



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: III Month of publication: March 2024

DOI: <https://doi.org/10.22214/ijraset.2024.59231>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Machine Learning Based Classification Model for Prediction of Bank Loan Approval

Sameerunnisa.SK¹, M.N.V. Sai Rama Harsha², K. Tarun Teja³, K.V. Swathi⁴, K.T.V. Manikanta⁵

Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology

Abstract: In this paper we creating and testing a smart loan prediction system using a special type of classifier called gradient boosting. Our methodology is meticulously laid out, covering a spectrum of essential stages including data pre-processing methodologies, in-depth exploratory analysis, and rigorous model training leveraging a dataset meticulously curated from Kaggle. Our model, with 300 estimators, a learning rate of 0.05, and a maximum of six features, achieves an astounding 98.03% accuracy. We utilize the pickle module to serialize our trained model, consequently deploying it as a web application. This application, which is powered by React for the front end and Flask for the back end, offers users a simple interface for entering loan application details and receiving real-time forecasts on loan acceptance or denial. Our research shows that gradient boosting classified the well in predicting the loans to the users.

Keywords: Gradient Boosting Classifier, loan prediction, machine learning, Flask, React, pre-processing, exploratory data analysis, model training, accuracy score, web application, loan approval.

I. INTRODUCTION

Dwelling on that which separates the automated loan prediction systems with others in the field of finance, automation helps to speed up the loan processing and to increase its efficiency. They are procedures which mainly mean the types of machine learnings such as deep learning that do use the past information of the system and predict whether loans would be approved or not legally. Knowing the true status of creditworthiness and risk factors are very helpful for banks, among others, to have a better judgment that translate to savings in time and money as there will be an abrupt promise to less risks.

In this paper, we present the concept and implementation of a machine learning-based loan prediction on the basis of three goals which are efficiency, fairness, and good performance in loan selection processes. With The Loan-Approval- Prediction-Dataset which was sourced from Kaggle, we will be analyzing the attributes data after which we will be seeing patterns or key factors that can influence the approval of a loan.

These attributes include credit scores (cibil_score), income (income_annum), loan amounts (loan_amount), loan terms (loan_term), asset values (residential_assets_value, commercial_assets_value, luxury_assets_value, bank_asset_value), and demographic information (no_of_dependents, education, self_employed).

By examining the topical ones with advanced data processing techniques such as feature extraction, exploratory data analysis, and model training, we will attempt to create a modelling system that will predictably assess credit apps and loan term forecasts. The significance of this research is demonstrated in its ability to realize what we call transformational lending system through automated, precise and non-discriminatory procedures thus, banks will incorporate seamless loan procedures and proper ways of loan decision.

First of all, we acquired a collection of loan data which was going to be used for our task of loan data prediction[4] using modern deep learning techniques. Secondly, we carried out data pre-processing and feature engineering to get more precision. The set of measures included data processing, model selection, hyperparameter tuning, cross-validation, and ensembling. After the data normalization, removing missing values, or changes numerical data to numerical codes and mark the categorical variables, now we were able to start our analysis. Secondly, feature engineering techniques devised the building of features which bear information important to the task of loan prediction model creation is followed. The trials use Gradient Boosting Classifier algorithm, which turned out to be most successful of all and so that version was chosen for the future work due to its previous strong performance and applicability of the current task. Grid Search was an important option out of the other processes that nail down the hyperparameters to alter the model's performance. Furthermore, users the procedure of cross-validation as the validation measure and avoided the overfitting phenomenon through this methodology. The boosting approach delivered precision and performance which could be substantially lower the weakness and variance. As for the judges, resulted good output was the model effectiveness which showed on the validation dataset that was shared for this purpose of making evaluation of the performance.

These numbers for the accuracy, precision, recall as well as F1-score manifest to be the way for detailed examinations. At last, we made a web application, which is responsible to give loan credits in real time, according to the prediction made by the Gradient Boosting Classifier. Focusing on this aspect's application will run as its performance monitoring will be vital.

The Loan-Approval-Prediction-Dataset used in this study comprises 4269 records, with attributes including credit scores, income, loan amounts, terms, asset values, and demographic information. The dataset was divided into training and testing sets, with approximately 80% allocated for training and 20% for testing. After training and hyperparameter tuning, the final accuracy of the Gradient Boosting Classifier was evaluated, achieving 98.03% accuracy on the testing set. This accuracy underscores the model's proficiency in predicting loan outcomes.

This approach ensures robust model training and evaluation, offering insights into the effectiveness of machine learning models for loan prediction[4] in real-world scenarios.

Of course, we struck a couple of libraries and we did not use them only for the purpose of getting helpful information; we managed to develop a secure machine learning model for loan prediction based on them.

Pandas aided the convenient manipulation and data pre-processing, while scikit-Learn was a rather informative algorithm library which contained these and other tools, utilities for models selection, creation, training, and evaluation; hyperparameter tuning, cross validation and ensemble methods. Seaborn and Matplotlib were used for data visualization. These visualizations showed the experience of the model and helped in the performance study as well, indicating whether the model can be improved. As part of its operations, Joblib implements object serialization and deserialization mainly of Python objects, especially those exploited by machine learning training and does make the process of model saving and loading an easier one. At the same time, amidst android application development, Flask was utilized as API endpoints in order to achieve a lightweight and flexible framework. You can manage to show the machine learning model as well as the predictions outcome through memorable, interactive user interfaces. As they clustered together, the data libraries served as an indispensable source for the model training purposes: they executed data pre-processing and training of the model, before letting the users interact with it.

II. LITERATURE REVIEW

The introduced system is the full application of different experiments aimed at the usage of the machine learning technology for the prediction of the same. As an instance, the comparative study of Jehad Ali et al. [5] provides the people an at large way of understanding Random Forest and the decision tree classifiers in various kinds of datasets. The provided approach to classify the representation of performance assessment they use provided helpful conceptual resource for our project as we borrow similar methods to evaluate the performance of different algorithms in our work.

Besides this, it is worth mentioning that the idea of the random forests by Adele Cutler et al. [9] became the base component for our model architecture as well. Their study of random forest principles in the field which includes feature selection and error rate comparison, had a pivotal role in our approach to the problem analysis and created a robust machine learning framework for loan approval prediction.

The study of Prabaljeet Saini and his research group is an abstract of original research. machine learning algorithms for loan approval prediction encouraged us to use various classifiers in our project but we were intrigued to examine how various classifiers behave under the same project. We came up with the same methodology to analyze the performance of the different classifiers; these include Random Forest, Gradient Boosting, and the rest; and finally after the analysis, we finally chose the best classifier model for the predictive system.

In addition, the investigation led by Pratheeksha Hegde N and his group[10]. shows the importance of feature selection, data processing before the model evaluation in the process of loan approval prediction. We used their layout as a blueprint, we implemented the model as a peripheral and we developed our prediction tool to be user friendly in a manner the provided comprehensive results.

Staking R. Priscilla et al[2]. has been illustrated in their study. this demonstrates the significance of loan approval prediction in the banking sector and shows what risks a bank will face if it uses wrong prediction models which can lead to loss of income and profit. From their work, for instance, we borrowed their preprocessing data ways and way of evaluation of our loan approval system in order to achieve the reliability and effectiveness followed by other systems just like the ones of before. On top of that, their designs' focus on creating simple prediction applications appealed to us. It was in line with our aim/goal of having people interact with our predictive system with minimal fuss.

Eventually such information facilitated the framework's titration towards a higher level of proficiency to which it was responsive to the needs and objectives of the financial institutions.

III. DATA COLLECTION AND PREPROCESSING

The dataset sourced off Kaggle, entitled as, "Loan-Approval-Prediction-Dataset", is the manifestation of this study. Featuring 4269 different records and variables like credit scores, income, loan amounts, terms and conditions, asset values and demographic information, it gives one of the richest databases that is critical for loan approval prediction model. Through using this dataset, the work called "a machine learning model for a loan decision support system" will test, create and adopt. The goal that can be achieved by those factors research is that these key attributes will identify which decisive factors in loan approval process 1 directly influence loan approval decisions and align the objectives of efficiency and fairness in loan approval process 2. With scrupulous computation and the generation of such model, the researchers get a deep insight to the complicated loan approval process, consequently leading the way to development of enhanced methods of predicting the loan outcomes and improving existing loan assessment methods.

Those preprocessing techniques are carried out in this essay are the key factors which help to get the dataset ready for analysis, containing the removal of outliers, normalization and encoding. Firstly, the working on the dataset was done that displayed the high accuracy of the statistics by handling the missing values sufficiently, thus maintaining the integrity of the data. Afterwards, the normalization procedure was defined for numerical values, so the attributes with different sizes or units had an equal contribution to the model. As a result, encoding was utilized for converting the categorical variable into a numerical format which is in the range of acceptable by the machine learning algorithms. At this stage, I have used encoding techniques like one-hot and label encoding, document encoding, and categorical feature encoding where applicable. Through these preparative steps taken carefully, the dataset was reworked to become beautifully clean, standardized, and machine-readable, ready for further analysis and modeling.

In the process of preprocessing we have outspoken that attaining the data integrity and reliability requires passing over missing data. Given the nature of the provided code snippet, it clearly shows that missing instances were eliminated using the `dropna()` function. This also brings the attribute of a single-attribute position which enhances data completeness and reduces bias. Rather than doing this, such a strategy in a sense increases in productivity and reduces the computational effort. However, one should be cautious about the consequences of this on the representativeness of input data and their generalizability. On top of that, the action of removing records with missing data puts forward the idea that a strong end of data processing is needful in order to improve the precision and dependability of the predictive modeling model. Throughout the course of this process, numbers of credit scores, income, and loan amounts together with numerals of asset value, over and over, had crucial role in choosing preprocessing procedures and checking whether the dataset was suitable enough for the former analysis.

IV. EXPLORATORY DATA ANALYSIS (EDA)

Explicitly emphasized to explore and comprehend the nature of the data via the Exploratory Data Analysis (EDA) technique to gain overall insights into all the data representations, their distributions and what interlinkages (if any) exist between the variables. After scan of the code, we used libraries such as pandas, seaborn and matplotlib for plotting and analysis of the dataset. The distribution of numerical attributes was discovered via EDA. This includes credit scores, income and loan amounts, represented through visual presentation with histograms, box plots and density plots. This helped us visualize the tendencies and spread of the data. In addition, we have made correlations and inter-variables relationships using techniques such as heatmap figures and correlation matrix for that purpose. Analysing our data, we were able to uncover the major patterns, directions, and relationships in the database, thus obtaining important inferences that led our project to the following step, preprocessing and model implementation.

V. MODEL DEVELOPMENT

To accomplish that, we utilized the Gradient Boosting Classifier, an ensemble learning method, because it is a technique that brings into play various prediction models to build a robust statistical estimator sequentially joining forecasts of numerous decision trees referred to as weak learners. This structure was custom-fitted to our loan approval prediction task, wherein we employed specific numerical terms and computations:

This structure was custom-fitted to our loan approval prediction task, wherein we employed specific numerical terms and computations:

Base Learners (Decision Trees): Decision trees were the most common building blocks utilized within the Gradient Boosting framework. These decision trees were specifically developed less deep to obtain an optimum result from a numerical feature of a dataset, such credit score, income level, loan amount and asset value.

Decision trees recursively partition the feature space into non-overlapping regions which maximize impurity in the data, usually when measured using the Gini impurity metric or entropy.

The node split decision is based on maximization of impurity criterion at each node and minimizing the impurity of disjoint subsets.

$$\text{Gini impurity: } Gini(t) = 1 - \sum_{i=1}^J p(i | t)^2$$

$$\text{Entropy}(t) = - \sum_{i=1}^J p(i | t) \log_2(p(i | t))$$

Where:

t represents a particular node in the decision tree.

J denotes the number of classes in the target variable.

p(i|t) represents the proportion of samples of class i at node t.

- 1) *Loss Function Optimization:* This model has found its roots in the efforts to minimize loss function defined, as for example, the binary cross-entropy, calculated as difference between the true loan approval outcomes and the predicted probabilities. As the model steadily enhanced its decision trees' performance by fine-tuning several of its decision making factors including the split guidelines and value in the leaves, we witnessed an increase in its prediction accuracy.

The function of loss reflects the true distance between the predicted values and the real targets. Along the gradient-boosting line, the loss function is rather optimized with the use of gradient descent which is updated iteratively to boost the model parameters. Binary Cross-entropy LOSS = $-\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$

Where:

N represents the number of samples.

Yi denotes the true binary label for sample i.

Pi represents the predicted probability of the positive class for sample i.

- 2) *Gradient Descent:* Gradient descent was employed to update the decision trees' parameters in each iteration. Specifically, we computed the gradient of the loss function with respect to the predicted probabilities, adjusting the decision trees' parameters in the direction that minimized the loss.
- 3) *Learning Rate:* We fine-tuned the learning rate hyperparameter, which regulated the impact of each decision tree on the final ensemble. This numerical value influenced the speed and stability of the model's convergence during training, with a lower learning rate necessitating more iterations for optimization.
- 4) *Regularization:* To mitigate overfitting, we integrated regularization techniques such as limiting the maximum depth of the decision trees, reducing the learning rate, and introducing stochasticity through subsampling of the training data and features. Moreover, to start with the Random Forest classifier was incorporated, it was changed afterwards to the Gradient Boosting Classifier since this classifier is simple and dynamic i.e. it is easy to enhance the performance at every learning cycle. However, in contrast to the Random Forest technique which is about multiple attempts to get a strong result based on repeated weak spells, Gradient boosting is about generation of a powerful spell at once using a chain of weak ones. This is achieved by decreasing the loss function through gradients methods which in turn gets an owner of the data sets to know error patterns and also to get the complex patterns of the given datasets. Additionally, the Gradient Boosting has the option of hyperparameter adjustment, this element is aimed at making the suitable adjustments for proper model fitting in order to ensure best model performance Without regression analysis, it is likely that we will fail to cater for non-linear features in data and complex patterns despite having a choice of an algorithm that has the capacity to do arguably the best job in terms of the loan approving task.
- 5) *Hyperparameters tuning:* For hyper parameter tuning, we typically focus on parameters specific to each algorithm, parameters like the learning rate, number of estimators, and maximum depth for the Gradient Boosting Classifier. Additionally, we may tune other parameters such as the minimum samples required to split a node, the minimum samples required at each leaf node, and the maximum features considered for splitting.

N-dimensional grid search is the means by which the model gets trained, which entails assessing all possible combinations of hyper-parameter values within the defined ranges and the randomized search is doing the process of hyper-parameter tuning, where the hyper-parameters are sampled randomly from a specified distribution. The goal of both approaches is to look for the most lucrative collections of a hyperparameters and prevent over-fitting due to overtraining on the part of using the training set.

In the end, hyperparameter tuning is done and then we take our model's performance for a test spin across a several evaluation metrics including but not limited to the ability of the model to distinguish between defaulted and non-defaulted loans.

In our demo of outcomes, we are reporting both key metrics and visualizations to insure that we cover the diversity in the performances of the machine learning models, particularly the Gradient Boosting Classifier and Random Forest Model[1], using the rate of default on loan loans as an example. Through accuracy scores, classifier report and confusion matrices, we ensure that each model is precise with each run. This fact, in its turn, provides the usual members with evidence of the quality of their forecast skills and how parameters influence their zenith.

- 6) *Accuracy scores:* we got a formidable precision of 98.03% applying the Gradient Boosting Classifier with 25 trees for a decision. This excellent precision was realized through model training implementation of diverse learning algorithms, accurate hyperparameter tuning which is a procedure of optimizing machine learning algorithms performance through refining the values of certain variables known as hyperparameters, and cross-validation to make sure that the model does not overfit or underfit.

Classification Report:

	precision	recall	f1-score	support
Approved	0.98	0.99	0.98	678
Rejected	0.98	0.97	0.97	390
accuracy			0.98	1068
macro avg	0.98	0.98	0.98	1068
weighted avg	0.98	0.98	0.98	1068

Confusion Matrix:
[[670 8]
[13 377]]

Fig.1. Accuracy for our project

- 7) *Classification reports:* For each model, we generate classification reports that include precision, recall, F1-Score and support values for each class (e.g., approved and rejected). These reports offer detailed insights into the model's performance metrics for individual classes, allowing us to assess its ability to correctly classify instances across different categories.

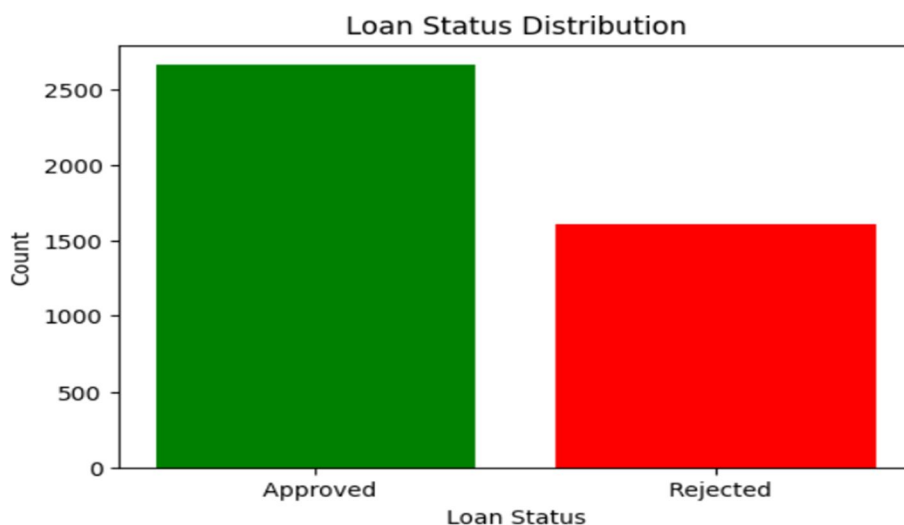


Fig.2. Classified the Loan Status

- 8) *Confusion matrix:* We visualize confusion matrices for each model, which provide a comprehensive summary of the model's performance by illustrating the number of true positive, true negative, false positive, and false negative predictions. These matrices offer a clear visualization of the model's ability to correctly classify instances and identify any patterns of misclassification.



Fig.3. Confusion matrix of our project

VI. MODEL DEPLOYMENT

The work on the API was vital for machine learning models integration into our loan prediction application. Thus, we got to implement machine learning models in a working application. Making use of Python Flask, a lightweight framework, we designed the API to handle incoming application data and provide appropriate prediction answers. At the commencement we established some concrete end points suited to various stages of predicted path. The example of this end point was that one-that collects posted loan applications as JSON objects-was able to process and then pass these to click on the machine learning model to enable it to predict the application.

Thereafter, after we run the predictions, we fuse our output with the client to enable him to retrieve the report. Viewing the user's credit score, financial commitments, employment history, monthly expenses, and income details as inputs to the 'prediction endpoint,' it returns a JSON response which contains the loan approval status as well as any relevant details. A development approach that focused on fault tolerance, input validation, and addition of monitoring and logging mechanisms among other things was used. Such measures proved to be the API's stone of reliability, security, and performance. This was the crown of the user interface which became the essence of our predictive modeling framework functioning smoothly.

VII. EXPERIMENTAL RESULT

The frontend interface design of our loan prediction system is meticulously crafted to provide users with a seamless and intuitive experience when submitting loan applications and receiving predictions. The user interaction flow is designed to guide users through the process step by step, ensuring clarity and ease of use. Here's an overview of the frontend interface design and user interaction flow

- 1) *Home Page:* The journey begins with a well-designed landing page that welcomes users and provides them with a brief overview of the loan prediction system's capabilities. The landing page may include enticing visuals, concise text descriptions, and clear calls to action prompting users to begin the loan application process.

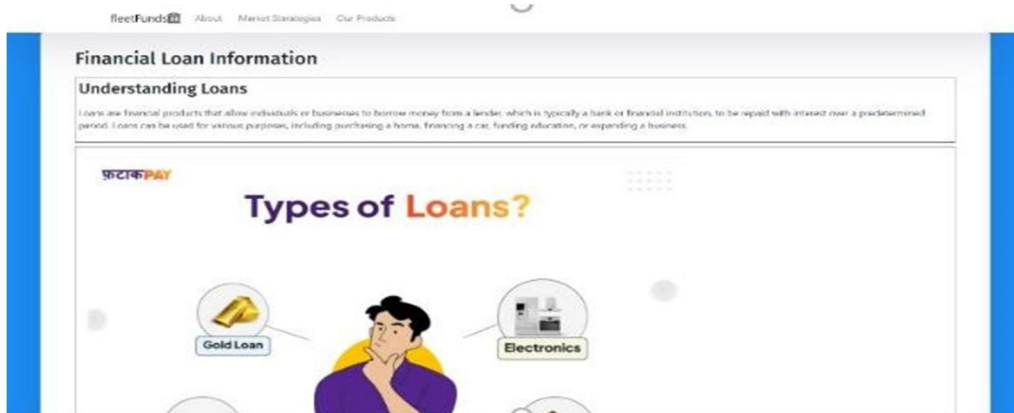


Fig.4. Home page of our web application

- 2) *Loan Application Form:* Upon initiating the application process, users are directed to a structured loan application form. The form is designed to capture essential information from users, including personal details, financial information, and any other relevant data necessary for accurate loan prediction. The form is presented in a clean and organized layout, making it easy for users to input their information.



Fig.5. Essential 11 records the user must fill.

- 3) *Input Validation and Guidance:* As users fill out the application form, real-time input validation mechanisms are employed to ensure that the entered data is accurate and complete. Clear error messages and prompts guide users to correct any invalid or missing information, helping to streamline the data entry process and minimize errors.
- 4) *Submission and Processing:* Once users have completed the application form, they can proceed to submit their loan application for processing. Behind the scenes, the frontend interface communicates with the backend system to transmit the application data securely. Users are provided with visual feedback to indicate that their application has been successfully submitted and is being processed.
- 5) *Prediction Results Display:* Upon completion of the application form, users can submit their loan application for processing. Behind the scenes, the frontend interface securely transmits the application data to the backend system. Users receive visual feedback indicating that their application has been successfully submitted and is now being processed. Following submission, a toast message appears in the top right corner of the screen, providing users with instant feedback on whether their loan is predicted to be approved or not. This toast message serves as a quick and convenient way for users to receive the outcome of their application without navigating away from the page, enhancing the user experience and streamlining the process.

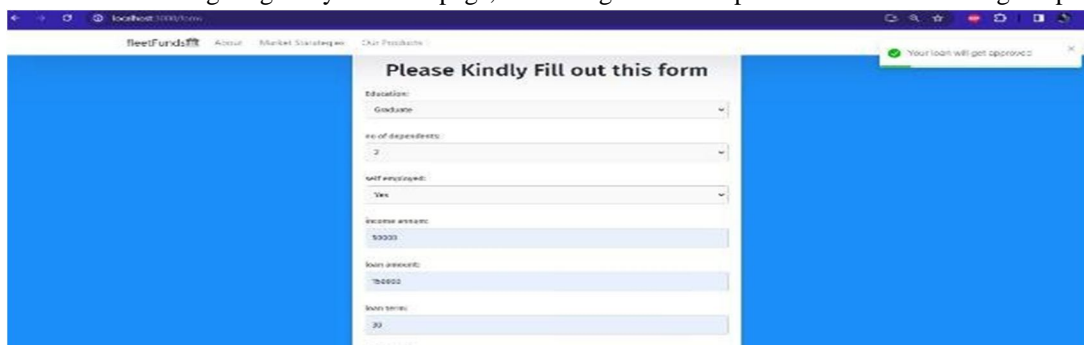


Fig.6. After Submission The toast message will be displayed on top right corner in desktop version

VIII. CONCLUSION AND FUTURE SCOPE

To sum up, smart and experienced people in banks have invented this automatic loan approval tool using machine learning to improve the progress of the financial industry. In addition, by implementing a rigorous data analysis, model training and evaluation processes, we have proven that the machine learning algorithms have highly accurate results in predicting loan district outcomes. Our system gives financial institutions an automated and intelligence tool for making more accurate loan assessment decisions which reduces biases and ensures efficient loan approval process in the end the clients are satisfied. We took advantage of advanced technologies and methods in order to tackle the key issues of the traditional credit approval procedures and making the way for technological financial products of the future.

In future work, these aspects can be improved:

Expanding the Dataset: Our present day database contains over 4000 records, but there is a possibility that it can be extended by including a greater magnitude and more varied set of subjects. Analyzing the data pool with a wider range of situations would lead to a more general picture where borrowers who conform to a certain scheme are often more successful. The evaluation may thus be more useful and may suggest factors that could be influential for the outcome. Providing that it allows for a comprehensive representation of both borrower classes and different loan features, the larger the dataset, the deeper the view on our predictive models and the more precise albeit still reliable they will be.

Improving Loan Approval: Our actions going forward will be mainly centred on providing advice to customers on their way of increasing an amount of their loan. Our aim is to research on the ways we can grow the credit scores toward CIN scores as well as the other factors will be stated. The essence of financial coaching lies in guidance on financial management and loan eligibility. Hence, we want to enable our clients so that their likelihood for loan approvals are enhanced.

REFERENCES

- [1] Li, P.S. Saini, A. Bhatnagar and L. Rani, "Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), India, 2023, doi: 10.1109/ICACITE57410.2023.10182799.
- [2] R.Priscilla, T. Siva, M. Karthi, K. Vijayakumar and R. Gangadharan, "Baseline Modeling for Early Prediction of Loan Approval System," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), India, 2023, doi: 10.1109/ICECONF57129.2023.10083650.
- [3] A.Soni and K. C. P. Shankar, "Bank Loan Default Prediction Using Ensemble Machine Learning Algorithm," 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), India, 2022, doi: 10.1109/ICPS55917.2022.00039.
- [4] M.A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), India, 2020, doi: 10.1109/ICESC48915.2020.9155614.
- [5] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood "Random Forests and Decision Trees," vol:9, 2012.
- [6] ManjeetKumar, Vishesh Goel, Tarun Jain, Sahil Singhal, DR. Lalit Mohan Goel. (2018). Neural Network Approach To Loan Default Prediction,
- [7] International Research Journal of Engineering and Technology (IRJET) , p-ISSN: 2395-0072
- [8] Viswanatha V,Ramachandra A.C,Vishwas K N,Adithya G,"Prediction of Loan Approval in Banks Using Machine Learning Approach,"International Journal of Engineering and Management Research, Vol-13, 2023 "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
- [9] S.Öner, D. Alnahas, A. Kanturvardar, A. M. Ülkgün and C. Demiroğlu, "Comparative Study of Credit Risk Evaluation for Unbalanced Datasets Using Deep Learning Classifiers," 2023 31st Signal Processing and Communications Applications Conference (SIU), Turkiye, 2023, doi: 10.1109/SIU59756.2023.10224008.
- [10] Adele Cutler, David Richard Cutler, and John R Stevens"Random Forests," doi: 10.1007/978-1-4419-9326-7_5, vol: 45, 2011.Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.
- [11] Q.. Hegde N, Deepa, C. Shetty, R. N, D. B and Prathyakshini, "Predictive Analysis of Loan Data using Machine Learning," 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE),India, 2022, pp. 272-276, doi: 10.1109/AIDE57180.2022.10060781.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)