Heart Disease Classification using Logistic Regression

William Lin, Kaiyu Yokoi, Ishani Parekh

Introduction

For this project we will create a logistic regression model in R to predict the occurrence of heart disease. The source of the data set is Kaggle, submitted by user Kamil Pytlak; the page is titled "Personal Key Indicators of Heart Disease." The original data set itself is the 2020 annual CDC survey data of 400,000 adults aged 18 and older related to their health status, with 279 predictors of heart disease. This data set was cleaned by the user and reduced to 319,795 observations with 18 of the most significant predictors.

Data Description:

Feature Name	Description
HeartDisease	Whether respondent ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
BMI	Body Mass Index (BMI)
Smoking	If respondent reported smoking at least 100 cigarettes in their entire life (5 packs = 100 cigarettes)
AlcoholDrinking	Respondents who are categorized as heavy drinkers (note that this is categorized as adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
Stroke	Was respondent ever told that they had a stroke
PhysicalHealth	How many days in the last 30 days did the respondent report physical illness and injury
MentalHealth	How many days during the past 30 days did respondent report their mental health to be good
DiffWalking	Whether the respondent reported having a

	serious difficulty walking or climbing stairs
Sex	Respondent reports whether they are male or female
AgeCategory	Respondents report what age category they are in. Categories include "18-24," "25-29," "30-34," "35-39," "40-44," "45-49," "50-54," "55-59," "60-64," "65-69," "70-74," "75-79," "80 or older,"
Race	Race/ethnicity reported by respondent. Categories include "White," "American Indian/Alaskan Native," "Asian," "Black," "Hispanic," "Other"
Diabetic	Whether the respondent was ever told they had diabetes. Categories include "No," "No, borderline diabetes," "Yes," "Yes (during pregnancy)"
PhysicalActivity	Respondent adults who reported reported doing physical activity or exercise during the past 30 days other than their regular job
GenHealth	General health reported by respondent. Categories include "Poor," "Fair," "Good," "Very good," and "Excellent"
SleepTime	Amount of sleep respondent reported getting on average in a 24-hour period
Asthma	Whether respondent was ever told they had asthma
KidneyDisease	Whether respondent was ever told they had kidney disease, no including kidney stones, bladder infection or incontinence
SkinCancer	Whether respondent was ever told they had skin cancer

Data Cleaning

The data was mostly clean. We converted ordinal variables to numerical variables using scores. This greatly reduces the number of predictors in the model, since each level (minus the reference) would be a predictor in the model. Additionally, we releveled GenHealth such that health decreases as score increases, and set White as the reference in the race variable.

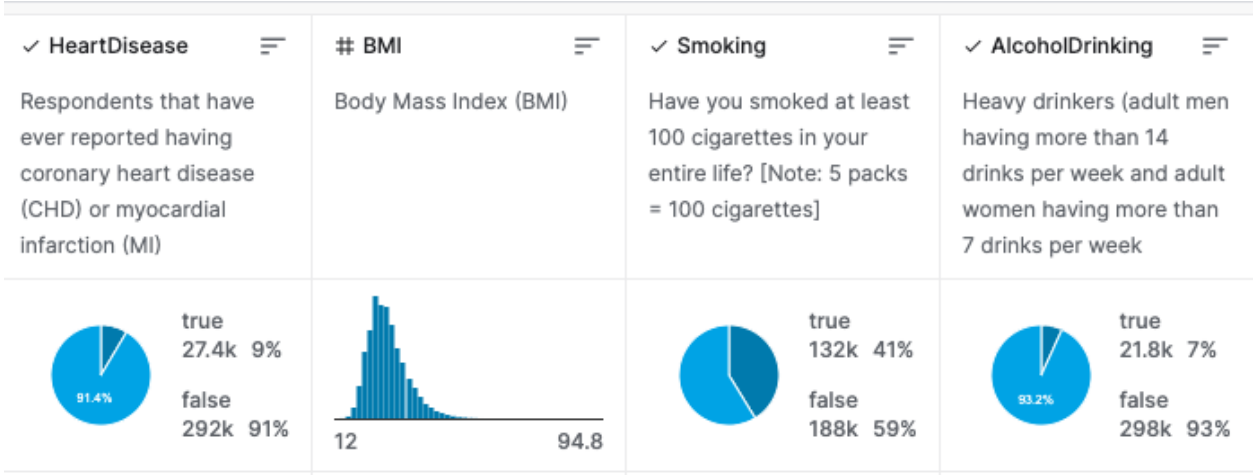
AgeScore: (21,27,32,37,42,47,52,57,62,67,72,77,90) -> (The midpoints of each "bin")

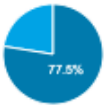
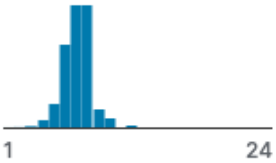
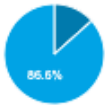
DiabeticScore: (0, 1, 2, 3) -> (No, No borderline, Yes, Yes pregnant)

GenHealthScore: (0, 1, 2, 3, 4) -> (Excellent, Very good, Good, Fair, Poor)

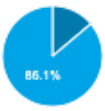
Visualization of the Data

We performed histograms and pie charts in order to analyze our data further and perform visualizations. We created pie charts for several categorical variables in our dataset and created histograms for a few non-categorical variables. From this we analyze the proportions of the different categories in the categorical variables, and see the numeric distribution in the non categorical variables. A few notable observations where, only 9% of the respondents reported being told they had diabetes, vs 91% of the respondents reported not being told they had diabetes. We also see that nearly half the respondents (41%) reported smoking at least 100 cigarettes in their lifetime, while 59% reported not smoking at least 100 cigarettes in their lifetime. Also another interesting thing to note is 78% of the respondents reported being told they had asthma, while only 22% of respondents reported not being told they had asthma.

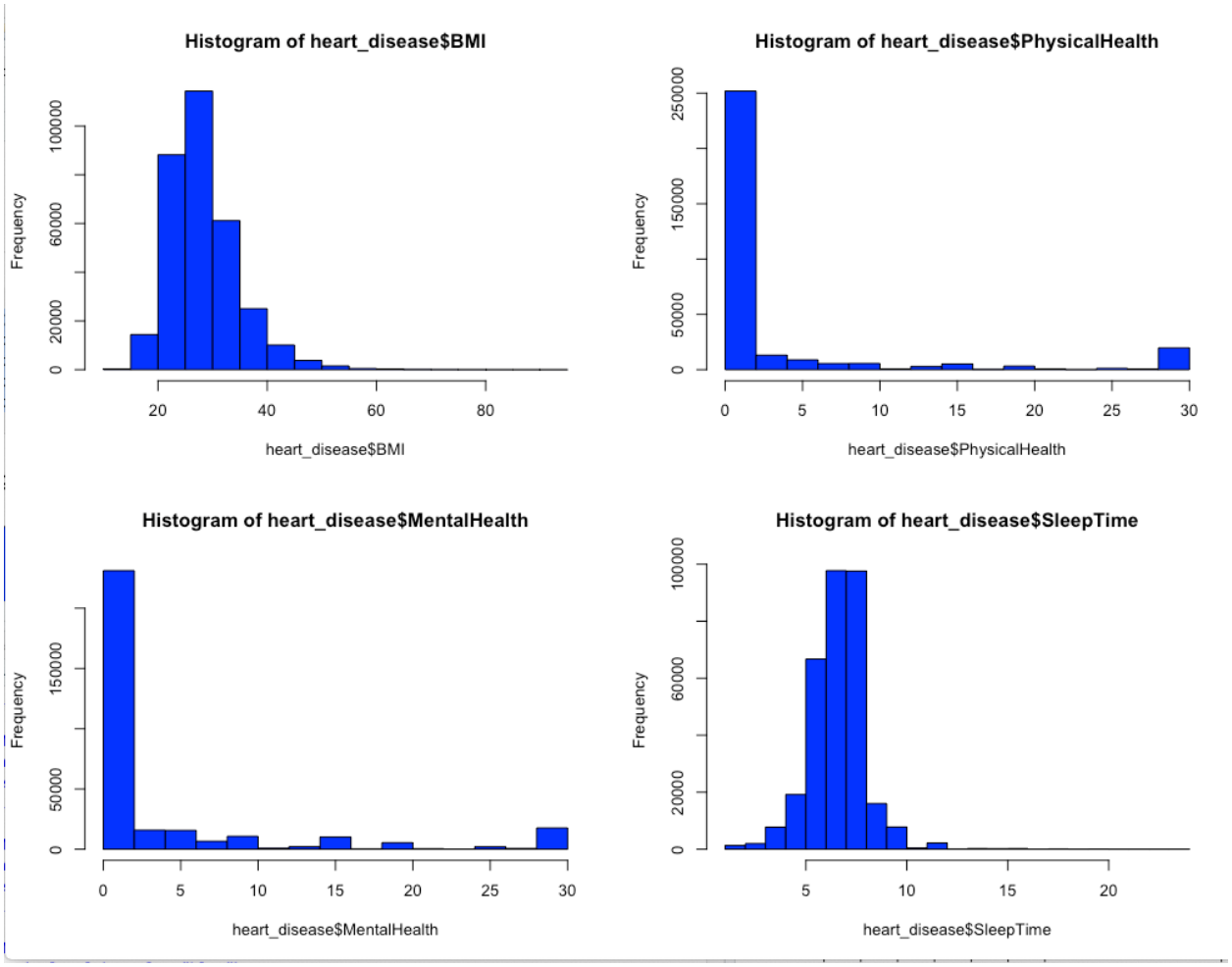


PhysicalActivity	GenHealth	SleepTime	Asthma
Adults who reported doing physical activity or exercise during the past 30 days other than their regular job	Would you say that in general your health is...	On average, how many hours of sleep do you get in a 24-hour period?	(Ever told) (you had) asthma?
 true 248k 78% false 71.8k 22%	Very good 36% Good 29% Other (112808) 35%		 true 42.9k 13% false 277k 87%

Sex	AgeCategory	Race	Diabetic
Are you male or female?	Fourteen-level age category	Imputed race/ethnicity value	(Ever told) (you had) diabetes?
Female 52% Male 48%	65-69 11% 60-64 11% Other (251958) 79%	White 77% Hispanic 9% Other (47137) 15%	No 84% Yes 13% Other (9340) 3%

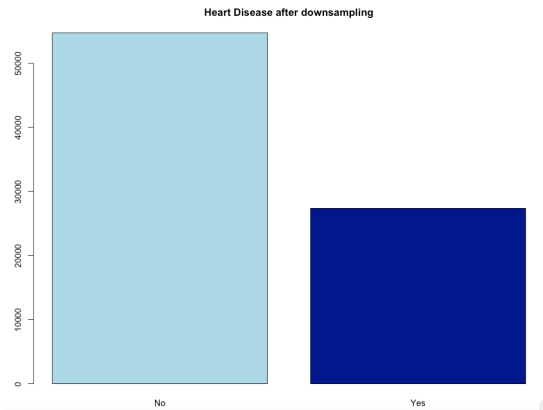
Stroke	PhysicalHealth	MentalHealth	DiffWalking
(Ever told) (you had) a stroke?	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?	Do you have serious difficulty walking or climbing stairs?
 true 12.1k 4% false 308k 96%			 true 44.4k 14% false 275k 86%

✓ KidneyDisease	✓ SkinCancer
Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?	(Ever told) (you had) skin cancer?
 true 11.8k 4% false 308k 96%	 true 29.8k 9% false 290k 91%



Unbalanced Response Variable and Resampling

The response variable is the binary factor HeartDisease with levels Yes and No. The proportion of Yes occurrences in the data is 27373/292422 = 8.5%, so the data is very unbalanced. We are concerned that training a model on unbalanced data will result in bias towards the majority class, so we downsample the No observations such that there are twice as many No's as Yes's.



Fitting the Logistic Model

First we split the data into a training and test set using an 80/20 split. That is, we train our model on 80% of the data and fit it on the remaining 20% of observations.

Now we fit the logistic regression model with all 18 predictors and produce the following output:

```
Call:
glm(formula = HeartDisease ~ ., family = binomial, data = heart_disease2,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8980  -0.7138  -0.3600   0.7560   3.1481

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -6.0388315  0.0936400 -64.490  < 2e-16 ***
BMI                                0.0118250  0.0016776   7.049 1.81e-12 ***
SmokingYes                         0.3860132  0.0206389  18.703  < 2e-16 ***
AlcoholDrinkingYes                -0.2400483  0.0459991  -5.219 1.80e-07 ***
StrokeYes                          1.1683305  0.0396899  29.437  < 2e-16 ***
PhysicalHealth                     0.0037299  0.0012813   2.911 0.003601 **
MentalHealth                       0.0061786  0.0013234   4.669 3.03e-06 ***
DiffWalkingYes                     0.2551703  0.0277258   9.203  < 2e-16 ***
SexMale                            0.7500774  0.0210167  35.690  < 2e-16 ***
RaceAmerican Indian/Alaskan Native 0.0611385  0.0758883   0.806 0.420451
RaceAsian                         -0.4207052  0.0892276  -4.715 2.42e-06 ***
RaceBlack                         -0.2422708  0.0414717  -5.842 5.16e-09 ***
RaceHispanic                      -0.1159969  0.0432923  -2.679 0.007376 **
RaceOther                          0.0218306  0.0577385   0.378 0.705360
PhysicalActivityYes                0.0458036  0.0238888   1.917 0.055191 .
SleepTime                         -0.0242179  0.0065171  -3.716 0.000202 ***
AsthmaYes                          0.2817197  0.0287569   9.797  < 2e-16 ***
KidneyDiseaseYes                   0.6174908  0.0413905  14.919  < 2e-16 ***
SkinCancerYes                      0.1433762  0.0297403   4.821 1.43e-06 ***
Age                                0.0526551  0.0007741  68.024  < 2e-16 ***
DiabeticScore                      0.2499736  0.0123329  20.269  < 2e-16 ***
GenHealthScore                     0.5126802  0.0120933  42.394  < 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: 83691  on 65694  degrees of freedom
Residual deviance: 60464  on 65673  degrees of freedom
AIC: 60508

Number of Fisher Scoring iterations: 5
```

Interestingly, physical activity is the least significant predictor in the model with an associated p-value of 0.055, yet difficulty walking is significant and positively correlated with heart disease. All predictors besides physical activity are statistically significant at the 0.05 level. Each coefficient has a magnitude less than 1. Alcohol drinking and sleep time appear to have a negative effect on the odds of heart disease. Blacks, Hispanics, and Asians also seem to have lower odds of heart disease compared to Whites. Women have lower odds than men to have heart disease. After applying backward subset selection using the AIC, PhysicalActivity is dropped from the model. This narrows down the number of predictors to 17. The likelihood ratio test comparing the models shows that the models can be assumed to be the same.

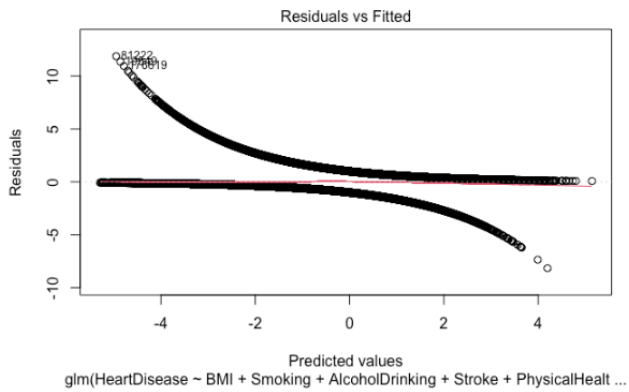
```
Analysis of Deviance Table

Model 1: HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalHealth +
    MentalHealth + DiffWalking + Sex + Race + SleepTime + Asthma +
    KidneyDisease + SkinCancer + Age + DiabeticScore + GenHealthScore
Model 2: HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalHealth +
    MentalHealth + DiffWalking + Sex + Race + PhysicalActivity +
    SleepTime + Asthma + KidneyDisease + SkinCancer + Age + DiabeticScore +
    GenHealthScore
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     65674      60468
2     65673      60464  1    3.683  0.05497 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here is the new model:

```
Call:
glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
    Stroke + PhysicalHealth + MentalHealth + DiffWalking + Sex +
    Race + SleepTime + Asthma + KidneyDisease + SkinCancer +
    Age + DiabeticScore + GenHealthScore, family = binomial,
    data = heart_disease2, subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9024  -0.7139  -0.3600   0.7555   3.1484

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -5.9830230  0.0889318 -67.277  < 2e-16 ***
BMI                                0.0115774  0.0016725   6.922 4.44e-12 ***
SmokingYes                         0.3842467  0.0206177  18.637  < 2e-16 ***
AlcoholDrinkingYes                -0.2408655  0.0460045  -5.236 1.64e-07 ***
StrokeYes                          1.1675848  0.0396865  29.420  < 2e-16 ***
PhysicalHealth                     0.0035364  0.0012773   2.769 0.005629 **
MentalHealth                       0.0061228  0.0013230   4.628 3.70e-06 ***
DiffWalkingYes                     0.2472590  0.0274155   9.019  < 2e-16 ***
SexMale                            0.7521179  0.0209918  35.829  < 2e-16 ***
RaceAmerican Indian/Alaskan Native  0.0593601  0.0758835   0.782 0.434066
RaceAsian                         -0.4203880  0.0891970  -4.713 2.44e-06 ***
RaceBlack                         -0.2434697  0.0414627  -5.872 4.31e-09 ***
RaceHispanic                      -0.1197466  0.0432327  -2.770 0.005609 **
RaceOther                          0.0224874  0.0577338   0.390 0.696905
SleepTime                         -0.0243946  0.0065157  -3.744 0.000181 ***
AsthmaYes                          0.2819941  0.0287570   9.806  < 2e-16 ***
KidneyDiseaseYes                   0.6160914  0.0413814  14.888  < 2e-16 ***
SkinCancerYes                      0.1454886  0.0297218   4.895 9.83e-07 ***
Age                                0.0525515  0.0007719  68.084  < 2e-16 ***
DiabeticScore                      0.2492847  0.0123272  20.222  < 2e-16 ***
GenHealthScore                     0.5101072  0.0120163  42.451  < 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: 83691  on 65694  degrees of freedom
Residual deviance: 60468  on 65674  degrees of freedom
AIC: 60510

Number of Fisher Scoring iterations: 5
```

Now that we have our model, we perform diagnostic checking to ensure the model meets the assumptions of logistic regression. These are: 1) the data has a linear relationship with the logit of the response, 2) there is no multicollinearity among the predictors, and 3) there are no highly influential observations on the data.

Analyzing the residuals, we see U-shaped patterns, which suggests our numerical predictors are not linear with the logit. This violates the first assumption of logistic regression.

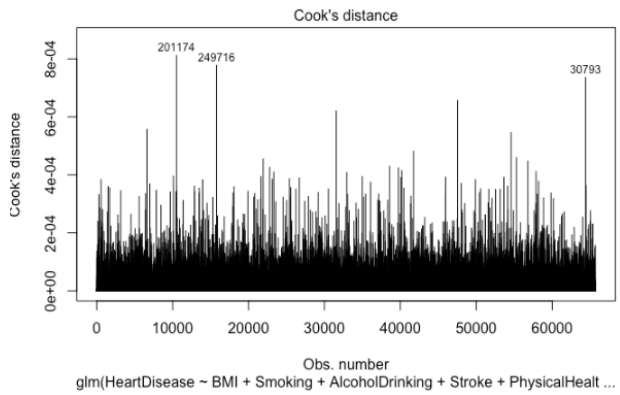


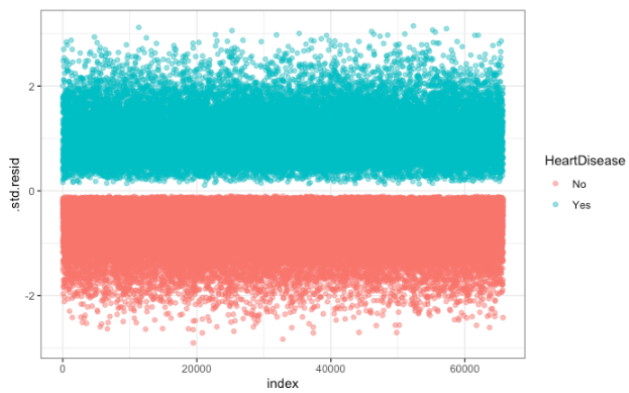
Here is the correlation matrix:

	BMI	PhysicalHealth	MentalHealth	SleepTime	Age
BMI	1.00000000	0.10978754	0.06413057	-0.051822254	-0.00987085
PhysicalHealth	0.10978754	1.00000000	0.28798667	-0.061386632	0.10992440
MentalHealth	0.06413057	0.28798667	1.00000000	-0.119716788	-0.15543072
SleepTime	-0.05182225	-0.06138663	-0.11971679	1.000000000	0.10684451
Age	-0.00987085	0.10992440	-0.15543072	0.106844508	1.00000000
DiabeticScore	0.20247228	0.15136118	0.03294478	0.000449238	0.19037529
GenHealthScore	0.23071978	0.48269718	0.24162528	-0.063071012	0.18860567

	DiabeticScore	GenHealthScore
BMI	0.202472284	0.23071978
PhysicalHealth	0.151361181	0.48269718
MentalHealth	0.032944777	0.24162528
SleepTime	0.000449238	-0.06307101
Age	0.190375286	0.18860567
DiabeticScore	1.000000000	0.26801834
GenHealthScore	0.268018341	1.00000000

There is no apparent multicollinearity in the predictors. Now we look for influential points.



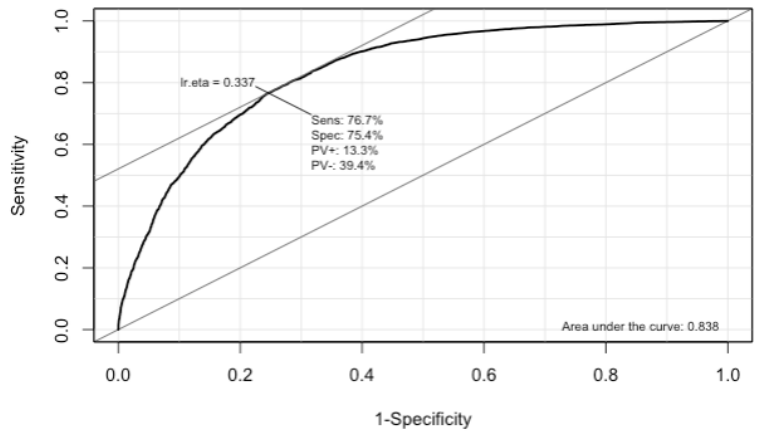


We see that most of the standardized residuals are within 0 +/- 2, with almost all within 0 +/- 3. There were 8 observations in the dataset with standardized residuals with magnitudes greater than 3.

Two of three logistic regression assumptions were met.

Making Predictions

We use the model trained on the training set to classify the occurrences of heart disease in the test set. If the probability is above a certain cutoff, the model assigns “Yes” to the predicted value. To find the best cutoff for π_0 we draw the ROC curve and find the point that maximizes the area under the curve (AUC).



The best cutoff is $\pi_0 = 0.337$. The following is the confusion matrix after assigning predictions:

```
glm.pred   No  Yes
     No  8226 1234
     Yes 2766 4198

glm.pred         No        Yes
     No  0.50085241 0.07513395
     Yes 0.16841208 0.25560156
[1] 0.756454
```

The model accuracy is 0.756; the test error is 0.234.

Cross-validation

Leave-one-out cross-validation is too computationally expensive and thus is not performed. We perform k-fold cross-validation with k = 10 folds. The resulting error was 0.151.

Comparison with Probit and Identity Links

```
Call:
glm(formula = HeartDisease ~ . - PhysicalActivity, family = binomial(link = "probit"),
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1277  -0.7307  -0.3497   0.7809   3.5007

Coefficients:
                                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)                       -3.4319522  0.0497633 -68.966 < 0.0000000000000002 ***
BMI                                0.0060573  0.0009688   6.252       0.000000000405 ***
SmokingYes                         0.2201179  0.0119938  18.353 < 0.0000000000000002 ***
AlcoholDrinkingYes                -0.1356333  0.0262001  -5.177       0.000000225693 ***
StrokeYes                          0.6744269  0.0228605  29.502 < 0.0000000000000002 ***
PhysicalHealth                     0.0022140  0.0007505   2.950             0.003176 **
MentalHealth                       0.0035969  0.0007674   4.687       0.000002768341 ***
DiffWalkingYes                     0.1536681  0.0163138   9.420 < 0.0000000000000002 ***
SexMale                            0.4306214  0.0120918  35.613 < 0.0000000000000002 ***
RaceAmerican Indian/Alaskan Native  0.0378650  0.0441706   0.857             0.391310
RaceAsian                         -0.2153245  0.0493014  -4.368       0.000012567123 ***
RaceBlack                         -0.1304751  0.0239399  -5.450       0.000000050340 ***
RaceHispanic                      -0.0485090  0.0244890  -1.981             0.047608 *
RaceOther                          0.0191942  0.0333636   0.575             0.565087
SleepTime                         -0.0126798  0.0037992  -3.337             0.000845 ***
AsthmaYes                          0.1654748  0.0167214   9.896 < 0.0000000000000002 ***
KidneyDiseaseYes                   0.3645147  0.0242235  15.048 < 0.0000000000000002 ***
SkinCancerYes                      0.0918728  0.0175938   5.222       0.000000177117 ***
Age                                0.0298507  0.0004270  69.905 < 0.0000000000000002 ***
DiabeticScore                      0.1478860  0.0072999  20.259 < 0.0000000000000002 ***
GenHealthScore                     0.2953130  0.0069288  42.621 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83691  on 65694  degrees of freedom
Residual deviance: 60494  on 65674  degrees of freedom
AIC: 60536
```

```
Call:
glm(formula = HeartDisease ~ . - PhysicalActivity, family = gaussian(link = "identity"),
    data = new_train, start = strt)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.23407  -0.28805  -0.07864   0.31917   1.25755

Coefficients:
                                     Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                       -0.36848765  0.01191588 -30.924 < 0.0000000000000002 ***
BMI                                0.00023972  0.00025442   0.942             0.346092
SmokingYes                         0.05524216  0.00322313  17.139 < 0.0000000000000002 ***
AlcoholDrinkingYes                -0.04019780  0.00652230  -6.163 0.000000000717340622 ***
StrokeYes                          0.21839791  0.00619089  35.277 < 0.0000000000000002 ***
PhysicalHealth                     0.00103439  0.00020892   4.951 0.000000739340876353 ***
MentalHealth                       0.00059619  0.00020307   2.936             0.003327 **
DiffWalkingYes                     0.07209013  0.00469694  15.348 < 0.0000000000000002 ***
SexMale                            0.11692315  0.00313893  37.249 < 0.0000000000000002 ***
RaceAmerican Indian/Alaskan Native  0.00247543  0.01177348   0.210             0.833470
RaceAsian                         -0.03876712  0.01080354  -3.588             0.000333 ***
RaceBlack                         -0.03683535  0.00615024  -5.989 0.000000002119038341 ***
RaceHispanic                      -0.01593969  0.00594821  -2.680             0.007370 **
RaceOther                          0.00310952  0.00861923   0.361             0.718276
SleepTime                         -0.00052752  0.00101790  -0.518             0.604292
AsthmaYes                          0.03982851  0.00445880   8.933 < 0.0000000000000002 ***
KidneyDiseaseYes                   0.12191934  0.00670183  18.192 < 0.0000000000000002 ***
SkinCancerYes                      0.04027963  0.00499504   8.064 0.000000000000000751 ***
Age                                0.00703427  0.00009782  71.911 < 0.0000000000000002 ***
DiabeticScore                      0.05016606  0.00208135  24.103 < 0.0000000000000002 ***
GenHealthScore                     0.08165127  0.00181718  44.933 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1545494)

    Null deviance: 14613  on 65694  degrees of freedom
Residual deviance: 10150  on 65674  degrees of freedom
AIC: 63789
```

We produce confusion matrices for the probit and identity link models.

```
ident.pred    0    1
         0 7611  950
         1 3381 4482
[1] 0.7363005
```

```
probit.pred    No   Yes
        No  10296  3362
        Yes   696  2070
[1] 0.7529226
```

```
glm.pred   No  Yes
     No  8226 1234
     Yes 2766 4198

glm.pred         No        Yes
     No  0.50085241 0.07513395
     Yes 0.16841208 0.25560156
[1] 0.756454
```

The logit model has the highest prediction accuracy. The probit model seems to have a much lower sensitivity (true positive rate) but a much higher specificity (true negative rate) than the others. The identity link appears to have a higher sensitivity but lower specificity than the logit model.

Conclusion

Each predictor with the exception of physical activity is significant in the logistic regression model at the 0.05 level. We note that physical activity had a p-value of 0.055. The logit model seems to predict the occurrence of heart disease moderately well, with an accurate rate of a little over 75%. However, the linearity assumption of the predictors to the logit of the response was not met. It is possible that polynomial regression or other nonlinear methods could improve our model. Furthermore, we undersampled the majority class, so results could differ based on how we choose to resample the data.