Modeling Heart Attack Risk

Alice Ding
DATA621
Fall 2023

Agenda

Introduction

Problem Statement

Literature Review

Methodology

Results

Conclusions

Introduction

- Heart attacks occur when blood flow to part of the heart muscle is blocked, often due to the formation of a blood clot or other health issues within the body

- Myocardial infarctions, or heart attacks, have grown increasingly problematic within the eyes of global and public health

- By shedding light on the web of factors contributing heart attacks, we can empower individuals, healthcare professionals, and policy makers with the knowledge to implement proactive and preventative measures against this issue

Problem Statement

- The purpose of this project is to see if there is a way to predict whether a patient is at risk of having a heart attack given certain attributes ranging from:
    - Age
    - Cholesterol levels
    - Blood pressure
    - Smoking habits
    - Exercise patterns
    - Income
    - And more

Literature Review

Adhikary, Barman, Ranjan, & Stone (2022)

- Different countries had a higher prevalence of CVD (cardiovascular diseases) due to the coexistence of multiple risk factors that were present in certain areas compared to others
- "... potential risk factors for coronary artery diseases are hypertension, obesity, and physical inactivity."
- This analysis focuses only on English-language papers and the available ones they chose from are mainly from Asia, Europe, Africa, and the United States

Berry, Dyer, Cai, Garside, Ning, Thomas, Greenland, Van Horn, Tracy, & Lloyd-Jones (2012)

- Those with a total cholesterol level of < 180 mg per deciliter, < 120 mm Hg systolic and 80 mm Hg diastolic blood pressure, and those with nonsmoking and nondiabetic statuses have substantially lower risks of death from cardiovascular disease through the age of 80 years than those with two or more major risk factors
- This analysis however was only done for black and white individuals at the ages of 45, 55, 65, and 75, thus missing a large portion of the global population

Fuchs and Whelton (2019)

- As a society, we have been artificially inflating what is deemed as a normal blood pressure (BP) and thus underrepresenting the impact of what high BP on the likelihood of cardiovascular disease (CVD)
- This hypothesis does its best to meet Occam's razor promise (has the least amount of assumptions)
- The paper focuses just on one variable that can lead to heart attacks rather than a myriad of them, although it does address that CVD is impacted by other factors

Methodology

- Taking a Kaggle dataset of patient information for almost 9000 individuals, this source was explored, cleaned, prepared, and modeled on to predict heart attack risk

- Exploration was conducted by graphing and showing distributions and correlations for all of the fields individually and in relation to the target variable
  - It should be noted that a majority of our sample size came from Asia and Europe, leading to a majority of our participants being located in the Northern Hemisphere
  - Other disproportionate categorical variable observations were a large proportion of our population being non-smoking, diabetic, and male. Whether this is proportional to the current state of the entire world's population, this is unknown.

- Data preparation occurred to address nulls or outliers if they existed and normalization for numerical fields was applied as well

- Two sets of models were created after splitting the training data
  - One using binomial logistic regression
  - The other using a decision tree

Results: Binomial Regression

- Two iterations, one with all variables and the second with only those that were statistically significant

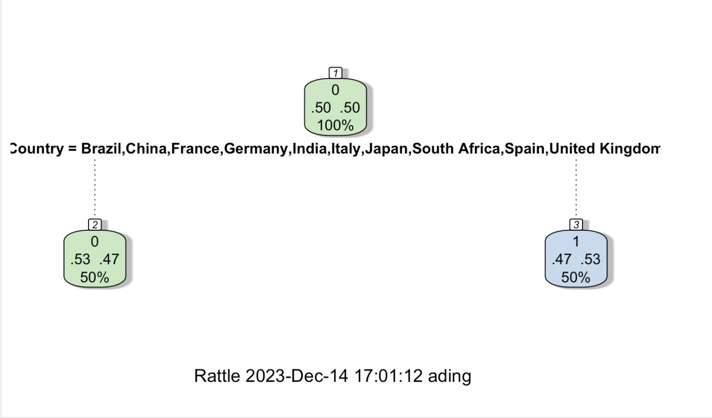
- AIC slightly improved on the second iteration, however residual deviance increased slightly

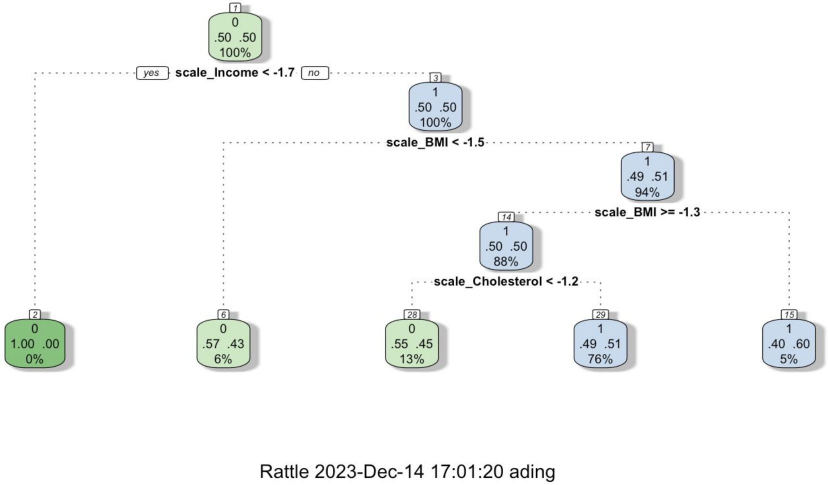
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4633  -1.1702   0.0011   1.1695   1.4582

Coefficients: (6 not defined because of singularities)
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                        0.053534   0.139266   0.384   0.7007
scale_Age                          0.009552   0.025472   0.375   0.7077
scale_Cholesterol                  0.045511   0.022767   1.999   0.0456 *
scale_Systolic                     0.014525   0.022679   0.640   0.5219
scale_Diastolic                   -0.045992   0.022855  -2.012   0.0442 *
scale_Heart.Rate                  -0.036065   0.022717  -1.588   0.1124
scale_Exercise.Hours.Per.Week      0.024426   0.022741   1.074   0.2828
scale_Stress.Level                -0.015471   0.022807  -0.678   0.4975
scale_Sedentary.Hours.Per.Day     -0.019075   0.022676  -0.841   0.4002
scale_Income                       0.025201   0.022768   1.107   0.2683
scale_BMI                          0.021018   0.022661   0.927   0.3537
scale_Triglycerides                0.033461   0.022761   1.470   0.1415
scale_Physical.Activity.Days.Per.Week -0.008817   0.022876  -0.385   0.6999
scale_Sleep.Hours.Per.Day         -0.057157   0.022748  -2.513   0.0120 *
SexMale                            0.066304   0.059397   1.116   0.2643
Diabetes                           0.086989   0.047924   1.815   0.0695 .
Family.History                     0.001359   0.045434   0.030   0.9761
Smoking                           -0.015688   0.097074  -0.162   0.8716
Obesity                           -0.077636   0.045375  -1.711   0.0871 .
Alcohol.Consumption               -0.113210   0.046242  -2.448   0.0144 *
DietHealthy                        0.048182   0.055632   0.866   0.3865
DietUnhealthy                      0.052812   0.055738   0.948   0.3434
Previous.Heart.Problems            0.007883   0.045442   0.173   0.8623
Medication.Use                     0.037574   0.045438   0.827   0.4083
CountryAustralia                  -0.001026   0.139269  -0.007   0.9941
CountryBrazil                     -0.158623   0.138870  -1.142   0.2534
CountryCanada                     -0.065960   0.137993  -0.478   0.6327
CountryChina                      -0.189355   0.139951  -1.353   0.1761
CountryColombia                    0.108872   0.140201   0.777   0.4374
CountryFrance                     -0.284705   0.140916  -2.020   0.0433 cm1a
CountryGermany                    -0.124688   0.136523  -0.913   0.3611
CountryIndia                      -0.351068   0.143135  -2.453   0.0142 *
CountryItaly                      -0.243914   0.140499  -1.736   0.0826 .
CountryJapan                      -0.215431   0.143172  -1.505   0.1324
CountryNew Zealand                -0.048437   0.141892  -0.341   0.7328
CountryNigeria                     0.100661   0.138966   0.724   0.4688
CountrySouth Africa               -0.144091   0.142027  -1.015   0.3103
CountrySouth Korea                 0.075428   0.141390   0.533   0.5937
CountrySpain                      -0.216156   0.140416  -1.539   0.1237
CountryThailand                    0.012666   0.141987   0.089   0.9289
CountryUnited Kingdom             -0.217065   0.139187  -1.560   0.1189
CountryUnited States              -0.063086   0.140351  -0.449   0.6531
CountryVietnam                     0.109014   0.141692   0.769   0.4417
ContinentAsia                            NA         NA      NA       NA
ContinentAustralia                       NA         NA      NA       NA
ContinentEurope                          NA         NA      NA       NA
ContinentNorth America                   NA         NA      NA       NA
ContinentSouth America                   NA         NA      NA       NA
HemisphereSouthern Hemisphere            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10916  on 7873  degrees of freedom
Residual deviance: 10843  on 7831  degrees of freedom
AIC: 10929

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = Class ~ scale_Cholesterol + scale_Diastolic + scale_Sleep.Hours.Per.Day + 
    Diabetes + Obesity + Alcohol.Consumption + Country, family = "binomial", 
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.40905  -1.17034   0.02631   1.17056   1.42089

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.13896    0.10799   1.287  0.19819
scale_Cholesterol          0.04550    0.02271   2.003  0.04516 *
scale_Diastolic           -0.04480    0.02280  -1.964  0.04949 *
scale_Sleep.Hours.Per.Day -0.05929    0.02270  -2.612  0.00901 **
Diabetes                   0.08773    0.04785   1.833  0.06673 .
Obesity                   -0.07727    0.04532  -1.705  0.08816 .
Alcohol.Consumption       -0.11235    0.04613  -2.435  0.01489 *
CountryAustralia          -0.00453    0.13900  -0.033  0.97400
CountryBrazil             -0.15068    0.13848  -1.088  0.27657
CountryCanada             -0.06569    0.13769  -0.477  0.63333
CountryChina              -0.19045    0.13964  -1.364  0.17263
CountryColombia            0.10699    0.13991   0.765  0.44442
CountryFrance             -0.28204    0.14047  -2.008  0.04465 *
CountryGermany            -0.12357    0.13622  -0.907  0.36432
CountryIndia              -0.34929    0.14277  -2.446  0.01443 *
CountryItaly              -0.23456    0.14012  -1.674  0.09412 .
CountryJapan              -0.20932    0.14287  -1.465  0.14290
CountryNew Zealand        -0.04882    0.14152  -0.345  0.73014
CountryNigeria             0.10587    0.13869   0.763  0.44523
CountrySouth Africa       -0.14224    0.14171  -1.004  0.31551
CountrySouth Korea         0.07448    0.14113   0.528  0.59765
CountrySpain              -0.21125    0.13999  -1.509  0.13128
CountryThailand            0.01535    0.14171   0.108  0.91373
CountryUnited Kingdom     -0.21416    0.13889  -1.542  0.12308
CountryUnited States      -0.05179    0.14006  -0.370  0.71155
CountryVietnam             0.11132    0.14142   0.787  0.43120
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10916  on 7873  degrees of freedom
Residual deviance: 10856  on 7848  degrees of freedom
AIC: 10908

Number of Fisher Scoring iterations: 3
```

Results: Decision Tree

- Two iterations, one with all variables and the with country removed





Results: Confusion Matrix Analysis

Description: df [4 × 5]

	Accuracy <dbl>	Precision <dbl>	Specificity <dbl>	Recall <dbl>	F1 <dbl>
Model 1A	0.4985181	0.4984967	0.5056313	0.4914049	0.4949254
Model 1B	0.5014819	0.5014854	0.5026675	0.5002964	0.5008902
Model 2A	0.4967398	0.4967742	0.4914049	0.5020747	0.4994104
Model 2B	0.5109662	0.5067444	0.1979846	0.8239478	0.6275395

4 rows

- While the second iteration of the decision tree model (Model 2B) performed the best in accuracy, precision, recall, and F1, its specificity was extremely low; recall on the other hand was very high
    - This model was over capturing positives. Implication wise, this would mean that people would be deemed as at risk for a heart attack when in reality, they're not.
- To avoid inducing stress on patients and going through several false-positives, Model 1B (the second iteration of binomial regression) is the next best-performing and was thus deemed as the best fit

Conclusions

- While the models created for this project were pruned down to the best fitting one, the reality is that accuracy is still only 50% overall for the final decision; this means that flipping a coin has the same results as running through the selected model

- In the future, perhaps it would be best to create models based on data from certain regions of the world given the reality that this is a global issue and not every human is the same

- Another idea could be to get more data from various countries and regions throughout the world, enough to create well-fitted models, and then also track these individuals over the course of their lives rather than just being a snapshot

References

Adhikary, D., Barman, S., Ranjan, R., & Stone, H. (2022, October 10). A systematic review of major cardiovascular risk factors: A growing global health concern. Cureus. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9644238/

Berry, J. D., Dyer, A., Cai, X., Garside, D. B., Ning, H., Thomas, A., Greenland, P., Van Horn, L., Tracy, R. P., & Lloyd-Jones, D. M. (2012, January 26). Lifetime risks of cardiovascular disease. New England Journal of Medicine. https://www.nejm.org/doi/full/10.1056/NEJMoa1012848

Fuchs, F. D., & Whelton, P. K. (2019, December 23). High blood pressure and cardiovascular disease | hypertension. AHA Journals. https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.119.14240