Jack Yan
BrainStation, Data Science

Diabetes Predictions with Machine Learning

Motivation

The American Diabetes Association estimates about 37 million Americans has diabetes in 2019. Of the 37.3 million adults with diabetes, 28.7 million were diagnosed, and 8.5 million were undiagnosed. The cost of undiagnosed diabetes are associated with the following:

- Reduced productivity while at work (26.9 billion) for the employed population
- Increased absenteeism (3.3 billion)
- Reduced productivity for those not in the labor force (2.3 billion)
- Lost productive capacity due to early mortality (19.9 billion)

Our goal of the project is to accurately predict diabetes status in patients given their basic medical information, and to find factors that is most indicative of diabetes. Diabetes is typically diagnosed using blood tests or an eye exam. Our project can act as a basic information test run on clinic databases to identify at risk patients. There is already a risk test on the ADA website. However, it is often the case that individuals not diagnosed are simply not aware that they have diabetes so they may not see a reason to look up diabetes.

For more information, please visit the American Diabetes Association website:
https://www.diabetes.org/about-us/statistics

Data

The data is the 2015 Behavioral Risk Factor Surveillance System survey from the CDC. It is collected by phone yearly across the US. The data collected is quite extensive. We believe that overall, the data did a great job at showing the variables to predict diabetes. However, there were a number of entries with missing values which we had to deal with in the data processing stage.

Preprocessing

There were 330 columns in the dataset. We had to narrow down all the columns that were relevant to the problem space. The codebook provided by the CDC detailed explanations on all the columns. The procedure for column selection was to:

- look through the codebook for appropriate columns
- select between certain duplicated columns
- transform the columns into binary or ordinal
- write a description for the column and what transformations took place
- rename into a name that is easier to understand

This process took some time to complete as we had to research relevant factors to diabetes and find them in the codebook.

EDA

During our exploratory data analysis, we plotted the histogram and found the correlations of each variable. The histograms revealed the distributions of every feature in the data. We found that our target feature, the diabetes column was quite imbalanced. We used a heatmap to showcase the correlation between variables. It hinted about which variables our model would predict diabetes the most. Lastly, we exported the dataset to Tableau and created a few visualizations to showcase relationships between diabetes and certain variables.

More information can be found in the EDA section in the first notebook.

Methodology

With our clean data we proceeded to scale, split, and model the data. We only scaled non-binary columns. Binary values are arbitrary, so scaling binary columns do not make sense. After scaling, we split the data into testing and training sets at an 80/20 split. The next transformation that we performed was to upscale the imbalanced target data. We used SMOTE to bring up the number of samples in the training dataset that had diabetes. This is to prevent our models from always classifying as the overrepresented sample, in this case no diabetes.

We performed a gridsearch on 3 models, logistic regression, decision tree, and XGBoost. Each model had different hyperparameters for searching that included common hyperparameters to adjust. We set the best model according to the grid search results and constructed a confusion matrix. Then we used the matrix to analyze our precision and recall scores. We then adjusted the probability for the model to classify for diabetes to improve our recall score at the expense of accuracy and f1 scores.

We used SHAP to determine which factors affected the decision making of our model the most. Using a force plot and summary plot we can illustrate the magnitude and direction of the shift different variables have on the prediction of the model.

Findings

Of the 3 business questions we aimed to answer we confidently can answer the first 2 with this project. The last question, “Can we use the most predictive features to create a questionnaire for determining diabetes?” can be used as the guiding question for the next part of the project.

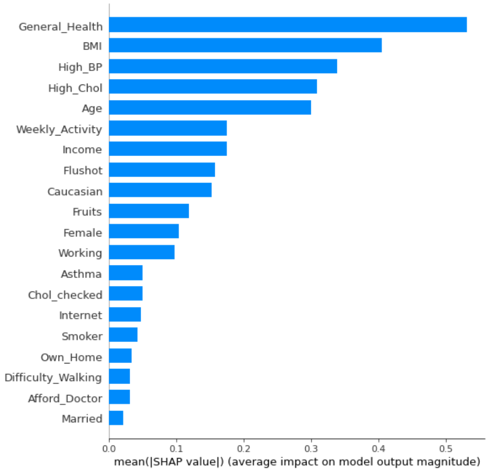
Have we trained a model to accurately predict diabetes?

Our best model predicts 79% accuracy with a 0.8 f1-score and a 0.65 recall score for the diabetes class. If we select for increased recall by lowering the threshold the model predicts diabetes, we get a 70% accuracy score with a 0.74 f1-score and a 0.81 recall score for the diabetes class at a threshold of 0.35.

We would say this is a good prediction of diabetes but not ideal. The sample size is good. Perhaps during the data processing step, we could have included more factors into the dataframe used to train the model or coded certain characteristics such as marriage status differently. There were also additional machine learning models not explored in this project. Training a neural network can be the next step for this project in terms of modeling.

Which factors are most predictive of diabetes?

General health, BMI, blood pressure status, cholestrol status and age affect diabetes status the most. Weekly activity, income, flu shot status, race, fruit intake, sex and working status affect diabetes status somewhat.



Future Work

The future roadmap for this project would be to improve the overall accuracy of the model, then to install this model in a clinical setting. Ideally, we would want either to create a questionnaire for determining diabetes using the most predictive features or embed the elements of a diabetes patient in a system and patients meeting the characteristics are automatically flagged for a test recommendation.