JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REVIEW PAPER ON LOAN PREDICTION SYSTEM USING MACHINE LEARNING

[1]Prof.S. R. Kale,

[2]Sarang R. Pawar, [3]Tushar M. Behare, [4]Sakshi H. Hingwe,

[5]Shivraj A. Khuje,

[1]Assistant Professor, Department of Computer Science and Engineering, P. R. Pote(Patil) Collage of Engineering Amravati,
[2]Department of Computer Science and Engineering, P. R. Pote(Patil) Collage of Engineering Amravati,
[3]Department of Computer Science and Engineering , P. R. Pote(Patil) Collage of Engineering Amravati
[4]Department of Computer Science and Engineering , P. R. Pote(Patil) Collage of Engineering Amravati,
[5]Department of Computer Science and Engineering , P. R. Pote(Patil) Collage of Engineering Amravati

Abstract: This review paper explores a machine learning-based loan prediction system, emphasizing the challenges in manually assessing loan eligibility and the benefits of automating the process. Comparing various algorithms, including Decision Tree, Random Forest, and Regression, the study highlights Random Forest's superior accuracy in handling extensive datasets. The architecture involves data input, pre-processing, classifier finalization, model training, and application to test data. Additionally, the paper introduces Naive Bayes as an efficient classification algorithm. The literature review examines related works, and the conclusion underscores the significance of Naive Bayes in loan classification, suggesting avenues for future enhancements. Overall, the paper offers valuable insights into advancing loan prediction systems.

Keywords: Loan, Machine Learning, Random Forest and Classification, Naïve Bayes.

I. INTRODUCTION

In the contemporary financial landscape, a surge in loan applications has created a daunting task for banking officials to discern deserving candidates for approvals. The traditional approach, often prone to biases, has paved the way for the imperative integration of automation into the loan approval system. The burgeoning population has further exacerbated the need for efficient handling of vast applicant information, necessitating the application of data science and machine learning techniques.

Addressing this challenge involves employing classification algorithms to sift through extensive datasets according to predefined bank criteria. Among these algorithms, the Random Forest algorithm emerges as a robust contender, demonstrating superior efficiency in the classification of loan applicants based on both bank criteria and the information provided by applicants.

The fundamental steps of this classification algorithm involve the meticulous selection of the dataset for classification, followed by a comprehensive pre-processing of the data. The training of the dataset is then carried out utilizing the Random Forest algorithm, a powerful ensemble learning technique. The culmination of this process is the application of the trained model to a testing dataset, empowering banks to make informed decisions based on the output generated.

The primary objective of this research is to leverage data classification techniques for the analysis of training data, ultimately facilitating decision-making regarding loan approval. A pivotal focus lies in predicting whether a loan can be sanctioned or not, crucial for optimizing the efficiency of the loan approval system. To determine the best-suited

classification model, a comparative analysis was conducted, evaluating algorithms such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests. Significantly, the Random Forest algorithm exhibited the highest accuracy among the contenders.

This automation initiative not only expedites the identification of deserving applicants but also serves as a streamlined solution for maintaining cash flow within the banking system and reducing Non-Performing Assets (NPAs). By amalgamating cutting-edge technology with financial decision-making, this approach offers a swift, efficient, and equitable means of assessing loan eligibility, thereby fostering a more resilient and responsive banking ecosystem.

II. LITERATURE REVIEW

In recent years, substantial research efforts have been directed towards developing and refining classification algorithms for the effective analysis of large datasets, particularly in the context of loan approval systems. Several noteworthy works in this domain have been consulted for the development of the proposed automation system.

An Approach for Prediction of Loan Approval Using Machine Learning Algorithm:

Authors: Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar.

The study emphasizes the pivotal role loans play in a bank's profit or loss, emphasizing the importance of predicting loan defaulters to reduce Non-Performing Assets (NPAs) and maintain Capital to Risk-Weighted Assets Ratio (CRAR). Logistic regression is explored as a predictive analytics approach for studying and comparing different algorithms.

Loan Default Forecasting using Data Mining:

Authors: Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal.

This work underscores the significance of constructing a model for assessing the likelihood of default on a debt. Data mining algorithms are employed to predict potential defaulters from a dataset containing information about home loan applications, aiding banks in making informed decisions and mitigating losses.

Prediction Defaults for Networked-guarantee Loans:

Authors: Dawei Cheng, Zhibin Niut, Yi Tu, Liqing Zhang.

Focusing on networked guarantee loans and associated systemic risks, the authors propose an imbalanced network risk diffusion model. The positive weighted k-nearest neighbors (pwKNN) algorithm is introduced, demonstrating superior performance in predicting enterprise default risks compared to conventional credit risk methods.

Overdue Prediction of Bank Loans Based on LSTM-SVM, Random Forest:

Authors: Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, Huakang Li.

Addressing the accuracy limitations of traditional loan risk prediction models, this study employs a combination of LSTM and SVM algorithms for dynamic and static analysis of user information, respectively. The LSTM-SVM model yields enhanced efficiency compared to traditional algorithms in predicting user delinquency.

Credit Data Prediction Using Min-Max Normalization and K Nearest Neighbor (K-NN) Classifier.

This paper advocates for a more accurate prophetic modeling system in the banking industry, specifically focusing on predicting credit defaulters. The proposed model integrates the Min-Max normalization and K-NN classifier for credit scoring, demonstrating its effectiveness in predicting loan status with high sensitivity.

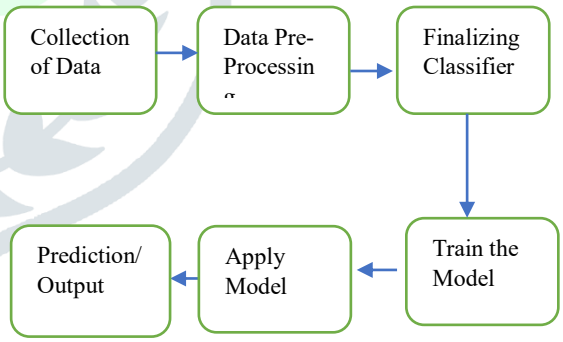
III. ARCHITECTURE DIAGRAM AND PROCESS

Input Data:

Download the train and test dataset consisting of the loan applicant's information such as application ID, Name, Loan ID Gender, Married, Dependents Education, Self-employed, Application loan amount term, Co-applicant income, Loan Amount, and Credit History.

Pre-Processing:

The CSV file contains some null values and irrelevant information that needs to be cleansed. The iloc() and shape() functions can be used to remove the null values. Noise cancellation and data cleansing are done in this step.



Architecture Diagram

Finalize Classifier:

Comparing classification algorithms involves evaluating various metrics such as Application id, Name, Loan ID Gender, Married, and Dependents Education Each algorithm has its strengths and weaknesses, and the choice depends on the specific characteristics of the dataset and the problem at hand.

After a thorough analysis, the Random Forest algorithm emerges as a suitable choice due to its robust performance across a wide range of scenarios.

Train Model:

Train the dataset using finalized classification algorithms.

Apply Model:

Apply this built model to test the dataset.

Output:

Classify the applicants based on the applicant's information and the bank's criteria using the Random Forest classifier, Naive Bayes classification.

IV. Naive Bayes

Naive Bayes is a supervised learning algorithm commonly used for classification tasks. It is based on Bayes' theorem and assumes independence between features, making it particularly effective for high-dimensional datasets. Naive Bayes classifiers are simple yet powerful, and they perform well in various scenarios.

Naive Bayes, a classification algorithm, assumes feature independence given the class label. Despite potential real-world deviations from this assumption, the algorithm often performs well. It calculates probabilities for features within each class based on the training dataset, estimating the likelihood of feature occurrence given a class. The algorithm determines prior class probabilities and utilizes Bayes' theorem to compute posterior probabilities based on observed features. During the classification of a new data point, it calculates these posterior probabilities and assigns the class with the highest probability as the predicted class. Naive Bayes excels in simplicity, efficiency, and handling of missing data, making it advantageous for various applications, particularly in text classification and spam filtering.

Algorithm Steps:

1. Data Preparation Organize the dataset with features and corresponding class labels.

2. Calculate Class Priors Estimate: the prior probability of each class based on the training data.

3. Calculate Feature Probabilities: For each feature and class, calculate the likelihood of observing that feature given the class.

4. Make Classification Decision: For a new data point, calculate the posterior probabilities and assign the class with the highest probability as the predicted class.

V. CONCLUSION

In the dynamic realms of data science and artificial intelligence, the Naive Bayes algorithm has emerged as the optimal choice for classifying loan aspirants, following rigorous research. This paper primarily concentrates on assessing loan aspirants, with the system predominantly utilizing Naive Bayes, showcasing efficient operations. Acknowledging its adaptability, there's a call for future improvements to enhance dependability, security, and accuracy. The system's current dataset training ensures its continued relevance. In addressing potential challenges like computer glitches, the paper advocates ongoing efforts to fortify software security. While the focus is on Naive Bayes, integrating additional algorithms holds promise for heightened predictive capabilities in the evolving technological landscape.

VI. REFERENCES

[1]. Yadav, O. P., Soni, C., Kandakatla, S. K., Sswanth, S. (2019). Loan prediction using decision tree. International Journal of Information and Computer Science, 6(5).

[2]. Arutjothi, G., Dr. Senthamarai, C. (2017). Comparison of feature selection methods for credit risk assessment. International Journal of Computer Science, 5(I), No 5.

[3]. Aida Krichene," Using a naive Bayesian classifier methodology for loan risk assessment," Journal of Economics, Finance and Administrative Science, 2017.

[4]. Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash, "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2019.

[5] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[7] A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

[8] G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".

[9] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".