Determining what Lifestyle Variables are Valuable in Predicting the Odds an Individual has Diabetes or Prediabetes

Reymond Ramirez
STA4504
Project #2

Table of Contents

I. Abstract

The dataset that was used in this research was based on Alex Teboul's version of the data originally collected by the CDC in 2015 and was uploaded to the website Kaggle to have a binary response variable where 1 was if an individual had diabetes/prediabetes and 0 if they did not. The data, which contains the responses of 253,680 Americans used factors such as age, BMI score, and physical health to determine if how influential some parts of an individual's lifestyle have on making them eventually become diabetic/prediabetic. By building this logistic regression model, we can determine the probability of an individual having diabetes/prediabetes through the lifestyle variables. After testing to determine if some variables were better as ordinal or not and conducting a model selection process based on AIC values, the final model had consisted of 21 predictors, 4 of which were numerical and the rest categorical. Additionally, the model was found to have an accuracy of 72.65%, an AUC value of 0.8246, and sensitivity and specificity values of 7.74% and 71.82%, respectively. These are valuable in showcasing the reliability in the model as it highlights its performance and how highly accurate it is in calculating the odds that an individual will have diabetes/prediabetes through these predictors.

II. Introduction and Research Question

According to the CDC, around 11.6% of the US population has diabetes and about 22.8% of adults with diabetes are undiagnosed (Centers for Disease Control and Prevention). As a result, many Americans, especially those in lower economic classes have been shown to have an increased risk in diabetic cases due to a variety of socioeconomic factors that cause them unable to afford and access healthcare (Liu et al.). It is important to note however, the distinct differences between an individual who has type 1 and type 2 diabetes as type 1 is not preventable and happens as a result of the immune system attacking and destroying insulin-producing cells within the pancreas while type 2 is a result of the pancreas making less insulin than used to, and the body becoming resistant to insulin and could be prevented through lifestyle changes such as a change to a healthier diet. For individuals who are in danger of having type 2 diabetes, they are known to be prediabetic and are highly encouraged to lead a healthier lifestyle through exercise and diet.

The dataset that was obtained from the online data platform, Kaggle, transformed a dataset created by the Behavioral Risk Factor Surveillance System (BRFSS) in 2015 in which 21 variables were used to predict if an individual has diabetes/prediabetes (1 if yes, 0 if no) (Teboul). While the dataset is from 10 years ago, the same predictors can be used to help predict the odds of a person becoming diabetic or prediabetic. By using this dataset, we are able to build a logistic regression model in which we can use to answer the question which health factors are significant in contributing diabetes and prediabetes to Americans. Using this model, we can educate Americans, especially those who are more prone to becoming diabetic, on which health and lifestyle factors are significantly causing them to eventually become diabetic. This research could help in educating and influencing these individuals on the effects these factors can have on their life and emphasize the importance of maintaining their health.

III. Data Description

The dataset that was uploaded to Kaggle had originally originated as a dataset from 2015 from the BRFSS. The dataset from Kaggle contains a total of 21 variables, most of which are categorical and are used to help in predicting the response variable, if an individual has diabetes/prediabetes (1 if yes, 0 if no). Some of the variables included in the dataset include variables such as BMI, sex (1 if male, 0 if female), and if they have high cholesterol (1 if yes, 0 if no). By using these variables, we can construct a logistic regression model using the link function, logit, to create a regression model in which we can determine the log odds of a person having diabetes/prediabetes based on an individual's health and lifestyle and use it to calculate the odds of them having diabetes/prediabetes.

With a total of 21 variables, there might be a potential issue with collinearity between certain variables as some health issues that person has, might cause more to occur. For instance, if an individual has high cholesterol, they might also have coronary heart disease or myocardial infarction as they are associated with high cholesterol levels. As a result, testing must be done between certain variables to prevent collinearity from making the model unreliable. Furthermore, the inclusion of first order interaction terms might be considered to keep the model as simple as possible while also acknowledging how some factors interact with each other. Interaction terms such as BMIHighChol and AnyHealthcareNoDocbcCost might be included due to the close relations that these variables have with each other which could influence the odds of an individual having diabetes or prediabetes.

Using these different variables and interaction terms we can create a logistic model to calculate the log odds of a person having diabetes/prediabetes. With these different variables, we can then use backwards selection to then determine what variables and interaction terms are

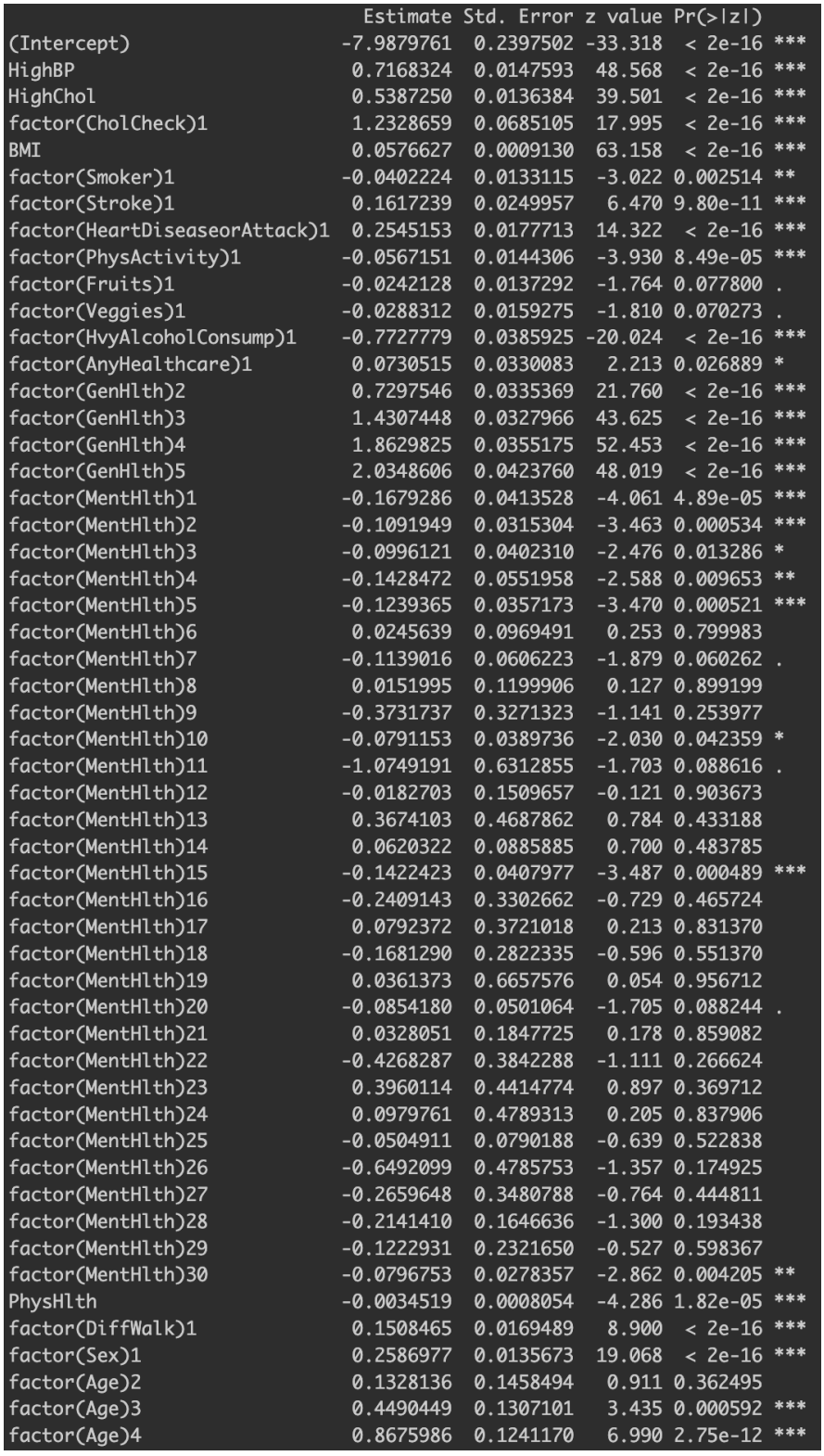
significant to calculate the log odds and odds. These significant predictors will provide meaningful insight into what factors are significantly causing diabetes and prediabetes to rise amongst Americans.

IV. Model Selection

To create the final logistic model that will be needed to predict and interpret the odds of an individual having diabetes/prediabetes, testing must first be done to the predictors MentHlth, PhysHlth, Age, and Income to determine if the variables should be ordinal instead. To do this, a likelihood-ratio test was conducted with the initial full model where the predictors were all factors and another model in which those predictors were not. The results of the test indicated that the complex (full) model was a better fit and so each predictor was individually tested to determine if there was possibly at least one predictor that could not be treated as ordinal. After conducting these tests, PhysHlth was the only predictor that did not have to be treated as ordinal, making a logistic model with a current AIC value of 160893.3.

To finalize the process, the final model was created using the `stepAIC()` function to further simplify the model and potentially lower the AIC value of the logistic model. After the program had completed the model selection, the predictor NoDocbcCost was removed, and the AIC value of the model had essentially no difference from the previous model as it had a new value of 160892. While the model selection process only removed one variable, it did indicate that an individual's inability to visit a doctor due to costs in the last 12 months was not a significant predictor in determining if an individual has prediabetes/diabetes. A summary of the model and its coefficients are listed below:

Figure 1: Screenshot of the Final Model and its Predictor Values



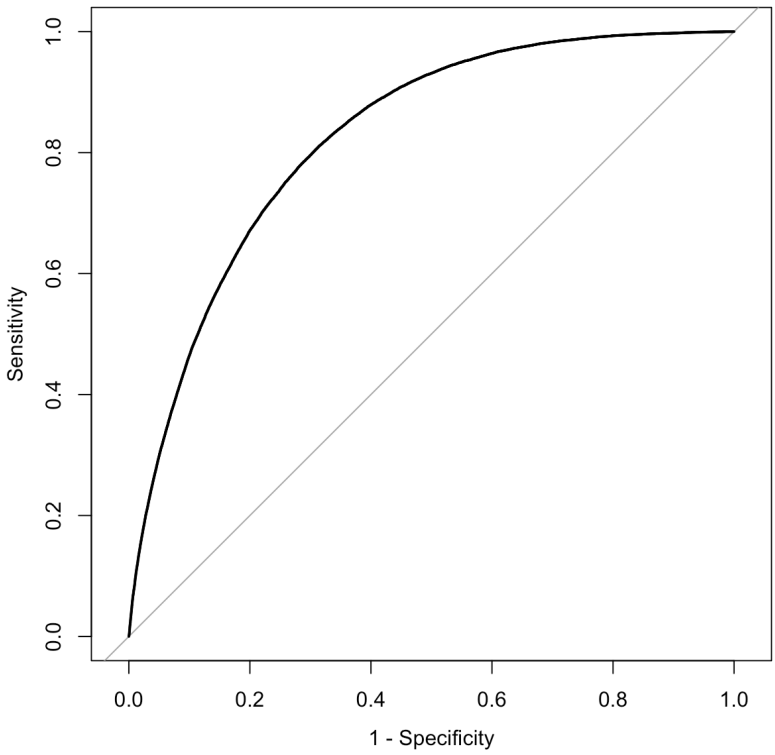
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.9879761	0.2397502	-33.318	< 2e-16	***
HighBP	0.7168324	0.0147593	48.568	< 2e-16	***
HighChol	0.5387250	0.0136384	39.501	< 2e-16	***
factor(CholCheck)1	1.2328659	0.0685105	17.995	< 2e-16	***
BMI	0.0576627	0.0009130	63.158	< 2e-16	***
factor(Smoker)1	-0.0402224	0.0133115	-3.022	0.002514	**
factor(Stroke)1	0.1617239	0.0249957	6.470	9.80e-11	***
factor(HeartDiseaseorAttack)1	0.2545153	0.0177713	14.322	< 2e-16	***
factor(PhysActivity)1	-0.0567151	0.0144306	-3.930	8.49e-05	***
factor(Fruits)1	-0.0242128	0.0137292	-1.764	0.077800	.
factor(Veggies)1	-0.0288312	0.0159275	-1.810	0.070273	.
factor(HvyAlcoholConsump)1	-0.7727779	0.0385925	-20.024	< 2e-16	***
factor(AnyHealthcare)1	0.0730515	0.0330083	2.213	0.026889	*
factor(GenHlth)2	0.7297546	0.0335369	21.760	< 2e-16	***
factor(GenHlth)3	1.4307448	0.0327966	43.625	< 2e-16	***
factor(GenHlth)4	1.8629825	0.0355175	52.453	< 2e-16	***
factor(GenHlth)5	2.0348606	0.0423760	48.019	< 2e-16	***
factor(MentHlth)1	-0.1679286	0.0413528	-4.061	4.89e-05	***
factor(MentHlth)2	-0.1091949	0.0315304	-3.463	0.000534	***
factor(MentHlth)3	-0.0996121	0.0402310	-2.476	0.013286	*
factor(MentHlth)4	-0.1428472	0.0551958	-2.588	0.009653	**
factor(MentHlth)5	-0.1239365	0.0357173	-3.470	0.000521	***
factor(MentHlth)6	0.0245639	0.0969491	0.253	0.799983	
factor(MentHlth)7	-0.1139016	0.0606223	-1.879	0.060262	.
factor(MentHlth)8	0.0151995	0.1199906	0.127	0.899199	
factor(MentHlth)9	-0.3731737	0.3271323	-1.141	0.253977	
factor(MentHlth)10	-0.0791153	0.0389736	-2.030	0.042359	*
factor(MentHlth)11	-1.0749191	0.6312855	-1.703	0.088616	.
factor(MentHlth)12	-0.0182703	0.1509657	-0.121	0.903673	
factor(MentHlth)13	0.3674103	0.4687862	0.784	0.433188	
factor(MentHlth)14	0.0620322	0.0885885	0.700	0.483785	
factor(MentHlth)15	-0.1422423	0.0407977	-3.487	0.000489	***
factor(MentHlth)16	-0.2409143	0.3302662	-0.729	0.465724	
factor(MentHlth)17	0.0792372	0.3721018	0.213	0.831370	
factor(MentHlth)18	-0.1681290	0.2822335	-0.596	0.551370	
factor(MentHlth)19	0.0361373	0.6657576	0.054	0.956712	
factor(MentHlth)20	-0.0854180	0.0501064	-1.705	0.088244	.
factor(MentHlth)21	0.0328051	0.1847725	0.178	0.859082	
factor(MentHlth)22	-0.4268287	0.3842288	-1.111	0.266624	
factor(MentHlth)23	0.3960114	0.4414774	0.897	0.369712	
factor(MentHlth)24	0.0979761	0.4789313	0.205	0.837906	
factor(MentHlth)25	-0.0504911	0.0790188	-0.639	0.522838	
factor(MentHlth)26	-0.6492099	0.4785753	-1.357	0.174925	
factor(MentHlth)27	-0.2659648	0.3480788	-0.764	0.444811	
factor(MentHlth)28	-0.2141410	0.1646636	-1.300	0.193438	
factor(MentHlth)29	-0.1222931	0.2321650	-0.527	0.598367	
factor(MentHlth)30	-0.0796753	0.0278357	-2.862	0.004205	**
PhysHlth	-0.0034519	0.0008054	-4.286	1.82e-05	***
factor(DiffWalk)1	0.1508465	0.0169489	8.900	< 2e-16	***
factor(Sex)1	0.2586977	0.0135673	19.068	< 2e-16	***
factor(Age)2	0.1328136	0.1458494	0.911	0.362495	
factor(Age)3	0.4490449	0.1307101	3.435	0.000592	***
factor(Age)4	0.8675986	0.1241170	6.990	2.75e-12	***

```
factor(Age)5                1.1112912  0.1214071   9.153  < 2e-16 ***
factor(Age)6                1.3073836  0.1196940  10.923  < 2e-16 ***
factor(Age)7                1.5285453  0.1185158  12.897  < 2e-16 ***
factor(Age)8                1.6131228  0.1181403  13.654  < 2e-16 ***
factor(Age)9                1.8221417  0.1179155  15.453  < 2e-16 ***
factor(Age)10               1.9840815  0.1179496  16.821  < 2e-16 ***
factor(Age)11               2.0274158  0.1183442  17.132  < 2e-16 ***
factor(Age)12               1.9381198  0.1190177  16.284  < 2e-16 ***
factor(Age)13               1.7509865  0.1191887  14.691  < 2e-16 ***
factor(Education)2         -0.0326502  0.1953625  -0.167 0.867271
factor(Education)3         -0.1516157  0.1934478  -0.784 0.433184
factor(Education)4         -0.2086445  0.1920380  -1.086 0.277269
factor(Education)5         -0.1650256  0.1921215  -0.859 0.390360
factor(Education)6         -0.2532680  0.1922426  -1.317 0.187691
factor(Income)2            -0.0129110  0.0355302  -0.363 0.716320
factor(Income)3            -0.0399078  0.0342972  -1.164 0.244591
factor(Income)4            -0.0603400  0.0336464  -1.793 0.072916 .
factor(Income)5            -0.1446746  0.0332363  -4.353 1.34e-05 ***
factor(Income)6            -0.2273607  0.0327518  -6.942 3.87e-12 ***
factor(Income)7            -0.2489409  0.0331022  -7.520 5.46e-14 ***
factor(Income)8            -0.3903472  0.0327983 -11.901  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204847  on 253679  degrees of freedom
Residual deviance: 160744  on 253606  degrees of freedom
AIC: 160892

Number of Fisher Scoring iterations: 6
```

V. Model Reliability

To determine if the model is reliable in determining the odds of an individual that contains diabetes/prediabetes, different metrics were used to determine will be reliable to use. To determine the reliability of the model, a ROC plot was created and shown in Figure 1 with an AUC value of 0.8246. Since the AUC value is greater than 0.5, the model is shown to be better at predicting the odds of a person having diabetes/prediabetes than if we were to randomly guess.

Figure 2: ROC Plot of Final Logistic Model



The model also appeared to an accuracy of about 72.65% which is very well considering that the dataset contains 253,680 responses, causing a high number of accurate predictions. Furthermore, the sensitivity and specificity of the model was also calculated and was found to have values of 77.74% and 71.82%, respectively. Within the context of the dataset, the model is 77.74% accurate in identifying individuals who have diabetes/prediabetes and 71.82% accurate

of identifying those who do not. These percentages indicated that the model is highly reliable, making it useful to predict the likelihood that an individual has diabetes/prediabetes and also highlight how useful the predictors in the model are with influencing the values of the probabilities.

VI. Conclusion

While working with this dataset and trying to create an accurate and reliable model, I have learned many important things about constructing a general linear model and the process of doing so. Specifically, the amount of trial and error that must be done to construct and finalize a model that has important implications within health sciences. Though there were originally many methods and ideas that could be implemented in the model such as originally stating that interaction models would be implemented, it was best not to include them within the final model to keep it as simple as possible, especially considering the vast amounts of ordinal variables within the dataset. Nevertheless, the final model that was created is very useful in predicting the odds of an individual having diabetes/prediabetes.

Despite the model being accurate with its predictions, some limitations within the model in the dataset were present that could have prevented from an even better model to be formed. For instance, the dataset is originally from 2015 and is based on values that are 10 years old, which could cause some predictors that may have been influential back then, to possibly not be as impactful in present day compared to other variables and vice versa. However, there are some variables that despite are a decade old, are still pivotal in predicting diabetes/prediabetes for individuals in the current day such as age as there is evidence that older individuals are more likely to have type 2 diabetes (“Diabetes in Older People | National Institute on Aging”).

Additionally, other model selection processes could have been implemented to potentially create a simpler model that is still accurate though model selection.

Building this model, allows us to analyze the implications that these predictors have on a person and their health. This model reveals just how impactful certain lifestyle choices that some predictors have more than others such as how a person's BMI score causes the odds of having diabetes/prediabetes to increase by 1.059358, holding all other variables constant. We can use this information to educate Americans, especially those in lower classes and areas who are prone to health issues, to have a more fit and active lifestyle through changes such as increasing their water intake and an increase in exercise to better their chances of not becoming diabetic or prediabetic. Diabetes is very prevalent issue within the United States, and with the help of this model, we can show Americans how their lifestyle can greatly impact their future.

Works Cited

Centers for Disease Control and Prevention. “National Diabetes Statistics Report.” *Centers for Disease Control and Prevention*, https://www.cdc.gov/diabetes/php/data-research/index.html. Accessed 31 March 2025.

“Diabetes in Older People | National Institute on Aging.” *National Institute on Aging*, 10 April 2024, https://www.nia.nih.gov/health/diabetes/diabetes-older-people. Accessed 16 April 2025.

Liu, Ce, et al. “Diabetes risk among US adults with different socioeconomic status and behavioral lifestyles: evidence from the National Health and Nutrition Examination Survey.” *Frontiers in public health*, vol. 11, no. 1197947, 2023. *National Library of Medicine*, https://pmc.ncbi.nlm.nih.gov/articles/PMC10477368/#abstract1. Accessed 1 April 2025.

Teboul, Alex. *Diabetes Health Indicators Dataset*. Dataset. 2022. *Kaggle*, https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. Accessed 1 April 2025.