Comparative Analysis of Classification Models

Abstract

Background

Accurate loan approval decisions are important for minimizing risk and maximizing profitability for lending institutions. “One of the important steps for banks to decide if a loan has to be authorized is to ensure that the candidate to borrow has the capacity of paying back the loan in the proposed terms. The advancement of technology like machine learning, computer science and other science is playing an important role by supporting banks to predict the probability of defaulting for a given customer based on his past behavior.” [1]. This study aims to develop and compare different machine learning algorithms for predicting loan approval status using a dataset of loan applications.

Objective

The main objective of this assignment was to identify the best machine learning model for loan approval prediction based on two metrics: accuracy and area under curve.

Methods

I took the following steps to preprocess the dataset [2][3]:

- Handling missing values: I replaced categorical data such as gender and credit history with the mode and numerical such as LoanAmount with the median.
- Encoding and standardization: Categorical variables were encoded using LabelEncoder, while Numerical features were standardized and scaled using StandardScaler).

Afterward, five machine learning algorithms were selected, implemented, and evaluated using accuracy and Area Under the Curve (AUC) metrics. The models are:

- Artificial Neural Network (MLPClassifier)
- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Support Vector Machine (SVC)

Feature selection was performed using the SelectKBest method with f_classif scoring function. Finally, hyperparameter tuning was conducted on the best-performing model using RandomizedSearchCV.

Results

Results showed that the Random Forest Classifier consistently outperformed other models' accuracy and AUC scores: (Accuracy: 85.7%, AUC: 88.2%). Feature selection slightly reduced the model performance but simplified the feature set and model computational efficiency. Hyperparameter tuning further improved results: (Accuracy: 87.9%, AUC: 90.1%)

Conclusion

In conclusion, the Random Forest algorithm was very effective for this loan approval prediction, likely due to its ability to handle complex relationships in the data. It also shows the importance of model selection, feature engineering, preprocessing, and hyperparameter tuning in achieving the best model performance, especially in ensemble learning[4]. This suggests that merely training a model isn't enough to achieve the best performance. Fine-tuning it over time will create the best outcome.

References

1. T. Ndayisenga, “Bank Loan Approval Prediction Using Machine Learning Techniques,” Thesis, 2021. Accessed: Nov. 21, 2024. [Online]. Available: http://dr.ur.ac.rw/handle/123456789/1437
2. “Data Preprocessing in Machine Learning.” Accessed: Nov. 20, 2024. [Online]. Available: https://kaggle.com/code/alirezahasannejad/data-preprocessing-in-machine-learning
3. “Data Preprocessing Techniques in Machine Learning [6 Steps],” Scalable Path. Accessed: Nov. 20, 2024. [Online]. Available: https://www.scalablepath.com/data-science/data-preprocessing-phase.
4. “What is ensemble learning? | IBM.” Accessed: Nov. 21, 2024. [Online]. Available: https://www.ibm.com/topics/ensemble-learning