

Phase 5b: Random Forest Model

Masheia Dzimba and Peter Mangoro

2025-12-06

Contents

1	Introduction	1
2	Load Data	1
3	Train/Test Split	1
4	Build Random Forest Model	2
5	Variable Importance	2
6	Model Evaluation	3
7	ROC Curve	3
8	Summary	5

1 Introduction

This document presents Phase 5b: Random Forest Model. We build an ensemble decision tree model using Random Forest, analyze variable importance, and evaluate model performance.

2 Load Data

Dataset: 24996 observations, 18 variables

3 Train/Test Split

Training set: 17498 observations (70%)

Testing set: 7498 observations (30%)

4 Build Random Forest Model

```
ntree      OOB      1      2
 50:  41.85% 33.96% 50.75%
100:  41.07% 32.54% 50.69%
150:  41.08% 32.42% 50.84%
200:  40.90% 32.15% 50.76%
250:  40.52% 31.58% 50.59%
300:  40.66% 31.68% 50.77%
350:  40.42% 31.44% 50.53%
400:  40.43% 31.25% 50.78%
450:  40.39% 31.23% 50.72%
500:  40.33% 31.11% 50.72%
```

Random Forest model fitted successfully!

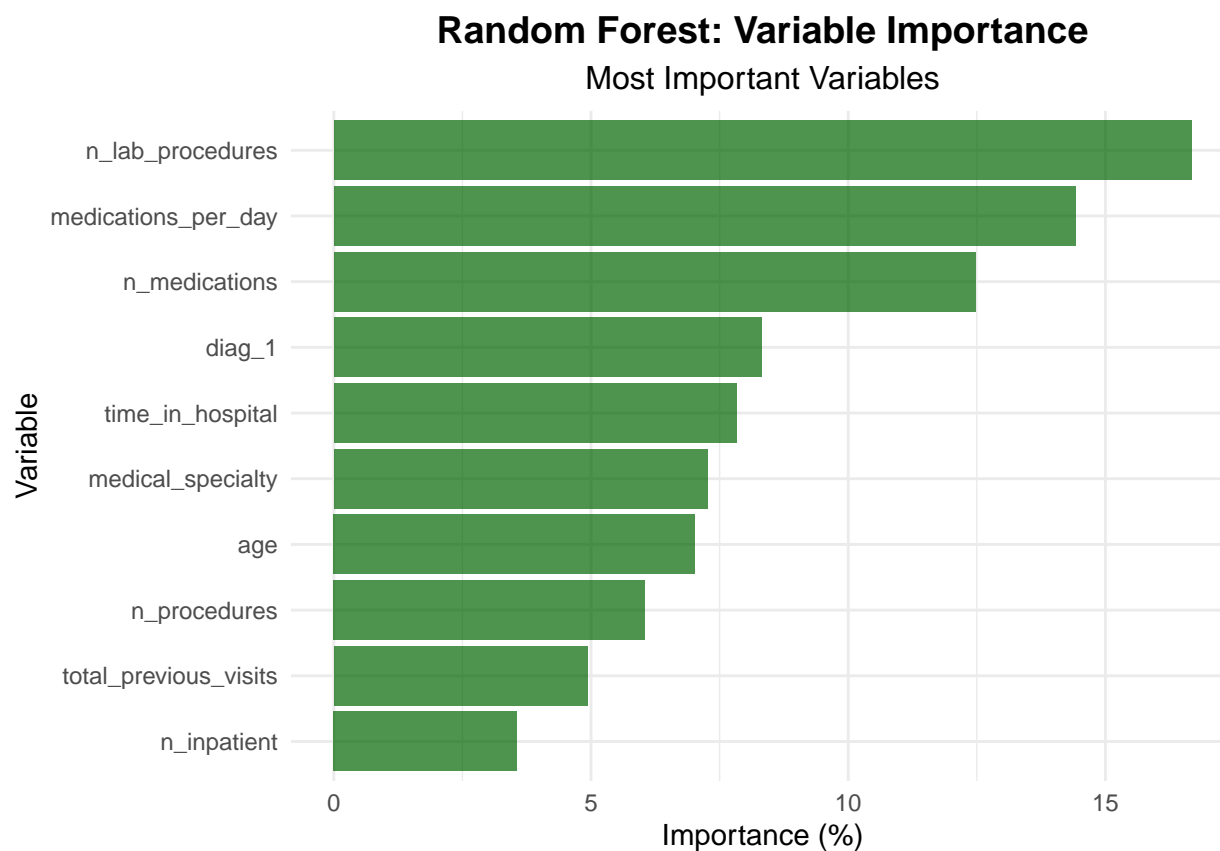
Number of trees: 500

Features per split (mtry): 4

5 Variable Importance

Table 1: Random Forest: Most Important Variables

Variable	MeanDecreaseGini	MeanDecreaseAccuracy	Importance_Percent
n_lab_procedures	1296.6613	8.444919	16.673656
medications_per_day	1121.3911	11.660038	14.419873
n_medications	969.7102	15.914413	12.469420
diag_1	646.2206	11.015168	8.309695
time_in_hospital	608.5389	13.999558	7.825149
medical_specialty	564.4217	7.138881	7.257850
age	545.8038	13.289090	7.018445
n_procedures	470.4693	12.364674	6.049724
total_previous_visits	383.1704	44.075157	4.927155
n_inpatient	276.9337	31.919932	3.561066



6 Model Evaluation

Table 2: Random Forest: Confusion Matrix

	Not_Readmitted	Readmitted
Not_Readmitted	2803	1720
Readmitted	1170	1805

Table 3: Random Forest: Performance Metrics

Metric	Value	Percentage
Accuracy	0.61	61.46
Precision	0.61	60.67
Recall (Sensitivity)	0.51	51.21
Specificity	0.71	70.55
F1-Score	0.56	55.54

7 ROC Curve

Area Under the Curve (AUC): 0.6482

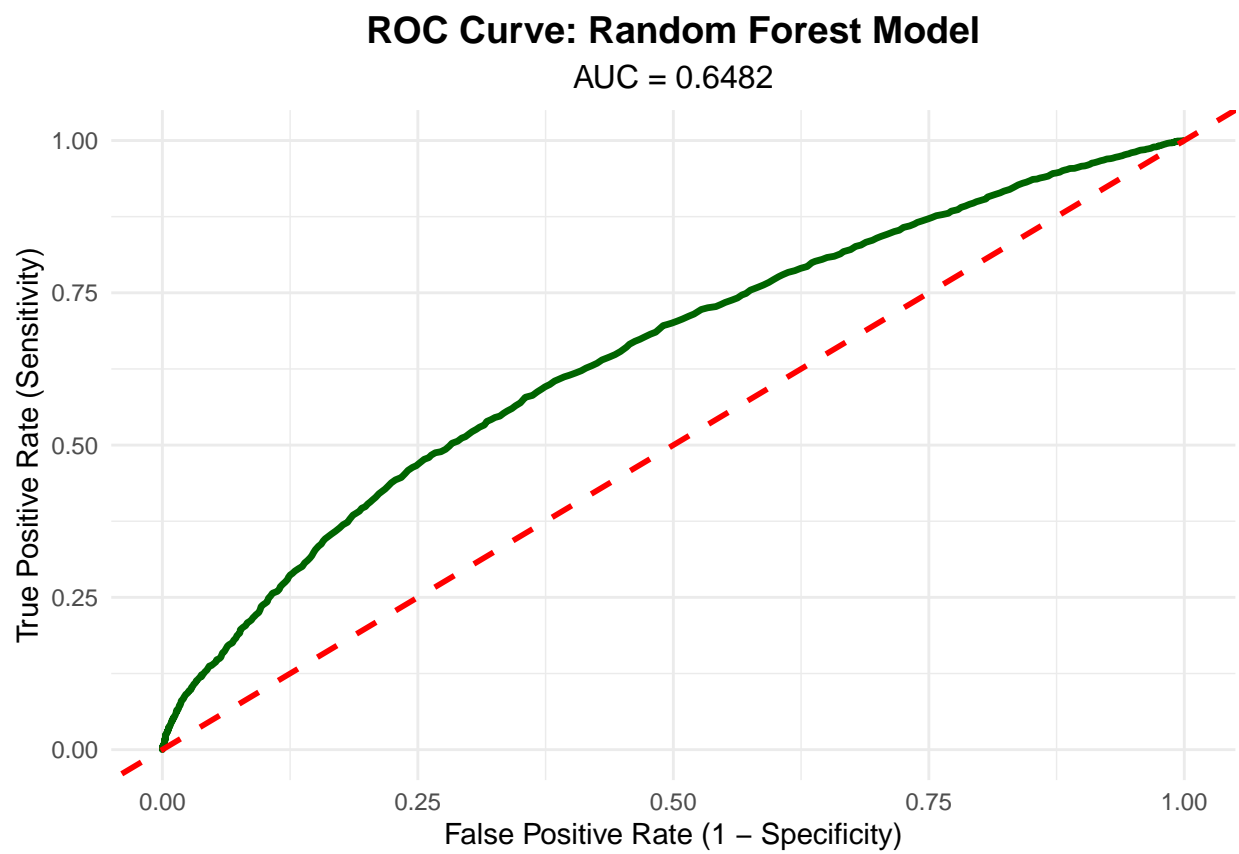


Figure 1: ROC Curve: Random Forest Model

8 Summary

This phase successfully built and evaluated the Random Forest model:

- **Accuracy:** 61.46%
- **AUC:** 0.648
- **Number of trees:** 500
- **Features per split (mtry):** 4
- **Top predictor:** n lab procedures (16.67% importance)
- **Interpretability:** Lower than CART (ensemble of 500 trees)