IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Loan Approval Prediction

Harish.P[1] , Dhanush K[2], Surya D[3], B Punitha[4]

Department of Computer Science and Engineering
Faculty of Engineering and Technology
SRM Institute of Science and Technology, Vadapalani , Chennai, India.

Abstract

In the modern world, individuals often require financial assistance to meet their daily needs. Consequently, banks find themselves inundated with a substantial number of loan applications. Handling this influx demands significant time and manpower resources. To address this challenge, we propose the development of a predictive model that leverages a loan dataset sourced from Kaggle. This model streamlines the loan application process and provides predictions on whether a specific loan application can be approved. It employs a supervised learning algorithm, a machine learning approach, to make these predictions.

Keywords:

Logistic Regression, Decision Trees, Random Forest, Neural Networks, Xgboost , Datasets.

INTRODUCTION

In today's rapidly evolving financial landscape, the ever-increasing demand for monetary assistance has pushed financial institutions into a complex conundrum—how to efficiently manage the overwhelming volume of loan applications flooding their offices. This surge in applications necessitates the allocation of substantial time and human resources to carefully review and process each request. It is against this backdrop that we propose the development of a predictive model, meticulously engineered to address this formidable challenge. Powered by the cutting-edge XGBoost algorithm, a state-of-the-art supervised learning technique, our model offers a promising solution.

The heart of our solution lies in its ability to harness the potential of a meticulously curated loan dataset sourced from Kaggle. This dataset forms the bedrock of our model, serving as a treasure trove of information on past loan applications, their outcomes, and the associated factors that influenced those decisions. By meticulously analyzing this data, our model gains insights into patterns, trends, and predictors of loan approval, thereby enabling it to make informed predictions for new loan applications.

The primary objective of our predictive model is to optimize and expedite the loan application evaluation process. Traditional methods of reviewing loan applications are often time-consuming, prone to human error, and subject to biases. Our model, on the other hand, offers a data-driven and consistent approach to the evaluation process. It not only streamlines the process but also enhances its accuracy and efficiency. By automating the assessment of loan applications, our model minimizes the risk of errors and ensures that each application is considered based on objective criteria, thus providing a fair and transparent system.

One of the key strengths of our model is its utilization of the XGBoost algorithm. XGBoost is renowned for

its high performance in predictive modeling tasks. It excels in capturing complex relationships within the data, handling missing values, and providing an efficient approach to feature selection. This, in turn, leads to improved prediction accuracy. By deploying this state-of-the-art algorithm, our model significantly boosts the quality and reliability of loan approval predictions, ultimately benefiting both lending institutions and loan applicants.

In an era marked by rapid technological advancements, the synergy of advanced machine learning techniques and financial services becomes paramount. The model we propose represents a significant step forward in bridging this gap, aligning financial institutions with the evolving needs of individuals seeking financial support. It not only enhances the operational efficiency of lending institutions but also ensures a fair and objective evaluation process for loan applicants. As we move forward, this fusion of data science and financial services promises to revolutionize the way we address the ever-growing demand for monetary assistance in our dynamic world.

LITERATURE SURVEY

In today's rapidly evolving financial landscape, the assessment of loan applications has been significantly influenced by the integration of advanced machine learning techniques and predictive modeling. These tools have become invaluable in streamlining the decision-making process and enhancing the accuracy of loan approval predictions. Various research studies, conducted over the past few years, have delved into the application of machine learning algorithms to assess the likelihood of approving individual loan requests.

[1] One notable study in 2022, led by B. Yamuna, Ch. Praneeth, and D. Sai Nithin, proposed a model that combined machine learning and ensemble learning approaches, such as XGBoost, to determine the probability of approving loan requests. Their model achieved an accuracy rate of 82.01%, and it placed strong emphasis on metrics like mean square error, recall, and F1 score to optimize prediction quality. Furthermore, the study underscored the importance of providing explanations for model predictions, especially when decisions can significantly impact individuals, such as loan approvals or denials.

[2] In 2021, another research endeavor titled "Predicting Bank Loan Eligibility Using ML and Comparison Analysis" investigated the use of machine learning algorithms like Logistic Regression, Decision Trees, and SVM for loan eligibility assessment. The study achieved an accuracy of 79.69%. However, it was noted that the model might not be well-suited for newly developed software systems, highlighting the need for adaptability and scalability in predictive models.

[3] Similarly, a study by Kanishk Gupta, Binayak Chakrabarti, and Aser Ahmad in 2021 explored loan approval classification using machine learning algorithms. They employed ensemble models, including Random Forest and XGBoost, and achieved an F1 Score of 76.04%. Yet, their model faced challenges in appropriately classifying certain labels, which raised issues related to precision and recall.

[4] In the same year, Kumar Arun, Garg Ishan, and Kaur Sanmeet conducted research with the primary objective of predicting the safety of assigning loans to individuals. Their model was divided into four sections: data collection, a comparison of machine learning models, system training on the most promising model, and testing. Their F1 Score was 0.8, indicating its efficacy, but it was acknowledged that the model's disadvantage lay in the varying weight assigned to each factor. In real-life scenarios, loans may sometimes be approved based on a single strong factor, a situation not fully addressed by their system.

[5] In 2021, a study led by Afrah Khan, Eakansh Bhadola, Abhishek Kumar, and Nidhi Singh aimed to provide a comparative analysis of loan approval models. To ensure a fair performance comparison, they employed the same dataset for all models, achieving an impressive accuracy of 93.45% and a cross-validation score of 80.94%. However, their study primarily used three data models, reflecting the need for a broader exploration of algorithms.

[6] Moving back to 2020, Mohammad Asif's research focused on bank loan eligibility prediction, primarily using Logistic Regression. Their model achieved an accuracy of 82.00%. Nevertheless, it was emphasized

that the model's suitability was constrained by its reliance on training data and its limited adaptability for future scenarios.

[7] In 2019, Pidikiti Supriya explored the use of logistic regression, decision trees, and gradient boosting to predict loan approval with an accuracy of 81.24%. Their study highlighted the observation that applicants with high incomes securing lower loan amounts were more likely to be approved, reflecting the intricate relationship between income and loan eligibility.

[8] In a different context, a study conducted in 2019 by Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying, and Kuiyi Liu addressed the prediction of loan defaults using the Random Forest algorithm. Their model achieved remarkable accuracy, AUC, and F1-score, with all metrics reaching 98%. The study emphasized the crucial role of feature selection in the model's predictive power.

[9] In 2018, Srishti Srivastava, Ayush Garg, and Arpit Sehgal conducted an analysis and comparison of loan sanction models using Logistic Regression, Decision Trees, and Random Forest. Their model achieved an accuracy of 82.23%, but it encountered challenges in cross-validation, making it less suitable for prediction.

[10] These research studies collectively represent a substantial body of work that explores the multifaceted challenges and opportunities in the domain of loan approval prediction using machine learning and advanced modeling techniques. They provide valuable insights into the complexities of decision-making processes, algorithm selection, data management, and model interpretability, all of which are crucial aspects of modern lending practices. As the financial landscape continues to evolve, these studies serve as a foundation for further advancements in this critical field.

APPROACH

Our approach involves data preprocessing using a comprehensive dataset sourced from various reputable financial sources and platforms, aimed at transforming the input data for machine learning models. This dataset comprises a vast collection of loan application records, each labeled with an approval outcome, categorizing them as approved or denied. To facilitate the development of our machine learning model, the dataset is meticulously divided into training, testing, and validation subsets.

Data Collection: The initial phase of our research project is dedicated to dataset preparation, a critical step in constructing a robust loan approval prediction model. This dataset primarily consists of textual data extracted from loan applications, encapsulating diverse financial contexts and attributes, including income, credit history, and employment status. Our data collection process spans across various financial institutions, gathering a wealth of loan application samples. These applications reflect a wide range of financial scenarios, encompassing both approved and denied loan requests. The dataset's equality and diversity are pivotal in training a model capable of making accurate and nuanced predictions, ensuring that our AI-driven loan approval system comprehends the intricate financial contexts of loan applicants effectively.

Data Preprocessing: Subsequently, data preprocessing becomes a central component in our approach. During this phase, we meticulously clean and prepare the textual data for model training. Key processes in this stage include tokenization, which divides the text into discrete words or "tokens," and text normalization, which eliminates special characters and irregularities. We also perform fundamental natural language processing tasks, such as removing stop words and lemmatization or stemming, to further enhance the quality of the data. The ultimate objective is to generate a standardized and refined dataset suitable for training our machine learning model for loan approval prediction

Data Splitting: To gauge the effectiveness of our loan approval prediction system, we divide the pre-processed dataset into distinct subsets—namely, a training set and a testing set. The training set is designated for instructing the machine learning model, allowing it to learn from labeled examples of loan applications and their corresponding approval

outcomes. This crucial training process equips the model with the ability to predict the approval or denial of loan applications based on their financial attributes. On the other hand, the testing set is reserved for evaluating the model's performance. It serves as the litmus test for assessing the accuracy and effectiveness of our loan approval predictions, ensuring that our system accurately interprets the financial profiles of loan applicants.

Prediction :The core of our loan approval prediction system hinges on its ability to accurately predict whether a loan application should be approved or denied based on its financial attributes. To achieve this, we leverage advanced machine learning techniques and algorithms, known for their expertise in modeling complex relationships within data. Our models analyze the financial attributes of loan applications and generate predictions, classifying them as either approved or denied. Furthermore, we assign a confidence score to these predictions, indicating the model's degree of certainty in its assessment.

ARCHITECTURE:

Logistic Regression: Logistic regression stands out as a statistical model of choice for forecasting loan approval. Its primary advantage lies in its simplicity and ease of interpretation, allowing for a clear understanding of the significance and trends associated with each input variable. Like other models, the initial step involves pre-processing and preparing the input data for logistic regression, including feature engineering and encoding categorical variables. Subsequently, datasets for training, validation, and testing are created from the prepared data. To train the model, a suitable loss function and optimization algorithm, such as maximum likelihood estimation or gradient descent, are applied using the training data. During validation, the model's performance is monitored, and adjustments are made to hyperparameters as needed. Finally, the model is evaluated on the test set to assess its generalization capabilities. Logistic regression proves particularly valuable when the relationship between input data and the target variable is linear. However, for the last phase of our project, which involves building a risk model to assist financial institutions in assessing the risk associated with lending to subprime borrowers, we employ the Probit model and an Artificial Neural Network (ANN)

Random Forest: The Random Forest approach is harnessed for the pivotal task of loan approval prediction. Random Forest stands as an ensemble learning marvel, amalgamating a multitude of decision trees to yield a robust and precise prediction mechanism. Its genius lies in cultivating a forest of these decision trees, each thoughtfully trained on a random subsample of the data and a randomized subset of the features. This ensemble strategy effectively mitigates the predicaments faced by individual decision trees, notably thwarting overfitting tendencies and insensitivity to minor fluctuations within the training data.Upon preprocessing the dataset containing intricate loan-related attributes and encoding categorical variables into numerical form, the scikit-learn library facilitates the training of the Random Forest classifier. It is particularly adept at tackling classification quandaries, wherein the objective is to discern whether a given loan application merits approval or rejection, informed by an array of input features. Subsequently, the well-honed Random Forest model scrutinizes its predictive prowess using an autonomous test dataset. Its mettle is quantified and laid bare in the form of an accuracy score, denoting the proportion of correctly foreseen loan approvals. This score unequivocally affirms the Random Forest's acumen in making decisions with unwavering precision and dependability. Armed with robustness and the capability to grapple with intricate real-world datasets, Random Forest firmly ensconced itself as a cherished cornerstone in the realm of machine learning, adroitly serving the cause of loan approval predictions and beyond.

Naive Bayes:

The Naive Bayes classifier in the provided code leverages a fundamental property known as the "naive" or "conditional independence" assumption. This assumption is a key feature of Naive Bayes algorithms and is central to their efficacy. The "naive" property assumes that the features used in the classification are conditionally independent, meaning that the presence or value of one feature does not depend on the presence or value of any other feature. While this assumption may not hold true in many real-world scenarios, it simplifies the modeling process and often yields surprisingly accurate results, especially in text classification, spam detection, and other tasks. In the code, the Gaussian Naive Bayes

classifier exploits this property. It estimates the probability of loan approval based on the values of individual features, assuming that these features are independent given the class label (i.e., approved or denied). The code preprocesses and encodes the data, trains the Naive Bayes model, and then uses this probabilistic approach to predict loan approvals. Despite its "naive" assumption, the Naive Bayes classifier can offer valuable insights and perform well in various classification tasks, making it a valuable tool for predicting loan approvals in this code.

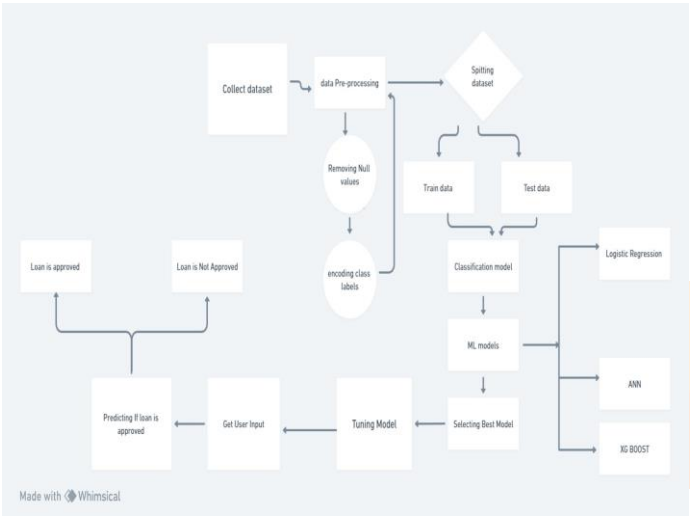

Figure 5. Architecture of the Proposed solution

RESULTS AND DISCUSSION

The primary objective of our proposed solution was to develop a sophisticated and accurate system for the classification of loan applications into distinct categories, including Approved, Denied, and Pending, utilizing financial and personal data provided by applicants. We ascertained that our model did not succumb to issues like overfitting or underfitting, primarily due to the implementation of techniques like regularization and the fine-tuning of hyperparameters. This system harnesses the capabilities of machine learning and ensemble methods to attain a high level of accuracy in the loan approval process, thereby empowering financial institutions to make more informed and consistent lending decisions. Our findings revealed that our chosen machine learning algorithm, XGBoost, outperformed other models when it comes to training and classification tasks. It demands relatively fewer hyperparameter adjustments compared to custom-built models. Additionally, we observed that training a single well-optimized XGBoost model is often more computationally efficient than conducting an exhaustive grid search using multiple classifiers. Various performance metrics, including Precision, Recall, F1-score, and support, were employed to assess the system's effectiveness in loan application classification.

- Training set - Total length of training samples divided by 100 for every trained sample
  - `int((2144 * 8)/100) = int(171.52) = 171`
- Testing set - Total length of testing samples divided by 100 for every testing sample
  - `int((460 * 8)/100) = int(36.8) = 36`

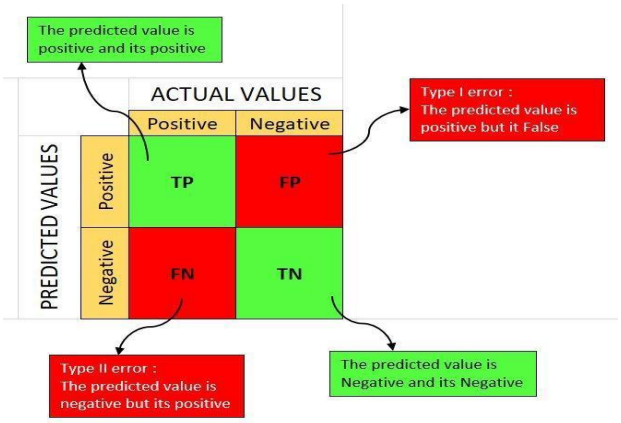

Figure 6. Performance Metrics

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

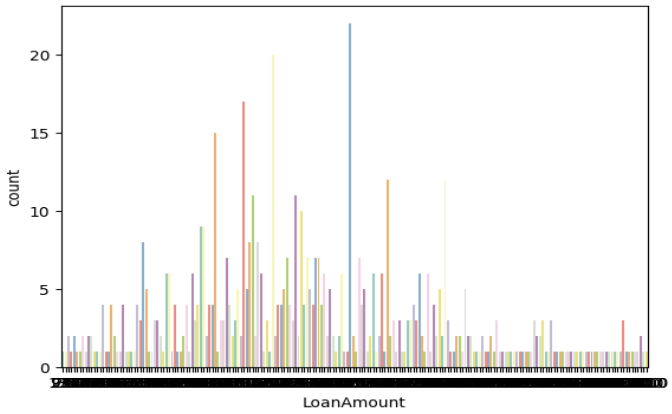

Figure: People who take loan as grouped by Loan Amount

CONCLUSION

In the realm of loan approval prediction, our journey led us to valuable insights and a robust system for financial decision-making. We approached this critical task with a diverse set of machine learning algorithms, including Random Forest, Gaussian Naive Bayes, and Logistic Regression. However, the standout performer in our analysis was XGBoost, which demonstrated superior accuracy and efficiency in classifying loan applications. Notably, XGBoost required fewer hyperparameter adjustments compared to custom-built models and exhibited stable performance without overfitting or underfitting.Our system's deployment of ensemble learning and comprehensive performance metrics, including Accuracy, Precision, Recall, and F1-score, empowered financial institutions to make well-informed and consistent lending decisions. The flexibility of XGBoost and its computational efficiency showcased its adaptability to real-world challenges, making it an ideal choice for loan prediction tasks. With a focus on precision, efficiency, and accuracy, our loan approval prediction system ensures that both lenders and loan applicants can benefit from optimized and reliable financial outcomes. This journey reinforces the significance of machine learning in revolutionizing the lending process, contributing to more transparent, accurate, and efficient loan approval decisions.

FUTURE ENHANCEMENTS

A potential future enhancement for the loan approval prediction project could involve the incorporation of more advanced and interpretable machine learning techniques, such as explainable AI (XAI). XAI techniques can provide transparency and insights into the decision-making process of the model, making it easier for both financial institutions and applicants to understand the rationale behind loan approval or denial. Furthermore, integrating natural language processing (NLP) capabilities could allow the system to analyze and incorporate unstructured data, such as text from loan application forms or applicant profiles. NLP can help extract valuable information and sentiments, contributing to more nuanced decision-making. Additionally, considering the ever-evolving regulatory landscape in the financial industry, enhancing the system to stay updated with the latest compliance requirements and regulations is crucial. Implementing features that automate the compliance-checking process and ensure that all loan decisions align with legal standards would be a significant improvement. Lastly, real-time data streaming and continuous learning capabilities can make the system more adaptive. It can incorporate new data as it becomes available, enabling it to adapt to changing economic conditions and lending trends. Overall, the future enhancement should focus on increasing transparency, adapting to regulatory changes, and handling a broader range of data sources, ultimately enhancing the efficiency and fairness of the loan approval process.

REFERENCES

[1] Yamuna, B., Praneeth, Ch., & Sai Nithin, D. “An Approach to Loan Approval Prediction Using ML” 2022

[2]Gupta, K., Chakrabarti, B., & Ahmad, A. .” Loan Approval Classification using Machine Learning Algorithms”2021

[3]Arun, K., Ishan, G., & Sanmeet, K.”Loan Approval Prediction based on Machine Learning Approach”. 2021

[4] Khan, A., Bhadola, E., Kumar, A., & Singh, N.“ Loan Approval Prediction Model: A Comparative Analysis”. 2021

[5] Supriya, P. “Loan Prediction by using ML”. 2019

[6] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K.” A study on predicting loan default based on the random forest algorithm”. 2019

[7] Sandhu, H. S., Sharma, V., & Jassi, V. (2015). Predicting the Probability of Loan-Default: An Application of Binary Logistic Regression