Springboard Project 8.6 (Apply Inferential Statistics)

Capstone Project on Diabetes Readmission - Exploratory Data Analysis

This part of the project investigates relationships between variables in the (cleaned) data set through the use of plots and descriptive statistics (e.g., means, proportions, distributions). In particular, I sought to identify relationships like correlation and dependence since they are often early indicators of predictive value. Additionally, I tested for significance of emergent trends where appropriate to make sure they did not occur by chance alone. The final issue of concern in my exploratory analysis is multicollinearity. In a few cases, I looked to see if independent variables were highly correlated such that one might need removal to prevent bias in final analyses.

To start, I checked the balance of the dependent variable and found that the ratio of negative (not readmitted within 30 days) to positive (readmitted within 30 days) outcomes is roughly 10:1. Knowing the overall likelihood of being readmitted within 30 days provides useful context in looking at the independent variables because if the balance is much different from 10:1 within a particular feature, that feature may have significant predictive value.

Next, I checked for multicollinearity, starting with a correlation matrix for number of procedures, number of medications, number of emergencies, and number of visits to a medical facility -- all features that could correlate highly with each other. None of the plots in the matrix showed much correlation, so multicollinearity is not likely an issue for any pairs of these features. That said, it is worth noting that number of emergency admissions and number of medications show a curved, downward relationship, which indicates that these two variables may be exponentially -- though not linearly -- related.

I was also interested in the distribution of patient ages for context in my analyses, as it is one of the few demographic features available. Because many other medical variables could correlate with age, I looked at age by itself before checking its relationships with other features. The histogram showed patients skewing older, meaning that other variables with higher values for older patients could correlate highly with age and/or each other. When I checked age against number of medications, the scatter plot indicated that there are more people in higher age groups taking higher

numbers of medications, but all age groups showed heavy clustering in lower numbers, too. It seemed unlikely that age and number of medications were so highly correlated that one variable would need removal.

I moved on to see if an association existed between the number of days a patient spends in the in hospital and the number of procedures he or she has. A grouped bar graph showed that as the number of days in the hospital increased, frequency rose quickly before dropping swiftly for patients with lower numbers of procedures performed. However, the converse was not true: the frequency of patients with higher numbers of procedures did not increase with the number of days spent in the hospital. There was not much evidence for correlation or multicollinearity.

The initial researchers using this data set were interested in determining if patients who had HbA1c levels tested were significantly less likely to be readmitted within 30 days. As such, I also checked for correlation between HbA1c testing and readmission. HbA1c result has four categories: 3 different levels and none (meaning that the test wasn't performed). The initial researchers were interested in whether or not merely having this test done was significant, but I kept the 4 levels separated initially to see if a more nuanced approach showed any correlation. A paired bar graph showed a somewhat steady balance in readmissions across the levels of HbA1c tests. The test results may be a significant predictor of readmissions, which the original researchers found to be the case with HbA1c results in a binary format. A bit more number crunching showed that 8.6% of tested patients were readmitted within 30 days, compared to 9.2% of non-tested patients, and a Chi-square test for independence indicated that the difference was not due to chance.

Finally, since a new variable -- the number of visits by one patient to a medical facility -- was created during data wrangling, I previewed its relationship with the outcome variable using another paired bar chart. Overall, the graph showed an exponential, negative trend between readmissions frequency and number of hospital visits. This newly created variable may well have some predictive power. I will learn more about it as I proceed to final analysis methods like logistic regression and tree-based approaches.