Predicting Hospital Readmissions Using Machine Learning

Team 11
Radhika Govindarajan & Hannah Doerr
MSCHA ‘24, MSIT-BIDA ‘24
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Carnegie Mellon University
Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Introduction

The primary objective of this project is to leverage machine learning techniques to develop a predictive model for hospital readmission among patients with chronic conditions.

- What are the most common health conditions that result in a readmission ?
- Effect of specific chronic conditions on readmission
- Enable providers to determine the group of individuals to be focused on to avoid readmissions

Key Findings

- **Logistic Regression trained on the SMOTE-balanced dataset** was concluded as the preferred model
  - Demonstrated the best performance, achieving a good balance between precision (60.7%) and recall (74.4%) for predicting readmissions

- **Feature Importance Insights**
  - 'n_inpatient,' 'n_medications,' 'n_lab_procedures,' 'time_in_hospital,' and 'Age' were observed to be the most determinant features of hospital readmission

Value Proposition

Adds substantial value to healthcare

- Enhanced Patient Outcomes
  - Early Identification
  - Targeted Intervention
- Financial Impact
  - Reducing Healthcare Costs
  - Alignment with Hospital Readmissions Reduction Program (HRRP)
- Advancing Healthcare Analytics

Motivation

Dual challenges of avoidable hospital readmissions

- Adverse effect on patient well-being
- Economic burden faced by healthcare providers and payers

Improve patient management using predictive models

- Optimized treatment plans
- Post-discharge follow-ups
- Patient education

Financial considerations

Efficient resource allocations

Problem Statement

$P(T, E+\Delta E) > P(T,E)$

Task (T): Determine the risk of readmission to a hospital after a patient is initially hospitalized

Experience (E): Improve upon prior readmission risk scores, such as the HOSPITAL score

Performance (P): Precision and Recall

Dataset

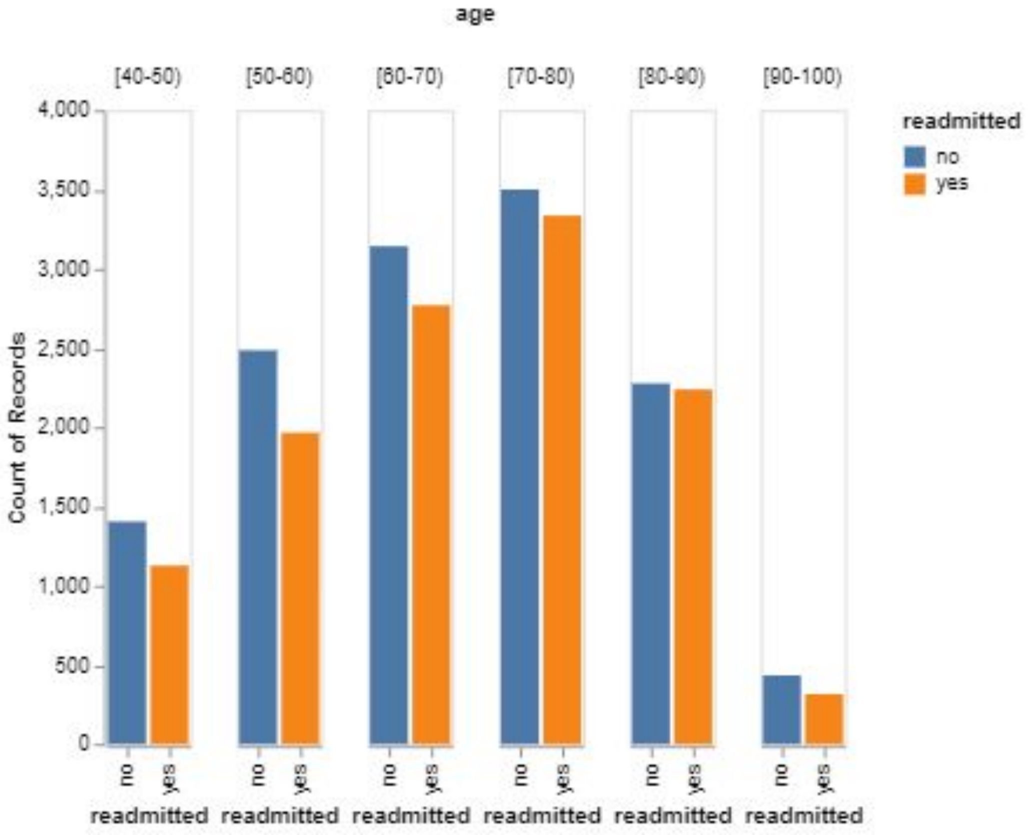
- Hospital Readmission Dataset - 10 year history of hospital readmission data, delineated by various measures of diabetes diagnosis

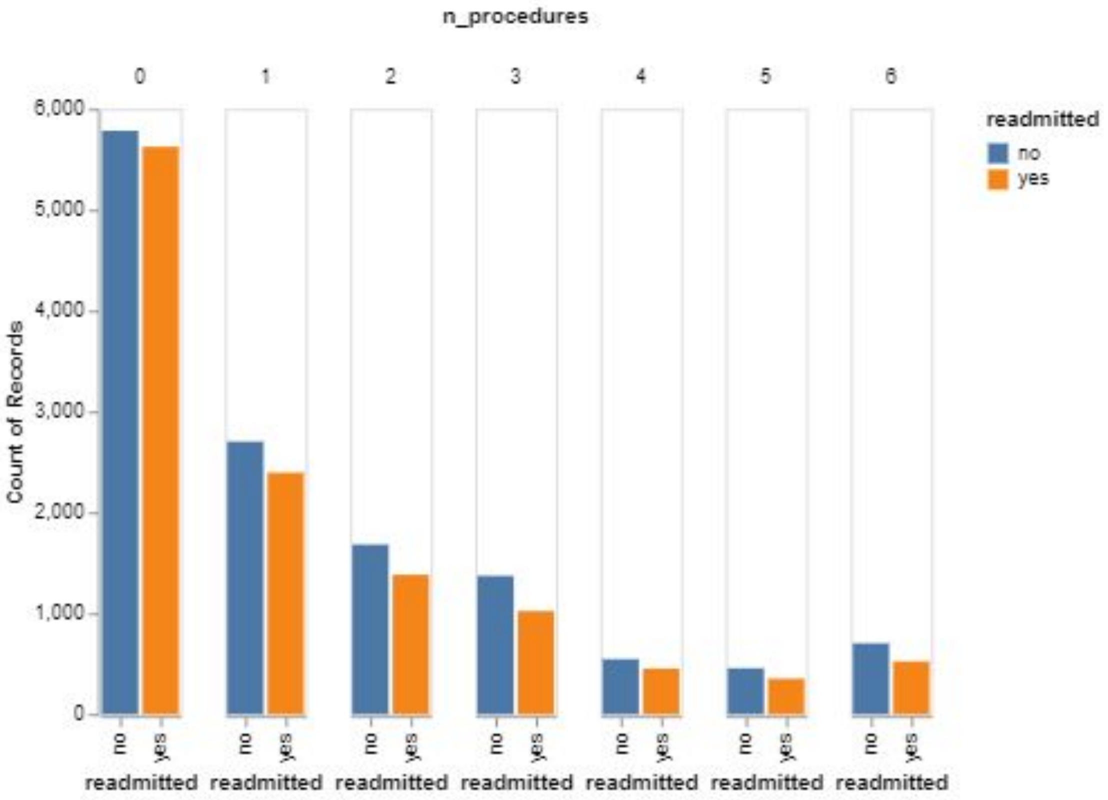
- Source: Kaggle
  - 25000 records, 17 columns
  - Fields are numerical and categorical
  - No duplicate rows
  - No sparse columns
  - No outliers

Dataset

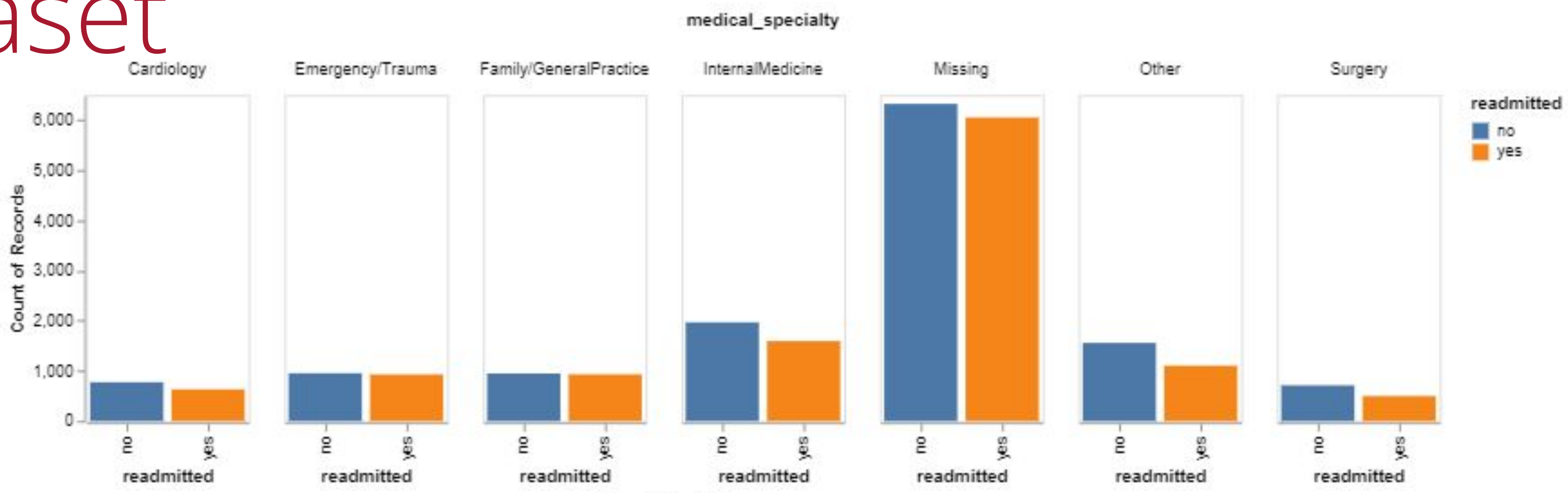
age: The age of the patient (non-null, object type).
time_in_hospital: The duration of the patient's stay in the hospital (non-null, integer type).
n_lab_procedures: The number of laboratory procedures performed for the patient (non-null, integer type).
n_procedures: The number of additional medical procedures performed (non-null, integer type).
n_medications: The number of distinct medications administered to the patient (non-null, integer type).
n_outpatient: The number of outpatient visits by the patient (non-null, integer type).
n_inpatient: The number of inpatient visits by the patient (non-null, integer type).
n_emergency: The number of emergency room visits by the patient (non-null, integer type).
medical_specialty: The medical specialty of the admitting physician (non-null, object type).
diag_1, diag_2, diag_3: Primary, secondary, and tertiary diagnoses for the patient (non-null, object type).
glucose_test: Whether the patient had a glucose test (non-null, object type).
A1Ctest: Whether the patient had an A1C test (non-null, object type).
change: Indicates if there was a change in the patient's medications (non-null, object type).
diabetes_med: Indicates if the patient is on diabetes medication (non-null, object type).
readmitted: The target variable indicating if the patient was readmitted (non-null, object type).

Dataset

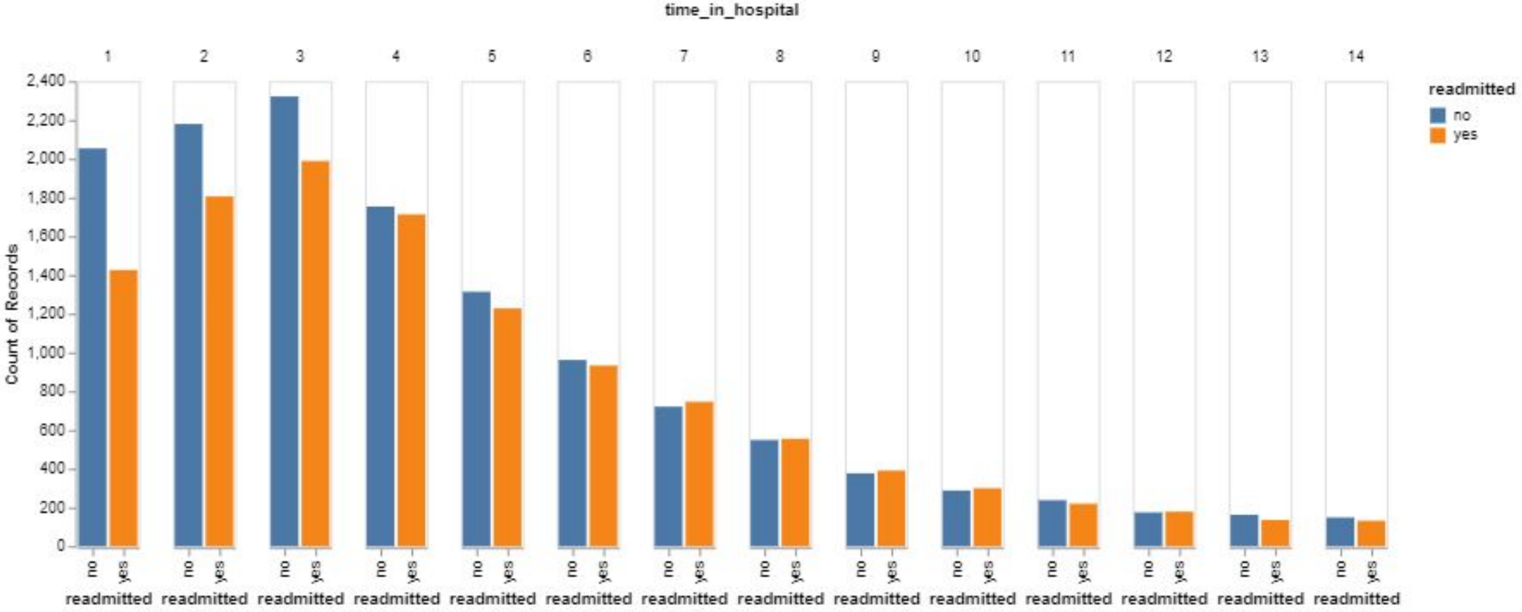

age
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
Count of Records
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
no
yes
readmitted
readmitted
no
yes


n_procedures
0
1
2
3
4
5
6
Count of Records
6,000
5,000
4,000
3,000
2,000
1,000
0
no
yes
readmitted
readmitted
no
yes

Dataset



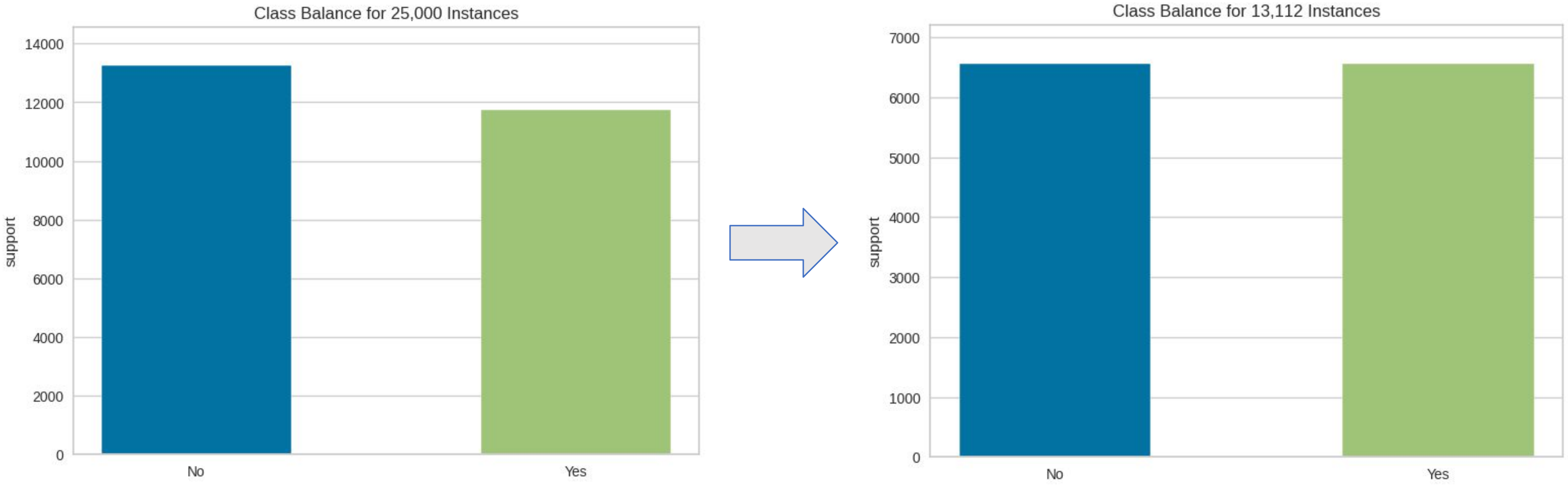
medical_specialty
Cardiology
Emergency/Trauma
Family/GeneralPractice
InternalMedicine
Missing
Other
Surgery
Count of Records
0
1,000
2,000
3,000
4,000
5,000
6,000
no
yes
readmitted
readmitted
no
yes



time_in_hospital
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Count of Records
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
2,400
no
yes
readmitted
readmitted
no
yes

Data Imbalance

Synthetic Minority Oversampling Technique (SMOTE)

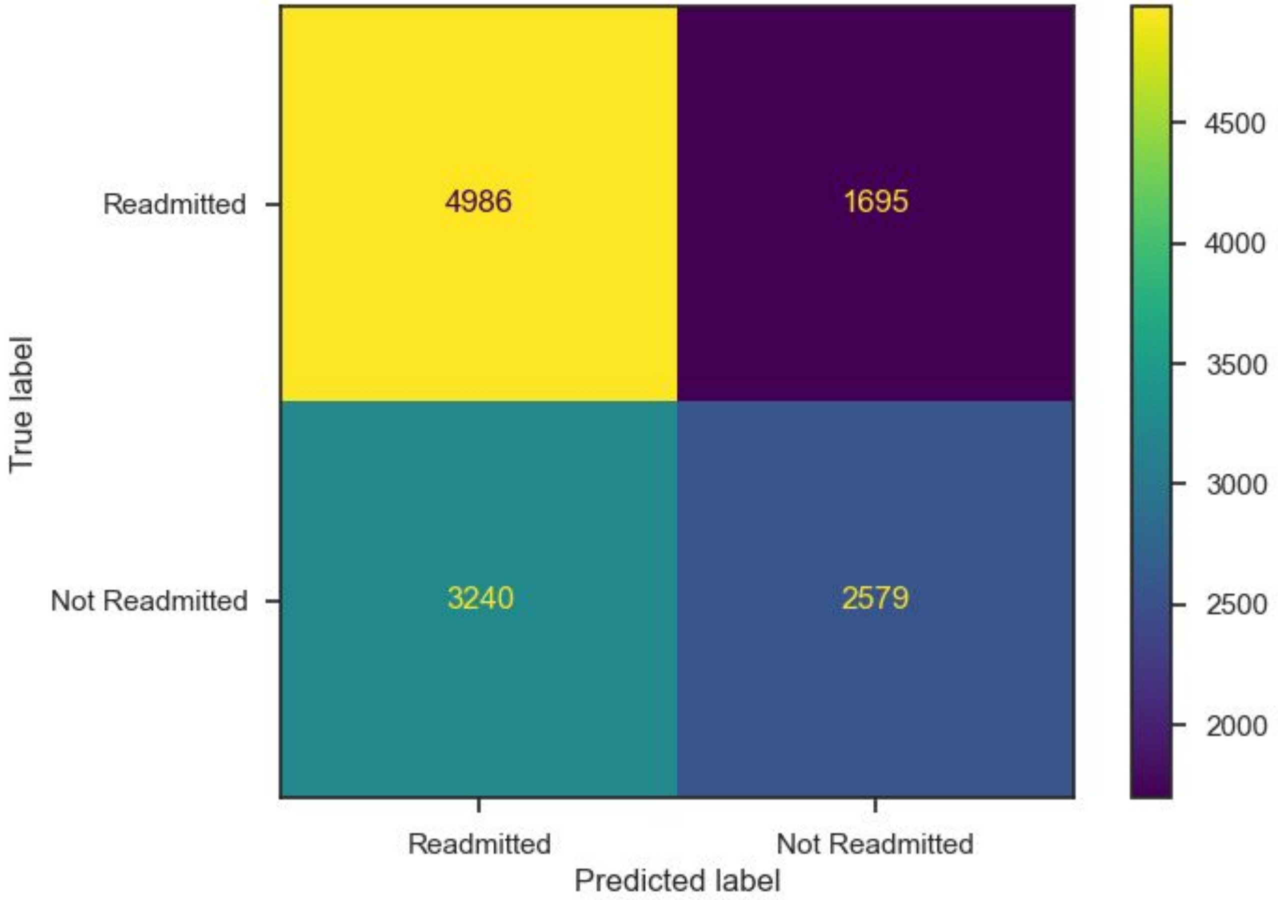


Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

ML Pipeline - Architecture

- Data Processing
  - Cleaning
  - Encoding
  - Scaling
- Training and testing the models
- Finding optimal hyperparameters for each model
- Tuning the model
- Testing the tuned model

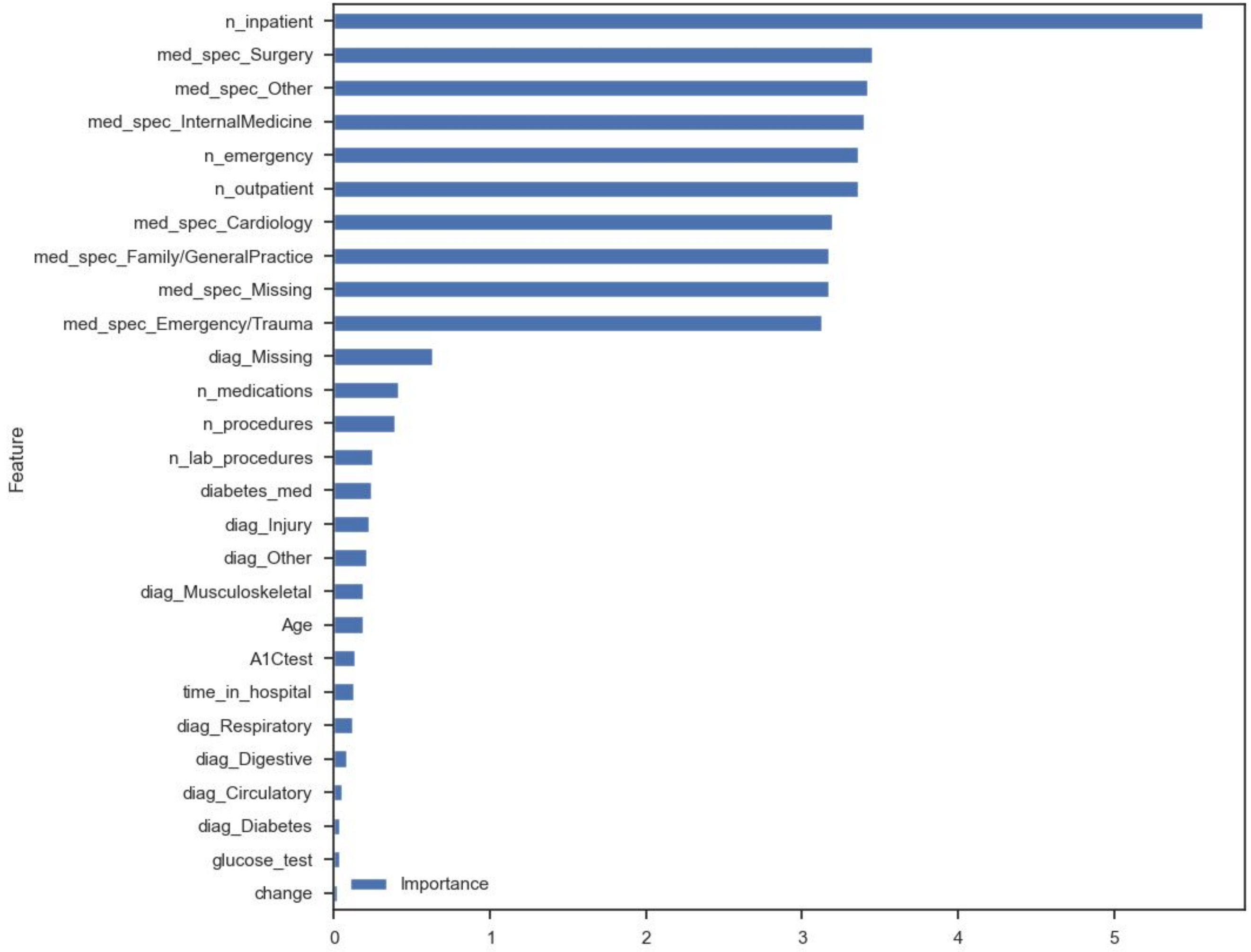
Logistic Regression Model



Classification Report:

	precision	recall	f1-score	support
Readmitted	0.606	0.746	0.669	6681
Not Readmitted	0.603	0.443	0.511	5819
accuracy			0.605	12500
macro avg	0.605	0.595	0.590	12500
weighted avg	0.605	0.605	0.595	12500

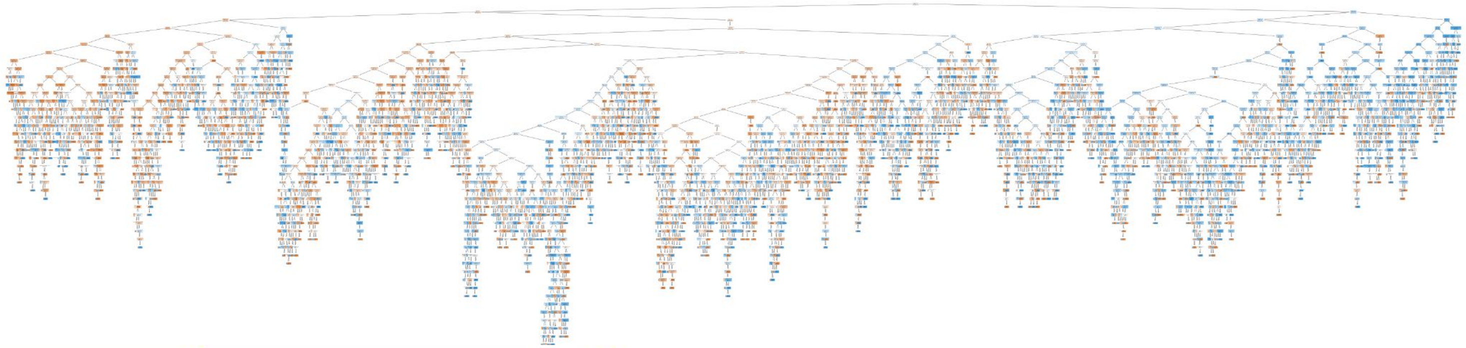
Feature Importance

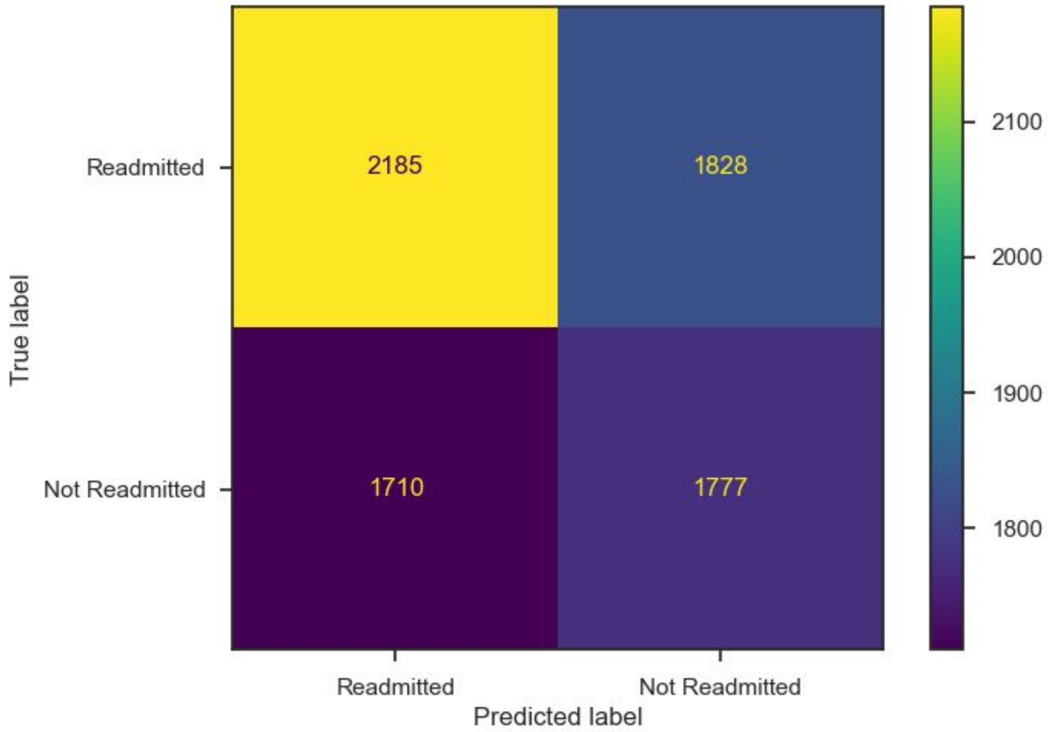


Feature Importance

1. n_inpatient
2. med_spec_surgery
3. med_spec_Other
4. med_spec_InternalMedicine
5. n_emergency
6. n_outpatient
7. med_spec_Cardiology

Decision Tree Model





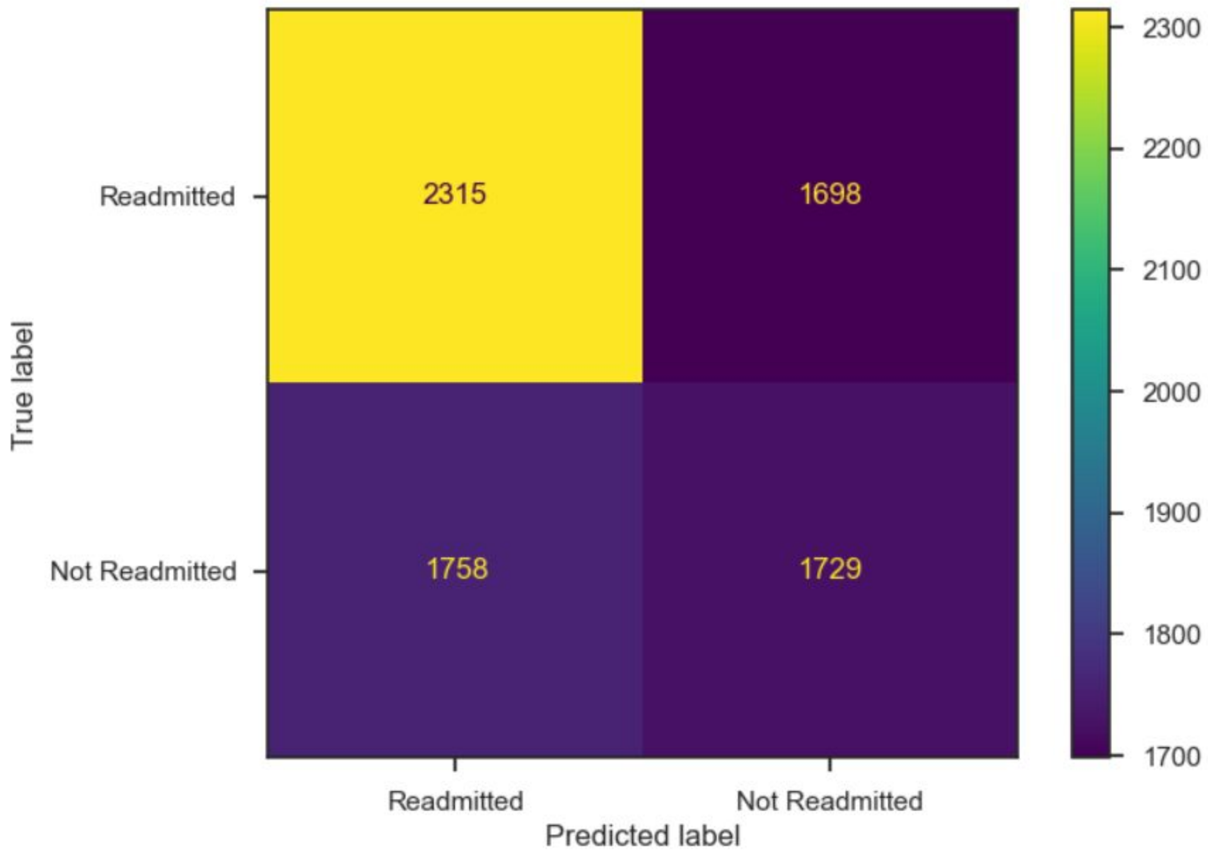

```
Classification Report:

                precision    recall  f1-score   support

Not Readmitted      0.561     0.544     0.553      4013
    Readmitted      0.493     0.510     0.501      3487

      accuracy                          0.528      7500
     macro avg      0.527     0.527     0.527      7500
  weighted avg      0.529     0.528     0.529      7500
```

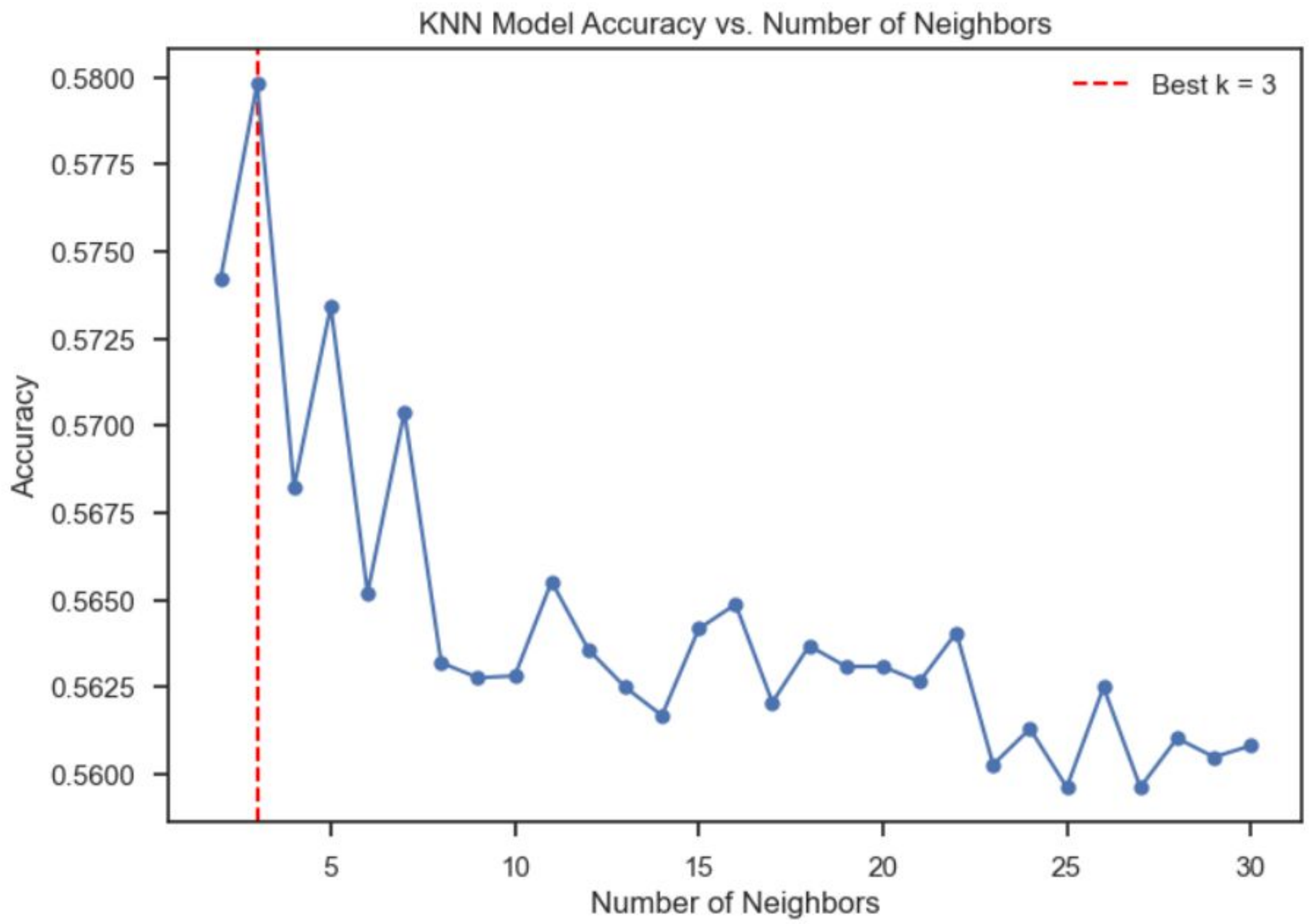
K Nearest Neighbour Model



Classification Report:

	precision	recall	f1-score	support
Not Readmitted	0.568	0.577	0.573	4013
Readmitted	0.505	0.496	0.500	3487
accuracy			0.539	7500
macro avg	0.536	0.536	0.536	7500
weighted avg	0.539	0.539	0.539	7500

Refining the model with optimal k-value



Classification Report:

	precision	recall	f1-score	support
Readmitted	0.559	0.560	0.560	4013
Not Readmitted	0.493	0.491	0.492	3487
accuracy			0.528	7500
macro avg	0.526	0.526	0.526	7500
weighted avg	0.528	0.528	0.528	7500

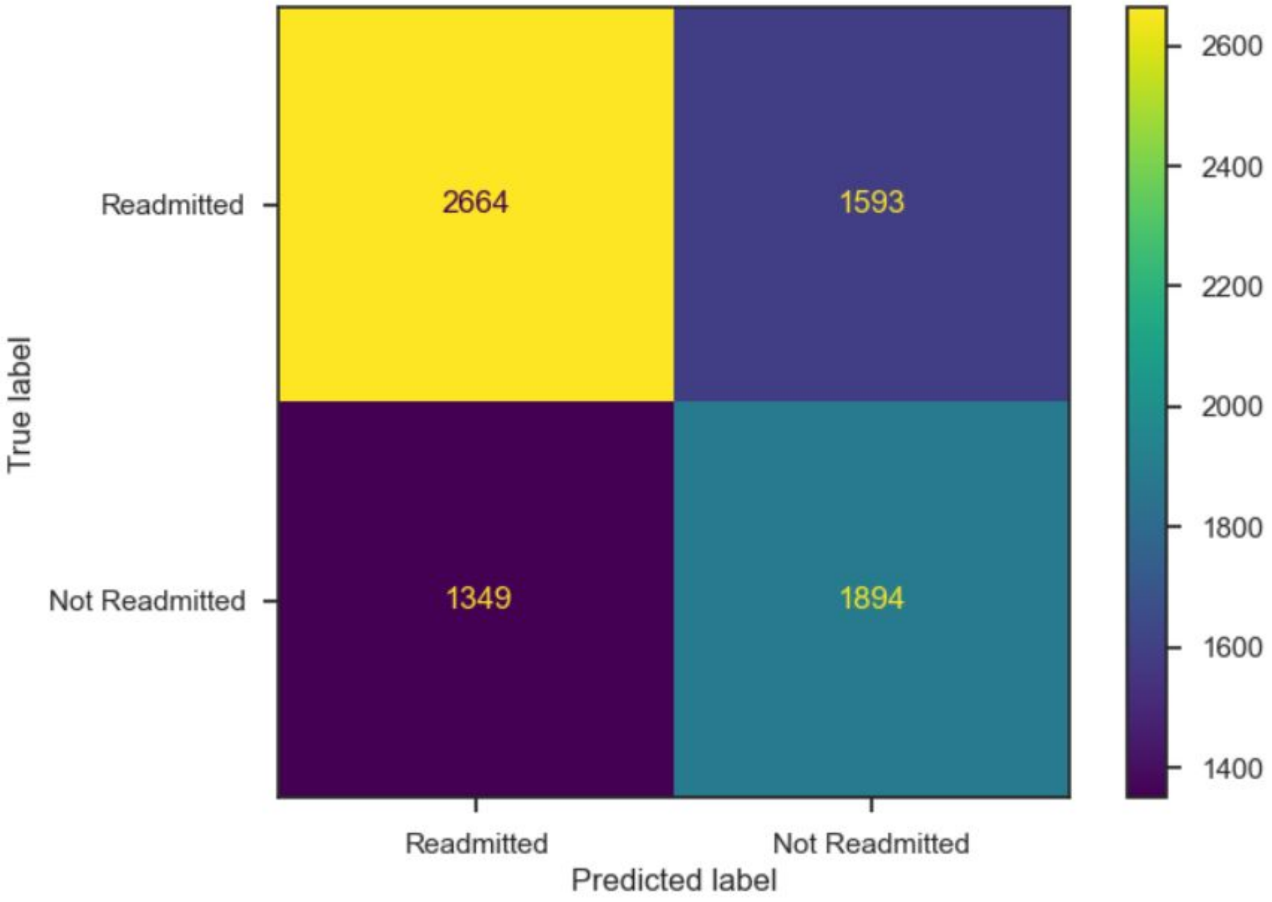
Random Forest Model

- The initial instance of random forest model trained on the balanced dataset obtained from SMOTE resulted in overfitting

- Model was tuned to obtain optima value of hyperparameters

  - Best number of trees: 200
  - Best parameters: {'max_depth': 15, 'min_samples_split': 10, 'n_estimators': 170}

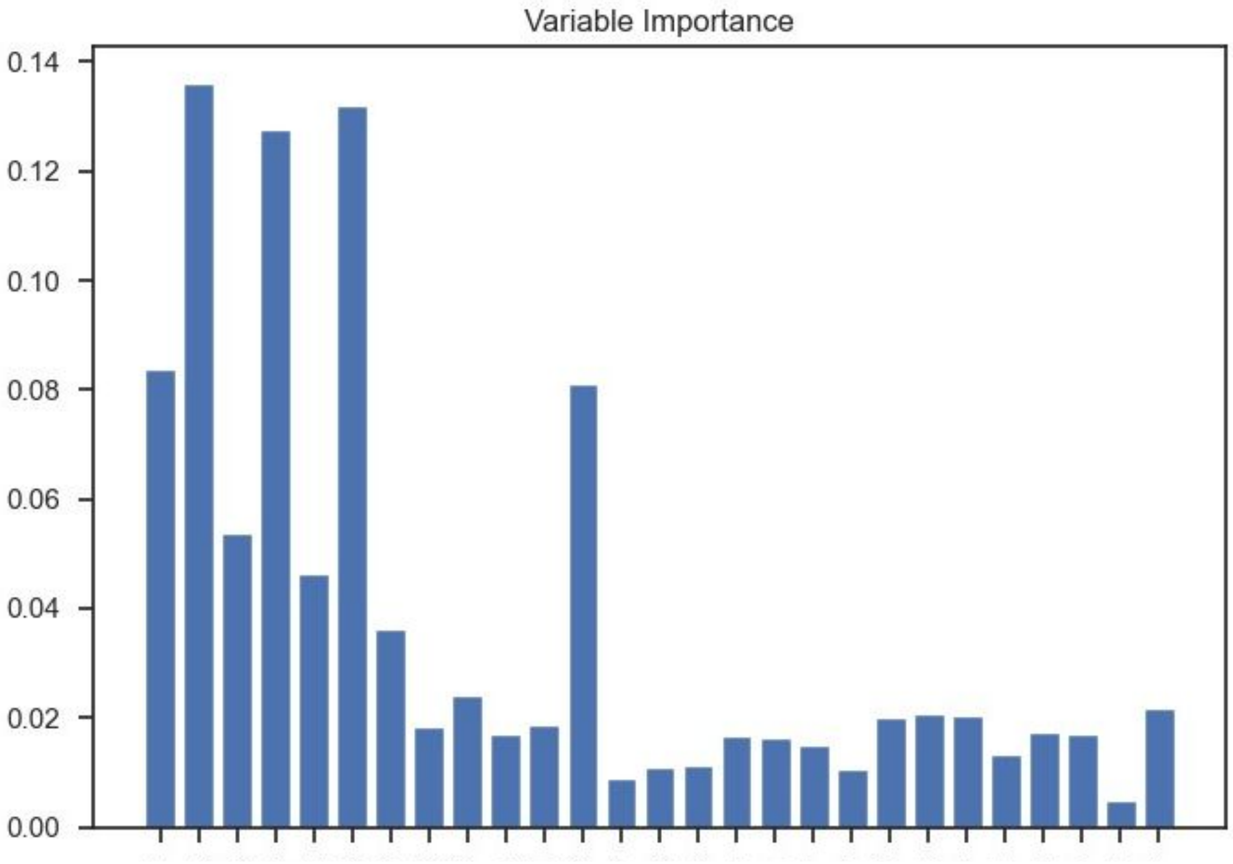
Results of the tuned model



Classification Report:

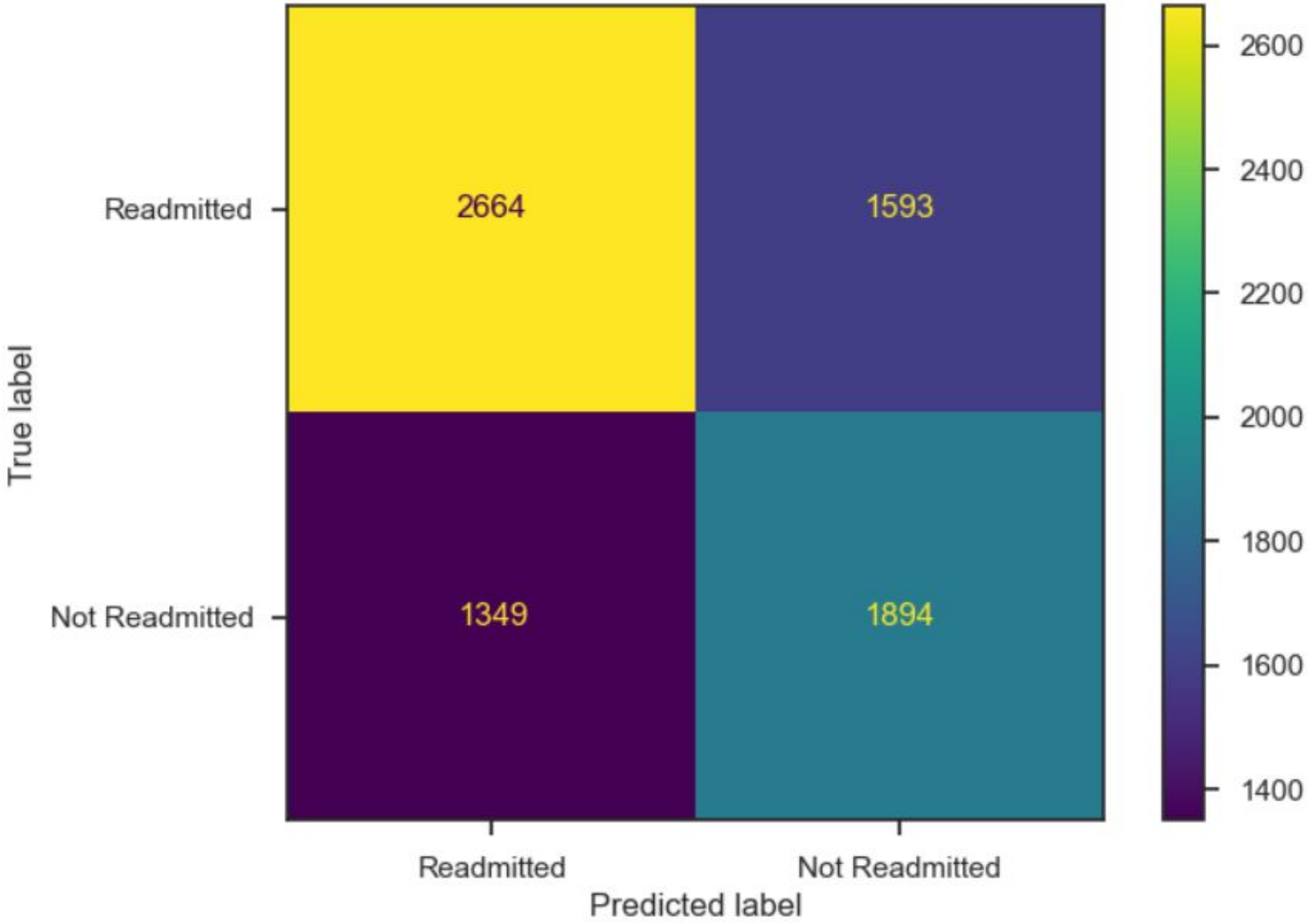
	precision	recall	f1-score	support
Readmitted	0.664	0.626	0.644	4257
Not Readmitted	0.543	0.584	0.563	3243
accuracy			0.608	7500
macro avg	0.604	0.605	0.604	7500
weighted avg	0.612	0.608	0.609	7500

Feature Importance



Feature Importance
1.n_lab_procedure(0.136050)
2. n_inpatient(0.132138)
3. n_medications (0.127699)
4. time_in_hospital (0.083753)
5. Age (0.080990)

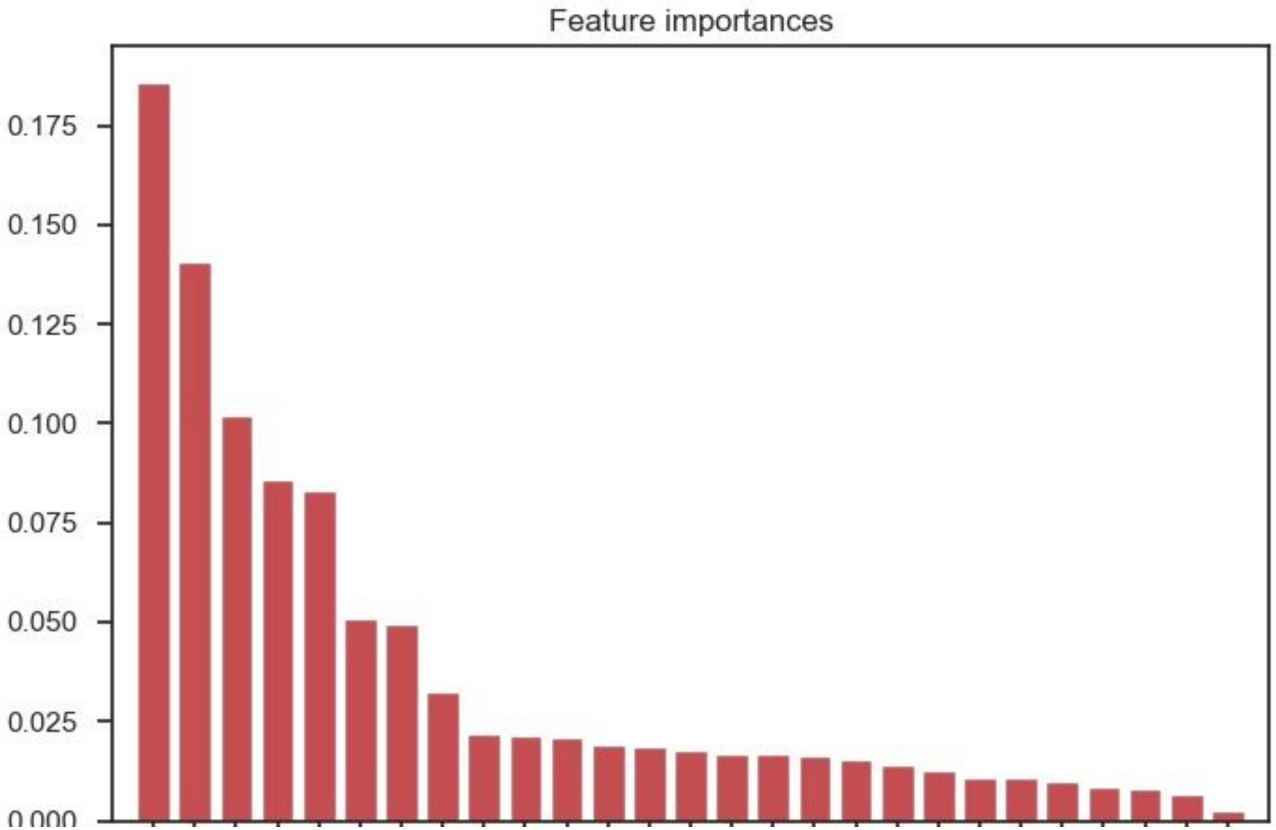
Adaboost Classifier



Classification Report:

	precision	recall	f1-score	support
Readmitted	0.62	0.68	0.65	4013
Not Readmitted	0.58	0.51	0.54	3487
accuracy			0.60	7500
macro avg	0.60	0.60	0.60	7500
weighted avg	0.60	0.60	0.60	7500

Feature Importance



Feature ranking:
1. n_lab_procedures (0.185823)
2. n_medications (0.140568)
3. n_inpatient (0.101933)
4. time_in_hospital (0.086020)
5. Age (0.083277)

Overall Analysis and Summary

- Logistic Regression, Decision Tree Classifier, KNN, and Random Forest were selected as potential models for prediction.
- Logistic Regression trained on the SMOTE-balanced dataset demonstrated the best performance, achieving a good balance between precision (60.7%) and recall (74.4%) for predicting readmissions.
- Feature-selected Logistic Regression showed higher precision (59.6%) but lower recall (78.0%).
- Decision Tree and KNN models performed less optimally, with lower precision and recall values.
- Random Forest achieved a balanced performance with moderate precision (54.3%) and recall (58.4%).

Future Work

- Model Refinement and Ensemble Methods
- Temporal Analysis
- Integration with Electronic Health Records (EHR)
- Validation on Diverse Datasets
- Ethical and Bias Considerations

Conclusion

The **Logistic Regression model** trained on the SMOTE-balanced dataset is recommended for predicting hospital readmissions in this context.

- Balanced Precision and Recall
- Interpretability of Logistic Regression
- Consideration of Healthcare Resource Allocation
- Robustness and Generalization

The common inclusion of 'n_inpatient,' 'n_medications,' 'n_lab_procedures,' 'time_in_hospital,' and 'Age' across Logistic Regression, Random Forest, and AdaBoost indicates their universal importance in predicting readmission.