Final Report

Prediabetes Health Indicators

Daniel Ellis Schwartz

Problem Statement

Prediabetes is a disease that can be understood as early stage diabetes, characterized by some symptoms of type 2 diabetes being present before a diagnosis is made. Progression to type 2 diabetes is not inevitable however and prediabetes can be reversed with lifestyle changes. The goal of this project was to develop a model that can identify individuals who have or are at risk of developing prediabetes.

Background

In 2019 an estimated 463 million people have diabetes, with type 2 diabetes making up approximately 90% of cases.[1] While the cause of type 1 diabetes is unknown type 2 diabetes is largely preventable by making lifestyle changes. Obesity and lack of exercise are the primary lifestyle factors but there are several others.

Prediabetes itself is widely prevalent in the US adult population. Approximately 1 in 3 US adults have prediabetes and at least 8 of 10 don't know that they have it.[2] It is generally advisable for everyone to maintain a healthy weight and to exercise regularly but providing individuals with an easy and straightforward way to evaluate their likelihood of having prediabetes can be very beneficial. A simple questionnaire could be completed online that would serve this function.

Dataset

The data used for this project is sourced from a Kaggle user who selected and organized a portion of survey data collected by the US Centers for Disease Control and Prevention (CDC). This is a subset of telephone public health surveys taken between 2011 and 2015 which includes responses from approximately 400,000 individuals.

The relevant portion of the survey involved 22 questions, including a question of whether the respondent had diabetes, prediabetes, or neither. Approximately two thirds of the questions had yes or no answers while the rest had responses categorized into at most 13 categories. One question had a numeric response. Some of these questions were:

- What is your age? (Binned into 13 categories)
- Do you have high blood pressure? (Yes or no)
- What is your BMI?
- Have you smoked at least 100 cigarettes in your entire life? (Yes or no)
- Would you say that in general your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

Some of these questions required information that a respondent would have received from a doctor but many others did not.

---

[1] https://www.diabetesatlas.org/upload/resources/material/20200302_133351_IDFATLAS9e-final-web.pdf
[2] https://www.cdc.gov/diabetes/basics/prediabetes.html

It is important to note that this self reported health survey data has some significant limitations. People tend to bias their responses towards how they are feeling at the time of collection and are generally disposed to present a better picture of their own health.[3] [4] Survey's are also limited by the respondent's knowledge of their own medical status which may be incomplete.

Additionally the lack of hard health metrics (resting heart rate, blood pressure, blood glucose levels, etc.) limits precision of the data. The questions with yes or no responses could hide a wide range of relevant health information. For example, a question on physical activity asked if the respondent had any physical activity in the past 30 days, excluding their job. Someone who got 10 minutes of exercise once a month and someone who exercises 5 days a week would both have answered yes to this question.

Self reported health survey data is invaluable in understanding public health but its limitations must be considered in any analysis using that data.

Data Wrangling

The data used in this project was already processed so very little data wrangling was required. There were 253,680 rows and 22 columns. The columns and their descriptions are shown in the table below.

Column Name	Description
Diabetes_012	0 = no diabetes, 1 = prediabetes, 2 = diabetes
HighBP	0 = no high BP, 1 = high BP
HighChol	0 = no high cholesterol, 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
BMI	Body Mass Index
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no, 1 = yes
Stroke	(Ever told) you had a stroke? 0 = no, 1 = yes
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI) 0 = no, 1 = yes
PhysActivity	Physical activity in the past 30 days - not including job. 0 = no, 1 = yes
Fruits	Consume Fruit 1 or more times per day. 0 = no, 1 = yes
Veggies	Consume Vegetables 1 or more times per day. 0 = no, 1 = yes
HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no, 1 = yes

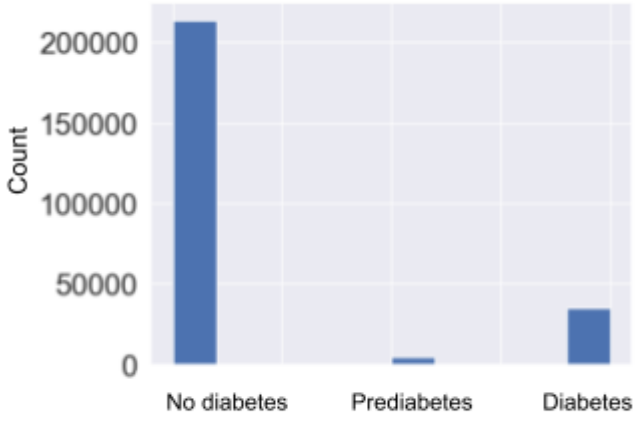
[3] https://www.sciencedirect.com/science/article/pii/S2210600612000226
[4] https://www.sciencedaily.com/releases/2019/10/191008140406.htm

AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no, 1 = yes
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? Scale 1-30 days
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? Scale 1-30 days
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes
Sex	0 = female, 1 = male
Age	13 level age category: 1 = 18-24, 2 = 25-29, 3 = 30-34, etc, 9 = 60-64, 13 = 80 or older
Education	Education level, scale 1-6. 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Income scale 1-8. 1 = less than $10k, 2 = less than $15k, 3 = less than $20k, 4 = less than $25k, 5 = less than $35k, 6 = less than $50k, 7 = less than $75k, 8 = $75k or more

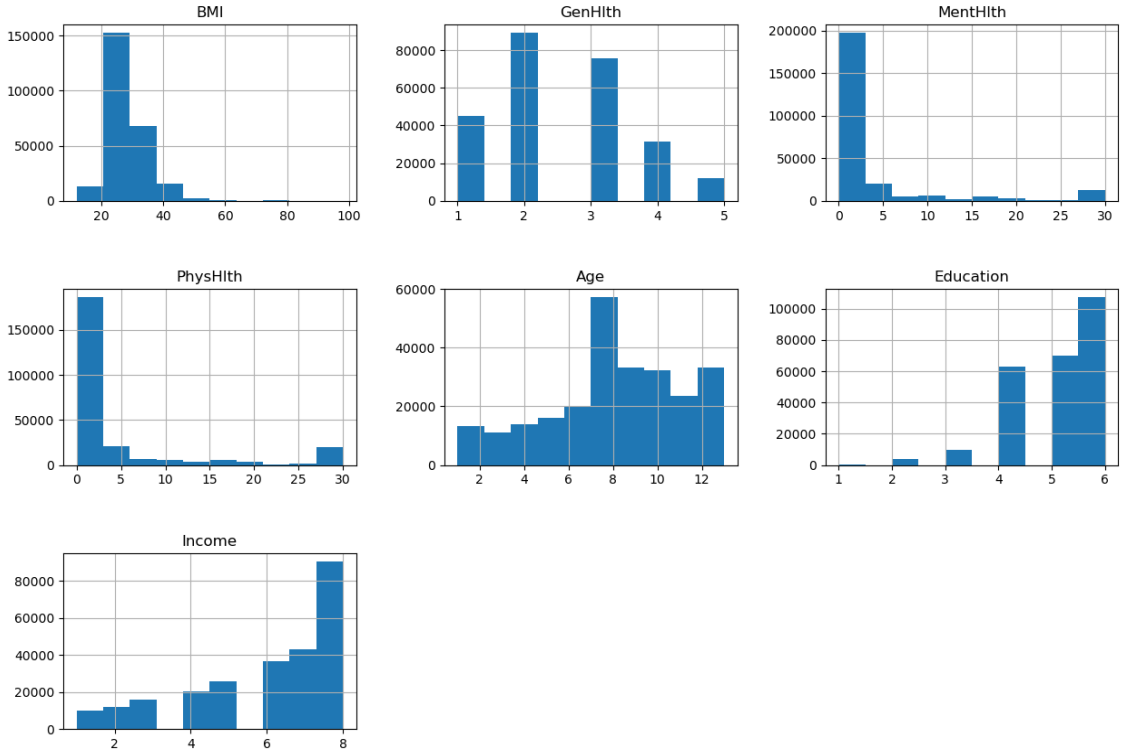
EDA

The target variable, 'Diabetes_012', has three possible values. 0 for no diabetes, 1 for prediabetes, and 2 for diabetes. There is significant imbalance between the counts of each class.

Value	Count	Percentage
No diabetes (0)	213703	84%
Prediabetes (1)	4631	2%
Diabetes (2)	35346	14%

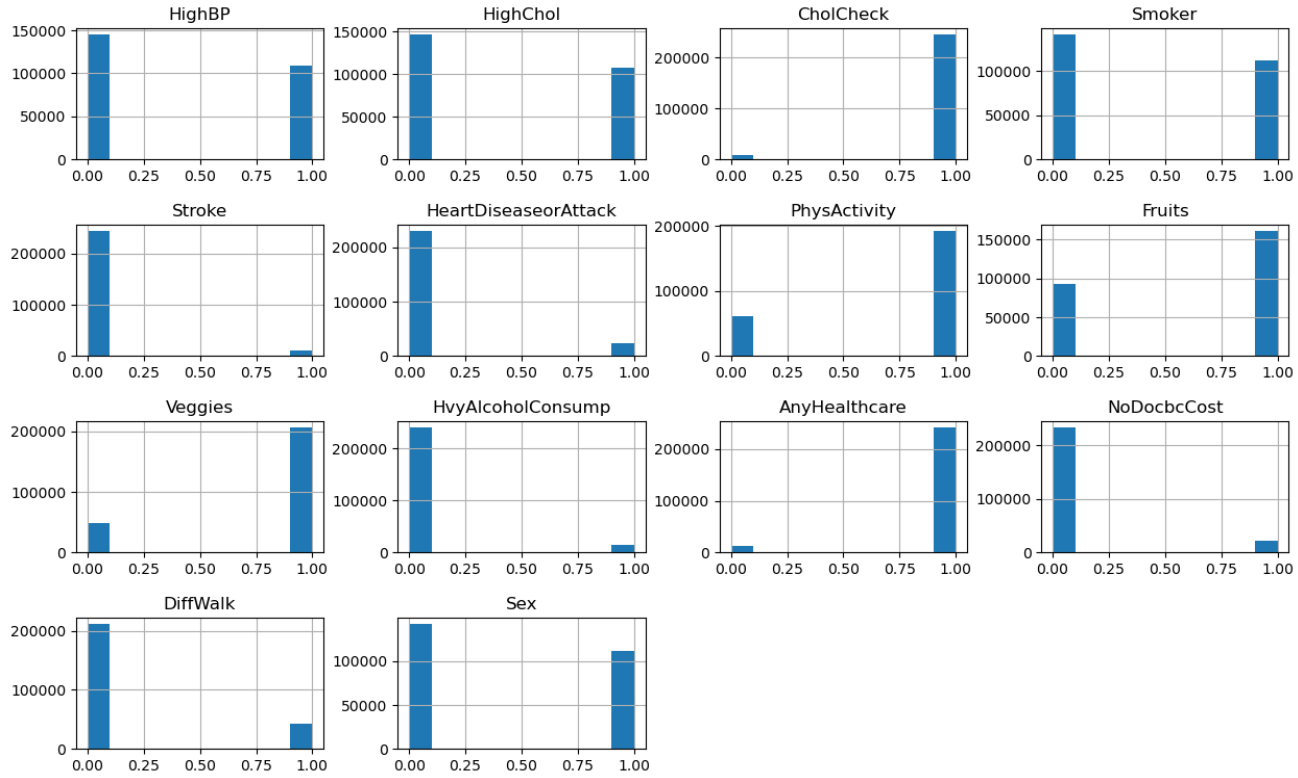


The features can broadly be divided into binary and non-binary features. The non-binary feature distributions are shown below. Note that a lower value indicates better health for GenHlth, MentHlth, and PhysHlth.



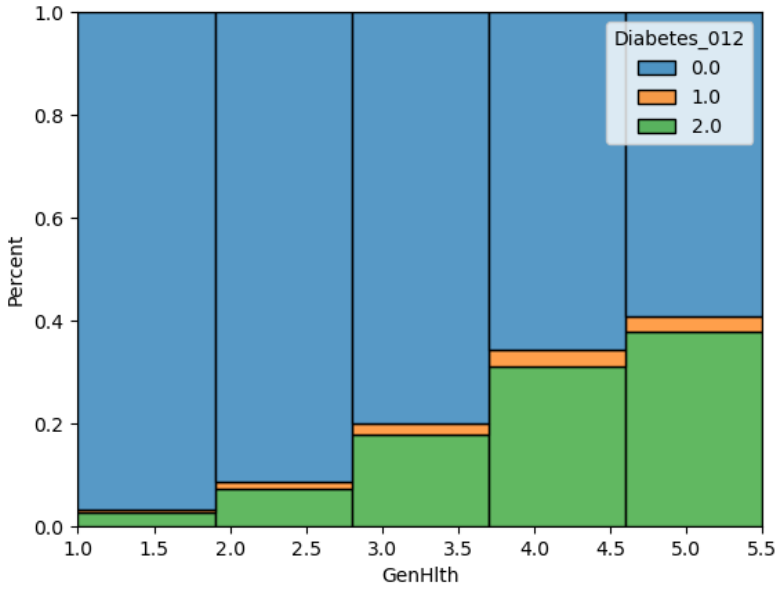
Generally speaking the majority of respondents rated themselves as healthy. There are however peaks at the high end for MentHlth and PhysHlth, indicating that a minority of people have low health most days.

Distributions for the binary variables are shown below.

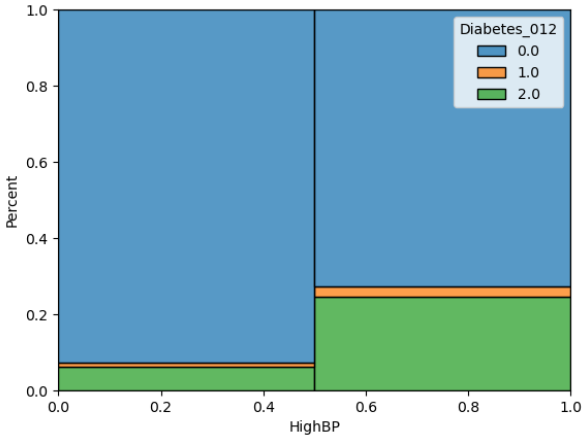


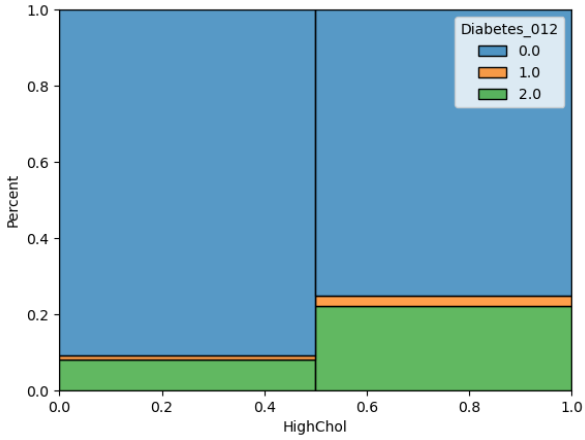
Here we can see some potential inconsistencies with the self reported health shown in the non-binary variables. While the vast majority of respondents rated themselves as being healthy 43% had high blood pressure, 42% had high cholesterol, and 44% have smoked at least 100 cigarettes.

In looking at the distributions for diabetes within each variable some expected trends can be seen. For example, plotting the percentage of people with or without diabetes vs their self reported general health we can see that people with worse health are more likely to have diabetes.



Similarly people with high blood pressure or high cholesterol are more likely to have diabetes or prediabetes.





All other variables show expected trends relating to health and likelihood of having diabetes or prediabetes.

Examining the correlation coefficients with the target variable also yields expected results. The 5 features with the highest correlation coefficients to the variable 'Diabetes_012' are shown in the table below.

Variable name	GenHlth	HighBP	BMI	DiffWalk	HighChol
Correlation coefficient	0.30	0.27	0.22	0.22	0.21

While not every variable was shown here there were no unexpected trends in relating each feature to the target variable.

Modeling

To handle the class imbalance three different resampling methods were evaluated. These were then fed into four different models. Some combinations of model and resampling method were rejected based on excessive training time or insufficient improvement in performance. The resampling methods used were:

- Undersampling of the non-minority class.
- Synthetic minority oversampling (SMOTE) of the non-majority class.
- A combination of undersampling and SMOTE.

The model types used include:

- Random Forest Classifier (RFC)
- Random Forest Classifier using One Vs Rest Classification (OVRC)
- Histogram Gradient Boosting Classifier (HGBC)
- Histogram Gradient Boosting Classifier using One Vs Rest Classification

The recall score was the primary metric used to evaluate these models since false negatives are more expensive than false positives in this case. Accuracy and precision were also used to avoid classifying every respondent as having diabetes or prediabetes regardless of their actual condition.

Random search cross validation and grid search cross validation were used for parameter tuning.

Unfortunately there was no clear best model. None of them performed sufficiently well in predicting prediabetes from the survey questions. The table below summarizes some of the most significant models. Model metrics shown are calculated from the test set. Precision and recall scores have three values, one for each class of target variable.

Model Type	Resampling Strategy	Best Parameters	Recall	Precision	Accuracy
1: Random Forest	Undersample	class_weight: balanced max_depth: 50 max_features: 2 min_samples_split: 30	No Diab: 0.62 **Prediab: 0.3** Diab: 0.59	No Diab: 0.95 Prediab: 0.03 Diab: 0.35	0.61
2: Random Forest	Undersample & oversample	class_weight: balanced max_depth: 50 max_features: 4 min_samples_split: 100	No Diab: 0.71 **Prediab: 0.01** Diab: 0.79	No Diab: 0.95 Prediab: 0.02 Diab: 0.30	0.71
3: Random Forest, One Vs Rest	Undersample	class_weight: balanced max_depth: 50 max_features: 6 min_samples_split: 47	No Diab: 0.62 **Prediab: 0.32** Diab: 0.61	No Diab: 0.96 Prediab: 0.03 Diab: 0.34	0.61
4: Histogram Gradient Boosting	Undersample	class_weight: balanced max_leaf_nodes: 3 max_depth: 100 learning_rate: 0.01 L2_regularization: 0.001	No Diab: 0.62 **Prediab: 0.33** Diab: 0.59	No Diab: 0.96 Prediab: 0.03 Diab: 0.35	0.61
5: Histogram Gradient Boosting, One vs Rest	Undersample	class_weight: {0: 1, 1: 2, 2: 2} learning_rate: 0.01 max_depth: 6 max_iter: 400 max_leaf_nodes: 100 min_samples_leaf=250	No Diab: 0.63 **Prediab: 0.31** Diab: 0.60	No Diab: 0.96 Prediab: 0.03 Diab: 0.34	0.61

6: Histogram Gradient Boosting, One vs Rest	Undersample	class_weight: balanced learning_rate: 0.01 max_depth: 6 max_iter: 400 max_leaf_nodes: 100 min_samples_leaf: 250	No Diab: 0.62 **Prediab: 0.30** Diab: 0.60	No Diab: 0.95 Prediab: 0.03 Diab: 0.34	0.61

The recall score for prediabetes was the primary focus for evaluating model performance. The highest recall score was found in model 4 yet this was only 33%. The precision was even worse at 3%. There were only minor differences in recall performance between these models. The exception being model 2 with an abysmal 1% recall, which is why oversampling was not used for any other models.

A higher weighting of the target class (prediabetes) and the diabetes class in model 5 did not lead to a significant improvement in model performance. Higher or similar class weighting in other models not shown here lead to an unacceptable decrease in overall model performance.

While not a particularly useful metric, the accuracy scores for all of these models are remarkably consistent but not identical. (Rounding to two digits hides minor variation in the scores.) The highest accuracy was achieved in model 2, using undersampled & oversampled data in a random forest, but this came at an unacceptable cost to the recall score. In a case like this with highly imbalanced multiclass data a higher accuracy might be expected so these low values are noteworthy.

No model was found that had sufficient predictive value to be useful.

Discussion

The goal of predicting prediabetes using this data set was not achieved. The question then is what went wrong?

It is unlikely that a different choice of model would lead to a better result. Several other models were attempted, including a K-nearest-neighbor and a support vector machine, that were not shown here due to their worse performance. The best performing models of each type in the table above all had similar performance metrics, indicating that only minor improvements could be made using the above methods.

The quality of the data and the choice of target class are likely the primary limiting factors. As discussed earlier, the data comes from a telephone health survey which carries with it significant chances of biases and errors due to the method and type of collection. A lack of precision in the responses to survey questions and the tendency to self-report healthier lifestyles can significantly decrease the accuracy of predictions made using those survey responses.

The choice of prediabetes as the target class was likely the most significant factor. Firstly, prediabetes is by definition a condition that develops before diabetes and as such it is much more likely for a person to have undiagnosed prediabetes than undiagnosed diabetes. The CDC estimates that one in three US Americans have prediabetes and 80% don't know they have it while only 2% of survey respondents indicated that they have prediabetes.[5] This indicates that the data is highly polluted and the performance of any model based on that data would be flawed.

[5] https://www.cdc.gov/diabetes/basics/prediabetes.html

Secondly, there was an order of magnitude decrease in the number of respondents with prediabetes compared to the number with diabetes, and a further order of magnitude difference between the number with diabetes and the number who had neither condition. This means that the most polluted class was the minority class, amplifying the impact of the data pollution.

This conclusion is supported by the better performance of every model on the non-target classes. Without exception the performance metrics for every model were higher when predicting diabetes or no diabetes compared to predicting prediabetes. None of these models were optimized for either of those classes since predicting the other two classes was not the goal. However better performance in predicting diabetes could be expected based on this discussion and the model metrics shown above.

Possible Next Steps

The best way to achieve the desired goal of predicting prediabetes would be to use better quality data. Ideally this would be data collected by medical professionals but I cannot speak to the practicality of acquiring such an ideal dataset.

A reevaluation of the primary goal could lead to better results using only the data in this project. If the goal is to enable individuals to assess their own risk then predicting prediabetes is not necessarily the best nor the only way to achieve this. Ignoring the minority polluted class and creating models to only predict diabetes or no diabetes would likely lead to better results. Then individuals who have prediabetes yet no diagnosis would be included when using the model to evaluate their own risk. Reframing the goal in this way should be the first next step to attempt due to its low cost and the immediate availability of the necessary data.