Capstone Project 1 Final Report

1. Project Proposal

1.1 Problem Statement

Diabetes, particularly type 2 diabetes, is a leading contributor to mortality rate in the United States and has known associations with heart disease, kidney disease, and other health problems. According to the American Diabetes Association (ADA), diabetes was the seventh leading cause of death among Americans in 2017. Additionally, the ADA reported $327 billion in medical costs and $90 billion in reduced productivity associated with diabetes for that same year.

Since type 2 diabetes is a preventable disease, identifying its risk factors would allow for earlier detection, prevention, and treatment of at-risk patients by health professionals. This in turn would improve the health of the general United States population and help to reduce the financial burden that treatment places on the economy.

1.2 Data

The data I will use for this project is from the 2014 Health and Behavioral Risk Factor Surveillance System (BRFSS), a yearly survey conducted by phone in the United States. The data is made readily available to the public by the Center for Disease Control. The data as provided includes 279 features and 464,664 observations.

1.3 Approach

All of the data in the BRFSS are labelled, including whether the participant has been diagnosed with diabetes, so a supervised learning approach will be taken. Although the data are labelled, participants' responses are encoded in digit form, likely to reduce the dataset's memory requirements, with a code book provided as a separate file. In order to more effectively communicate my findings, these codes will need to be converted into more informative labels.

Following the guidance of other studies on diabetes risk factors, twenty-five predictive variables were selected from the dataset. Those features are general health; mental

health; age; body mass index (BMI); time since last checkup; income; race; sex; metropolitan status code; employment status; marital status; education; health coverage; average hours of sleep per night; whether a person exercises regularly, received a flu shot in the past 12 months, rents or owns their home, smokes, or uses any medical equipment; and the presence of a depressive disorder, blindness, difficulty concentrating, angina or coronary heart disease, or kidney disease.

The training data for the model will be based primarily on a holdout of the main dataset, though some weighting or oversampling will be needed in order to overcome the unbalanced nature of the data -- that is, since far fewer people have diabetes than not, we need to ensure that this discrepancy does not result in a biased model.

1.4 Deliverables

Deliverables for this project will include the code used to preprocess the data and to construct the model, a full report covering the process and the findings, and a slide deck summarizing those findings.

2. Wrangling the Data

2.1 Data Summary

The source of the data for this study on risk factors for type 2 diabetes is the 2014 BRFSS Survey Data and Documentation**,** which constructed its data from the survey responses of about 138,000 individuals living in the United States and its territories. The data organized into a table where each entry is a respondent and each feature is the response to a survey question, so the data is already constructed in a way that is useful for analysis. However, the data itself is encoded as numeric values; determining a participant's actual responses requires referencing a codebook that is provided alongside the dataset.

2.2 Cleaning the Data

Luckily, using the codebook is fairly intuitive; searching for columns by name allows one to quickly find a table of values that associates each possible value with a label describing the content of its response. Since most of the features had some form of

encoding for missing data, I decided that the first step would be to convert all of these values to NaN and to drop all of the rows that contained at least one NaN. After this, I replaced all of the encoded data with the labels provided in the codebook.

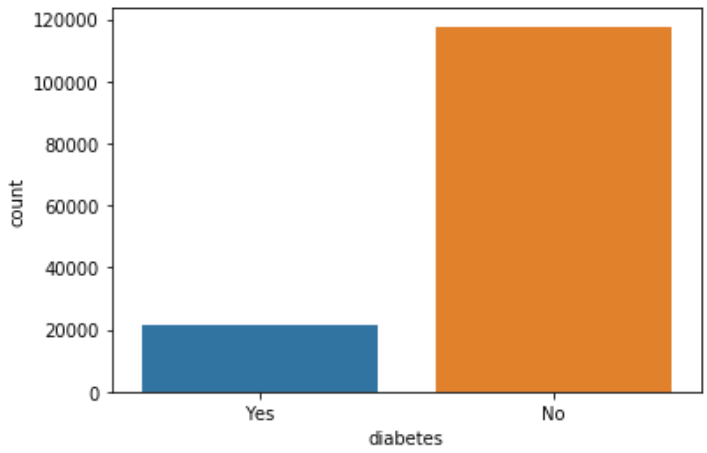
The CDC study I am using as a reference point, Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques, provided reasons to drop additional rows in order to focus the study on adults who had developed chronic type two diabetes, since the data collected in the BRFSS does not distinguish between types one and two. First, all adults below age 30 were excluded since it was more likely that their cases of diabetes were type one. Second, adults who only had diabetes while pregnant or who only had prediabetes were excluded in order to limit the data to those who tested strictly positive or negative for diabetes.

Finally, I condensed the variables for age, mental health, and sleep time into categories that matched the categories used in the CDC study. I also assigned more human-readable names to the columns before saving the resulting dataframe to a new .csv file.

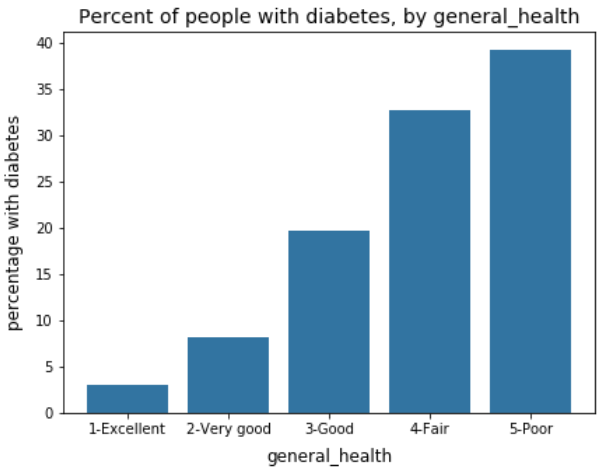
3. Data Analysis

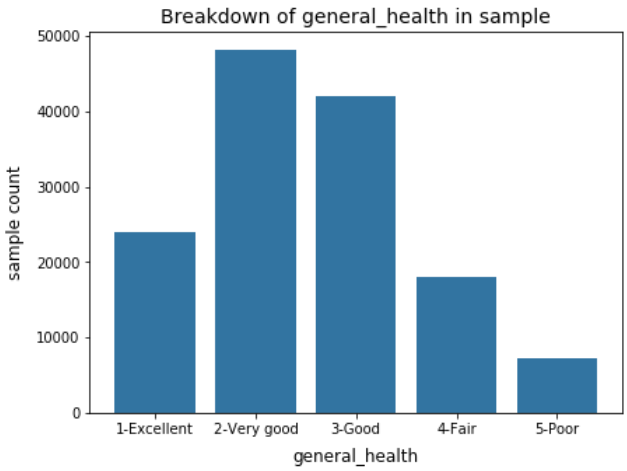
3.1 A Look at the Target Variable

First, I create a countplot of the diabetes feature to get a feel for the data I will ultimately try to predict.Plotted are 117679 “No” observations and 21587 “Yes” observations, so about 15.5% of the observations retained after wrangling the data have received a diagnosis of diabetes. In 2020, the CDC reported a diabetes rate of 10.5% for the general population and a rate of 13.0% was also reported for adults 18 and over. Bearing in mind that my data are restricted to adults at least 30 years old and that diabetes is known to be more prevalent among older individuals, my observations seem to be well in line with what is reported by the CDC.



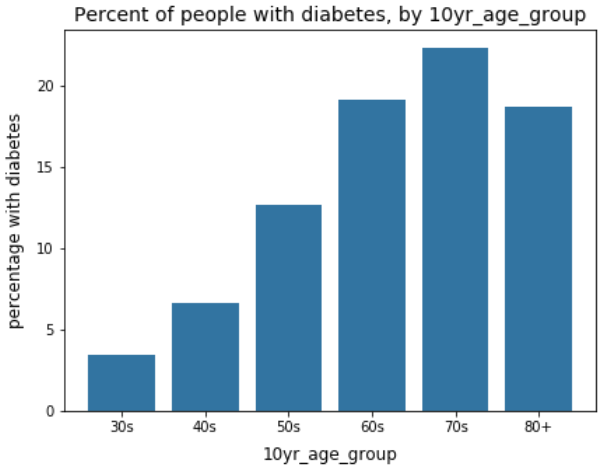
3.2 A Look at Selected Features

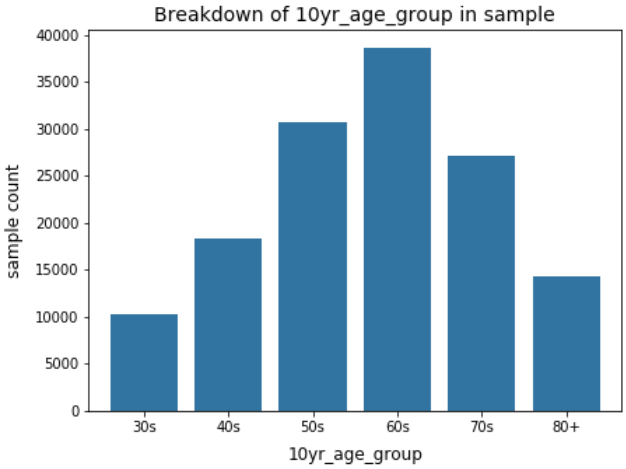




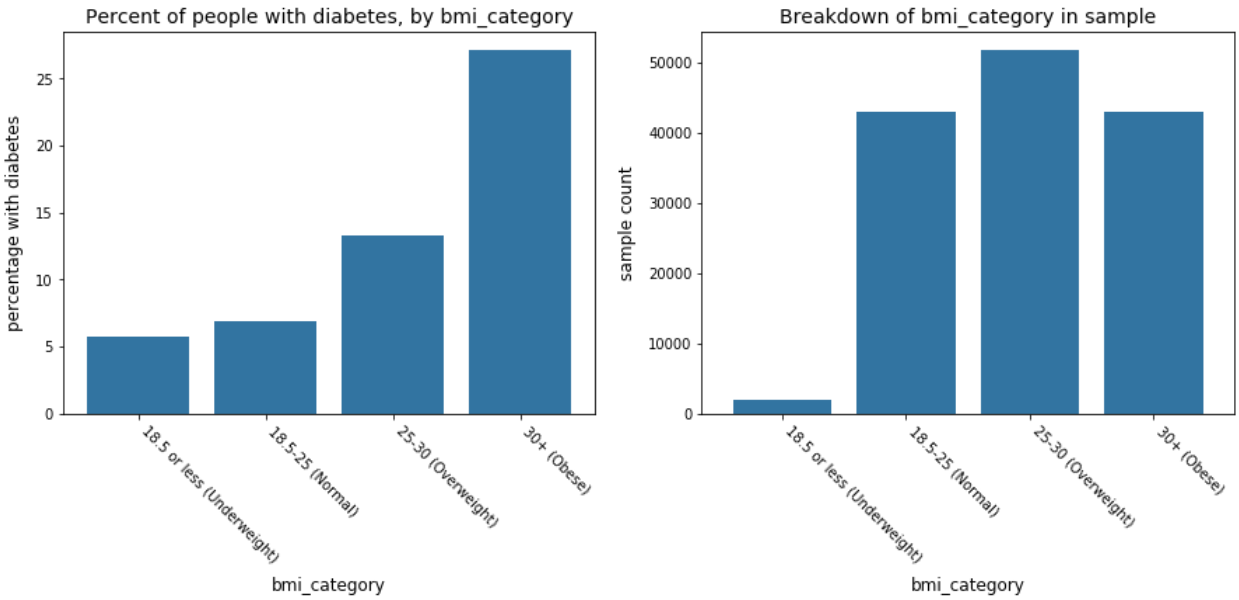
As one would expect, diabetes has a strong association with poor general health. The general health measurement is based on the subjective responses of survey participants, which probably explains the right skew of the breakdown of `general_health` -- if data on other health measurements like BMI are any indication, I might expect a more self-aware sample to have a left skew instead. In other words, it seems that people may tend to overestimate their own healthiness. Despite this, those who self-report excellent health are far less likely to report having diabetes than the average participant, and the reverse is true for those who self-report poor health, so `general_health` looks to be a strong predictor of whether a person has diabetes.

The rate of diabetes increases non-linearly with BMI. Since most Americans are either overweight or obese, high BMI is likely one of the most prominent risk factors

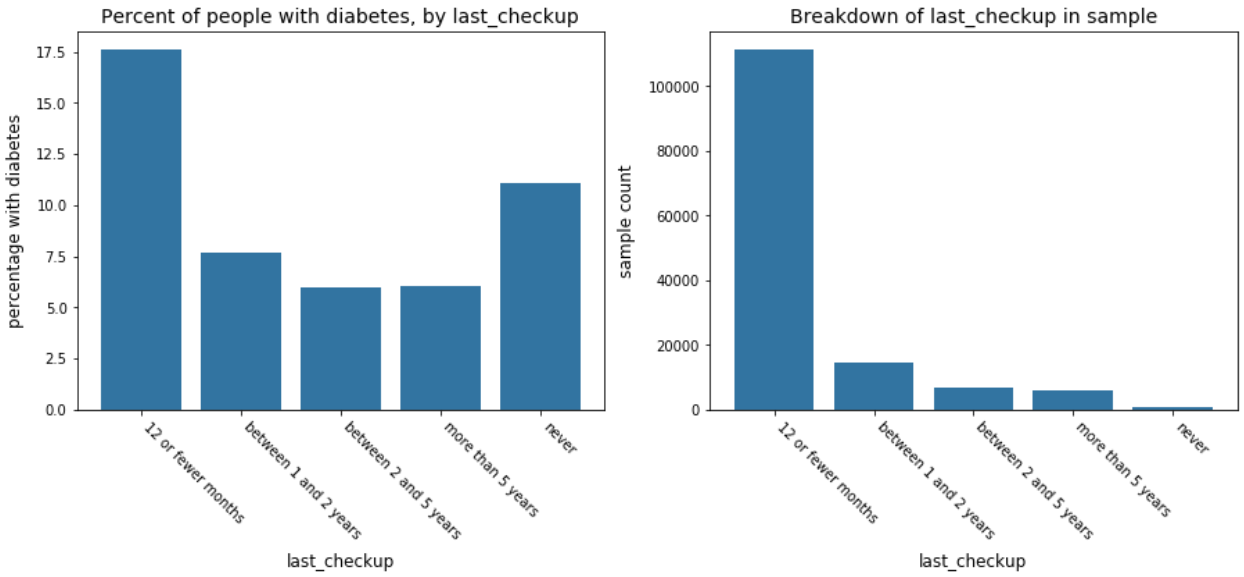




associated with diabetes and poor health outcomes in general, especially as BMI approaches the obese (30+) category.



A person's risk of diabetes tends to increase with age, which is expected since health tends to decline as people get older. Interestingly, a lower risk of diabetes is found in those over 80 years old than those in their 70s, contrary to the pattern established in the rest of the graph. The World Bank reported an average life expectancy of 78.8 years in the US in 2014; therefore, one possible reason for this change in the pattern is that those who live longer than expected have tended to maintain better health than those who do not.



I found the associations in the last_checkup feature to be rather surprising because I associate regular checkups with good health, yet the group that has had a checkup in the last 12 months has the highest rate of diabetes. Those who have never had a checkup were the next most likely to have been diagnosed with diabetes, but the rate of incidence is still below that of the whole sample. Perhaps those who consider themselves to be in good health are more likely to skip their regular checkups. It may also be worth noting that one has to actually visit a doctor in order to be diagnosed with diabetes.

3.3 Conclusions

A number of factors including general heath measurements, lifestyle indicators, and demographic information have value as predictive features. Those features included in this report -- general_health, mbi_category, 10yr_age_group, and last_checkup -- are those that appear to have the strongest associations with diabetes based on a broad analysis of the data. However, even these risk factors are associated with a diabetes rate of around 25% to 30%, so in order to construct an accurate predictive model, I expect we will need to be able to identify individuals with multiple risk factors.

4. Statistical Inference

4.1 Chi-Squared Test

As seen in the previous section, all of the data presented is provided in categorical form. Many of the features either do not have a consistent interval between categories or do not have intervals at all, so when conducting statistical inference and building the machine learning model, the most intuitive way to approach the data will be to use methods well-suited to categorical data. For statistical inference, that means starting with the chi-squared (χ^2) test of significance.

My null hypothesis is that the two selected variables are independent of each other, and my chosen significance level is α=0.01. I opted for a lower value of α because I believed the relatively large sample size I'm using would otherwise make it easier to detect false positives. I chose to test the relationships of all predictive variables to diabetes as well as all predictive variables to each other, since I plan to use their χ^2 values in the next step. The test is performed using the SciPy library in Python.

Four pairs of predictive variables fail to reject the null hypothesis, those being (flushot, any_exercise), (flushot, blindness), (flushot, trouble_concentrating), and (sex, kidney_disease), so it seems likely that there is some degree of association between most pairs of predictive variables. All predictive variables as related to diabetes successfully reject the null hypothesis, so I am confident that all selected variables will have some predictive value.

4.2 Cramér's V

An additional test I can run after computing the χ^2 statistics and their p-values is Cramér's V, which uses the value of χ^2 to compute the strength of association between two variables. This measure of association ranges from 0 to 1 and serves a similar purpose to measuring the correlation between two continuous variables, so it seems like a potentially useful tool to use on data that is entirely categorical (whether ordinal or nominal). Again, I am most interested in the strength of association between diabetes and all other variables, but I have also computed Cramér's V for all other pairs since it may be useful to know how strongly associated two features are to each other when it comes time to build the machine learning model. The bias correction operation is included in the computation of Cramér's V.

Although most pairs of features were found to likely have some association by their low p-values on the χ^2 test, very few pairs of features have a strength of association higher than 0.25. Perhaps the large sample size made it possible to be reasonably confident that even weak associations were statistically significant (recall that this is related to the reason I selected a very low p-value). Three of the four pairs of features that failed to reject the null hypothesis on the chi-squared test were also found to have a Cramér's of 0.00. The fourth, flushot and blindness, has a Cramér's V of 0.01.

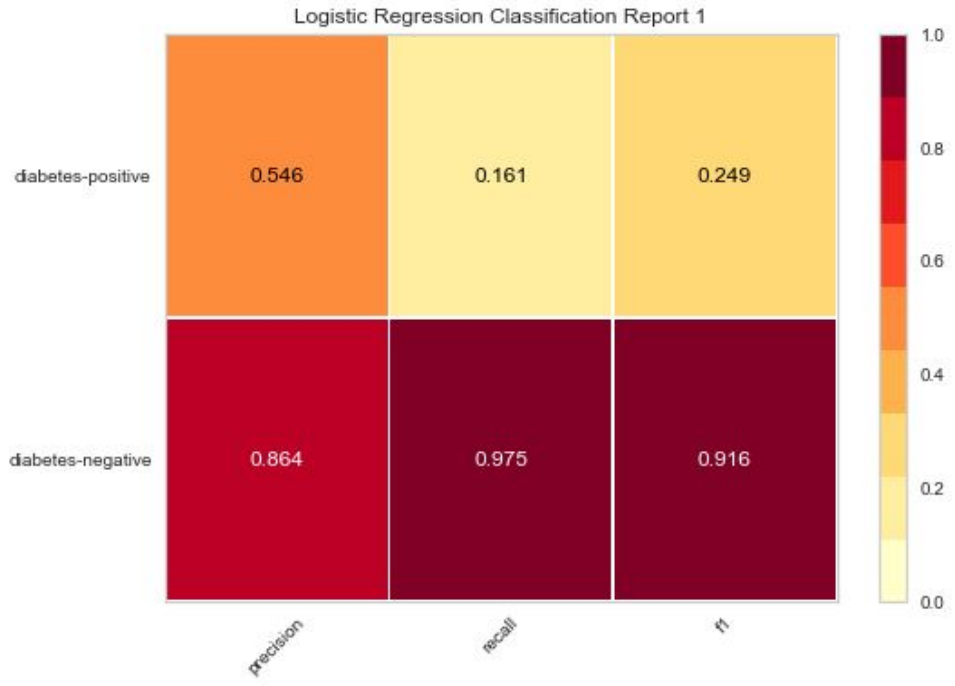
On diabetes, the strongest association, 0.30, is to general_health, and the lowest, 0.02, is to mscode. The next-highest associations to diabetes are to bmi_category at 0.23, employed at 0.20, income at 0.18, and age_category at 0.17.

Some pairs of features, such as mental_health_days_per_month and depressive_disorder, have a moderate degree of association to each other. However, no pair of features has an association higher than 0.44, so at this point it does not seem especially likely that any of the dependent variables will be too strongly associated with each other, so all of the features recommended by the scientific literature on diabetes will likely be retained by the model.

5. Machine Learning - Logistic Regression

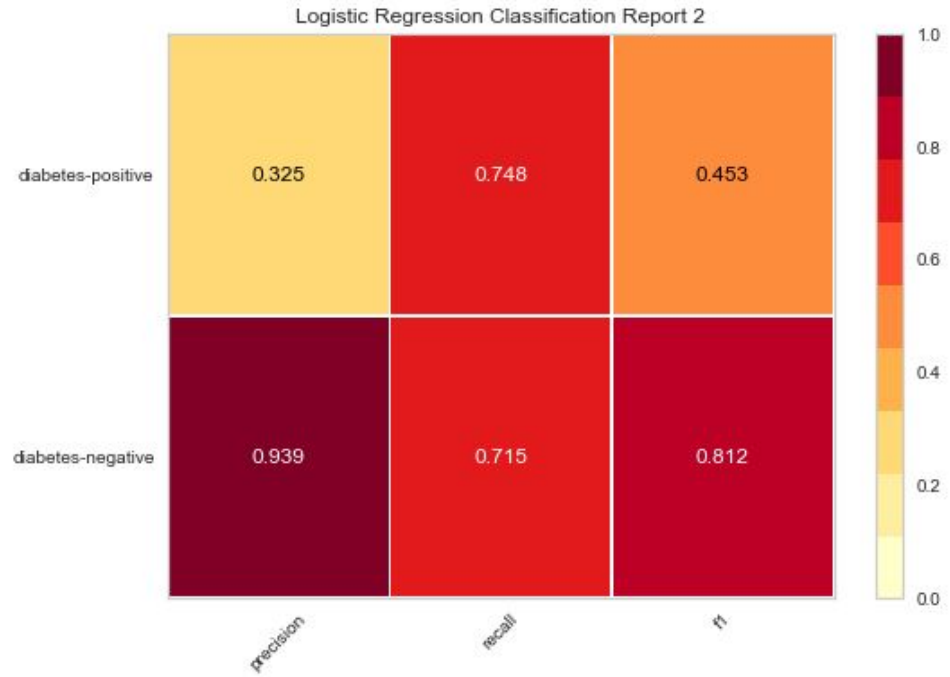
5.1 Building the Model

I set out to build a logistic regression model using Python's scikit-learn library (classification reports are plotted using the yellowbrick library). First, I split the data into training and testing sets, using 25% of the data as the training set and the remaining data as the testing set. Next, I created a simple logistic regression model to use as a control when testing changes to the model's hyperparameters. The classification report for this simple model is shown below.

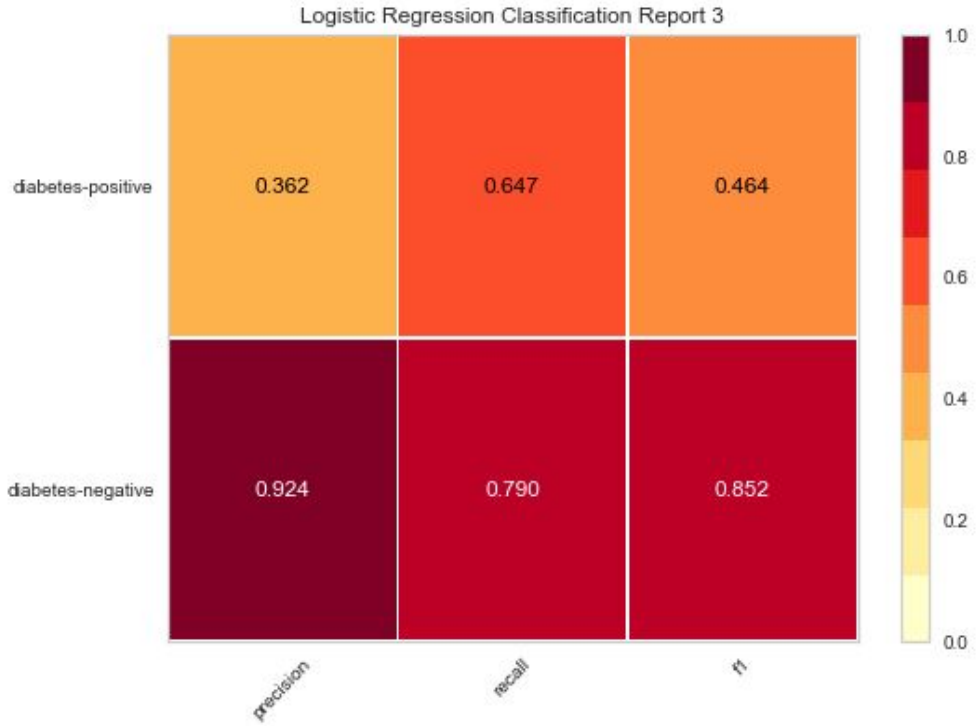


As one might expect for data whose classes are unbalanced, a simple model does a poor job of predicting cases that fit the minority class, which in this case is the diabetes-positive class. The model's AUC score is 0.80, which may seem fine in isolation. However, we find a recall of only 16% for those with diabetes; in other words, the remaining 84% of those with diabetes were reported as false negatives by the model, which has a strong negative effect on the f_1 score (0.25).

Since I am primarily interested in identifying those with diabetes, I re-tune the model by assigning class weights inversely proportional to their frequency in the data.



Already, we see a marked improvement in the rate of recall and the f_1 score for diabetes-positive, and the AUC score remains about the same at roughly 0.80. Although the f_1 score for diabetes-negative drops slightly, I more strongly consider regard the increased performance on diabetes-positive and consider this a net improvement.

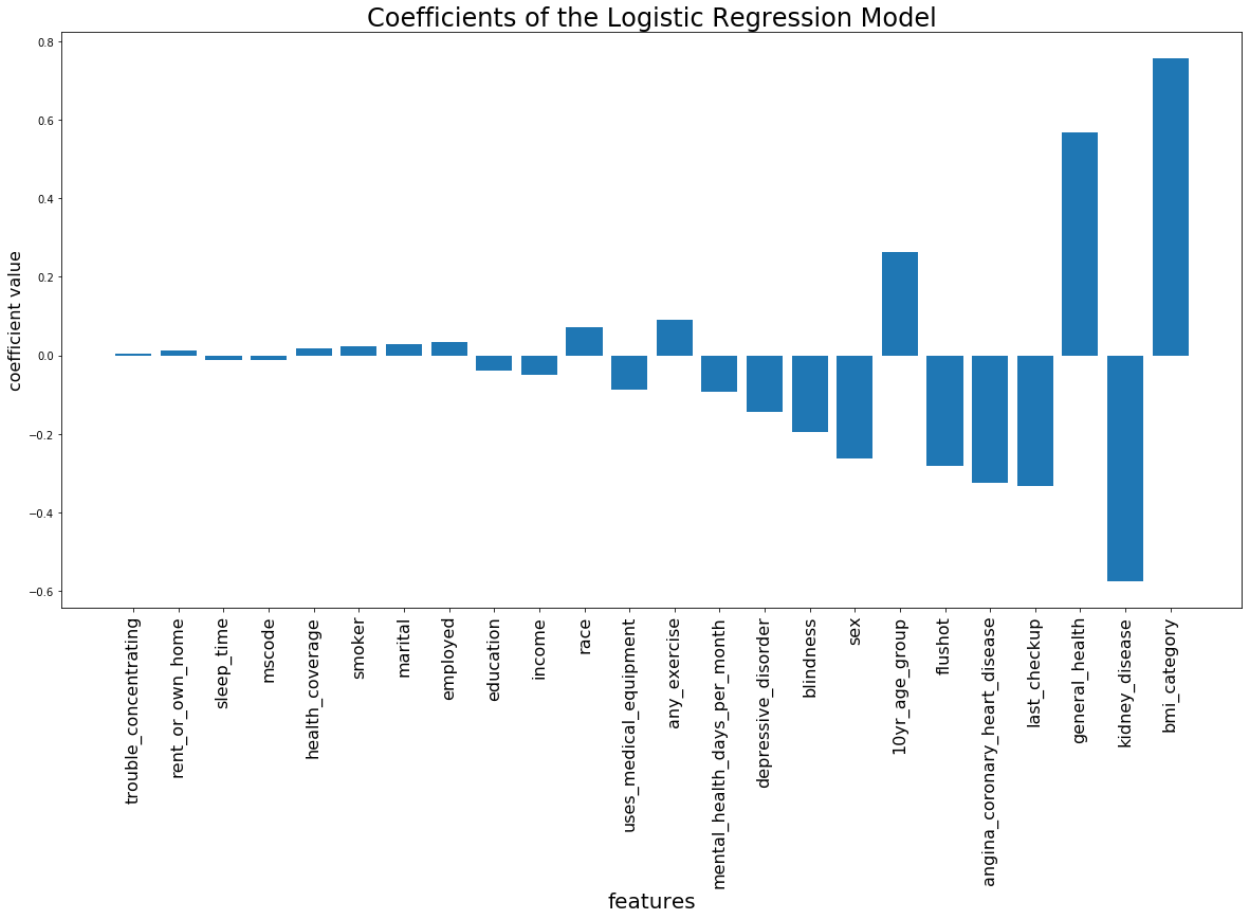


I next attempt to tune the model further by adjusting the regularization parameter, C. I also provide a range of possible weights, centered around the standard weights, to allow the model more flexibility in choosing parameters. (The resulting classification report is shown on the previous page.)

This updated model uses a value of 100 for C instead of the default 1 and adjusts the class weights from 16:84 to 20:80. The AUC score is once again about 0.80, and both f_1 scores are slightly higher in this version of the model compared to the second version. I believe this version of the model is sufficient to move forward with a final analysis of the predictor variables.

5.2 Analysis and Conclusions

Below is a plot of the logistic regression model's coefficients. The features of the model are sorted by the absolute values of their coefficients.



The features are ordered from left to right as follows: trouble_concentrating, rent_or_own_home, sleep_time, mscode, health_coverage, smoker, marital, employed, education, income, race, uses_medical_equipment, any_exercise, mental_health_days_per_month, depressive_disorder, blindness, sex, 10yr_age_group, flushot, angina_coronary_heart_disease, last_checkup, general_health, kidney_disease, and bmi_category.

The exploratory data analysis I performed earlier offers some insight into which of these features' values would contribute to diabetes. High BMI, advanced age, poor general health, being male, and the presence of kidney disease, heart disease, or blindness all have an association with diabetes. Yearly checkups and never having had a checkup both have an association; the former is more plausibly a result of people in poor health more closely monitoring their condition, while the latter could indicate that a lack of preventative care allowed a case of diabetes to develop.

Based on these results, I would consider the following recommendations to patients to be the most pressing:

1. Those in poor health, particularly those who are overweight, should be made aware of their risk of diabetes and encouraged to improve their health before a serious condition develops.
2. Elderly patients, particularly male patients, should be made aware of their risk of diabetes since these factors are outside of a person's control.
3. Kidney disease, heart disease, and blindness should be indicated as possible complications that may result from diabetes.

This is not to say that other factors are unimportant; a sedentary lifestyle, race, factors affecting access to care, etc. should all be noted by healthcare providers as well. The three recommendations listed above are meant to highlight the strongest associations to diabetes found by the logistic regression model.

Due to the cross-sectional nature of the data, it is not possible for the model itself to draw conclusions about causality; inferences I make about the direction of causality are drawn from the scientific literature on type two diabetes. As such, a possible avenue for future study would be to reaffirm the relationship between identified risk factors and diabetes.