Personal Key Indicators to Prevent Heart Disease for Americans

Written by:

Maria Ahumada

Yi Huang

Justin Zelin Jiang

Table of Contents

Executive Summary

As reported by the Centers for Disease Control and Prevention (CDC), heart disease is the leading cause of death in the United States. Therefore, the objective of this report is to act as data scientists and researchers to advocate younger generation of individuals with supporting evidence on the topmost vital factors that could lead to getting heart disease in the future. This information serves as an act of awareness for millennials and Gen-Z individuals who need to be aware of these health care preventions. Our statistical model will facilitate social media influencers in the healthcare and fitness industry to identify areas of improvement for visibility and attention of the existing factors that one should be routinely checked upon.

A dataset from Kaggle was considered where real data was collected from the CDC on U.S. residents' health status. A total of 319,795 data points are available in this dataset. The data contains 18 features such as Body Mass Index (BMI), Race, Diabetic condition, Sex, SleepTime, etc. From the respondents' information, 91.4% did not have heart disease, while 8.6% had heart disease. All of these features were considered while building 3 different machine learning models to predict heart disease and identify the most important variables contributing to this cardiac problem. The best model predicts heart disease with an accuracy of **76%** and AUC-ROC (Area under the curve using the Receiving Operator Curve) of **0.8417.** After performing our statistical analysis, we can conclude that the top 3 features that have the most impact on heart disease are *Difficulty in Walking*, *Age*, and lastly, *Stroke*. Therefore, we highly recommend the younger generation from the U.S. population routinely check based on these attributed features that can significantly impact obtaining a cardiac problem as they age.

Introduction

Americans are at risk of heart disease! According to CDC Heart Disease Facts report, heart disease is the top cause of mortality for men and women of all races in the United States, including but not limited to White, Black, Hispanic, and Asian Americans. It is a commonly heard term, but it encompasses a range of different heart conditions. According to CDC Heart Disease report, the most common type of heart disease is coronary artery disease or CAD. To make it more relevant, the CDC has presented statistics that mention: “About 18.2 million adults aged 20 and older suffer from CAD and roughly 2 out of 10 deaths happen for those less than 65 years old.” These numbers are definitely worth pointing out because there is no better solution than preventing heart diseases. Some of the following factors might lead to someone becoming vulnerable to heart attacks: high blood pressure or cholesterol, smoking, excessive alcohol drinking, or lack of physical activity. Our team, as data scientists, want to urge the younger generation of U.S. residents to become aware of the risks that certain conditions might possess in the years to come. We want to advocate and spread awareness through data on the importance of self-care, routine checks with a doctor, and the need to visit a cardiologist. We need to help the upcoming generations to diminish heart disease cases!

Through this project, we will develop distinct machine learning models based on the knowledge acquired in this course throughout the Spring quarter. Our objective is to predict accurately heart disease and obtain a set of features related to heart disease that will be the most important ones in leading to heart disease. We will perform a comparison of model performance and consequently use the most optimal to make recommendations. We are confident we can spread

awareness of early stages of self-prevention with the final objective of diminishing future heart disease cases in the United States.

Data Characteristics

Our dataset was downloaded from Kaggle website. It contains 319,795 data entries, which corresponds to actual data collected by the CDC from US residents via telephone surveys. It has 18 columns where 17 columns are the predictor variables and 1 is the predictand variable (heart disease). For this problem, the dependent variable is a binary outcome as “Yes” or “No”. Therefore, this is a classification problem. The following table displays the number of categorical and continuous variables in our dataset (excluding the dependent variable).

Table.1 Dataset Overlook

Type of Variables	Number of Variables
Categorical	13
Continuous	4

a) Data Cleaning

During data preprocessing, we have found that the dataset is **highly imbalanced** - with only **8.5%** of the samples are with heart disease. In order to solve this imbalance, we applied **Bootstrapping** to build up a sample dataset which has the same proportion of both ‘heart disease’ and ‘Not Heart Disease’ for the modeling.

Considering there are many categorical variables which we could not directly put into the model, we have used label-encoding in Python Scikit-learn package to convert the labels into a numeric form.

b) Univariate Analysis

Based on the **Exploratory Data Analysis (EDA),** we found an interesting summary of the following relationships:

1. BMI (Body Mass Index) vs Heart Disease distribution

Usually, the normal BMI range is regarded as 18.5~24.9. From the BMI distribution plot of people who have heart disease (see **Appendix 1.1**), we can see that the median BMI is nearly 30, which is higher than the normal range. Besides, we have calculated the percentage of heart disease people with an abnormal BMI (out the range of 18.5~24.9) is as high as 77%.

2. Frequency of demographic traits

Next, we take a look at the demographic traits of people who have heart disease, be it gender, age, and race (see **Appendix 1.2-1.4**). From the frequency plots, we can see that among all the people with heart disease, there is a higher percentage of male. As for race, we can see that white people takes the largest portion. Furthermore, the age frequency plot shows that older people take a large proportion of the ones who have heart diseases.

However, as all the datapoints come from 2020 CDC survey, there is selection bias as we only have obtained responses to the survey for individuals who were interested in participating.

3. Frequency of sleep times

For sleep times (see **Appendix 1.5**), we can see that for people who have heart disease, most of them actually have fair sleep times (~8hours/day).

4. Frequency of other diseases

For people who have heart disease, we want to know if they tend to have other diseases as well. From the plots in **Appendix 1.6-1.9**, we can see that the rates distribution of Asthma, Kidney disease, Skincancer, and Diabetic is not showing a linkage with heart disease.

c) Bivariate Analysis

1. Correlation plot of input features

It is worth pointing out that among the correlation plot (see **Appendix 2**), we found that there is a positive correlation of **0.44** between the variable **DiffWalking** and **Physical Health**.

Because our data contains categorical variables, we need to manipulate them. For Logistic Regression, dummy variables were created for the categorical variables while for Random Forest and XGBoost algorithms label encoders was utilized.

Model Development & Results

We built model to predict heart disease using CDC 2020 dataset. The cleaned sample dataset were splited into train and test dataset to develop the prediction models. We start from simple supervised model - logistic regression, then step into ensembled tree-based models Random Forest and XGBoosting. As our target is to **accurately identify the person with heart disease**, apart from accuracy, we would also pay attention to the **precision** and **recall** metrics, in order to validate the model according to our target.

Model 1: Logistic Regression

As the heart disease prediction is a binary problem, we started by a simple supervised learning model - logistic regression. We used all features from the sample dataset into predicting heart disease and validated the model performance using cross-validation.

The out-of-sample precision and recall are both 76%, which are all fairly high. And the average cross-validation accuracy is 74.2%, which is also acceptable.

Table.2 Logistic Regression OOS Performance Metrics

OOS Accuracy	ROC-AUC	OOS Recall	OOS Precision
76%	0.8323	76%	76%

Model 2: Random Forest

On top of the logistic regression, we built a random forest model (n_estimator=20, max_depth=6), which is a tree-based model to capture the non-linear trend in the dataset. The out-of-sample precision and recall are 74% and 81%, respectively, which are all fairly high. The ROC-AUC is as high as 0.8344, which is better than 0.8323 in logistic regression.

The average cross-validation accuracy is 75.4%.

Table.3 Random Forest OOS Performance Metrics

OOS Accuracy	ROC-AUC	OOS Recall	OOS Precision
76%	0.8344	81%	74%

Model 3: XGBoosting

Apart from the ‘Bagging’ model, we also built a XGBoosting model (n_estimator=20,

max_depth=6). The out-of-sample precision and recall are 74% and 82%, respectively, which are high. The ROC-AUC is as high as 0.8376, which is highest among three models (0.8323 in logistic regression, 0.8344 in random forest).

The average cross-validation accuracy is 75.9%, also higher than random forest.

Table.4 XGBoosting OOS Performance Metrics

OOS Accuracy	ROC-AUC	OOS Recall	OOS Precision
76%	0.8376	82%	74%

Overall, we can see that the XGBoosting has a best OOS performance among three models, according to ROC-AUC, recall, precision, and accuracy. Therefore, **we choosed XGBoosting as the best model, and tune it to further improve the performance.**

We applied a **Grid-Search method** in Python to tune the hyperparameters of XGBoosting model, including learning rate, maximum depth, number of estimators, etc. Then, we used the best parameter from grid search result to build the best XGBoosting model. With the best XGBoosting model, we have achieved a **ROC-AUC as high as 0.8417**, and **PR-AUC 0.8223**.

Feature Importance

We extracted and visualized the feature importance from the best prediction model using feature importance attribute from Python Scikit-learn package. From the feature importance plot (see **Appendix 3**), we can see that the most important feature to predict heart disease is: **whether the person has difficulty to walk.** The second most important attribute is the person's **age** and the third most important one is **stroke**.

Recommendations

After performing in-depth analysis, we identified the factors had the most significant impact on a person getting ill with heart disease. The most important was **Difficulty Walking**. The second one was **Age** and the third one was **Stroke** (where the brain tissues do not receive enough oxygen and nutrients). These factors are important information, especially to younger generations in the U.S. population, to take health care checkups seriously at an early stage. In particular, we recommend that **both children and adults to maintain a healthy and balanced diet accompanied by physical activity**. This will help to maintain a healthy weight (normal range BMI) that would prevent one from having difficulty walking. Additionally, **people should take charge of their medical conditions and perform routinely checks (blood pressure, cholesterol, manage diabetes if applicable, etc.) to prevent strokes**. Then, they can decrease the probability of getting heart disease in the future. By following these two recommendations, the future generations can curtail the heart disease cases in the United States.

Summary & Conclusion

As is well known, Americans are at risk of heart disease. For our project, we first explored how certain features related to heart disease, developed three machine learning models to predict whether a person becomes ill with heart disease accurately, and finally analyzed what factors are more likely to contribute to heart disease.

As the data is highly imbalanced, we leveraged Bootstrapping method to build up the sample dataset. Then, we established Logistic Regression, Random Forest, and XGBoosting models and

performed model comparison. In terms of model selection, we focused on precision and recall metrics, as these metrics are appropriate for measuring heart disease prediction. Precision refers to the True Positive predictions (TP) of a person having heart disease divided by all positive predictions (P*) whereas recall has the same numerator (TP) divided by actual cases of heart diseases (P). The Logistic model had a precision and recall of 76%, The Random Forest and XGBoost had a precision of 74%, and the recall for Random Forest was of 81% and 82% for XGBoost. As for the AUC-ROC, XGBoost was the winner (0.8376) followed by Random Forest (0.8344) and concluding with Logistic Regression (0.8323). Among the three competitors, XGBoost slightly outperformed the others. Hence, we chose the XGBoost model and performed hyperparameter tuning with Grid Search method. The ROC-AUC increased to 0.8417. Based on this model, one can predict whether a person can get heart disease using all the features from the dataset. We then performed feature importance to see which factors had the most significant impact on getting ill, and provided meaningful and actionable suggestions according to the analysis and modeling result.

In conclusion, our analysis helps raise awareness of the risks that certain conditions might pose in the years to come. We are pleased to leverage the data analysis tactics and machine learning knowledge to help future generations prevent heart disease.

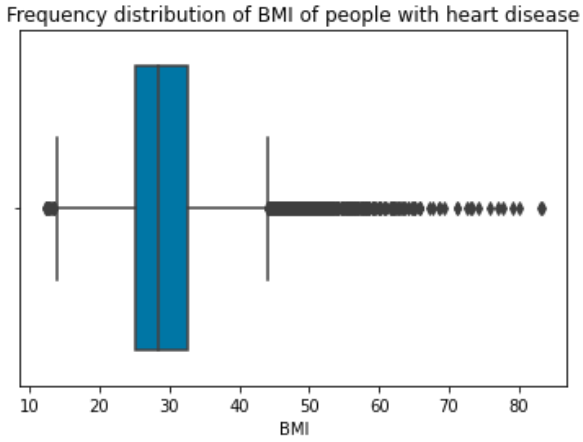
References

1. Centers for Disease Control and Prevention. (2022, February 7). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved March 13, 2022, from https://www.cdc.gov/heartdisease/facts.htm
2. Pytlak, K. (2022, February 16). *Personal key indicators of heart disease*. Kaggle. Retrieved March 13, 2022, from https://www.kaggle.com/kamilpytlak/personal-key-indicators-of-heart-disease?select=heart_2020_cleaned.csvkaggle.comkaggle.com

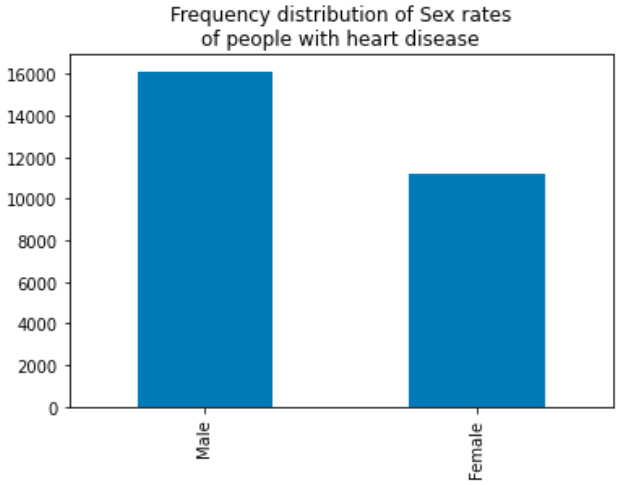
Appendix

1. Univariate analysis

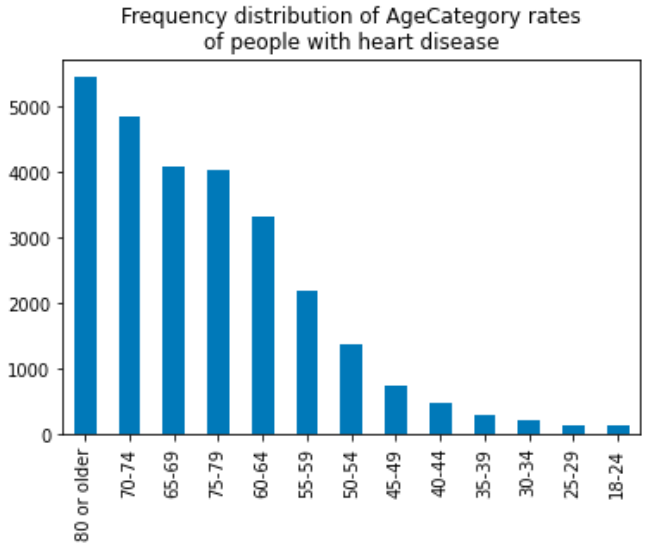
1.1 BMI distribution plot of people with heart disease



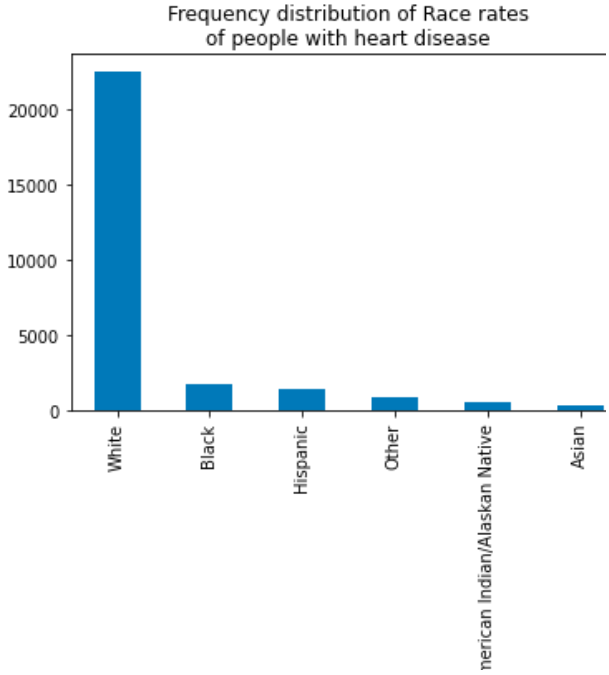
1.2 Gender distribution plot of people with heart disease



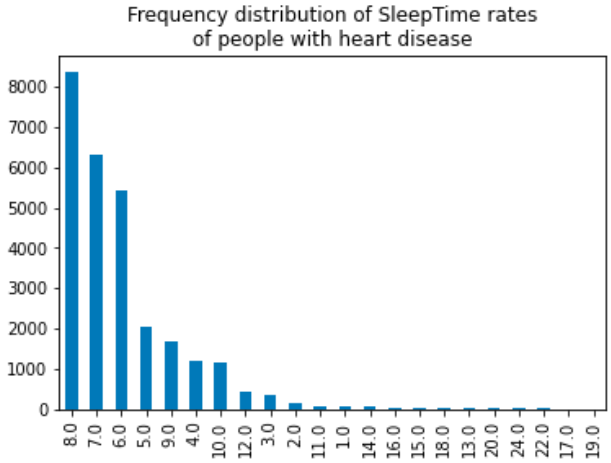
1.3 Age distribution plot of people with heart disease



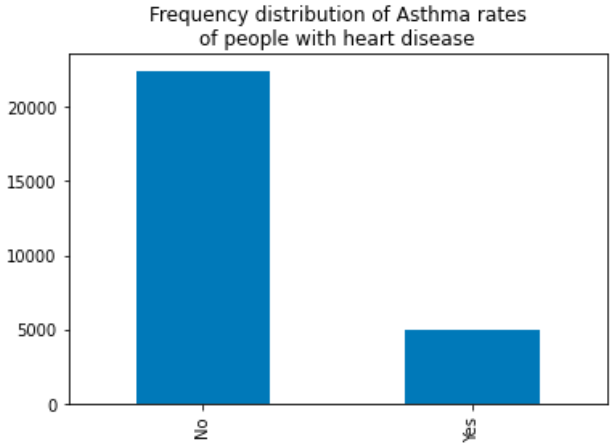
1.4 Race distribution plot of people with heart disease



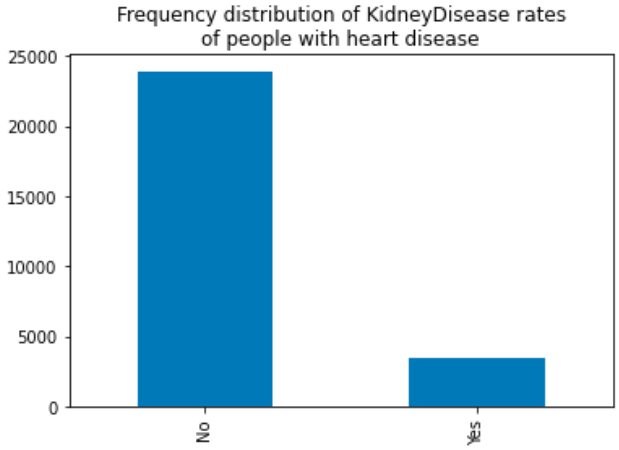
1.5 Sleep time distribution plot of people with heart disease



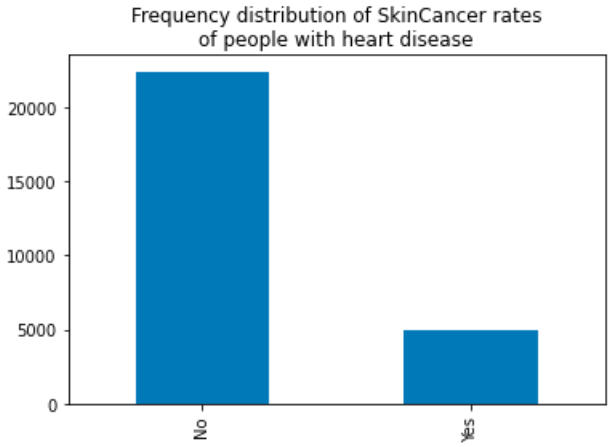
1.6 Asthma distribution plot of people with heart disease



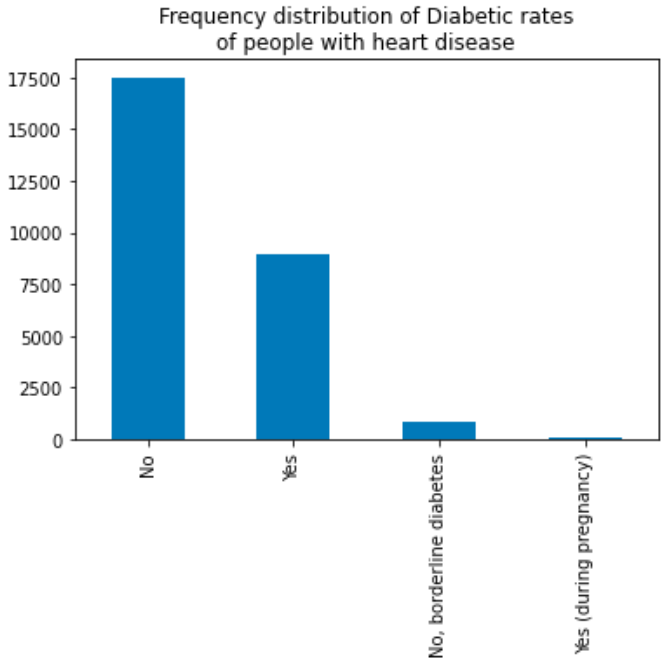
1.7 Kidney disease distribution plot of people with heart disease



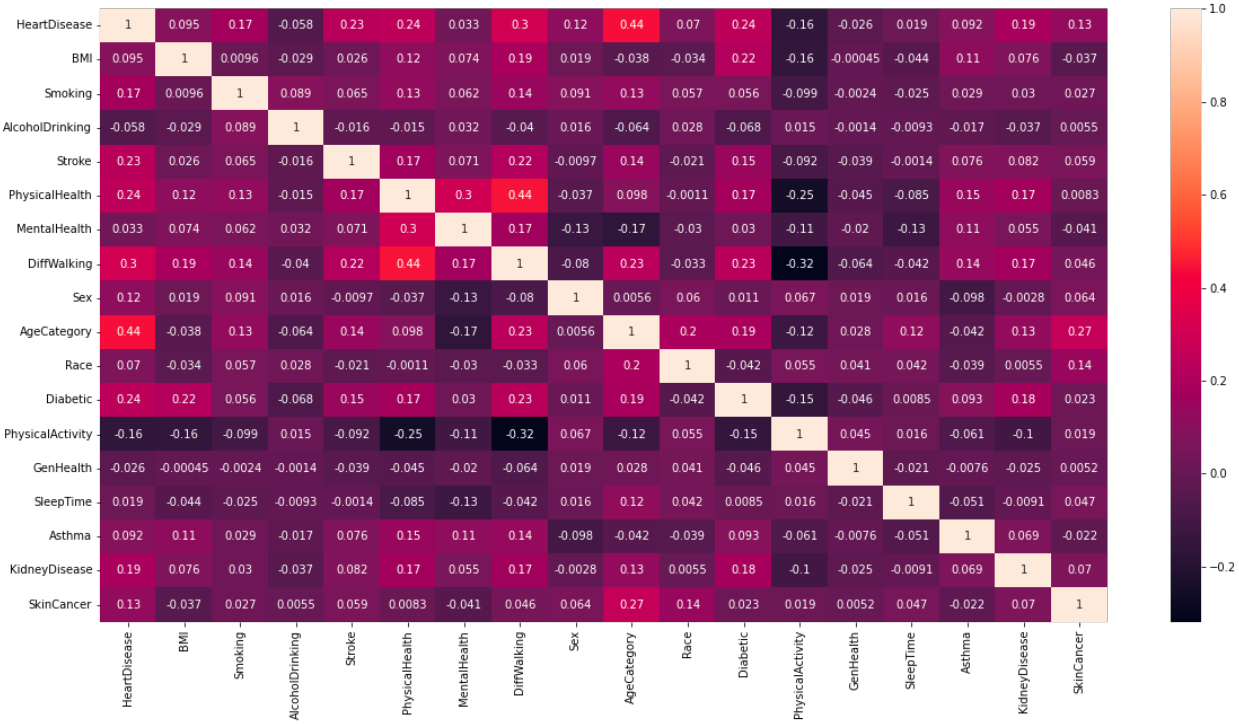
1.8 Skincancer distribution plot of people with heart disease



1.9 Diabetic distribution plot of people with heart disease



2. Bivariate analysis



3. Feature Importance of XGBoosting