Loan Approval Prediction Using Machine Learning

Jingting Zhan

Northeastern University

12/7/2025

Abstract

This report presents a complete analytics workflow for developing a loan approval prediction model. The analysis includes data quality assessment, exploratory data analysis (EDA), and the implementation of three supervised machine learning approaches: Logistic Regression, Random Forest, and Gradient Boosting. The goal is not only to build accurate predictive models but also to understand the underlying behavioral patterns that influence loan decisions.

Introduction

Predictive modeling has become a fundamental tool in financial services, especially for credit scoring and loan approval decisions. The purpose of this project was to build machine learning models that can classify loan applications as approved or rejected using financial and demographic features. The workflow follows the CRISP-DM structure: data understanding, preparation, modeling, and evaluation.

Data Quality Assessment

Several data quality issues were identified. First, some asset-related variables contained negative values, which are not logically valid. These were corrected by clipping values to zero, following recommended approaches for handling invalid outliers (Ghosh & Vogt, 2012). Second, many numeric variables—including income, loan amount, and asset values—showed heavy right-skewness and extreme outliers, which were mitigated using winsorization at the 1st and 99th percentiles. Third, missing values were handled with median and most-frequent imputation,

consistent with best practices for machine learning data preparation (Kim & Kim, 2018). Finally, the target variable exhibited moderate imbalance, addressed using class weighting and stratified sampling (He & Garcia, 2009). Exploratory Data Analysis (EDA) EDA revealed meaningful behavioral patterns. Higher income levels were strongly associated with loan approval. CIBIL score demonstrated one of the clearest separations between approved and rejected applicants, aligning with established credit-scoring literature (Khandani et al., 2010). Education level and employment type also showed predictive value: self-employed applicants had a lower approval rate, while applicants with higher education levels were more likely to be approved. Asset variables such as residential and commercial assets were positively correlated with approval probability.

Modeling Methods

Three supervised learning algorithms were applied. Logistic Regression served as a baseline model due to its interpretability (Hosmer et al., 2013). Random Forest, known for capturing nonlinear relationships and interactions, has strong performance in credit scoring (Breiman, 2001). Gradient Boosting, a powerful ensemble method, is well-established as a top-performing algorithm for tabular credit datasets (Friedman, 2001; Lessmann et al., 2015). A unified preprocessing pipeline ensured consistent imputation, scaling, and encoding.

Model Evaluation

Performance was assessed using accuracy, recall, precision, F1 score, ROC-AUC, confusion matrices, and cross-validation. Random Forest and Gradient Boosting achieved the strongest performance, with high discriminatory ability shown in their ROC curves. Logistic Regression performed reasonably well but lack ednonlinear modeling capability. Patterns Illustrated by the Models Feature importance results revealed that CIBIL score was the strongest predictor across all models. Income, loan amount, loan term, and asset values were also influential. The models uncovered risk-driven interactions, such as the combination of low CIBIL and high loan amount significantly reducing approval probability. Self-employed applicants had systematically lower predicted approval rates even after controlling for income, supporting EDA findings. These insights align with real-world credit risk considerations (Provost & Fawcett, 2013).

Conclusion

This project developed a comprehensive predictive modeling workflow for loan approval classification. Throughrigorous EDA, data cleaning, and model comparison, the analysis identified key financial factors influencing approval decisions. Ensemble models provided the strongest predictive capability and most useful managerial insights. Ensuring data quality was essential to obtaining reliable and interpretable results.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. Advances in Intelligent and Soft Computing, 166, 89–96.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787.

Kim, S., & Kim, H. (2018). A data quality framework for machine learning. Journal of Intelligent Information Systems, 50(1), 1–20.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking classification algorithms for credit scoring. European Journal of Operational Research, 247(1), 124–136.

Provost, F., & Fawcett, T. (2013). Data science for business. O'Reilly Media.