Loan Approval Prediction Analysis

Authors: Pinkesh Tandel and Dennis Darko
Northeastern University
Course: ALY6020: Predictive Analytics
Date: October 22, 2024
Under the guidance of Dr. Nina Rajabi Nasab

Abstract
This report outlines a predictive analysis for loan approval using various machine learning models, including Logistic Regression, SVM, Decision Tree, Random Forest, and Gradient Boosting. The models were evaluated based on performance metrics such as precision, recall, and F1-score. The Gradient Boosting model provided the highest F1-score, making it the most effective for predicting loan approval. Key factors like **Income_log**, **Family**, and **CCAvg_log** were identified as significant contributors to loan approval decisions.

Introduction
The task of loan approval involves assessing numerous factors, making it ideal for predictive modeling. In this analysis, various machine learning models were trained to predict loan approvals based on a range of customer data. The dataset includes variables such as age, income, family size, and credit card spending. This report presents the steps taken to preprocess the data, train models, and evaluate their performance.

Methods

Data Preprocessing

The dataset was checked for missing values and cleaned. Multicollinearity was assessed using Variance Inflation Factor (VIF). Features such as **Experience** were removed due to high multicollinearity, and log transformations were applied to skewed variables such as **Income** and **CCAvg** to handle outliers.

Model Training

Five machine learning models were trained:

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Decision Tree**
- **Random Forest**
- **Gradient Boosting**

The data was split into training and testing sets, and the features were standardized before training.

Model Evaluation

The models were evaluated using **Precision**, **Recall**, **F1-Score**, and **Processing Time**. These metrics provide insight into each model's ability to correctly approve or reject loans while minimizing false positives and false negatives.

Results
The following table summarizes the performance of each model:

Model	Precision	Recall	F1-Score	Processing Time (s)
Logistic Regression	0.8246	0.4476	0.5802	0.0291
SVM	0.9630	0.4952	0.6541	0.2043
Decision Tree	0.8222	0.7048	0.7590	0.0131
Random Forest	0.9412	0.6095	0.7399	0.3765
Gradient Boosting	0.9077	0.5619	0.6941	0.4140

The **Gradient Boosting** model performed the best, achieving an F1-score of **0.6941** and a precision of **0.9077**.

Feature Importance

The Gradient Boosting model identified the following features as the most significant:

1. **Income_log**: The highest contributor to loan approval decisions.
2. **Family**: Family size also played a significant role.
3. **CCAvg_log**: Average credit card spending had a moderate influence on loan approval decisions.

Discussion
This analysis identified the key factors influencing loan approval. Of the three most significant variables, **Income_log** had the most positive influence on loan acceptance. The best-performing model was **Gradient Boosting**, as it provided the highest F1-score. The F1-score was chosen as the most appropriate KPI, balancing precision and recall to minimize both false positives and false negatives.

Conclusion
In conclusion, the **Gradient Boosting** model is recommended for loan approval predictions, offering the best balance between accuracy and processing time. Further improvements could involve exploring more advanced feature engineering techniques and fine-tuning model parameters to enhance performance.

References

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.