Part l: Research Question

A.1. One question relevant to a real world organization is can we use the provided information to predict whether or not a patient will be readmitted within 30 days of their initial discharge? I will answer it using random forests.

A.2. One goal of my data analysis is to predict whether or not someone will be readmitted from the provided information, generally taken at check in.

Part II: Method justification

B.1. The prediction method I chose for this dataset is the random forest classifier. It uses decision trees which are a series of selections describing something’s attributes. The algorithm records each outcome and the input variables. When predicting future outcomes, it compares the input to each decision tree from the training data. I expect that we will be able to make predictions with at least 80 % accuracy.

B.2. There are no distributional assumptions for the random forest method. “Random forests are non-parametric and can thus handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal,” (Yiu, 2019).

B.3. The libraries I will be using in this analysis and their benefits are as follows:

* Pandas - I used the pandas library to read my csv and to concatenate columns with dummy variables
* Numpy - I used the numpy library to convert my lists into arrays
* Matplotlib - I used matplotlib to create my visualizations of data
* Seaborn - I used seaborn to create a heatmap with my correlations between variables
* SKlearn random forest classifier - I used the random forest classifier to execute the algorithm
* SKlearn standard scaler - I used StandardScaler to standardize my data
* SKlearn preprocessing - I used this library to binarize columns
* SKlearn train\_test\_split - I used this package to split the data into X and Y and train and test groups
* SKlearn classification\_report - I used this to view my accuracy statistics
* SKlearn mean\_squared\_error - I used this to calculate my MSE for the model

Part III: Data Preparation

C.1. One goal for the preprocessing step is to get dummies for the categorical variables before dropping the original columns.

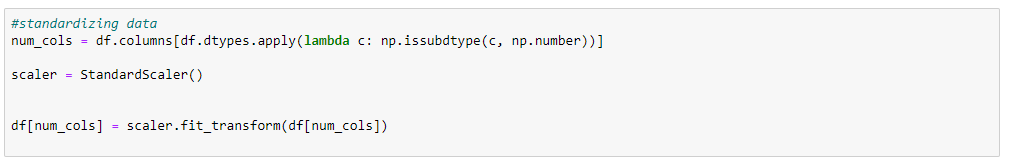
C.2. Here are the variables I will be including in my first model:

* Age - continuous
* Income - continuous
* ReAdmis - categorical
* VitD\_levels - continuous
* Doc\_visits - categorical
* Soft\_drink - categorical
* HighBlood - categorical
* Stroke - categorical
* Overweight - categorical
* Arthritis - categorical
* Diabetes - categorical
* Hyperlipidemia - categorical
* BackPain - categorical
* Anxiety - categorical
* Allergic\_rhinitis - categorical
* Reflux\_esophagitis - categorical
* Asthma - categorical
* Initial\_days - continuous
* TotalCharge - continuous
* Additional\_charges - continuous
* Area - categorical
* Marital - categorical
* Gender - categorical
* Initial\_admin - categorical
* Complication\_risk - categorical
* Services - categorical

C.3. The first step I took for data preparation is to drop the columns I won’t be using.



Next I standardized the data using StandardScaler.



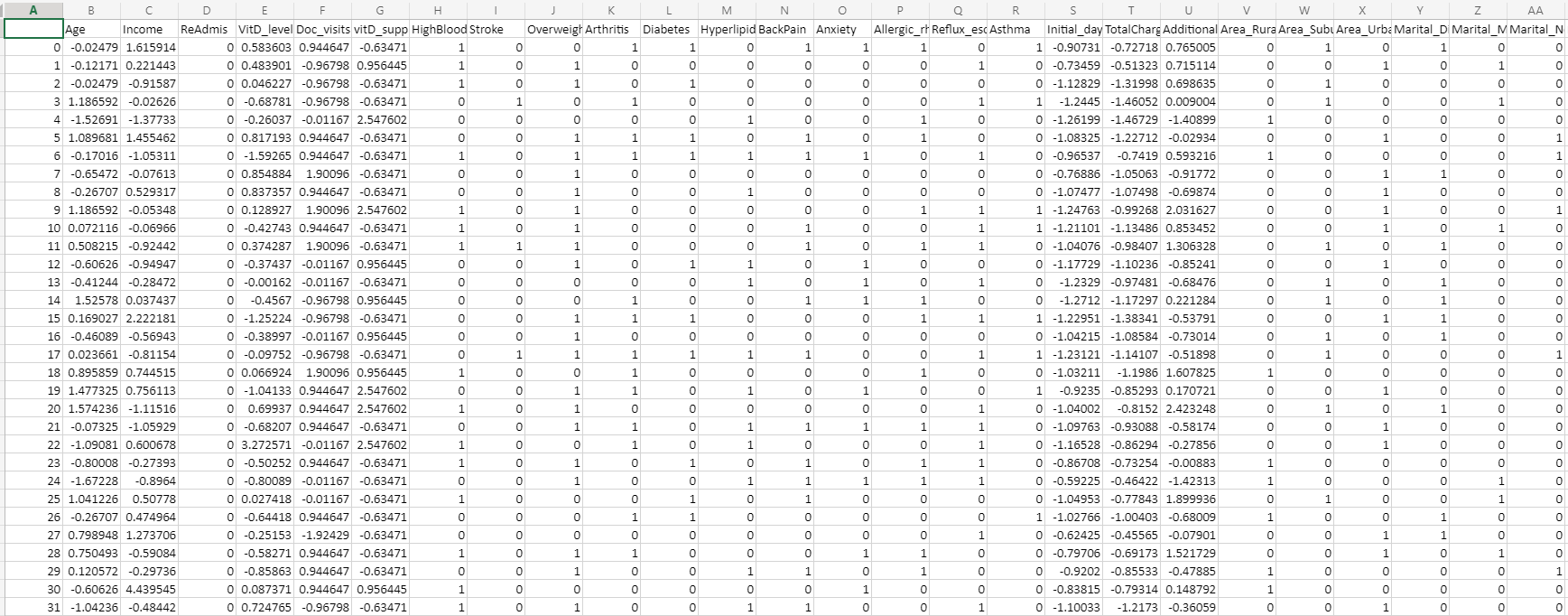
After that I got dummies for the categorical variables.



Finally, I wrote the prepared data to a csv.

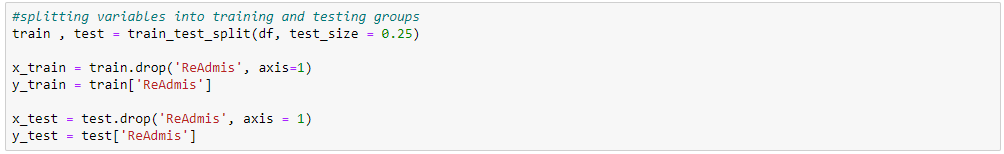


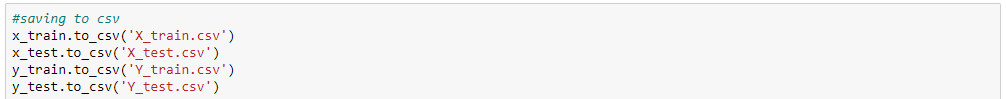
C.4. Below is a snippet of the cleaned dataset included in this submission.



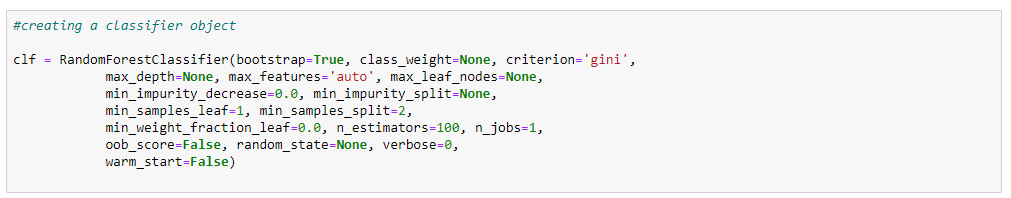
Part IV: Analysis

D.1.Below, I split the data into testing and training data. The files are included in this submission.





D.2. The analysis technique I used here is random forest classification. After cleaning and splitting my data I created the classifier object.



After that I made predictions on the testing data.



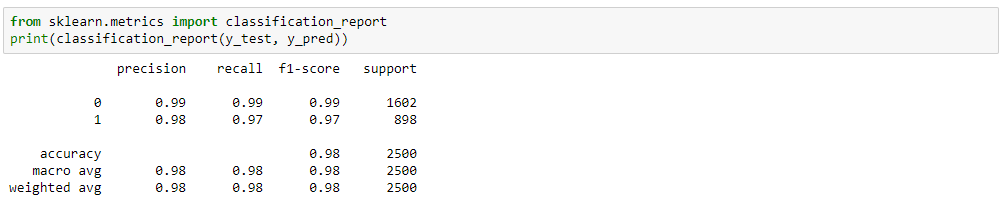
Here I am viewing the predicted and actual values.



Next I checked the accuracy of the model. It got a score of .98 which is really great.



After that I viewed the summary of the classification report.



Finally, I used SKlearn to calculate the MSE.



D.3. The code used to perform the analysis is included in the .ipynb file of this submission.

Part V: Data Summary and Implications

E.1. The accuracy of my model is about 98.1%. For mean squared error, I got a value of 0.0192. We want the MSE to be as close to 0 as possible so our model is pretty reliable and performing well.

E.2. The result of this analysis is that we can predict with confidence whether or not someone will be readmitted to the hospital within 30 days of their initial discharge based on the information given. The implication of the accuracy is that this model is accurate enough to be used in a real world scenario.

E.3. One limitation of this analysis is the quantity of data. There are only 10,000 entries which is a very small percentage of hospital admissions on any scale.

E.4. I recommend collecting data to increase the number of observations. Hospital wide data would be a much larger scope and provide a better basis for predictions.

Part VI: Demonstration

F. A Panopto recording is included in this submission.

G. Third party code referenced

*Random forests classifiers in python*. DataCamp Community. (n.d.). https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

H. Works Cited

Yiu, T. (2019, August 14). *Understanding random forest*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.