SCMA 648 BUSINESS DATA ANALYSIS

PROJECT SUBMISSION

GROUP 3

Morgen Kiser
Ernest Washington
Saddam Hussain
Charmi Kanani

Contents

Executive Summary

The purpose of the report entails providing recommendations for updating the Montanaro policies to maximize appropriate assignment into either the observation unit or in-patient care. Over the past year, the hospital involuntarily turned away or dealt with balking patients approximately 1,900 times, resulting in an estimated $1.3 million loss in revenue. Appropriate care unit assignment results in minimized wasted resources and reduced length of stay for flipped patients, yielding less overcrowding in the emergency department and the ability to treat more patients.

We first clustered our patients into categories enabling visualization of patients and similarities in diagnostic data. The analysis uncovered the following major factors that contributed to the grouping: Age, gender, as well as private and Medicare other insurance categories. Next, we produced five models based on the historical patient data to accurately predict which patients flip. Through analysis, we concluded that our support vector machine (SVM) - radial weights model performed the best with a true positive rate of .69. One of the major insights involves the decision boundary-based models seemingly performing better than the rule-based models. We also noticed that gender and diagnostic code played a major factor in whether or not a patient flips.

Utilizing the predictive data provided we fed the data through the SVM - radial wights model to determine the probability of a patient flipping. The SVM yielded probabilities for each patient's likelihood to flip.

Based on our analysis, we recommend that the hospital disregards the current policy on placement of patients. Provided with the model and the diagnostic data from each patient allows increased accuracy regarding placement for care. If a patient has a flip probability of over .50, they need to proceed to in-patient care directly increasing hospital throughput.

Problem Introduction

The Montanaro medical facility consistently encounters overcrowding in the emergency department. Over the course of the year, the emergency department turned away roughly 1,900 patients from being seen due to the significant wait time in the emergency department. Turning away approximately 1,900 patients resulted in a $1.3 million loss in revenue per year based on the average $700 emergency room visit.

A considerable number of patients that arrive at the emergency department developed complex, but not life-threatening symptoms in which such patients require an average observation time of 48 or fewer hours. Due to the shorter length in stay and fewer resources required per patient, patients with non-life-threatening systems stay in smaller observation rooms and require considering less equipment and personnel for care and treatment.

On average, the classification proportion breakdown of patients in the observation unit are 1/3 post-surgery and 2/3 medicine service patients. On a weekly basis, the observation unit handles 44 medicine service patients per week and an approximate 45% flip to in-patient ratio takes place. The flip patients not only utilize highly coveted resources and bed space needed by other patients in the emergency department, but also utilize valuable resources to successfully hand over their care to the in-patient provider. On the contrary, on average 115 observation patients per week receive in-patient placement either due to lack of bed space in the observation unit or the observation unit "exclusion list" status of their preliminary diagnosis.

Updating the "exclusion list" based on historical data affords the hospital opportunities to reduce the number of flipped patients, resulting in increased hospital productivity and patient care. Given patient information and vitals at the time of patient check-in to the emergency department and

whether or not a patient flipped or ended up correctly placed into the observation unit, our group built a series of models to accurately predict whether or not a patient ends up flipping. By appropriately placing patients at check in, the medical facility successfully reduces the number of flipped patients, directly leading to an increase in patients serviced while maintaining the same utilization rate.

Data Pre-Processing

To conduct our analysis, we started with reading in the model data file, which obtains data from 556 observed patients in the following categories: age, gender, primary insurance, care category/flip, length of stay, diagnosis related groups, blood pressure, pulse, pulse oximetry, respiration, and temperature. Based on the listed categories, we read in the data with gender, primary insurance, care category, diagnosis related group as factors and the remaining categories as numeric.

After data read-in, we summarized the data frame observing outliers in blood pressure, respiration, pulse and length of stay. We dropped the outliers from such categories and maintained 434 observations to conduct our analysis. Of the remaining data points, we maintained one observation in no recording of temperature occurred; for modeling purposes, we changed that observation to the median temperature.

Clustering

Clustering Pre-Processing

To discover patterns within the data we conducted clustering and principal component analysis. During principal component analysis we filtered our data frame to focus on the following

categories: blood pressure upper, DRG01, blood pressure lower, gender, age, primary insurance category, blood pressure diff, pulse, pulse oximetry, respirations, temperature and length of stay. To utilize K-Means we must first convert gender, primary insurance category and DRG01 to dummy variables and scale our data to ensure uniform scaling.

Clustering Analysis

We utilized K-Means and clustered our data into three distinct groups (Exhibit 1). We then visualized each group separately as shown in Exhibit 2, 3 and 4. To determine the variables of importance for the variation in the data we applied principal component analysis (PCA). The rotation matrix (Exhibit 5) indicated an association of the first principal component (PC) with patients that have high upper blood pressure, and either private or Medicare other medical insurance. PC2 associates with a patient's gender - male or female.

Group 1 (Exhibit 2) tends to be female patients due their upper spectrum location on PC2. The grouping leans towards the positive side of PC1 so they tend to be older patients and have a higher upper blood pressure and Medicare other insurance.

Group 2 (Exhibit 3) tends to be male patients due their lower spectrum location on PC2. The grouping leans towards the positive side of PC1 so they tend to be older patients and have a higher upper blood pressure and Medicare other insurance.

Group 3 (Exhibit 4) tends to be a mixture of males and females. Group 3 is positioned on the lower spectrum of PC1 and tents to have private insurance and a younger age.

In conclusion, three distinct clusters of Black, Red, and Green clusters located at bottom-center, top-right, and almost equally distributed on top-left and bottom-left of the plot respectively appear. We discovered the following major findings obtained when comparing the clusters:

- *Black cluster patients tend to have a higher number of female patients than green and red clusters.*
- *The patients of the green cluster were found to be younger in age as compared to that of black and red.*
- *Also, the patients falling in green clusters prefer private insurance whereas the red and black clusters tend to prefer Medicare other insurance.*
- *The patients of clusters red and black have higher upper blood pressure as compared to that of patients falling under green clusters.*

Models

Model Pre-Processing

To predict future flip patients, we conducted tests on five different models to include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine - Linear Weights, and Support Vector Machine - Radial Weights. Our model data frame is based on the following variables: blood pressure upper, DRG01, blood pressure lower, gender, age, primary insurance category, blood pressure diff, pulse, pulse oximetry, respirations and temperature with a prediction of the flipped column. We maintained the same outlier parameters from above and converted "NA" observations to the median of the category. We partitioned our data into training and testing rows with 70% of our data utilized for training.

Logistic Regression

Through our logistic regression model the variables that provided the most importance was diagnostic code 786, primary insurance private, diagnostic code 780, and gender male. We received 77 accurate results and 54 inaccurate results, as shown below in the lr_predict_class confusion matrix, with a misclassification rate of .41. When conducting a ROC curve for the model the area under the curve was .66.

	lr_predict_class	
	0	1
0	49	22
1	32	28

Decision Tree

While testing the decision tree model the variables of importance were diagnostic codes, blood pressure difference, and primary insurance category. We yielded 70 accurate results and 61 inaccurate results, as shown in the rpart_predict confusion matrix below, with a misclassification rate of .46. When conducting a ROC curve for the model the area under the curve was .53.

	rpart_predict	
	F	DNF
F	32	28
DNF	33	38

Random Forest

Although the random forest model yielded better results than the decision tree model, the logistic regression model demonstrated stronger accuracy. We found that the main variable of importance when conducting random forest was DRG01. The model analysis resulted in receiving 73 accurate results and 58 inaccurate results, as shown in the predict_rf confusion matrix below, with a

misclassification rate of .44. Analysis of the ROC curve led to discovering a .58 area under the curve.

	predict_rf	
	F	DNF
F	26	34
DNF	24	47

Support Vector Machine – Linear Weights

Although the support vector machine - linear weights model yielded more accurate results than the previous two models, the logistic regression model performed stronger. The support vector machine - linear weights model creation utilized the best parameters through hypermeter tuning. Analysis of the model indicated 75 accurate results and 56 misclassifications, shown below in the predict_svm confusion matrix, with a misclassification rate of .43. Conducting ROC analysis, we noted an area under the curve of .64.

	predict_svm	
	F	DNF
F	26	34
DNF	22	49

Support Vector Machine – Radial Weights

We also created a support vector machine - radial weights model. The model yielded 82 accurate results and 49 inaccurate results, as shown in the predict_svm_rbf confusion matrix below, with a misclassification rate of .37. We used hyperparameter tuning to tune the parameters. Through producing a ROC curve the area under the curve was .69.

	predict_svm_rbf	
	F	DNF
F	28	32
DNF	17	54

Results

Based on the misclassification rate and the area under the curve the support vector machine model - radial weights performed the best of all four models we created. Through testing the model based on the area under the curve, we determined that the model provides accurate results 69% of the time. As demonstrated in exhibit six, the rankings of the models from best to worst is the following: Support vector machine - radial weights, logistic regression, support vector machine - linear weights, random forest, and decision tree. According to the models, we found that the variables of gender and diagnostic code demonstrated the most importance in determining a patient's flip status. Exhibit 7 indicates a higher likelihood of male patients flipping than female patients. Of interest, however, involves the observation via Exhibit 8 that more females on average enter into the emergency department than males. Therefore, although females do not flip as consistently as males, a higher number of females in comparison to males flip and go in-patient during their stay. Furthermore, via Exhibit 9, initial diagnosis codes of 558, 577 and 599 have the highest rates of flipping.

Conclusions

We observed that the decision boundary-based models (SVM RBF, SVM Linear, Logistic Regression) tend to perform better than that of rule-based models (Classification Tree, Random Forest). The superior performance indicates the separability of data with decision boundaries.

Prediction Pre-Processing

While conducting predictions on data we conducted the same preprocessing as we did in the earlier data set. We received "NA" data in the temperature, pulse oximetry, blood pressure difference, pulse, respiration and blood pressure upper data; we utilized the median for each category to

replace the “NA” data. We then created dummy variables for gender, primary insurance category, and the diagnostic category and scaled the data in order to ensure our SVM - radial weights model functions correctly.

Prediction Conclusion

We conducted our prediction analysis using the SVM - radial weights model. For each observation we provided a prediction value of whether or not a patient flipped. We assume that the prediction of .50 or higher yields a higher likelihood that a patient flips. Thus, our recommendation for future policy involves utilizing the vitals data received when a patient enters the emergency department. The staff needs to be instructed to run the data through the SVM - radial weights model and if a predicted flipped value of greater than .50 occurs, the patient needs to be directed to in-patient rather than the observation unit. As a result, the hospital minimizes wasted resources due to flipped patients and affords the emergency department increased time and resources to successfully treat more patients.

Appendices

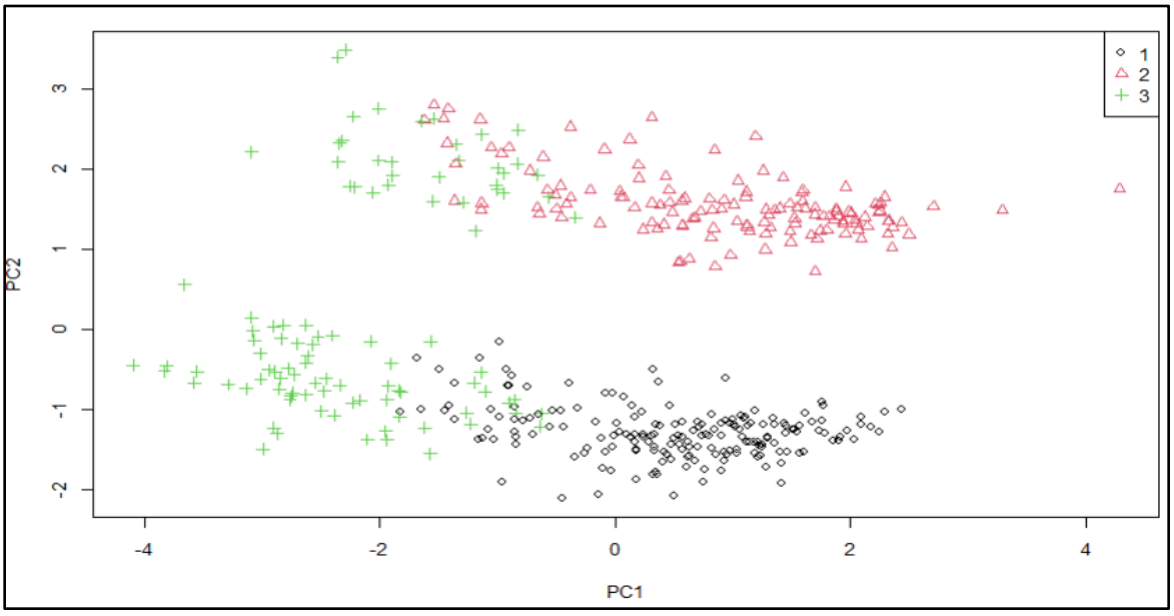

Exhibit 1: K-Means Principal Component Analysis (PCA) Plot

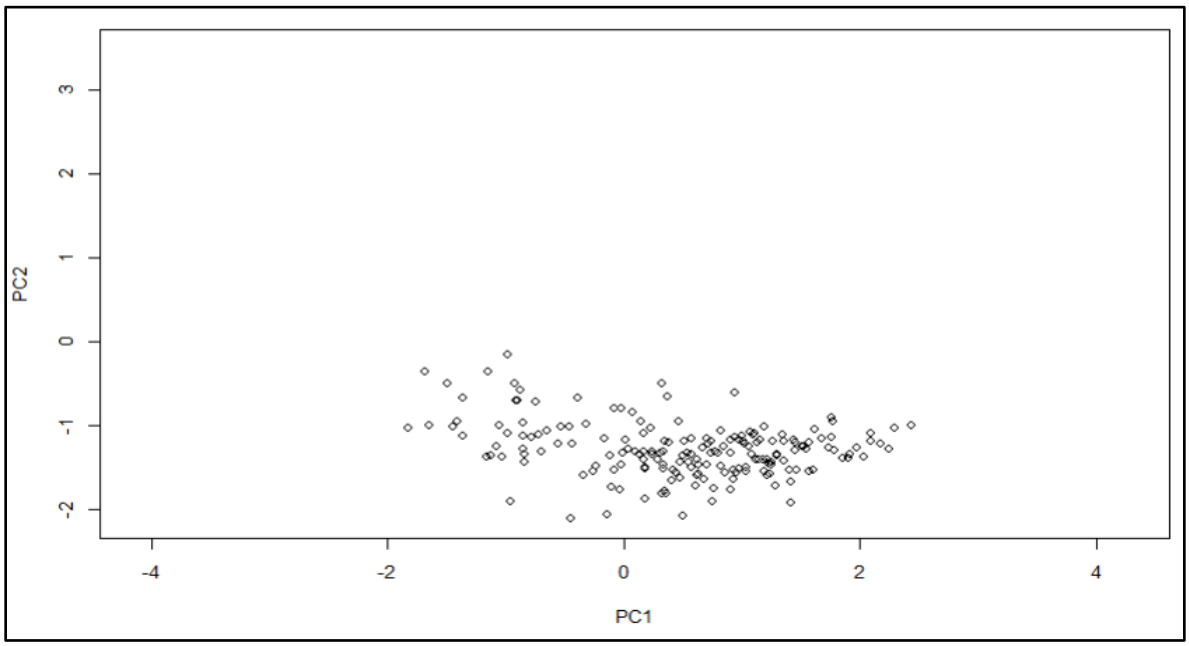

Exhibit 2: K-Means PCA Group 1

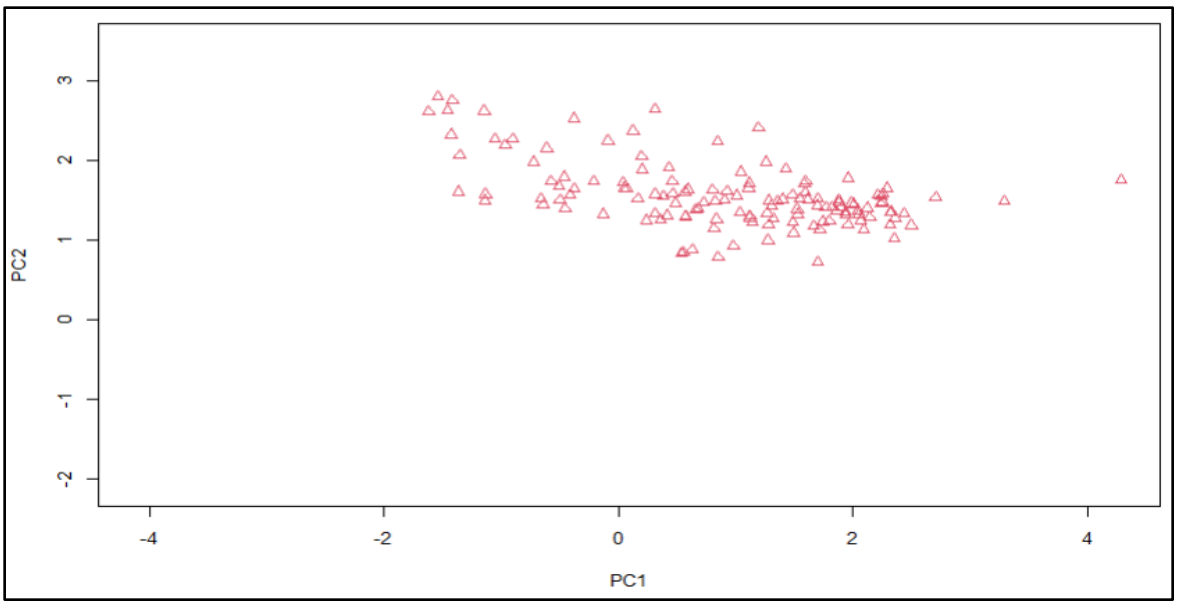

Exhibit 3: K-Means PCA Group 2

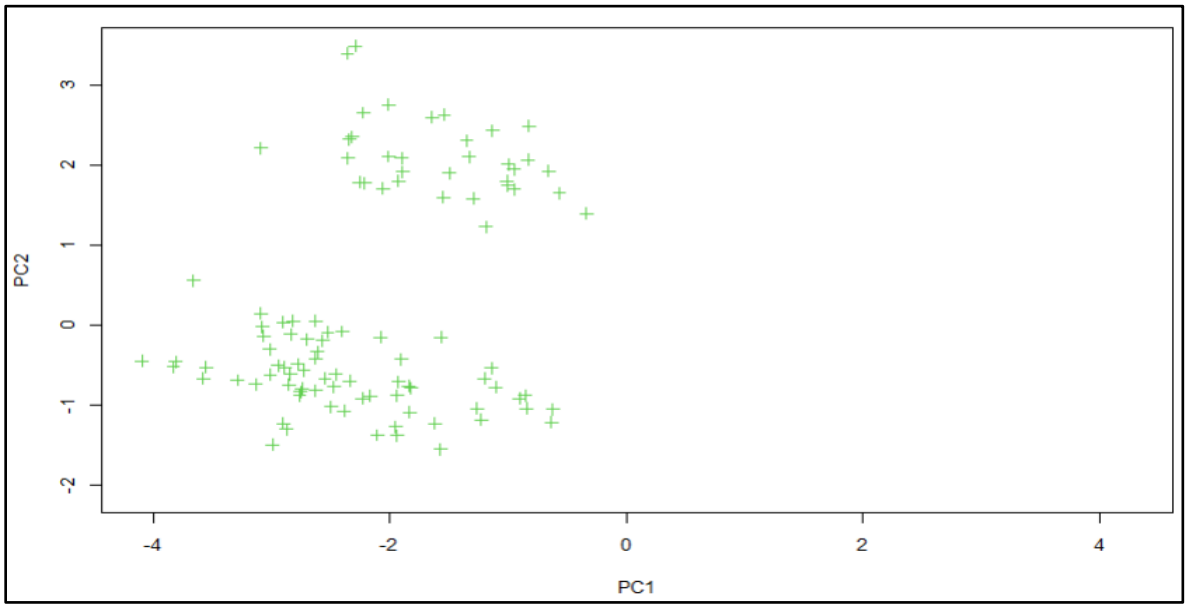

Exhibit 4: K-Means PCA Group 3

	PC1	PC2
BloodPressureUpper	0.078115416	-0.069547057
DRG01276	0.033794274	-0.124461003
DRG01428	0.036704399	0.014591325
DRG01486	0.078169572	0.006567542
DRG01558	-0.085185323	-0.084182204
DRG01577	-0.107689537	0.137109465
DRG01578	-0.033872537	-0.025320437
DRG01599	0.100360574	0.041191339
DRG01780	0.246270432	-0.018692090
DRG01782	-0.006554491	-0.023331491
DRG01786	-0.064966220	0.003757293
DRG01787	-0.116815150	0.069566337
DRG01789	-0.277172863	0.052777314
BloodPressureLower	-0.051161515	-0.047718153
GenderFemale	-0.135556268	-0.656658566
GenderMale	0.135556268	0.656658566
Age	0.534143383	-0.161378471
PrimaryInsuranceCategoryMEDICAID OTHER	-0.236998388	0.110147048
PrimaryInsuranceCategoryMEDICAID STATE	-0.137931997	0.154040940
PrimaryInsuranceCategoryMEDICARE	0.017282119	-0.098135246
PrimaryInsuranceCategoryMEDICARE OTHER	0.436556753	-0.051944577
PrimaryInsuranceCategoryPrivate	-0.348977229	-0.007472628
BloodPressureDiff	0.023282925	-0.026184543
Pulse	-0.088829971	-0.023455778
PulseOximetry	-0.102383608	-0.003593568
Respirations	0.038656434	-0.014050503
Temperature	-0.042941645	-0.002741128
OU_LOS_hrs	0.255051995	0.047329926

Exhibit 5: Rotation Matrix

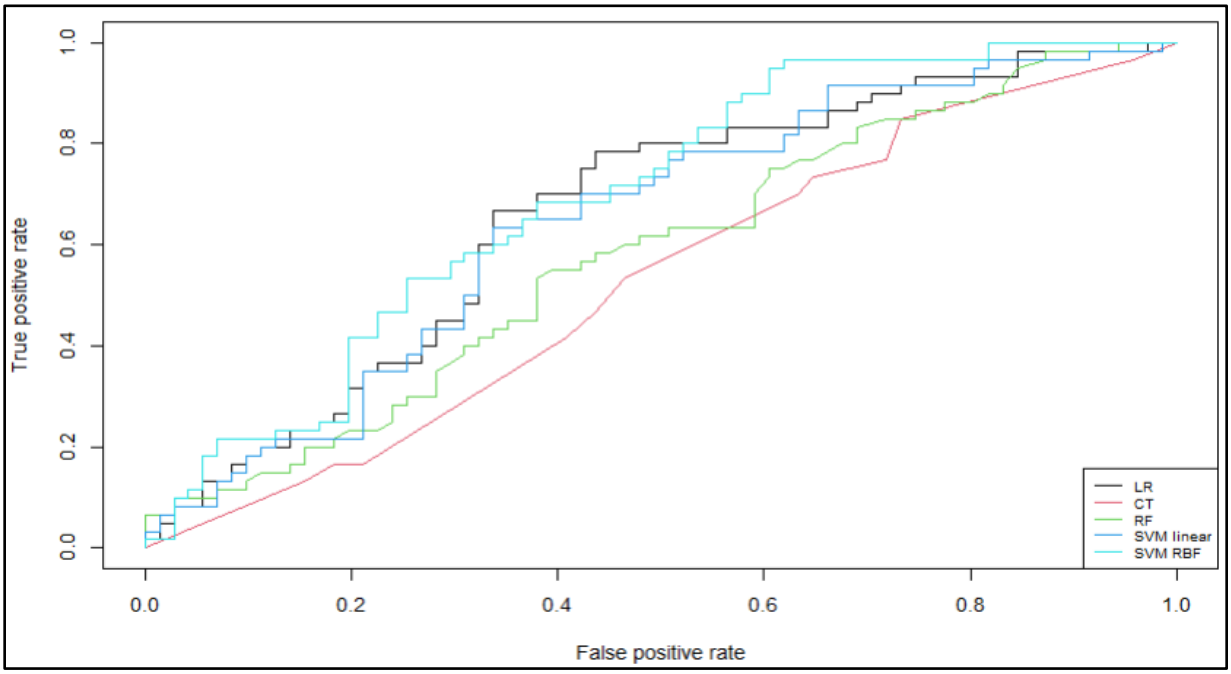

Exhibit 6: ROC Curve for Models

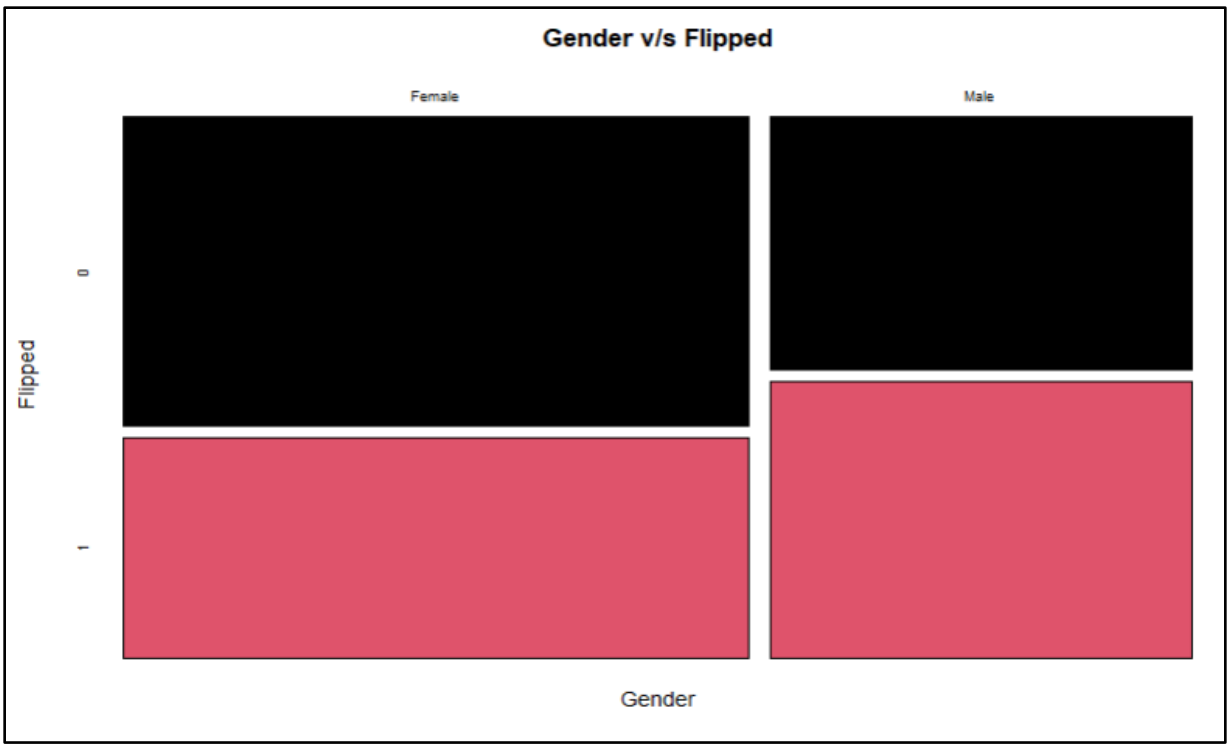

Exhibit 7: Gender vs. Flipped

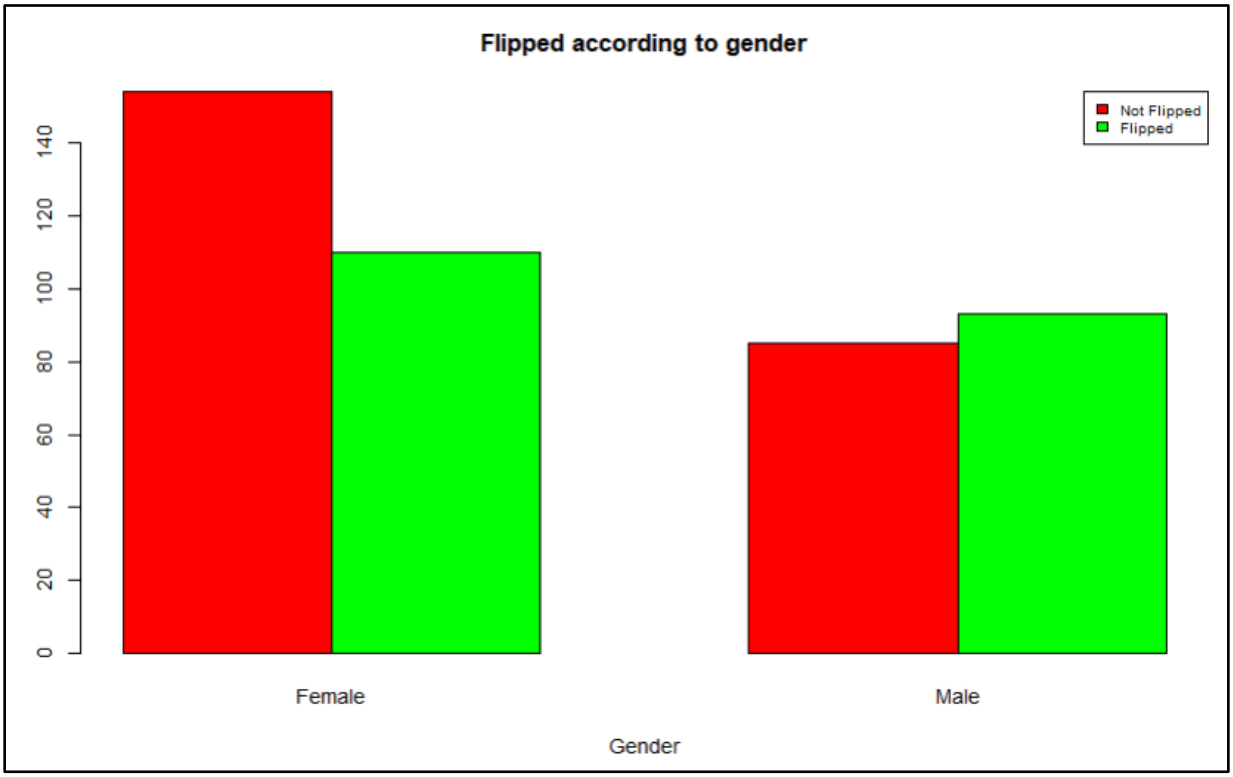

Exhibit 8: Bar Chart (Flipped According to Gender)

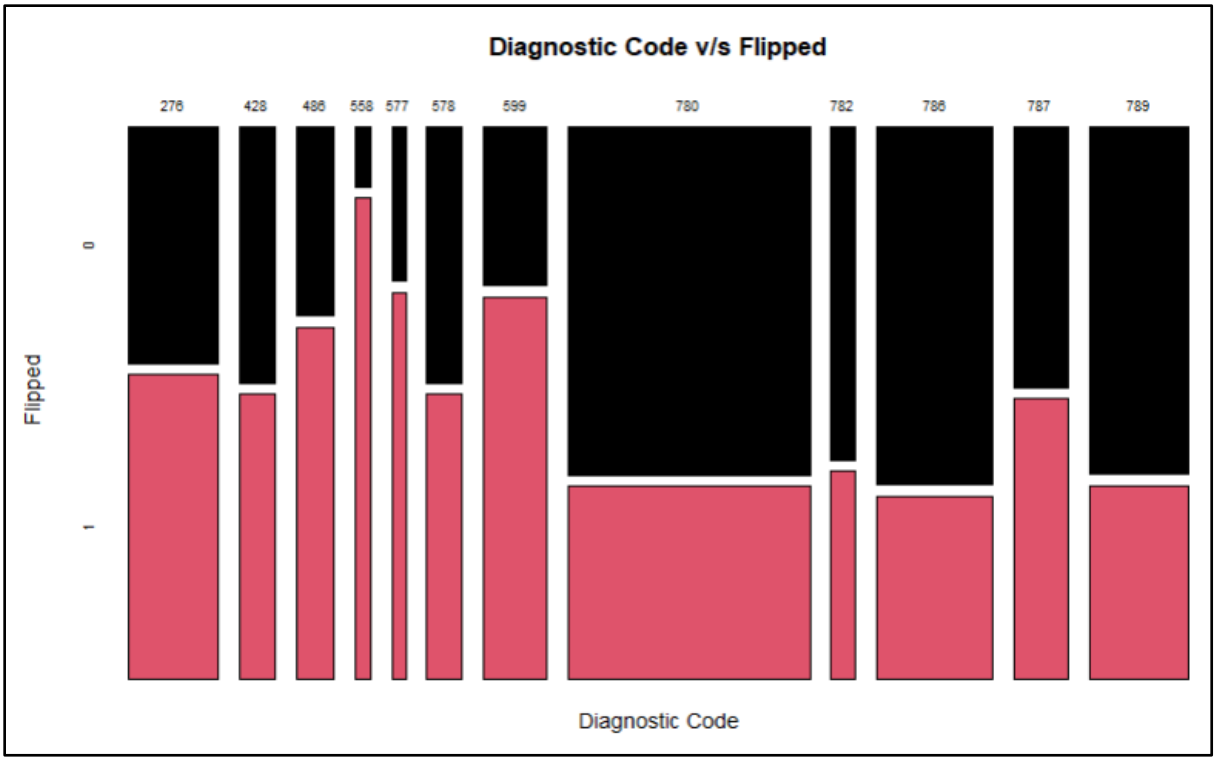

Exhibit 9: Diagnostic Code vs. Flipped