Predictive Analysis on Loan Approval

Bidya Bhattarai
Shriju Maharjan

Department of Business Analytics

University of South Dakota

DSCI 726: Operational Analytics

Dr. Yi Liu

Project Overview

The purpose of this memorandum is to present predictive analysis results from "Loan approval Prediction and Optimization" project. The focus is on predicting loan approval outcomes using financial and personal attributes such as income, credit score, loan amount, and education feature. The analysis targets applicant's income between $50k and $100k, aiming to optimize the approval status.

Target Variable and Predictor Variable

Predictors: Based on descriptive analysis, the most influential features are: Cibil Score and Loan Amount. Other features like Education and Income Annum had less influence , so it will be served as secondary variable.

Data Preprocessing

Education and Loan Status were converted into factors with binary levels 1 and 0 where 1 being graduate and approved for education and loan status respectively and 0 being not graduate and rejected for education and loan status respectively to fit the models properly. No other transformations were needed as the data was clean and ready for analysis without extra modification.

Methods Tried and Evaluated

The methods we will try are logistic regression and decision tree.

1. **Logistic Regression**
   Logistic Regression is a simple yet effective method for binary classification problems such as predicting loan approval status. We did not find this method useful as it was limited by low sensitivity, specificity, and F1 score, suggesting that the model was not complex enough to capture all the patterns in the dataset.

2. **Decision Tree**
   Decision trees can handle both numerical and categorical data effectively, allowing for nonlinear relationship.

3. **Random Forests**
   Random forests help address the overfitting issues found in decision trees by combining multiple trees to improve generalization.

   Evaluation:
   Logistic Regression: It was chosen because of its ability to provide probability estimated.
   Decision Trees: It was used for its ease of interpretation.
   Random Forest: Random forest was tested to improve accuracy and minimize overfitting compared to single decision trees.

Data Transformations:

a. **Binning:**
   We did not use binning in this analysis and kept the continuous variable like loan amount, income and cibil score in their original continuous form to retain their detailed numerical relationships with the loan approval status.

b. **Variables Added:**
   We have added loan term as predictor variable because there was noticeable correlation between the loan term and the likelihood of loan approval. Loans with shorter terms were more likely to be approved than those with longer terms. The duration of loan is an important factor in determining whether the loan is approved.

c. **Variables Removed:**
   We removed education level as there was no significant difference in loan approval rates between graduates and non-graduates. Both educational level had similar proportions of approved and rejected loans, which suggested that education was not a strong determinant of loan approval. Since the variable did not provide meaningful difference, it was necessary to exclude it to avoid overfitting.

d. **Creation of Synthetic Variable:**
   We did not create any synthetic variables in this analysis. In this analysis, the original variables like loan amount, income, cibil score and loan term were sufficient to model the loan approval prediction.

Data Splitting or Cross Validation Choice

a. **Data Splitting:**
   We used data splitting to train and test the logistic regression model. The technique we used is train-test method, where 80% of the data was used for training the model and 20% was used for testing it. This random split ensures that the training

and test sets are independent of each other to ensure unbiased evaluation of model's performance on unseen data.

We used createDataPartition() function from the caret package to split the data randomly into training and test sets based on the target variable loan_status.

b. **Cross Validation:**

We used cross validation as part of evaluation process. 10 fold cross validation was used for decision trees and random forests model. Here, the data is divided into 10 equal parts or folds. We trained the model on 0 of these folds and tested on the remaining fold, repeating the process for 10 times with each fold being used as the test set once.

Collinearity

We checked collinearity in the analysis through correlation matrix. We examined correlation matrix to assess potential multicollinearity between predictors.

No significant multicollinearity was identified between the key predictors like loan amount, income, cibil score, loan team).

Predictors Removed:

Yes, we removed one of the predictor during the model refinement process.

a. **Techniques Used:**

We evaluated the significance of each predictor variable using p-values from the logistic regression model. The predictor with high p-value which showed that it lacks statistical significance was considered for removal.

b. **Variable Removed:**

The variable "education" was removed as it did not contribute significantly to the model's predictive power with high p-value. It showed weak association with the loan status based on descriptive analysis.

Predictive Analysis Report: Model Wise Results

1. **Logistic Regression**

Method:

Logistic regression was chosen to model the loan approval status using features such as loan amount, income, CIBIL score, education, and loan term. To evaluate the model's performance robustly, we used 10 fold cross-validation. Cross-validation involves splitting the dataset into several folds, training the model on a subset of the data, and testing it on the remaining fold. This process is repeated multiple times, and the results are averaged, ensuring the model's performance is consistent and not reliant on a particular split of the data. This approach also helps in minimizing overfitting and gives a better estimate of the model's generalizability.

Evaluation Metrics:

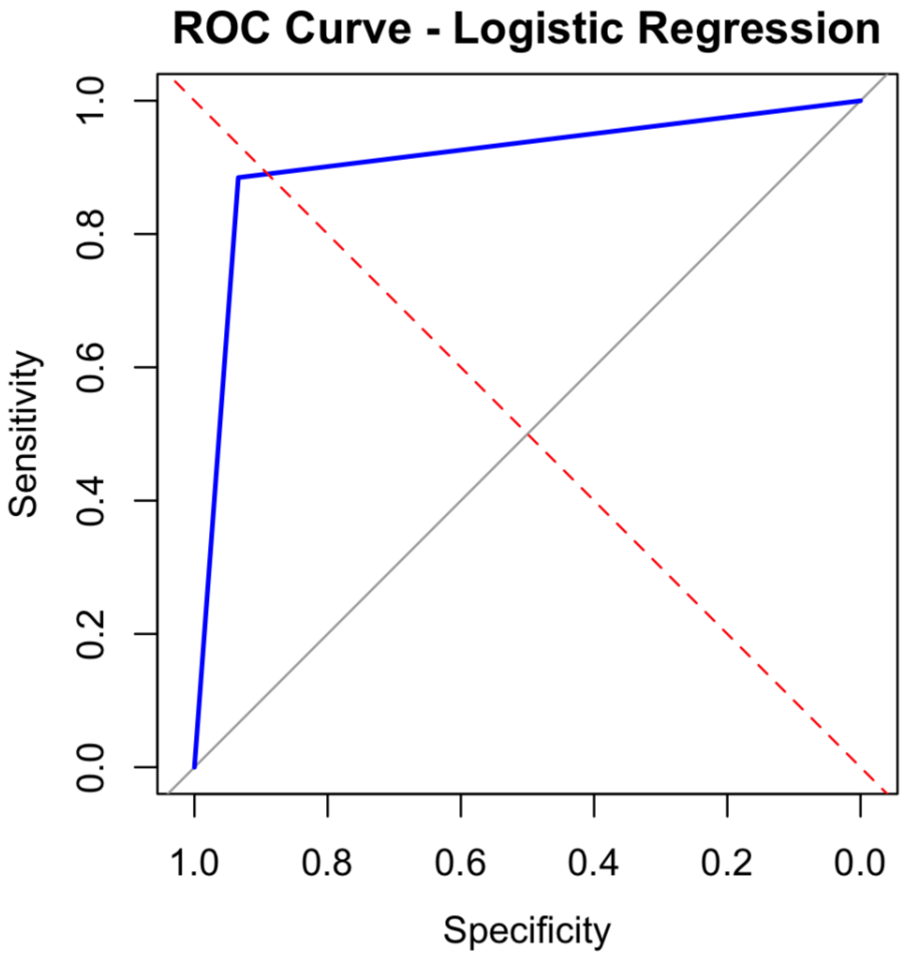
- **Accuracy:** 0.9154
- **Precision:** 0.9303

- **Recall:** 0.9341
- **F1 Score:** 0.9322

The accuracy of 0.9154 sindicates strong overall performance. The precision of 0.9303 shows that most predicted approved loans were correct, while the recall of 0.9341 reflects the model's effectiveness in capturing nearly all eligible applicants. The F1 score of 0.9322 highlights a well-balanced trade-off between precision and recall.

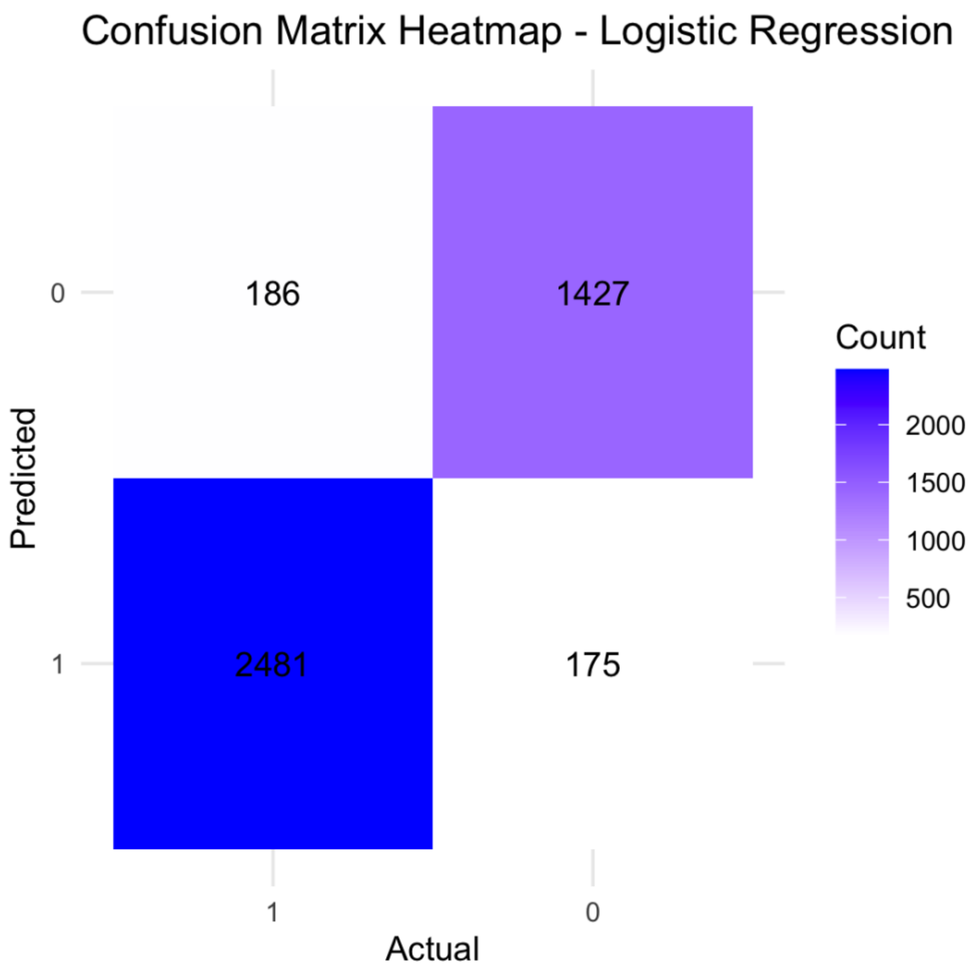
Motivation for Metrics:
Accuracy and F1 score were prioritized as they provide insights into the model’s ability to correctly classify loan approvals while maintaining a balance between precision (minimizing false positives) and recall (minimizing false negatives). These metrics give a comprehensive understanding of the model's performance across various dimensions.
Visualization of results:

1. **ROC Curve - Logistic Regression**
   The **Receiver Operating Characteristic (ROC) curve** visualizes the model's performance across different classification thresholds. The y-axis represents **sensitivity** (or recall), and the x-axis represents **1 - specificity** (false positive rate). The blue curve represents the trade-off between true positive rate and false positive rate for various thresholds.



- **Interpretation**: The curve is close to the top-left corner, which indicates that the model performs well in distinguishing between classes. The steeper the curve, the better the performance of the classifier.

2. **Confusion Matrix Heatmap - Logistic Regression**
   The confusion matrix heatmap shows the number of actual and predicted classifications for both classes (0 and 1). The matrix is color-coded, with darker colors representing higher counts.



- **Interpretation**:
    - **True Negatives (TN):** 1427 (Actual: 0, Predicted: 0) – The number of correctly predicted rejections.
    - **False Positives (FP):** 175 (Actual: 0, Predicted: 1) – The number of incorrect predictions where the model predicted approval, but the actual result was a rejection.
    - **False Negatives (FN):** 186 (Actual: 1, Predicted: 0) – The number of incorrect predictions where the model predicted rejection, but the actual result was an approval.
    - **True Positives (TP):** 2481 (Actual: 1, Predicted: 1) – The number of correctly predicted approvals.

This confusion matrix indicates that the model performs reasonably well but still misclassifies a few number of loan applications, particularly in the false negative and false positive cases.

Evaluation of Methods:

- **Bootstrap, Boosting, Random Forests:** We did not apply bootstrap or boosting in this model, as logistic regression is inherently suited for linear relationships, and more complex ensemble methods were not necessary for this specific dataset and goal.
- **Criteria for Choosing Metrics:** Metrics like accuracy, precision, recall, and F1 score were selected because they capture both correct and incorrect classifications, which is crucial in loan approval scenarios.

2. Decision Tree

Method:
The decision tree model was selected due to its simplicity, interpretability, and ability to capture non-linear relationships between features. Here as well 10 fold cross-validation was used to evaluate the model, ensuring that its performance was consistent across different subsets of the data. Cross-validation divides the data into several folds, allowing the model to be trained and tested on various subsets, which minimizes the risk of overfitting and provides a more reliable performance estimate. This method ensures that the model is robust and performs well across different segments of the dataset.

Evaluation Metrics:

- **Accuracy:** 0.9642
- **Precision:** 0.9937
- **Recall:** 0.9484
- **F1 Score:** 0.9705

The decision tree achieved a high accuracy of 0.9642 and an impressive precision of 0.9937, indicating a strong ability to correctly predict approved loans. Its recall of 0.9484 was slightly lower, meaning some eligible loans were missed, but the F1 score of 0.9705 shows a strong balance between precision and recall.

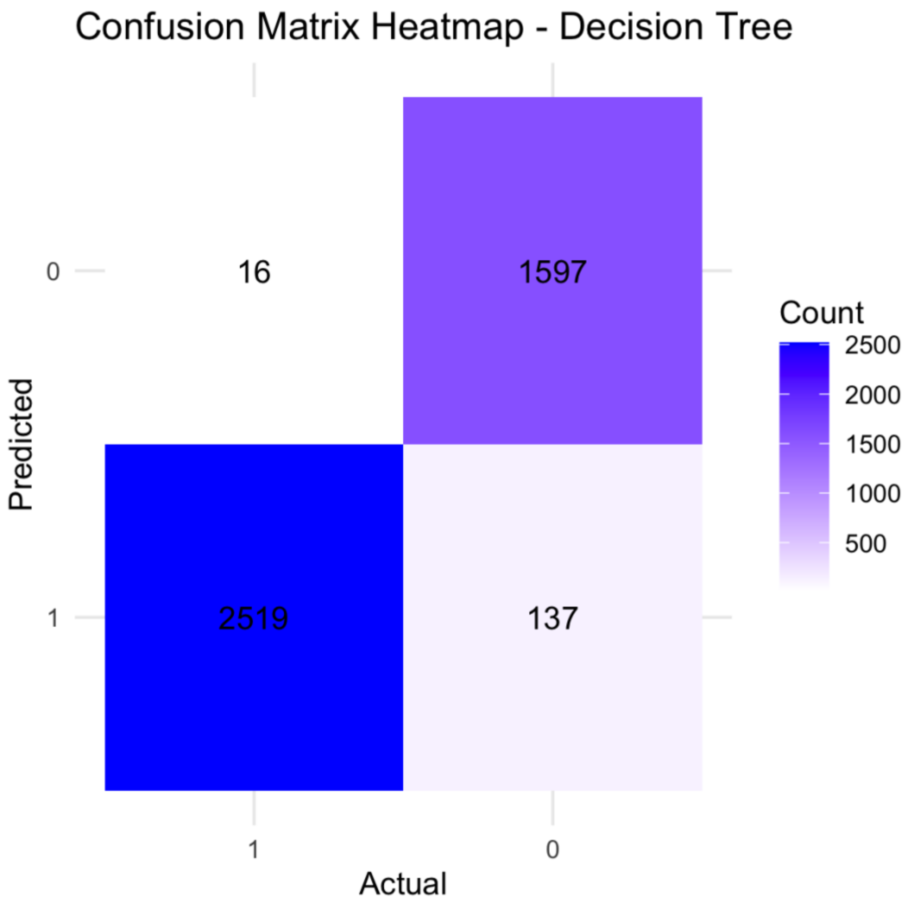
Motivation for Metrics:

We prioritized precision to minimize incorrect loan approvals, as incorrect approvals can have significant financial consequences. However, recall was also an important consideration, as missing eligible loan applicants can negatively impact loan approval processes.

Visualization of results:

1. **Confusion Matrix Heatmap - Decision Tree**

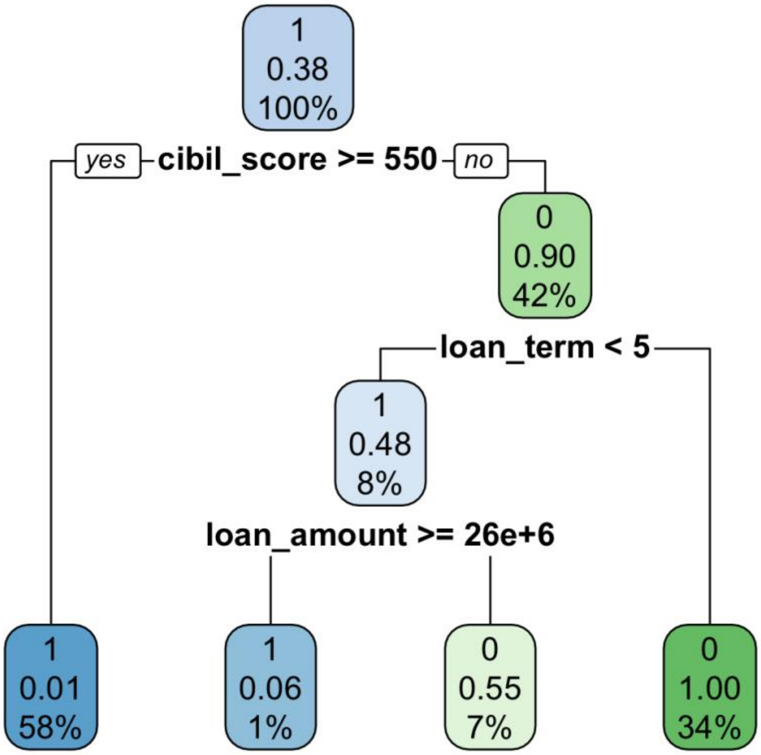
The confusion matrix heatmap shows the number of actual and predicted classifications for both classes (0 and 1). The matrix is color-coded, with darker colors representing higher counts.



- **Interpretation:**
  - **True Negatives (TN):** 1597 (Actual: 0, Predicted: 0) – The number of correctly predicted rejections.
  - **False Positives (FP):** 137 (Actual: 0, Predicted: 1) – The number of incorrect predictions where the model predicted approval, but the actual result was a rejection.
  - **False Negatives (FN):** 16 (Actual: 1, Predicted: 0) – The number of incorrect predictions where the model predicted rejection, but the actual result was an approval.
  - **True Positives (TP):** 2519 (Actual: 1, Predicted: 1) – The number of correctly predicted approvals.

This confusion matrix indicates that the decision tree model performed very well, with a low number of false negatives (16) and false positives (137). However, there are still some misclassifications, indicating room for improvement in predicting rejections.

2. **Decision Tree Plot**
   This plot visualizes the structure of the decision tree, showing the conditions for splitting at each level. The decision tree starts by checking if the **CIBIL score** is greater than or equal to 550. From there, the model checks additional features such as the **loan term** and **loan amount** to make further decisions.



- **Interpretation:**
  - If the CIBIL score is below 550, the model predicts a rejection with a probability of 90%.
  - If the CIBIL score is 550 or above, the model uses additional features like loan term and loan amount to make the final prediction.
  - This structure highlights how certain features are prioritized over others in making the loan approval decision, and the model provides a clear and interpretable decision-making process.

Evaluation of Methods:

- **Bootstrap, Boosting, Random Forests:**
  We did not apply bootstrap or boosting in the decision tree model. The decision tree was selected for its interpretability and simplicity, making more complex techniques unnecessary for this specific task. However, we did perform random

forest analysis as a more advanced ensemble technique to further evaluate model performance which will be discussed below.

- **Criteria for Choosing Metrics:**
  Precision, recall, and F1 score were emphasized as they provide valuable insight into both the accuracy of approvals and the model's ability to avoid false positives and negatives.

3. Random Forest

Method:
Random forest was selected for its ensemble learning capabilities, where multiple decision trees are combined to improve accuracy and generalization. Each tree is trained on a random subset of the data, reducing the risk of overfitting and capturing more complex patterns within the dataset. Cross-validation was used to ensure stable and robust performance across different folds, minimizing bias from any single data split.

Evaluation Metrics:

- **Accuracy:** 1.0000
- **Precision:** 1.0000
- **Recall:** 1.0000
- **F1 Score:** 1.0000

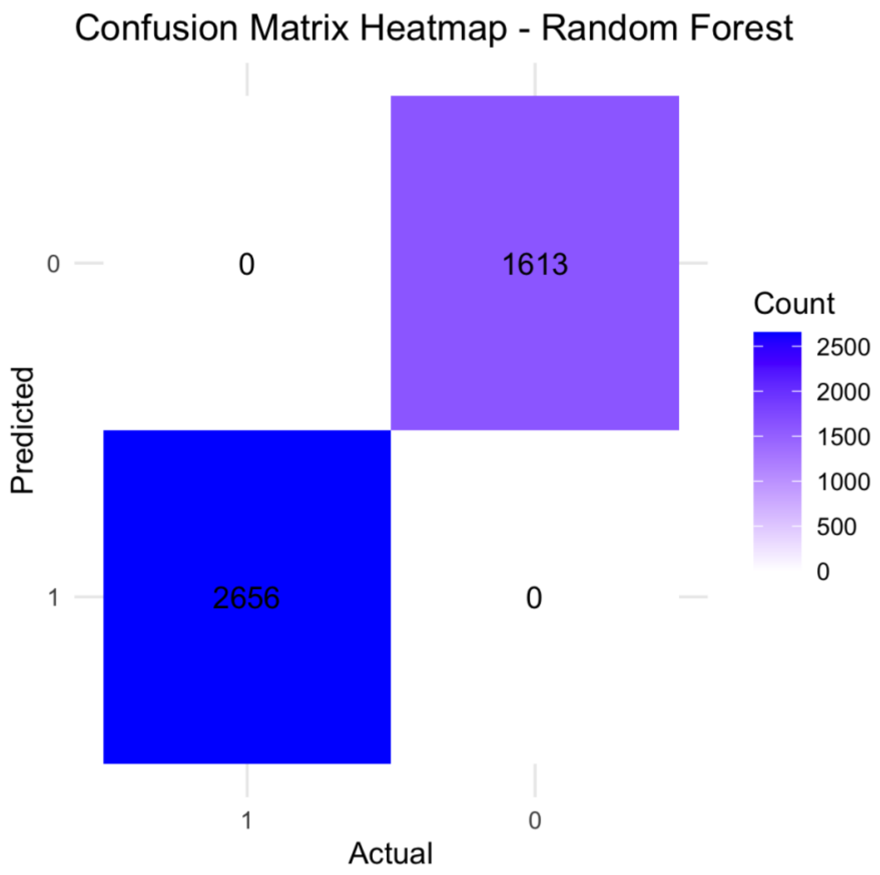
The random forest model achieved perfect scores across all metrics, with an accuracy, precision, recall, and F1 score of 1.0000. This indicates that the model classified all cases correctly, making it the best-performing model in terms of evaluation metrics.

Motivation for Metrics:

The perfect scores for all evaluation metrics reflect the random forest model's ability to handle high-dimensional data and complex interactions between features. Both precision and recall were given equal importance to ensure accurate loan approvals while avoiding false positives and false negatives.

Visualization of results:

1. **Confusion Matrix Heatmap - Random Forest**
   The confusion matrix heatmap shows the number of actual and predicted classifications for both classes (0 and 1). The matrix is color-coded, with darker colors representing higher counts.



- **Interpretation:**
  - **True Negatives (TN):** 1613 (Actual: 0, Predicted: 0) – The number of correctly predicted rejections.
  - **False Positives (FP):** 0 (Actual: 0, Predicted: 1) – The model made no false predictions where it predicted approval when it should have predicted rejection.
  - **False Negatives (FN):** 0 (Actual: 1, Predicted: 0) – The model made no false predictions where it predicted rejection when it should have predicted approval.

True Positives (TP): 2656 (Actual: 1, Predicted: 1) – The number of correctly predicted approvals.

The confusion matrix for the random forest model shows perfect classification performance with no false positives or false negatives. This indicates that the model has achieved excellent predictive power on this dataset. This makes random forest the most accurate model in this analysis, reinforcing its strength in ensemble learning by combining multiple decision trees.

Evaluation of Methods:

- **Bootstrap, Boosting, Random Forests:** Random forests were applied in this model due to their strength in handling large, complex datasets and their ensemble nature, which improves overall predictive performance.
- **Criteria for Choosing Metrics:** Since random forest is a more complex model, accuracy, precision, recall, and F1 score were crucial in evaluating its overall effectiveness in predicting loan approvals.

Comparison of Methods

Logistic regression, decision tree, and random forest were all evaluated using accuracy, precision, recall, and F1 score. While logistic regression provided a strong balance across all metrics, random forest outperformed the other models, achieving perfect scores. The decision tree offered a balance between interpretability and performance, although its recall was slightly lower. Random forest's ensemble approach made it the most powerful model in terms of predictive capability.

Prescriptive Goal

Based on the analysis, we will focus our prescriptive model on credit score and loan amount to optimize loan approval processes. Applicants with mid-level income ($50k–$100k) but strong credit scores should be prioritized for approval to maximize profitability while minimizing risk.

Optimize Loan Approval

Implementing prescriptive analytics to automate the decision-making process for borderline applicants which will reduce processing times and improve customer satisfaction.