CapstoneProject_Group3
Kwadwo Asante
Marco Lopez
Malachai Cravens
Yarely Vargas

Heart Disease Predictor Analysis
Executive Summary

Introduction:
According to the CDC, heart disease is one of the leading causes of death for most races in the US. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking.The data we selected from Kaggle comes from the 2020 annual cdc survey data of 400k adults. Personal Key indicators include: BMI, Smoking, Alcohol Drinking, Stroke, Physical health, Mental health, Difficulty Walking, Sex, Age Category, Race, Diabetes, Physical Activity, Gen Health, Sleep Time, Asthma, Kidney Disease, Skin Cancer.

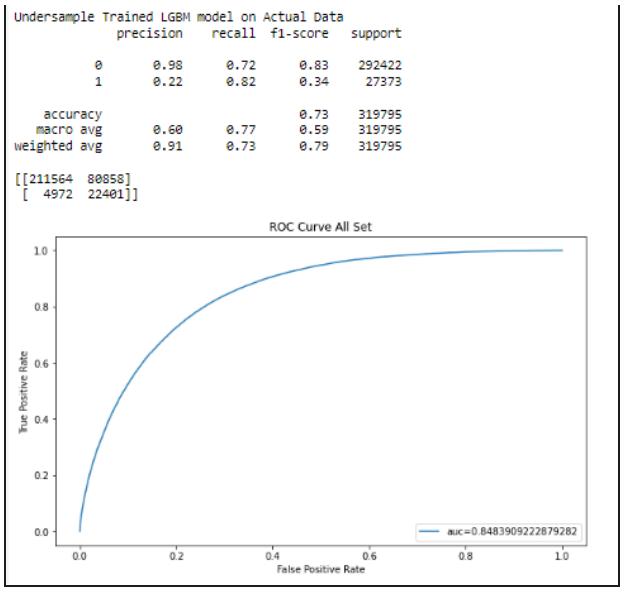
Machine Learning Models & Data Cleaning:
After loading our data, we found a count of 319,795 rows of data that were heavily sided with respondents who did not have heart disease (Figure 1). This was an issue we saw early in the process, as imbalanced datasets cause most machine learning techniques to have poor performance on the minority class (which the minority class was the most important part in our modeling).After searching through various imbalanced classification methods, we opted to train our models using random undersampling.

```
#Class counts for whether respondent had heart disease
df_og.HeartDisease.value_counts()
```

```
No     292422
Yes     27373
Name: HeartDisease, dtype: int64
```

(Figure 1: Snapshot of class imbalance)

After training different models, and seeing its performance on testing data.. Specifically, our group focused on the heart disease recall percentage and the AUC. As a result, our undersampled data performed best with the LGBM model with an AUC of .8491 and a heart disease predictor recall of 82%.The AUC was .0086 higher when running it on the real data than the results given when training the model (Figure 2).



(Figure 2: Snapshot of Logistic Regression model results)

Tableau

Tableau Public was used to visualize our data so that we understood what our data set contained as far as values and trends. Our color scheme utilizes a marron-like red to black/grayish color scale that helps us see how many of the Residents within our visuals that did say they had heart disease to those that do or do not have a Heart Disease. Our “Demographic Dashboard” dashboard can be used by the filters set on the page, where users can filter by, whether “Heart Disease " is yes or no, the kind of “Race” or ethnicity that is found in our data set, as well as the BMI Level of our Residents. Our “Health Scaled Dashboard” can be filtered and the Mental and Physical Health can be filtered by the number of days reported that the residents did not feel well. We were able to see that of the People that did have a heart disease, a majority of those people were Males, and most of the residents that responded for the survey were white people by over 80%. We also found that based on the BMI index of our residents, we found a lot of them were obease or overweight. And most of them were of the age 65 and older as well that did have a heart disease.

Website Deployment.

The website was designed as a resource for visitors who are curious to explore the health status of U.S. residents and those who wish to explore whether our model will predict heart disease. The questionnaire on the website which predicts whether someone has heart disease based on the answers is similar to a survey form. At the end there is a “Make Prediction” button that displays prediction. Design of the website was inspired by the heart disease website, really bringing out the shades of red and gray.

Conclusions

In the future, we hope to be able to use a model trained with oversampling with our data. However, due to deployment to Heroku and limited space we had to use the model trained on undersampled data. It would be interesting to see how well our chosen LGBM model would perform on the dataset of 2021. We hope that visitors to the website gain value in determining their health and explore how their health compares to the health status of respondents using our tableau dashboard.