International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 11 | Nov 2024 www.irjet.net p-ISSN: 2395-0072

A Survey on Loan Default Prediction using Machine Learning Techniques

Adwait Mandge[1], Rohan Fatehchandka[2], Kunal Goudani[3], Tanaya Shelke[4], Prof. Pramila M. Chawan[5]

[1]1B. Tech Student, Dept of Computer Engineering, and IT, VJTI College, Mumbai, Maharashtra, India
[2]1B. Tech Student, Dept of Computer Engineering, and IT, VJTI College, Mumbai, Maharashtra, India
[3]1B. Tech Student, Dept of Computer Engineering, and IT, VJTI College, Mumbai, Maharashtra, India
[4]1B. Tech Student, Dept of Computer Engineering, and IT, VJTI College, Mumbai, Maharashtra, India
[5]Associate Professor, Dept of Computer Engineering, and IT, VJTI College, Mumbai, Maharashtra, India

---

Abstract - *Loan default prediction is a critical challenge in the banking sector, where inaccurate assessments can lead to significant financial losses. Recent advances in machine learning, particularly in ensemble methods and deep learning, present valuable opportunities for automating and improving loan default predictions. This paper presents a novel Weighted Voting Ensemble approach by training multiple models—Random Forest, XGBoost, and Neural Networks—and assigning weights to their predictions based on performance metrics. Neural networks capture complex, high-dimensional patterns, while traditional models like Random Forest and XGBoost handle simpler but crucial features. This hybrid method enhances prediction robustness, optimizing loan default predictions by leveraging the strengths of each model. Additionally, data imbalance is addressed using SMOTE to improve model performance.*

Key Words: Machine Learning, Deep Learning, Ensemble, Loan Default Prediction

1.INTRODUCTION

In today's financial landscape, loan default is a significant risk that banks and lending institutions constantly face. The challenge of ensuring loan repayment while minimizing default rates is crucial for maintaining profitability and operational stability. Traditionally, credit assessments were performed manually, requiring extensive human intervention and time, which often resulted in inefficiencies and inaccuracies. As the volume of loan applications increased, it became evident that manual methods could no longer meet the needs of modern financial systems, particularly in dealing with large datasets.

The prevalence of Machine Learning has made it easier for banks to access advanced technologies that can automate and improve credit risk management. Machine learning models are capable of analyzing data in bulk and identify and recognize structures that are perplexing for humans to detect through manual analysis. By leveraging these models, financial institutions can not only predict loan defaults with greater precision but also significantly reduce the resources required for credit assessments.

This study aims to examine the potency of several algorithms in predicting loan defaults. Using a dataset sourced from Kaggle, we compare the performance of individual and ensemble models, with a particular focus on addressing the challenge of data imbalance. The objective is to determine which model provides the highest accuracy and reliability for lenders, thereby helping in making informed decisions and loss prevention.

2. LITERATURE REVIEW

2.1 Machine Learning Algorithms

Logistic Regression calculates the probability that a given input is a member of a specific class. The output is a score in the range [0,1] which is then threshold-based to produce a binary forecast (yes/no, spam/not spam, etc.).

Naive Bayes is a classification algorithm that supposes that attributes are conditionally independent of the given class label provided,derived from Bayes' Theorem and works well even when this assumption is violated in practice, especially with high-dimensional data. The advantages of this algorithm are its speed, simplicity, and effectiveness, particularly for text classification and large datasets.

Decision Tree is a rule based technique used for classification and regression of data points.. The data points are divided based on feature values,each node represents a test on attributes, and each leaf node represents a final outcome. The algorithm offers ease of interpretation, supports both categorical and numerical data, and does not require data scaling. It is capable of incorporating missing values and provides seamless visualization, providing insights into feature importance.

Random Forest learning technique is used to solve various problems of regression and classification problems. It builds several trees during training and aggregates the outcomes to increase prediction accuracy.

Features are randomly selected at every node(feature bagging), and each tree in the forest is trained on an arbitrary subclass of the data (bootstrapping). Random Forest reduces variation and overfitting and averages the outcomes of these trees, which enhances generalization. The advantages of this algorithm are handling large datasets, reducing overfitting, and maintaining good accuracy even in the presence of missing data.

XGBoost (Extreme Gradient Boosting) is a scalable and extremely effective gradient boosting implementation. It constructs decision trees sequentially, aiming to fix the mistakes caused by preceding trees with each new tree. XGBoost uses gradient descent to minimize the overall loss function in order to optimize the model. Performance and speed are well-known for XGBoost, particularly with tabular or structured data.The advantages of this algorithm are its speed, scalability, handling of missing data, and regularization to minimize overfitting.

AdaBoost (Adaptive Boosting) is an ensemble learning method that combines weak classifiers, typically decision trees, into a stronger classifier. It ensures that the subsequent classifier focuses more on fixing the errors of the preceding one by assigning larger weights to misclassified examples and smaller weights to correctly classified ones. AdaBoost continues this process until the overall error is minimized.The advantages of this algorithm are its ability to improve weak classifiers and adaptability to challenging data.

Bagging (Random Forest) is an ensemble technique where multiple models are trained on arbitrary subclasses of the data. For classification, majority voting is used for determining the final output and that of regression are averaged. Random Forest is a well-known bagging method. The advantages of this algorithm are the reduction of variance and overfitting, along with improved model robustness. Moreover, bagging enhances predictive accuracy by allowing each model to correct the errors of others. It also helps in minimizing the influence of noise and outliers present in the dataset. By training on completely randomized shares of data, it introduces diversity among models, making the final model highly resistant to overfitting. This technique is particularly useful when discrepancy is observed in large amounts in the base models, as it effectively stabilizes predictions, providing a more reliable and generalized performance across unseen data.

Boosting is an ensemble learning technique where models are built sequentially to correct the errors of the preceding model. By adjusting the weights of weak learners, boosting helps improve their accuracy. Weak learners are often decision trees.The advantages of this algorithm are its ability to reduce bias.

Stacking (LR, SVM, NN) uses many base models (e.g., SVM, Neural Networks, and Logistic Regression) to create a meta-model that synthesizes their predictions. Each base model is trained independently, and the meta-learner combines their outputs to make a final prediction.The advantages of this algorithm are its ability to reduce overfitting and improve the accuracy by using various models.

2.2 Deep Learning Algorithms

Neural Networks replicate the composition of the human brain. Neural networks comprise multiple bands of neurons connected by edges with weights that are updated during training. Neural networks are widely used in deep learning processes as they can interpret occurrences of perplexing repetitive sequences. The advantages of this algorithm are its flexibility, it can interpret information from huge datasets, and its potential to replicate lateral relationships.

Multi-Layer Perceptron (MLP) is used in deep learning tasks. MLPs consist of multiple layers which are interconnected to each other, every neuron of one layer acts as an input to the neuron in the next layer. MLPs are used in classification, regression, and serve as the foundation for more intrinsic neural networks. Its capability to handle both regression and classification problems, and its role as a foundation for more advanced neural networks is advantageous

3. PAPERS REVIEWED

Vijay Kumar, Rachna Narula, and Akanksha Kochhar [1] developed a model to analyze bank credit data for loan approval decisions. Their results showed that K-NN outperformed Logistic Regression, helping businesses reduce risks associated with loan defaults

Jinchen Lin[2] applied machine learning to analyze credits associated with loans using Kaggle dataset, comparing various algorithms. XGBoost showed the highest accuracy, while Logistic Regression performed the worst, highlighting XGBoost's potential to boost profitability for lenders.

Wanjun Wu's[3] study applies Random Forest and XGBoost algorithms to predict loan default cases, achieving high accuracy scores of 0.90657 and 0.90635, respectively. The author has employed feature engineering techniques, including the variance threshold method and Variance Inflation Factor (VIF), to remove irrelevant features before applying the models. The results show minimal differences in prediction performance between the two algorithms, indicating that both are effective for this task.

Platur Gashi[4] highlights the effectiveness of various algorithms, including Neural Networks (NN), Naive Bayes, as well as ensemble methods like Bagging (Random Forest), Boosting (Decision Trees), and Stacking (LR, SVM, NN) in predicting loan defaults. The results reveal that ensemble models outperform individual classifiers, with Boosted Decision Trees achieving maximum accuracy at 84.9% whereas 83.1%. Among individual classifiers, Neural Networks performed the best, reaching an accuracy of 80.3%, while Naive Bayes lagged significantly at 46.5%. These findings underscore the superior predictive power of ensemble approaches in this domain.

Qingyong Chu, Ping Hu, Xinchang Song, Xu Zhu and Lu Peng[6] utilized algorithms to determine the likelihood that the loan will default. With an accuracy greater than 0.8, LightGBM and XGBoost obtained the best results. Crucial elements impacting forecasts, like loan length, grade, and credit rating, were explained using the LIME technique.

Author	Technique Used	Accuracy
Vijay Kumar, Rachna Narula, Akanksha Kochhar [1]	LR and KNN	KNN performed better than LR
Jinchen Lin[2]	Random Forest, Logistic Regres-sion, XGBoost, AdaBoost	XGBoost - 93.26% AdaBoost - 92.37% RF - 92.72%
Wanjun Wu[3]	Random Forest, XGBoost	RF - 0.90657, XGBoost - 0.90635
Platur Gashi[4]	LR, DT SVM, NN, NB, Bagging - RF, Boosting - DT, Stacking - LR, SVM, NN	LR - 76.9% DT - 75.3% SVM - 75.9% NN - 80.3% Naive Bayes - 46.5% Bagging RF - 83.1% Boosting DT - 84.9% Stacking - LR, SVM, NN - 73.7%
Lili Lai[5]	XGBoost, AdaBoost. KNN, RF, MLP	RF - 50.1% AdaBoost - 100% XGBoost - 71.6% KNN - 50.3% MLP - 50%
Qingyong Chu,Ping Hu, Xinchang Song,Xu Zhu,Lu Peng[6]	DT, LR, LightGBM, XGBoost LIME	Logistic Regression- 65.55%, Decision Tree- 63.17%, XGBoost - 80%, LightGBM- 81.04%

4. PROPOSED SYSTEM

Problem Statement: "To predict loan default using machine learning techniques."

Problem Elaboration: For financial organizations, loan failure poses a serious problem since it can result in large losses and elevated risk. Conventional credit evaluation techniques are frequently ineffective and have trouble identifying subtle trends in borrower behavior, which leads to imprecise forecasts. The volume of loan data has increased due to the growth of digital financial services, necessitating the use of more advanced algorithms to forecast defaults. Loan default prediction can be automated and data-driven with machine learning; nevertheless, choosing the best algorithm can be difficult, especially in cases when the datasets are unbalanced and defaults are few. In order to determine which machine learning model performs best, this study compares the efficacy of several algorithms in forecasting loan defaults. Assisting financial institutions in managing risk better and making more informed loan decisions is the aim.

Proposed Methodology:

Proposed Method: Weighted Voting Ensemble Model for Loan Default Prediction

The Weighted Voting Ensemble model combines the strengths of multiple algorithms to improve the accuracy as well as robustness of loan default predictions. In this approach, individual models are trained on the loan dataset after it has been preprocessed for cleaning and scaling. Each model generates its predictions based on the input features, reflecting its unique learning approach. The diversity among the models helps capture various patterns in the data, which is crucial for complex financial datasets that may not be easily understood by a single model.

After generating predictions, the ensemble method assigns weights to each model based on metrics such as precision, accuracy and recall that can provide insights into the performance of the model. This weighting system ensures that better-performing models contribute more significantly to the final prediction. The predictions from each model are then combined through a weighted voting mechanism, leading to a consensus prediction that is expected to be more reliable than those generated by individual models. This method harnesses the collective knowledge of different algorithms, ultimately enhancing the ability to predict loan defaults effectively in the financial sector.

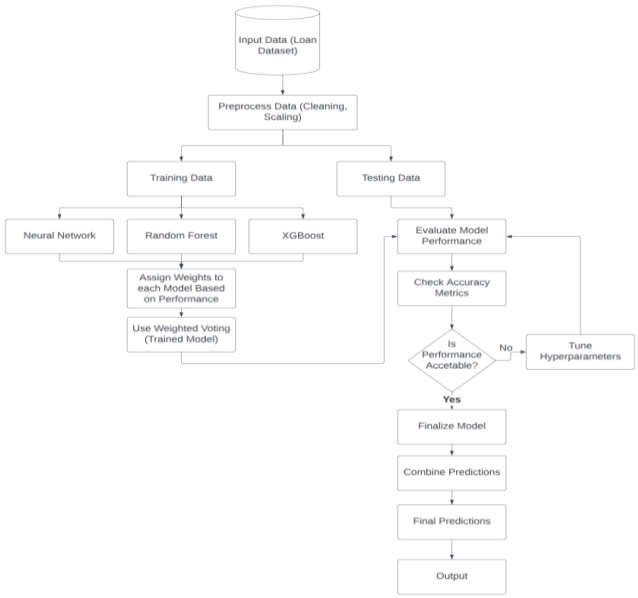


Fig.1 System Architecture

5. CONCLUSION

In this paper, we explored various machine learning models for loan default prediction and found that ensemble learning methods generally yield high accuracy due to their ability to combine multiple model predictions effectively. Building on this insight, we proposed a more complex model, essentially a "double ensemble" approach—where we apply ensemble learning on top of individual ensemble models like Random Forest, XGBoost, and Neural Networks. By assigning weights to each model's prediction based on performance, this layered ensemble method further enhances prediction accuracy and robustness, making it a highly reliable tool for loan default risk assessment.

6. REFERENCES

[1] Vijay Kumar, Rachna Narula, Akanksha Kochhar "Loan Default Prediction using Machine Learning Models" 2024 DOI: 10.5281/zenodo.8337054

[2] Jinchen Lin "Research on loan default prediction based on logistic regression, randomforest, xgboost and adaboost" 2023 pp DOI: 10.1051/shsconf/202418102008

[3] Wanjun Wu "Machine Learning Approaches to Predict Loan Default 2022 DOI: 10.4236/iim.2022.145011

[4] Platur Gashi, "Loan Default Prediction Model" 2023, DOI:10.13140/RG.2.2.22985.01126

[5] Lai, L, "Loan default prediction with machine learning techniques". In: 2020 International Conference on Computer Communication and Network Security (CCNS). pp. 5–9. IEEE (2020)

[6] Xu Zhu, Qingyong Chu,Xinchang Song, Ping Hu,Lu Peng, Explainable prediction of loan default based on machine learning models DOI:10.1016/j.dsm.2023.04.003

[7] Loan Default Dataset Kaggle - http://bit.ly/3BA9WeX

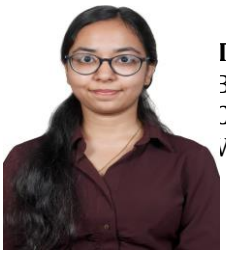
BIOGRAPHIES

Prof. Pramila M. Chawan holds the position of Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She pursued B.E. (Computer Engineering) and M.E. (Computer Engineering) from VJTI College of Engineering.She has guided 100+ M. Tech. and 150+ B. Tech. projects in her 31 years of profession. In Peer-reviewed International journals, International conferences & Symposiums she has published 181 papers. She has been on the planning committees for six faculty development programs and 29 international conferences. She is consulting editor on 9 scientific research journals. The Society of Innovative Educationalist & Scientific Research Professional, Chennai (SIESRP) awarded her with 'Innovative & Dedicated Educationalist Award Specialization: Computer Engineering & I.T.'

Adwait Mandge,
B. Tech Student, Dept. of Computer Engineering and IT, VJTI, Mumbai, Maharashtra, India

Rohan Fatehchandka,
B-Tech Student, Dept. of Computer Engineering and IT, VJTI, Mumbai, Maharashtra, India

Kunal Goudani,
B-Tech Student, Dept. of Computer Engineering and IT, VJTI, Mumbai, Maharashtra, India



Tanaya Shelke,
B. Tech Student, Dept. of Computer Engineering and IT, VJTI, Mumbai, Maharashtra, India