PROJECT REPORT DATA SCIENCE FALL 2023

Predicting Loan Approval Status Using Machine Learning

Haris Hassaan (23L-2642), Fardeen Mirza (23L-2576), Zuraiz Anjum (23L-2547)

[1]Department of Computer Science, National University of Computer and Emerging Sciences, Lahore

November 27, 2024

Abstract

This study focuses on developing a predictive model for determining the risk score associated with loan applications. The risk score is crucial in assessing the likelihood of a borrower defaulting on a loan, and accurate predictions can help lenders make informed decisions. The dataset used for this analysis includes various features such as credit score, employment status, education level, and others. A combination of machine learning models, including XGBoost, Random Forest, Support Vector Regression (SVR), and Linear Regression, was employed to identify the most effective approach for risk prediction. We began by cleaning the data, eliminating irrelevant features, and performing dimensionality reduction using Recursive Feature Elimination (RFE). This resulted in selecting the most influential features. Subsequently, four different models were trained and evaluated on the test data. The performance of each model was assessed using key metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), and Root Mean Squared Error (RMSE). The results indicated that Support Vector Regression (SVR) with hyperparameter tuning outperformed other models, delivering the lowest MSE and RMSE, with an R^2 score of 0.6855. To further enhance the predictive power, a Stacking Regressor model was built by combining XGBoost, Random Forest, and the tuned SVR model as base learners, with Linear Regression as the meta-learner. This ensemble model achieved a R^2 score of 0.7009, outperforming individual models, and presented the best overall performance with an MSE of 0.2868 and RMSE of 0.5355. The results demonstrate the potential of stacking techniques in improving predictive performance, especially in complex tasks like loan risk prediction. Future work could explore additional feature engineering, the incorporation of more advanced algorithms, and real-time prediction applications to optimize loan approval processes.

1 Introduction

The prediction of loan risk is a critical area in financial services, as it enables lenders to assess the likelihood of default and make informed decisions. Machine learning has emerged as a powerful tool in this domain, allowing for the analysis of vast datasets and the extraction of patterns that are not immediately obvious. Traditional credit scoring models often rely on predefined rules, but machine learning models can learn from data, adapting and improving over time. This study seeks to compare the performance of multiple machine learning algorithms in predicting loan risk based on a variety of borrower features. The primary objective of this study is to identify the most accurate model for predicting loan risk and to explore whether combining models in a stacking ensemble method can improve prediction accuracy. The dataset used in this study consists of features such as credit score, loan duration, bankruptcy history, and employment status, among others. By comparing the performance of different models, including XGBoost, Random Forest, Support Vector Regression (SVR), and Linear Regression, this research aims to determine which model or combination of models provides the best results for predicting loan risk.

2 Methodology

The methodology adopted for this study involved several key steps, starting with data preparation and cleaning. The dataset was split into training and testing sets using an 80/20 ratio. We then removed unnecessary features, such as the "ApplicationDate" column, which did not contribute to predicting the loan risk score. Dimensionality reduction was performed using Recursive Feature Elimination (RFE) to select the most important features. After feature selection, four models were trained: XGBoost, Random Forest, SVR, and Linear Regression. These models were evaluated using multiple performance metrics, including MSE, MAE, R^2, and RMSE. Additionally, hyperparameter tuning was performed for the SVR model using RandomizedSearchCV to find the best combination of hyperparameters. The best-performing models were then used in a Stacking Regressor ensemble, which combined XGBoost, Random Forest, and the tuned SVR model as base learners, with Linear Regression as the meta-learner. The final performance of the Stacking Regressor was compared to the individual models.

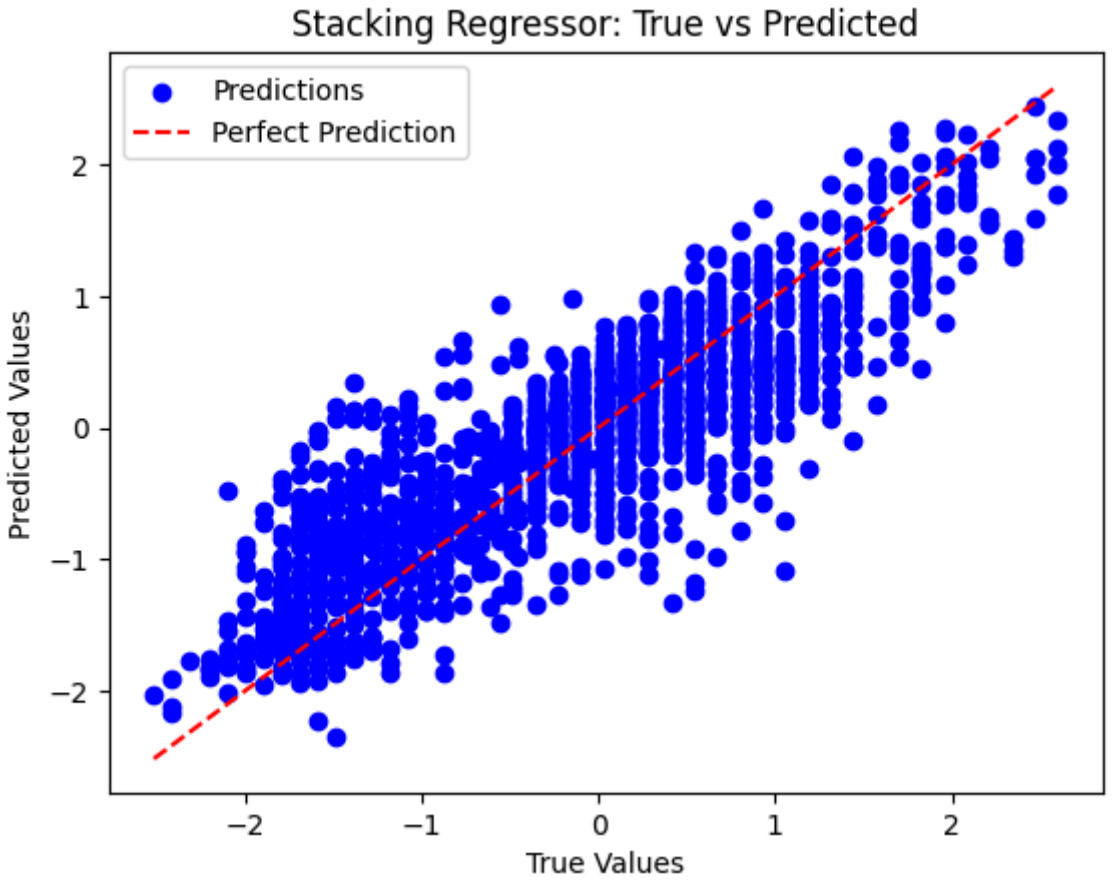

Figure 1: Scatter plot of Regressor stacking (True vs Predicted) Values.

3 Experiments

The experiment started by preparing the dataset, followed by splitting the data into training and test sets. The training set was used to train the models, and the testing set was used to evaluate their performance. The models were assessed based on the following metrics: **Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), Root Mean Squared Error (RMSE).** After training and evaluating the initial models, the best performing model (SVR) was selected for hyperparameter tuning using RandomizedSearchCV. The hyperparameters were fine-tuned to improve the model's performance. Following the optimization, the SVR model was retrained, and its performance was compared to the other models. To improve prediction accuracy, a Stacking Regressor model was constructed, combining the best performing models (XGBoost, Random Forest, and SVR). The stacking model was then evaluated based on the same metrics used for individual models, and the results were compared to determine the best-performing approach.

4 Results & Discussion

The results showed that the SVR model, after hyperparameter tuning, outperformed other models in terms of prediction accuracy. The final SVR model achieved an MSE of 0.3015 and an R^2 score of 0.6855. However, the Stacking Regressor model, which combined multiple models, yielded even better results, with an MSE of 0.2868 and an R^2 score of 0.7009. These results highlight the importance of model combination techniques, such as stacking, in improving predictive performance. The individual models provided useful insights into which features were important for predicting loan risk. For example, features such as credit score, employment status, and loan duration were found to have a significant impact on the prediction. The performance of the models was also consistent across the evaluation metrics, with the Stacking Regressor consistently outperforming other models in terms of all metrics. The results suggest that the use of machine learning models, particularly ensemble methods like stacking, can significantly improve loan risk prediction accuracy. This has important implications for financial institutions, as more accurate predictions can lead to better loan approval decisions and reduced financial risk.

Conclusion and Future Work

In conclusion, this study demonstrates that machine learning, particularly ensemble methods like stacking, can enhance the prediction of loan risk, providing more accurate and reliable predictions compared to individual models. The findings suggest that a combination of XGBoost, Random Forest, and SVR models, with Linear Regression as the meta-learner, is the most effective approach for predicting loan risk in the given dataset. Future work could explore additional techniques such as feature engineering, the inclusion of new data sources, or the application of more advanced algorithms. Additionally, real-time prediction models could be developed to improve decision-making processes in the financial industry. Further research could also look into the impact of different hyperparameters, such as kernel functions in SVR or tree depths in XGBoost, on model performance.

References

[1] Lorenzo Zoppelletto. *Financial Risk for Loan Approval Dataset*. Available at: https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval/data. Accessed November 27, 2024.