Michael Erwin

Data Analytics

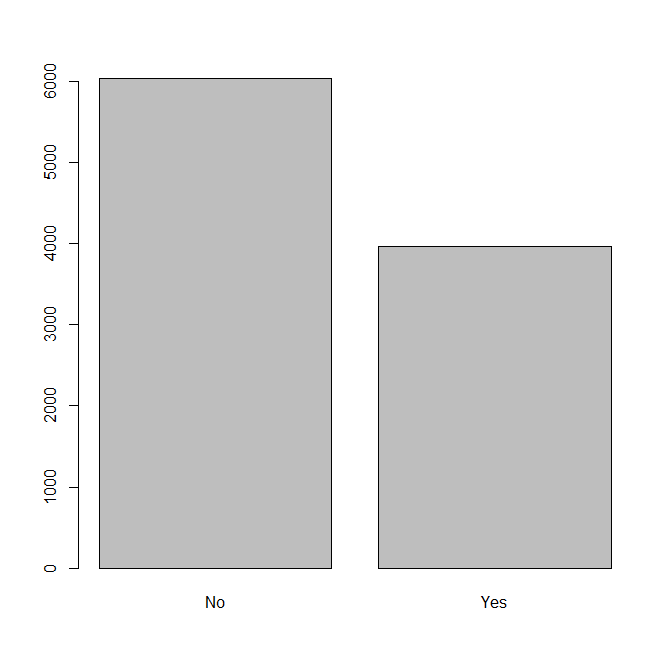
Final Report

I used the slightly preprocessed version of the dataset. Before I did preprocessing, there were several NA values, including “?”, “Not Available”, and simply blank fields, for example. Columns dropped include Weight, Payer Code, Medical Specialty, Change of Medications, and 23 Features for Medications.

I used kNN imputation for the NA values, which basically means kNN guessed what the values should be.

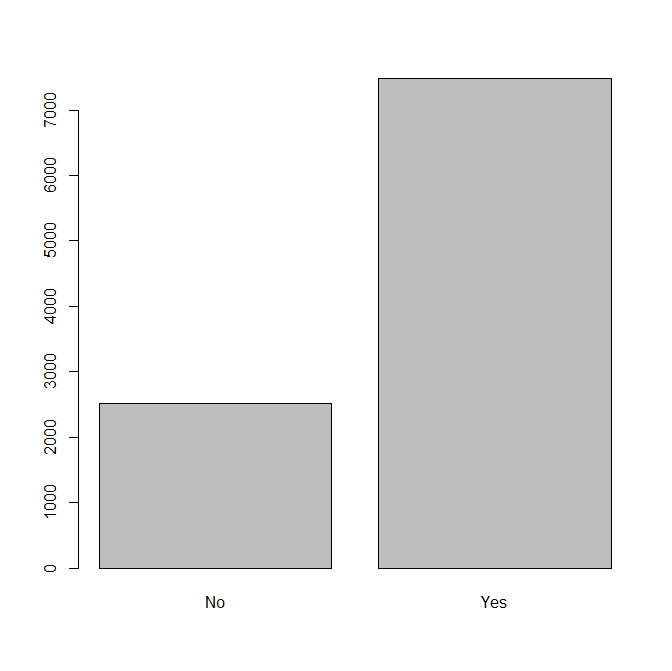
I then took a subset of the data using relevant columns. Columns include "age", "race", "gender", "admission\_type\_id", "discharge\_disposition\_id", "diabetesMed", and "readmitted.”

Here is a graph showing the amount of readmitted patients vs patients who weren’t readmitted:



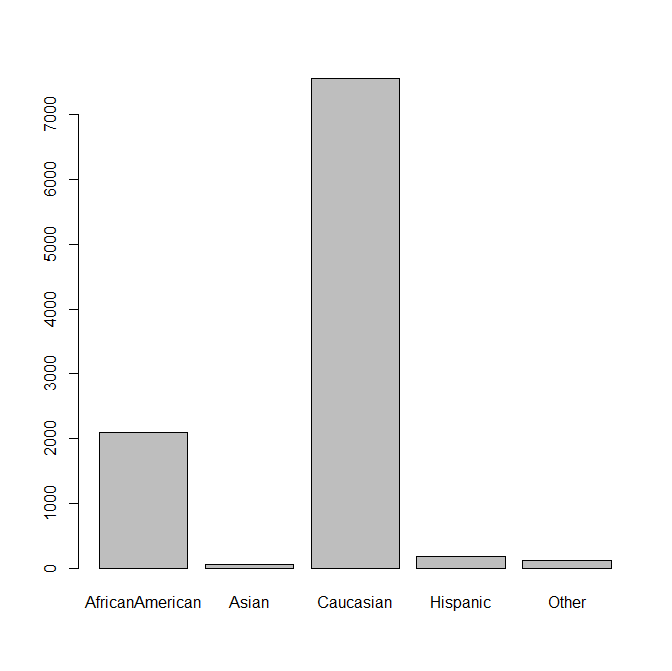
60.35% of patients were not readmitted, and 39.65% of patients were readmitted.

Here is a graph showing the number of patients who had diabetes medicine prescribed or not:



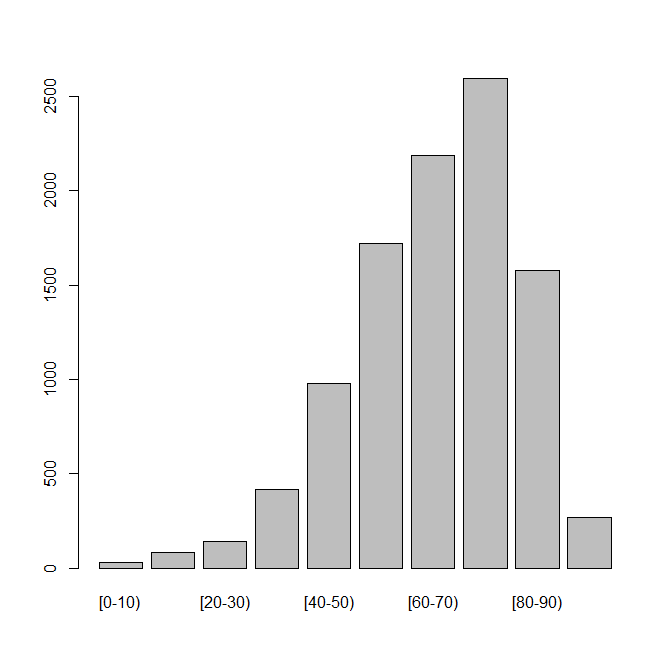
25.22% of patients weren't taking diabetes medicine, and 74.78% of patients were taking diabetes medicine.

Here is a graph showing how many patients of each race there were:



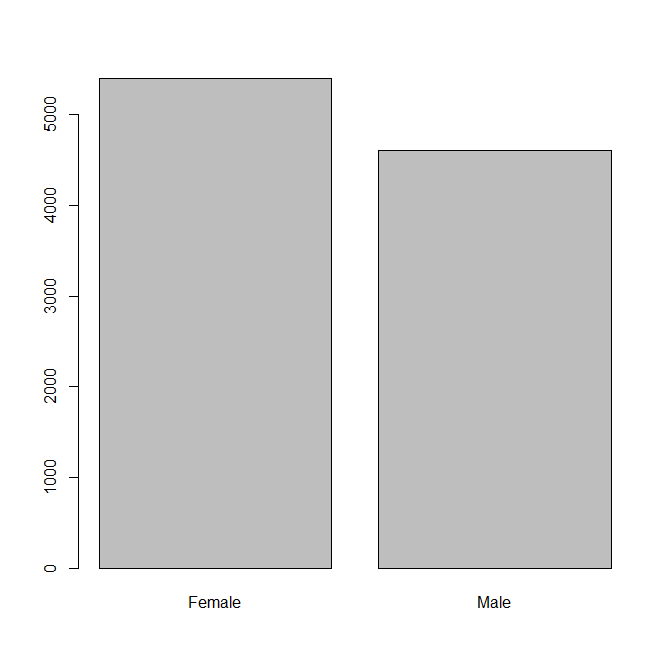
We might be able to assume that this hospital was located in an area where Caucasians made up a majority of the population.

Here is a graph comparing the age groups of the patients:



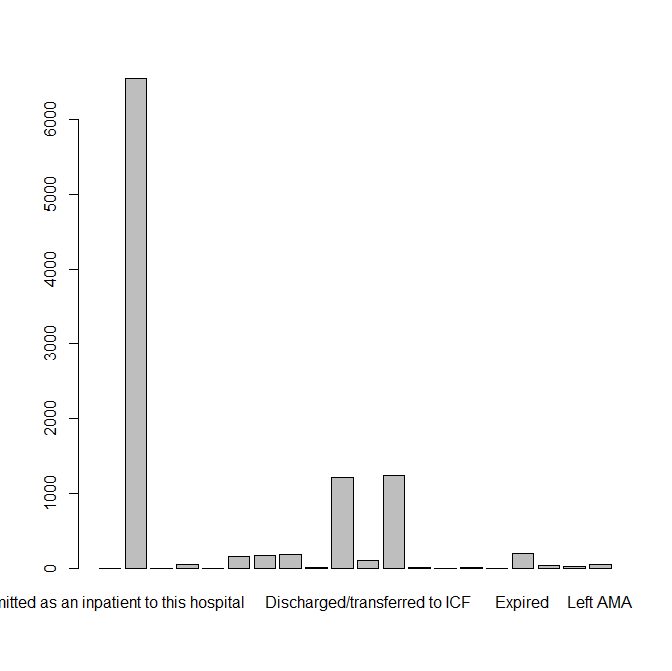
As we should expect, a majority of the patients were in the older age ranges.

Here is a plot comparing the amount of male patients vs female patients:



53.98% of patients were female, and 46.02% of patients were male.

Here is a graph comparing the discharge disposition ids:



This isn’t especially useful. Since there seems to be many different types, I simply calculated the number of patients who expired compared to everyone else.

Of all 10,000 patients, 1.96% of them passed away.

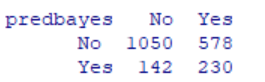
Next, I did some pattern mining. I determined the rules that would cause a patient to not be readmitted. Obviously, if a patient has deceased then they aren’t able to be readmitted. Therefore, patients with the status of expired have a 1.0 (100%) confidence to not be readmitted.

Here are some rules for not being readmitted. Being in the age group 30-40, female, and admission type being elective (not emergency) is the rule with the highest lift (1.423985) that after the rules involving the patient being deceased. This rule has a support of 0.0055 and a confidence of 0.859375. The next rule group is the same and has the same values and rules, except instead of ‘gender=female,’ there is a ‘race=Caucasian’ rule. The next few are similar with admission type being elective. A few rules down after these are interesting, however. This rule says age group is 40-50, race is Caucasian, gender is Male, and admission type=Urgent. This rule group has a lift of 1.325601, a support of 0.006, and a confidence of 0.8. This rule makes sense because this age group is when men can begin to have more serious health problems, hence the admission type of urgent.

Then I looked for rules for patients who were readmitted. The rule with the highest lift (1.611321) had patients who were Caucasian, female, had an admission type of urgent, and were discharged to home with health services. This rule has a support of 0.0069 and a confidence of 0.6388889. The next rule is the same, except the patient also is taking diabetes medicine. This rule has a lift of 1.587969, a support of 0.0051, and a confidence of 0.6296296. This makes sense. If someone has diabetes, then they will likely have continuing health problems.

Next, I looked at rules for patients who have deceased. The rule with the highest lift entails that the patient be in the age group of 80-90 and not be readmitted. This rule has a support of 0.0052, a confidence of 0.05922551, and a lift of 3.0063712. This rule makes obviously makes sense because people in that age group are at the end of a normal human’s lifespan. The next highest rule entails that the admission type is an emergency, that the patient not be taking diabetes medicine, and that the patient not be readmitted. This rule has a support of 0.0057, a confidence of 0.05864198, and a lift of 2.9767500. This rule is interesting. Admission type being an emergency definitely makes sense, and perhaps the patient was not taking diabetes medicine even though should have been.

Next, I trained some models to predict whether or not a patient would be readmitted. For this I used the whole dataset, not the subset. I split the data into 80% for training the model and 20% for validating the model. I used the naïve Bayes model first. Here is a comparison of the naïve Bayes model’s prediction of the readmitted variable compared to the actual readmitted variable in the testing data:



True negatives are when we predict a ‘no’ and the actual value was ‘no,’ false negatives are when we predict a ‘no’ but the actual value was a ‘yes,’ true positives are when we predict a ‘yes’ and the actual value is a ‘yes,’ and false positives are when we predict a ‘yes’ but the actual value is a ‘no.’ The values in green are “good” as in we want them to be high, and the red values are “bad” as in we want them to be low.

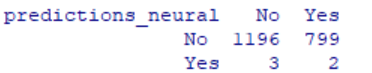
**True negatives: 1050**

**True positive: 230**

**False negatives: 578**

**False positive: 142**

I feel that the number of false negatives given by the naïve Bayes model has the potential to give concern. Next, I tried a neural net model. Here is the comparison of the model’s predictions to the actual values.



**True negatives: 1196**

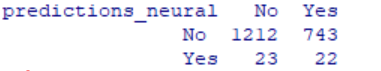
**True positive: 2**

**False negative: 799**

**False positive: 3**

We can see that the number of true positives was extremely low, and the number of false negatives was very high. This model did not perform as well as the naïve Bayes model.

Next, I tried to train a neural net model with just the subset of the data used earlier with the thought that the full data set might be causing the curse of dimensionality. Here is the table comparing predicted vs actual values.



**True negatives: 1212**

**True positive: 22**

**False negative: 743**

**False positive: 23**

This is better than the previous neural net model, but the naïve Bayes model still had a significantly higher number of true positive values. This model doesn’t seem to be a good option for predicting whether or not a patient will be readmitted.