Final Report

Will Gullion, Brian Nalley, Nate Nellis, McKade Thomas

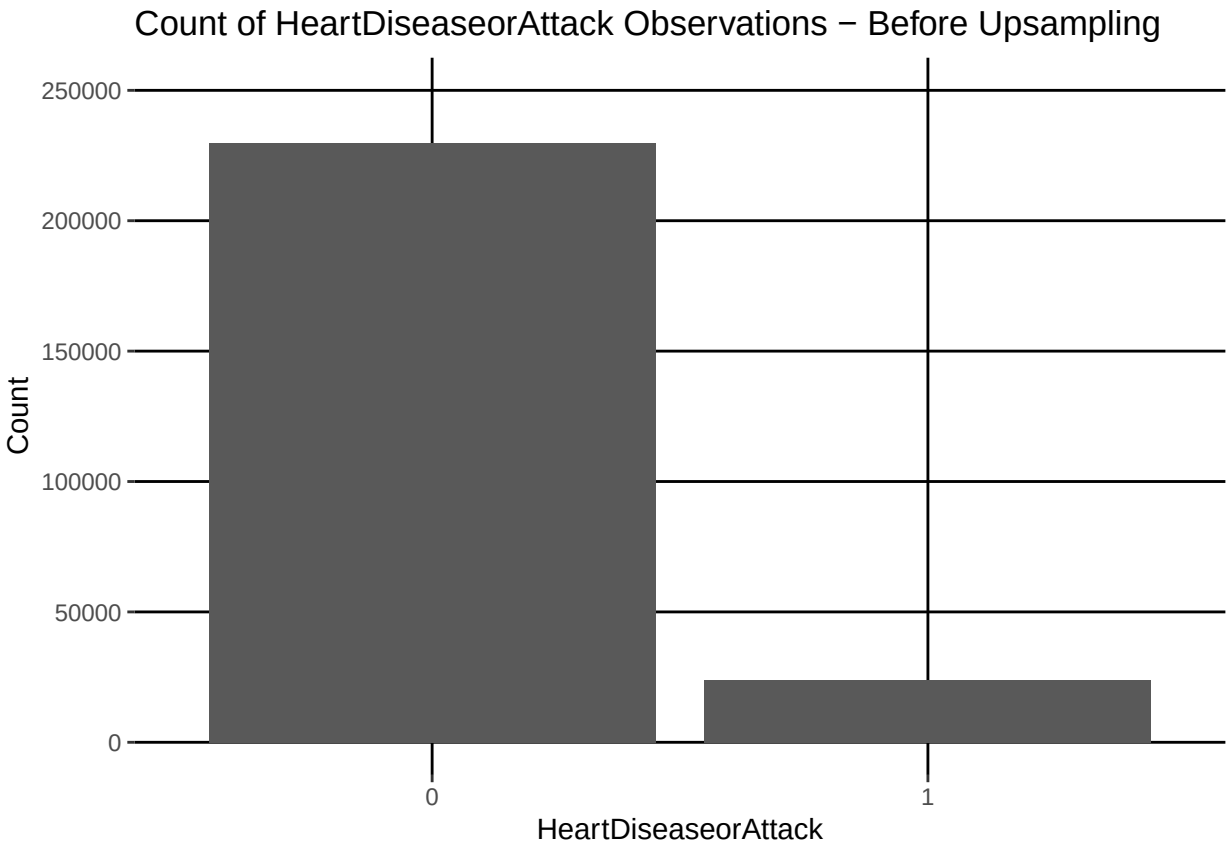
STAT 5650

Introduction

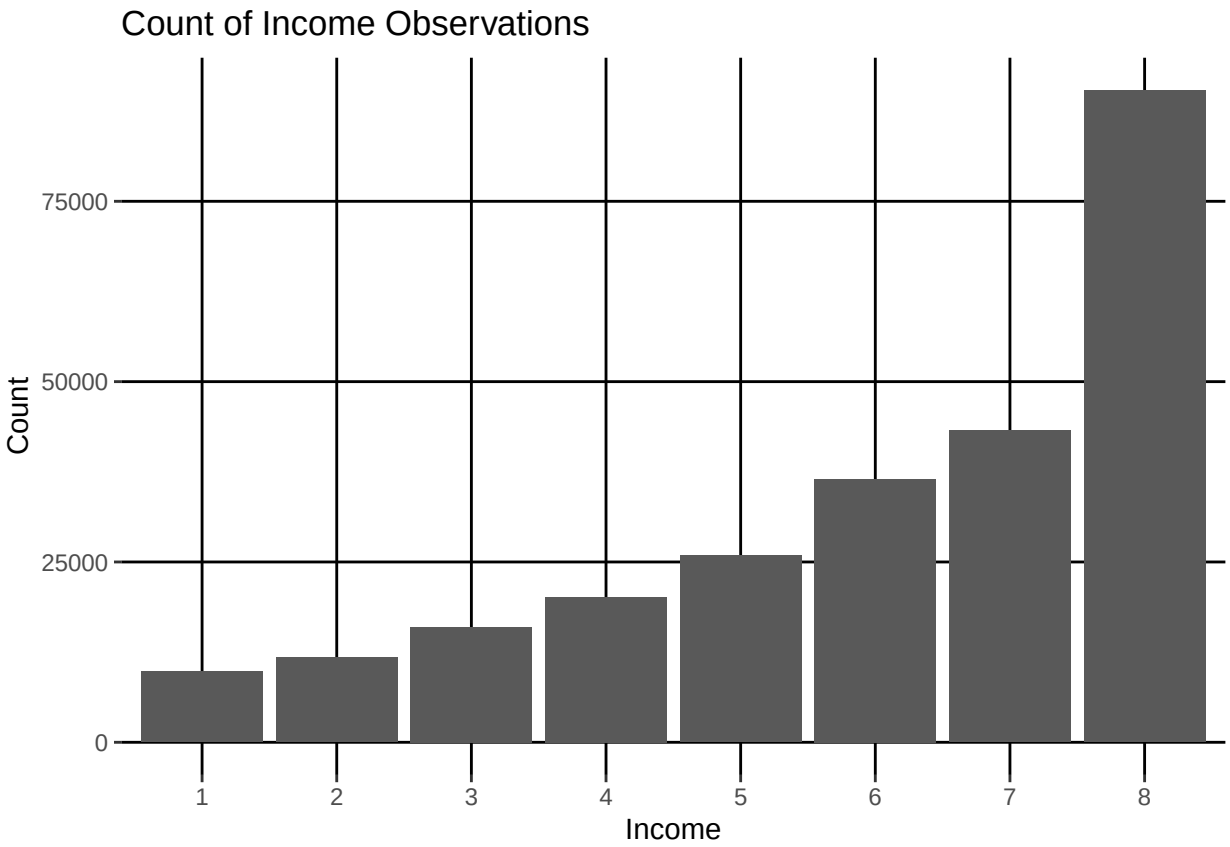
We chose to focus on data having to do with heart disease/heart attach indicators to understand which factors impact a person's probability to have a heart attach or summer from strong levels of heart disease. Data was found on Kaggle and was modified from a CDC survey. See the *Data Dictionary* section for more details at the end of the document. The dataset contains 252,680 observations. The response variable we have is a binary one telling us if the person being interviewed suffered from a heart attack or not. There are a total of 20 available predictor variables.

Exploratory Data Analysis

Taking a closer look at our predictor variable, we noticed that only 9.4% of participants in the survey had heart disease or a heart attack, 23,893 responded yes versus 229,787 who responded no. This means that we could just predict a 0 for all observations and we could get an accuracy of over 90%. Since we aren't concerned in this instance with a high accuracy, but actually a decent sensitivity (getting accurate results when someone is at risk of a heart attack or heart disease), we utilized Upsampling in some of our methods to skew the model to predict the results we wanted. This decreased overall accuracy, but resulted in a much better prediction accuracy of finding those people with heart disease or who had a heart attack.



When it comes to the predictor variables there are 12 binary, 3 numerical, and 5 factor variables. Having fewer numerical variables is less than desireable, especially when it comes to how the factor variables are coded/binned. A terrible example of this is the **Income** variable where the CDC survey sets the highest bucket capped at \$75,000+. Due to how low that is for the maximum response allowed, a full third of the observations fall into that category.



Data Dictionary

Data found by us at https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

Description of variables found in https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Response Variable

HeartDiseaseorAttack, binary

Predictor Variables

HighBP, binary
- Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

HighChol, binary
- Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?

CholCheck, binary
- Cholesterol check within past five years

BMI, numerical

Smoker, binary
- Have you smoked at least 100 cigarettes in your entire life?

Stroke, binary

Diabetes, factor
- 0 is no diabetes,
- 1 is pre-diabetes
- 2 is diabetes
PhysActivity, binary
- During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
Fruits, binary
- Consume Fruit 1 or more times per day
Veggies, binary
- Consume Vegetables 1 or more times per day
HvyAlcoholConsum, binary
- Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
AnyHealthcare, binary
- Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?
NoDocbcCost, binary
- Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
GenHlth, factor
- Would you say that in general your health is:
- 1 = Excellent
- 2 = Very Good
- 3 = Good
- 4 = Fair
- 5 = Poor
MentHlth, numerical
- Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
PhysHlth, numerical
- Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
DiffWalk, binary
- Do you have serious difficulty walking or climbing stairs?
Sex, binary
- 0 = Female
- 1 = Male
Age, factor
- 1 Age 18 to 24
- 2 Age 25 to 29
- 3 Age 30 to 34
- 4 Age 35 to 39
- 5 Age 40 to 44
- 6 Age 45 to 49
- 7 Age 50 to 54
- 8 Age 55 to 59
- 9 Age 60 to 64
- 10 Age 65 to 69
- 11 Age 70 to 74
- 12 Age 75 to 79
- 13 Age 80 or older
- 14 Don't Know / Refused to answer (I removed these as well)
Education, factor
- What is the highest grade or year of school you completed?

- 1 Never attended school or only kindergarten
- 2 Grades 1 through 8 (Elementary)
- 3 Grades 9 through 11 (Some high school)
- 4 Grade 12 or GED (High school graduate)
- 5 College 1 year to 3 years (Some college or technical school)
- 6 College 4 years or more (College graduate)

Income, factor
- Is your annual household income from all sources
- 1 Less than $10,000
- 2 Less than $15,000 ($10,000 to less than $15,000)
- 3 Less than $20,000 ($15,000 to less than $20,000)
- 4 Less than $25,000 ($20,000 to less than $25,000)
- 5 Less than $35,000 ($25,000 to less than $35,000)
- 6 Less than $50,000 ($35,000 to less than $50,000)
- 7 Less than $75,000 ($50,000 to less than $75,000)
- 8 $75,000 or more