk forecasting methods before proposing a prediction methodology based on LSTM and SVM.

Several authors contributed to build a machine learning model to predict approval of loans to customers using techniques such as Random Forest, Logistic Regression, XGBoost, Support Vector Machine, K-Nearest Neighbors, and Decision Trees[11-15]. These algorithms have shown promising results in predicting loan approvals, indicating the potential for automated loan processing systems. However, there is a need for further research to explore and compare the performance of different algorithms and their combinations to develop more accurate and reliable loan approval prediction models.

3.METHOD:

In this study, we explore the effectiveness of the random forest algorithm in combination with the grid search cross-validation technique to predict bank loan approval. The objective is to demonstrate how fine-tuning the hyperparameters of the random forest model through grid search CV can enhance its performance and reliability. The study's methodology is illustrated in Figure 1, which presents the architectural diagram of the proposed approach. The figure provides an overview of the data flow and the steps involved in building and evaluating the model. The results of the study indicate that the proposed approach can significantly improve loan approval prediction accuracy.

The method section of the study comprises four main steps. Firstly, data set collection and pre-processing, which involves collecting relevant data and preparing it for analysis. The second step is feature selection, in which the researchers select the appropriate variables to be included in the model. The third step involves model building and hyper parameter tuning to optimize its performance. Finally, the model is trained and evaluated in the fourth step.

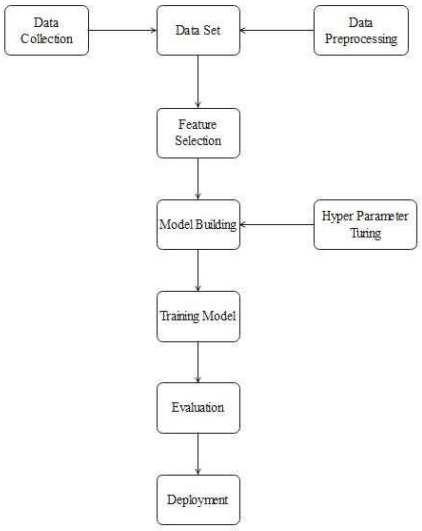

FIGURE 1: ARCHITECTURE DIAGRAM

1.Data set collection and pre-processing:

The study shows that combining Random Forest with Grid Search CV can result in a highly optimized and robust model. The study utilized a publicly available dataset from Kaggle comprising 614 instances and 13 features. The target variable is the loan status, indicating approval or rejection. The dataset was analyzed in Jupyter Notebook using pandas, NumPy, matplotlib, and sci-kit-learn libraries. The 13 features analyzed in the study are shown in figure 2.

Column name	Datatype
Loan id	object
Gender	object
Married	object
Dependents	object
education	object
self_employed	object
ApplicantIncome	int64
coapplicantincome	float64
LoanAmount	float64
Loan_amount_term	float64
Credit_history	float64
Property_Area	object
Loan_Status	object

Figure 2:Data Describtion

The study identified outliers in some numerical columns using box plots and scatterplots. Missing values in these numerical columns were replaced with the median value to prevent adverse effects on the model's performance. For categorical columns, missing values were filled using the mode. Categorical variables were encoded using label encoder to represent them as numerical values for the model training.

2.Feature selection:

In this study, SelectKBest[16] was used as the feature selection technique, which selects the top K features based on a given scoring function. The f_regression scoring function was used in this study, which calculates the F-value between each feature and the target variable. The SelectKBest method was applied with k=3 to select the top three important features. The features selected by this method were then used for model building and evaluation.

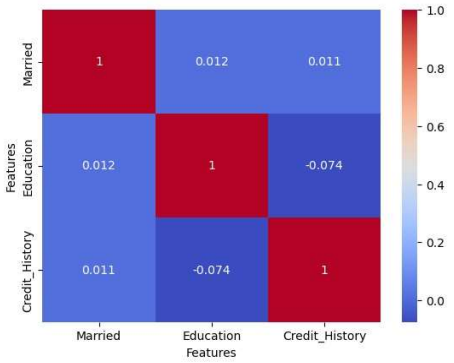

FIGURE 3: HEAT MAP

The Figure 3 shows that the selected features ('married', 'education', and 'credit_history') had weak pairwise correlations with coefficients ranging from -0.074 to 0.012. This suggests that using all three features together may lead to better performance in predicting the target variable y.

3.Model building and Hyper parameter tuning:

In this study, the Random Forest algorithm with Grid Search CV was utilized to predict bank loan approval. The impact of hyperparameter tuning on model performance was evaluated, and the results showed that combining Random Forest with Grid Search CV can result in a highly optimized and robust model.

4.Model Training and Evaluation:

During the model training phase, the dataset was split into a training set and a testing set.The training set was used to train the Random Forest model with the selected hyperparameters using Grid Search CV.The testing set was then used to evaluate the model's performance. The

accuracy, precision, recall, F1 score, and sensitivity were used as performance metrics to evaluate the model's performance on the testing set.
Classification Algorithms:

Random forest:

Random forest is a machine learning algorithm that combines multiple decision trees to improve the accuracy and robustness of the model[17].
XG Boost:

XGBoost is a gradient boosting machine learning algorithm that uses an ensemble of decision tree models to improve the predictive performance of the model[18].

4.RESULTS:

Our study found that Random Forest outperforms XGBoost in predicting bank loan approvals with an accuracy of 0.7888 compared to 0.7777. Fine-tuning the hyperparameters of Random Forest using Grid Search CV significantly improves accuracy to 0.8111. Table 1 shows a table comparing the accuracy, Precision, F1-score, Recall, and Sensitivity of the three algorithms - XGBoost, Random Forest, and Random Forest with Grid Search CV.

Algorithm	Accuracy	Precision	F1- score	Recall	Sensitivity
XG boost	0.7777	0.7971	0.8461	0.9016	0.9016
Random Forest	0.7888	0.7837	0.8592	0.9508	0.9508
Random Forest with Grid search CV	0.8111	0.7820	0.8776	1.0	1.0

Table 1: Comparison of Model Performance Metrics

Figure 4 shows that Random Forest with Grid Search CV has the highest accuracy among the three algorithms. These findings highlight the importance of hyperparameter tuning in improving machine learning model performance.

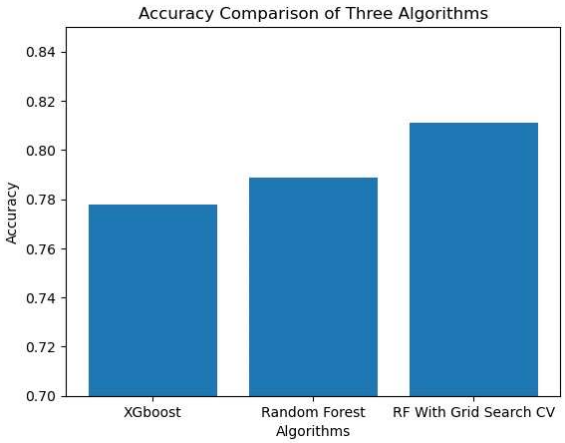

Figure 4: Comparison of the accuracies

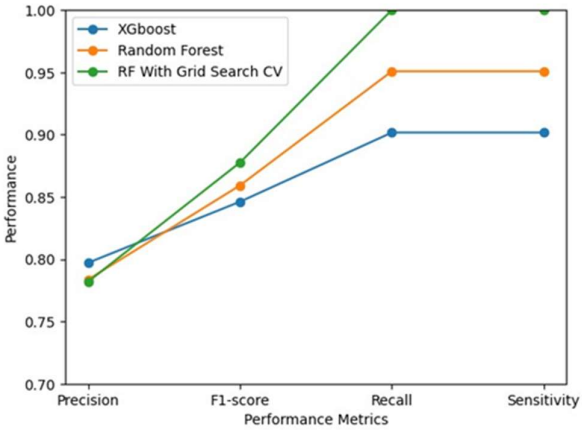

Figure 5: Comparison of Performance Metrics

Figure 5 represents a line graph comparing the Precision, F1-score, Recall, and Sensitivity of the three algorithms. The graph demonstrates that the Random Forest model with Grid Search CV outperforms the other two algorithms in terms of F1-score, Recall, and Sensitivity. This provides further evidence of the effectiveness of hyperparameter tuning using Grid Search CV in enhancing the performance of the Random Forest model.

5.DISCUSSIONS:

Grid search offers an automated and systematic approach to hyperparameter tuning, which can significantly enhance the model's performance. In contrast, manual tuning can be a tedious and time-consuming process that may result in suboptimal model performance. Our study demonstrated that implementing grid search for tuning hyperparameters in a random forest model increased accuracy from 0.7888 to 0.8111. This improvement was achieved by systematically exploring a range of hyperparameters and selecting the optimal set that maximized accuracy. Grid search is advantageous over manual tuning as it can search through all possible combinations of hyperparameters in a systematic manner, leading to better accuracy. Additionally, grid search can be easily parallelized, allowing for faster optimization of hyperparameters. In conclusion, employing grid search for hyperparameter tuning in machine learning models is a preferable approach as it provides a more efficient and systematic way of optimizing model performance. One limitation of using Grid Search CV for hyperparameter tuning is that it can be computationally expensive, especially when dealing with a large number of hyperparameters or when using a large dataset. In such cases, a Randomized Search CV can be a more efficient alternative as it searches only a subset of the hyperparameters randomly.

6.CONCLUSION:

The study developed a machine learning model for loan approval prediction using historical applicant data. The Random Forest algorithm with Grid Search CV achieved the highest accuracy of 0.8111. The study identified marital status, education, and credit history as important predictors of loan approval using the SelectKBest method. These features can be used to streamline the loan approval process and reduce evaluation time and effort in the

banking sector. The study highlights the potential of machine learning algorithms to automate and improve the loan approval prediction process. Future research can focus on adding additional data columns and developing explainable and interpretable machine-learning models for loan approval prediction.

7.REFERENCES:

[1] Bank Loan Prediction System using Machine Learning by Anshika Gupta, Vinay Pant, Sudhanshu Kumar and
Pravesh Kumar Bansal. Proceedings of the SMART–2020, IEEE Conference ID: 50582 9th International Conference on System Modeling & Advancement in Research Trends, 4th–5th, December 2020 Faculty of Engineering & Computing Sciences, Teerthanker Mahaveer University, Moradabad, India .

[2] An Approach for Prediction of Loan Approval using Machine Learning Algorithm by Mohammad Ahmad Sheikh, Amit Kumar Goel and Tapas Kumar, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4 .

[3] Bank Loan Approval Prediction Using Data Science Technique (ML) by Subhiksha R, Vaishnavi L, Shalini B, Mr. N. Manikandan,DOI Link: https://doi.org/10.22214/ijraset.2022.43665.

[4] The loan prediction using Machine Learning by Dr.C K Gomathy, Ms.Charulatha,Mr.AAkash ,Ms.Sowjanya, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056

[5] Customer Loan Prediction Using Supervised Learning Technique by L. Udaya Bhanu , Dr. S. Narayana, DOI: 10.29322/IJSRP.11.06.2021.p11453

[6] Prediction of Modernized Loan Approval System Based on Machine Learning Approach by Vishal Singh, Ayushman
Yadav and Rajat Awasthi,2021 International Conference on Intelligent Technologies (CONIT) Karnataka, India. June 2R5-27, 2021 .

[7] Predicting Default Loans Using Machine Learning(OptiML) by Zoran Erieiz,member,IEEE.

[8] Design and Simulation of Loan Approval Prediction Model using AWS Platform by Ramachandra H V, Balaraju G, Divyashree and Harish Patil,2021 International Conference on Emerging Smart Computing and Informatics (ESCI) AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021 .

[9] Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal "Loan default forecasting using data mining" Department of Information Technology, St. Francis Institute of Technology, Mumbai, India (2020)

[10] Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li "Overdue Prediction of Bank Loans Based on LSTM-SVM"Jiangsu Key Lab of Big Data and Security and Intelligent Processing Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.

[11] "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier", G. Arutjothi, Dr. C. Senthamarai department of computer applications government arts college (Autonomous) Salem, India (2017.)

[12] "Credit Risk Evaluation using Hybrid Feature Selection Method", Arutjothi .G, Dr. C. Senthamarai,Software engineering and technology (2017)

[13] "An Improved Algorithm for Efficient Mining of Frequent Item Sets on Large Uncertain Databases", Arutjothi .G, Dr. C. Senthamarai in International Journal of Computer Applications, Volume 73, No. 12 July 2013.

[14]A novel ensemble decision tree classifier using hybrid feature selection measures for parkinson's disease prediction", Bala brahmeswara kadaru et al, Int. J. Data science (IJDS), ISSN: 2053-082X, Vol.3, No.4,2018.

[15] "Data mining techniques to analyze risk giving loan (bank)",] Mrunal Surve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit, Internation Journal of Advance Research and Innovative Ideas in Education Volume 2 Issue 1 2016 Page 485-490.

[16] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[17Random forests. Machine learning, Breiman, L. (2001). , 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

[18] XGBoost: A scalable tree boosting system, Chen T & Guestrin C (2016). In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).