621-FinalProject

Ahmed Elsaeyed

2022-11-30

Probability of Cardiovascular Disease

This research centers around building a model that predicts the probability of cardiovascular disease given some prior health information.

Abstract

For our final assignment, we will be looking at a dataset of physical characteristics, behavioral characteristics, and health results of patients aged 30-59 in the 1940's in the United States. The data was gathered by the Framingham Heart Study, which is a long-term ongoing study. 5209 subjects were monitored to see if cardiac health can be influenced by lifestyle and environmental factors. It is mainly due to studies like this one, and indeed this one in particular due to its large scope, that we take that conclusion for granted. Before this study, the contributing factors of cardiac health were not clear at all.

A result of this study is the 10-year cardiovascular risk score was developed, which is an estimation of how likely it is that someone will have a coronary disease based on the findings of the study. The objective of this project is to build a model to predict this to take their research a step further.

Keywords - Coffy

Introduction - Coffy

Literature Review

Framingham Study is a study initiated by the United State Public Health Service, that started in 1948 and has been ongoing since. The origin of the study has been closely linked with the cardiovascular health of President Franklin D. Roosevelt and his death from hypertensive heart disease and stroke in 1945. It aims to investigate the epidemiology and factors that are involved in the development of cardiovascular disease. The plan of the study was to track a large cohort of patients over time. The study initially recruited the original cohort of 5209 participants from the town if Framingham, MA. The age group for this study ranges from 30-59 years old. The patients were given questionnaires and exams every 2 years and these questions and exams expanded with time. This study was continued over three generations of the original participants.

The data collected by the study was used to develop the Framingham Risk Score. The Framingham Risk Score is a sex-specific algorithm that is used to estimate the 10 year cardiovascular risk of an individual. This score gives an estimate probability of a person developing cardiovascular disease within a specified amount of time, the range of which can be from 10 to 30 years.

The study collect a range of 16 variables measurements from the patients in order to create the database. The variables include: Male, Age, Education, Current Smoker, Cigs Per Day, BP Medications, Prevalent Stroke, Prevalent Hypertension, Diabetes, Total Cholesterol, Systolic Blood Pressure, Diastolic Blood Pressure, Heart

Variable	Description
Sex	Participant Sex (Male or Female)
Age	Age at exam (years)
Education	Attained Education
Current Smoker	Whether or not the patient is a current smoker
Cigs Per Day	The number of cigarettes that the person smoked on average in one day
BP Meds	Whether or not the patient was on blood pressure medication
Prevelant Stroke	Whether or not the patient had previously had a stroke
Prevalant Hyp	Whether or not the patient was hypertensive
Diabetes	Whether or not the patient had diabetes
Tot Chol	Total cholesterol
Sys BP	Systolic blood pressure
Dia BP	Diastolic blood pressure
BMI	Body Mass Index
Heart Rate	Heart rate
Glucose	Glucose level
Ten Year CHD	10 year risk of coronary heart disease, 'TARGET: 1 = Yes | 2 = No'

Rate, Glucose, CHD in 10 years. Most of the variable are categorized by their role is a person's cardiovascular health.

Demographic Risk Factors: - Male - Age - Education

Behavioral Risk Factors: - Current Smoker - Cigs Per Day

Medical History Factors: - BP Medications - Prevalent Stroke - Prevalent Hypertension - Diabetes

Physical Exam Risk: - Total Cholesterol - Systolic Blood Pressure - Diastolic Blood Pressure - Heart Rate - Glucose

Methodology

Our methodology for creating a predictive model for future coronary heart disease begins with cleaning and preprocessing our subject dataset. The dataset itself contains 16 fields, one of which is the target: `TenYearCHD`. While the dataset was relatively complete to begin with, we needed to employ multiple operations in order to transform it into a format which subsequent models could interpet. In this vain, we addressed missing values, multicollinearity, outliers, and mis-shaped data using a combination of self-developed functions and transformations. After cleaning / preprocessing, we split the new dataframe into training and testing sets in order to create and test various machine learning models. While developing models, we decided to try multiple approaches, including Binary Logistic Regression, standard and with log transforms, and Step-AIC. After each model, we calculated the relvant performance metrics and stored them in a tracker. Ultimately, our selection criteria was heavily influenced by the models' recall and f1 scores. With medical diagnosis, it's imperative that individuals with an illness be properly classified.

Experimentation and Results

```
[1] 4240   16
```

Data Exploration

Our data set had 4,240 observations and 16 variables. `Sex`, `education`, `currentSmoker`, `BPMeds`, `prevalentStroke`, `prevalentHyp`, `diabetes` and `TenYearCHD` in the analysis carried out in this report are categorical variables, while the rest are numeric. The histograms for the numeric variables revealed that `BMI`, `glucose`, and `totChol` were skewed, and a Box-Cox transformation was undertaken for these variables in

order to create new variables that contained the transformed versions of `BMI`, `glucose`, and `totChol`. An analysis into the correlation was done in order to remove variables that were highly correlated with another. Also, the data had several observations with missing data. The methodology that was undertaken to account for these missing values is shown in the "*Missing Values*" section. For comparison purposes, two different datasets were created; one containing the original data with observations with missing values omitted, and another containing manipulated data, where the data went through several transformations as explained in this report.

```
     Sex            age        education   currentSmoker   cigsPerDay
 female:2420   Min.   :32.00   1   :1720   No :2145      Min.   : 0.000
 male  :1820   1st Qu.:42.00   2   :1253   Yes:2095      1st Qu.: 0.000
               Median :49.00   3   : 689                 Median : 0.000
               Mean   :49.58   4   : 473                 Mean   : 9.006
               3rd Qu.:56.00   NA's: 105                 3rd Qu.:20.000
               Max.   :70.00                             Max.   :70.000
                                                         NA's   :29
  BPMeds     prevalentStroke prevalentHyp diabetes      totChol
 0   :4063   0:4215          0:2923       No :4131   Min.   :107.0
 1   : 124   1:  25          1:1317       Yes: 109   1st Qu.:206.0
 NA's:  53                                           Median :234.0
                                                     Mean   :236.7
                                                     3rd Qu.:263.0
                                                     Max.   :696.0
                                                     NA's   :50
     sysBP            diaBP            BMI          heartRate
 Min.   : 83.5   Min.   : 48.0   Min.   :15.54   Min.   : 44.00
 1st Qu.:117.0   1st Qu.: 75.0   1st Qu.:23.07   1st Qu.: 68.00
 Median :128.0   Median : 82.0   Median :25.40   Median : 75.00
 Mean   :132.4   Mean   : 82.9   Mean   :25.80   Mean   : 75.88
 3rd Qu.:144.0   3rd Qu.: 90.0   3rd Qu.:28.04   3rd Qu.: 83.00
 Max.   :295.0   Max.   :142.5   Max.   :56.80   Max.   :143.00
                                 NA's   :19      NA's   :1
    glucose       TenYearCHD
 Min.   : 40.00   0:3596
 1st Qu.: 71.00   1: 644
 Median : 78.00
 Mean   : 81.96
 3rd Qu.: 87.00
 Max.   :394.00
 NA's   :388
```

Multicollinearity Analysis

Correlation refers to the strength of a relationship between two or more variables. Problems arise when there significant correlation between two or more predictor variables within a dataset, as this can lead to solutions that are wildly varying and possibly numerically unstable and there is redundancy between predictor variables. One solution to this is to drop one variable that is highly correlated with the other. Calkins explains that we can describe the correlation values between variables in the follwing manner:

- Correlation coefficients between 0.9 and 1 = very highly correlated
- Correlation coefficients between 0.7 and 0.9 = highly correlated
- Correlation coefficients between 0.5 and 0.7 = moderately correlated
- Correlation coefficients under 0.5 = low correlation

The correlations between the variables in the dataset were calculated. What was found was that the vast majority of the variables had a correlation that was less than 0.25. `sysBP` and `diaBP` had a correlation of

0.79, meaning that these two variables were highly correlated. One of these variables was removed from the dataset. When a binary logistic regression model with the original data was constructed, it was found that `sysBP` had a p-value of 0.00015 indicating a high degree of statistical significance, while `diaBP` had a p-value of 0.55. Therefore, the `diaBP` variable was removed from the dataset. The second highest correlation value after this was 0.38, and that was for the `BMI` and `diaBP` variables, but with the removal of the `diaBP` variable, and the fact that 0.38 would be considered a low correlation, `BMI` was kept in.

Missing Values

The dataset itself contains several observations containing missing values. The percentage of missing data for each of the predictor variables is shown in Table 1.

Variable Name	Percentage of Missing Data
education	2.48
cigsPerDay	0.68
BPMeds	1.25
totChol	1.18
BMI	0.45
heartRate	0.02
glucose	9.15

Table 2: Percentage of missing data for each of the variables that had missing data.

Ultimately, it was decided to remove the observations with the missing values instead of imputing. The reason for this is explained in the "*Removing Outlier Values*" section.

Data Preparation

Removing Outlier Values

The boxplots, summaries, and histograms that were generated revealed that some of these heart rates, diastolic, and systolic values as well as total cholesterol do not seem realistic. Imputing the values would not be appropriate because this is medical data and would harm the authenticity of the dataset, and so we have decided to simply remove the bad data.

We conducted research on suspected variables to further examine the authenticity of the data points in the dataset. Using multiple credited medical sources we created plausible ranges for each variable and omitted extreme or impossible values. Brief discussions and analysis of the process for each variable are below.

Heart Rate

An article by Cardiologist Jim Liu of Ohio State University states a normal resting heart rate, measured when the person is not exercising, is between 60 and 100 bpm (beats per minute). The further you move from 100 bpm increases the level of risk on the high sided. On the decreasing side going lower than 60 deviates from the safe levels. The risk levels are assigned by taking into account the heart rate along with the heart rhythm. Taking this into account we removed records with heart rates over a 100.

Systolic Blood Pressure

The FDA states that a normal blood pressure reading is 120/80 with 120 being the systolic measurement. If the systolic reading is 130, the blood pressure is considered high at Stage 1, and a reading of 140 is considered Stage 2. A systolic reading of 180 or greater is considered 'hypertensive crisis' and needs immediate medical attention. Taking into account the different stages of risk we declared a range of 100 to 160 an acceptable Systolic Blood Pressure reading.

Total Cholesterol

The classification of the levels of total cholesterol are as follows. Less than 200 is optimal, between 200 and 240 is elevated, and greater than 240 is high. Upon further researched that higher levels of total cholesterol, such as 500 are possible. For this reason, we decided to cap our total cholesterol maximum at 500.

Box-Cox Transformation for Skewed Variables

For the variables that exhibited skewness as explained in the "*Data Exploration*" section of this paper, A Modern Approach to Regression with R explains the following:

> . . . if the skewed predictor can be transformed to have a normal distribution conditional on Y, then just the transformed version of X should be included in the logistic regression model. (Sheather, 284)

In order to address this skewness and attempt to normalize these features for future modeling, we employed Box-Cox transformations. Because some of these values include 0, we replaced any zero values with infinitesimally small, non-zero values.

The λ's that were used to transform the skewed variables are shown on Table 3.

Column Name	λ
BMI	-0.309
glucose	-1.279
totChol	0.069

Table 3: λ's for transforming skewed variables to normal distribution.

Split Data Into Testing and Training

The data was into testing and training subsets such that 70% of it will be used to train, and 30% to test. The first row shows the split for the testing data while the second row shows the split for the training data. Tables 4 and 5 show the distributions of the testing and training data using the original and modified datasets.

	Number of Observations where `TenYearCHD` = 1	Number of Observations where `TenYearCHD` = 0
Test	930	167
Train	2171	390

Table 4: Distribution of testing and training data using original dataset

	Number of Observations where `TenYearCHD` = 1	Number of Observations where `TenYearCHD` = 0
Test	736	109
Train	1716	255

Table 5: Distribution of testing and training data using modified/transformed dataset

Building Models

Binary Logistic Regression Model with Original Data

A binary logistic regression model was generated using the original dataset. The observations with missing values were omitted. The final binary logistic regression takes on the following form:

$$logit(p) = -9.155 + (0.504 * Sex_{male}) + (0.066 * age) +$$
$$(-0.290 * education_2) + (-0.223 * education_3) + (-0.120 * education_4) +$$
$$(0.086 * currentSmoker_{Yes}) + (0.019 * cigsPerDay) + (0.123 * BPMeds_1) +$$
$$(0.982 * prevalentStroke_1) + (0.091 * prevalentHyp_1) + (-0.435 * diabetes_{Yes}) +$$
$$(0.002 * totChol) + (0.018 * sysBP) + (-0.005 * diaBP) + (0.02 * BMI) +$$
$$(-0.005 * heartRate) + (0.01 * glucose)$$

Equation 1: Binary logistic regression model using original data.

In this model, the variables that had a p-value greater than 0.05 were `Sex` `age`, `cigsPerDay`, `sysBP` and `glucose`. The sex of the person definitely seems to play a significant factor in determining whether or not a person will develop heart disease. Figure XX revealed that men that were surveyed on average smoked more cigarettes per day than women. In fact, Figure XX reveals that within the sample, women on average do not smoke. Since smoking increases the formation of plaque in the blood vessels which leads to coronary heart disease according to the CDC, the fact that men smoke more than women is why the sex variable in particular is of great statistical significance. The National Institute on Aging points out the following:

> People age 65 and older are much more likely than younger people to suffer a heart attack, to have a stroke, or to develop coronary heart disease (commonly called heart disease) and heart failure. Heart disease is also a major cause of disability, limiting the activity and eroding the quality of life of millions of older people. (NIH)

This explains why `age` was statstically significant in this model as well. The Brisighella Heart Study conducted in 2003 was able to conclude that within their sample that systolic blood pressure was a strong predictor of coronary heart disease, while diastolic blood pressure was not statistically significant. This result in this study is also seen in this model; `sysBP` has a much lower p-value (0.0001) than `diaBP` (0.55). `cigsPerDay` is tied to `Sex` as shown on Figure xx, which is why the `cigsPerDay` variable is also statistically significant. Park was able to conclude the following about glucose levels and coronary heart disease based on a study which had a sample size of 1,197,384 adults:

> Both low glucose level and impaired fasting glucose should be considered as predictors of risk for stroke and coronary heart disease. The fasting glucose level associated with the lowest cardiovascular risk may be in a narrow range. (Park)

This provides evidence as to probably why `glucose` is also statistically significant within the model.

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial(link = "logit"),
    data = original_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1920  -0.5860  -0.4192  -0.2732   2.7225

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -9.154885   0.877043 -10.438  < 2e-16 ***
Sexmale           0.503829   0.132129   3.813 0.000137 ***
age               0.066400   0.008136   8.161 3.32e-16 ***
education2       -0.290393   0.149348  -1.944 0.051846 .
education3       -0.223186   0.184932  -1.207 0.227486
education4       -0.119852   0.199401  -0.601 0.547800
currentSmokerYes  0.086271   0.189409   0.455 0.648768
cigsPerDay        0.018727   0.007488   2.501 0.012384 *
BPMeds1           0.123212   0.286340   0.430 0.666977
```

```
prevalentStroke1  0.982234   0.595212   1.650 0.098897 .
prevalentHyp1     0.090940   0.168521   0.540 0.589447
diabetesYes      -0.434817   0.409077  -1.063 0.287818
totChol           0.001996   0.001396   1.431 0.152566
sysBP             0.017671   0.004672   3.783 0.000155 ***
diaBP            -0.004660   0.007811  -0.597 0.550789
BMI               0.021327   0.015395   1.385 0.165957
heartRate        -0.004900   0.005042  -0.972 0.331105
glucose           0.011013   0.002930   3.759 0.000171 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2185.3  on 2560  degrees of freedom
Residual deviance: 1903.2  on 2543  degrees of freedom
AIC: 1939.2

Number of Fisher Scoring iterations: 5
```

Binary Logistic Regression Model with Modified Data

A binary logistic regression model was generated using the modified dataset. The observations with missing values were omitted. The final binary logistic regression model that used modified data takes on the following form:

$$\begin{aligned} logit(p) = -128.3 + (0.52 * Sex_{male}) + (0.073 * age) + \\ (-0.105 * education_2) + (-0.55 * education_3) + (-0.074 * education_4) + \\ (0.277 * currentSmoker_{Yes}) + (0.013 * cigsPerDay) + (0.091 * BPMeds_1) + \\ (0.651 * prevalentStroke_1) + (0.145 * prevalentHyp_1) + (-0.026 * diabetes_{Yes}) + \\ (0.009 * totChol) + (0.013 * sysBP) + (0.18 * BMI) + (-0.009 * heartRate) + \\ (0.013 * glucose) + (-13 * tf\_BMI) + (-140.9 * tf\_glucose) + (-1.39 * tf\_totChol) \end{aligned}$$

Equation 2: Binary logistic regression model using modified data.

The parameters in this model that had p-values less than 0.05 were `Sex`, `age`, `education_3`, `sysBP` and `glucose`. None of the Box-Cox transformed variables had p-values less than 0.05. We see the same variables (except `cigsPerDay`) that were statistically significant when the original data was used.

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial(link = "logit"),
    data = modified_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9142  -0.5564  -0.3987  -0.2737   2.8641

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.283e+02  1.349e+02   0.951 0.341801
Sexmale           5.164e-01  1.550e-01   3.331 0.000865 ***
age               7.270e-02  8.949e-03   8.123 4.54e-16 ***
education2       -1.051e-01  1.683e-01  -0.625 0.532142
```

```
education3       -5.491e-01  2.282e-01  -2.406 0.016110 *
education4       -7.443e-02  2.193e-01  -0.339 0.734331
currentSmokerYes  2.774e-01  2.160e-01   1.284 0.199058
cigsPerDay        1.302e-02  8.515e-03   1.529 0.126229
BPMeds1           9.097e-02  5.050e-01   0.180 0.857039
prevalentStroke1  6.507e-01  6.103e-01   1.066 0.286348
prevalentHyp1     1.450e-01  1.959e-01   0.740 0.459116
diabetesYes      -2.551e-01  4.898e-01  -0.521 0.602544
totChol           8.562e-03  9.697e-03   0.883 0.377252
sysBP             1.310e-02  6.409e-03   2.044 0.040943 *
BMI               1.805e-01  1.129e-01   1.599 0.109869
heartRate        -9.006e-03  7.019e-03  -1.283 0.199437
glucose           1.331e-02  5.245e-03   2.538 0.011138 *
tf_BMI           -1.300e+01  9.534e+00  -1.364 0.172658
tf_glucose       -1.409e+02  1.738e+02  -0.811 0.417489
tf_totChol       -1.390e+00  1.854e+00  -0.750 0.453482
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1653.4  on 2127  degrees of freedom
Residual deviance: 1467.5  on 2108  degrees of freedom
AIC: 1507.5

Number of Fisher Scoring iterations: 5
```

Step-AIC Binary Linear Regression Model with Original Data

Step AIC works by deselecting features that negatively affect the AIC, which is a criterion similar to the R-squared. It selects the model with not only the best AIC score but also a model with less predictors than the full model, since the full model may have predictors that do not contribute or negatively contribute to the model's performance. The direction for the Step AIC algorithm was set to `both`, because this implements both forward and backward elimination in order to decide if a predictor negatively affects the model's performance. The final step-AIC binary logistic regression model that used the original data takes on the following form:

$$logit(p) = -9.589 + (0.503 * Sex_{male}) + (0.072 * age) + (0.02 * cigsPerDay) + (1.004 * prevalentStroke_1) + (0.017 * sysBP) + (0.023 * BMI) + (0.009 * glucose)$$

Equation 3: Step-AIC binary logistic regression model using the original data.

The p-values for this model, which is far more parsimonious than the original binary logistic regression model that used all of the predictors, were all below 0.05 except for `prevalentStroke1` and `BMI`, which had p-values of 0.0861 and 0.1179 respectively.

```
Call:
glm(formula = TenYearCHD ~ Sex + age + cigsPerDay + prevalentStroke +
    sysBP + BMI + glucose, family = binomial(link = "logit"),
    data = original_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2231  -0.5892  -0.4224  -0.2792   2.6812
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -9.589258   0.603218 -15.897  < 2e-16 ***
Sexmale           0.503148   0.127044   3.960 7.48e-05 ***
age               0.072072   0.007718   9.339  < 2e-16 ***
cigsPerDay        0.020838   0.005012   4.158 3.21e-05 ***
prevalentStroke1  1.003640   0.584843   1.716   0.0861 .
sysBP             0.017106   0.002706   6.321 2.60e-10 ***
BMI               0.022856   0.014619   1.563   0.1179
glucose           0.008856   0.002112   4.193 2.75e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2185.3  on 2560  degrees of freedom
Residual deviance: 1912.5  on 2553  degrees of freedom
AIC: 1928.5

Number of Fisher Scoring iterations: 5
```

Step-AIC Binary Logistic Regression Model with Modified Data

The final step-AIC binary logistic regression model that used the modified data takes on the following form:

$$logit(p) = -14.41 + (0.52 * Sex_{male}) + (0.073 * age) + (-0.09 * education_2) + (-0.55 * education_3) + (-0.056 * education_4) + (0.02 * cigsPerDay) + (0.015 * sysBP) + (0.2 * BMI) + (0.009 * glucose) + (-14.76 * tf\_BMI)$$

Equation 4: Step-AIC binary logistic regression model using the modified data.

The p-values for this model, which is far more parsimonious than the original binary logistic regression model that used all of the predictors, were all below 0.05 except for `Intercept`, `education2`, `education4`, `BMI` and `tf_BMI`; 3 more than the amount of predictors that had values above 0.05 for the step-AIC binary logistic regression model using the original predictors.

```
Call:
glm(formula = TenYearCHD ~ Sex + age + education + cigsPerDay +
    sysBP + BMI + glucose + tf_BMI, family = binomial(link = "logit"),
    data = modified_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7476  -0.5552  -0.4010  -0.2753   2.7893

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  14.410817  14.969749   0.963 0.335717
Sexmale       0.520105   0.150793   3.449 0.000562 ***
age           0.072950   0.008761   8.326  < 2e-16 ***
education2   -0.093929   0.167136  -0.562 0.574121
education3   -0.545300   0.226890  -2.403 0.016245 *
education4   -0.056454   0.217162  -0.260 0.794892
```

```
cigsPerDay    0.020256   0.005666   3.575 0.000350 ***
sysBP         0.015889   0.004996   3.180 0.001471 **
BMI           0.198747   0.109916   1.808 0.070579 .
glucose       0.009194   0.002376   3.870 0.000109 ***
tf_BMI      -14.756198   9.247248  -1.596 0.110547
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1653.4  on 2127  degrees of freedom
Residual deviance: 1474.6  on 2117  degrees of freedom
AIC: 1496.6

Number of Fisher Scoring iterations: 5
```

Model Selection

The various metrics that were used to determine the best model are shown in Table 6. Figure XX is a bar chart which plots all of the metrics for all of the models. Table 6 and Figure xx reveal that the accuracy, precision, recall, AUC, and F-score are higher when the original data was used as opposed to the modified. Table 6 revealed that the step-AIC binary logistic regression model using the original data has the second highest AIC, but it also has the highest precision and F-score out of all of the models. Taking all of this into consideration, and the fact that the step-AIC binary logistic regression model using the original data is the most parsimonious out of all of the models, this model is the best performing model out of all of them. The AUC curve for the step-AIC binary logistic regression with original data is provided in Figure xx.

Model	Precision	Recall	AIC	AUC	F-score	Accuracy	Error
Bin. Log. w/ Original Data	0.68	0.11	1939.25	0.71	0.195	0.86	0.14
Bin. Log. w/ Modified Data	0.67	0.02	1507.5	0.7	0.033	0.87	0.13
Step AIC Bin. Log. w/ Original Data	0.7	0.11	1928.55	0.71	0.196	0.86	0.14
Step AIC Bin. Log. w/ Modified Data	0.67	0.02	1496.61	0.69	0.033	0.87	0.13

Table 6: Model metrics for binary logistic regression models

Discussion and Conclusions

In this paper, 4 different binary logistic regression models were generated in order to predict the 10-year risk of chronic heart disease. The results from the analysis carried out in this report indicate that the transformation of the skewed variables to a normal distribution and the removal of outliers resulted in worse model performance when comparing the metrics to these models and the models that used the original unaltered dataset. With that being said, the AUC's for all of the models were relatively the same. The final model that was selected in order to predict the 10-year risk of chronic heart disease was the best performing and the most parsimonious. We were able to test the validity of this particular model from when the data was split into testing and training datasets.

References

Center for Drug Evaluation and Research. (2021a, January 21). High Blood Pressure– Understanding the Silent Killer. U.S. Food And Drug Administration. https://www.fda.gov/drugs/special-features/high-blood-pressure-understanding-silent- killer

Center for Drug Evaluation and Research. (2021b, January 21). High Blood Pressure– Understanding the Silent Killer. U.S. Food And Drug Administration. https://www.fda.gov/drugs/special-features/high-blood-

pressure-understanding-silent- killer

Framingham Study | Boston Medical Center. (n.d.). https://www.bmc.org/stroke-and- cerebrovascular-center/research/framingham-study

High cholesterol - Symptoms and causes. (2021, July 20). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms- causes/syc-20350800

Liu, J., MD. (2022a, July 19). What's a dangerous heart rate? What's a Dangerous Heart Rate? | Ohio State Health & Discovery. Retrieved December 5, 2022, from https://health.osu.edu/health/heart-and-vascular/what-is-dangerous-heart-rate

NCBI - WWW Error Blocked Diagnostic. (n.d.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159698/

NHS website. (2022, July 4). Low blood pressure (hypotension). nhs.uk. https://www.nhs.uk/conditions/low-blood-pressure-hypotension/

Tachycardia: Symptoms, Causes & Treatment. (n.d.). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/22108-tachycardia

Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE, comment = NA)

library(tidyverse)
library(reshape2)
library(faraway)
library(ggplot2)
library(mice)
library(caTools)
library(MASS)
library(corrplot)
library(car)
library(PRROC)
library(pROC)
library(kableExtra)

tabledata <- data.frame(matrix(c("Sex","Participant Sex (Male or Female)", "Age", "Age at exam (years)

# Specify the column names and join
colnames(tabledata) <- c("Variable", "Description")

# Select style and print table
kbl(tabledata) %>%
  kableExtra::kable_minimal(c( "striped"), full_width = F, position = "center")

#my_git_url <- getURL("https://raw.githubusercontent.com/peterphung2043/data_621_hw_1/main/FinalProject
heart_history <- read.csv("https://raw.githubusercontent.com/peterphung2043/data_621_hw_1/main/FinalPro
heart_history_original <- heart_history

dim(heart_history)
heart_history <- heart_history %>%
    mutate(
    Sex = as.factor(Sex),
    education = as.factor(education),
    BPMeds = as.factor(BPMeds),
```

```
    prevalentStroke = as.factor(prevalentStroke),
    prevalentHyp = as.factor(prevalentHyp),
    diabetes = as.factor(diabetes),
    TenYearCHD = as.factor(TenYearCHD),
    currentSmoker = as.factor(currentSmoker)
    )
heart_history_original <- heart_history %>%
    mutate(
    Sex = as.factor(Sex),
    education = as.factor(education),
    BPMeds = as.factor(BPMeds),
    prevalentStroke = as.factor(prevalentStroke),
    prevalentHyp = as.factor(prevalentHyp),
    diabetes = as.factor(diabetes),
    TenYearCHD = as.factor(TenYearCHD),
    currentSmoker = as.factor(currentSmoker)
    )

heart_history_original <- heart_history_original[complete.cases(heart_history_original), ]
summary(heart_history)
heart_history %>% dplyr::select(-TenYearCHD) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density(col = 'red') +
    geom_histogram(aes(y = stat(density)))

melt(heart_history) %>%
  ggplot(aes(x = TenYearCHD, y = value, fill = TenYearCHD)) +
  geom_boxplot() +
  facet_wrap(variable~., scales = "free")

melt(heart_history) %>%
  ggplot(aes(x = Sex, y = value, fill = Sex)) +
  geom_boxplot() +
  facet_wrap(variable~., scales = "free")

corrplot(cor(dplyr::select_if(heart_history, is.numeric), use = "na.or.complete"),
         method = 'number',
         type = 'lower',
         diag = FALSE,
         col = 'black',
         number.cex = 0.75,
         tl.cex = 0.5)
# Removal of diaBP
heart_history <- heart_history %>% dplyr::select(-diaBP)
tabl <- "
|Variable Name|Percentage of Missing Data|
|---------------|-------------:|
|education|2.48|
|cigsPerDay|0.68|
|BPMeds|1.25|
```

```
|totChol|1.18|
|BMI|0.45|
|heartRate|0.02|
|glucose|9.15|
"
cat(tabl) # output the table in a format good for HTML/PDF/docx conversion
heart_history  %>%
  summarise_all(list(~is.na(.)))%>%
  pivot_longer(everything(),
               names_to = "variables", values_to="missing") %>%
  count(variables, missing) %>%
  ggplot(aes(y=variables,x=n,fill=missing))+
  geom_col()+
  scale_fill_manual(values=c("skyblue3","gold"))+
  theme(axis.title.y=element_blank()) + theme_classic()

# Removing outlying sysBP values
heart_history <- heart_history[heart_history$sysBP < 160 & heart_history$sysBP > 100,]

## Removing outlying diaBP variables
# heart_history <- heart_history[heart_history$diaBP < 100 & heart_history$diaBP > 65,]

# Removing outlier heart rates
heart_history <- heart_history[heart_history$heartRate < 100 & heart_history$heartRate > 25,]

# Removing observations with NAs
heart_history <- heart_history[complete.cases(heart_history), ]
heart_history %>% dplyr::select(-TenYearCHD) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_density(col = 'red') +
    geom_histogram(aes(y = stat(density)))

tabl <- "
|Column Name|$\\lambda$|
|---------------|-------------:|
|BMI|-0.309|
|glucose|-1.279|
|totChol|0.069|

"
cat(tabl) # output the table in a format good for HTML/PDF/docx conversion
heart_history$cigsPerDay[heart_history$cigsPerDay == 0] <- 1e-6
skewed_vars <-   c("BMI", "glucose", "totChol")
lambdas <- powerTransform(eval(parse(text = paste("cbind(",toString(skewed_vars),")", "~ 1"))), heart_hi
transformed_data <- bcPower(lambdas$y, coef(lambdas))

colnames(transformed_data) <- sprintf("tf_%s", skewed_vars)
heart_history <- cbind(heart_history, transformed_data)
as.data.frame(transformed_data) %>%
  keep(is.numeric) %>%
```

```
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins = 35)
orig_tabl <- "
||Number of Observations where `TenYearCHD` = 1|Number of Observations where `TenYearCHD` = 0|
|---------------|---------------|-------------:|
|Test|930|167|
|Train|2171|390|

"

cat(orig_tabl)
mod_tabl <- "
||Number of Observations where `TenYearCHD` = 1|Number of Observations where `TenYearCHD` = 0|
|---------------|---------------|-------------:|
|Test|736|109|
|Train|1716|255|

"

cat(mod_tabl)
set.seed(123)

original_split <- sample.split(heart_history_original$TenYearCHD, SplitRatio = 0.7)
original_train <-  subset(heart_history_original, original_split == TRUE)
original_test <- subset(heart_history_original, original_split == FALSE)

modified_split <- sample.split(heart_history$TenYearCHD, SplitRatio = 0.7)
modified_train <-  subset(heart_history, modified_split == TRUE)
modified_test <- subset(heart_history, modified_split == FALSE)


# table(original_test$TenYearCHD)
# table(original_train$TenYearCHD)
#
# table(modified_test$TenYearCHD)
# table(modified_train$TenYearCHD)


# Accuracy Function
accuracy <- function(true_values, predictions){
  TP <- sum(true_values == 1 & predictions == 1)
  TN <- sum(true_values == 0 & predictions == 0)
  round((TP + TN)/length(true_values), 4)
}

## Classification error rate function
error <- function(true_values, predictions){
  FP <- sum(true_values == 0 & predictions == 1)
  FN <- sum(true_values == 1 & predictions == 0)
  round((FP + FN)/length(true_values), 4)
}
```

```r
## F-score function
fscore <- function(precision_score, recall_score) {
  (2* precision_score * recall_score)/(precision_score + recall_score)
}

## Precision function
##The precision contains 2 values corresponding to the classes 0, and 1.
##In binary classification tasks, we will look at the values of the
##positive class (1) for reporting metrics.
precision <- function(true_values, predictions){
  TP <- sum(true_values == 1 & predictions == 1)
  FP <- sum(true_values == 0 & predictions == 1)

  round(TP/(TP+FP), 4)
}

## Sensitivity function
##The sensitivity/recall contains 2 values corresponding to the classes 0, and 1.
##In binary classification tasks, we will look at the values of the positive
##class (1) for reporting metrics.
recall <- function(true_values, predictions){
  TP <- sum(true_values == 1 & predictions == 1)
  FN <- sum(true_values == 1 & predictions == 0)

  round(TP/(TP+FN), 4)
}

## ROC function for step 10
ROC <- function(x, y){
  x <- x[order(y, decreasing = TRUE)]
  TPR <- cumsum(x) / sum(x)
  FPR <- cumsum(!x) / sum(!x)
  df <- data.frame(TPR, FPR, x)

  FPR_df <- c(diff(df$FPR), 0)
  TPR_df <- c(diff(df$TPR), 0)
  area_under_curve <- sum(df$TPR * FPR_df) + sum(TPR_df * FPR_df)/2

  plot(df$FPR, df$TPR, type = "l",
       main = "ROC ",
       xlab = "FPR",
       ylab = "TPR")
  abline(a = 0, b = 1)
  legend("center", legend= c("AUC", round(area_under_curve, 4)))

}
#create data frame with 0 rows and 3 columns
tracker <- data.frame(matrix(ncol = 8, nrow = 0))

#provide column names
colnames(tracker) <- c("Model", "Precision", "Recall", "AIC", "AUC", "F-score", "Accuracy", "Error")

#create function to update the tracker
```

```r
update_tracker <- function(tracker, model_name, true_values, predictions, model_object, df){
  accuracy = accuracy(true_values, predictions)
  error = error(true_values, predictions)
  recall = recall(true_values, predictions)
  precision = precision(true_values, predictions)
  aic = model_object$aic
  auc = as.numeric(str_extract(roc(true_values,  predict(model_object, df, interval = "prediction"))$au
  f_score = fscore(precision, recall)


  tracker[nrow(tracker) + 1,] <- c(model_name, round(precision, 2), round(recall, 2), round(aic, 2), ro
  return(tracker)
}

bin_log_orig <- glm(TenYearCHD ~ ., data = original_train, family = binomial(link = "logit"))
summary(bin_log_orig)

bin_log_orig_predictions <- predict.glm(bin_log_orig, original_test, type = "response")

bin_log_orig_predictions_binary <- ifelse(bin_log_orig_predictions > 0.5, 1, 0)

tracker <- update_tracker(tracker, "Bin. Log. w/ Original Data", original_test$TenYearCHD, bin_log_orig
bin_log_modified <- glm(TenYearCHD ~ ., data = modified_train, family = binomial(link = "logit"))
summary(bin_log_modified)

bin_log_modified_predictions <- predict.glm(bin_log_modified, modified_test, type = "response")

bin_log_modified_predictions_binary <- ifelse(bin_log_modified_predictions > 0.5, 1, 0)

tracker <- update_tracker(tracker, "Bin. Log. w/ Modified Data", modified_test$TenYearCHD, bin_log_modi
step_aic_bin_log_orig <- stepAIC(bin_log_orig, direction = "both", trace = FALSE)
summary(step_aic_bin_log_orig)

step_aic_bin_log_orig_predictions <- predict.glm(step_aic_bin_log_orig, original_test, type = "response
step_aic_bin_log_orig_predictions_binary <- ifelse(step_aic_bin_log_orig_predictions > 0.5, 1, 0)

tracker <- update_tracker(tracker, "Step AIC Bin. Log. w/ Original Data", original_test$TenYearCHD, ste
step_aic_bin_log_modified <- stepAIC(bin_log_modified, direction = "both", trace = FALSE)
summary(step_aic_bin_log_modified)

step_aic_bin_log_modified_predictions <- predict.glm(step_aic_bin_log_modified, modified_test, type = "
step_aic_bin_log_modified_predictions_binary <- ifelse(step_aic_bin_log_modified_predictions > 0.5, 1, 

tracker <- update_tracker(tracker, "Step AIC Bin. Log. w/ Modified Data", modified_test$TenYearCHD, ste
knitr::kable(tracker)
# ggplot(tracker, aes(x=factor(Model, level=c('Simple', 'Transformed', 'Negative Bimodal', 'Reduced Tra
#   geom_bar(stat = "identity") +
#   ylab("Precision") +
#   xlab("Model") +
#   theme(axis.text.x = element_text(angle = 90))

plt <- melt(tracker[,colnames(tracker)],id.vars = 1)
```

```
ggplot(plt, aes(x=factor(Model, level=tracker$Model), y = value)) +
  geom_bar(aes(fill = variable),stat = "identity",position = "dodge") +
  xlab("Model") +
  ylab("Score") +
  theme(axis.text.x = element_text(angle = 90))
plot(roc(original_test$TenYearCHD,  predict(step_aic_bin_log_orig, original_test, interval = "predictio
```