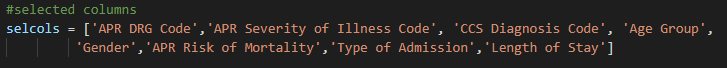
\*\*THIS DOCUMENT IS INTENDED AS GENERAL NOTES AND THOUGHTS\*\*

To improve accuracy, we decided to undersample and oversample our data in order to obtain a more balanced dataset. After running the random forest tree model, we determined that the accuracy was not improving at a meaningful level. Consequently, we turned to feature engineering to determine the most important features for the model. Based on the features originally selected, we determined that the APR\_DRG Code and CSS Diagnosis Code were by far the most influential for the model.

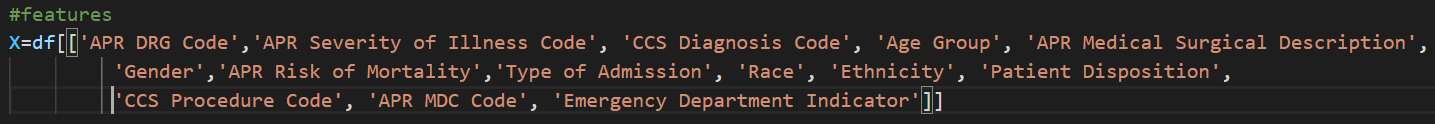


The least influential were Gender, Age, and Type of Admission, therefore we removed them from the dataframe. After running the Random Forest model with only 4 features, we saw an 1% increase in accuracy. (Based on APR DRG Code, APR Severity of Illness Code, CCS Diagnosis Code, and APR Risk of Mortality)

Graphical user interface, application, table, Excel

Description automatically generated

The next step was to preprocess all columns and run the feature selection again across a more diverse dataframe.





Ranked from most influential feature to least

|  |  |  |
| --- | --- | --- |
| 1 | APR DRG CODE | 768300 |
| 2 | CCS DIAGNOSIS CODE | 746800 |
| 3 | CCS Procedure Code | 240300 |
| 4 | Patient Disposition | 49310 |
| 5 | APR MDC Code | 20770 |
| 6 | APR Risk Mortality | 18810 |
| 7 | APR Severity of Illness | 17960 |
| 8 | Age Group | 10530 |
| 9 | Emergency Department Indicator | 2326 |
| 10 | Type of Admission | 1510 |
| 11 | APR Medical Surgical Description | 364.9 |
| 12 | Gender | 341 |
| 13 | Race | 140 |
| 14 | Ethnicity | 10.24 |

The Code Responsible for obtaining the above data

Text

Description automatically generated

34 Columns make up the excel sheet, however only 14 features were analysed. This is because many of the columns used in the excel spreadsheet are simply “Descriptions” of medical codes, as well as data that would have been added after the patient had already been discharged. Other columns like “Health Service Area,” “Hospital County,” “Zip Code,” etc, were also not included because they are all location-based features that aren’t applicable for the individual patient.

The CCS Procedure Code, a feature we have not yet used for training, turned out to be #3 most influential in the models predictive ability. Though not as influential, Patient disposition ranked #4 best, another feature we hadn’t included originally. Patient disposition will not be included in the dataframe, because it refers to the location the patient is being discharged. In other words, we only are interested in features that will be available during the patients stay at the hospital. The next step was to filter the dataframe with only the best features and remove the bottom 7. We may remove or add more features to increase accuracy later.

After determining the top 7 features to be used, we re-ran the model to see if the Random Forest model increased in accuracy. Below are the results:

Graphical user interface, application, table, Excel

Description automatically generated

Table

Description automatically generated

Comparing the results with the table data below, we found a similar increase in accuracy using the new 7 features, compared to when we only used the top 4 of the original feature set. Undersampling and Oversampling datasets from more than 2 groups appears to negatively affect predictive accuracy. Though there is still an increase in accuracy for two groups, the predictive model is most effective using the original, un-altered dataset.

Table

Description automatically generated

Below is a table detailing the original 7 features we used before feature selection, as well as the new features we decided to continue with.

|  |  |
| --- | --- |
| Features | |
| Original | After Feature Selection |
| APR DRG Code | APR DRG Code |
| APR Severity of Illness Code | APR Severity of Illness Code |
| CCS Diagnosis Code | CCS Diagnosis Code |
| Age Group | APR Risk of Mortality |
| Gender | Age Group |
| APR Risk of Mortality | CCS Procedure Code |
| Type of Admission | APR MDC Code |

When we originally did our feature selection, we saw that accuracy increased when we removed the 3 least important features from the dataframe. The next thing to test is if accuracy improves if we only use 4/5/6 of the best features from the new dataframe. These tests will only be on the original data, un-altered by under or oversampling as we have determined it to have a negative impact on accuracy. The results of testing only the best 4/5/6 features are shown below.

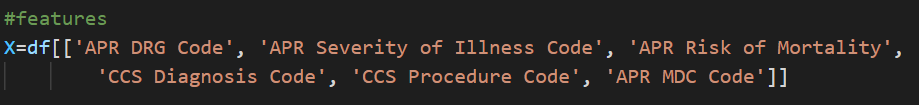
|  |  |  |  |
| --- | --- | --- | --- |
| # of Features | Two Groups | Three Groups | Feature Removed |
| 7 | .76495 | .66385 | N/A |
| 6 | .76779 | .66608 | Age Group |
| 5 | .76602 | .65842 | APR Sev. Of Illness |
| 4 | .74606 | .63092 | APR Risk of Mortality |

The next thing we can examine is how the accuracy is affect only using the top 6 features, across all prediction models.

Table

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Ranges: 1-3, 4- | | Ranges: 1-3, 4-8, 9- | |
| Classifier | Accuracy | Classifier | Accuracy |
| LR | 0.7420 | LR | 0.6388 |
| DT | 0.7659 | DT | 0.6602 |
| RF | 0.7679 | RF | 0.6658 |
| GBC | 0.7593 | GBC | 0.6845 |
| ABC | 0.7548 | ABC | 0.6703 |



Using only 6 features, we found an overall increase in accuracy among classifiers. Random Forest had the highest accuracy at 76.79% for two groups and the Gradient Boosting classifier had the highest accuracy of 68.45% for 3 groups.