Predicting Heart Disease

Raghuram Palaniappan
University of California, Davis
Faculty of Undergraduate Studies
rpalani@ucdavis.edu

Abstract

Heart Disease is extremely prevalent in the United States, being the number 1 cause of death. The aim of this study is to pick out the strongest predictors to Heart Disease. As a result, a good model is sought after to identify these predictors. Data analysis will be carried out on the model to ensure that it is a good fit.

1 Introduction

Every 36 seconds one person dies from Heart Disease in the United States. It is the leading cause of death in the country and 1 in every 4 deaths are that of Heart Disease [1]. However, what if we could detect the factors that lead to Heart Disease? This study uses Linear Modeling and Logistic Modeling to discover the leading predictors that influence Heart Disease.

2 Data and Methodology

2.1 Data

The data in this project comes from a Kaggle Dataset. The dataset has 21 predictors and 1 explanatory variable.

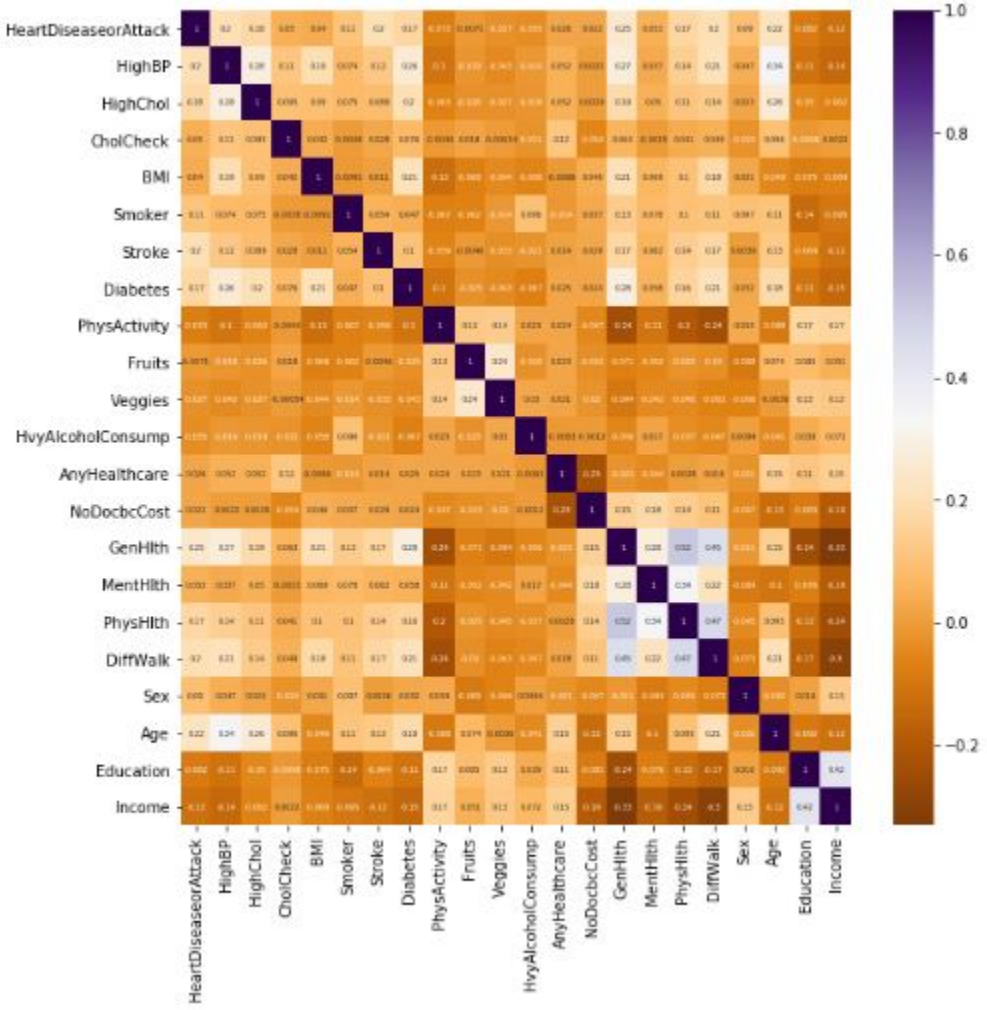
Variable	Description	Data Type
HeartDiseaseorAttack	If the person has heart disease	Binary
HighBP	If the person has high blood pressure	Binary
HighChol	If the person has high cholestrol	Binary
CholCheck	If the person checks cholestrol	Binary
BMI	Person's Body Mass Index	Continous
Smoker	If the person is a smoker	Binary
Stroke	Has the person had a stroke	Binary
Diabetes	If the person is diabetic	Binary
PhysActivity	If the person is physically active	Binary
Fruits	If the person eats fruits	Binary
Veggies	If the person eats vegetables	Binary
HvyAlcoholConsump	If the person consumes alchohol heavy	Binary
AnyHealthcare	If the person has access to healthcare	Binary
NoDocbcCost	If the persons cost of healthcare is free	Binary
GenHlth	Score of General Health	Continous
MentHlth	Score of Mental Health	Continous
PhysHlth	Score of Physical Health	Continous
DiffWalk	Person's ability to walk	Binary
Sex	Person's Sex at Birth	Binary
Age	Person's age in years	Continous
Education	Person's level of education	Continous
Income	Person's level of income	Continous

2.2 Methodology

We start by trying to gain a better understanding of how the variable relates to Heart Disease and each other through exploratory visualizations. We consider the summary statistics of the variables and the correlations they may hold. To analyze the relationship between our predictors and our explanatory variable, the data is modeled into Linear and Logistic Regressions. Then the data is trained and tested through Machine learning and used to test the accuracy of the models created. The analysis is described formally in greater detail in the following section.

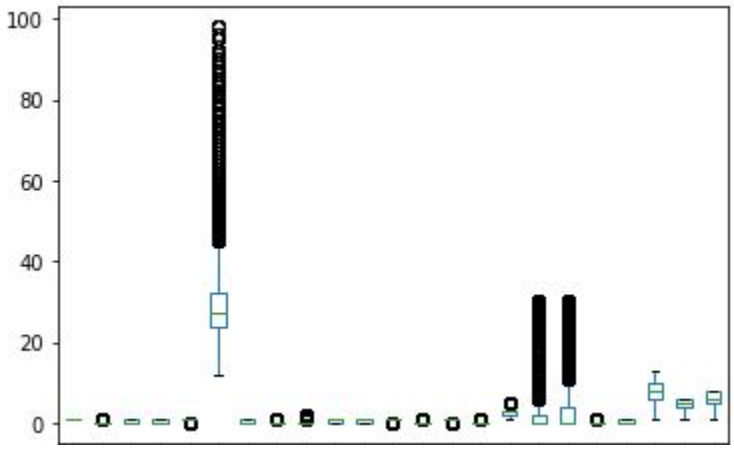
3 Data Visualization

3.1 Correlation Heatmap



The heatmap shows the correlations of the predictors with each other and the predictor variable of Heart Disease. The predictors do not have strong correlations with each other proving that our model will not have multicollinearity which would violate assumptions of our Linear and Logistic Models.

3.2 Boxplot



The boxplot shows that there are outliers and duplicates in the data that need to be removed and scaled so that the data is ready to be used to model.

4 Results

4.1 Linear Regression

In this model, Heart Disease is regressed on all available predictor variable. The Durbin-Watson test is passed with a score near 2 meaning no issues of multicollinearity, however the Jarque Bera test is extremely high meaning the data is not distributed correctly. Along with this the R-squared of the Linear model is extremely low at 0.145. Due to the failed assumption and low R-squared the Linear Model is not a good fit for our data.

OLS Regression Results

Dep. Variable:	HeartDiseaseorAttack	**R-squared:**	0.145
Model:	OLS	**Adj. R-squared:**	0.145
Method:	Least Squares	**F-statistic:**	1862.
Date:	Sat, 07 May 2022	**Prob (F-statistic):**	0.00
Time:	21:25:43	**Log-Likelihood:**	-34562.
No. Observations:	229781	**AIC:**	6.917e+04
Df Residuals:	229759	**BIC:**	6.940e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1648	0.006	-26.874	0.000	-0.177	-0.153
HighBP	0.0345	0.001	25.527	0.000	0.032	0.037
HighChol	0.0416	0.001	32.686	0.000	0.039	0.044
CholCheck	0.0149	0.003	4.901	0.000	0.009	0.021
BMI	-0.0012	9.24e-05	-12.636	0.000	-0.001	-0.001
Smoker	0.0242	0.001	19.878	0.000	0.022	0.027
Stroke	0.1853	0.003	63.452	0.000	0.180	0.191
Diabetes	0.0223	0.001	25.179	0.000	0.021	0.024
PhysActivity	0.0042	0.001	2.986	0.003	0.001	0.007
Fruits	0.0042	0.001	3.301	0.001	0.002	0.007
Veggies	0.0045	0.002	2.986	0.003	0.002	0.008
HvyAlcoholConsump	-0.0206	0.002	-8.282	0.000	-0.026	-0.016
AnyHealthcare	0.0072	0.003	2.625	0.009	0.002	0.013
NoDocbcCost	0.0085	0.002	3.948	0.000	0.004	0.013
GenHlth	0.0340	0.001	46.961	0.000	0.033	0.035
MentHlth	-0.0002	8.41e-05	-1.856	0.064	-0.000	8.77e-06
PhysHlth	0.0009	8.19e-05	10.617	0.000	0.001	0.001
DiffWalk	0.0475	0.002	25.634	0.000	0.044	0.051
Sex	0.0585	0.001	47.920	0.000	0.056	0.061
Age	0.0120	0.000	54.539	0.000	0.012	0.012
Education	0.0016	0.001	2.348	0.019	0.000	0.003
Income	-0.0035	0.000	-10.331	0.000	-0.004	-0.003

Omnibus:	89001.018	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	288431.430
Skew:	2.050	Prob(JB):	0.00
Kurtosis:	6.648	Cond. No.	349.

4.2 Logistic Regression

The Logistic Model follows the assumptions, but the only significant predictor variable is BMI. The BMI variable produces the highest Z-score with 3.85.

Logit Regression Results

Dep. Variable:	HeartDiseaseorAttack	**No. Observations:**	229781
Model:	Logit	**Df Residuals:**	229759
Method:	MLE	**Df Model:**	21
Date:	Sat, 07 May 2022	**Pseudo R-squ.:**	1.000
Time:	21:33:27	**Log-Likelihood:**	-0.00013574
converged:	False	**LL-Null:**	-76308.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-12.4934	704.308	-0.018	0.986	-1392.911	1367.924
HeartDiseaseorAttack	57.1148	6075.191	0.009	0.992	-1.19e+04	1.2e+04
HighBP	1.3466	222.898	0.006	0.995	-435.526	438.219
HighChol	0.8583	221.419	0.004	0.997	-433.115	434.831
CholCheck	-6.9655	644.805	-0.011	0.991	-1270.761	1256.830
BMI	-0.0005	14.210	-3.85e-05	1.000	-27.851	27.850
Smoker	-0.4268	198.700	-0.002	0.998	-389.871	389.017
Stroke	0.7098	431.141	0.002	0.999	-844.310	845.730
Diabetes	0.5261	150.903	0.003	0.997	-295.238	296.290
PhysActivity	-1.0718	222.711	-0.005	0.996	-437.578	435.434
Fruits	-0.1858	208.568	-0.001	0.999	-408.972	408.600
Veggies	-1.1191	259.274	-0.004	0.997	-509.287	507.049
HvyAlcoholConsump	0.3373	343.890	0.001	0.999	-673.675	674.350
AnyHealthcare	-3.1666	324.380	-0.010	0.992	-638.939	632.606
NoDocbcCost	-0.2398	252.732	-0.001	0.999	-495.586	495.106
GenHlth	-0.4026	98.397	-0.004	0.997	-193.257	192.452
MentHlth	-0.0220	10.130	-0.002	0.998	-19.877	19.833
PhysHlth	0.1751	10.919	0.016	0.987	-21.226	21.576
DiffWalk	-9.4140	6022.288	-0.002	0.999	-1.18e+04	1.18e+04
Sex	-0.8453	217.348	-0.004	0.997	-426.839	425.148
Age	-0.1019	46.110	-0.002	0.998	-90.475	90.272
Education	-0.6647	82.213	-0.008	0.994	-161.799	160.470

4.3 Training and Testing the Logistic Model

The Heart Disease Data was separated into a training and testing set. The trained data was used in the Logistic model to see if it accurately predict the test data set aside. The Logistic Model using the Machine Learning data responded with over 89.9 percent accuracy.

```
# Training and Testing Sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify= y, random_state=31)
# Scale
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.fit_transform(X_test)
```

In [202]:
```
# Testing Acccuracy of Model
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
classification_model = LogisticRegression(max_iter = 10000)
logmodel = classification_model.fit(X_train, y_train.values.ravel())
predictions = classification_model.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
accuracy
```

Out[202]: 0.8990360554431316

5 Conclusion

The aim of this study is to find the predictor that affected the explanatory variable of Heart Disease the greatest. The Linear Model did not end up meeting assumptions and failed to produce a high R-squared. The logistic model on the other hand highlighted that BMI was an extremely strong predictor. The Logistic Model also achieved a strong accuracy percentage of 89.9. Since the explanatory variable was a binary variable, the Logistic Model was effective. Further studies could be conducted on the variables that influence BMI since it is shown to be an influential predictor variable.

References

[1] Centers for Disease Control and Prevention. (2022, February 7). Heart disease facts. Centers for Disease Control and Prevention. Retrieved May 7, 2022, from https://www.cdc.gov/heartdisease/facts.html