BST 263 FINAL PROJECT

Predicting Diabetes Status with the Help of Machine Learning

Willow Duffell, Rindala Fayyad, Daniel Herrera, Bhawesh Kumar, Lauren Mock

Department of Biostatistics, Harvard TH Chan School of Public Health

May 2022

1. Introduction

37.3 million Americans have diabetes, totaling 11.3% of the US population. Its prevalence is high, and its impact on the individuals it effects can be devastating. Uncontrolled, diabetes can lead to severe complications such as limb loss, blindness, heart disease and renal disease. Additionally, the treatments for diabetes can be burdensome financially and logistically, often requiring several self-administered injections and invasive glucose level tests. Early detection of prediabetic and diabetic individuals affords them the opportunity to make significant and consequential lifestyle changes before further serious medical treatment is required, and many millions of Americans are currently undiagnosed diabetics (Centers for Disease Control and Prevention, 2022). As a result, understanding the risk factors for diabetes and identifying individuals who may be undiagnosed diabetics is highly important.

One of the main issues with diabetes is that so many diabetics are undiagnosed. It is estimated by the CDC that 1 in 5 diabetics and about 8 in 10 prediabetics are unaware of their status (Centers for Disease Control and Prevention, 2021). This could be due to the lack of knowledge of the topic, lack of doctors' visits to test glucose levels, or a lack of healthy lifestyle habits. Filling this knowledge gap would potentially result in people taking the disease more seriously and realizing what healthy lifestyle choices are necessary moving forward. Furthermore, there is evidence that those who are of lower socioeconomic status are more at risk of diabetes and experience more of the burdens that the disease provides (Hill *et al*, 2013). Regardless of status, those diagnosed with diabetes still experience financial burdens. It is estimated that diagnosed diabetes costs about $327 billion dollars per year, and the total costs for undiagnosed diabetes and prediabetes are nearing $400 billion annually (American Diabetes Association, 2018).

The first goal of this study is to build an interpretable model that predicts diabetes status based on easily collected feature variables. The second goal of this study is to determine which factors are most predictive of diabetes risk. To achieve these goals, we will consider both classification and regression models by fitting LDA, QDA, logistic, LASSO, neural networks, and boosting models.

2. Methods

We are using a publicly available dataset from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS) containing 20 predictor variables and 253,680 complete survey responses. The 20 predictor variables in the dataset are: High Blood Pressure Indicator, High Cholesterol Indicator, Cholesterol Check in the Past 5 Years Indicator, BMI, Smoker Indicator, Stroke Indicator, Heart Disease or Heart Attack Indicator, Physical Activity in the Past 30 Days Indicator, At Least one Fruit Consumption per Day Indicator, At Least one Vegetable Consumption per Day Indicator, Heavy Alcohol Consumption Indicator, Presence of Any Health Care Indicator, Could Not See a Doctor Because of Cost Indicator, General Health Status, Frequency of Bad Mental Health in Past 30 Days, Frequency of Bad Physical Health in Past 30 Days, Difficulty Walking Indicator, Sex, Age, Highest Education Level, Income Level.

The outcome variable in our study is a two-class Diabetes indicator, where the first class corresponds to those who self-identified as non-diabetic or prediabetic, and the second class corresponds to those who self-identified as diabetic.

Initially, our outcome variable consisted of three classes that were in the original survey data: 0=non-diabetic, 1=prediabetic, and 2=diabetic. However, this dataset was very imbalanced, as there were around 84% non-diabetic individuals, 2% prediabetic individuals, and 14% diabetic individuals. After randomly splitting the dataset into 75% training set and 25% test set, we fit all six models on the training set and then computed the train and test error, accuracy, and confusion matrices for each model. Due to the very

low number of prediabetic individuals, most of the six models did not predict a single individual to be pre-diabetic; all individuals were classified as non-diabetic or diabetic. Brownlee (2019) explains that this is because most machine learning algorithms used for classification were designed under the assumption of an equal number of observations in each class. Thus, when we provide the model with an imbalanced dataset, the model will perform poorly in predicting the minority class—in this case, the pre-diabetic class. Figure 1 below shows this poor predictive performance for the prediabetic class (outcome = 1).

Figure 1: Confusion matrices for 3 classes (0 = non-diabetic, 1 = pre-diabetic, and 2 = diabetic) for 4 different models. (From left to right, top to bottom: LDA, QDA, Multinomial, LASSO)

To avoid this issue of poor predictive performance in the prediabetic class, we chose to combine the non-diabetic and prediabetic classes into a single class. We made this decision because many of the people who self-identified as non-diabetic were likely unaware that they were actually pre-diabetic. In fact, one

third of the US population is pre-diabetic according to a review of current data from the Centers for Disease Control and Prevention (CDC) done by Blue Cross Blue Shield, so a large percentage of individuals in this dataset have incorrectly self-identified as non-diabetic. Hence, our new dataset resulted in two classes: 1 = non-diabetic or prediabetic, and 2 = diabetic.

Before beginning our analyses on the modified dataset with two classes for the outcome variable, we set the seed to 356. We used the same train/test split of 75%/25% that we used on the dataset with the 3-class outcome. We then re-fit LDA, QDA, logistic, LASSO, neural networks, and boosting models, where all the 20 variables in the dataset were included as predictor variables. For the logistic regression model, we used the backward selection algorithm to select the most important variables, which resulted in the elimination of the "Fruits" and "Veggies" variables. The LASSO model eliminated the smoking status and the indicator for not being able to see a doctor due to cost.

Concerning the LASSO model, we used K-fold cross-validation (CV) to choose the best tuning parameter. We initially attempted K = 10, but because of the very large dataset that we have, the CV algorithm would not run in R. We then had to reduce the number of folds to K = 5. After getting the "best" tuning parameter/lambda, we used that value in our training and test predictions and in our calculations of the training and test error rates.

We proceeded differently for the boosting model. To find the best shrinkage parameter, we created a vector of possible values for the shrinkage parameter ranging from 10e-10 to 1, with small increments. We then fit a Boosting model on the training set for each value in the shrinkage vector and used that model to make predictions on the training and test sets. We chose the shrinkage parameter value that minimized the test error rate. We used 500 trees in each boosting model. Because of a technical issue similar to what we faced with the CV mentioned above, we were not able to run the model with 1000 trees, which would have been ideal, and thus had to limit the number of trees to 500.

In order to fit a standard feed-forward neural network, data processing was required. Specifically, categorical variables were transformed using one hot encoding. Additionally, a slightly different approach was used for the training, testing, and validation data splits since all of these are standard for a neural network to update the parameters accordingly. 60% of the data was set aside for training, 20% of the data was set aside for testing and 20% for validation. Several considerations for network architecture were considered, including the number of hidden layers, the number of hidden neurons and the choice of activation functions. Ultimately, a model with three hidden layers of sizes 256, 128, and 32 neurons was selected. This model included dropout as a regularization technique to prevent overfitting.

Moving forward, we realized that while our various model fits had reasonable accuracy (around 86%), the recall values were very low. While we hoped to predict all diabetics and non-diabetics correctly, our primary goal is to identify those who may be diabetic so they can get tested. For this reason, we adjusted the probability threshold needed to classify an individual as diabetic and accepted some loss of overall accuracy in favor of higher recall. We tested a range of cutoff values for each model and chose the best cutoff value for each model based on a balance of accuracy and recall.

We also looked at the models' performance on certain groups of people in order to identify any bias or unfairness. For instance, we recalculated these performance metrics separately for males and females, those who had completed some college education and those who hadn't, those who earned at most $35,000 per year and those who earned more than $35,000 per year, those who had healthcare and those who did not, and those under and over age 65.

To compare the six models' predictive abilities, we computed the training error rate, test error rate, training accuracy, test accuracy, test precision, test recall, and test specificity. We also considered the

interpretability of each model. The usefulness of our final model depends in part on physicians being willing to trust our model and incorporate it into their clinical decisions, and non-data scientists will be more likely to use a tool they can understand. Logistic regression and LASSO regression are easily interpretable; we can look at the model coefficients to understand the relationship between various risk factors and diabetes, and we know exactly why the model predicts a given person to be non-diabetic or diabetic. LDA, QDA, boosting, and neural networks make the relationship between these risk factors and predictions more complicated. We used Shapley to gain insight into feature importance for the logistic, LASSO, LDA, QDA, and boosted models. In the feed-forward neural network model, local interpretable model-agnostic explanation was used to approximate a local understanding of the model's individual predictions. By perturbing one observation, a respondent who was considered low income, we were able to gain insight about the model's decision-making process when considering low-income respondents.

In selecting the best overall model, our primary considerations were overall accuracy, ability to identify true diabetics, and interpretability.

3. Results

Table 1, which can be found in the Appendix, displays certain characteristics of the 253,680 survey respondents broken down by self-reported status as non-diabetic/pre-diabetic or diabetic. Some notable differences between the two classes are that a much larger percentage of diabetics are at least 60 years old, have high blood pressure and cholesterol, and make no more than $35,000 per year compared to non- or pre-diabetics.

Table 2 below displays the initial test set metrics for each of the six models we fit with the default probability cutoff of 0.50. Here we see that all six models have reasonable accuracy, but the recall is quite low (apart from QDA). Since our primary goal is to correctly predict diabetics, this led us to lower the probability threshold needed to predict an individual to be diabetic.

Overall Metrics for Various Models

At the default cutoff of 0.5

Model	Accuracy	Precision	Recall	Specificity
LDA	0.86	0.49	0.20	0.97
QDA	0.77	0.31	0.55	0.80
Boosting	0.87	0.59	0.13	0.99
Logistic	0.86	0.58	0.06	0.99
Lasso	0.86	0.52	0.15	0.98
Neural Network	0.86	0.70	0.03	1.00

Table 2: Overall metrics for each of the 6 models at the default 0.5 cutoff

We then explored a range of lower probability cutoff points for each of the six models. We chose to sacrifice some overall accuracy for large gains in recall, so we primarily focused on balancing these two metrics. We immediately eliminated logistic regression and neural networks because these models still had low recall values, and we eliminated QDA because it had low overall accuracy. The remaining models (LDA, LASSO, and boosting) all had similar test metrics. Because LASSO is by far the most interpretable of these models, we selected LASSO as the best overall model. Table 3 displays the various cutoff values we looked at for the LASSO model; these same metrics for the five other models can be found in the appendix. We ultimately selected a value of 0.15 instead of the original 0.5 because this new model had a high recall (0.75, as opposed to 0.15 in the initial model) with only a slightly lower overall accuracy (0.74, as opposed to the initial accuracy of 0.86).

Cutoff	Accuracy	Precision	Recall	Specificity
LASSO				
0.10	0.65	0.27	0.86	0.62
0.15	0.74	0.31	0.75	0.74
0.20	0.79	0.35	0.63	0.82
0.25	0.82	0.39	0.52	0.87
0.30	0.84	0.42	0.43	0.91
0.35	0.85	0.45	0.34	0.93
0.40	0.86	0.48	0.27	0.95
0.45	0.86	0.50	0.20	0.97
0.50	0.86	0.52	0.15	0.98

Table 3: Overall metrics for the LASSO model at different cutoffs

We also conducted a sub-group analysis to explore our model's performance on various groups of people. Table 4 displays the sub-group analysis results for the LASSO model; the sub-group analysis results for the other five models can be found in the appendix. Notably, the LASSO model has lower accuracy among low-income people relative to high-income people, less-educated people relative to highly educated people, and older people relative to younger people. However, the LASSO model also has higher recall for these groups. This is likely because diabetes is more prevalent in these groups (Centers for Disease Control and Prevention), making both true positives and false positives more common.

Model	Variable	Value	Precision	Recall	Accuracy	Fraction of Dataset
LASSO	Sex	Female	0.32	0.74	0.76	0.56
	Sex	Male	0.30	0.75	0.71	0.44
	Health Coverage	No Health Coverage	0.29	0.69	0.78	0.05
	Health Coverage	At least Some Health Coverage	0.31	0.75	0.74	0.95
	Income	<35000 USD p.a	0.33	0.84	0.62	0.33
	Income	>=35000 USD p.a	0.29	0.65	0.80	0.67
	Education	No College	0.32	0.82	0.64	0.30
	Education	At least Some College	0.31	0.69	0.78	0.70
	Age	<60 Years Old	0.30	0.64	0.84	0.51
	Age	>=60 Years Old	0.32	0.80	0.63	0.49

Table 4: Overall metrics for LASSO model within specified subgroups using optimal cutoff value (0.15).

We also used the Shapfast package in R to find the most important features in the LDA, QDA and Boosting models. For Logistic Regression and LASSO, we used the absolute value of coefficients as feature importance. The feature importance for the LASSO model is shown below in Figure 2, and the feature importance for the other 5 models can be found in the Appendix. Here we see that a Cholesterol check in the past 5 years, blood pressure, heavy alcohol consumption, cholesterol levels and general health status are the most important variables in this dataset when classifying an individual as diabetic or non-diabetic.

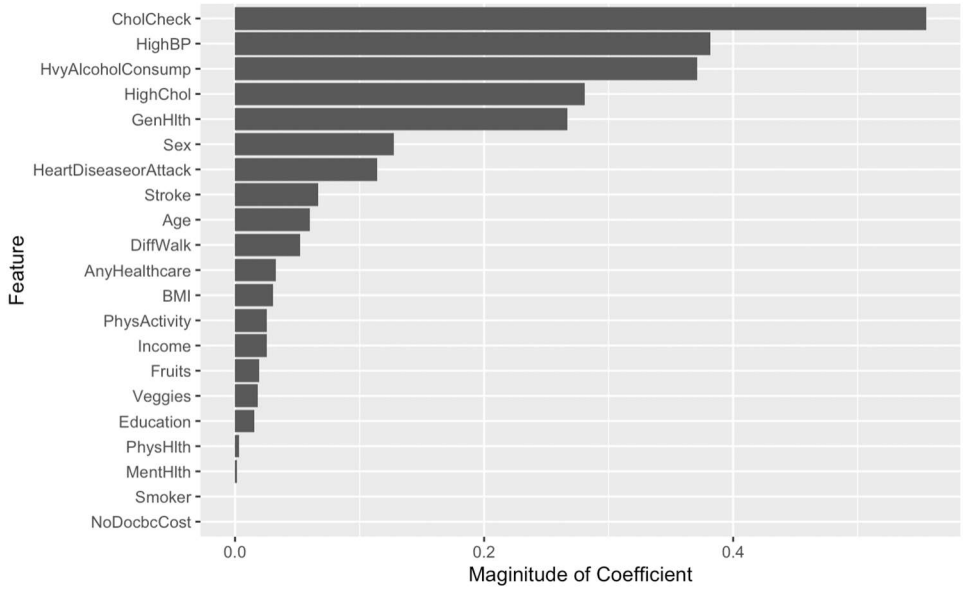


Figure 2: Feature importance for LASSO model.

4. Discussion

Our analysis began with our initial dataset which included 3 classes: non-diabetic, prediabetic and diabetic. Ideally, we could have taken advantage of all three available classes, but due to the imbalance of the dataset and the inaccuracies of self-reported survey data, we combined the non-diabetic and prediabetic classes. After fitting LDA, QDA, logistic, LASSO, neural networks, and boosted models with the default probability threshold of 0.5, we chose to lower this threshold in order to improve recall while accepting a reduction of overall accuracy since the main goal of our analysis was to correctly predict diabetics. False negatives are more costly than false positives here because it would be less problematic to predict someone as having a higher risk of diabetes when they don't, instead of telling someone that they have a lower risk of diabetes when they actually are at high risk of diabetes. We determined the LASSO model to be the best overall model given its interpretability and balance of accuracy (74%) with recall (75%) at a threshold of 0.15.

In further analysis with Shapley, we also identified the most important features for each of the six models in predicting diabetes. For the LASSO model, cholesterol check in the past 5 years, blood pressure, heavy alcohol consumption, cholesterol levels and general health status were most important when classifying an individual as diabetic or non-diabetic. This is applicable in real life when physicians assess a patient's characteristics that would most likely contribute to the onset of diabetes. This knowledge is important for physicians in determining who is most at risk for diabetes and therefore who should be tested. We also investigated the performance of our models within different subgroups as seen in Table 4 and determined that our model does perform differently within these groups. When using this model in the real world, we need to be mindful of these biases.

One of the main issues surrounding diabetes is the lack of awareness and lack of early detection of the disease. Early detection of the disease could result in quicker lifestyle changes and avoiding the extreme consequences of undiagnosed diabetes, such as limb loss, blindness, heart disease and renal disease. Having the ability to accurately predict those who are at a higher risk of diabetes will allow these lifestyle changes to be made earlier on in diabetic progression. This will in turn lead to an improved life of the individual and a decrease in prevalence of the disease, financial burden and extreme consequences of the disease if it is ignored and can progress.

Throughout our analysis, we did face some limitations. First, since we are relying on a dataset that has self-reported diabetes status as labels, we are biasing our model since people with low access to healthcare are likely to be unaware of their true diabetes status. Since we use these labels for training our model, we are likely to build a model that may underestimate diabetes risk in low-income population (especially since we also use income as a predictor for diabetes.) Second, our model does not achieve a very high precision. This means that we may falsely classify a lot of non-diabetic cases as diabetes. This is to some extent unavoidable given that we rely on a noisy label to build our model and try to maximize the recall at some expense of precision. Due to the imbalance of the original data set, we were also unable to create models that predicted the prediabetic class. Thus, we had to change our dataset and our original goal from correctly predicting prediabetics to predicting those who are diabetic.

Another limitation included the complexity of some of our models. For example, it would have been ideal to use 10-fold cross validation (CV) to find the best LASSO tuning parameter, but due to the complexity of the model and the computational burden, we were unable to do so and had to use 5-fold CV instead. A similar issue arose when fitting our boosting model. Using 1,000 trees would have been ideal in our model creation, but due to the computing power and the software being used, we had to use 500 trees instead. Lastly, for our neural networks model, a different standard was used for analysis. This process included a particular approach of creating training, testing and validation data splits, which was different compared

to the other fitted models. In essence, this is a limitation since we did not compare the same data in our neural network model compared to our other 5 models due to the complexity and structure of the neural network model, which may lead to varied results

Despite the limitations listed above, we were able to fit six models to predict diabetes status that confirmed many of the previously known risk factors for this disease. Our models generally agree on important risk factors such as high blood pressure, heavy alcohol consumption, and high cholesterol. With our selected final model, the LASSO model, we were able to obtain an interpretable model with reasonable accuracy and recall, and we identified the potential biases in our model's performance on various sub-groups, which is important if we plan to use it in a clinical decision-making setting.

Each group member's contributions:

We all contributed **equally** to the project:

Rindala contributed to writing the code, working on the presentation, writing the methods section, creating some tables, editing and reviewing the paper, and recording a section of the final presentation.

Lauren contributed to the code, writing the results section, writing some of the methods section, editing the paper for clarity, and recording a section of the final presentation.

Daniel contributed to code and writing relevant to the neural network, writing the introduction, revisions of paper, presentation slides, and recording a section of the final presentation.

Willow contributed to writing some of the results section, writing the discussion section, creating some tables and figures, and recording a section of the final presentation.

Bhawesh contributed to analyzing the model performance across subgroups, creating feature importance for models, creating tables and figures, writing part of the discussion and result sections, and recording a section of the final presentation.

References

American Diabetes Association. (2018, March 22). *Economic costs of diabetes in the U.S. in 2017*. American Diabetes Association. Retrieved from https://diabetesjournals.org/care/article/41/5/917/36518/Economic-Costs-of-Diabetes-in-the-U-S-in-2017

Brownlee, J. (2020, January 14). *A gentle introduction to imbalanced classification*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/what-is-imbalanced-classification/

Centers for Disease Control and Prevention. (2022, January 18). *National Diabetes Statistics Report*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html

Centers for Disease Control and Prevention. (2021, December 29). *Risk Factors for Diabetes-Related Complications*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/diabetes/basics/risk-factors.html

Centers for Disease Control and Prevention. (2021, December 21). *Prediabetes - your chance to prevent type 2 diabetes*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/diabetes/basics/prediabetes.html

Daily Sentinel. (2022, April 5). *Report: One third of adults are prediabetic, most don't know it.* Retrieved from https://romesentinel.com/stories/one-third-of-adults-are-prediabetic-most-dont-know-it,132221

Hill, J., Nielsen, M., & Fox, M. H. (2013). *Understanding the social factors that contribute to diabetes: A means to informing health care and social policies for the chronically ill*. The Permanente journal. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3662286/

Appendix

Table 1:

	Non-Diabetic or Prediabetic (N=218334)	Diabetic (N=35346)	Total (N=253680)
Sex Assigned at Birth			
Female	123563 (56.6%)	18411 (52.1%)	141974 (56.0%)
Male	94771 (43.4%)	16935 (47.9%)	111706 (44.0%)
Age			
Less Than 60 Years Old	120064 (55.0%)	11302 (32.0%)	131366 (51.8%)
At Least 60 Years Old	98270 (45.0%)	24044 (68.0%)	122314 (48.2%)
Blood Pressure			
No High BP	136109 (62.3%)	8742 (24.7%)	144851 (57.1%)
High BP	82225 (37.7%)	26604 (75.3%)	108829 (42.9%)
Cholesterol Level			
No High Cholesterol	134429 (61.6%)	11660 (33.0%)	146089 (57.6%)
High Cholesterol	83905 (38.4%)	23686 (67.0%)	107591 (42.4%)
Cholesterol Check in past 5 Years			
No	9229 (4.2%)	241 (0.7%)	9470 (3.7%)
Yes	209105 (95.8%)	35105 (99.3%)	244210 (96.3%)
BMI			
Mean (SD)	27.8 (6.29)	31.9 (7.36)	28.4 (6.61)
Median [Min, Max]	27.0 [12.0, 98.0]	31.0 [13.0, 98.0]	27.0 [12.0, 98.0]
Smoked At Least 100 Cigs in Lifetime			
No	124228 (56.9%)	17029 (48.2%)	141257 (55.7%)
Yes	94106 (43.1%)	18317 (51.8%)	112423 (44.3%)
Ever Told Had a Stroke			
No	211310 (96.8%)	32078 (90.8%)	243388 (95.9%)
Yes	7024 (3.2%)	3268 (9.2%)	10292 (4.1%)
CHD or Myocardial Infarction			
No	202319 (92.7%)	27468 (77.7%)	229787 (90.6%)
Yes	16015 (7.3%)	7878 (22.3%)	23893 (9.4%)
Physical Activity in Past 30 Days			
No	48701 (22.3%)	13059 (36.9%)	61760 (24.3%)
Yes	169633 (77.7%)	22287 (63.1%)	191920 (75.7%)
Consume At Least 1 Fruit per Day			
No	78129 (35.8%)	14653 (41.5%)	92782 (36.6%)
Yes	140205 (64.2%)	20693 (58.5%)	160898 (63.4%)
Consume At Least 1 Vegetable per Day			
No	39229 (18.0%)	8610 (24.4%)	47839 (18.9%)
Yes	179105 (82.0%)	26736 (75.6%)	205841 (81.1%)

Heavy Drinker			
No	204910 (93.9%)	34514 (97.6%)	239424 (94.4%)
Yes	13424 (6.1%)	832 (2.4%)	14256 (5.6%)
Have Any Kind of Health Care Coverage			
No	10995 (5.0%)	1422 (4.0%)	12417 (4.9%)
Yes	207339 (95.0%)	33924 (96.0%)	241263 (95.1%)
Could Not See a Doctor in Past 12 Months Due To Cost			
No	200722 (91.9%)	31604 (89.4%)	232326 (91.6%)
Yes	17612 (8.1%)	3742 (10.6%)	21354 (8.4%)
General Health Status			
Excellent	44159 (20.2%)	1140 (3.2%)	45299 (17.9%)
Very Good	82703 (37.9%)	6381 (18.1%)	89084 (35.1%)
Good	62189 (28.5%)	13457 (38.1%)	75646 (29.8%)
Fair	21780 (10.0%)	9790 (27.7%)	31570 (12.4%)
Poor	7503 (3.4%)	4578 (13.0%)	12081 (4.8%)
Nb Days Mental Health Not Good in Past 30 Days			
Mean (SD)	2.98 (7.11)	4.46 (8.95)	3.18 (7.41)
Median [Min, Max]	0 [0, 30.0]	0 [0, 30.0]	0 [0, 30.0]
Nb Days Physical Health Not Good in Past 30 Days			
Mean (SD)	3.64 (8.06)	7.95 (11.3)	4.24 (8.72)
Median [Min, Max]	0 [0, 30.0]	1.00 [0, 30.0]	0 [0, 30.0]
Serious Difficulty Walking or Climbing Stairs			
No	188780 (86.5%)	22225 (62.9%)	211005 (83.2%)
Yes	29554 (13.5%)	13121 (37.1%)	42675 (16.8%)
Highest Education Level			
Never Attended School	127 (0.1%)	47 (0.1%)	174 (0.1%)
Grades 1 Through 8 (Elementary)	2860 (1.3%)	1183 (3.3%)	4043 (1.6%)
Grades 9 Through 11 (Some High School)	7182 (3.3%)	2296 (6.5%)	9478 (3.7%)
Grade 12 or GED (High School Graduate)	51684 (23.7%)	11066 (31.3%)	62750 (24.7%)
1 to 3 Years of College (Some College or Technical School)	59556 (27.3%)	10354 (29.3%)	69910 (27.6%)
4 or More Years of College (College Graduate)	96925 (44.4%)	10400 (29.4%)	107325 (42.3%)
Annual Income			
At Most $35,000	66011 (30.2%)	17595 (49.8%)	83606 (33.0%)
More Than $35,000	152323 (69.8%)	17751 (50.2%)	170074 (67.0%)

Table 1: Summary statistics for survey respondents.

Different Cutoff Metrics for Each Model:

Cutoff	Accuracy	Precision	Recall	Specificity
LASSO				
0.10	0.65	0.27	0.86	0.62
0.15	0.74	0.31	0.75	0.74
0.20	0.79	0.35	0.63	0.82
0.25	0.82	0.39	0.52	0.87
0.30	0.84	0.42	0.43	0.91
0.35	0.85	0.45	0.34	0.93
0.40	0.86	0.48	0.27	0.95
0.45	0.86	0.50	0.20	0.97
0.50	0.86	0.52	0.15	0.98
LDA				
0.10	0.67	0.27	0.83	0.65
0.15	0.75	0.32	0.72	0.76
0.20	0.80	0.36	0.60	0.83
0.25	0.82	0.39	0.51	0.87
0.30	0.84	0.41	0.43	0.90
0.35	0.85	0.44	0.36	0.93
0.40	0.85	0.46	0.30	0.94
0.45	0.86	0.48	0.24	0.96
0.50	0.86	0.49	0.20	0.97

QDA				
0.10	0.66	0.26	0.82	0.63
0.15	0.69	0.28	0.76	0.68
0.20	0.71	0.29	0.72	0.71
0.25	0.73	0.29	0.68	0.74
0.30	0.74	0.30	0.64	0.76
0.35	0.75	0.30	0.61	0.77
0.40	0.76	0.30	0.59	0.78
0.45	0.76	0.31	0.57	0.79
0.50	0.77	0.31	0.55	0.80
Boosting				
0.10	0.64	0.26	0.87	0.61
0.15	0.74	0.31	0.76	0.73
0.20	0.79	0.36	0.64	0.81
0.25	0.82	0.40	0.54	0.87
0.30	0.84	0.43	0.43	0.91
0.35	0.86	0.47	0.34	0.94
0.40	0.86	0.51	0.25	0.96
0.45	0.87	0.54	0.18	0.98
0.50	0.87	0.59	0.13	0.99

Neural Networks				
0.10	0.60	0.92	0.25	0.54
0.15	0.71	0.81	0.30	0.69
0.20	0.77	0.72	0.34	0.78
0.25	0.80	0.65	0.37	0.82
0.30	0.83	0.52	0.42	0.88
0.35	0.86	0.31	0.50	0.95
0.40	0.86	0.17	0.56	0.98
0.45	0.86	0.08	0.61	0.99
0.50	0.86	0.03	0.70	1.00
Logistic				
0.10	0.86	0.53	0.13	0.98
0.15	0.86	0.54	0.12	0.98
0.20	0.86	0.54	0.11	0.99
0.25	0.86	0.55	0.10	0.99
0.30	0.87	0.56	0.09	0.99
0.35	0.87	0.57	0.09	0.99
0.40	0.87	0.58	0.08	0.99
0.45	0.87	0.59	0.07	0.99
0.50	0.86	0.58	0.06	0.99

Table 3: Overall metrics for each model at different cutoffs

Overall metrics for the LDA, QDA, Logistic and Boosting models within specified subgroups using the default cutoff of 0.5

Model	Variable	Value	Precision	Recall	Accuracy	Fraction of Dataset
LDA	Sex	Female	0.50	0.21	0.87	0.56
	Sex	Male	0.48	0.18	0.85	0.44
	Health Coverage	No Health Coverage	0.52	0.15	0.89	0.05
	Health Coverage	At least Some Health Coverage	0.49	0.20	0.86	0.95
	Income	<35000 USD p.a	0.49	0.29	0.79	0.33
	Income	>=35000 USD p.a	0.49	0.10	0.90	0.67
	Education	No College	0.48	0.25	0.81	0.30
	Education	At least Some College	0.51	0.16	0.88	0.70
	Age	<60 Years Old	0.47	0.17	0.91	0.51
	Age	>=60 Years Old	0.50	0.21	0.81	0.49

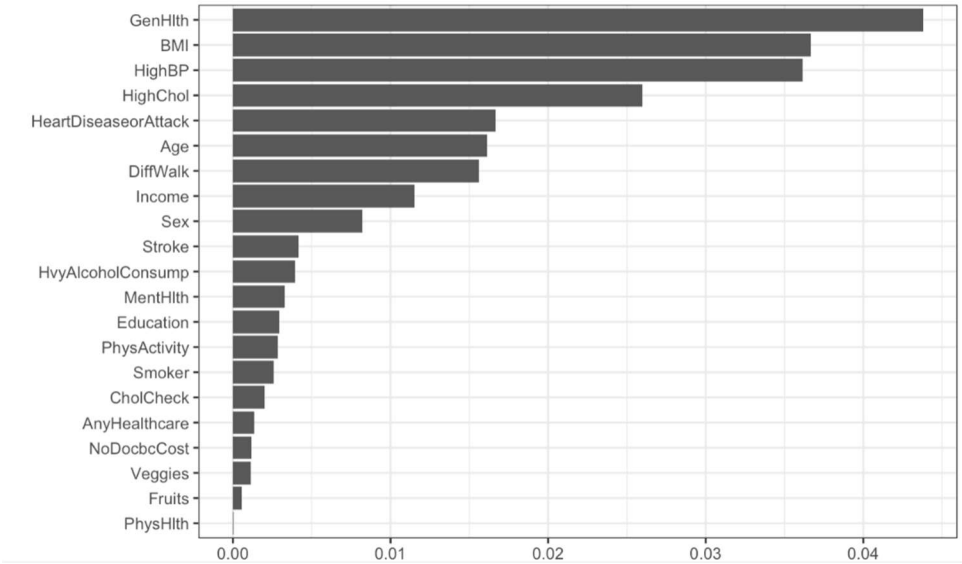
Model	Variable	Value	Precision	Recall	Accuracy	Fraction of Dataset
QDA	Sex	Female	0.31	0.58	0.78	0.56
	Sex	Male	0.31	0.52	0.76	0.44
	Health Coverage	No Health Coverage	0.35	0.39	0.85	0.05
	Health Coverage	At least Some Health Coverage	0.31	0.56	0.76	0.95
	Income	<35000 USD p.a	0.34	0.69	0.66	0.33
	Income	>=35000 USD p.a	0.27	0.42	0.82	0.67
	Education	No College	0.33	0.64	0.69	0.30
	Education	At least Some College	0.29	0.49	0.80	0.70
	Age	<60 Years Old	0.27	0.48	0.85	0.51
	Age	>=60 Years Old	0.33	0.59	0.68	0.49

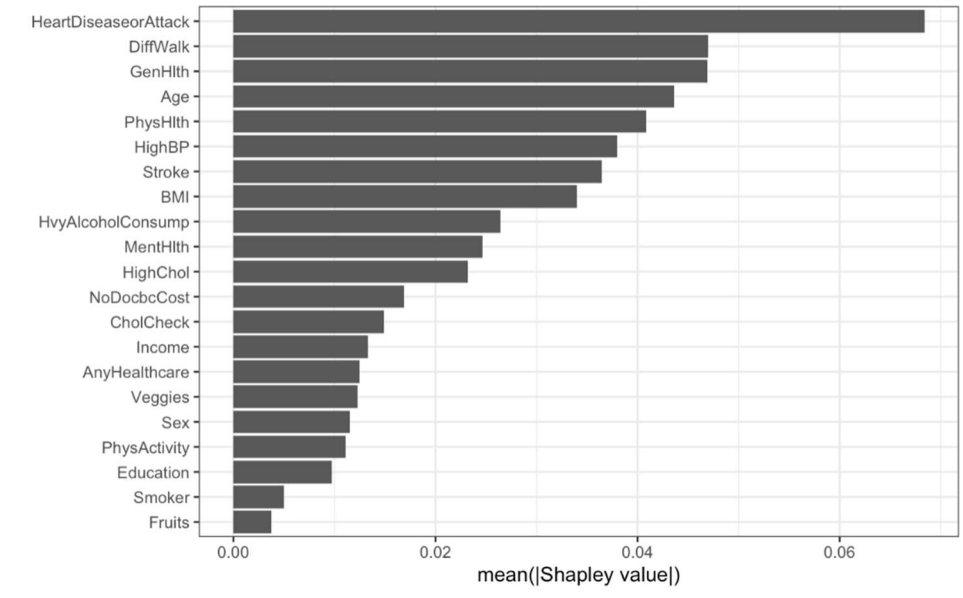
Model	Variable	Value	Precision	Recall	Accuracy	Fraction of Dataset
LR	Sex	Female	0.59	0.06	0.87	0.56
	Sex	Male	0.57	0.06	0.85	0.44
	Health Coverage	No Health Coverage	0.52	0.03	0.89	0.05
	Health Coverage	At least Some Health Coverage	0.59	0.06	0.86	0.95
	Income	<35000 USD p.a	0.59	0.10	0.80	0.33
	Income	>=35000 USD p.a	0.56	0.03	0.90	0.67
	Education	No College	0.58	0.09	0.82	0.30
	Education	At least Some College	0.59	0.05	0.89	0.70
	Age	<60 Years Old	0.52	0.05	0.92	0.51
	Age	>=60 Years Old	0.61	0.07	0.81	0.49

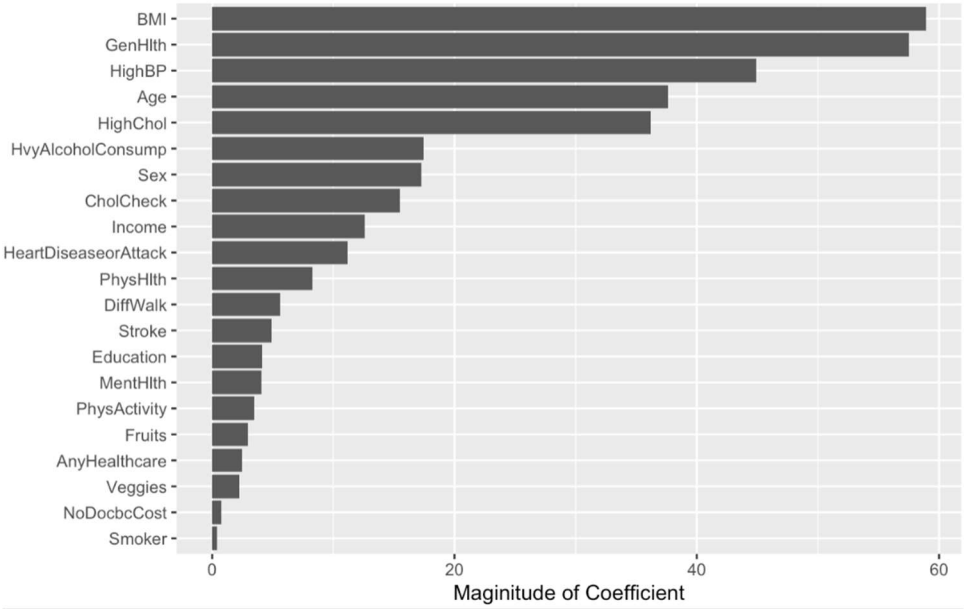
Model	Variable	Value	Precision	Recall	Accuracy	Fraction of Dataset
XGBoost	Sex	Female	0.59	0.13	0.88	0.56
	Sex	Male	0.58	0.12	0.86	0.44
	Health Coverage	No Health Coverage	0.62	0.08	0.89	0.05
	Health Coverage	At least Some Health Coverage	0.59	0.13	0.87	0.95
	Income	<35000 USD p.a	0.58	0.18	0.80	0.33
	Income	>=35000 USD p.a	0.59	0.07	0.90	0.67
	Education	No College	0.56	0.15	0.82	0.30
	Education	At least Some College	0.62	0.10	0.89	0.70
	Age	<60 Years Old	0.62	0.07	0.92	0.51
	Age	>=60 Years Old	0.58	0.15	0.81	0.49

Table 4: Overall metrics for the LDA, QDA, Logistic and Boosting models within specified subgroups using default cutoff of 0.5

Feature Importance:







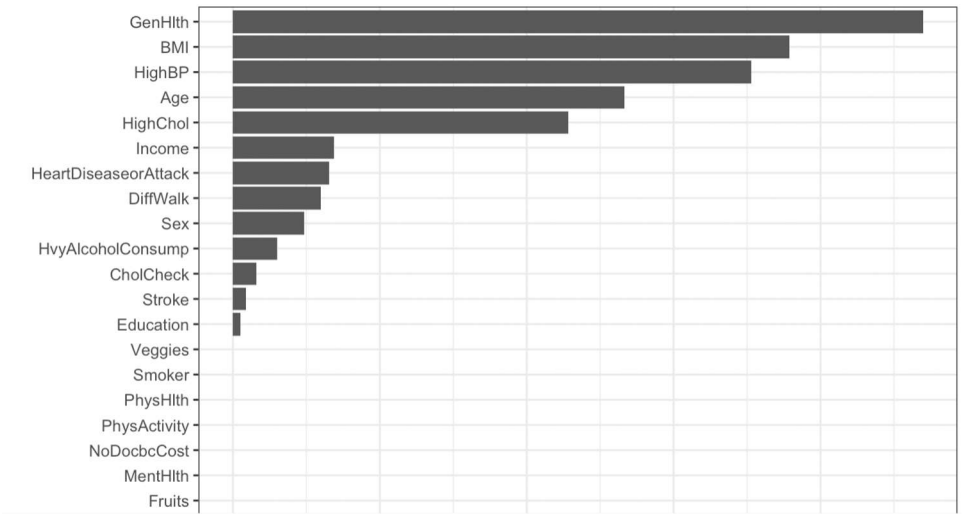


Figure 2: Feature importance for LDA, QDA, Logistic and Boosting models

GitHub Repository:

https://github.com/rindalafayyad17/263-Project.git