Comparative Analysis of Machine Learning Models for Loan Default Prediction: A Mathematical Perspective

Michael Miller

Abstract

This study presents a comprehensive analysis of three machine learning models—Logistic Regression, Decision Tree, and Random Forest—for predicting loan defaults. We explore the mathematical foundations of these models, their implementation using Python's scikit-learn library, and their performance in predicting loan defaults and estimating expected losses. The study also addresses class imbalance using SMOTE and investigates feature importance across models.

1. Introduction

Accurate prediction of loan defaults is crucial for financial institutions to manage risk effectively. This research compares three popular machine learning models in their ability to predict loan defaults and estimate potential losses. We focus on the mathematical underpinnings of these models and their practical implementation.

2. Methodology

2.1 Data Preparation and Preprocessing

The dataset contains various features related to loan applicants, including financial indicators and employment information. We perform the following preprocessing steps:

1. Removal of non-predictive features (e.g., customer_id)
2. Splitting the data into training and test sets (70-30 split)
3. Addressing class imbalance using Synthetic Minority Over-sampling Technique (SMOTE)
4. Feature scaling using StandardScaler for Logistic Regression

2.2 Model Descriptions and Mathematical Foundations

2.2.1 Logistic Regression Logistic Regression models the probability of default using the logistic function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}}$$

where Y is the binary outcome (default or not), X_i are the input features, and β_i are the model coefficients.

The model is trained by minimizing the log-likelihood function:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

where y_i are the true labels and p_i are the predicted probabilities.

2.2.2 Decision Tree Decision Trees make predictions by learning decision rules inferred from the data features. The algorithm aims to maximize information gain at each split, typically using metrics like Gini impurity or entropy.

For a binary classification problem, the Gini impurity for a node is calculated as:

$$Gini = 1 - (p_0^2 + p_1^2)$$

where p_0 and p_1 are the proportions of class 0 and class 1 samples at the node.

2.2.3 Random Forest Random Forest is an ensemble of Decision Trees. It uses bagging (bootstrap aggregating) and feature randomness to create a robust classifier. The final prediction is typically the mode of the predictions from individual trees.

For a Random Forest with T trees, the probability of default is:

$$P(Y = 1|X) = \frac{1}{T}\sum_{t=1}^{T} P_t(Y = 1|X)$$

where $P_t(Y = 1|X)$ is the prediction of the t-th tree.

2.3 Expected Loss Calculation

The expected loss for each loan is calculated as:

$$E[Loss] = Loan\_Amount \times P(Default) \times (1 - Recovery\_Rate)$$

where $P(Default)$ is the predicted probability of default, and we assume a fixed Recovery_Rate of 0.10.

3. Results and Analysis

3.1 Model Performance

The ROC curves for all three models are presented in Figure 1:

All models show excellent performance, with AUC scores close to 1.0. The Random Forest and Logistic Regression models slightly outperform the Decision Tree model.

3.2 Feature Importance

Figure 2 shows the feature importance for each model:

Key observations:
1. Logistic Regression: 'credit_lines_outstanding' and 'total_debt_outstanding' are the most important features.
2. Decision Tree: 'credit_lines_outstanding' dominates, with other features having minimal importance.
3. Random Forest: Provides a more balanced importance distribution, with 'credit_lines_outstanding' still being the most important.

3.3 Expected Loss Estimation

The average expected loss per loan for each model:

1. Logistic Regression: $807.24
2. Decision Tree: $802.60
3. Random Forest: $807.97

These results provide valuable insights into the economic implications of each model's predictions. The expected loss estimates are remarkably close across all three models, with a difference of less than $6 between the highest and lowest estimates. This suggests that all three models are relatively consistent in their risk assessment, despite their different methodologies.

The Random Forest model shows the highest average expected loss, which might indicate a slightly more conservative approach to risk assessment. The Decision Tree model, on the other hand, shows the lowest average expected loss, potentially suggesting a more optimistic view of loan risks.

The Logistic Regression model falls between the other two, providing a middle-ground estimate. This could be seen as a balance between the more complex Random Forest model and the simpler Decision Tree model.

It's important to note that while these differences are small, they could translate to significant amounts when applied to a large portfolio of loans. For instance, in a portfolio of 10,000 loans, the difference between the Decision Tree and Random Forest models would amount to about $54,000 in expected losses.

4. Discussion

The results demonstrate the high predictive power of all three models in loan default prediction. The similar performance across models suggests that the relationship between the features and loan default risk is well-captured by both linear (Logistic Regression) and non-linear (Decision Tree, Random Forest) approaches. This is further reinforced by the close alignment of expected loss estimates across all three models, indicating a consistent assessment of risk despite the different methodological approaches.

The feature importance analysis reveals that 'credit_lines_outstanding' is consistently the most important predictor across all models. This suggests that the number of credit lines a customer has open is a strong indicator of their likelihood to default on a loan.

The expected loss calculations provide a practical application of these models in risk management. By quantifying the potential loss for each loan, financial institutions can make more informed decisions about loan approvals and pricing.

The consistency in expected loss estimates across models provides additional confidence in their reliability. However, the slight differences between models could be significant when applied to large loan portfolios. Financial institutions might consider using an ensemble approach, combining predictions from multiple models to achieve a more robust risk assessment.

5. Conclusion

This study provides a comprehensive comparison of Logistic Regression, Decision Tree, and Random Forest models for loan default prediction. All models demonstrate high predictive accuracy, with Random Forest showing slightly better performance in terms of AUC score. The analysis of feature importance across models offers valuable insights into the key factors influencing loan default risk. Moreover, the consistent expected loss estimates across models provide a reliable basis for financial risk assessment in loan portfolios.

The close alignment of expected loss estimates, despite the different methodologies employed by each model, suggests that these approaches are capturing fundamental patterns in the data related to default risk. This consistency adds credibility to the models' predictions and provides a solid foundation for risk management strategies in lending institutions.

Future work could explore more advanced ensemble methods, incorporate additional features, or investigate the models' performance on different datasets to assess their generalizability. Additionally, research into the economic implications of using these models for decision-making in real-world lending scenarios could provide valuable insights for financial institutions.

References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
2. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Appendix

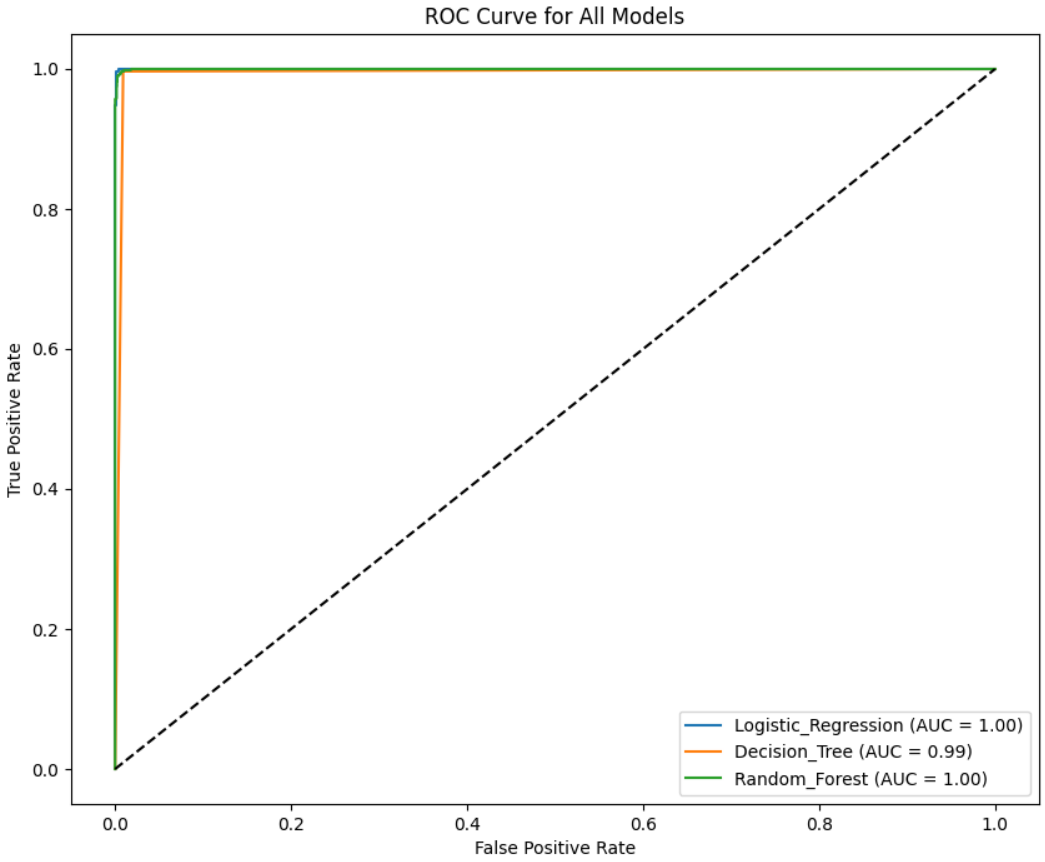

Figure 1: ROC Curve for All Models

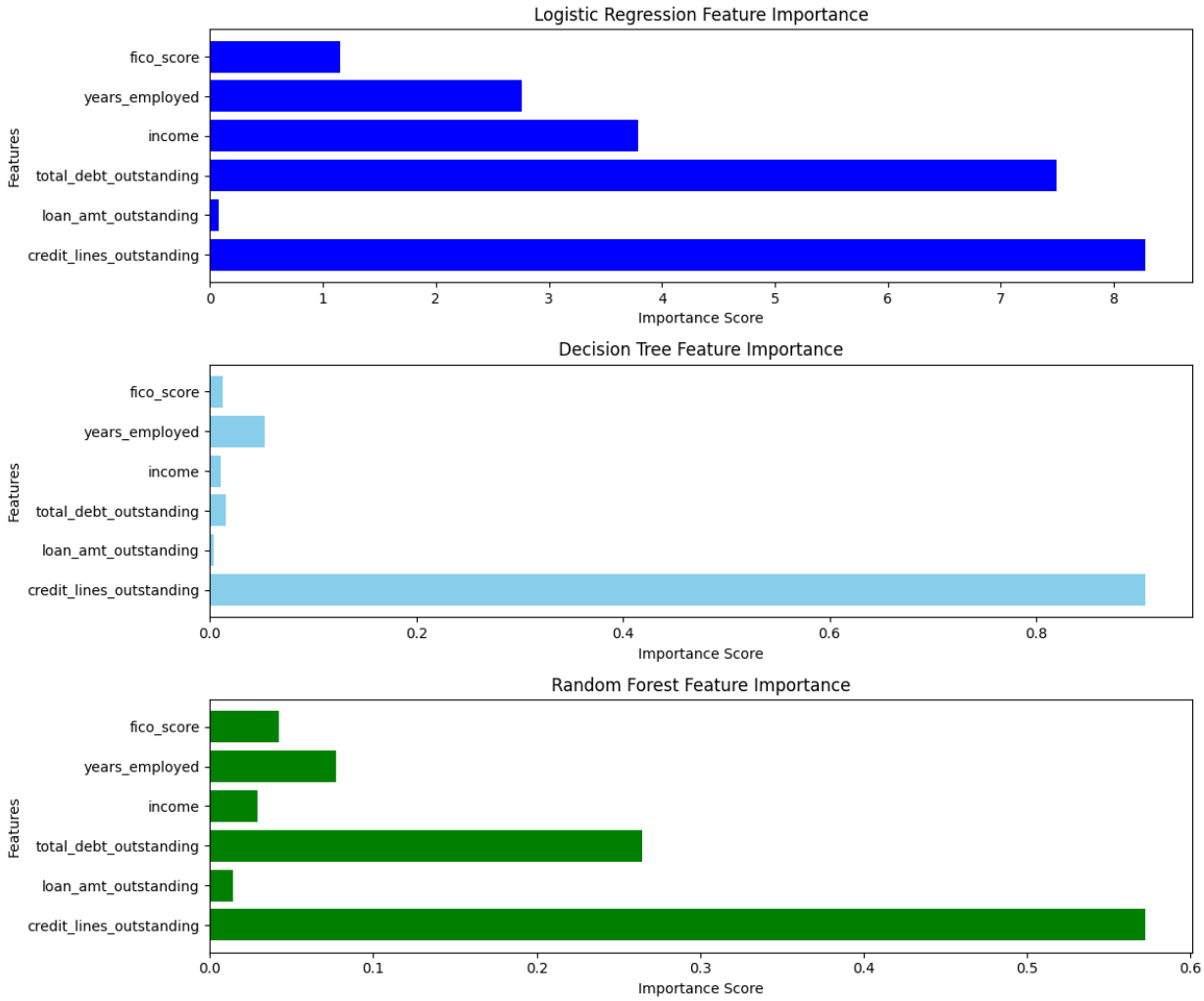

Figure 2: Feature Importance Comparison