Predictive Analytics for Heart Disease: A Comprehensive Analysis of Risk Factors


| | email | Contribution | SID |
|---|---|---|---|
| Jieyi Xu | jieyixu@uw.edu | Model selection, relationship between heart attack risk and smoking | 2066532 |
| Fangzhou Xie | xfz2020@uw.edu | Logistic regression 1, heart rate distribution, income distribution | 2063441 |
| Tao Zhang | tzhang39@uw.edu | Analysis of dataset, Logistic regression with changes, ROC curve | 2138482 |
| Gefei Shen | gefeis3@uw.edu | Relationship between risk and age, heart rate and exercise hour | 2163384 |

Department of Statistics, University of Washington


STAT423: Applied Regression and Analysis of Variance

Dr. Emanuela Furfaro

Mar 2024

# Contents

# 1. Introduction

In today's society, heart attack has become a major health problem on a global scale;  its morbidity and mortality continue to be high. This dataset was designed to collect and analyze a variety of information related to heart disease risk, including patient demographics, health indicators, lifestyle habits, and geographic location. Through in-depth analysis of these data, researchers can better understand which factors may increase the risk of heart disease, thereby providing a scientific basis for heart disease prevention and intervention. This will not only improve the quality of life of patients, but also reduce the burden on the healthcare system, which has a positive impact on society and the economy.

## 1.1 Dataset Introduction

This dataset is from Kaggle(2023) and was updated five months ago. It contains 8763 observations, a response variable and 24 meaningful independent variables. The response variable is heart attack risk. The independent variables range from categorical variables such as continent, country, and smoking to continuous variables such as hours of exercise per week and BMI. More details of each variable can be found in the appendix.

## 1.2 Research Questions

1. Which factors are significant in predicting the diagnosis?
2. At the 5 levels of significance, is there a statistically significant association between heart attack risk (response variable) and smoking, controlling for diet (categorical variable)? At the 5 levels of significance, is there a statistically significant association between heart attack risk (response variable) and diabetes, controlling for sex?
3. Do age, heart rate, and exercise hours per week (numerical variable) have an effect on heart attack risk (response variable)?
4. What is the predictive model for heart attack risk?

# 2. Methodologies

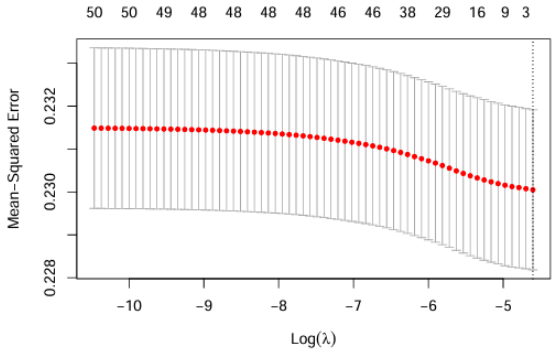
## 2.1 Model Selection

In this analysis, we would utilize forward-step selection to determine the optimal set of parameters for our model. This process will progressively add predictors from the empty model to the full model. We would choose the set of parameters that would generate the lowest Akaike Information Criterion (AIC) value. From Figure 1 below, we can see that the variables we should use are cholesterol, sleep hours per day, higher blood pressure limit, and whether or not the patient has diabetes. However, the AIC for the best model is very large, making us think that this cannot be a good model. Adding or removing any variables to the best model will also not cause a significant change in the AIC.

```
Step:  AIC=11985.61
Heart.Attack.Risk ~ Cholesterol + Sleep.Hours.Per.Day + blood_high +
    Diabetes
```

(Figure 1 Forward Step Selection Result)

Concerning the high AIC value, we plan to perform a LASSO regression to see which variables will be chosen in the model. Here, we propose using LASSO because unlike forward stepwise, it eliminates all variables that have a small correlation with the response variable, only choosing variables that have a convincingly clear correlation. The result is shown in Figure 2. We conclude that none of the given variables correlates with the outcome. Figure 3 also shows that the less weight we put towards each variable during the regression, the better result we gain, meaning that the variables do not have a strong effect on the outcome.




```
## (Intercept)                      0.3582107
## (Intercept)                      .
## Age                              .
## SexMale                          .
## Cholesterol                      .
## Heart.Rate                       .
## Diabetes1                        .
## Family.History1                  .
## Smoking1                         .
## Obesity1                         .
## Alcohol.Consumption1             .
## Exercise.Hours.Per.Week          .
## DietHealthy                      .
## DietUnhealthy                    .
## Previous.Heart.Problems1         .
## Medication.Use1                  .
## Stress.Level2                    .
## Stress.Level3                    .
## Stress.Level4                    .
## Stress.Level5                    .
## Stress.Level6                    .
## Stress.Level7                    .
## Stress.Level8                    .
## Stress.Level9                    .
## Stress.Level10                   .
## Sedentary.Hours.Per.Day          .
## Income                           .
## BMI                              .
## Triglycerides                    .
## Physical.Activity.Days.Per.Week  .
## Sleep.Hours.Per.Day              .
## CountryAustralia                 .
## CountryBrazil                    .
## CountryCanada                    .
## CountryChina                     .
## CountryColombia                  .
## CountryFrance                    .
## CountryGermany                   .
## CountryIndia                     .
## CountryItaly                     .
## CountryJapan                     .
## CountryNew Zealand               .
## CountryNigeria                   .
## CountrySouth Africa              .
## CountrySouth Korea               .
## CountrySpain                     .
## CountryThailand                  .
## CountryUnited Kingdom            .
## CountryUnited States             .
## CountryVietnam                   .
## ContinentAsia                    .
## ContinentAustralia               .
## ContinentEurope                  .
## ContinentNorth America           .
```

(Figure 2: result table of LASSO regularization, as shown right)
(Figure 3: plot between regression MSE and log(λ), as shown above)

## 2.2 Analysis of the Relationship

### 2.21 Relationship between Heart Attack Risk (HAR) and Smoking

In this section, we find if there is a relationship between heart attack risk and various lifestyle factors, especially smoking habits, while adjusting for dietary patterns. To be more specific, we utilized stratified analyses by segmentation of data based on dietary categories: Healthy, Average, and Unhealthy. From Figure 2, there is not a statistically significant relationship between heart attack risk and smoking within the control of Diet. To determine whether Diet acts as a confounder in the relationship between smoking and heart attack risk, we applied the Cochran Mantel

Haenszel (CMH) Test. As indicated in Figure 3, with a p-value of 0.7364, which is larger than the significance level of 0.1, we do not have sufficient evidence to reject the null hypothesis. This suggests that there is no significant evidence that Diet is a confounding variable that affects the relationship between smoking and heart attack risk. Because the p-value was greater than 0.05, we concluded that the dietary habits of the different diets did not cause an effect. Then, since both smoking and risk are categorical variables, we used the chi-square test to find if there is a relationship between smoking and risky tests for different diets. (In fact, if we pass the Mantelhaen.test, we do not need to test for different diets, but use smoking and risk, again to prevent potential effects of diet). Then, according to Figure 4,

(i) H_0: there is no association between variables smoking and Heart.Attack.Risk among Healthy Diet people.
H_a: there is an association between variables smoking and Heart.Attack.Risk among Healthy Diet people.

In **Healthy diet** people sample, from the chi-squared test, with p-value > 0.1, so we failed to reject the H_0 and there is not sufficient evidence to support that there is an association between variables smoking and heart attack risk.

(ii) H_0: there is no association between variables smoking and Heart.Attack.Risk among Average Healthy diet people.
H_a: there is an association between variables smoking and Heart.Attack.Risk among Average Healthy diet people.

In **Average Healthy diet** people sample, from the chi-squared test, with p-value > 0.1, so we fail to reject the H_0 and there is not sufficient evidence to support that there is an association between variables smoking and heart attack risk.

(iii) Healthy diet
H_0: there is no association between variables smoking and heart attack risk among unhealthy diet people.
H_a: there is an association between variables smoking and heart attack risk among unhealthy diet people.

In **Unhealthy Diet** people sample, from the chi-squared test, with p-value > 0.1, so we fail to reject the H_0 and there is not sufficient evidence to support that there is an association between variables smoking and heart attack risk. They are all irrelevant. We then tested heart attack risk and Diabetes, controlling for Sex. Sex went to test the confounding variables, and then measured the effect of heart attack risk and Diabetes, and there was no relationship. Our initial thought was

that there might be some things that could be improved with this dataset. We potentially think that they may not be randomly sampling the data.


(Figure 4: Proportional Distribution of Heart Attack Risk by Smoking Status Across Dietary Groups)

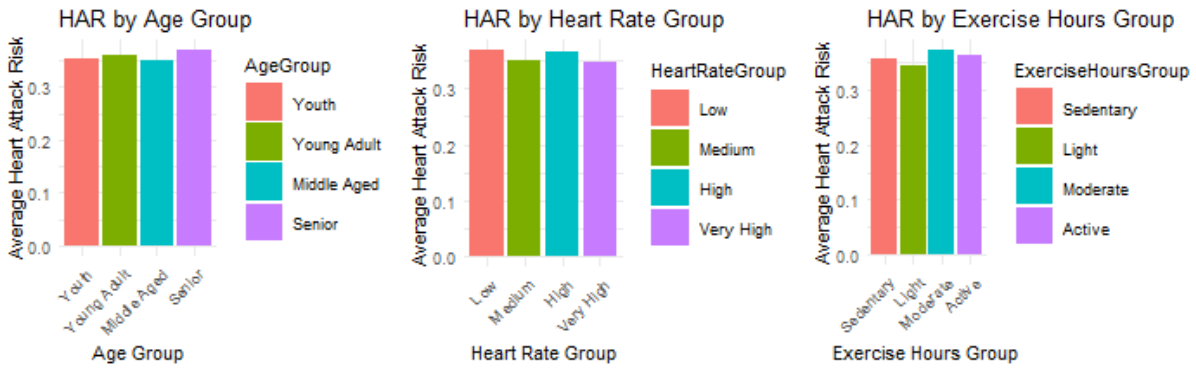
```
           not smoking smoking
  not risky         183    1703
  risky             103     923

, ,  = Healthy


           not smoking smoking
  not risky         186    1695
  risky             120     959

, ,  = Unhealthy


           not smoking smoking
  not risky         206    1651
  risky             106     928


        Mantel-Haenszel chi-squared test with continuity correction

data:  data1$Heart.Attack.Risk and data1$Smoking and data1$Diet
Mantel-Haenszel X-squared = 0.11333, df = 1, p-value = 0.7364
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.8433229 1.1228579
sample estimates:
common odds ratio
        0.9731042
```

（Figure 5: Result of CMH test ）

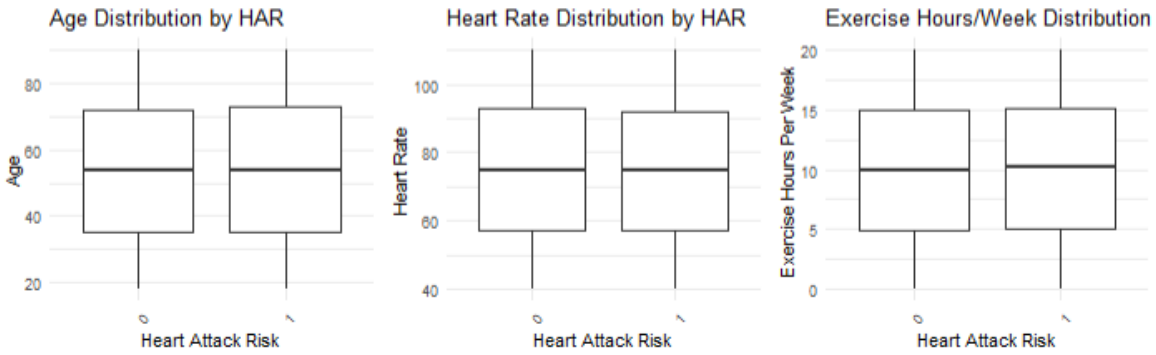
```
        Pearson's Chi-squared test with Yates' continuity correction

data:  Health$Heart.Attack.Risk and Health$Smoking
X-squared = 0.99562, df = 1, p-value = 0.3184


        Pearson's Chi-squared test with Yates' continuity correction

data:  Average$Heart.Attack.Risk and Average$Smoking
X-squared = 0.050979, df = 1, p-value = 0.8214


        Pearson's Chi-squared test with Yates' continuity correction

data:  Unhealth$Heart.Attack.Risk and Unhealth$Smoking
X-squared = 0.40525, df = 1, p-value = 0.5244
```

(Figure 6: Chi-squared Test)

## 2.22 Relationship between Heart Attack Risk( HAR) and Other Factors




(Figure 7: Comparative Analysis of Heart Attack Risk Across Different Demographic and Behavioral Groups)




(Figure 8: Boxplot Distribution of Age, Heart Rate, and Exercise Hours per Week by Heart Attack Risk)

Initially, we used the cut() function to categorize continuous variables into predefined groups for analysis. The method has segmented the population into distinct groups based on age, heart rate, and exercise hours per week. For age, individuals are categorized into 'Youth', 'Young Adult',

'Middle Aged', and 'Senior' groups. Heart rates are classified into 'Low', 'Medium', 'High', and 'Very High' categories. Exercise hours per week are divided into 'Sedentary', 'Light', 'Moderate', and 'Active' levels of physical activity. The bar charts display the average heart attack risk within these groups, providing a visual comparison of how each factor may be related to heart attack risk. Then, To construct the visualizations, we created bar charts where the height of each bar corresponds to the mean heart attack risk within each category. These representations allow us to observe potential patterns and draw comparisons across different demographic and behavioral segments. As shown in Figure 5 above, the charts appear to show a relatively uniform distribution of heart attack risk across the different categories within each group, without any striking differences. This could suggest that within the granularity of the chosen groupings, there appears to be a weak relationship between the categorized age, heart rate, or exercise level and the average heart attack risk.

Similarly, we also used Box-Plot to visualize the distribution of age, heart rate, and exercise hours per heart attack risk, as displayed in Figure 6. The overlapping IQRs suggest there may not be a significant difference in the distribution of age, heart rate, and exercise hours between the different heart attack risk groups. This indicates that, in isolation, none of these factors show a distinct correlation with heart attack risk within the observed dataset. Further analysis could involve investigating these variables in combination or with additional factors to identify more complex relationships.

## 2.3 Logistic regression

We used many variables and decided to use logistic regression to test which variables were significant in the first place. Since there are two categorical variables, stress level and country, we use the ANOVA test to test whether these two categorical variables are significant.

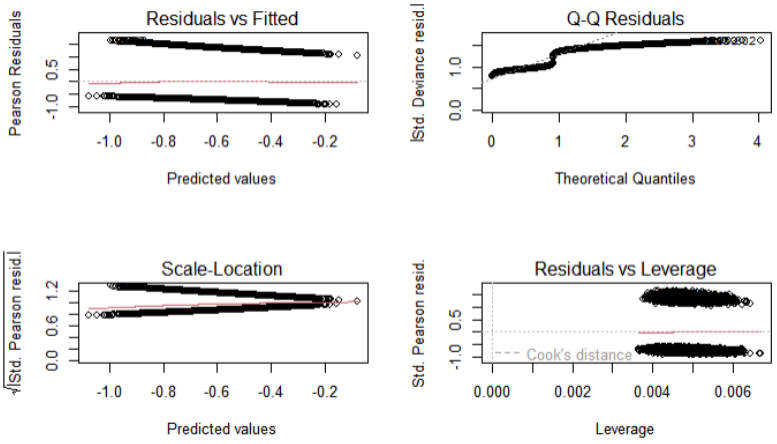
```
Analysis of Deviance Table

Model 1: as.factor(Heart.Attack.Risk) ~ Age + Sex + Cholesterol + Heart.Rate +
    Diabetes + Family.History + Smoking + Obesity + Alcohol.Consumption +
    Exercise.Hours.Per.Week + as.factor(Stress.Level) + Income +
    BMI + Triglycerides + Country
Model 2: as.factor(Heart.Attack.Risk) ~ Obesity + Alcohol.Consumption +
    Exercise.Hours.Per.Week + Stress.Level + Income + BMI + Triglycerides
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1      8721      11395
2      8755      11427 -34  -32.574   0.5375
Analysis of Deviance Table

Model 1: as.factor(Heart.Attack.Risk) ~ Age + Sex + Cholesterol + Heart.Rate +
    Diabetes + Family.History + Smoking + Obesity + Alcohol.Consumption +
    Exercise.Hours.Per.Week + as.factor(Stress.Level) + Income +
    BMI + Triglycerides + Country
Model 2: as.factor(Heart.Attack.Risk) ~ Age + Sex + Cholesterol + Heart.Rate +
    Diabetes + Family.History + Smoking + Obesity + Alcohol.Consumption +
    Exercise.Hours.Per.Week + Income + BMI + Triglycerides +
    Country
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      8721      11395
2      8730      11400 -9  -5.0326   0.8315
```

(continued to next page)

```
Call:
glm(formula = as.factor(Heart.Attack.Risk) ~ Age + Sex + Cholesterol +
    Heart.Rate + Diabetes + Family.History + Smoking + Obesity +
    Alcohol.Consumption + Exercise.Hours.Per.Week + as.factor(Stress.Level) +
    Income + BMI + Triglycerides + Country, family = binomial,
    data = data)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -7.557e-01  2.227e-01  -3.393  0.00069 ***
Age                         1.128e-03  1.179e-03   0.957  0.33863
SexMale                     4.278e-02  5.856e-02   0.730  0.46510
Cholesterol                 4.772e-04  2.770e-04   1.723  0.08492 .
Heart.Rate                 -4.243e-04  1.088e-03  -0.390  0.69657
Diabetes                    8.123e-02  4.713e-02   1.724  0.08480 .
Family.History             -8.997e-03  4.477e-02  -0.201  0.84072
Smoking                    -9.209e-02  9.595e-02  -0.960  0.33715
Obesity                    -5.877e-02  4.475e-02  -1.313  0.18909
Alcohol.Consumption        -5.548e-02  4.558e-02  -1.217  0.22353
Exercise.Hours.Per.Week     4.114e-03  3.868e-03   1.064  0.28748
as.factor(Stress.Level)2    3.700e-02  9.951e-02   0.372  0.71002
as.factor(Stress.Level)3    5.704e-02  1.006e-01   0.567  0.57086
as.factor(Stress.Level)4   -4.275e-02  1.002e-01  -0.427  0.66946
as.factor(Stress.Level)5    5.828e-02  1.008e-01   0.578  0.56303
as.factor(Stress.Level)6    1.058e-01  1.007e-01   1.051  0.29321
as.factor(Stress.Level)7    8.149e-02  9.953e-02   0.819  0.41290
as.factor(Stress.Level)8    1.165e-02  1.006e-01   0.116  0.90789
as.factor(Stress.Level)9   -5.199e-02  1.008e-01  -0.516  0.60619
as.factor(Stress.Level)10  -1.729e-02  1.025e-01  -0.169  0.86612
Income                      2.555e-07  2.777e-07   0.920  0.35756
BMI                        -2.356e-04  3.540e-03  -0.067  0.94693
Triglycerides               8.964e-05  9.993e-05   0.897  0.36968
CountryAustralia            1.728e-02  1.368e-01   0.126  0.89947
CountryBrazil              -6.915e-02  1.366e-01  -0.506  0.61284
CountryCanada              -4.618e-02  1.380e-01  -0.335  0.73798
CountryChina               -6.081e-02  1.386e-01  -0.439  0.66089
CountryColombia             3.899e-02  1.382e-01   0.282  0.77778
CountryFrance              -8.064e-02  1.379e-01  -0.585  0.55874
CountryGermany             -3.874e-02  1.352e-01  -0.287  0.77446
CountryIndia               -2.504e-01  1.431e-01  -1.749  0.08026 .
CountryItaly               -2.457e-01  1.411e-01  -1.741  0.08175 .
CountryJapan               -1.694e-01  1.400e-01  -1.210  0.22619
CountryNew Zealand         -9.668e-02  1.391e-01  -0.695  0.48710
CountryNigeria              1.145e-01  1.361e-01   0.841  0.40013
CountrySouth Africa        -1.348e-01  1.403e-01  -0.961  0.33674
CountrySouth Korea          1.197e-01  1.392e-01   0.860  0.38999
CountrySpain               -9.304e-02  1.394e-01  -0.667  0.50451
CountryThailand             2.551e-02  1.383e-01   0.184  0.85369
CountryUnited Kingdom      -8.335e-02  1.371e-01  -0.608  0.54327
```

(Table 1: Logistic Regression Summary Table)

According to Table 1, we find that neither of these two categorical variables is significant. Since none of the variables are significant, we use all of them for now and transform the value variables. We use a number of methods, including logistic transformation, but almost none of the variables have been changed. So we use the original model; that is, we use all the variables and plot their residual in the untransformed state.



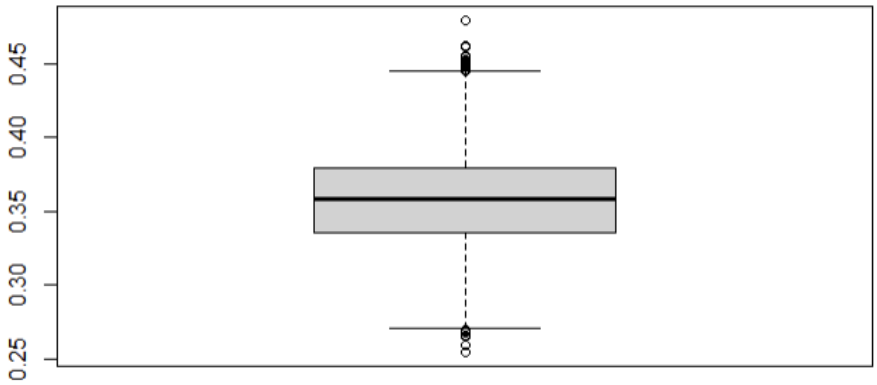

(Figure 9: Diagnostic Plots for Regression Model Analysis)

However, the residual vs. fitted plot shown in Figure 9 has no meaning in logistic regression, driving us to divide those who predicted correctly and those who predicted incorrectly into two groups and perform a logistic regression again as shown in Table 2. However, none of the variables in their model were significant for those who succeeded and those who failed, and the p-value is very large.

```
Call:
glm(formula = as.factor(Heart.Attack.Risk) ~ Age + Sex + Cholesterc
    Heart.Rate + Diabetes + Family.History + Smoking + Obesity +
    Alcohol.Consumption + Exercise.Hours.Per.Week + Stress.Level +
    Income + BMI + Triglycerides + Country, family = binomial,
    data = success)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.123e+03  1.074e+04   0.104    0.917
Age                     -1.547e+01  1.433e+02  -0.108    0.914
SexMale                 -6.448e+02  5.968e+03  -0.108    0.914
Cholesterol             -6.793e+00  6.301e+01  -0.108    0.914
Heart.Rate               5.286e+00  5.113e+01   0.103    0.918
Diabetes                -1.210e+03  1.125e+04  -0.108    0.914
Family.History           1.914e+02  1.880e+03   0.102    0.919
Smoking                  1.408e+03  1.303e+04   0.108    0.914
Obesity                  8.541e+02  7.894e+03   0.108    0.914
Alcohol.Consumption      9.131e+02  8.480e+03   0.108    0.914
Exercise.Hours.Per.Week -6.078e+01  5.635e+02  -0.108    0.914
Stress.Level             6.185e+01  5.859e+02   0.106    0.916
Income                  -3.868e-03  3.589e-02  -0.108    0.914
BMI                      3.203e+00  4.649e+01   0.069    0.945
Triglycerides           -1.421e+00  1.323e+01  -0.107    0.914
CountryAustralia        -4.195e+02  4.166e+03  -0.101    0.920
CountryBrazil            8.923e+02  8.425e+03   0.106    0.916
CountryCanada            5.315e+02  5.078e+03   0.105    0.917
CountryChina             7.543e+02  8.316e+03   0.091    0.928
CountryColombia         -5.989e+02  6.448e+03  -0.093    0.926
CountryFrance            9.762e+02  9.242e+03   0.106    0.916
CountryGermany           3.774e+02  4.115e+03   0.092    0.927
CountryIndia             4.280e+03  4.254e+04   0.101    0.920
CountryItaly             3.418e+03  9.355e+04   0.037    0.971
CountryJapan             2.422e+03  2.514e+04   0.096    0.923
CountryNew Zealand       1.217e+03  1.263e+04   0.096    0.923
CountryNigeria          -1.889e+03  1.962e+04  -0.096    0.923
CountrySouth Africa      1.744e+03  1.636e+04   0.107    0.915
CountrySouth Korea      -1.492e+03  1.696e+04  -0.088    0.930
CountrySpain             1.247e+03  1.187e+04   0.105    0.916
CountryThailand         -4.459e+02  4.367e+03  -0.102    0.919
CountryUnited Kingdom    9.977e+02  9.375e+03   0.106    0.915
CountryUnited States    -1.382e+03  1.572e+04  -0.088    0.930
CountryVietnam           1.153e+03  1.080e+04   0.107    0.915
```

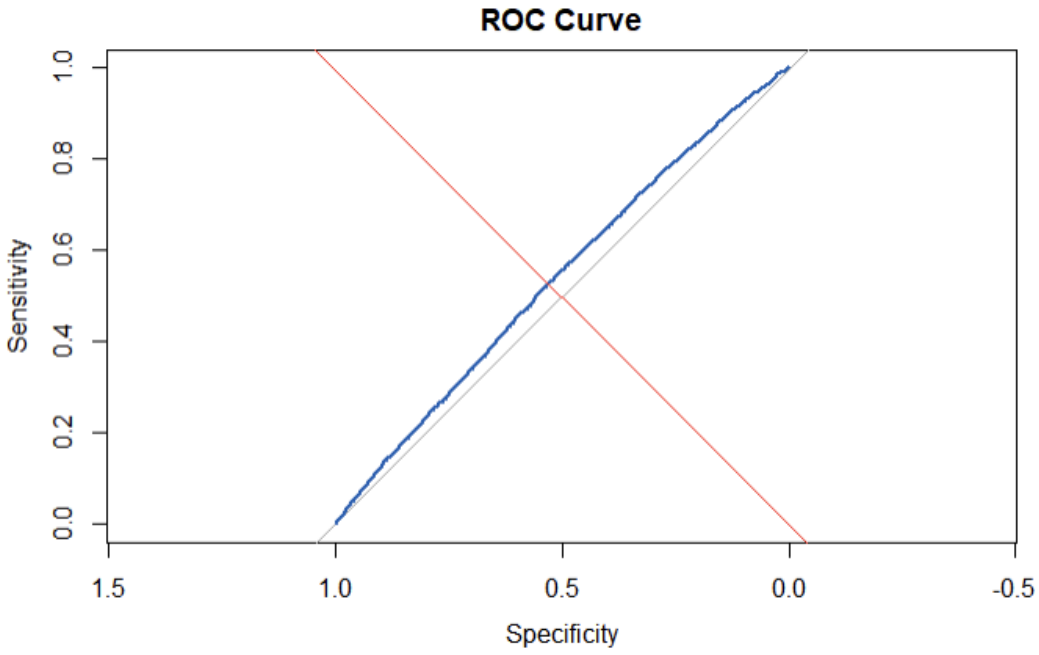
（Table 2: Logistic Regression Model with Changes）

We explore the results shown in Figure 10. Due to the fact that using our model results in a probability distribution of model predictions ranging from 0.3 to 0.45, so basically all the data is predicted to be not risky.



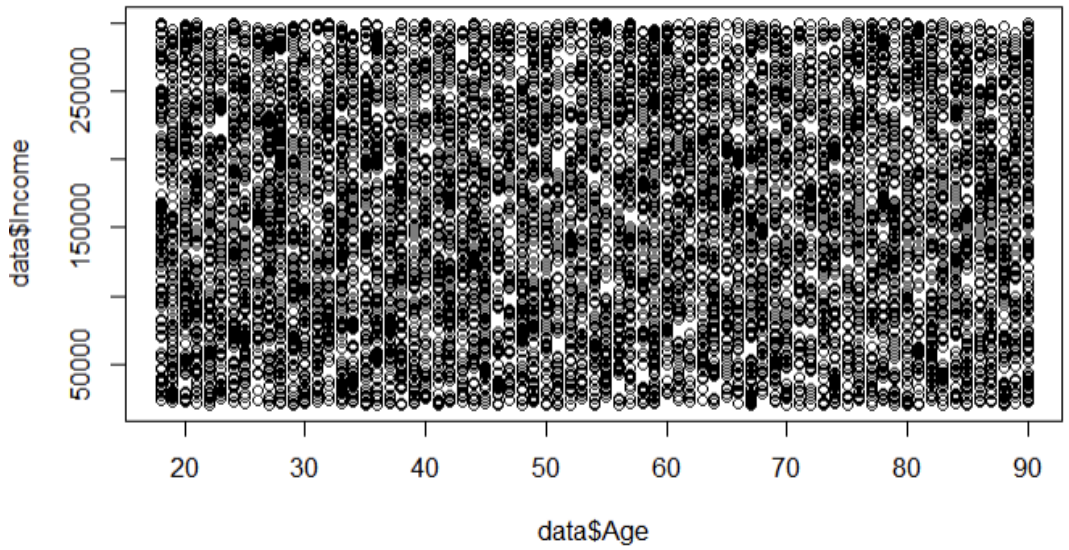

(Figure 10: Boxplot Plots of Probability Predictions)

We explored the ROC curve shown in Figure 11, which was odd, with an area under the curve of 0.5365, meaning our model was only correct about 50% of the time. This means the model takes a random guess. We summarized the reasons, and finally, we guessed that this may be a fake dataset or that this is definitely not a random dataset.



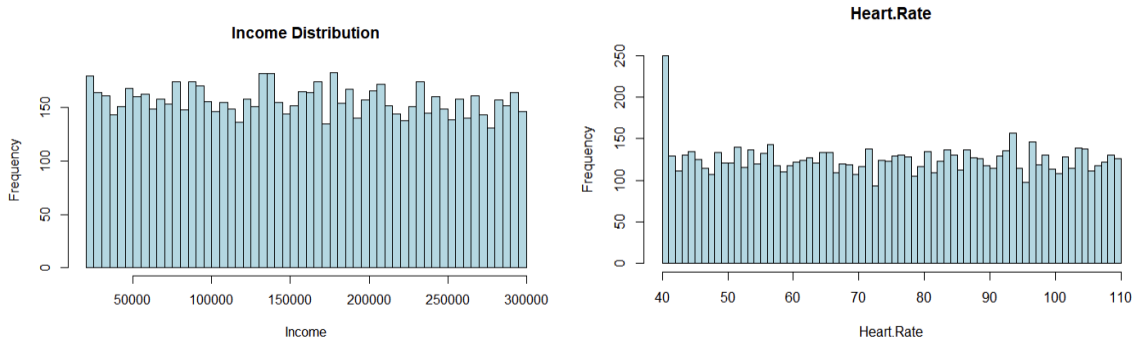

(Figure 11: ROC Curve for Model Performance Evaluation)

We try to explore the data set. There seems to be no relationship between age and income in this dataset, according to Figure 12.




(Figure 12: Scatter Plot of Income Distribution Across Different Ages)

The results shown in Figure 13 contradict what one would typically expect, as middle-aged individuals generally have higher earnings compared to their younger years. Furthermore, when examining income distribution in Figure 13, we found that it shows a uniform pattern, which does not make sense. In addition to income, we investigated factors such as heart rate in Figure 13, which surprisingly showed uniform distributions as well. Notably, a significant number of individuals have a heart rate of 40, a finding that also defies logical expectations.




(Figure 13: Histograms of Income Distribution and Heart Rate Frequencies)

## 3. Conclusion and Discussion

We can basically say that humans manually create this data set after using all tests, model selection, and logistic regression. This actually makes sense because, in the real world, we would not have that much real data accessible when it concerns public health and safety, especially for patient data. Fabricating data for reports related to diseases could have fatal consequences for patients. It is imperative to refrain from resorting to false data in such critical matters.

# 4. Appendix

| Variable | Describe | Type | Content |
|---|---|---|---|
| Age | Age of the patient | Ordinal and Numerical | |
| Sex | Gender of the patient | Categorical | Male/Female |
| Cholesterol | Cholesterol levels of the patient | Numerical | |
| Blood Pressure | Blood pressure of the patient | Numerical | systolic/diastolic |
| Heart Rate | Heart rate of the patient | Numerical | |
| Diabetes | Whether the patient has diabetes | Categorical | Yes/No) |
| Family History | Family history of heart-related problems | Categorical | 1: Yes, 0: No |
| Smoking | Smoking status of the patient | Categorical | 1: Smoker, 0: Non-smoker |
| Obesity | Obesity status of the patient | Categorical | 1: Obese, 0: Not obese |
| Alcohol Consumption | Level of alcohol consumption by the patient | Categorical | None/Light/Moderate/Heavy |
| Exercise Hours Per Week | Number of exercise hours per week | Numerical | |
| Diet | Dietary habits of the patient | Categorical | Healthy/Average/Unhealthy |
| Previous Heart Problems | Previous heart problems of the patient | Categorical | 1: Yes, 0: No |
| Medication Use | Medication usage by the patient | Categorical | 1: Yes, 0: No |
| Stress Level | Stress level reported by the patient | Categorical/Numerical | 1-10 |
| Sedentary Hours Per Day | Hours of sedentary activity per day | Numerical | |
| Income | Income level of the patient | Numerical | |
| BMI | Body Mass Index (BMI) of the patient | Numerical | |

Reply to proposal:

1. We did not use Bayesian Model Averaging because of the lack of association between independent and response variables.

2. Because forward stepwise results in a large AIC and only small change in AIC by adding or subtracting variables, we decided to use LASSO to discover the important variables

3. We have used stacked bar charts to visually demonstrate the relationship between categorical variables.

4. We employed the Mantel-Haenszel test to determine if a variable is a confounder, and the Chi-square test to assess whether there is an association between categorical variables.

5. We do not intend to use the Box-Cox transformation. Instead, we plan to divide the numerical variables into several categorical variables using the cut function, and then use bar plots to examine their relationship. Additionally, we will use box plots for numerical and categorical variables to observe their relationship.

6.We are not planning to review TA plots and QQ plots; rather, we are preparing to look at the relationship between residuals and leverages, and attempt to remove outliers and high leverage points to help refine the logistic regression model.