Determinants of Diabetes:
Identifying the Primary Risk Factors of a Diabetes Diagnosis

Aidan Gildea, Arjun Prabhakar, Hannah Long

STA 325: Machine Learning and Data Mining (Fall 2022)
Professor Simon Mak
18 December 2022

I. Introduction

Diabetes is a major public health issue in the United States. Over the past two decades, incidences of the chronic health condition have continually increased, in part due to negative trends in health behavior and resulting obesity. The Centers for Disease Control and Prevention (CDC) reports that 37.3 million people in the United States have diabetes, which is 11.3% of the population (CDC, 2022). This marks a one percent increase in the prevalence of diabetes since 2001 (CDC, 2022). Moreover, more than 8 million adults in the US have undiagnosed cases of diabetes, a statistic that is particularly concerning (CDC, 2022). Chronic diabetes affects the body's ability to produce insulin, and in turn, retain healthy levels of blood sugar. Without proper treatment, diabetes can cause serious health complications such as stroke, heart disease, and kidney failure (CDC, 2022). These conditions, among others, make diabetes the 8th leading cause of death in the US (CDC, 2022). Further analysis on the risks, behaviors, and demographics associated with diabetes is crucial in order to predict its incidence and combat its rise both in the US and around the world.

In addition to the immediate health concerns diabetes poses, the disease also exacts significant costs for diagnosis, treatment, and research. In 2021 alone, the National Institute of Health (NIH) spent $1.1 billion on diabetes research, with many other public and private institutions funneling funds into the space as well (Juvenile Diabetes Cure Alliance, 2022). This is an incredible amount of money, therefore, harnessing insights on the primary factors associated with diabetes helps ensure that funding is utilrized strategically and effectively. On the patient side, diabetes screenings can cost between $50-70, with an additional $100-200 for visit fees (Slobin, 2022). And once diagnosed with diabetes, individuals can expect to incur medical costs of $16,752 per year, on average (ADA, n.d.-a). With approximately 1.4 million new cases per year, it is pressing that we identify risk factors for diabetes to save both the patient and the healthcare industry from considerable costs (ADA, n.d.-b).

While this analysis will allow for the prediction of a diabetes diagnosis based on associated characteristics, our central goal is to infer the behaviors, demographics, and circumstances most related to being at risk for diabetes. Our specific project goals are stated below:

- Employ machine learning models and methods to predict respondents' diabetes diagnosis
- Conduct a comprehensive feature selection process to identify key risk factors
- Infer how certain behaviors, conditions, and demographics influence the risk of having diabetes
- Harness data insights to prescribe specific actions physicians, policymakers, and patients can take to reduce diabetes incidence

By achieving these goals, our analysis will allow for several beneficial outcomes:

- Empower medical professionals to more efficiently recommend diabetes screenings to patients
- Inform more strategic education efforts and policymaking on diabetes
- Allow people to identify modifiable behaviors and reduce their risk of diabetes

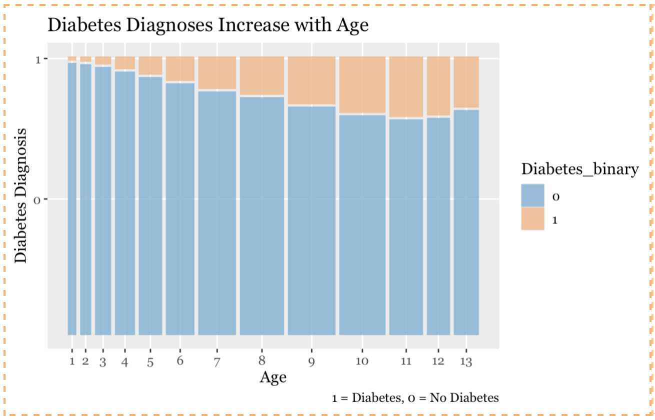
To achieve the proposed goals and their resulting outcomes, we have detail the following project road map:

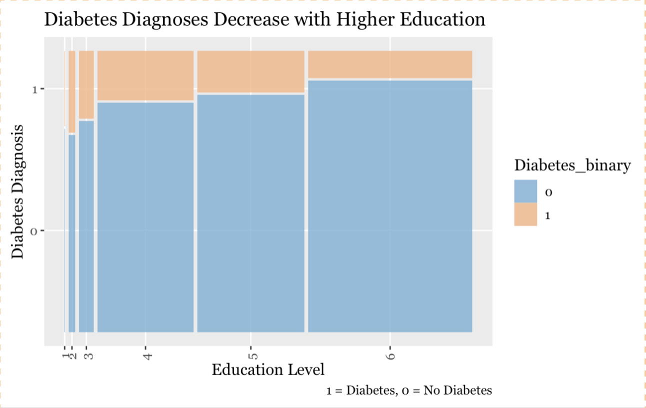
1. **Exploratory Data Analysis (EDA):** By using various visualization tools, we will illustrate how our predictors are associated with the response **Diabetes_binary,** as well as one another through interaction effects. These initial findings will prove helpful in directing our model building process.
2. **Model Building and Feature Selection:** With a binary response variable, we will try to build various classification models including logistic regression, lasso regression (regularization), generalized additive models (GAMs), decision trees, etc. Additionally, we will utilize different feature selection techniques to identify predictors that improve model metrics (misclassification rate, AIC, BIC, etc.).
3. **Analysis and Recommendations:** After selecting our final model, we will investigate its meaning in both quantitative and qualitative contexts. Using these insights, we will craft tangible recommendations for diabetes prevention.

II. Data

The data for this analysis is from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS), a collaborative survey distributed throughout all states and select territories of the US (CDC, 2022a). The goal of the BRFSS is to collect comprehensive data on the health practices and behaviors that may be linked to disease and injury. To explore the specific risk factors that may be associated with diabetes, we use a subset of the larger BRFSS data set posted by user Alex Teboul on Kaggle. Using background research on diabetes incidence, Teboul identified select variables that are known to have a relationship with the chronic disease. The dataset has 22 variables and 253,680 observations – and has already been cleaned. There is no missingness in the dataset.

In preparation for our analysis and model-building process, we convert all categorical variables to factor variables. In addition, we transform the variables **PhysHlth** and **MentHlth** from doubles to type integers, as they represent a number of days, and in turn, must be a whole number. Finally, to make our analysis more interpretable for particular demographics, we consolidate the number of categories in the **Age** and **Education** variables. Previously containing 13 different age levels, **Age** now is captured by three levels (18-39, 40-59, and 60+). **Education** previously had six levels and now is captured by four levels (Some High School or Less, High School Grad, Some College, College Graduate). To verify that these adjustments are appropriate, we ensure that all age and education categories being binned together share similar proportions of diagnoses (see figures below). This adoption will allow us to more succinctly interpret our model in the case **Age** is a significant predictor. The variables in our data set are described below:





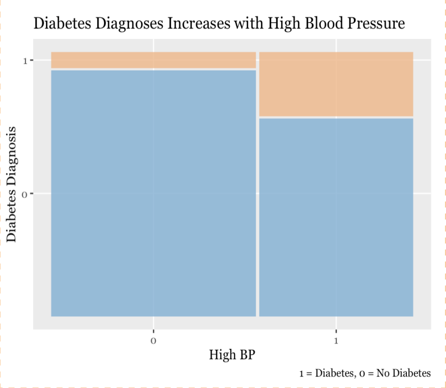
Variable	Type	Description
Diabetes_binary	fct	Respondent’s diabetes diagnosis status as indicated by levels: 0 = no diabetes, 1 = diabetes
HighBP	fct	Respondent has high blood pressure (0 = no, 1 = yes)
HighChol	fct	Respondent has high cholesterol (0 = no, 1 = yes)
CholCheck	fct	Respondent has had a cholesterol check in past 5 years (0 = no, 1 = yes)
BMI	dbl	Body Mass Index (BMI) of respondent
Smoker	fct	Respondent has smoked at least a 100 cigarettes in their life (0 = no, 1 = yes)
Stroke	fct	Respondent has ever had a stroke (0 = no, 1 = yes)
HeartDiseaseorAttack	fct	Respondent has had coronary heart disease or myocardial infarction (0 = no, 1 = yes)
PhysActivity	fct	Respondent has done physical activity in past 30 days (0 = no, 1 = yes)
Fruits	fct	Respondent consumes fruit once or more a day (0 = no, 1 = yes)
Veggies	fct	Respondent consumes vegetables once or more a day (0 = no, 1 = yes)
HvyAlcoholconsump	fct	Respondent is a heavy drinker (for men > 14 drinks/week, for women > 7 drinks/week) (0 = no, 1 = yes)
AnyHealthCare	fct	Respondent has any kind of health insurance (0 = no, 1 = yes)
NoDocbcCost	fct	Respondent needed to go to doctor in past 12 months but could not because of cost (0 = no, 1 = yes)
GenHlth	fct	Respondent’s ranking of their general health with levels: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor
MentHlth	int	Respondent’s answer to : How many days during the past 30 days was your mental health not good? (scale 1-30 days)
PhysHlth	int	Respondent’s answer to : How many days during the past 30 days was your physical health not good? (scale 1-30 days)
DiffWalk	fct	Respondent has serious difficulty walking or climbing stairs (0 = no, 1 = yes)
Sex	fct	Respondent’s sex (0 = female, 1 = male)
Age	fct	Respondent’s age category: 1 = 18-24, 9 = 60-64, 13 = 80 or older
Education	fct	Respondent’s highest education level on scale: 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	fct	Respondent’s income level on scale: 1 = less than $10,000, 5 = less than $35,000, 8 = $75,000 or more

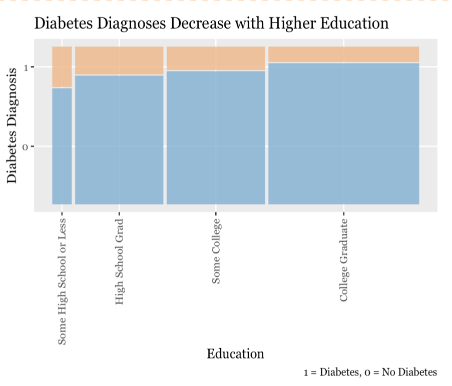
The response variable in our analysis is **Diabetes_binary**, as we aim to determine what health risk factors are associated with a diabetes diagnosis. There are 21 predictor variables; however, we can use exploratory data analysis (EDA) to determine which of these are more important to our model building process. By identifying the variables that are most significantly related to **Diabetes_binary**, we can develop a final model that is both robust and interpretable.

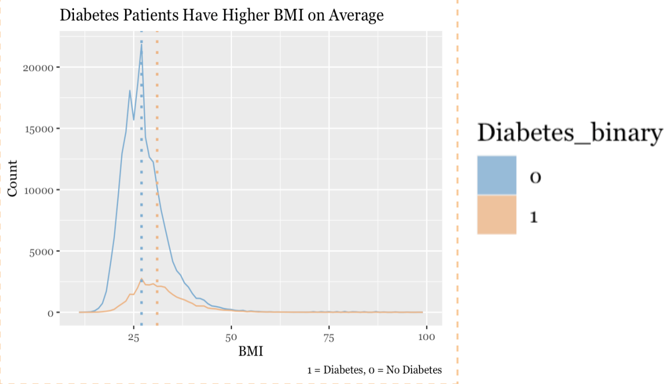
Note that the dataset almost exclusively contains categorical variables, which limits the precision of our results and the number of statistical modeling options. Also, variables rating patient health on a scale from 1 to 5 are self-evaluated, making them subjective and difficult to compare between observations.

III. Exploratory Data Analysis (EDA)

After initializing our data set, we perform exploratory data analysis on our predictor and response variables. To investigate the significance of our categorical predictors, we create mosaic plots. Mosaic plots are useful for visualizing how the proportion of our response **Diabetes_binary** changes when dependent on a particular explanatory variable. Based on these plots, we expect the predictors **HighBP**, **HighChol**, **BMI**, **GenHlth**, **Age**, **Education**, **Income**, **PhysActivity**, **PhysHlth**, and **DiffWalk** to have strong relationships with our response. Their mosaic plots demonstrated that the proportion of those with a diabetes diagnosis visibly changed when depended on (see left and center figs. below).



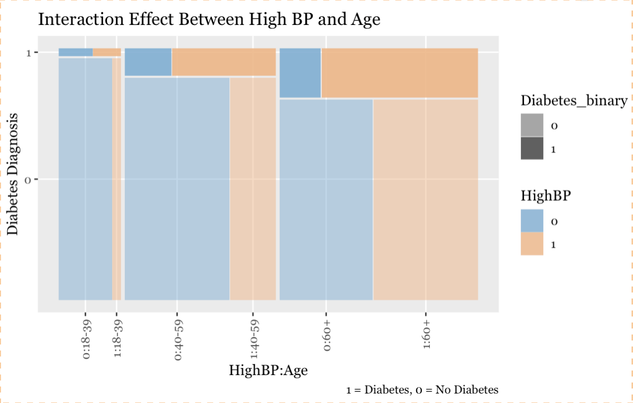




To better understand the relationship between our numerical predictors and the response, we create frequency polygon plots (see figures above). These plots are similar to histograms, but are more suitable when comparing across levels of a categorical variable. We expect the numerical variables **BMI** and **PhysHlth** to exhibit important relationships with **Diabetes_binary** because their medians vary dependent on a diabetes diagnosis. The median is used as the measure of center instead of the mean seeing as their distributions are skewed.

Based on additional background research, it is reasonable to assume that several of our predictors are interrelated with one another and therefore necessitate interaction terms. To visually explore this possibility, we create three-variable mosaic plots (see figure below). Testing multiple different combinations, we believe that the interaction terms **HighBP:Age**, **HeartDiseaseorAttack:Age**, **HighChol:Age** may be helpful to our model. Their mosaic plots demonstrate that the proportion of respondents with said health conditions varies as the age differs, meaning they share a relationship in context of a diabetes diagnosis.

Next, we check for multicollinearity between predictors by calculating the variance inflation factors (VIFs). All of the predictors in the logistic classifier model have VIFs between 1 and 5––and none exceed 5––so we do not need to remove any variables for multicollinearity risk.



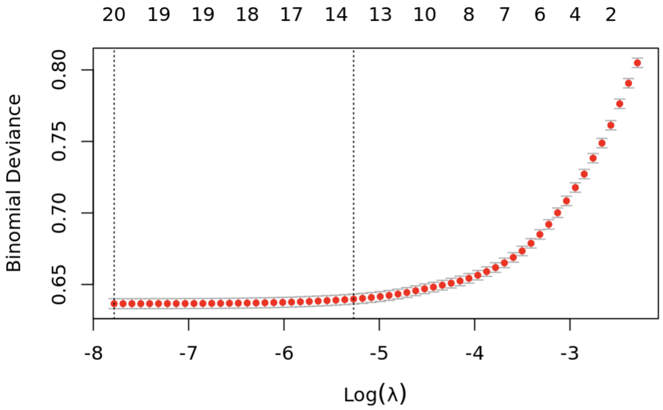
IV. Methodology

Logistic Classifier

We begin with a baseline model, logistic classification, to guide our investigation of the important determinants of diabetes. We begin constructing an initial model with all of the predictors in the dataset. Since the primary focus of our analysis is to infer which small set of factors most influences diabetes risk, we aim to select a more parsimonious model, so we are adopting a significance level of 0.0001. At this significance level, this model determines all predictors as statistically significant except **Smoker, Fruits, Veggies, and NoDocbcCost**. The baseline model yields a misclassification rate of 13.83 percent on the test set after a 70-30 train-test split of the binary diabetes data. To begin the feature selection process, we run a backward stepwise selection on the full model, which eliminates unnecessary predictors based on the Akaike Information Criterion (AIC). Backward selection only eliminates the variable **NoDocbcCost.** While AIC is commonly used to calculate prediction error, there are other metrics that may be more effective in a classification context.

Term	Coefficient
Intercept	-8.9109
HighBP	0.7503
HighChol	0.5064
CholCheck	0.5833
BMI	0.0539
Smoker	0
Stroke	0.0619
HeartDiseaseOrAttack	0.2287
PhysActivity	-0.0029
Fruits	0
Veggies	0
HeavyAlcoholConsump	-0.4644
AnyHealthcare	0
NoDocbcCost	0
GenHlth	0.4768
MentHlth	0
PhysHlth	0
DiffWalk	0.0902
Sex	0.1516
Age	0.4917
Education	-0.0203
Income	-0.0426

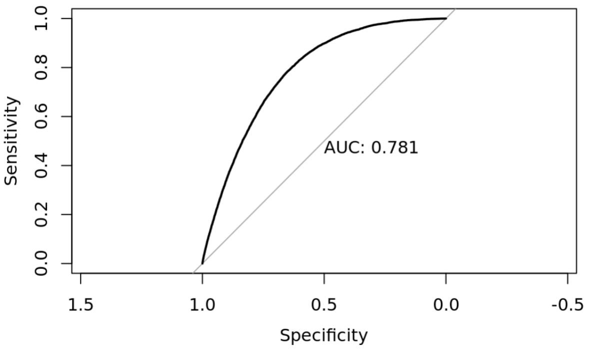
Since LASSO regularization is generally preferred over criterion-based methods like backward selection for variable selection in logistic classification, we proceed with LASSO regularization. Due to its L1-norm, LASSO can regularize variable coefficients to zero unlike ridge regression. LASSO eliminates the following variables (as shown in the figure to the left): **Smoker, Fruits, Veggies, AnyHealthcare, NoDocbcCost, MentHlth,** and **PhysHlth.**



Thus, the model after LASSO is:

$$\begin{aligned}&PredictedDiabetes(0|1) = -8.91 + 0.75 * HighBP + 0.51 * HighCol + 0.58 * CholCheck\\&+ 0.05 * BMI + 0.06 * Stroke + 0.23 * HeartDiseaseOrAttack - 0.003 * PhysActivity\\&- 0.46 * HeavyAlcoholConsump + 0.48 * GenHlth + 0.09 * DiffWalk + 0.15 * Sex\\&+ 0.49 * Age - 0.02 * Education - 0.04 * Income\end{aligned}$$

This model, however, yields a higher dfassification error on the test set than the baseline logistic classifier: 24.7 percent. Particularly, the false negative rate of 61.5 percent is extremely high, which is problematic because potential diabetes patients should be notified as quickly as possible to take necessary remediation steps. The AUC for this model, 0.781, is also lower than the AUC of the previous model (0.82).



Next, we add interactions to this model to see if we can lower the misclassification error. Logically, many health factors interact with demographic information, especially age. We can start with interactions between **Age-HighBP**, **Age-HeartDiseaseorAtatck,** and **Age-HighChol**, as these are all health conditions with risks that are exacerbated as age increases. We then explore interactions between **GenHlth** and factors that could have a strong effect on it: **BMI**, **Income**, **Age**, and **Sex.** These logical choices for interactions are further supported by our three-variable mosaic plots, which demonstrate that the incidence of diabetes is more prevalent when these factors are combined (see EDA section for example).

The final logistic classification model with interactions is as follows. The results and evaluation of this model are outlined in the Results section:

$$\begin{aligned}&PredictedDiabetes(0|1) = -8.74 + 1.09 * HighBP + 1.32 * HighChol + 1.25 * CholCheck +\\&0.07 * BMI + 0.15 * Stroke + 0.80 * HeartDiseaseOrAttack - 0.04 * PhysActivity -\\&0.78 * HvyAlcoholConsump + 0.64 * GenHlth + 0.09 * DiffWalk + 0.61 * Sex\\&+ 0.30 * Age - 0.06 * Education - 0.26 * Income - 0.04 * HighBP : Age -\\&0.05 * HeartDiseaseOrAttack : Age - 0.08 * HighChol : Age - 0.03 * GenHlth : Age\\&- 0.11 * GenHlth : Sex - 0.003 * GenHlth : BMI + 0.06 * GenHlth : Income\end{aligned}$$

Boosted Classification Tree

We then construct a boosted classification tree which classifies patients as either having diabetes or not having diabetes based on all the predictor variables in our dataset. Boosted trees are built by fitting a tree on the data with only the number of splits specified by the interaction depth parameter, subsequently fitting new trees on the residuals of the previous tree (still with the interaction depth number of splits), then adding all the trees together. By constructing many small trees based on previous trees' errors and learning slowly, boosted trees slowly improve in areas where they do not perform well, thus ultimately making them more accurate. We build the boosted tree using the training dataset, an interaction depth of 3, and 500 trees.

V. Results

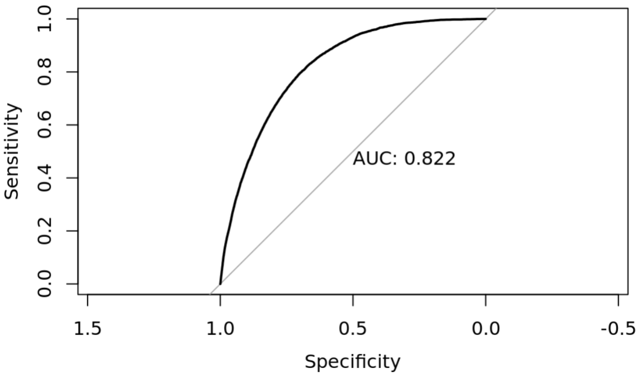
Logistic Classifier

The final model obtained from our logistic classification approach described in the methodology section is:

$$
\begin{aligned}
&\mathit{PredictedDiabetes}(0|1) = -8.74 + 1.09 * \mathit{HighBP} + 1.32 * \mathit{HighChol} + 1.25 * \mathit{CholCheck} + \\
&0.07 * \mathit{BMI} + 0.15 * \mathit{Stroke} + 0.80 * \mathit{HeartDiseaseOrAttack} - 0.04 * \mathit{PhysActivity} - \\
&0.78 * \mathit{HvyAlcoholConsump} + 0.64 * \mathit{GenHlth} + 0.09 * \mathit{DiffWalk} + 0.61 * \mathit{Sex} \\
&+ 0.30 * \mathit{Age} - 0.06 * \mathit{Education} - 0.26 * \mathit{Income} - 0.04 * \mathit{HighBP} : \mathit{Age} - \\
&0.05 * \mathit{HeartDiseaseOrAttack} : \mathit{Age} - 0.08 * \mathit{HighChol} : \mathit{Age} - 0.03 * \mathit{GenHlth} : \mathit{Age} \\
&- 0.11 * \mathit{GenHlth} : \mathit{Sex} - 0.003 * \mathit{GenHlth} : \mathit{BMI} + 0.06 * \mathit{GenHlth} : \mathit{Income}
\end{aligned}
$$

This model with interactions yields a misclassification error rate on the test set of 13.59 percent. Additionally, the false negative rate of 13.99 percent is much lower than that of the post-LASSO model (61.5 percent) described in the methodology section. The false positive rate is low, at only 1.80 percent. The AUC is 0.822, which is the highest of the models evaluated thus far. Logistic classifier models with an AUC above 0.80 are considered highly effective at discriminating between both cases of a binary response variable. Interactions clearly improve the accuracy of the model on test data in the case of predictive diabetes modeling. They account for more complex relationships in patient factors, such as age and late-onset health factors, than a model with solely individual predictors, such as the model after LASSO, can include.

Test Accuracy	86.41%
MC Test Error	13.59%
False Positive Rate	1.80%
False Negative Rate	13.99%



Overall, the logistic classifier on the binary diabetes data yielded a fairly accurate model after LASSO regularization and inclusion of interactions based on mosaic plots. The accuracy rate of 86.39 percent means that the vast majority of diabetes patients will be accurately diagnosed by this logistic classifier. There is a notable false negative risk (13.99 percent), so patients should treat this model as a fairly accurate indicator, rather than conclusive medical diagnosis. The logistic classifier can be used for both making diabetes predictions on new patients and drawing inferences, while the boosted classification tree is more useful for evaluating the relative importance of features.

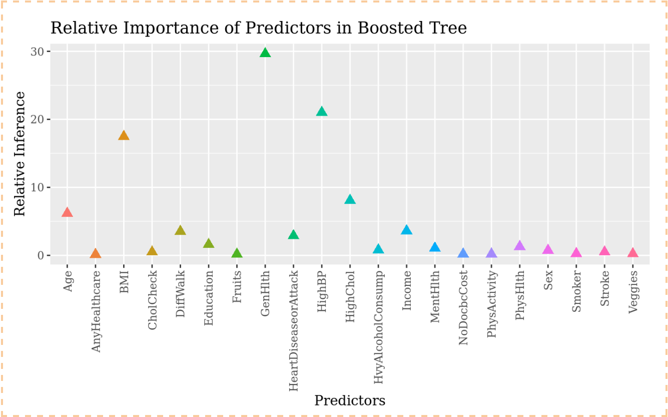
Boosted Classification Tree

The table and plot below show the relative influence of each variable in building the boosted tree.

Variable	Relative Influence
GenHlth	29.634
HighBP	21.016
BMI	17.47
HighChol	8.081
Age	6.176
Income	3.603
DiffWalk	3.52
HeartDiseaseorAttack	2.877
Education	1.596
PhysHlth	1.289
MentHlth	1.059
HvyAlcoholConsump	0.798
Sex	0.744
Stroke	0.491
CholCheck	0.483
Smoker	0.256
Veggies	0.221
PhysActivity	0.197
NoDocbcCost	0.193
Fruits	0.166
AnyHealthcare	0.132

General health, high blood pressure, and BMI have the greatest influence on the construction of the boosted tree. Thus, these predictor variables have the strongest association with diabetes diagnoses. The table below shows the test metrics for the boosted classification tree.

Test Accuracy	86.6%
False Positive Rate	12.1%
False Negative Rate	43.8%



While this overall accuracy is very high, we should be cautious of the high false negative rate, especially because false negatives are much more costly than false positives.

VI. Discussion

Overall, the boosted classification tree is most useful for determining the relative order of risk factors of diabetes, while the logistic classifier with LASSO regularization and interaction terms is better for patient diagnosis. Particularly, the logistic model has a much lower false negative rate (13.99 percent) than the boosted classification tree (43.8 percent), while both models have comparable misclassification test errors. The logistic model is better for patient diagnoses because it alleviates much of the false negative risk, which is problematic because it delays corrective health measures for those newly diagnosed with diabetes.
The boosted classification tree ranks the relative order of influence of the predictors in our dataset, offering important insight into the risk factors of diabetes. This relative influence metric ranks general health, high blood pressure, high BMI, and high cholesterol as the strongest predictors of diabetes incidence. While the boosted classification tree has high test accuracy, the false negative rate is extremely high and therefore dangerous to use in the context of diabetes.

Interpreting the coefficients of our models yields a lot of insight for how healthcare professionals should prioritize patients for diabetes screening, individuals should optimize their lifestyle habits to reduce the risk of diabetes, educators should inform the public about diabetes risk, and legislators should create policy to support public health with regard to diabetes.

Based on the logistic classifier, the controllable health factors of blood pressure, cholesterol, BMI, physical activity, and general health are all significant indicators of diabetes incidence. The significant interaction terms between age and these health conditions indicate that aging patients should be even more cautious about health indicators like blood pressure. By reversing these controllable factors to healthy levels, a patient can reduce his or her log-odds of diabetes incidence. These reductions are as follows:

Health Factor	**Reduction of Log-Odds of Predicted Diabetes**		
Age	Age 18-39	Age 40-59	Age 60+
High Blood Pressure	1.09	1.13	1.43
High Cholesterol	1.32	1.40	1.70
BMI*	0.15	0.15	0.15
Regular Physical Activity	0.04	0.04	0.04
General Health**	0.64	0.64	0.64

* For each unit reduction in BMI ** For each unit improvement in general health (on scale of 1- excellent to 5-poor)

According to these results, individuals should focus on creating lifestyle habits which optimize these variables. Exercising regularly and eating healthy foods reduce blood pressure, cholesterol, and BMI while increasing regular physical activity levels and general health.

The boosted classification tree shows that general health, high blood pressure, high BMI, high cholesterol, age, and income are the most influential variables for diabetes prediction. Therefore, medical professionals should be alert to patients with high blood pressure, high cholesterol, high BMI, poor general health, old age, and low income. Those individuals are at high risk for diabetes and should therefore be prioritized in screening for diabetes.

Many of these diabetes risk factors can be optimized by adopting healthy lifestyle habits from a young age. For example, by staying active daily, eating healthy regularly, not smoking, and not heavily consuming alcohol, almost all the greatest risk factors for diabetes can be prevented. So, it is crucial to establish education programs in schools that teach students about healthy habits to prevent issues like diabetes. In addition, because being of older age increases risk of diabetes even more, education programs should also be established at work and in retirement homes for adults and the elderly, respectively.

Our results show there are many ways public policy can reduce risk of diabetes for the general public. First more specifically, our logistic regression model shows a negative coefficient for income, meaning that each one-unit decrease in income increases an individual's log-odds of having diabetes. Therefore, the government should be creating policies which provide more health resources and support to low-income populations. Access to quality healthcare is important for every individual's general health, and it is especially an issue for low-income people. Thus, our public policy should be specifically targeting low-income individuals and families to provide them with education and health insurance.

There are a few limitations of our model that are important to note before it can be deployed in critical public health contexts. First, we used the reconfigured dataset, which had a binary response variable, instead of the original dataset, which had 3 categories for the response variable. This reconfiguration treated prediabetes patients equivalently to non-diabetics, which might have affected the conclusions of our analysis. Second, both of our models had relatively high false negative rates, especially the boosted classification model, so there is a predictive power limitation in diagnosing prospective diabetes patients. The imbalanced nature of the dataset, which had far more non-diabetic patients, could have contributed to a much higher false negative than false positive rate for both models. Finally, some of our variables were self-reported categorical values, such as patients evaluating their own health from 1 through 5. Different patients may view these vague health ratings differently, which threatens the reliability and consistency of the data.

VII. Conclusion

Diabetes is a critical health problem that afflicts millions of people in the United States and worldwide. The number of patients affected by diabetes is only increasing, which is a concerning trend that requires robust risk factor analysis to reverse. Patients need to be properly informed about their risk factors of diabetes and how they can adapt their lifestyle to minimize likelihood of diabetes. Our statistical investigation finds a narrow set of risk factors to help doctors with patient selection for screening and patients with preventative measures to incorporate into their lifestyle. These factors, especially general health, blood pressure, cholesterol, and BMI, should guide potential diabetes and prediabetes patients on how to live a healthy lifestyle. Moreover, even the uncontrollable factors that our analysis found to be predictive of diabetes, including age, income, and heart disease history, inform doctors of which patients should be prioritized for screening. To further expand this study, we would make predictions for three types of patients instead of two: diabetics, pre-diabetics, and non-diabetics; this would better inform treatment and diet options for all three groups. With the knowledge from our investigation and future research inquiries on this subject, healthcare providers and patients alike can have a better picture of the determinants of diabetes to boost patient health outcomes.

VIII. Bibliography

ADA. (n.d.-a). *Statistics About Diabetes*. Retrieved December 17, 2022, from https://diabetes.org/about-us/statistics/about-diabetes

ADA. (n.d.-b). *The Cost of Diabetes*. Retrieved December 17, 2022, from https://diabetes.org/about-us/statistics/cost-diabetes

CDC. (2022a, August 29). *BRFSS*. https://www.cdc.gov/brfss/index.html

CDC. (2022b, September 6). *Leading Causes of Death*. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

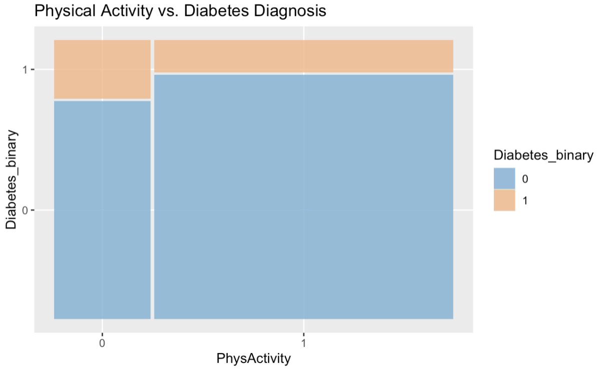
CDC. (2022, September 21). *Prevalence of Diabetes*. https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html

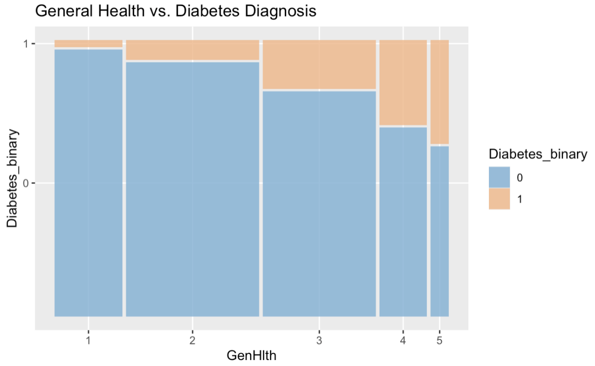
Juvenile Diabetes Cure Alliance (JDCA). (2022, September 18). *NIH Diabetes Research Funding FY2021: 2022*. https://www.thejdca.org/publications/report-library/archived-reports/2022-reports/nih-diabetes-research-funding-fy2021.html

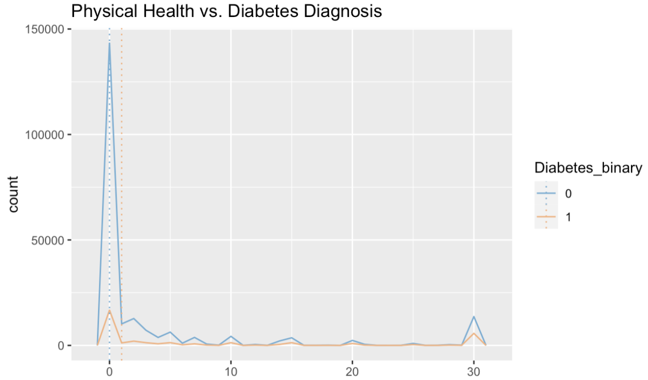
Slobin, J. (2022, August 23). *How Much Does a Diabetes Screening Test Cost Without Insurance in 2021? | Mira*. https://www.talktomira.com/post/what-is-a-diabetes-screening-test-and-how-much-it-costs

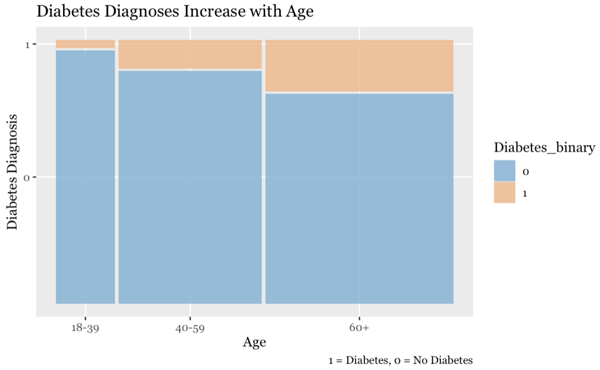
IX. Appendix

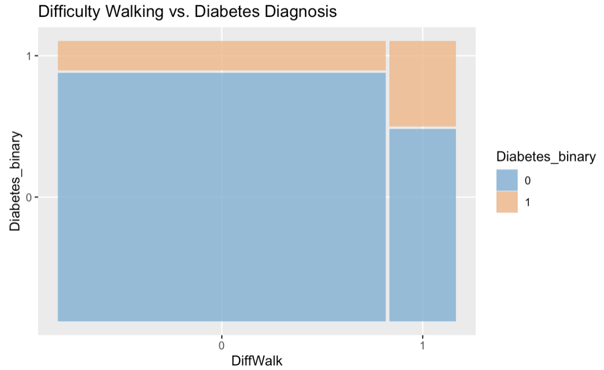
Appendix I - EDA Visualizations

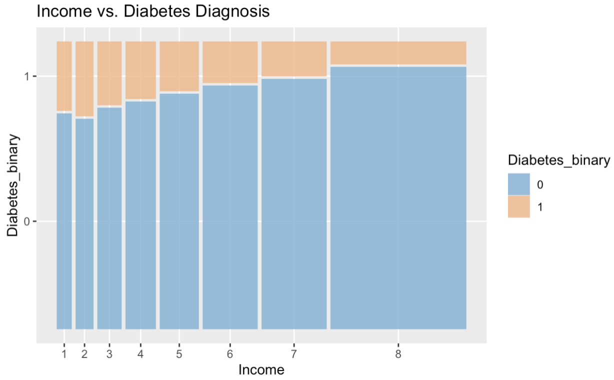

Physical Activity vs. Diabetes Diagnosis

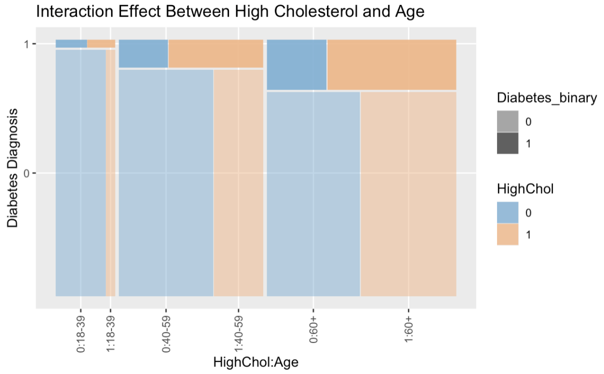

General Health vs. Diabetes Diagnosis

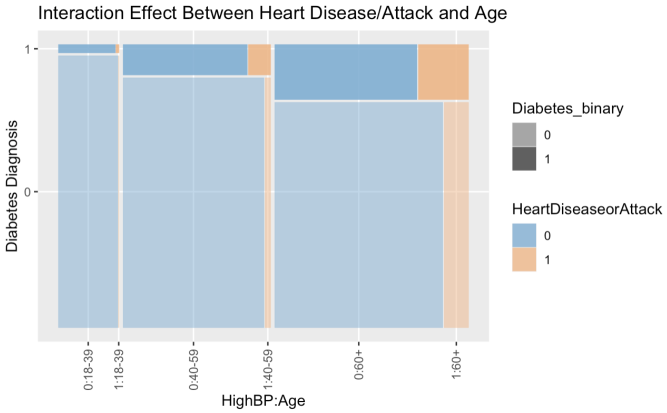

Physical Health vs. Diabetes Diagnosis


Diabetes Diagnoses Increase with Age


Difficulty Walking vs. Diabetes Diagnosis

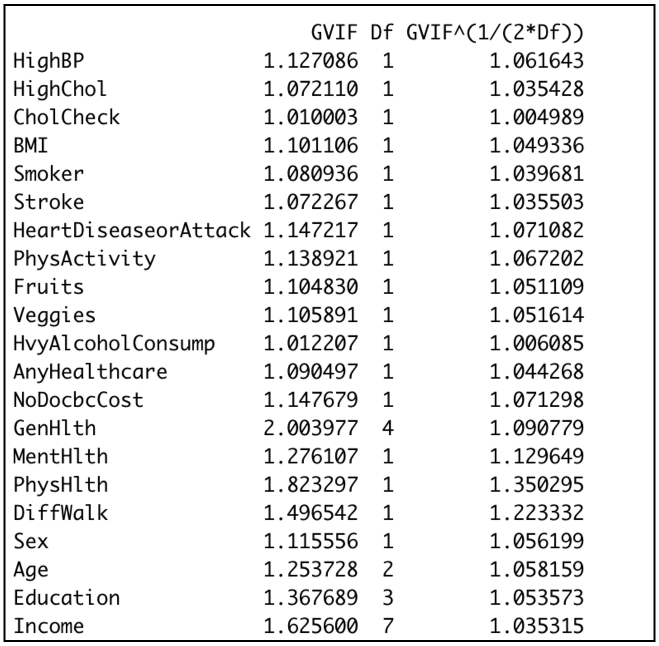

Income vs. Diabetes Diagnosis


Interaction Effect Between High Cholesterol and Age


Interaction Effect Between Heart Disease/Attack and Age

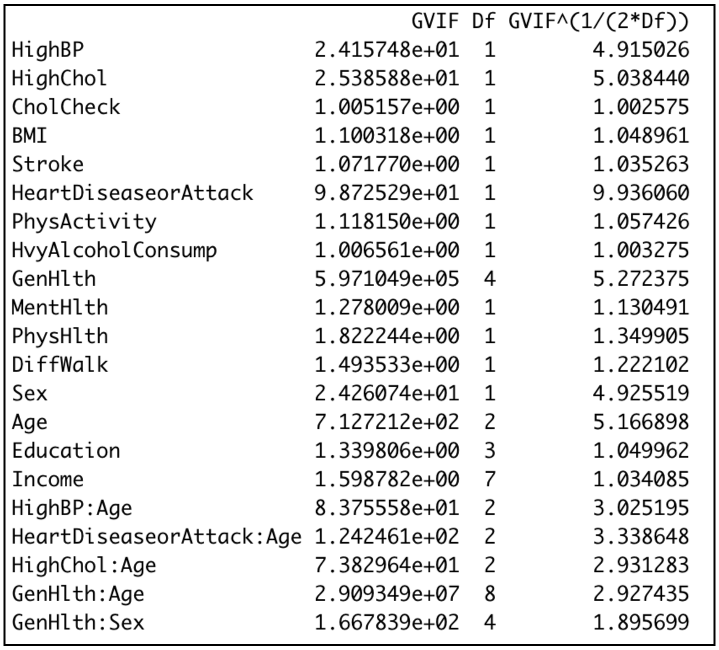
Appendix II - VIF Output

Logistic w/o Interactions



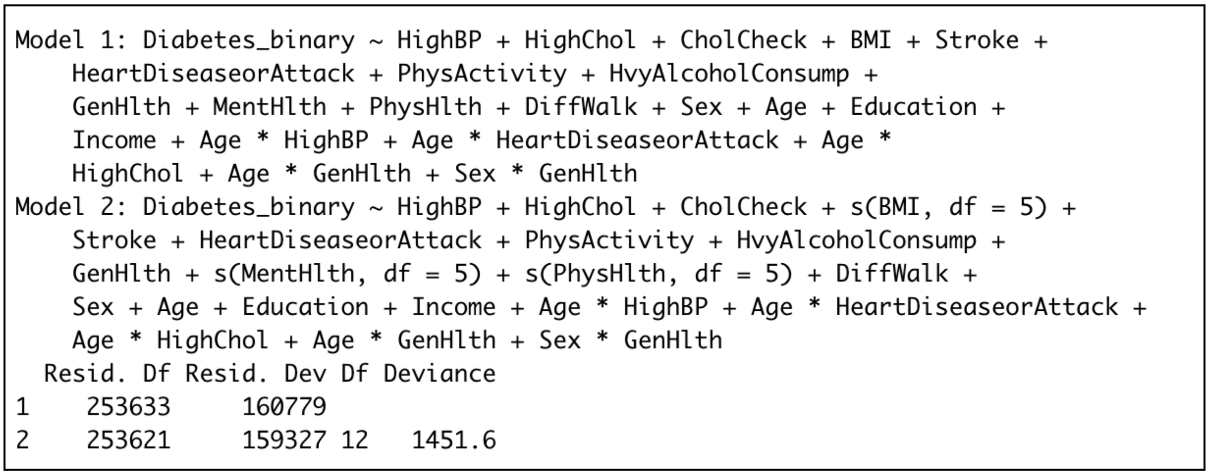
```
                         GVIF Df GVIF^(1/(2*Df))
HighBP               1.127086  1        1.061643
HighChol             1.072110  1        1.035428
CholCheck            1.010003  1        1.004989
BMI                  1.101106  1        1.049336
Smoker               1.080936  1        1.039681
Stroke               1.072267  1        1.035503
HeartDiseaseorAttack 1.147217  1        1.071082
PhysActivity         1.138921  1        1.067202
Fruits               1.104830  1        1.051109
Veggies              1.105891  1        1.051614
HvyAlcoholConsump    1.012207  1        1.006085
AnyHealthcare        1.090497  1        1.044268
NoDocbcCost          1.147679  1        1.071298
GenHlth              2.003977  4        1.090779
MentHlth             1.276107  1        1.129649
PhysHlth             1.823297  1        1.350295
DiffWalk             1.496542  1        1.223332
Sex                  1.115556  1        1.056199
Age                  1.253728  2        1.058159
Education            1.367689  3        1.053573
Income               1.625600  7        1.035315
```

Logistic w/ Interactions



```
                                 GVIF Df GVIF^(1/(2*Df))
HighBP                   2.415748e+01  1        4.915026
HighChol                 2.538588e+01  1        5.038440
CholCheck                1.005157e+00  1        1.002575
BMI                      1.100318e+00  1        1.048961
Stroke                   1.071770e+00  1        1.035263
HeartDiseaseorAttack     9.872529e+01  1        9.936060
PhysActivity             1.118150e+00  1        1.057426
HvyAlcoholConsump        1.006561e+00  1        1.003275
GenHlth                  5.971049e+05  4        5.272375
MentHlth                 1.278009e+00  1        1.130491
PhysHlth                 1.822244e+00  1        1.349905
DiffWalk                 1.493533e+00  1        1.222102
Sex                      2.426074e+01  1        4.925519
Age                      7.127212e+02  2        5.166898
Education                1.339806e+00  3        1.049962
Income                   1.598782e+00  7        1.034085
HighBP:Age               8.375558e+01  2        3.025195
HeartDiseaseorAttack:Age 1.242461e+02  2        3.338648
HighChol:Age             7.382964e+01  2        2.931283
GenHlth:Age              2.909349e+07  8        2.927435
GenHlth:Sex              1.667839e+02  4        1.895699
```

Appendix III - GAM Output



```
Model 1: Diabetes_binary ~ HighBP + HighChol + CholCheck + BMI + Stroke +
    HeartDiseaseorAttack + PhysActivity + HvyAlcoholConsump +
    GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Age + Education +
    Income + Age * HighBP + Age * HeartDiseaseorAttack + Age *
    HighChol + Age * GenHlth + Sex * GenHlth
Model 2: Diabetes_binary ~ HighBP + HighChol + CholCheck + s(BMI, df = 5) +
    Stroke + HeartDiseaseorAttack + PhysActivity + HvyAlcoholConsump +
    GenHlth + s(MentHlth, df = 5) + s(PhysHlth, df = 5) + DiffWalk +
    Sex + Age + Education + Income + Age * HighBP + Age * HeartDiseaseorAttack +
    Age * HighChol + Age * GenHlth + Sex * GenHlth
  Resid. Df Resid. Dev Df Deviance
1    253633     160779
2    253621     159327 12   1451.6
```