Submitted by: Ayush Bhotika

PGDM-Research & Business Analytics.

Abstract

This work builds predictive models to determine whether a loan application will be approved, focusing on Logistic Regression and Decision Tree classifiers as the core methods. To address the skewed target distribution, SMOTE is used to balance classes before training, enabling fairer learning and evaluation across approval outcomes. The objectives are to surface the most influential approval drivers and to compare models using accuracy, precision, recall, F1, and ROC-AUC to understand performance trade-offs.

Introduction

Loan approval prediction is central to risk management in lending because accurate estimates reduce defaults and improve portfolio quality. Applicant assessments typically span demographics, financial capacity, and credit behavior, making the problem well-suited to supervised machine learning for robust, repeatable decisions. By formalizing these variables into models, institutions can streamline approvals and enhance consistency at scale.

Dataset description

The dataset captures attributes such as applicant and co-applicant income, employment and education status, credit history, loan amount, term, property area, and the binary loan status target indicating approval or rejection. The target is imbalanced with one class underrepresented, necessitating synthetic oversampling to reduce bias in model training and evaluation. Variables combine numeric and categorical types aligned to common consumer credit features used in retail lending.

Methodology

- Data preparation includes handling missing values, encoding categorical fields, and partitioning the data into training and testing sets to ensure unbiased validation of results.
- Class imbalance is mitigated using SMOTE on the training set to synthesize minority samples and create a balanced learning signal for classifiers.
- Two baseline models are fit: Logistic Regression for transparent linear decision boundaries and Decision Trees for non-linear splits and interactions.

- Performance is assessed with accuracy, precision, recall, F1-score, and ROC-AUC to capture both threshold-based and ranking-based quality, enabling a broader view of model behavior.

Results and evaluation

Logistic Regression offers clear interpretability, with variables like credit history and applicant income emerging as high-impact factors for approval decisions in this dataset. Decision Trees capture complex relationships but are prone to overfitting unless regularized and tuned, which can lower generalization quality relative to simpler linear boundaries. Metric comparisons show expected trade-offs between overall accuracy and error balance across classes, with ROC-AUC complementing threshold metrics for a fuller diagnostic.

Conclusion and insights

Both models are viable for loan status classification, but Logistic Regression stands out for ease of interpretation and reliable generalization on this problem setup. Decision Trees can uncover non-linear structure but typically require stronger hyperparameter tuning or ensembles to consistently outperform a well-specified logistic baseline. Future enhancements should explore Random Forests or Gradient Boosting and apply targeted feature engineering to further raise ROC-AUC and class-balanced metrics.