Capstone 1 Data Story

Predicting Prediabetes with Machine Learning

NHANES Data 2007-2016

https://github.com/tcardwell/Capstone-1/blob/master/Capstone%201%20EDA.ipynb

In analyzing the NHANES data, I was curious whether I would see the same statistics regarding diabetes and prediabetes prevalence as reported by the American Diabetes Association (ADA), and whether I would find features that showed a positive correlation to prediabetes and/or diabetes diagnosis criteria.

In my initial exploration, I identified the quantities of various categories of my data. I had noticed a wide variation in participation of various tests, exams, and questionnaires in my data wrangling step. I wanted to make sure I had enough data to get reasonable results.

I then went on to look at the laboratory tests used to diagnose diabetes and prediabetes, starting with basic summary statistics, then continuing with distributions and whether there was any correlation among them.

Interestingly, I noticed that all three diagnostic test distributions are skewed right, with the means greater than the mode and very long right tails. These tails are due to very high abnormal test results in a small population and are valid data.

I was curious about the relationships among the three diagnostic tests, so plotted them in pairs and calculated their correlations. I was not surprised to note moderate to strong positive correlations between each pair of diagnostic test values. Doctors usually use two test results to confirm a diagnosis of diabetes: either from different samples or two different tests from one sample. I would expect when one lab test is abnormally high the others would also be elevated.

I was also curious whether there was a relationship between glycohemoglobin (HbA1c) and age. Type 2 diabetes used to be called adult onset diabetes long ago, and type 1 was called juvenile diabetes. This is because almost all new cases of type 2 diabetes used to be in adults. Most type 1 cases, which are caused by a different mechanism than type 2, were in children. Unfortunately, that has changed, and now many children are diagnosed with prediabetes and type 2 diabetes. Not surprisingly, there is an upward trend of glycohemoglobin with age.

I compared the NHANES data to the statistics published by the ADA. I first determined whether each NHANES participant met diabetes diagnosis criteria, prediabetes diagnosis criteria, or neither. I then plotted the total number of participants in each category and the percentage of my sample. The results were very close on the diabetes percentage and a bit low on the prediabetes percentage.

Next, I looked at those who were taking medication for diabetes or elevated blood sugar, which in the absence of a diagnosis of diabetes is prediabetes. The ADA states that about 24% of Americans with diabetes are undiagnosed and unaware of their condition. The NHANES data shows a ratio of about 3 people on medication for diabetes for every 4 who meet diagnosis criteria. I also looked at those taking medication for prediabetes, but this is a very small sample. Treatment for prediabetes is usually lifestyle

changes instead of medication. My data showed only about 4 people taking medication for elevated blood sugar for every 100 who meet the criteria for prediabetes.

I was curious whether those taking medication for diabetes or prediabetes had any differences in lab test values compared to those taking nothing for those conditions.

This was difficult to plot as the groups taking medication for diabetes or prediabetes had very few Oral Glucose Tolerance Test values (OGTT). Nonetheless, I was able to see that the range of test results for those not taking medication was generally narrower and somewhat lower than those on medication.

Finally, I looked at several blood markers and physical exam markers for those meeting diabetes diagnosis criteria, prediabetes diagnosis criteria, or neither. There were several markers that appear to have a positive correlation with diabetes.

It would be very interesting to see which features in the NHANES data have the strongest correlation with the diagnosis indicator, and which are most important in predicting prediabetes or diabetes. This could guide doctors to testing those who are undiagnosed so they can avoid the life-altering consequences of diabetes.

It would also be very interesting to check the trend shown in the plots by testing a null hypothesis that the means of the blood and exam markers are the same for each diagnosis indicator.