Phase 6: Model Comparison - Logistic Regression vs. CART

Masheia Dzimba and Peter Mangoro

2025-12-06

Contents

1 Introduction **1**

2 Load Results **1**

3 Performance Metrics Comparison **1**

4 Visualization **2**

4.1 Performance Metrics Comparison 2

4.2 AUC Comparison 2

5 Interpretability Comparison **2**

6 Model Selection **4**

7 Key Findings **5**

8 Summary **5**

1 Introduction

This document presents Phase 6: Model Comparison. We compare the performance of Logistic Regression, CART, and Random Forest models, analyze their interpretability, and justify model selection.

2 Load Results

3 Performance Metrics Comparison

Table 1: Side-by-Side Performance Comparison

Metric	Logistic_Regression	CART	Random_Forest
Accuracy	61.84	60.78	61.46
Precision	64.09	58.71	60.67
Recall (Sensitivity)	42.05	55.83	51.21
Specificity	79.26	65.16	70.55
F1-Score	50.78	57.23	55.54
AUC	64.81	60.50	64.82

4 Visualization

4.1 Performance Metrics Comparison

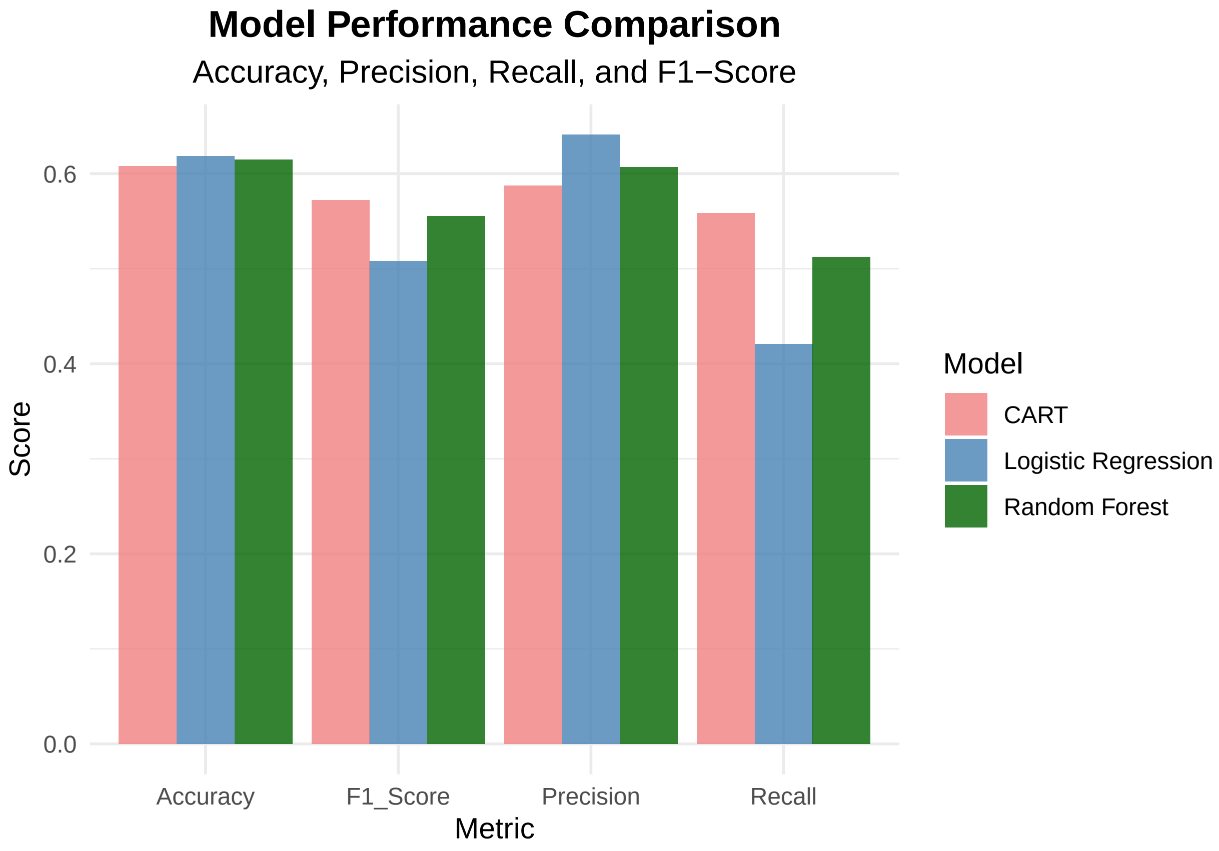

Figure 1: Model Performance Comparison

4.2 AUC Comparison

5 Interpretability Comparison

Logistic Regression:

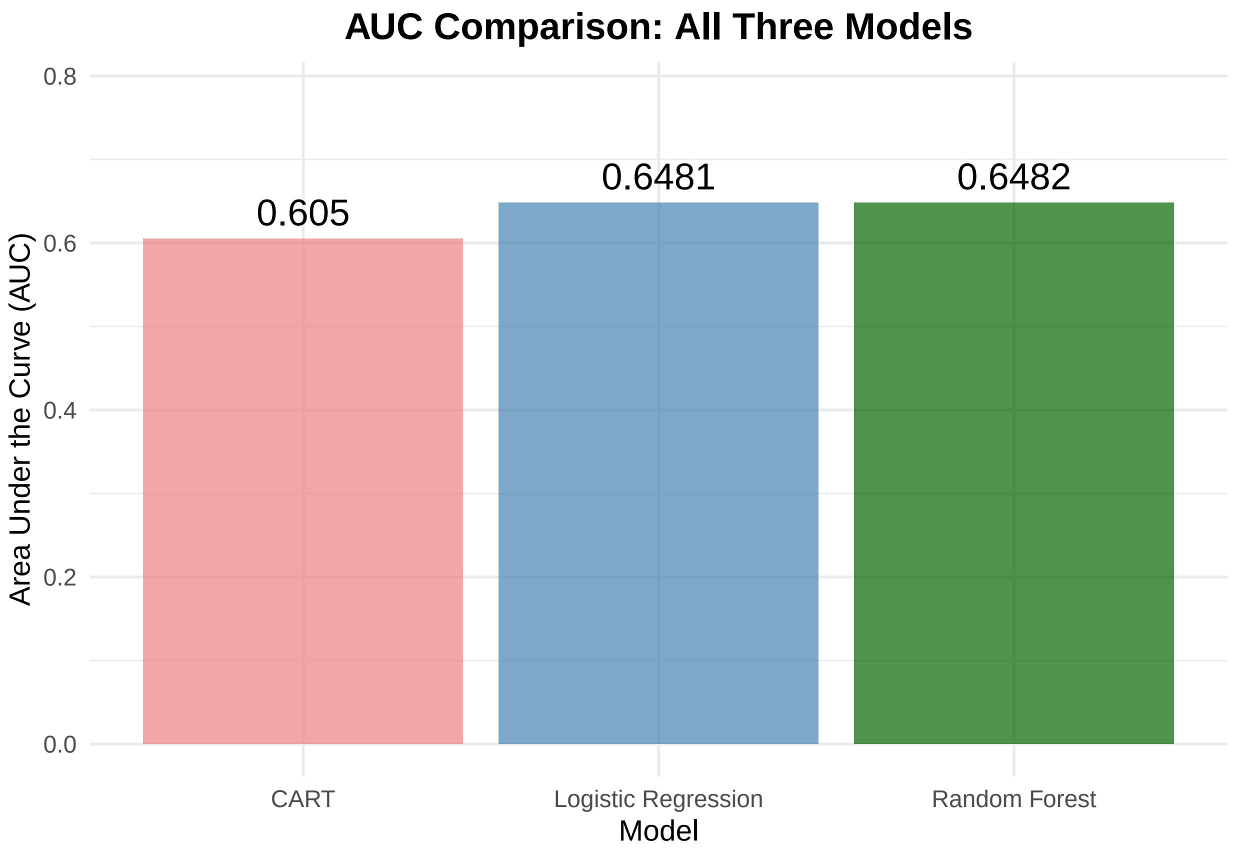

Figure 2: AUC Comparison: Logistic Regression vs. CART

- Provides coefficients and odds ratios for each variable
- Statistical significance testing (p-values)
- 29 parameters in the model
- 17 significant variables ($p < 0.05$)

CART:

- Simple decision tree with 6 variables considered
- Very interpretable: simple decision rule(s)
- Non-linear relationships captured
- Top variable: total_previous_visits (42.32% importance)

Random Forest:

- Ensemble of 500 decision trees
- Uses bootstrap sampling and feature randomization
- 17 variables considered
- Lower interpretability (ensemble effect)
- Non-linear relationships captured
- Top variable: n_lab_procedures (16.67% importance)

6 Model Selection

Performance Metrics Won: Logistic Regression: 3 metrics CART: 2 metrics Random Forest: 1 metrics

Recommended Model: Logistic Regression **Reason**: Best overall performance and statistical rigor

Table 2: Model Comparison: Detailed Justification

Criterion	Logistic_Regression	CART	Random_Forest
Accuracy	61.84%	60.78%	61.46%
AUC	64.81%	60.5%	64.82%
Precision	64.09%	58.71%	60.67%
Recall	42.05%	55.83%	51.21%
F1-Score	0.5078	0.5723	0.5554
Interpretability	High (coefficients, odds ratios)	Very High (simple tree, easy rules)	Low (ensemble of 500 trees)
Complexity	High (29 parameters)	Very Low (simple tree)	High (500 trees, complex ensemble)
Statistical Rigor	High (p-values, hypothesis tests)	Medium (no p-values, variable importance)	Medium (variable importance, no p-values)

7 Key Findings

1. **Performance:**
   - All three models show similar performance (accuracy ~61%)
   - Best AUC: Random Forest (0.648)
   - Logistic Regression: 0.648, CART: 0.605, Random Forest: 0.648
   - All models have fair to poor discrimination (AUC < 0.7)
2. **Interpretability:**
   - CART: Simplest (6 variables, single tree)
   - Logistic Regression: 29 parameters, detailed statistical insights
   - Random Forest: Most complex (500 trees, 17 variables)
   - All models identify similar key predictors
3. **Key Predictors:**
   - Logistic Regression: n_inpatient (OR: 1.47), age groups, medical specialty
   - CART: total_previous_visits (42.32% importance)
   - Random Forest: n_lab_procedures (16.67% importance)

8 Summary

This phase compared all three models:

- **Logistic Regression** wins 3 out of 6 performance metrics
- **CART** offers superior simplicity and interpretability
- **Random Forest** wins 1 out of 6 performance metrics
- **Recommended**: Logistic Regression (Best overall performance and statistical rigor)
- All models identify **total_previous_visits** as a key predictor