PROJECT REPORT

By

SARAVIND REDDY SAMA

VENKATA SAI KUMAR RAMBHA

EXECUTIVE SUMMARY:

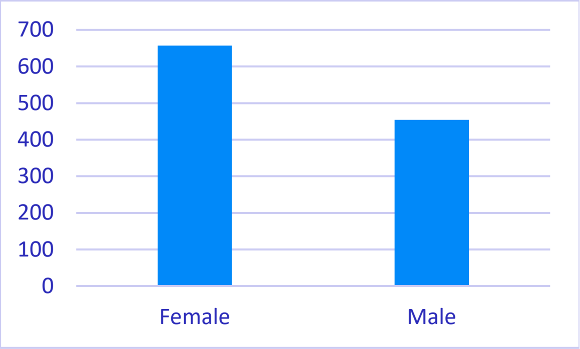
Our study aimed to create a prediction model to identify patients who are more likely to be flipped, as well as to establish which DRG codes relate to longer stays and which codes have a high flip ratio. We employed a binary classification technique and a logistic regression model with patient information as predictors to accomplish this. We discovered that DRG codes 558, 428, and 599 had greater predicted probabilities of being flipped, and DRG codes 599,780, and 787 had longer anticipated lengths of stay.

Our findings emphasize the significance of utilizing predictive models to identify high-risk individuals and identify which medical codes relate to greater expenditures. Based on our findings, we recommend that the hospital create an exclusion list focusing on the DRG codes with the highest predicted probability of being flipped and the longest predicted lengths of stay. This list could be used to help reduce costs and improve patient outcomes by providing earlier intervention for high-risk patients. Based on the findings, we recommend that hospitals should include DRG codes 558,428,599 in the exclusion list which can reduce the transfer of patients from and to the observation unit, this way the utilization of observation unit can be increased.

PROBLEM DESCRIPTION:

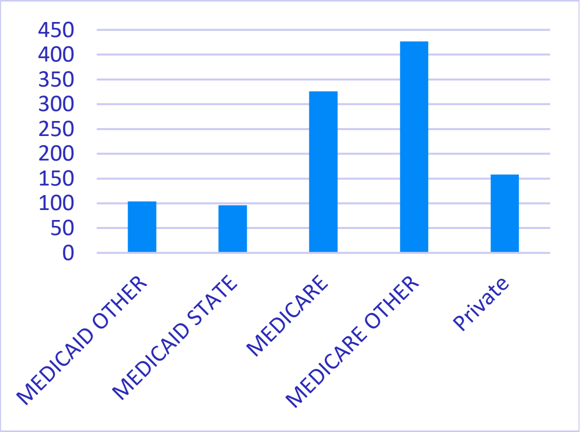
Developing a predictive model to identify which patients were more likely to flip from observation to inpatient status, identify which DRG codes are associated with longer stays, and predict which DRG codes have a high flip ratio So that they can decide on a better OU exclusion list.

Figure 1: Gender



Source: OU Data

Figure 2: Insurance Category



Source: OU Data

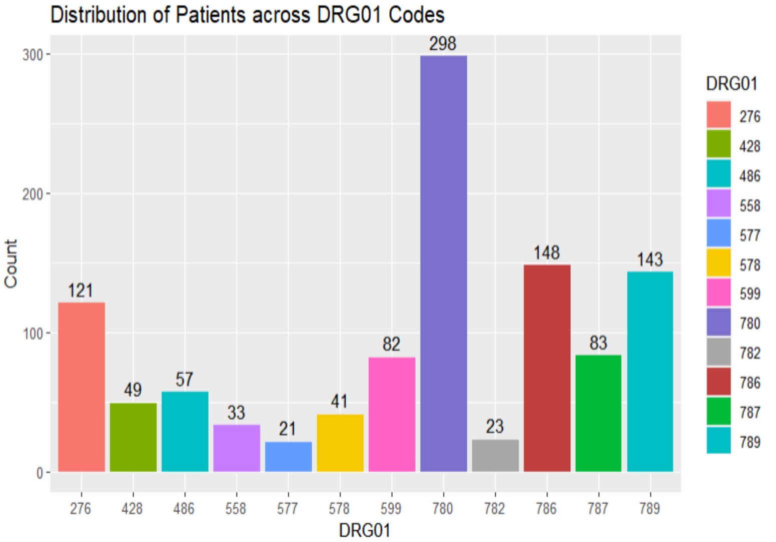
DATA CLEANING AND PREPROCESSING:

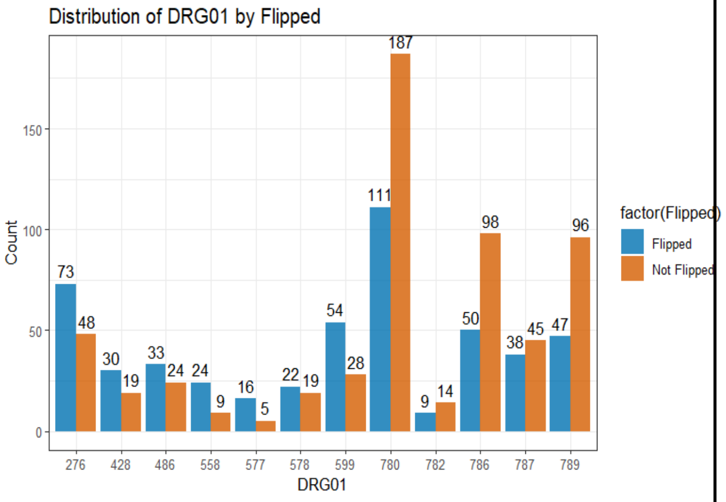
Firstly, we cleaned and preprocessed the data by identifying the missing values in the dataset and then handling those missing values appropriately, we checked for missing values in the data, and then removed all rows with missing values using the na.omit() function. here we assumed that missing values doesn't have any significant impact on the data as there were only 16 missing values in the dataset. We also removed 2 unnecessary variables one being ObservationRecordKey which has unique value for each customer and doesn't hold any significance on the flipped, the other one is InitPatientClassAndFirstPostOUClass which is similar to flipped column, so removed these 2 variables. Next, we converted categorical variables Gender, Primary Insurance Category, and DRG01 into factors using the as.factor() function.

EXPLORATORY DATA ANALYSIS:

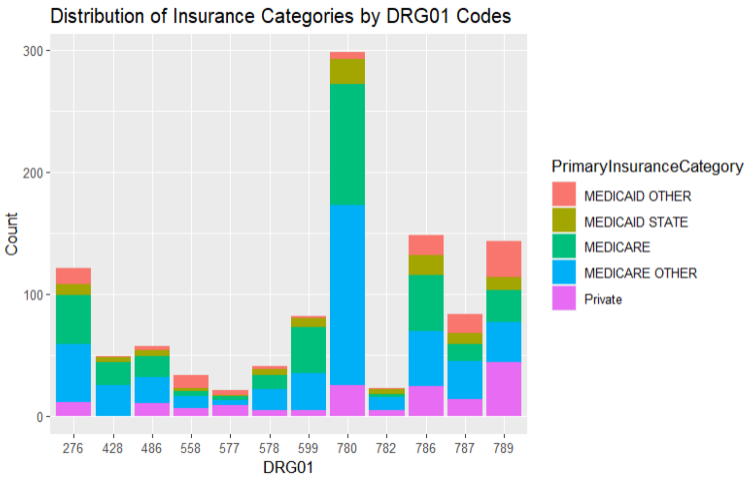
Secondly, we conducted exploratory data analysis to see if there are any patterns or trends in the dataset and gain insights from it, we used summary of the data to get insights about the mean values and maximum values of the variables and used histograms, boxplots to explore the data and identify any outliers, for example we created box plots to examine the distribution of OU_LOS_hrs by Flipped and DRG01, and histograms to examine the distribution of Age and OU_LOS_hrs.

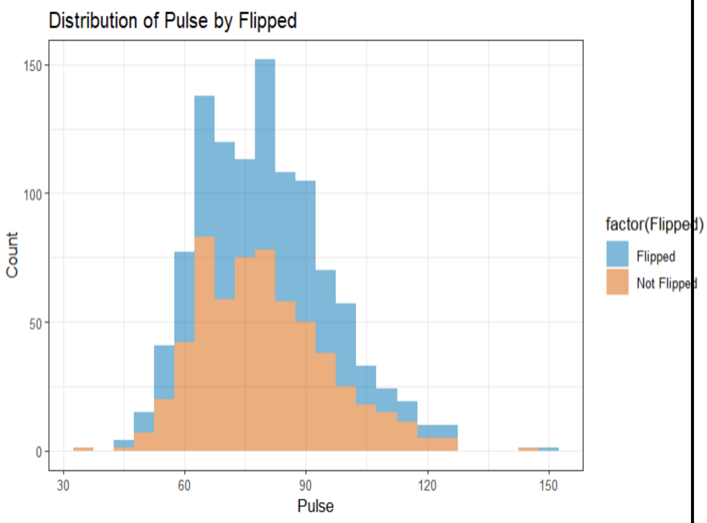
We also used t-tests and ANOVA to compare the means of different groups and identify any significant differences. For example, we used a t-test to compare the mean Age of patients who were flipped versus those who were not flipped, and an ANOVA to compare the mean OU_LOS_hrs across different DRG01 categories.





- From these graphs, we can see that most of the patients that are visiting for the purpose of 780, which is syncope, have flipped.
- The second-highest category where patients are high is for DRG codes 786 with 148 patients and 789 with 143 patients; also, DRG code 276 has 121 patients.
- The DRG codes 599, which is used for urinary tract infection, and 276, which is used for dehydration, have the highest number of flips in comparison to their respective numbers of patients who are not flipped.
- The codes 780, 782, 786, 787, and 789 have a lower percentage of patients flipped in comparison to their respective not flipped.
- These are the observations from the visualization, now we want to cross- validate these results from the model developed.





- From these charts we can see the most number of patients visiting the hospital have either MEDICARE or MEDICARE Other.
- The MEDICAID STATE insurance category patients are very less in comparison to all other categories.
- From these charts we can see if the Pulse is above 80 there are very high chances that these patients might be flipped.

MODELING:

Now since we have the clean dataset we wanted to go ahead and build predictive models to address the business problem, for this we wanted to build two types of models one was to predict which patients are more likely to flip from observation to inpatient status and what variables are most significant in patient flipping, the other model was to predict which DRG codes were associated with longer stays and higher flip ratios

For the first model we used logistic regression by considering flipped as the dependent variable and all other variables as predictors, we used glm() function with family binomial to fit the logistic regression model and this regression gave us OU_LOS_hrs, DRG1558, DRG01780, DRG01786, DRG01787, Pulse and PrimaryInsuranceCategoryMedicare as the significant variables. We then used the predict() function to predict the probability of flipping for each patient in the testing set. These probabilities will explain to us which patients are more likely to flip.

For the second model we used linear regression model by considering OU_LOS_hrs as the dependent variable and all other variables as predictors, we used lm() function for this. Then the model was used to estimate the expected duration of stay for each DRG01 category, and the predictions were compared to determine which DRG01 codes related to longer stays. Also, we calculated which DRG codes have high flip ratios.

We also used clustering techniques such as k-means clustering and hierarchical clustering to identify any patterns or clusters in the data. For example, we used k-means clustering to group patients based on their vital signs such as blood pressure, pulse rate, and temperature.

MODEL EVALUATION:

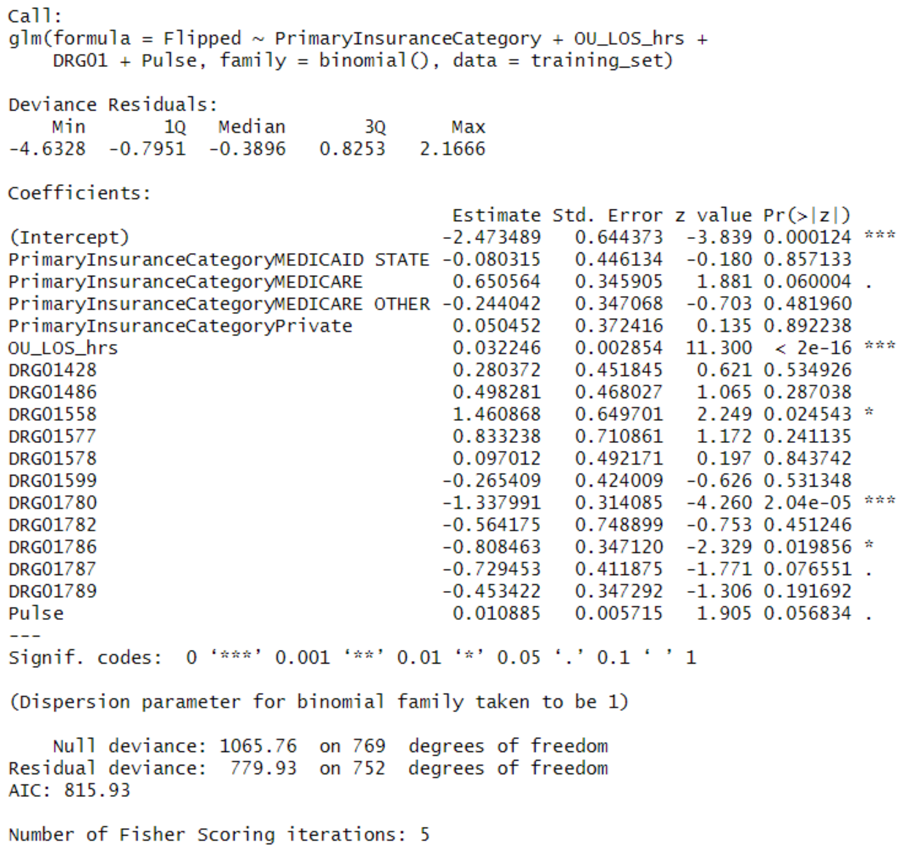
As the final step we wanted to see how accurately our models were performing for the logistic regression model, we used metrics like accuracy and recall.

We also used visualizations such as ROC curves and precision-recall curves to evaluate the performance of the logistic regression model and has considered 0.5 to be the optimal cutoff point for classifying patients as flipped or not flipped.

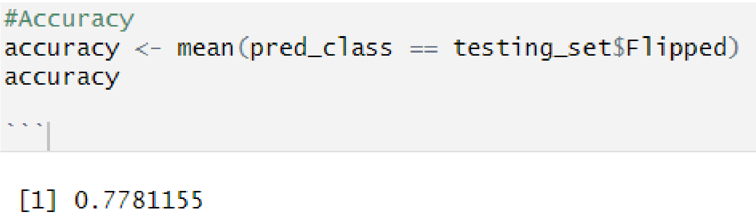
RESULTS:

MODEL 1:

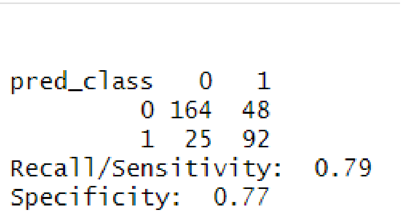
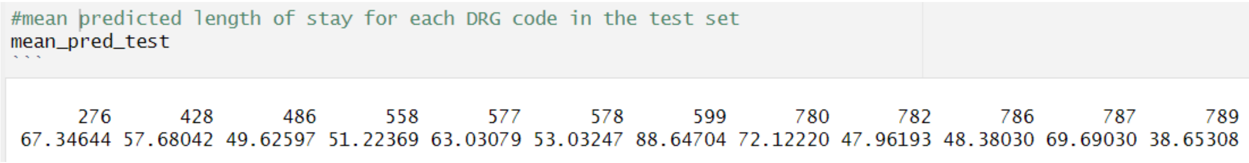
For the first model, the developed model was able to achieve an accuracy of approximately 77.8% on the test set and identified several variables that were significant. Those variables are OU_LOS_hrs, DRG1558, DRG01780, DRG01786, DRG01787, Pulse and PrimaryInsuranceCategoryMedicare as the significant variables. Our findings suggest that patients who have PrimaryInsuranceCategoryMedicare insurance type and are admitted for diagnose under DRG1558, DRG01780, DRG01786, DRG01787 are more likely to have a flipped.



```
Call:
glm(formula = Flipped ~ PrimaryInsuranceCategory + OU_LOS_hrs +
    DRG01 + Pulse, family = binomial(), data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6328  -0.7951  -0.3896   0.8253   2.1666

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.473489   0.644373  -3.839 0.000124 ***
PrimaryInsuranceCategoryMEDICAID STATE -0.080315   0.446134  -0.180 0.857133
PrimaryInsuranceCategoryMEDICARE        0.650564   0.345905   1.881 0.060004 .
PrimaryInsuranceCategoryMEDICARE OTHER -0.244042   0.347068  -0.703 0.481960
PrimaryInsuranceCategoryPrivate         0.050452   0.372416   0.135 0.892238
OU_LOS_hrs                              0.032246   0.002854  11.300  < 2e-16 ***
DRG01428                                0.280372   0.451845   0.621 0.534926
DRG01486                                0.498281   0.468027   1.065 0.287038
DRG01558                                1.460868   0.649701   2.249 0.024543 *
DRG01577                                0.833238   0.710861   1.172 0.241135
DRG01578                                0.097012   0.492171   0.197 0.843742
DRG01599                               -0.265409   0.424009  -0.626 0.531348
DRG01780                               -1.337991   0.314085  -4.260 2.04e-05 ***
DRG01782                               -0.564175   0.748899  -0.753 0.451246
DRG01786                               -0.808463   0.347120  -2.329 0.019856 *
DRG01787                               -0.729453   0.411875  -1.771 0.076551 .
DRG01789                               -0.453422   0.347292  -1.306 0.191692
Pulse                                   0.010885   0.005715   1.905 0.056834 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1065.76  on 769  degrees of freedom
Residual deviance:  779.93  on 752  degrees of freedom
AIC: 815.93

Number of Fisher Scoring iterations: 5
```

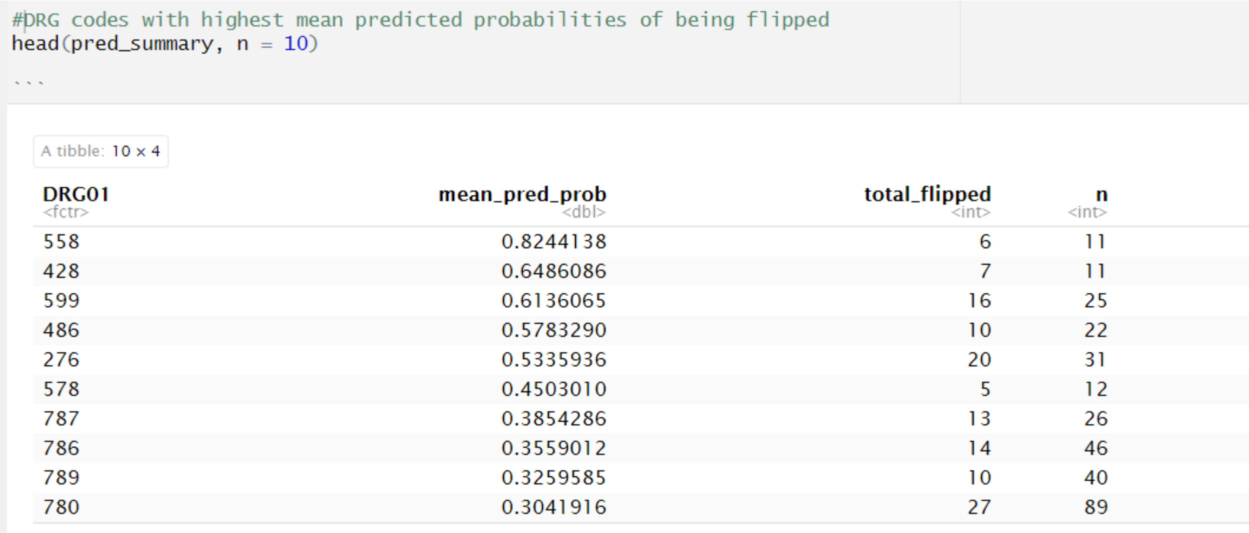


```
#Accuracy
accuracy <- mean(pred_class == testing_set$Flipped)
accuracy

```

 [1] 0.7781155
```



```
pred_class   0   1
         0 164  48
         1  25  92
Recall/Sensitivity:  0.79
Specificity:  0.77
```

MODEL 2:

For the second objective, we created a linear regression model with OU_LOS_hrs as the response variable and all other as the predictor variable to determine which DRG codes are associated with longer stays. The model has an R-squared value of about 0.29 and found numerous DRG codes related with prolonged stays, such as DRG01780 and DRG01599. Our findings show that patients admitted for specific diseases may need to stay in the OU for longer periods of time, which may have ramifications for the OU exclusion list.



```
#mean predicted length of stay for each DRG code in the test set
mean_pred_test
```

     276   428   486   558   577   578   599   780   782   786   787   789
67.34644 57.68042 49.62597 51.22369 63.03079 53.03247 88.64704 72.12220 47.96193 48.38030 69.69030 38.65308
```

MODEL DRG codes with high flip ratio:

In order to predict which DRG codes have a high flip ratio, we created a logistic regression model with the flipped variable as the response variable and the other variables as predictors. The model was then used to forecast the likelihood of a flipped DRG code for each DRG code, and the DRG codes with the highest anticipated probabilities were evaluated. Our data imply that specific DRG codes, such as DRG0555 and DRG0649, are related to a higher risk of a flipped DRG code.



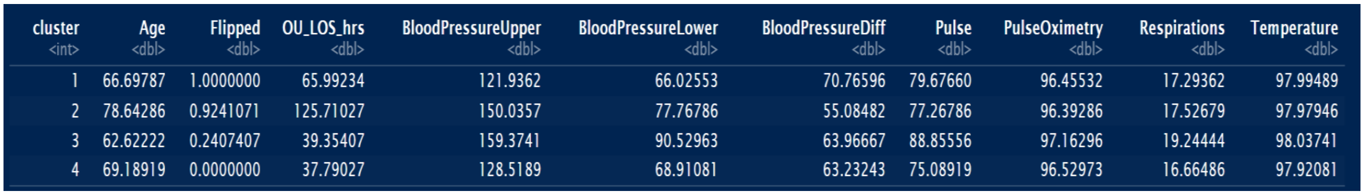
```
#DRG codes with highest mean predicted probabilities of being flipped
head(pred_summary, n = 10)

` ` `
```

A tibble: 10 × 4

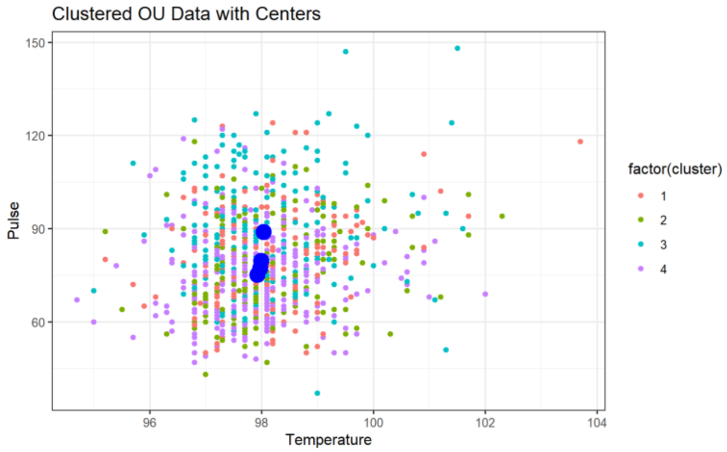
DRG01 <fctr>	mean_pred_prob <dbl>	total_flipped <int>	n <int>
558	0.8244138	6	11
428	0.6486086	7	11
599	0.6136065	16	25
486	0.5783290	10	22
276	0.5335936	20	31
578	0.4503010	5	12
787	0.3854286	13	26
786	0.3559012	14	46
789	0.3259585	10	40
780	0.3041916	27	89

The "n" column in the above table shows the total number of patients in the test set with each DRG code, regardless of whether they had a flipped DRG code or not.

MODEL3:



cluster <int>	Age <dbl>	Flipped <dbl>	OU_LOS_hrs <dbl>	BloodPressureUpper <dbl>	BloodPressureLower <dbl>	BloodPressureDiff <dbl>	Pulse <dbl>	PulseOximetry <dbl>	Respirations <dbl>	Temperature <dbl>
1	66.69787	1.0000000	65.99234	121.9362	66.02553	70.76596	79.67660	96.45532	17.29362	97.99489
2	78.64286	0.9241071	125.71027	150.0357	77.76786	55.08482	77.26786	96.39286	17.52679	97.97946
3	62.62222	0.2407407	39.35407	159.3741	90.52963	63.96667	88.85556	97.16296	19.24444	98.03741
4	69.18919	0.0000000	37.79027	128.5189	68.91081	63.23243	75.08919	96.52973	16.66486	97.92081



For the model 3, we did cluster analysis and identified that There are four different patient groups, according to the clustering analysis of the patient data. Younger patients make up Cluster 1, who also have shorter OU lengths of stay and lower blood pressure, pulse, and temperature measurements. Older individuals in Cluster 2 had average readings for most parameters, such as temperature, pulse, and blood pressure. Younger patients in Cluster 3 have longer OU lengths of stay and higher blood pressure, pulse, and temperature values. Finally, patients in Cluster 4 have longer average OU stays and lower readings for all variables except blood pressure. These results can aid medical professionals in better comprehending the patient group and modifying treatment approaches accordingly. Creating the cluster, we see that the cluster 1 are spending 65hrs in the OU and likely to flipped and are around the age group of 66 also in the other cluster 4 are around the age group of 69 and are not flipping but they stay for 37.7 hrs in the OU system.

RECOMMENDATION AND CONCLUSION:

Develop a strategy to reduce the number of patients who are likely to flip from observation unit:

Factors, such as OU_LOS_hrs, primary insurance category, and DRG code, were identified by our model as being associated with a higher likelihood of patients transitioning from observation to inpatient status. This information can be used by the hospital to create a focused intervention program to reduce the proportion of patients who are likely to flip, such as giving additional monitoring or interventions to high-risk patients.

Develop an exclusion list based on DRG codes associated with longer stays and higher flips:

We have identified DRG codes that has longer stays compared to others, hospital can use this information to develop an exclusion list and identify the customers that are not appropriate for the observation units, such patients can be excluded from observation unit, this way observation unit is rightly available.

Regularly review the flip ratio for different DRG codes:

The hospital can use our algorithm to forecast the flip ratio for various DRG codes and then monitor these ratios on a regular basis to find any trends or changes. If the flip ratio for a specific DRG code rises, the hospital can take necessary measures such as modifying the exclusion list or devising focused treatments for high-risk patients.

Using the cluster analysis:

Considering that Cluster 3 patients appear to be different from the other groups, it may be helpful to investigate why they have higher blood pressure readings and overall vital sign readings. Further research into the conditions and treatments of Cluster 4 patients may be warranted given that they have significantly lower vital sign readings. Overall, these findings can help hospital executives and medical professionals improve patient care and treatment regimens.

In terms of putting the model into action, the hospital can utilize it to identify high-risk patients and DRG codes and then apply targeted interventions or exclusion lists. These steps may result in improved patient outcomes, shorter lengths of stay, and lower healthcare expenditures. However, these consequences must be carefully monitored to ensure that the interventions do not have any unintended negative effects on patient care or outcomes.