BDA 640 - Final Project Report

John Pole Madhu

Executive Summary:

This report examines the challenge of high patient reclassification rates from the Observation Unit (OU) to the Emergency Unit at Dr. Kelly's hospital, where nearly 45% of OU patients are later admitted as inpatients. To address this, we leveraged predictive modeling techniques using historical data to enhance the OU exclusion criteria. Our analysis incorporated logistic regression, random forest, and decision tree models to evaluate patient outcomes, with logistic regression demonstrating the highest predictive accuracy. Based on our findings, implementing this model could potentially decrease the flipped rate to 20%, allowing the hospital to accommodate 572 more patients annually, generate an estimated $400,400 in additional revenue, and improve bed utilization by reducing unnecessary inpatient admissions by 3,328 patients each year.

Problem Description:

Observation Units (OUs) are designed to accommodate patients for shorter durations while utilizing fewer hospital resources compared to Emergency Units (EUs). However, Dr. Kelly observed that approximately 45% of patients initially admitted to the OU under her management eventually required transfer to the EU, prolonging their hospital stay. Minimizing patient hospitalization time benefits both individuals and the facility by expediting discharges and increasing the hospital's capacity to treat more patients throughout the year. To address this issue, Dr. Kelly aimed to refine the hospital's OU exclusion criteria using historical data and predictive modeling. This approach was intended to enhance system efficiency and ensure that patients were directed to the appropriate care setting, particularly in preparation for the flu season, when patient intake was expected to rise significantly.

Methodology:

Data Collection:

The dataset used for this analysis was sourced from the hospital's Electronic Health Record (EHR) system, which contains detailed records of patient visits to the Observation Unit (OU). Each record includes key attributes such as patient age, gender, insurance category, length of stay, and whether the patient was later transferred to inpatient care.

Data Preparation:

To ensure data accuracy, a thorough cleaning process was conducted to address missing values and properly format categorical variables like gender and insurance category for analysis. This step was essential for preserving the reliability of our predictive models. Additionally, the dataset was standardized, and any missing values were replaced using the median of the respective variable. Since both length of stay and flipped status were identified as target variables,

OU_LOS_hours was removed from the dataset to allow for separate investigation in subsequent analyses.

Exploratory Data Analysis (EDA):

To gain insights into the dataset, initial analyses were performed to examine the distribution of key variables. This included creating histograms and density plots to identify underlying trends and patterns. Additionally, a correlation matrix was generated to determine which variables had a stronger association with patients transitioning from the Observation Unit (OU) to inpatient status.

Model Development:

Logistic Regression Model (Model 1)

The first model utilized was logistic regression, a statistical technique designed to predict binary outcomes based on multiple predictor variables. In this analysis, logistic regression was applied to estimate the probability of a patient flipping from the Observation Unit (OU) to inpatient care, using all available variables from the cleaned dataset. This method was particularly suitable as it models the likelihood of an event occurring, making it ideal for categorical outcomes such as flipped vs. not flipped.

To enhance model performance, we experimented with different probability thresholds, adjusting them within the range of 0.4 to 0.8. The threshold value determines when a patient is classified as flipped or not flipped. While a 0.5 threshold is commonly used, it does not always yield the best results, especially when balancing sensitivity and specificity is crucial. By testing various threshold values, we aimed to achieve the best balance between predictive accuracy and an acceptable flipped rate, ensuring the model effectively distinguished between patients who required inpatient care and those who did not.

Random Forest Model (Model 2)

The second model applied was random forest, a machine learning technique particularly effective for classification tasks. Random forest operates by constructing multiple decision trees during training and aggregating their predictions to enhance accuracy and stability. This ensemble approach mitigates overfitting, a common issue with individual decision trees, making it well-suited for datasets with numerous predictor variables.

To assess its effectiveness in predicting the flipped rate, we ran the random forest model multiple times. Similar to logistic regression, we experimented with threshold adjustments between 0.4 and 0.8, aiming to identify the optimal balance between accuracy and misclassification rates.

Decision Tree Model (Model 3)

The third model employed was a decision tree, a widely used technique for both classification and regression tasks. Decision trees function by recursively splitting the dataset based on predictor variable values, creating a tree-like structure that categorizes observations into distinct groups. Each internal node represents a decision point based on a variable, while the terminal nodes (leaves) denote the predicted class or outcome.

For this analysis, the decision tree model was designed to enhance overall accuracy and improve predictions of the flipped rate. To ensure reliability, the model was tested multiple times, using the same probability thresholds (0.4 to 0.8) as in the logistic regression and random forest models.

Refined Models and Adjustments:

After testing the initial three models (logistic regression, random forest, and decision tree), we reran them using a refined set of predictor variables. This iteration included only variables that were statistically significant in the highest-performing logistic regression model, which was originally conducted with the full dataset. The two significant variables identified were "DRG01" (representing initial diagnosis) and "PrimaryInsuranceCategory" (indicating the patient's insurance type). These refined models were evaluated multiple times using thresholds between 0.4 and 0.8 to assess their impact on both model performance and flipped rate.

During the analysis, we discovered an initial misinterpretation in how we calculated the flipped rate. Initially, we incorrectly assumed that it could be directly derived from the confusion matrix, which primarily measures accuracy, precision, and recall. However, flipped rate specifically refers to the proportion of instances predicted as "Flipped" (or 1 in the binary outcome). To correct this, we recalculated the flipped rate by taking the mean of predicted classifications, ensuring a more accurate representation of the model's predictions. After identifying this discrepancy, we updated our code to ensure proper evaluation of flipped rate alongside other key performance metrics such as accuracy.

Alternative Model:

In addition to the primary models, we conducted a linear regression analysis to examine the relationship between predictor variables and the length of stay in the Observation Unit (OU), measured as OU_LOS_hours. Unlike previous models, this analysis excluded the Flipped variable since it was not the primary outcome of interest. Instead, we created a new dataset tailored for this model, replacing the target variable while retaining all relevant predictor variables—except RecordKey and Gender, which were either unnecessary or had low correlation based on our exploratory data analysis.

The adjusted R-squared value for this regression model was 6%, indicating that the model explained only a small portion of the variability in OU_LOS_hours. The analysis identified Age and DRG as the most significant variables. However, when rerunning the regression using only

these two predictors, the variability remained at 6%, suggesting that additional factors may be influencing length of stay in the OU.

Results:

We initially hypothesized that logistic regression would be the most effective model for predicting binary outcomes. After testing multiple models, we found that our original logistic regression model, which used Flipped as the dependent variable while excluding ObservationRecordKey, InitPatientClassAndFirstPostOUClass, and OU_LOS_hrs, delivered good accuracy and the lowest flipped rate. The model achieved an accuracy of 59.94% with an optimal cutoff threshold of 0.6. While this accuracy is not as high as desired, it remains the most effective predictive model in our analysis.

To calculate the accuracy, we summed the true negatives and true positives, then divided this by the total number of true negatives, false positives, false negatives, and true positives (148 true negatives, 102 false positives, 31 false negatives, and 51 true positives), resulting in an accuracy of 59.94%.

Through extensive trial and error, we determined that a cutoff level of 0.6 provided the best balance between flipped rate and accuracy. Applying this threshold resulted in a flipped rate of 24.7%, meaning fewer patients were incorrectly classified as needing inpatient care. The goal was to minimize the flipped rate while maintaining a strong overall predictive performance.

Although this model performed well, further refinements—such as incorporating additional predictive variables—could enhance its accuracy. While models tested with a 0.7 threshold resulted in a lower flipped rate, they also reduced accuracy or failed to outperform our chosen logistic regression model.

Overall, our analysis confirms that logistic regression is an effective approach for accurately classifying Flipped cases. While there is room for improvement, these results establish a strong foundation for further optimization, aiming to improve predictive accuracy and further reduce the flipped rate.

Recommendations:

To assess the impact of reducing the flipped rate to 20% on profitability and bed utilization, it is essential to first understand Dr. Kelly's current system and estimate the expected improvements. Currently, the flipped rate stands at 45%, meaning that nearly half of the Observation Unit (OU) patients are eventually reclassified as inpatients, leading to longer hospital stays and increased resource strain.

Dr. Kelly reports that the OU currently treats an average of 44 patients per week. If the flipped rate is reduced from 45% to 20%, as predicted by our logistic regression model, the total number of patients treated in the OU would increase. The expected number of new patients treated can be calculated using the formula:

New Patients = Current Patients + (Current Patients × (Old Flipped Rate − New Flipped Rate))

Applying the values from our analysis:

New Patients = 44 + (44 × (0.45 − 0.20)) = 44 + (44 × 0.25) = 44 + 11 = 55 patients per week

This means the OU could accommodate 11 additional patients per week. Over the course of a year, assuming continuous operations, the annual increase in patient capacity would be:

Additional Patients = 11 × 52 = 572 more patients per year

Financial Impact

With the hospital earning approximately $700 per emergency room visit, the additional revenue generated by treating more patients would be:

Additional Profit = 572 patients × $700 = $400,400 annually

Impact on Bed Utilization

To evaluate how reducing the flipped rate to 20% affects inpatient bed capacity, we calculated the expected reduction in inpatient admissions using the formula:

New Patients in Beds = Current Patients in Beds × (New Flipped Rate / Old Flipped Rate)

With 115 current inpatient cases, and the flipped rate adjusted from 45% to 20%, the calculation is: *New Patients in Beds = 115 × (0.20/0.45) = 115 × 0.4444 = 51 patients per week*

Thus, under this model, only 51 patients would require inpatient beds each week, compared to the current 115. This represents a reduction of 64 patients per week (115 - 51 = 64).

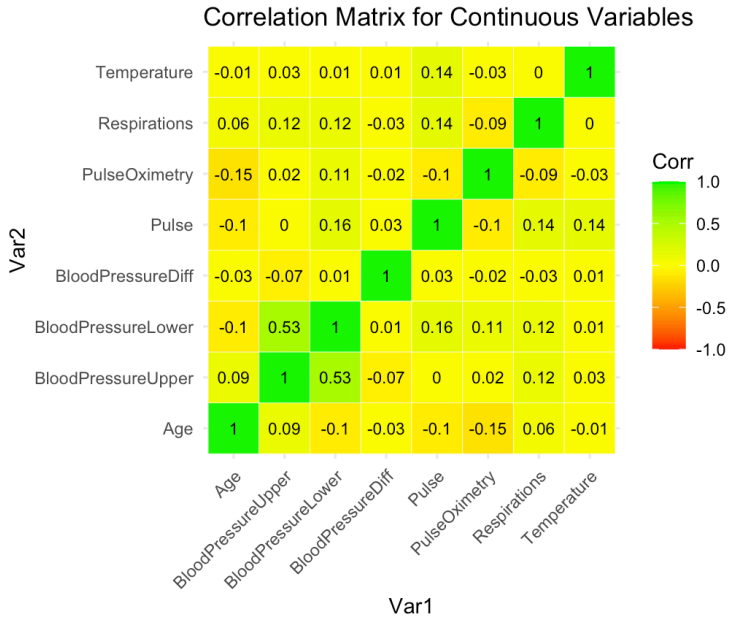
Annually, this reduction translates to: *64 patients per week × 52 weeks = 3,328 fewer inpatient admissions per year*

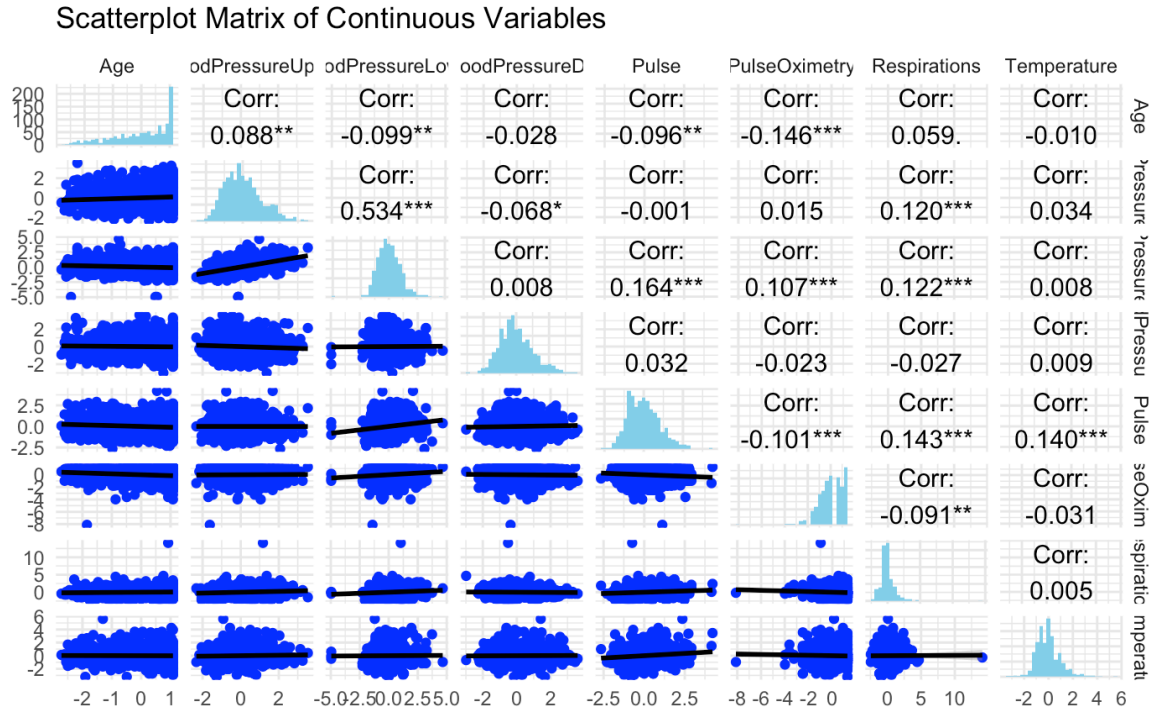
This significant decline in inpatient admissions would optimize hospital bed utilization, reducing overcrowding and improving patient flow.

Conclusion:

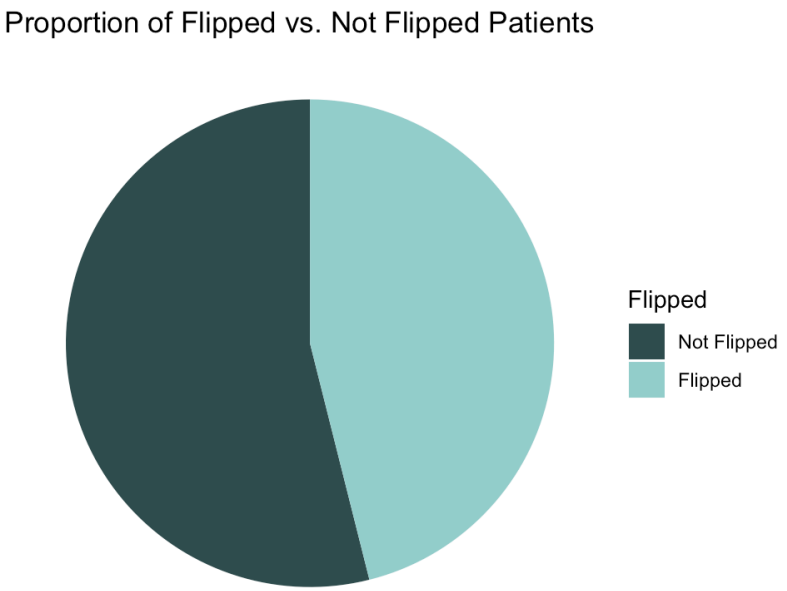
By implementing our logistic regression model, Dr. Kelly's hospital could increase the number of patients treated, enhance resource efficiency, and boost revenue while significantly reducing unnecessary inpatient admissions. Aligning with Dr. Kelly's goal of improving hospital operations, this strategy ensures better resource management and patient care. Therefore, we strongly recommend adopting the logistic regression model to achieve these operational improvements.

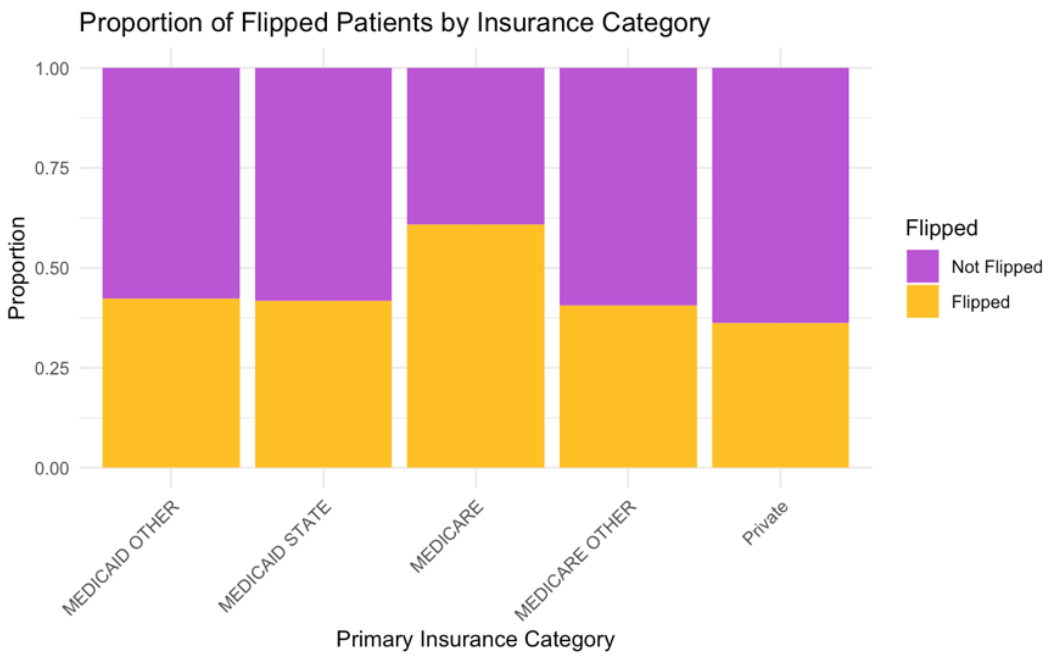
Appendix:

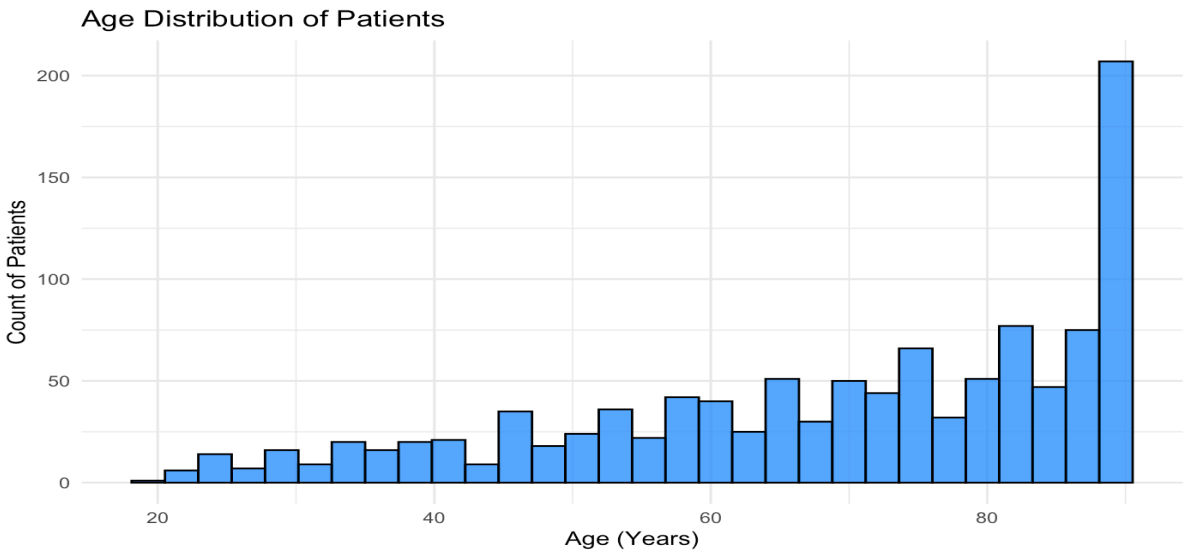


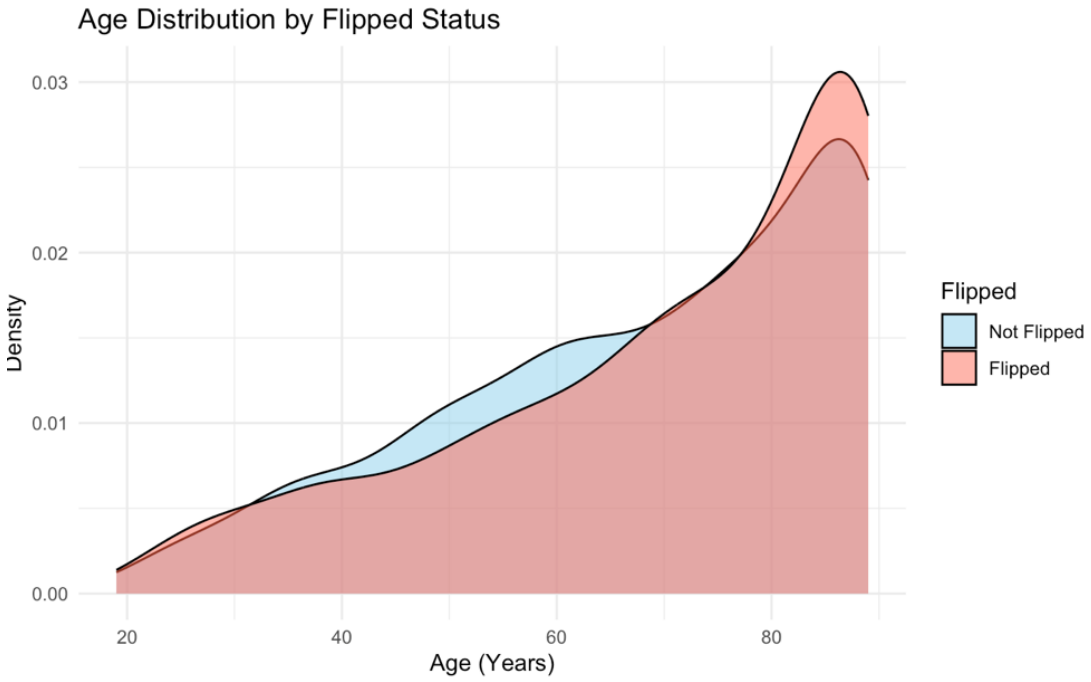


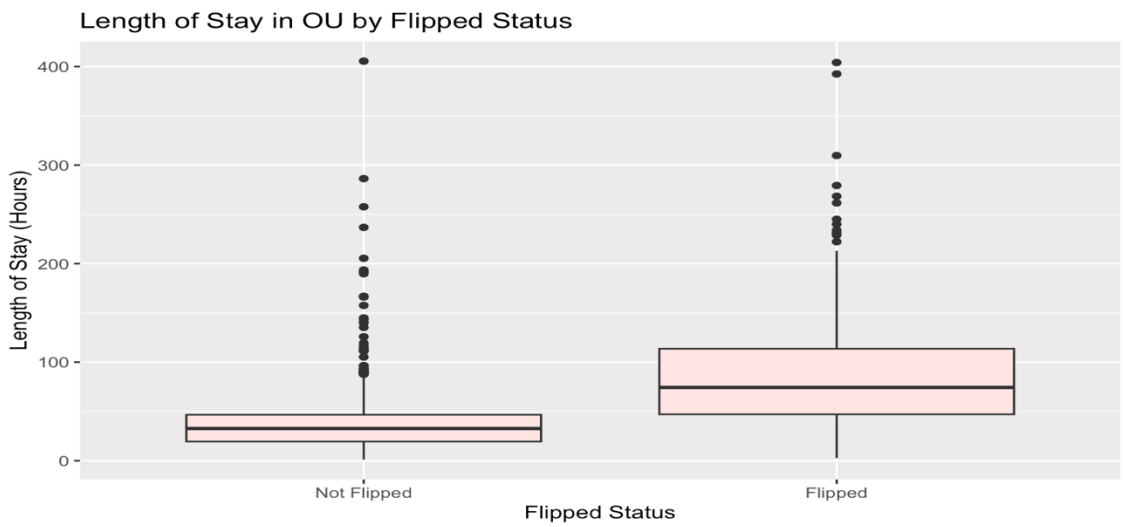
Flipped	Age_Mean	Age_SD	BloodPressureUpper_Mean	BloodPressureUpper_SD	BloodPressureLower_Mean	Bl
0	-0.05	0.99	0.01	0.98	0.05	
1	0.05	1.01	-0.02	1.02	-0.06	

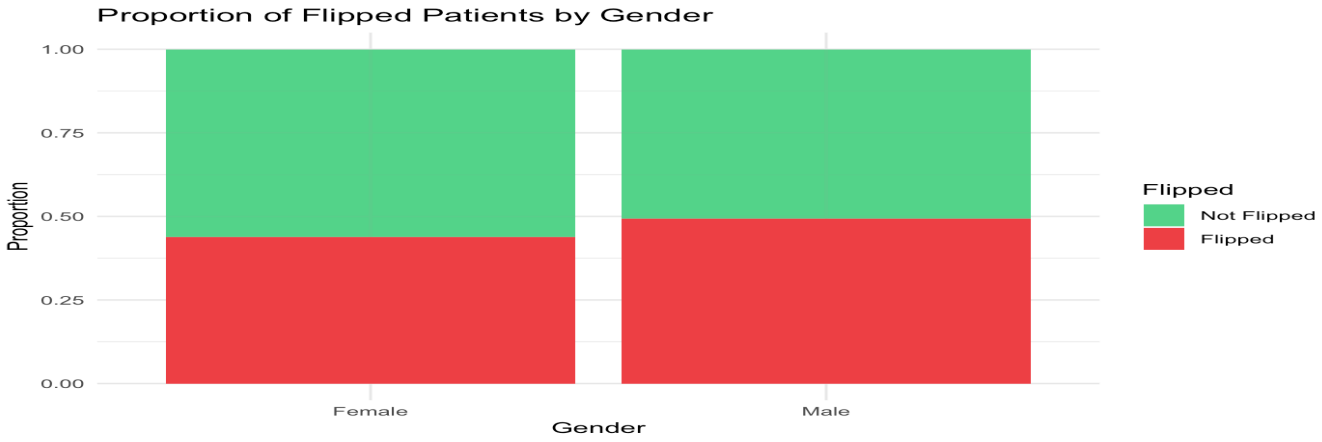

Proportion of Flipped vs. Not Flipped Patients
Flipped
Not Flipped
Flipped

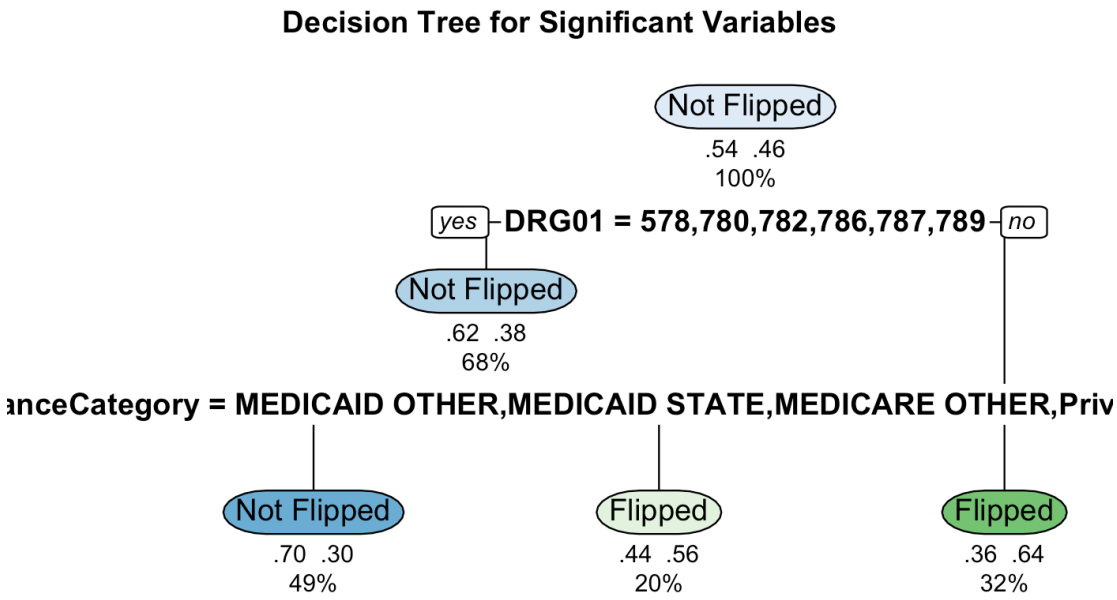

Proportion of Flipped Patients by Insurance Category
1.00
0.75
0.50
0.25
0.00
Proportion
MEDICAID OTHER
MEDICAID STATE
MEDICARE
MEDICARE OTHER
Private
Primary Insurance Category
Flipped
Not Flipped
Flipped

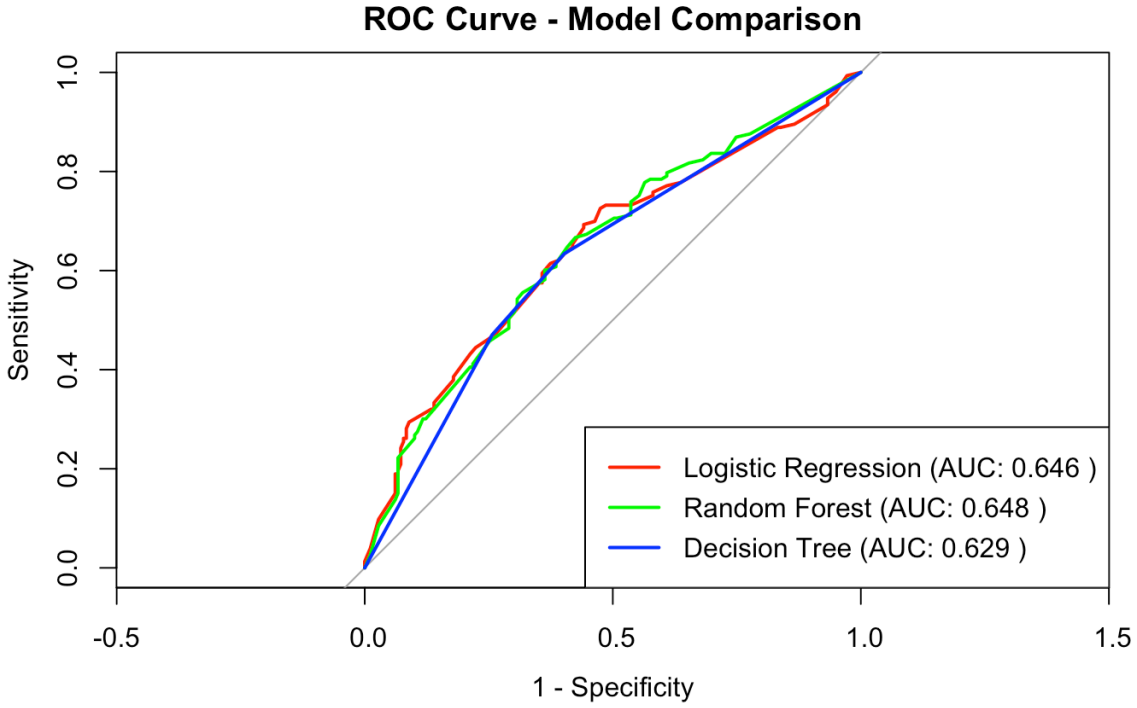

Age Distribution of Patients
Count of Patients
200
150
100
50
0
20
40
60
80
Age (Years)

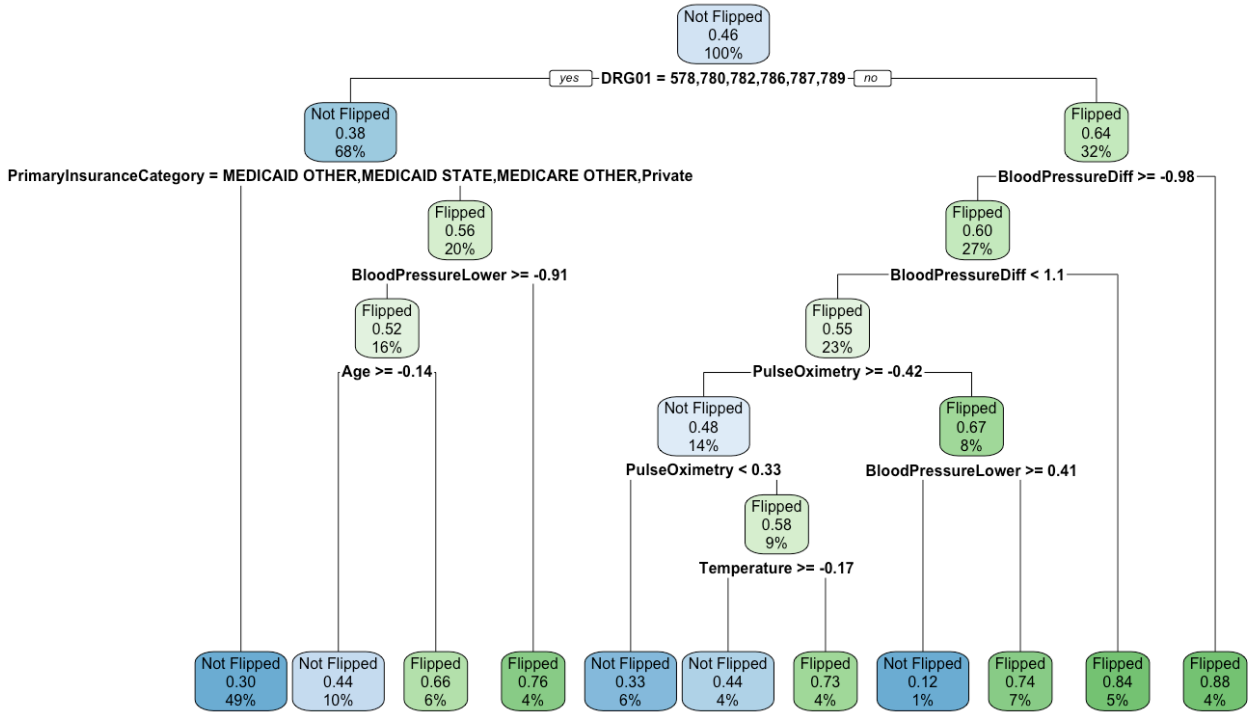

Age Distribution by Flipped Status
Density
0.03
0.02
0.01
0.00
20
40
60
80
Age (Years)
Flipped
Not Flipped
Flipped


Length of Stay in OU by Flipped Status
Length of Stay (Hours)
400
300
200
100
0
Not Flipped
Flipped
Flipped Status


Proportion of Flipped Patients by Gender
Proportion
1.00
0.75
0.50
0.25
0.00
Female
Male
Gender
Flipped
Not Flipped
Flipped


Decision Tree for Significant Variables
Not Flipped
.54 .46
100%
yes
DRG01 = 578,780,782,786,787,789
no
Not Flipped
.62 .38
68%
anceCategory = MEDICAID OTHER,MEDICAID STATE,MEDICARE OTHER,Priv
Not Flipped
.70 .30
49%
Flipped
.44 .56
20%
Flipped
.36 .64
32%


ROC Curve - Model Comparison
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
1 - Specificity
Logistic Regression (AUC: 0.646)
Random Forest (AUC: 0.648)
Decision Tree (AUC: 0.629)



Decision Tree for Flipped Rate Prediction