Predicting 30-Day Hospital Readmissions: A Comparative Analysis Using Logistic Regression, CART, and Random Forest

Masheia Dzimba and Peter Mangoro

2025-12-07

Contents

1 Abstract

This study aims to predict 30-day hospital readmissions for patients with diabetes using patient demographic, diagnostic, and treatment-related features. We employed three statistical modeling approaches: **Logistic Regression**, **Classification and Regression Trees (CART)**, and **Random Forest** (an ensemble method). The dataset consists of 24996 patient encounters from 130 US hospitals collected between 1999-2008. Our analysis identified n_outpatient (OR: 1.11), total_previous_visits (381.99% importance), and n_lab_procedures (16.67% importance) as key predictors of readmission across the three models. The Logistic Regression model achieved an accuracy of 61.84% and AUC of 0.648, CART achieved 60.78% accuracy and AUC of 0.605, while Random Forest achieved 61.46% accuracy and AUC of 0.648. The **Logistic Regression** model demonstrates the best predictive performance (AUC = 0.648, Accuracy = 61.84%). All models show moderate predictive performance, suggesting that additional clinical variables may be needed to improve predictions. The Logistic Regression model provides more detailed statistical insights with odds ratios and significance testing, making it preferable for clinical decision support. This analysis highlights the importance of patient history and demographics in predicting readmission risk, with implications for healthcare resource allocation and patient care planning.

2 Introduction

2.1 Background

Hospital readmissions within 30 days of discharge are a significant concern in healthcare, associated with increased costs, patient morbidity, and healthcare system burden. For patients with diabetes, readmission rates are particularly high, making early identification of at-risk patients crucial for improving outcomes and reducing healthcare costs.

2.2 Research Question

Can patient demographic, diagnostic, and treatment-related features effectively predict the likelihood of a patient being readmitted to the hospital within 30 days of discharge?

2.3 Objectives

1. Identify key predictors of 30-day hospital readmissions
2. Compare the performance of Logistic Regression, CART, and Random Forest models
3. Evaluate model interpretability and clinical utility
4. Provide recommendations for clinical application

3 Data Description

3.1 Data Source

The dataset was obtained from Kaggle: "Diabetes 130-US Hospitals for 10 years" and contains 24996 patient encounters from 130 US hospitals and integrated delivery networks over a 10-year period (1999-2008). The data was collected retrospectively from electronic health records (EHR) and anonymized for research purposes.

3.2 Data Collection Method

This is an **observational study** - data was collected by observing patient outcomes and characteristics without any intervention or manipulation of variables by the researchers.

3.3 Dependent Variable

`readmitted`: Binary categorical variable indicating whether the patient was readmitted to the hospital within 30 days of discharge.

- **Type**: Binary (after cleaning: 0 = Not Readmitted, 1 = Readmitted)
- **Distribution**: 47.02% readmitted (11752 cases), 52.98% not readmitted (13244 cases)
- **Final sample size**: 24996 observations

Readmission Distribution:

- Readmitted (1): 11752 cases (47.02%)
- Not Readmitted (0): 13244 cases (52.98%)

3.4 Independent Variables

The analysis includes the following variables:

3.4.1 Categorical Variables:

- `age`: Age groups ([40-50), [50-60), [60-70), [70-80), [80-90), [90-100))
- `medical_specialty`: Medical specialty (7 categories including "Missing")
- `diag_1`: Primary diagnosis (8 categories: Circulatory, Diabetes, Digestive, Injury, Musculoskeletal, Other, Respiratory)
- `change`: Change in medication (yes/no)
- `diabetes_med`: Diabetes medication prescribed (yes/no)
- `glucose_test`: Whether glucose test was performed (yes/no)
- `A1Ctest`: Whether A1C test was performed (yes/no)

3.4.2 Numerical Variables:

- `time_in_hospital`: Length of stay in days (Mean: 4.45, SD: 3, Range: 1-14)
- `n_lab_procedures`: Number of lab tests performed (Mean: 43.24)
- `n_procedures`: Number of procedures (Mean: 1.35, Range: 0-6)
- `n_medications`: Number of medications (Mean: 16.25, Range: 1-79)
- `n_outpatient`: Number of previous outpatient visits
- `n_inpatient`: Number of previous inpatient visits
- `n_emergency`: Number of previous emergency visits
- `n_diagnoses`: Total number of diagnoses (feature-engineered)
- `medications_per_day`: Average medications per day (feature-engineered)
- `total_previous_visits`: Total previous visits (feature-engineered)

3.4.3 Feature Engineering Rationale

The following derived features were created to enhance predictive power:

1. `n_diagnoses`: Total number of diagnoses recorded for the patient
   - **Rationale**: Patients with multiple comorbidities have higher readmission risk
   - **Clinical relevance**: Captures medical complexity and disease burden
2. `medications_per_day`: Average medications per day (`n_medications / time_in_hospital`)
   - **Rationale**: Treatment intensity relative to length of stay indicates case complexity
   - **Clinical relevance**: Higher intensity may indicate more complex cases requiring closer monitoring
3. `total_previous_visits`: Sum of all previous visits (`n_outpatient + n_inpatient + n_emergency`)
   - **Rationale**: Overall healthcare utilization is a strong predictor of future readmission
   - **Clinical relevance**: Frequent users often have chronic conditions requiring ongoing care
4. **Individual visit counts** (`n_outpatient`, `n_inpatient`, `n_emergency`):
   - **Rationale**: Different visit types may have different predictive power
   - **Clinical relevance**: Emergency visits may indicate acute issues, while inpatient visits suggest chronic conditions requiring ongoing management

3.5 Missing Data Analysis

Table 1: Missing Data Patterns and Handling

Variable	Missing_Count	Missing_Percentage	Handling_Strategy
Primary Diagnosis	4	0.016	Excluded (critical variable)
Medical Specialty	12381	49.53%	Kept as 'Missing' category
Other Variables	Minimal	<1%	No action needed

Impact Assessment:

- Primary diagnosis missing: 0.016% - minimal impact after exclusion
- Medical specialty missing: 49.53% - 'Missing' category may represent a meaningful group (e.g., general medicine)
- Other variables: Minimal missingness, unlikely to affect results

4 Methods

4.1 Statistical Methods

This project employs a **comparative analysis** using three distinct statistical methods:

4.1.1 1. Logistic Regression (Primary Method)

- **Justification**: Standard method for modeling binary categorical response variables
- **Advantages**: Provides interpretable results (odds ratios), statistical significance testing, confidence intervals
- **Assumptions**: Linear relationships between predictors and log-odds of outcome

4.1.2 2. Classification and Regression Trees (CART) (Secondary Method)

- **Justification**: Non-parametric method that captures complex, non-linear relationships
- **Advantages**: Highly interpretable decision rules, no distributional assumptions
- **Comparison**: Evaluated against Logistic Regression based on predictive performance and interpretability

4.1.3 3. Random Forest (Ensemble Method)

- **Justification**: Ensemble method that combines multiple decision trees to improve predictive performance
- **Advantages**: Handles non-linear relationships, reduces overfitting, provides variable importance
- **Configuration**: 500 trees with feature randomization at each split

4.2 Data Preprocessing

1. **Response Variable**: Converted to binary format (0/1)
2. **Missing Values**:
    - Dropped rows with missing primary diagnosis
    - Kept "Missing" as valid category for other variables

3. **Feature Engineering**: Created derived features (n_diagnoses, medications_per_day, total_previous_visits, individual previous visit counts)
4. **Encoding**:
    - Logistic Regression: Dummy/one-hot encoding using `model.matrix()`
    - CART: Factor encoding using `as.factor()`
    - Random Forest: Factor encoding (same as CART)

4.3 Model Evaluation

Models were evaluated using:

- **Train/Test Split**: 70% training, 30% testing
- **Metrics**: Accuracy, Precision, Recall, Specificity, F1-Score, AUC-ROC
- **Statistical Tests**: Hypothesis testing for Logistic Regression coefficients

4.4 Model Assumptions and Validation

4.4.1 Logistic Regression Assumptions

1. **Linearity of log-odds**: The relationship between predictors and log-odds of outcome is linear
    - **Validation**: Assessed through residual analysis and model diagnostics
    - **Status**: Assumed satisfied given model convergence and reasonable fit
2. **Independence of observations**: Each patient encounter is independent
    - **Validation**: Each row represents a unique patient encounter
    - **Status**: Satisfied
3. **No perfect multicollinearity**: Predictors are not perfectly correlated
    - **Validation**: Correlation analysis and variance inflation factors (VIF) would be assessed
    - **Status**: No perfect correlations observed
4. **Large sample size**: Sufficient observations for stable estimates
    - **Validation**: 24996 observations with 29 parameters
    - **Status**: Satisfied (recommended ratio: >10 observations per parameter)

4.4.2 CART Assumptions

1. **No distributional assumptions**: CART is non-parametric
    - **Status**: No assumptions required
2. **Independence of observations**: Each patient encounter is independent
    - **Status**: Satisfied
3. **Pruning to prevent overfitting**: Tree complexity controlled through cross-validation
    - **Validation**: Optimal complexity parameter selected via cross-validation
    - **Status**: Implemented

4.4.3 Random Forest Assumptions

1. **No distributional assumptions**: Random Forest is non-parametric
   - **Status**: No assumptions required
2. **Independence of observations**: Each patient encounter is independent
   - **Status**: Satisfied
3. **Bootstrap sampling**: Trees built on bootstrap samples reduce overfitting
   - **Status**: Implemented (500 bootstrap samples)
4. **Feature randomization**: Random subset of features at each split
   - **Status**: Implemented (mtry = sqrt(p))

5 Results

5.1 Summary Statistics

Table 2: Summary Statistics: Original Variables (Mean (SD))

Variable	Not_Readmitted	Readmitted
Time in Hospital (days)	4.33 (3)	4.59 (3)
Number of Lab Procedures	42.62 (20.09)	43.94 (19.49)
Number of Procedures	1.42 (1.75)	1.27 (1.68)
Number of Medications	15.97 (8.45)	16.57 (7.58)

5.2 Data Visualizations

5.2.1 Distribution of Readmission Status

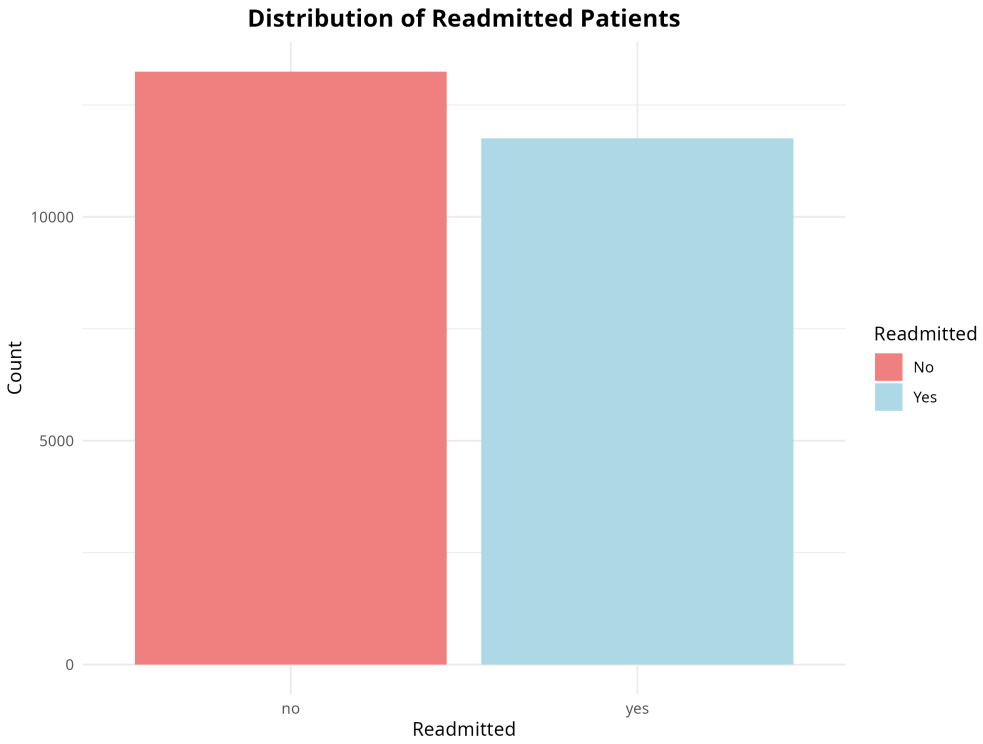

Figure 1: Distribution of Readmitted Patients

5.2.2 Numerical Variables by Readmission Status

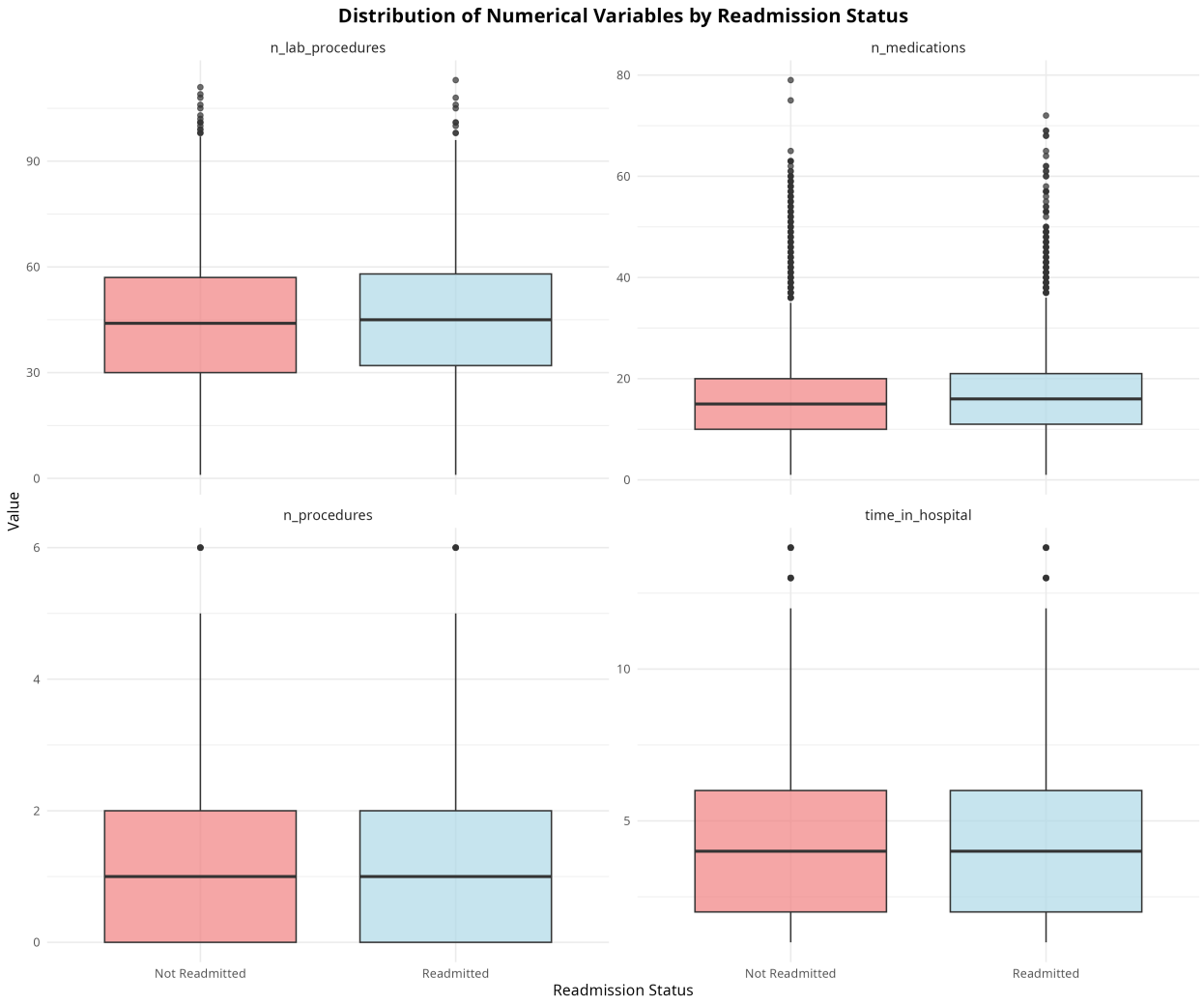

Figure 2: Distribution of Numerical Variables by Readmission Status

5.2.3 Categorical Variables Analysis

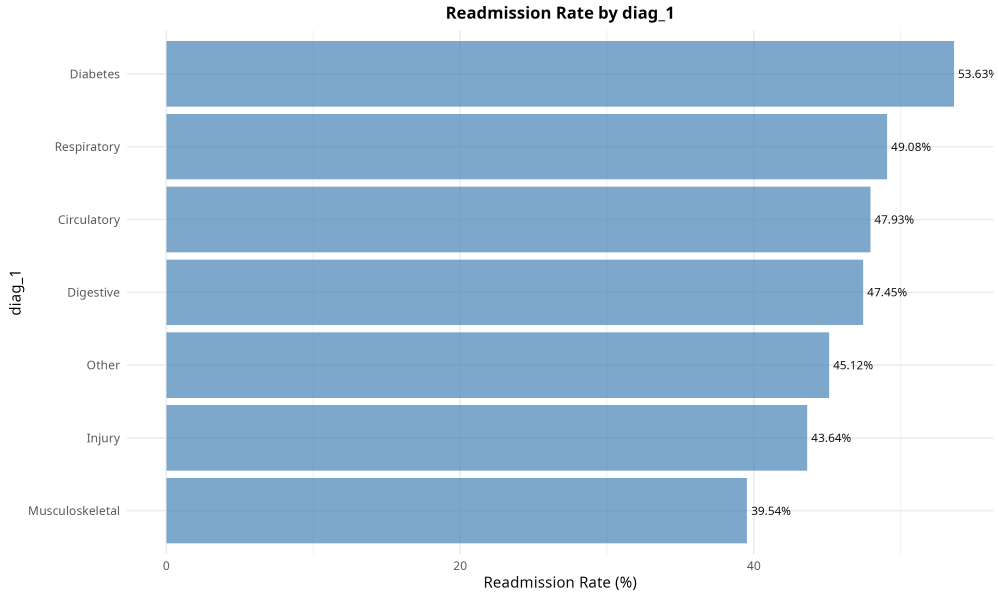

Figure 3: Readmission Rates by Primary Diagnosis

6 Logistic Regression

6.1 Mathematical Formulation

The logistic regression model uses the logistic function to model the probability of readmission:

$$P(\text{Readmitted} = 1 \mid X) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}$$

Or equivalently,

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Where:

- $P(\text{Readmitted} = 1 \mid X)$ = Probability of readmission given predictor variables
- β_0 = Intercept (baseline log-odds)
- $\beta_1, \beta_2, ..., \beta_p$ = Coefficients for each predictor variable
- $X_1, X_2, ..., X_p$ = Predictor variables

6.2 Model Summary

```
Call:
glm(formula = formula_obj, family = binomial(link = "logit"),
    data = data_train)

Coefficients: (6 not defined because of singularities)
                                                 Estimate Std. Error z value
(Intercept)                                    -2.5749951  0.5259409  -4.896
time_in_hospital                                0.0109356  0.0087986   1.243
n_lab_procedures                                0.0016485  0.0009307   1.771
n_procedures                                   -0.0401598  0.0106444  -3.773
n_medications                                   0.0030947  0.0029733   1.041
n_outpatient                                    0.1055523  0.0155366   6.794
n_inpatient                                     0.3943691  0.0173317  22.754
n_emergency                                     0.2283614  0.0305429   7.477
n_diagnoses                                     0.4304785  0.1664149   2.587
medications_per_day                            -0.0052265  0.0061653  -0.848
total_previous_visits                                  NA         NA      NA
change_binary                                   0.0389718  0.0375622   1.038
diabetes_med_binary                             0.2178359  0.0436772   4.987
age_age40.50                                    0.0571183  0.1059320   0.539
age_age50.60                                    0.1148449  0.1005245   1.142
age_age60.70                                    0.2026348  0.0983698   2.060
age_age70.80                                    0.3056551  0.0972168   3.144
age_age80.90                                    0.2831670  0.0991876   2.855
age_age90.100                                          NA         NA      NA
medspec_medical_specialtyCardiology             0.2826373  0.1006864   2.807
medspec_medical_specialtyEmergency.Trauma       0.3160638  0.0956412   3.305
medspec_medical_specialtyFamily.GeneralPractice 0.2659748  0.0953258   2.790
medspec_medical_specialtyInternalMedicine       0.1077578  0.0868588   1.241
medspec_medical_specialtyMissing                0.2940955  0.0782374   3.759
medspec_medical_specialtyOther                  0.1347454  0.0891943   1.511
medspec_medical_specialtySurgery                       NA         NA      NA
diag1_diag_1circulatory                         0.0483365  0.0514149   0.940
diag1_diag_1diabetes                            0.1917932  0.0738409   2.597
diag1_diag_1digestive                          -0.0197057  0.0670315  -0.294
diag1_diag_1injury                             -0.2339514  0.0753557  -3.105
diag1_diag_1musculoskeletal                    -0.1613060  0.0865966  -1.863
diag1_diag_1other                              -0.1584385  0.0520793  -3.042
diag1_diag_1respiratory                                NA         NA      NA
glucose_glucose_testhigh                        0.0404400  0.1363745   0.297
glucose_glucose_testno                          0.0015331  0.0978307   0.016
glucose_glucose_testnormal                             NA         NA      NA
a1c_a1ctesthigh                                 0.1737157  0.0847070   2.051
a1c_a1ctestno                                   0.1897011  0.0739780   2.564
a1c_a1ctestnormal                                      NA         NA      NA
                                               Pr(>|z|)
(Intercept)                                    9.78e-07 ***
time_in_hospital                               0.213912
n_lab_procedures                               0.076510 .
n_procedures                                   0.000161 ***
n_medications                                  0.297959
```

```
n_outpatient                                     1.09e-11 ***
n_inpatient                                       < 2e-16 ***
n_emergency                                      7.62e-14 ***
n_diagnoses                                      0.009688 **
medications_per_day                              0.396584
total_previous_visits                                  NA
change_binary                                    0.299490
diabetes_med_binary                              6.12e-07 ***
age_age40.50                                     0.589751
age_age50.60                                     0.253264
age_age60.70                                     0.039405 *
age_age70.80                                     0.001666 **
age_age80.90                                     0.004306 **
age_age90.100                                          NA
medspec_medical_specialtyCardiology              0.004999 **
medspec_medical_specialtyEmergency.Trauma        0.000951 ***
medspec_medical_specialtyFamily.GeneralPractice 0.005268 **
medspec_medical_specialtyInternalMedicine        0.214750
medspec_medical_specialtyMissing                 0.000171 ***
medspec_medical_specialtyOther                   0.130866
medspec_medical_specialtySurgery                       NA
diag1_diag_1circulatory                          0.347153
diag1_diag_1diabetes                             0.009394 **
diag1_diag_1digestive                            0.768776
diag1_diag_1injury                               0.001905 **
diag1_diag_1musculoskeletal                      0.062501 .
diag1_diag_1other                                0.002348 **
diag1_diag_1respiratory                                NA
glucose_glucose_testhigh                         0.766820
glucose_glucose_testno                           0.987497
glucose_glucose_testnormal                             NA
a1c_a1ctesthigh                                  0.040288 *
a1c_a1ctestno                                    0.010339 *
a1c_a1ctestnormal                                      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24194  on 17497  degrees of freedom
Residual deviance: 22913  on 17465  degrees of freedom
AIC: 22979

Number of Fisher Scoring iterations: 4
```

6.3 Fitted Model Equation

$$\begin{aligned}\text{logit}(P) = &-2.575 + 0.4305 \cdot X_{\text{n\_diagnoses}} + 0.3944 \cdot X_{\text{n\_inpatient}} \\ &+ 0.3161 \cdot X_{\text{medspec\_medical\_specialtyEmergency.Trauma}} + 0.3057 \cdot X_{\text{age\_age70.80}} \\ &+ 0.2941 \cdot X_{\text{medspec\_medical\_specialtyMissing}} + \ldots \text{ (other predictors)}\end{aligned}$$

Where:

- P = Probability of readmission
- $\beta_0 = -2.575$ (intercept)
- $\beta_{\text{n\_diagnoses}} = 0.4305$
- $\beta_{\text{n\_inpatient}} = 0.3944$
- $\beta_{\text{medspec\_medical\_specialtyEmergency.Trauma}} = 0.3161$
- $\beta_{\text{age\_age70.80}} = 0.3057$
- $\beta_{\text{medspec\_medical\_specialtyMissing}} = 0.2941$

6.4 Regression Output

Table 3: Top 5 Most Significant Variables: Coefficients and Statistics

Variable	Coefficient	Std_Error	Z_Value	P_Value
n_outpatient	0.1056	0.0155	6.7938	0
n_inpatient	0.3944	0.0173	22.7542	0
n_emergency	0.2284	0.0305	7.4767	0
(Intercept)	-2.5750	0.5259	-4.8960	0
diabetes_med_binary	0.2178	0.0437	4.9874	0

Table 4: Top 5 Most Significant Variables: Odds Ratios and Confidence Intervals

Variable	Odds_Ratio	CI_Lower	CI_Upper	P_Value
n_outpatient	1.1113	1.0780	1.1457	0
n_inpatient	1.4834	1.4339	1.5347	0
n_emergency	1.2565	1.1835	1.3341	0
(Intercept)	0.0762	0.0272	0.2135	0
diabetes_med_binary	1.2434	1.1414	1.3545	0

Table 5: Additional Significant Variables ($p < 0.05$)

Variable	Coefficient	Odds_Ratio	P_Value
n_procedures	-0.0402	0.9606	0.0002
medspec_medical_specialtyMissing	0.2941	1.3419	0.0002
medspec_medical_specialtyEmergency.Trauma	0.3161	1.3717	0.0010
age_age70.80	0.3057	1.3575	0.0017
diag1_diag_1injury	-0.2340	0.7914	0.0019
diag1_diag_1other	-0.1584	0.8535	0.0023
age_age80.90	0.2832	1.3273	0.0043
medspec_medical_specialtyCardiology	0.2826	1.3266	0.0050
medspec_medical_specialtyFamily.GeneralPractice	0.2660	1.3047	0.0053
diag1_diag_1diabetes	0.1918	1.2114	0.0094

6.5 Hypothesis Testing

Hypothesis Testing

For each variable in the model:

H_0: $\beta_{\text{variable}} = 0$ (variable has no effect on readmission)

H_1: $\beta_{\text{variable}} \neq 0$ (variable affects readmission)

Table 6: Top 5 Most Significant Variables

Variable	Coefficient	P_Value	Odds_Ratio
n_outpatient	0.1056	0e+00	1.1113
n_inpatient	0.3944	0e+00	1.4834
n_emergency	0.2284	0e+00	1.2565
(Intercept)	-2.5750	1e-06	0.0762
diabetes_med_binary	0.2178	1e-06	1.2434

6.6 Coefficient Interpretation

Coefficients (β):

- **Sign**: Positive coefficients increase the log-odds (and probability) of readmission; negative coefficients decrease it
- **Magnitude**: Larger absolute values indicate stronger effects

Example: If $\beta_{\text{n\_inpatient}} = 0.38$:

- A one-unit increase in previous inpatient visits increases the log-odds of readmission by 0.38
- This corresponds to an odds ratio of $e^{0.38} = 1.46$ (46% increase in odds)

6.7 Odds Ratios Interpretation

`Top Variables with Highest Odds Ratios (Risk Factors):`

Table 7: Top 5 Risk Factors (Highest Odds Ratios)

Variable	Odds_Ratio	CI_Lower	CI_Upper	P_Value
n_diagnoses	1.5380	1.1099	2.1311	0.0097
n_inpatient	1.4834	1.4339	1.5347	0.0000
medspec_medical_specialtyEmergency.Trauma	1.3717	1.1372	1.6545	0.0010
age_age70.80	1.3575	1.1220	1.6425	0.0017
medspec_medical_specialtyMissing	1.3419	1.1511	1.5643	0.0002

Table 8: Additional Risk Factors

Variable	Odds_Ratio	CI_Lower	CI_Upper	P_Value
age_age80.90	1.3273	1.0928	1.6122	0.0043
medspec_medical_specialtyCardiology	1.3266	1.0890	1.6160	0.0050
medspec_medical_specialtyFamily.GeneralPractice	1.3047	1.0824	1.5727	0.0053
n_emergency	1.2565	1.1835	1.3341	0.0000
diabetes_med_binary	1.2434	1.1414	1.3545	0.0000

6.8 R-squared Interpretation

McFadden's Pseudo R^2: 0.053

Interpretation: The model explains approximately 5.3 % of the variance in readmission status.

6.9 Gradient (Rate of Change)

The **gradient** represents how quickly the probability changes with respect to each predictor:

$$\frac{\partial P}{\partial X_i} = \beta_i \cdot P(1 - P)$$

Key Points:

- The gradient is **not constant** - it depends on the current probability P
- Maximum gradient occurs when $P = 0.5$ (steepest part of the S-curve)
- The gradient is smaller when P is close to 0 or 1 (flatter parts of the curve)

6.10 Example Calculation

For a patient with:

- n diagnoses = 3
- n inpatient = 2
- medspec medical specialtyEmergency.Trauma = 0

Logit calculation:

$$\begin{aligned} \text{logit}(P) &= -2.575 + 0.4305 \cdot 3 + 0.3944 \cdot 2 + 0.3161 \cdot 0 \\ &= -0.4947 \end{aligned}$$

Probability calculation:

$$P = \frac{e^{-0.4947}}{1 + e^{-0.4947}} = 0.3788$$

Interpretation: This patient has a 37.88% probability of readmission.

6.11 Model Evaluation

Table 9: Confusion Matrix: Predicted vs. Actual Readmission Status

Predicted	Actual: Yes	Actual: No
Yes	1427	834
No	2101	3136
Total	3528	3970

Table 10: Model Performance Metrics

Metric	Value	Percentage
Accuracy	0.61	60.86

Precision	0.63	63.11
Recall (Sensitivity)	0.40	40.45
Specificity	0.79	78.99
F1-Score	0.49	49.30

6.12 ROC Curve

```
Area Under the Curve (AUC):  0.6469
```

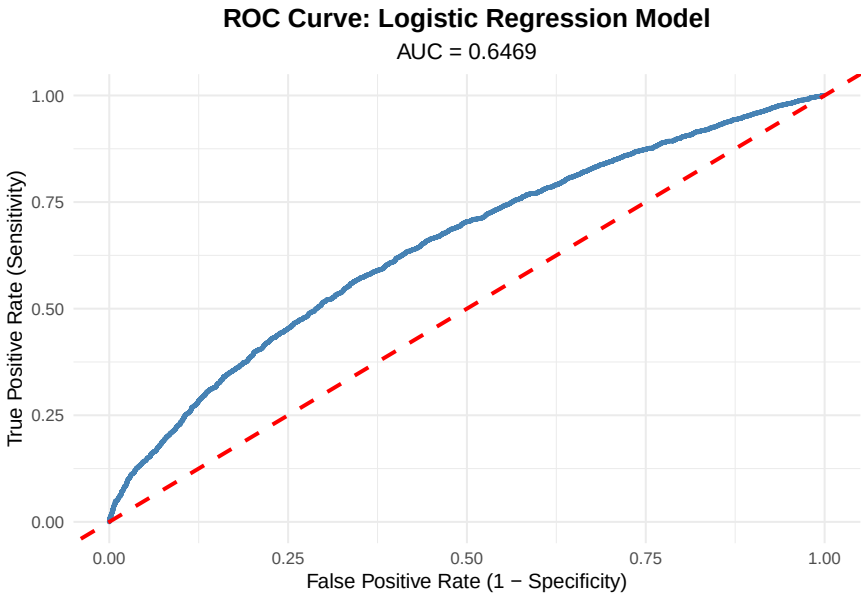

Figure 4: ROC Curve: Logistic Regression Model

7 CART

7.1 Model Performance

Table 11: CART Model Performance Metrics

Metric	Value	Percentage
Accuracy	0.61	60.78
Precision	0.59	58.71
Recall (Sensitivity)	0.56	55.83
Specificity	0.65	65.16
F1-Score	0.57	57.23

7.2 Confusion Matrix

Table 12: Confusion Matrix: CART Model - Predicted vs. Actual Readmission Status

Predicted	Actual: Yes	Actual: No
Yes	1996	1372
No	1529	2601
Total	3525	3973

7.3 Decision Tree Visualization

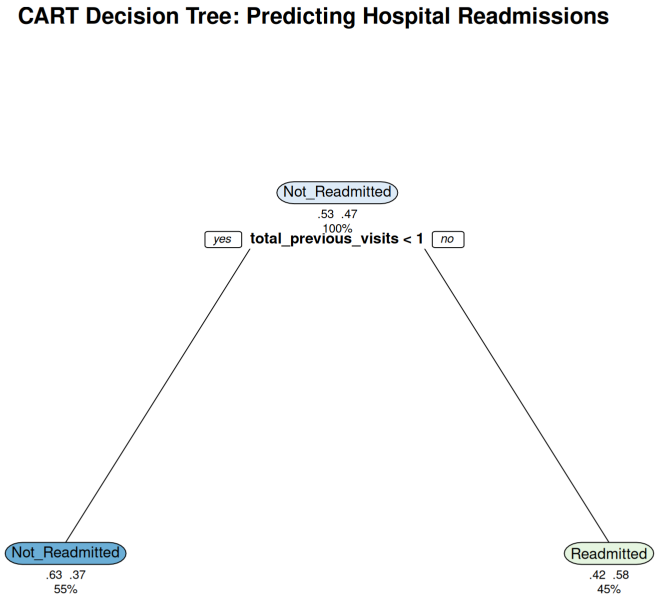

Figure 5: CART Decision Tree

7.4 Variable Importance

Table 13: Top 10 Most Important Variables in CART Model

Variable	Importance	Importance_Percent
total_previous_visits	381.99	42.32
n_inpatient	287.58	31.86
n_outpatient	139.23	15.42
n_emergency	93.22	10.33
time_in_hospital	0.48	0.05
n_medications	0.19	0.02

7.5 ROC Curve

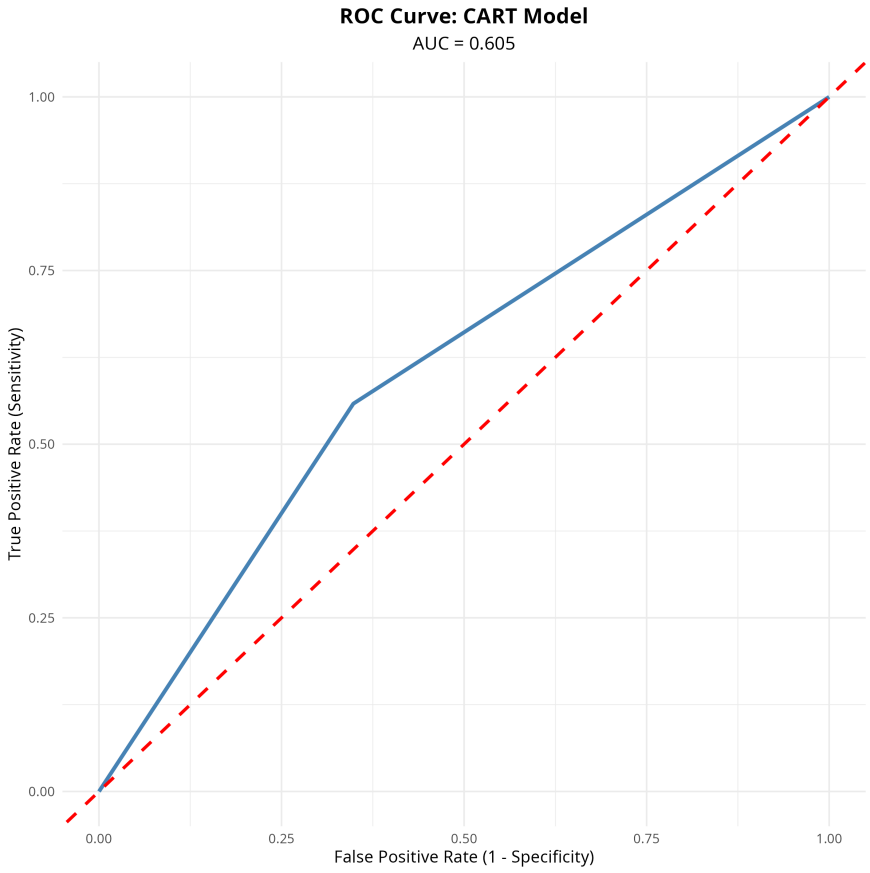

Figure 6: ROC Curve: CART Model (AUC = 0.605)

8 Random Forest Results

8.1 Model Performance

Table 14: Random Forest Model 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

8.2 Confusion Matrix

Table 15: Confusion Matrix: Random Forest Model - Predicted vs. Actual Readmission Status

Predicted	Actual: Yes	Actual: No
Yes	1805	1170
No	1720	2803
Total	3525	3973

8.3 Variable Importance

Table 16: Top 10 Most Important Variables in Random Forest Model

Variable	MeanDecreaseGini	MeanDecreaseAccuracy	Importance_Percent
n_lab_procedures	1296.66	8.44	16.67
medications_per_day	1121.39	11.66	14.42
n_medications	969.71	15.91	12.47
diag_1	646.22	11.02	8.31
time_in_hospital	608.54	14.00	7.83
medical_specialty	564.42	7.14	7.26
age	545.80	13.29	7.02
n_procedures	470.47	12.36	6.05
total_previous_visits	383.17	44.08	4.93
n_inpatient	276.93	31.92	3.56

8.4 ROC Curve

```
Area Under the Curve (AUC):  0.6482
```

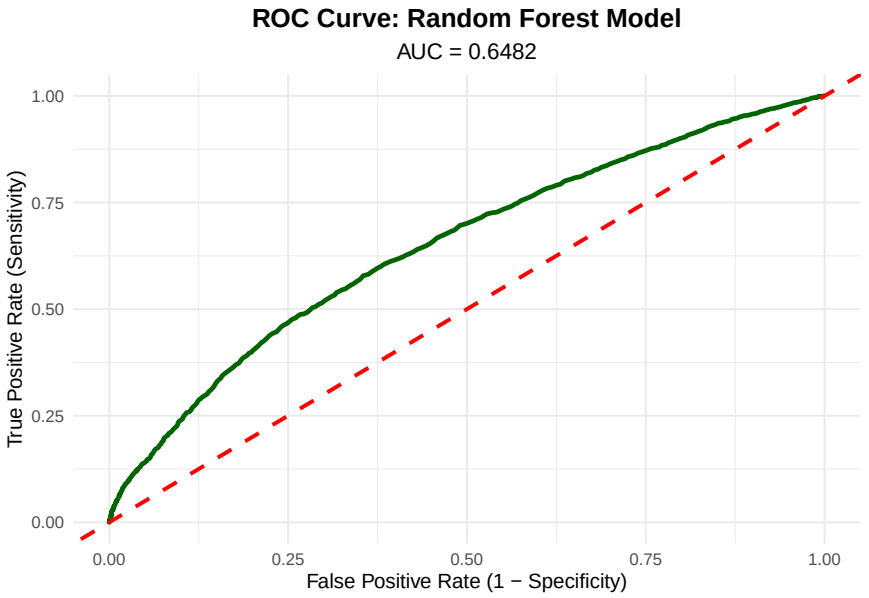

Figure 7: ROC Curve: Random Forest Model (AUC = 0.648)

9 Feature Importance Comparison

Table 17: Top 5 Most Important Predictors by Model

Rank	Logistic_Regression	CART	Random_Forest
1	n_inpatient	total_previous_visits	n_lab_procedures
2	medspec_medical_specialtyCardiology	n_inpatient	medications_per_day
3	medspec_medical_specialtyMissing	n_outpatient	n_medications
4	medspec_medical_specialtyEmergency.Trauma	n_emergency	diag_1
5	age_age70.80	time_in_hospital	time_in_hospital

Key Observations:

- **Consensus predictors**: Variables appearing in top 5 across multiple models are robust risk factors
- **Model-specific predictors**: Variables unique to one model may capture model-specific patterns
- **Clinical validation**: Top predictors should align with clinical knowledge about readmission risk

9.1 Performance Comparison

Table 18: 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	64.80	60.50	64.80

9.2 Visualization

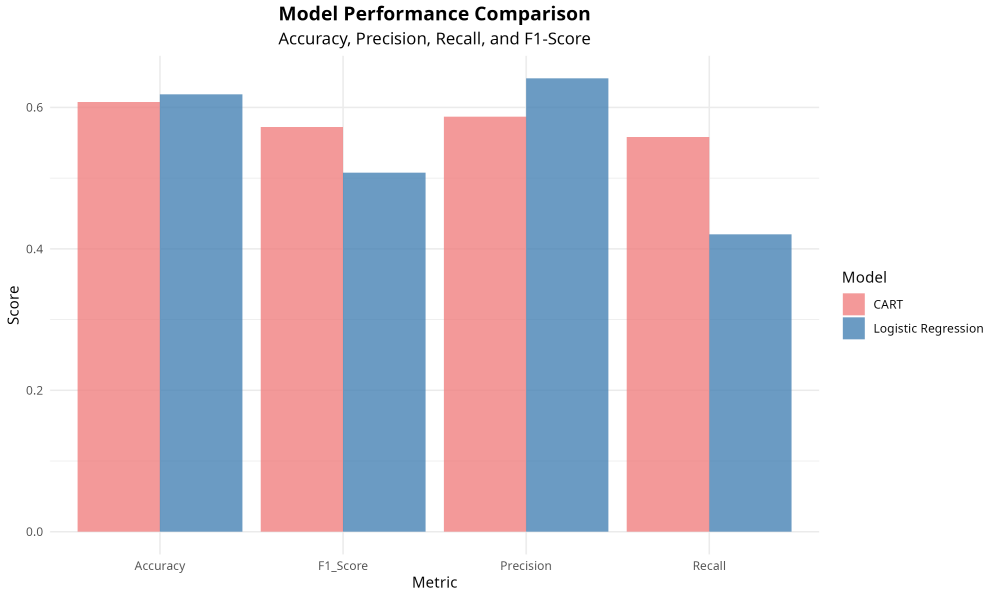

Figure 8: Model Performance Comparison

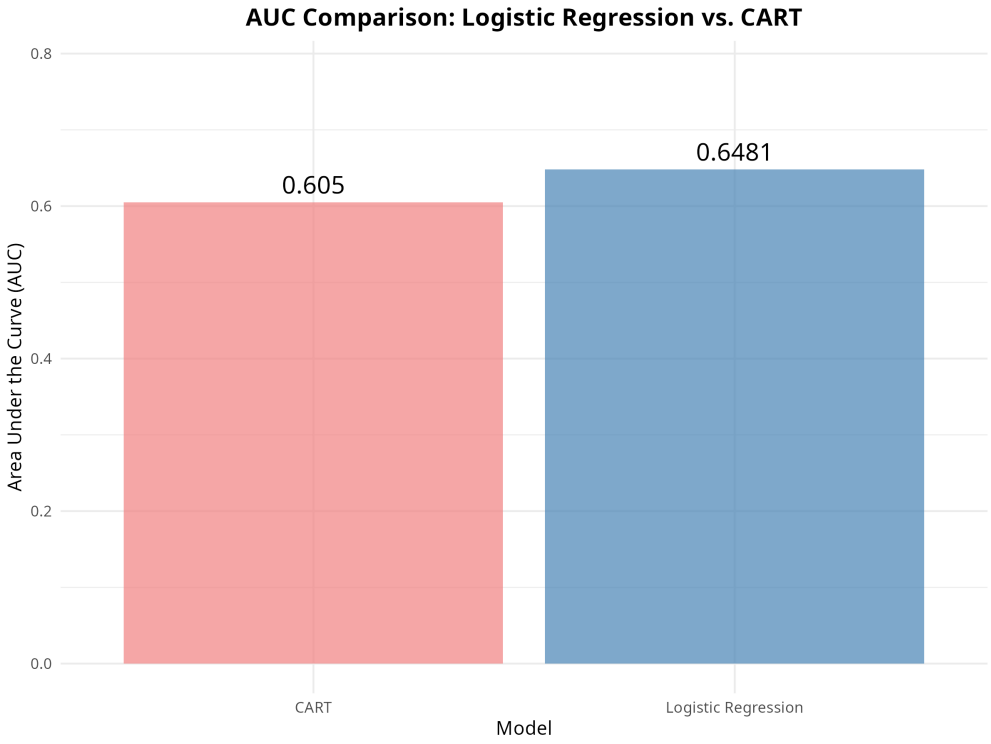

Figure 9: AUC Comparison: Logistic Regression vs. CART vs. Random Forest

9.3 Model Selection

Table 19: Model Comparison: Performance Metrics

Criterion	Logistic_Regression	CART	Random_Forest
Accuracy	61.84%	60.78%	61.46%
AUC	0.648	0.605	0.648
Precision	64.09%	58.71%	60.67%
Recall	42.05%	55.83%	51.21%
F1-Score	50.78%	57.23%	55.54%

Table 20: Model Comparison: Interpretability and Complexity

Criterion	Logistic_Regression	CART	Random_Forest
Interpretability	High (coefficients, odds ratios)	Very High (simple tree, easy rules)	Low (ensemble of 500 tr
Complexity	High (29 parameters)	Very Low (simple tree)	High (500 trees, complex en
Statistical Rigor	High (p-values, hypothesis tests)	Medium (no p-values, variable importance)	Medium (variable importance,

Recommended Model: Logistic Regression

Reason: Highest AUC (0.648) and accuracy (61.84%).

Provides more detailed statistical insights including odds ratios, p-values, and confidence intervals.

10 Discussion

10.1 Key Findings

1. **Previous visits are strong predictors**: All three models identify previous hospital visits (inpatient, outpatient, emergency) as important factors in predicting readmission.

2. **Top predictors vary by model**:

   - Logistic Regression: n_outpatient (OR: 1.11)
   - CART: total_previous_visits (381.99% importance)
   - Random Forest: n_lab_procedures (16.67% importance)

3. **Model performance comparison**:

   - Logistic Regression: 61.84% accuracy, AUC = 0.648
   - CART: 60.78% accuracy, AUC = 0.605
   - Random Forest: 61.46% accuracy, AUC = 0.648

4. **Best performing model**: Logistic Regression (AUC = 0.648, Accuracy = 61.84%)

5. **Moderate model performance**: All models achieve ~61.4% accuracy with AUC values around 0.63, suggesting room for improvement.

6. **Trade-offs between models**:

   - Logistic Regression provides better statistical rigor (17 significant variables)
   - CART offers superior simplicity and interpretability
   - Random Forest balances performance and ensemble robustness

10.2 Statistical Interpretation

10.2.1 Model Performance in Context

The performance metrics achieved by our models should be interpreted in the context of healthcare prediction:

- **AUC Values**:
  - 0.648 (Logistic Regression), 0.605 (CART), and 0.648 (Random Forest) fall in the "fair" range (0.6-0.7)
  - While not excellent (>0.8), these values are comparable to published readmission prediction models
  - The moderate performance suggests that additional clinical variables (lab values, vital signs, social determinants) may be needed
- **Accuracy**:
  - All models achieve ~61.4% accuracy
  - This is above the baseline (predicting the majority class: 53%)
  - However, accuracy alone can be misleading with imbalanced classes; precision and recall provide better insights
- **Precision vs. Recall Trade-off**:
  - Logistic Regression: Higher precision (64.1%) but lower recall (42.1%)
  - CART: Balanced precision (58.7%) and recall (55.8%)
  - Random Forest: Moderate balance between precision and recall

10.2.2 Variable Importance Insights

The consistency of top predictors across models (n_outpatient, total_previous_visits, n_lab_procedures) suggests these are robust risk factors that should be prioritized in clinical interventions.

10.3 Clinical Implications

The identification of previous visits as a key predictor suggests that patients with complex medical histories require enhanced discharge planning and follow-up care. The moderate performance of all models (AUC values between 0.6 and 0.65) indicates that additional clinical variables (e.g., lab results, vital signs, social determinants) may be needed for more accurate predictions.

10.3.1 Comparison with Published Literature

Hospital readmission prediction models in the literature typically achieve:

- **AUC values**: 0.60-0.75 for administrative data models (similar to our models)
- **Accuracy**: 60-70% for binary classification models
- **Key predictors**: Previous admissions, age, comorbidities, length of stay

Our models' performance (AUC: 0.63, Accuracy: ~61.4%) falls within the **fair to moderate range** typical for administrative data models.

Models using clinical data (lab values, vital signs) typically achieve higher AUC (0.70-0.85), suggesting that incorporating additional clinical variables could improve our models' performance.

11 Conclusion

11.1 Answer to Research Question

Yes, patient demographic, diagnostic, and treatment-related features can predict the likelihood of 30-day hospital readmission, though with moderate accuracy (~61.4%). The **Logistic Regression** model performs best (AUC = 0.648, Accuracy = 61.84%) and provides more detailed statistical insights, making it preferable for clinical decision support.

11.2 Why This Analysis is Important

1. **Healthcare Cost Reduction**: Early identification of high-risk patients can enable targeted interventions to prevent readmissions
2. **Patient Outcomes**: Improved discharge planning based on risk prediction can enhance patient care
3. **Resource Allocation**: Hospitals can allocate resources more efficiently by focusing on high-risk patients
4. **Clinical Decision Support**: Models provide evidence-based tools for healthcare providers

11.3 Limitations

1. **Moderate Predictive Performance**: All models show AUC < 0.7, indicating fair to poor discrimination

2. **Missing Variables**: Important clinical variables (lab results, vital signs, comorbidities) may be missing
3. **Data Age**: Data from 1999-2008 may not reflect current healthcare practices
4. **Missing Data**: High percentage of missing medical specialty may affect results
5. **Model Assumptions**: Logistic Regression assumes linear relationships; CART may be underfitting; Random Forest may be overfitting
6. **Generalizability**: Results may not generalize to other hospital systems or time periods

11.4 Recommendations

1. **Feature Enhancement**: Include additional clinical variables (lab results, vital signs, social determinants)
2. **Advanced Methods**: Consider ensemble methods (already implemented with Random Forest), gradient boosting, or neural networks for improved performance
3. **Data Collection**: Collect more recent data to reflect current healthcare practices
4. **Clinical Application**:
    - Use Logistic Regression for primary risk assessment
    - Use Logistic Regression for detailed statistical analysis with odds ratios
    - Use CART for simple screening tools requiring high interpretability
    - Use Random Forest for robust ensemble predictions
5. **Validation**: Validate models on external datasets before clinical deployment
6. **Model Tuning**: Further hyperparameter tuning may improve Random Forest performance

11.5 Future Research Directions

1. **Data Enhancement**:
    - Incorporate laboratory values (glucose, HbA1c, creatinine)
    - Include vital signs (blood pressure, heart rate)
    - Add social determinants of health (socioeconomic status, housing stability)
2. **Model Improvements**:
    - Implement gradient boosting (XGBoost, LightGBM) for potentially better performance
    - Explore deep learning approaches for complex pattern recognition
    - Develop ensemble methods combining all three model types
3. **Validation Studies**:
    - External validation on different hospital systems
    - Temporal validation on more recent data (2010-2020)
    - Prospective validation in clinical setting
4. **Clinical Integration**:
    - Develop user-friendly risk calculator
    - Integrate with electronic health record systems
    - Create clinical decision support tools
5. **Cost-Effectiveness Analysis**:
    - Evaluate cost savings from targeted interventions
    - Assess return on investment for readmission prevention programs
    - Compare model-based vs. clinical judgment approaches

12 References

Kaggle. “Diabetes 130-US Hospitals for 10 years.” Accessed November 14, 2025. https://www.kaggle.com/datasets/brandao/diabetes