Cardiovascular Disease Risk Prediction

STAT 183 | Nicole Carter

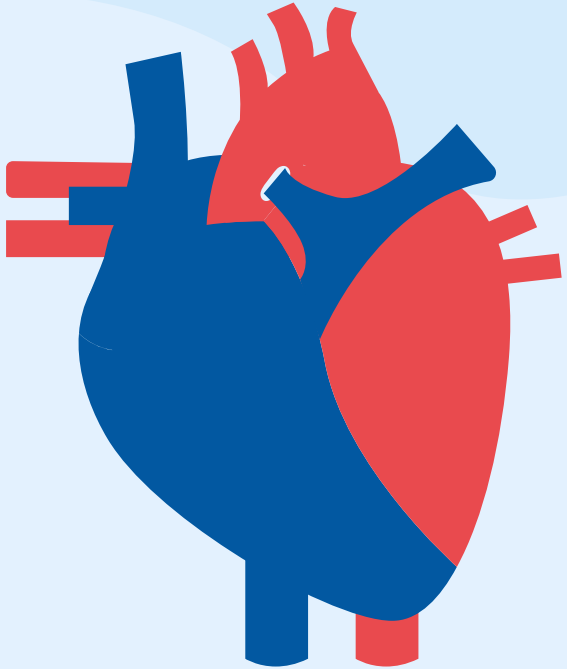


Table of Contents

Introduction 01

Data Description 02

EDA 03

04 Modeling

05 Model Analysis

06 Conclusions

01

Introduction

Introduction

- Heart Disease is one of the leading causes of death
  - Coronary heart disease
  - Stroke
  - Other minor CVD causes combined
  - High blood pressure
  - Heart failure
  - Diseases of the arteries
- On average, someone dies of CVD every 34 seconds in the U.S.
- There are about 1,905 deaths from heart disease, each day in the U.S., including heart attacks

https://www.heart.org/en/about-us/heart-and-stroke-association-statistics?uid=1740
https://newsroom.heart.org/news/more-than-half-of-u-s-adults-dont-know-heart-disease-is-leading-cause-of-death-despite-100-year-reign

02

Data Description

Data Description

- Dataset from Kaggle
  - Cardiovascular Diseases Risk Prediction Dataset
    - Data was already pre-processed, no missing values and a large amount of observations
- Data was collected via telephone surveys conducted for the Behavioral Risk Factor Surveillance System (BRFSS) to assess preventative services
- The target response is identified by respondents who answered that they have Coronary Heart Disease or Myocardial Infarction (heart attack history)

https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data

Data Description

- ➔ 308,854 observations x 19 variables
  - 9 Categorical variables
  - 7 Numerical variables
  - 3 Ordinal variables
    - General_Health
    - Checkup
    - Exercise
    - **Heart_Disease** (Target)
    - Skin_Cancer
    - Other_Cancer
    - Depression
    - Diabetes
    - Arthritis
    - Sex
    - Age_Category
    - Height_.cm.
    - Weight_.kg.
    - BMI
    - Smoking_History
    - Alcohol_Consumption
    - Fruit_Consumption
    - Green_Vegetables_Consumption
    - FriedPotato_Consumption

Data Description

- 19 Variables
  - **General_Health (5 levels)**
    - "Poor" "Very Good" "Good" "Fair" "Excellent"
  - **Checkup (5 levels)**
    - "Within the past 2 years" "Within the past year" "5 or more years ago" "Within the past 5 years" "Never"
  - Exercise
    - Yes / No, **recreational within the last month**
  - Heart_Disease
    - Yes / No
  - Skin_Cancer
    - Yes / No
  - Other_Cancer
    - Yes / No
  - Depression
    - Yes / No
  - **Diabetes (4 levels)**
    - Yes / No
    - No, pre-diabetes or borderline diabetes
    - Yes, but female told only during pregnancy
  - Arthritis
    - Yes / No
  - Sex
    - Female / Male
  - **Age_Category (13 levels)**
    - "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+"
  - Height_.cm.
    - Numerical, ranged from 91 to 241
  - Weight_.kg.
    - Numerical, ranged from 24.95 to 293.02
  - BMI
    - Numerical, ranged from 12.02 to 99.33
  - Smoking_History
    - Yes / No
  - **Alcohol_Consumption**
    - Numerical, # of days within the last month
  - **Fruit_Consumption**
    - Numerical, # of days within the last month
  - **Green_Vegetables_Consumption**
    - Numerical, # of days within the last month
  - **FriedPotato_Consumption**
    - Numerical, # of days within the last month

Glimpse of the Data

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_.cm.
1	Poor	Within the past 2 years	No	No	No	No	No	No	Yes	Female	70-74	150
2	Very Good	Within the past year	No	Yes	No	No	No	Yes	No	Female	70-74	165
3	Very Good	Within the past year	Yes	No	No	No	No	Yes	No	Female	60-64	163
4	Poor	Within the past year	Yes	Yes	No	No	No	Yes	No	Male	75-79	180
5	Good	Within the past year	No	No	No	No	No	No	No	Male	80+	191
6	Good	Within the past year	No	No	No	No	Yes	No	Yes	Male	60-64	183
7	Fair	Within the past year	Yes	Yes	No	No	No	No	Yes	Male	60-64	175

Weight_.kg.	BMI	Smoking_History	Alcohol_Consumption	Fruit_Consumption	Green_Vegetables_Consumption	FriedPotato_Consumption
32.66	14.54	Yes	0	30	16	12
77.11	28.29	No	0	30	0	4
88.45	33.47	No	4	12	3	16
93.44	28.73	No	0	30	30	8
88.45	24.37	Yes	0	8	4	0
154.22	46.11	No	0	12	12	12
69.85	22.74	Yes	0	16	8	0

Data Cleaning

- No missing values
- 80 duplicate observations in the data
- After removal:
  - Data set dimensions: 308,774 observations x 19 variables

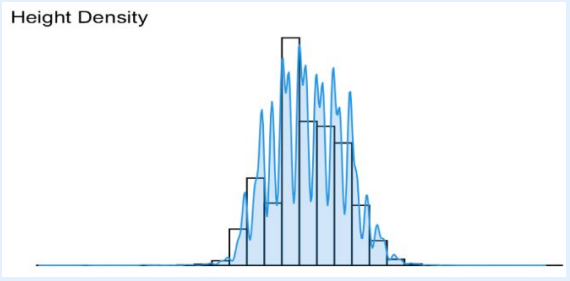
Analysis Goal

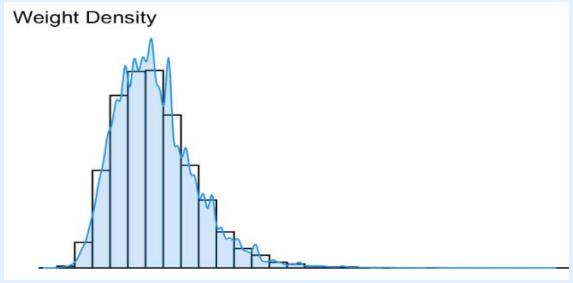
- ➔ The goal is to find out which variables have the most significance to predict the risk of cardiovascular diseases.
- ➔ With a person's health-related behaviors and chronic health conditions, what factors would be used to best identify a person who is at risk of having cardiovascular diseases?

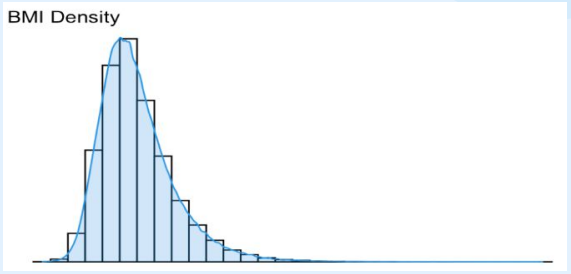
03

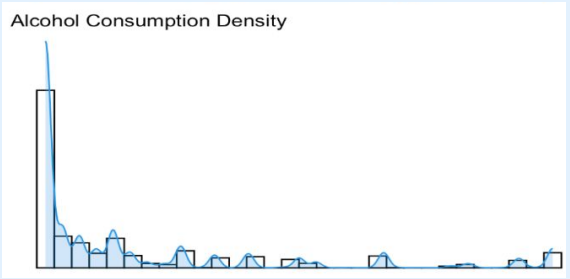
EDA

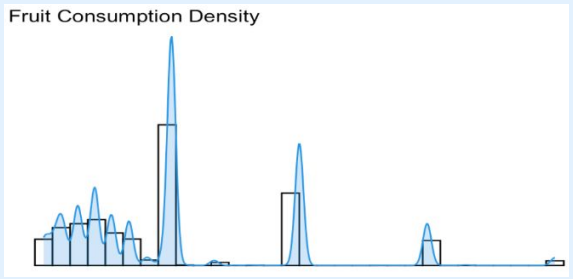
Distributions of Numerical Variables

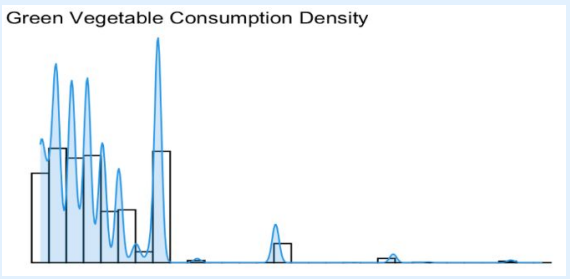


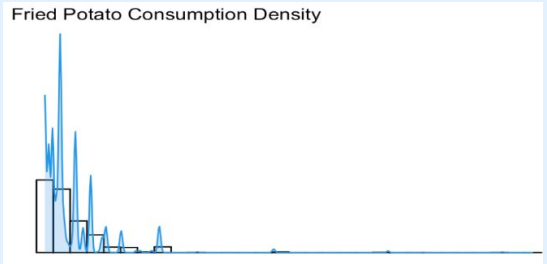












04

Modeling

Logistic Regression Model (1)

```
glm(formula = HD_Int ~ . - Heart_Disease, family = "binomial",
    data = cvd)

Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                 -6.272e+00  4.939e-01 -12.698
General_HealthFair                           1.707e+00  3.551e-02  48.059
General_HealthGood                           1.069e+00  3.362e-02  31.811
General_HealthPoor                           2.250e+00  3.995e-02  56.319
General_HealthVery Good                      5.127e-01  3.443e-02  14.890
CheckupNever                                 2.612e-01  1.531e-01   1.706
CheckupWithin the past 2 years               2.262e-01  6.353e-02   3.560
CheckupWithin the past 5 years               1.297e-01  7.485e-02   1.732
CheckupWithin the past year                  5.235e-01  5.787e-02   9.047
ExerciseYes                                 -1.805e-02  1.643e-02  -1.099
Skin_CancerYes                               1.213e-01  1.975e-02   6.142
Other_CancerYes                              4.802e-02  1.945e-02   2.468
DepressionYes                                2.448e-01  1.812e-02  13.509
DiabetesNo, pre-diabetes or borderline diabetes  1.462e-01  4.120e-02   3.548
DiabetesYes                                  5.333e-01  1.695e-02  31.461
DiabetesYes, but female told only during pregnancy  1.292e-01  1.091e-01   1.184
ArthritisYes                                 2.606e-01  1.534e-02  16.986
SexMale                                      8.324e-01  2.098e-02  39.672
Age_Category25-29                            3.063e-01  1.404e-01   2.182
Age_Category30-34                            6.373e-01  1.259e-01   5.063
Age_Category35-39                            7.547e-01  1.206e-01   6.260
Age_Category40-44                            1.085e+00  1.149e-01   9.441
Age_Category45-49                            1.463e+00  1.114e-01  13.139
Age_Category50-54                            1.777e+00  1.085e-01  16.381
Age_Category55-59                            2.116e+00  1.069e-01  19.789
Age_Category60-64                            2.356e+00  1.062e-01  22.184
Age_Category65-69                            2.598e+00  1.059e-01  24.523
Age_Category70-74                            2.858e+00  1.059e-01  26.991
Age_Category75-79                            3.097e+00  1.064e-01  29.104
Age_Category80+                              3.373e+00  1.062e-01  31.753
Height_.cm.                                 -5.191e-03  2.827e-03  -1.836
Weight_.kg.                                  3.153e-04  2.564e-03   0.123
BMI                                          8.352e-04  7.401e-03   0.113
Smoking_HistoryYes                           3.958e-01  1.481e-02  26.735
Alcohol_Consumption                         -9.856e-03  9.141e-04 -10.782
Fruit_Consumption                            3.171e-05  3.105e-04   0.102
Green_Vegetables_Consumption                 7.983e-04  5.295e-04   1.508
FriedPotato_Consumption                     -6.098e-04  8.664e-04  -0.704
                                              Pr(>|z|)
(Intercept)                                  < 2e-16 ***
General_HealthFair                           < 2e-16 ***
General_HealthGood                           < 2e-16 ***
General_HealthPoor                           < 2e-16 ***
General_HealthVery Good                      < 2e-16 ***
CheckupNever                                 0.087919 .
CheckupWithin the past 2 years               0.000371 ***
CheckupWithin the past 5 years               0.083194 .
CheckupWithin the past year                  < 2e-16 ***
ExerciseYes                                  0.271973
Skin_CancerYes                               8.15e-10 ***
Other_CancerYes                              0.013570 *
DepressionYes                                < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes  0.000388 ***
DiabetesYes                                  < 2e-16 ***
DiabetesYes, but female told only during pregnancy  0.236584
ArthritisYes                                 < 2e-16 ***
SexMale                                      < 2e-16 ***
Age_Category25-29                            0.029134 *
Age_Category30-34                            4.13e-07 ***
Age_Category35-39                            3.85e-10 ***
Age_Category40-44                            < 2e-16 ***
Age_Category45-49                            < 2e-16 ***
Age_Category50-54                            < 2e-16 ***
Age_Category55-59                            < 2e-16 ***
Age_Category60-64                            < 2e-16 ***
Age_Category65-69                            < 2e-16 ***
Age_Category70-74                            < 2e-16 ***
Age_Category75-79                            < 2e-16 ***
Age_Category80+                              < 2e-16 ***
Height_.cm.                                  0.066368 .
Weight_.kg.                                  0.902127
BMI                                          0.910157
Smoking_HistoryYes                           < 2e-16 ***
Alcohol_Consumption                          < 2e-16 ***
Fruit_Consumption                            0.918677
Green_Vegetables_Consumption                 0.131636
FriedPotato_Consumption                      0.481547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173465  on 308773  degrees of freedom
Residual deviance: 136957  on 308736  degrees of freedom
AIC: 137033

Number of Fisher Scoring iterations: 7
```

Full Model

- AIC: 137,033
- Non-Significant Factors
  - Checkup
    - Never & Past 5 yrs.
  - Exercise
  - Diabetes
    - Pregnancy
  - Height
  - Weight
  - BMI
  - Fruit Consumption
  - Green Vegetables Consumption
  - Fried Potato Consumption

Analysis of Deviance Test for Checkup + Diabetes

```
Analysis of Deviance Table (Type II tests)

Response: HD_Int
         Df Chisq Pr(>Chisq)
Checkup   4  2460  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Deviance Table (Type II tests)

Response: HD_Int
          Df  Chisq Pr(>Chisq)
Diabetes   3 9187.5  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Check for overall significance of Checkup and Diabetes to the response variable
- Both p-values are small, indicating significance to Heart Disease
- Certain levels can still be labeled as non-significant

Multicollinearity Analysis (Full Model)

- ➔ 3 high VIF values
- ➔ Indicates a multicollinearity problem

	GVIF	Df	GVIF^(1/(2*Df))
General_Health	1.293889	4	1.032731
Checkup	1.063265	4	1.007698
Exercise	1.171483	1	1.082351
Skin_Cancer	1.085002	1	1.041634
Other_Cancer	1.059698	1	1.029416
Depression	1.147095	1	1.071025
Diabetes	1.155402	3	1.024367
Arthritis	1.137940	1	1.066743
Sex	2.095054	1	1.447430
Age_Category	1.366381	12	1.013092
Height_.cm.	17.958276	1	4.237721
Weight_.kg.	61.822837	1	7.862750
BMI	48.305413	1	6.950210
Smoking_History	1.054379	1	1.026830
Alcohol_Consumption	1.079800	1	1.039134
Fruit_Consumption	1.104466	1	1.050936
Green_Vegetables_Consumption	1.100718	1	1.049151
FriedPotato_Consumption	1.036166	1	1.017922

Logistic Regression Model (2)

```
glm(formula = HD_Int ~ . - Heart_Disease - Height_.cm. - Weight_.kg. -
    BMI, family = "binomial", data = cvd)

Coefficients:
                                              Estimate Std. Error z value
(Intercept)                                -7.079e+00  1.227e-01 -57.689
General_HealthFair                          1.715e+00  3.540e-02  48.457
General_HealthGood                          1.076e+00  3.353e-02  32.083
General_HealthPoor                          2.258e+00  3.989e-02  56.592
General_HealthVery Good                     5.149e-01  3.440e-02  14.969
CheckupNever                                2.702e-01  1.530e-01   1.766
CheckupWithin the past 2 years              2.282e-01  6.352e-02   3.593
CheckupWithin the past 5 years              1.308e-01  7.485e-02   1.748
CheckupWithin the past year                 5.265e-01  5.784e-02   9.103
ExerciseYes                                -2.223e-02  1.635e-02  -1.360
Skin_CancerYes                              1.167e-01  1.973e-02   5.915
Other_CancerYes                             4.609e-02  1.945e-02   2.370
DepressionYes                               2.458e-01  1.811e-02  13.576
DiabetesNo, pre-diabetes or borderline diabetes  1.535e-01  4.111e-02   3.734
DiabetesYes                                 5.380e-01  1.664e-02  32.325
DiabetesYes, but female told only during pregnancy  1.343e-01  1.091e-01   1.231
ArthritisYes                                2.626e-01  1.523e-02  17.244
SexMale                                     7.601e-01  1.541e-02  49.309
Age_Category25-29                           3.113e-01  1.404e-01   2.218
Age_Category30-34                           6.429e-01  1.258e-01   5.109
Age_Category35-39                           7.612e-01  1.205e-01   6.318
Age_Category40-44                           1.092e+00  1.148e-01   9.505
Age_Category45-49                           1.470e+00  1.113e-01  13.207
Age_Category50-54                           1.785e+00  1.084e-01  16.461
Age_Category55-59                           2.124e+00  1.069e-01  19.866
Age_Category60-64                           2.363e+00  1.062e-01  22.259
Age_Category65-69                           2.606e+00  1.059e-01  24.606
Age_Category70-74                           2.868e+00  1.059e-01  27.091
Age_Category75-79                           3.109e+00  1.064e-01  29.223
Age_Category80+                             3.386e+00  1.061e-01  31.903
Smoking_HistoryYes                          3.931e-01  1.478e-02  26.593
Alcohol_Consumption                        -1.006e-02  9.129e-04 -11.022
Fruit_Consumption                           7.103e-06  3.105e-04   0.023
Green_Vegetables_Consumption                7.560e-04  5.297e-04   1.427
FriedPotato_Consumption                    -6.783e-04  8.667e-04  -0.783
                                              Pr(>|z|)
(Intercept)                                 < 2e-16 ***
General_HealthFair                          < 2e-16 ***
General_HealthGood                          < 2e-16 ***
General_HealthPoor                          < 2e-16 ***
General_HealthVery Good                     < 2e-16 ***
CheckupNever                                0.077397 .
CheckupWithin the past 2 years              0.000327 ***
CheckupWithin the past 5 years              0.080453 .
CheckupWithin the past year                 < 2e-16 ***
ExerciseYes                                 0.173913
Skin_CancerYes                              3.31e-09 ***
Other_CancerYes                             0.017778 *
DepressionYes                               < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes  0.000188 ***
DiabetesYes                                 < 2e-16 ***
DiabetesYes, but female told only during pregnancy  0.218496
ArthritisYes                                < 2e-16 ***
SexMale                                     < 2e-16 ***
Age_Category25-29                           0.026573 *
Age_Category30-34                           3.24e-07 ***
Age_Category35-39                           2.65e-10 ***
Age_Category40-44                           < 2e-16 ***
Age_Category45-49                           < 2e-16 ***
Age_Category50-54                           < 2e-16 ***
Age_Category55-59                           < 2e-16 ***
Age_Category60-64                           < 2e-16 ***
Age_Category65-69                           < 2e-16 ***
Age_Category70-74                           < 2e-16 ***
Age_Category75-79                           < 2e-16 ***
Age_Category80+                             < 2e-16 ***
Smoking_HistoryYes                          < 2e-16 ***
Alcohol_Consumption                         < 2e-16 ***
Fruit_Consumption                           0.981750
Green_Vegetables_Consumption                0.153487
FriedPotato_Consumption                     0.433863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173465  on 308773  degrees of freedom
Residual deviance: 136986  on 308739  degrees of freedom
AIC: 137056

Number of Fisher Scoring iterations: 7
```

Reduced Model 1

- Based off of Multicollinearity Results

➔ Removed Variables
- Height
- Weight
- BMI

➔ Non-Significant Factors
- Checkup (Never & Past 5 yrs.)
- Exercise
- Diabetes (Pregnancy)
- Fruit Consumption
- Green Vegetables Consumption
- Fried Potato Consumption

➔ AIC: 137,056 > Full Model
- More complex than Full

Multicollinearity Analysis (Reduced Model 1)

➔ No sign of multicollinearity

	GVIF	Df	GVIF^(1/(2*Df))
General_Health	1.281180	4	1.031457
Checkup	1.060951	4	1.007423
Exercise	1.159146	1	1.076637
Skin_Cancer	1.082643	1	1.040501
Other_Cancer	1.058826	1	1.028993
Depression	1.145372	1	1.070221
Diabetes	1.111833	3	1.017825
Arthritis	1.121720	1	1.059113
Sex	1.130849	1	1.063414
Age_Category	1.269022	12	1.009976
Smoking_History	1.051306	1	1.025332
Alcohol_Consumption	1.076278	1	1.037438
Fruit_Consumption	1.104214	1	1.050816
Green_Vegetables_Consumption	1.100307	1	1.048955
FriedPotato_Consumption	1.035179	1	1.017437

Logistic Regression Model (3)

```
glm(formula = HD_Int ~ General_Health + Checkup + Skin_Cancer +
    Other_Cancer + Depression + Diabetes + Arthritis + Sex +
    Age_Category + Height_.cm. + Weight_.kg. + Smoking_History +
    Alcohol_Consumption + Green_Vegetables_Consumption, family = "binomial",
    data = cvd)

Coefficients:
                                                Estimate Std. Error z value
(Intercept)                                   -6.2311969  0.2005056 -31.077
General_HealthFair                             1.7096980  0.0352880  48.450
General_HealthGood                             1.0702314  0.0335476  31.902
General_HealthPoor                             2.2560762  0.0394223  57.228
General_HealthVery Good                        0.5123969  0.0344188  14.887
CheckupNever                                   0.2623874  0.1530742   1.714
CheckupWithin the past 2 years                 0.2260931  0.0635142   3.560
CheckupWithin the past 5 years                 0.1293801  0.0748481   1.729
CheckupWithin the past year                    0.5234075  0.0578510   9.048
Skin_CancerYes                                 0.1207921  0.0197426   6.118
Other_CancerYes                                0.0479837  0.0194509   2.467
DepressionYes                                  0.2456192  0.0181038  13.567
DiabetesNo, pre-diabetes or borderline diabetes 0.1462278  0.0412005   3.549
DiabetesYes                                    0.5340382  0.0169280  31.548
DiabetesYes, but female told only during pregnancy 0.1283153  0.1091454   1.176
ArthritisYes                                   0.2607693  0.0153372  17.002
SexMale                                        0.8303160  0.0208292  39.863
Age_Category25-29                              0.3065668  0.1404123   2.183
Age_Category30-34                              0.6379956  0.1258801   5.068
Age_Category35-39                              0.7555672  0.1205478   6.268
Age_Category40-44                              1.0867726  0.1149207   9.457
Age_Category45-49                              1.4656747  0.1113442  13.163
Age_Category50-54                              1.7798942  0.1084610  16.410
Age_Category55-59                              2.1195917  0.1068943  19.829
Age_Category60-64                              2.3592314  0.1061230  22.231
Age_Category65-69                              2.6013223  0.1058589  24.573
Age_Category70-74                              2.8618893  0.1058114  27.047
Age_Category75-79                              3.1015121  0.1063255  29.170
Age_Category80+                                3.3787104  0.1060891  31.848
Height_.cm.                                   -0.0055565  0.0010177  -5.460
Weight_.kg.                                    0.0006379  0.0003995   1.597
Smoking_HistoryYes                             0.3962494  0.0147515  26.862
Alcohol_Consumption                           -0.0098946  0.0009135 -10.832
Green_Vegetables_Consumption                   0.0007387  0.0005111   1.445
                                                Pr(>|z|)
(Intercept)                                   < 2e-16 ***
General_HealthFair                            < 2e-16 ***
General_HealthGood                            < 2e-16 ***
General_HealthPoor                            < 2e-16 ***
General_HealthVery Good                       < 2e-16 ***
CheckupNever                                  0.086507 .
CheckupWithin the past 2 years                0.000371 ***
CheckupWithin the past 5 years                0.083886 .
CheckupWithin the past year                   < 2e-16 ***
Skin_CancerYes                                9.46e-10 ***
Other_CancerYes                               0.013628 *
DepressionYes                                 < 2e-16 ***
DiabetesNo, pre-diabetes or borderline diabetes 0.000386 ***
DiabetesYes                                   < 2e-16 ***
DiabetesYes, but female told only during pregnancy 0.239740
ArthritisYes                                  < 2e-16 ***
SexMale                                       < 2e-16 ***
Age_Category25-29                             0.029011 *
Age_Category30-34                             4.01e-07 ***
Age_Category35-39                             3.66e-10 ***
Age_Category40-44                             < 2e-16 ***
Age_Category45-49                             < 2e-16 ***
Age_Category50-54                             < 2e-16 ***
Age_Category55-59                             < 2e-16 ***
Age_Category60-64                             < 2e-16 ***
Age_Category65-69                             < 2e-16 ***
Age_Category70-74                             < 2e-16 ***
Age_Category75-79                             < 2e-16 ***
Age_Category80+                               < 2e-16 ***
Height_.cm.                                   4.77e-08 ***
Weight_.kg.                                   0.110351
Smoking_HistoryYes                            < 2e-16 ***
Alcohol_Consumption                           < 2e-16 ***
Green_Vegetables_Consumption                  0.148370
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173465  on 308773  degrees of freedom
Residual deviance: 136959  on 308740  degrees of freedom
AIC: 137027

Number of Fisher Scoring iterations: 7
```

- ➔ Stepwise Regression of Full Model
- ➔ Removed Variables
  - ◆ Exercise
  - ◆ BMI
  - ◆ Fruit Consumption
  - ◆ Fried Potato Consumption
- ➔ Non-Significant Factors
  - ◆ Checkup (Never & Past 5 yrs.)
  - ◆ Diabetes (Pregnancy)
  - ◆ Weight
  - ◆ Green Vegetables Consumption
- ➔ AIC: 137,027 < Full Model
  - ◆ Less complex than full

Multicollinearity Analysis (Stepwise Model)

➔ No sign of multicollinearity

	GVIF	Df	GVIF^(1/(2*Df))
General_Health	1.216653	4	1.024816
Checkup	1.062218	4	1.007573
Skin_Cancer	1.083744	1	1.041030
Other_Cancer	1.059274	1	1.029210
Depression	1.144614	1	1.069866
Diabetes	1.152001	3	1.023864
Arthritis	1.137527	1	1.066549
Sex	2.064662	1	1.436893
Age_Category	1.339493	12	1.012253
Height_.cm.	2.326569	1	1.525310
Weight_.kg.	1.502236	1	1.225657
Smoking_History	1.046645	1	1.023057
Alcohol_Consumption	1.077918	1	1.038228
Green_Vegetables_Consumption	1.025018	1	1.012432

05

Model Analysis

ANOVA (Full Model)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
General_Health	4	1429	357.3	5497.475	< 2e-16	***
Checkup	4	119	29.7	456.776	< 2e-16	***
Exercise	1	14	14.0	215.936	< 2e-16	***
Skin_Cancer	1	128	127.9	1968.513	< 2e-16	***
Other_Cancer	1	36	36.3	558.612	< 2e-16	***
Depression	1	8	8.2	126.209	< 2e-16	***
Diabetes	3	268	89.4	1375.067	< 2e-16	***
Arthritis	1	106	105.7	1627.085	< 2e-16	***
Sex	1	164	163.7	2519.636	< 2e-16	***
Age_Category	12	559	46.6	716.544	< 2e-16	***
Height_.cm.	1	1	0.9	14.016	0.000181	***
Weight_.kg.	1	1	1.5	23.032	1.59e-06	***
BMI	1	0	0.1	0.877	0.348892	
Smoking_History	1	44	44.0	677.229	< 2e-16	***
Alcohol_Consumption	1	10	9.8	150.987	< 2e-16	***
Fruit_Consumption	1	0	0.0	0.653	0.418894	
Green_Vegetables_Consumption	1	0	0.2	2.889	0.089176	.
FriedPotato_Consumption	1	0	0.1	0.841	0.359101	
Residuals	308736	20064	0.1			

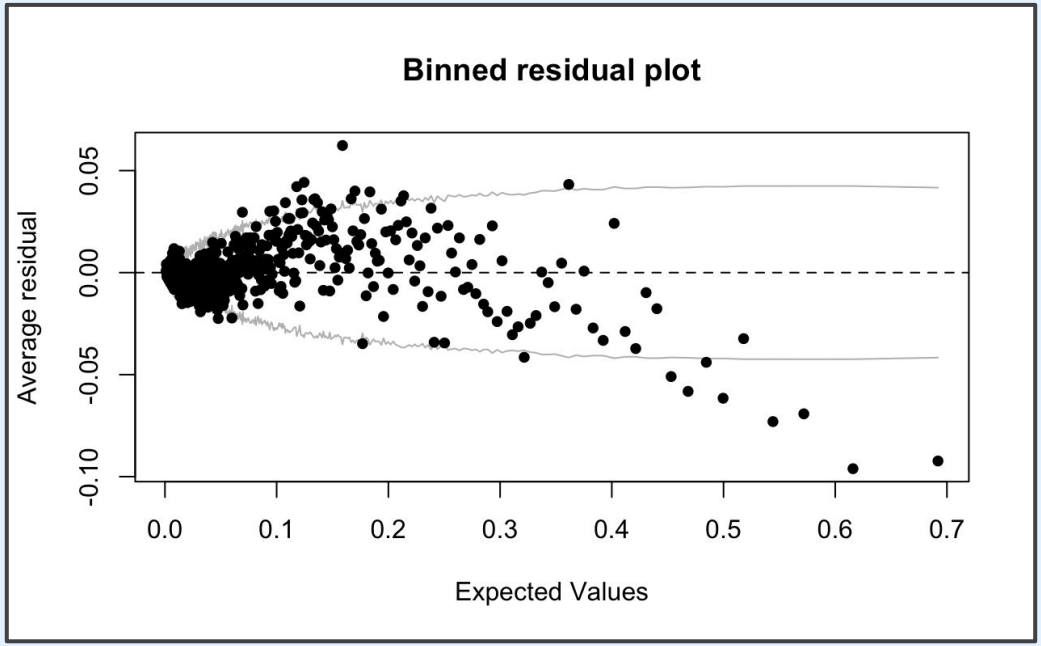
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

➔ Non-Significant Factors

- BMI
- Fruit Consumption
- Green Vegetables Consumption
- Fried Potato Consumption

Binned Residuals (Full Model)

- ➔ Many outliers
- ➔ Decreasing pattern



Model Accuracy (Full Model)

- Accuracy: 0.9193261
- Area Under Curve: 0.8351

Confusion Matrix	Predicted No Heart Disease	Predicted Heart Disease
No Heart Disease	282,311	1,492
Heart Disease	23,418	1,553

ANOVA (Reduced Model 1)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
General_Health	4	1429	357.3	5496.895	<2e-16	***
Checkup	4	119	29.7	456.728	<2e-16	***
Exercise	1	14	14.0	215.914	<2e-16	***
Skin_Cancer	1	128	127.9	1968.306	<2e-16	***
Other_Cancer	1	36	36.3	558.553	<2e-16	***
Depression	1	8	8.2	126.195	<2e-16	***
Diabetes	3	268	89.4	1374.922	<2e-16	***
Arthritis	1	106	105.7	1626.913	<2e-16	***
Sex	1	164	163.7	2519.370	<2e-16	***
Age_Category	12	559	46.6	716.469	<2e-16	***
Smoking_History	1	44	44.3	681.413	<2e-16	***
Alcohol_Consumption	1	10	9.6	148.111	<2e-16	***
Fruit_Consumption	1	0	0.1	0.847	0.3573	
Green_Vegetables_Consumption	1	0	0.2	3.163	0.0753	.
FriedPotato_Consumption	1	0	0.1	1.326	0.2495	
Residuals	308739	20067	0.1			

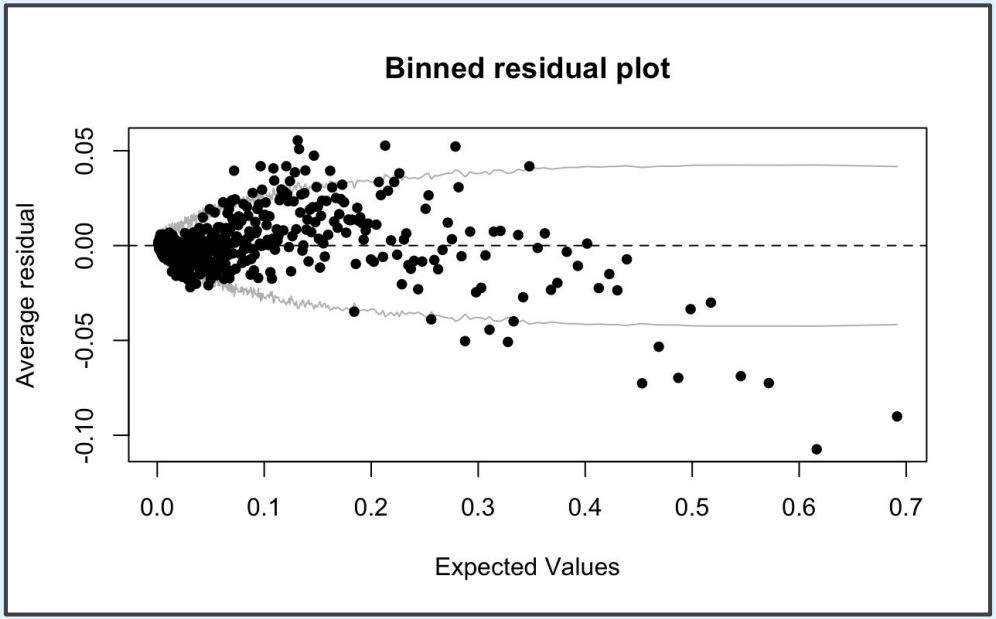
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

➔ Non-Significant Factors

- Fruit Consumption
- Green Vegetables Consumption
- Fried Potato Consumption

Binned Residuals (Reduced Model 1)

- ➔ Many outliers
- ➔ Not too different from Full Model
- ➔ Decreasing pattern



Model Accuracy (Reduced Model 1)

- Accuracy: 0.9193876
- Area Under Curve: 0.835

Confusion Matrix	Predicted No Heart Disease	Predicted Heart Disease
No Heart Disease	282,350	1,453
Heart Disease	23,438	1,533

ANOVA (Stepwise Model)

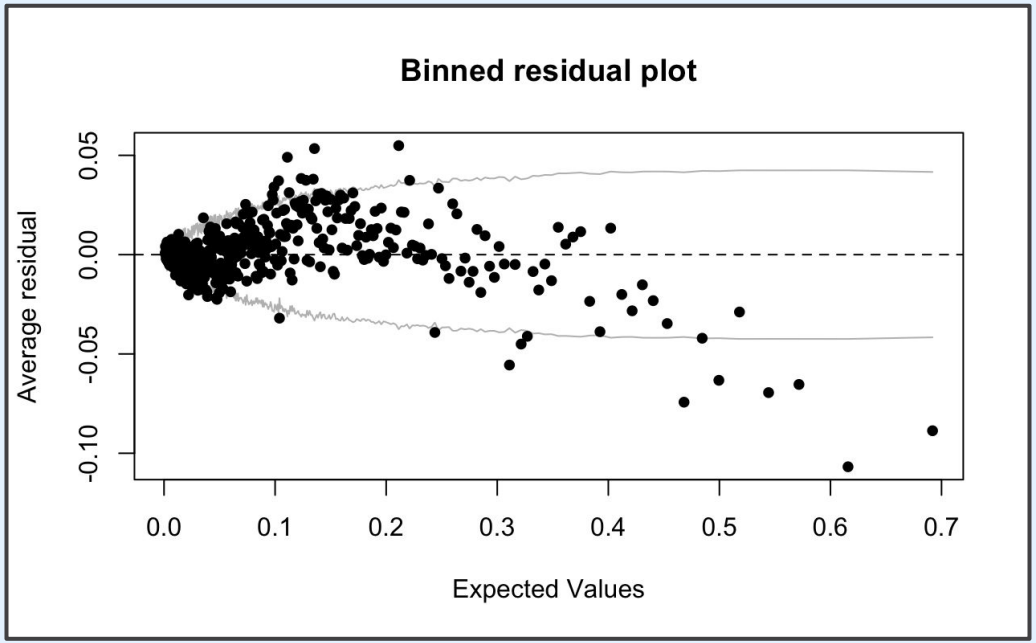
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
General_Health	4	1429	357.3	5497.361	< 2e-16	***
Checkup	4	119	29.7	456.767	< 2e-16	***
Skin_Cancer	1	127	127.1	1955.380	< 2e-16	***
Other_Cancer	1	37	36.9	567.510	< 2e-16	***
Depression	1	8	7.6	117.540	< 2e-16	***
Diabetes	3	276	91.9	1413.357	< 2e-16	***
Arthritis	1	108	108.2	1664.915	< 2e-16	***
Sex	1	160	160.3	2466.710	< 2e-16	***
Age_Category	12	566	47.2	725.694	< 2e-16	***
Height_.cm.	1	1	1.0	15.338	8.99e-05	***
Weight_.kg.	1	1	1.2	19.167	1.20e-05	***
Smoking_History	1	45	44.7	687.230	< 2e-16	***
Alcohol_Consumption	1	10	10.1	154.808	< 2e-16	***
Green_Vegetables_Consumption	1	0	0.2	2.468	0.116	
Residuals	308740	20065	0.1			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- ➔ Non-Significant Factor
  - ◆ Green Vegetables Consumption

Binned Residuals (Stepwise Model)

- ➔ Many outliers
- ➔ Looks similar to other models
- ➔ Decreasing pattern



Model Accuracy (Stepwise Model)

- Accuracy: 0.9193488
- Area Under Curve: 0.8351

Confusion Matrix	Predicted No Heart Disease	Predicted Heart Disease
No Heart Disease	282,317	1,486
Heart Disease	23,417	1,554

06

Conclusions

Conclusions

1. In terms of model accuracy, the best-to-worst of the 3 models are:
    a. Reduced Model 1
    b. Stepwise Model
    c. Full Model
2. For future use, it would be possible to obtain better performing results with other techniques like Ridge or Lasso Regression.
    a. Transform the data to have a more standardized pattern
    b. Target more specific characteristics to look into