

CapstoneProject_Group3

Kwadwo Asante

Marco Lopez

Malachai Cravens

Yarely Vargas

Heart Disease Predictor Analysis Report

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. Our group is interested in learning about what key indicators lead to heart disease. 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.

Inspiration:

Driven by the silent approach heart disease has on its victims, we took it upon ourselves to search for relations within the indicators given above. We all run the risk in a variety of ways, whether it be not getting enough sleep, balanced diets, polluted environments, etc. that could ultimately curate many forms of heart disease from cardiomyopathy, congenital heart disease, pericardial disease, to heart failure. The nature of the topic is critical which warranted our response and drove us to evaluate a combination of underlying reasons that would exacerbate this disease. Evaluating ourselves with the help of data led to our interest in building a model that could send us in the right direction if an individual runs the risk of heart disease to then begin taking proactive measures to combat this silent killer.

Hypotheses:

With the data set given, our group managed to curate numerous ideas as to what may be the leading factors that cause heart disease. Reviewing our data set we found that High Blood Pressure, Diabetes, Smoking, Obesity, Physical Inactivity are all factors that exacerbate heart disease. As humans we are prone to countless infections: having the wherewithal gives us data scientists the ability to form numerous variables to how we can take preventative measures towards this disease. We predicted that those who smoke, do not get enough exercise and do not have enough sleep were all indicators of someone who could have heart disease. Through our research, we found that there are so many more combinations of factors that could lead to having heart disease.

Machine Learning Models & Data Cleaning:

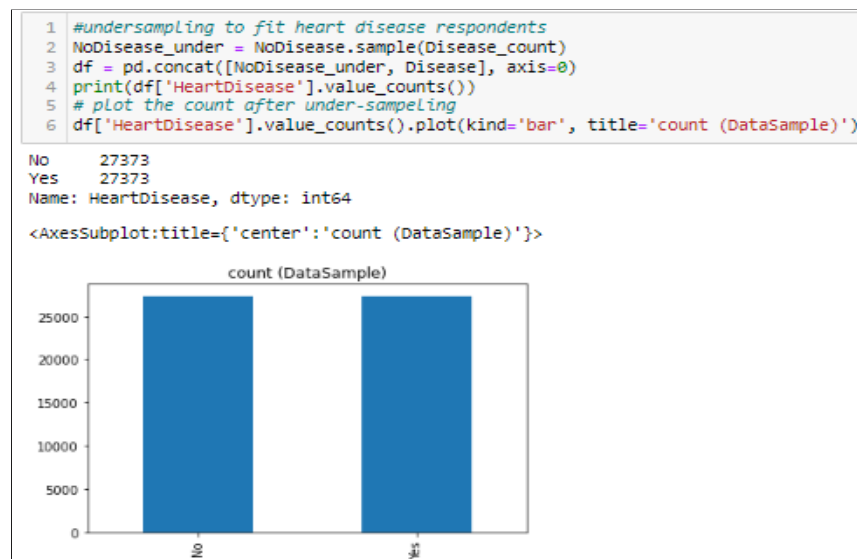
After loading our data, we found a count of 319,795 rows of data that were heavily sided with respondents who did not have heart disease (Figure 1). This was an issue we saw early in the process, as imbalanced datasets cause most machine learning techniques to have poor performance on the minority class (which the minority class was the most important part in our modeling). After searching through various imbalanced classification methods, we opted to try random undersampling, random oversampling, and SMOTE (Synthetic Minority Oversampling Technique).

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

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

(Figure 1: Snapshot of class imbalance)

Through our search, we found pros and cons for using each of the sampling methods. For example, random undersampling randomly deletes examples from the majority group up to the count of the minority group. Our team favored this method because it was very easy to understand, and the result was true to the data the team originally had minus some rows of data from the majority class (Figure 2). Random Oversampling randomly duplicates the minority group to match the majority group (Figure 3). Our team was hesitant for this approach because of the duplication aspect of the data. Our concern for this approach was that given the duplicated data to our ML models, the model would potentially perform really well with our dataset but not when predicting on new data. SMOTE uses an improved approach, which synthesizes new examples by selecting a minority class instance at random and finding its K nearest minority class in order to generate a combination of the two chosen instances (Figure 4). Our team was hopeful of this method as the idea of finding little groups within our minority class and creating new data (instead of duplicating) could give our ML models a large amount of data to train on and make accurate predictions.



(Figure 2: Snapshot of Undersampling)

```

1 from imblearn.over_sampling import RandomOverSampler
2 from collections import Counter
3
4 ros = RandomOverSampler(random_state=42)
5 X_resampled, y_resampled = ros.fit_resample(X_train, y_train)
6
7 Counter(y_resampled)
Counter({0: 219418, 1: 219418})

```

(Figure 3: Snapshot of over sampling with RandomOverSample)

```

1 # Resample the training data with SMOTE
2 from imblearn.over_sampling import SMOTE
3 X_resampled_smote, y_resampled_smote = SMOTE(random_state=1).fit_resample(
4     X_train, y_train
5 )
6 Counter(y_resampled_smote)
Counter({0: 219418, 1: 219418})

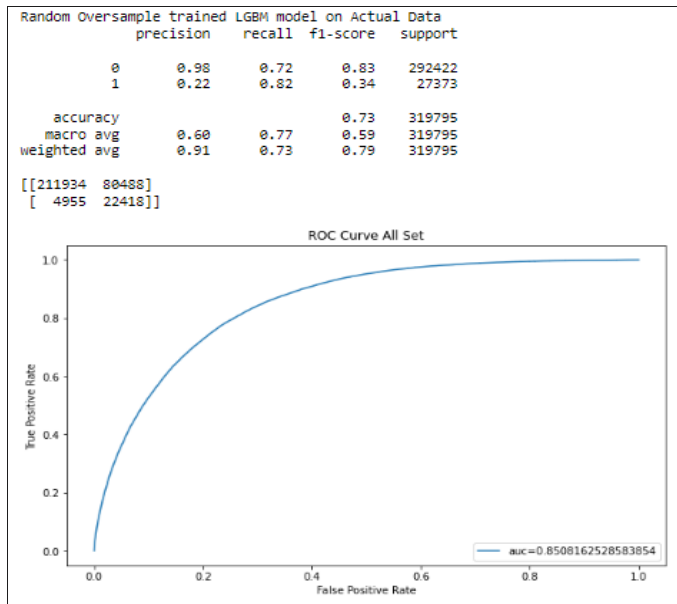
```

(Figure 4: Snapshot of over sampling with SMOTE)

Due to limited time, our group focused on using RandomForest, LogisticRegression, LGBM, XGBC, and ADABOOST models to train our data. After training the models, and seeing its performance on testing data, we chose the best models to run all of our data set and see how well it performed or if there would be discrepancies. Specifically, our group focused on the heart disease recall percentage and the AUC. Recall would give us how well the model performed when predicting heart disease against those that actually had heart disease. AUC would give us the probability of which the model correctly made the predictions.

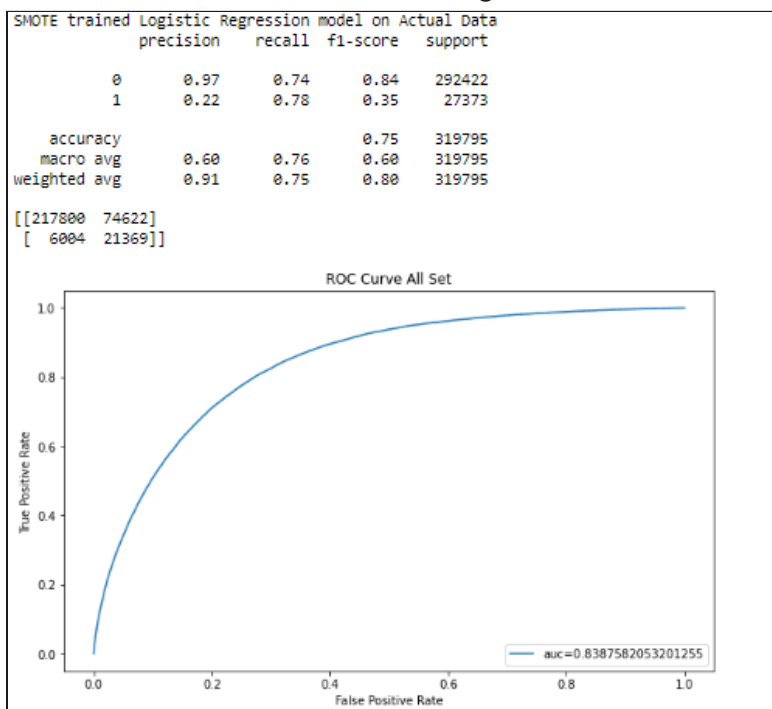
Below are the results:

- Random oversampled data performed best with the LGBM model, with an AUC of .8508, and a heart disease predictor recall of 82% on our real data (Figure 5). The AUC was .0112% higher when running it on the real data than the results given when training the model.



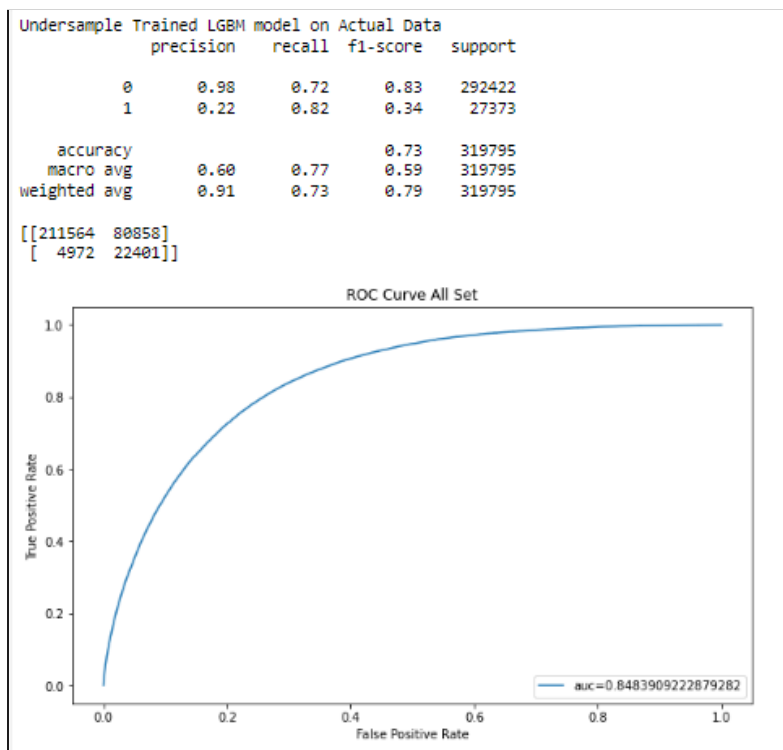
(Figure 5: Snapshot of LGBM model results)

- SMOTE sampled data performed best with the logistic regression model with an AUC of .84, and a heart disease predictor recall 78% recall (Figure 6). This model performed the same on the actual data as it did with the training data.



(Figure 6: Snapshot of Logistic Regression model results)

- 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.



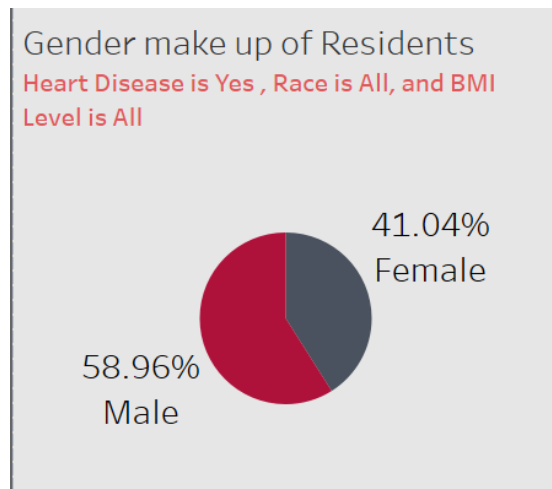
(Figure 6: Snapshot of Logistic Regression model results)

Overall, the models had an AUC between .77 and .84 with both undersampling and oversampling data. We found it interesting that the Randomforest model, when trained with random oversampling, had an AUC of .77 but when used with the actual data the AUC went up to .99. This is probably a result of overfitting. This model would probably fail to predict new data sets reliably. The model we chose for our website is the LGBM model trained with undersampled data. We decided to go with this model because it had the most consistent/highest results for heart disease recall and AUC during training and with our actual data.

Tableau

Tableau Public was used to visualize our data for easy initial analysis of our data set such as distributions of features and trends. Our color scheme utilizes a marron-like red to black/grayish color scale that helps us see how many of the respondents within our visuals said 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 respondents. 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 (Figure 7), and most of the respondents for the survey were white people by over 80% (Figure 8).

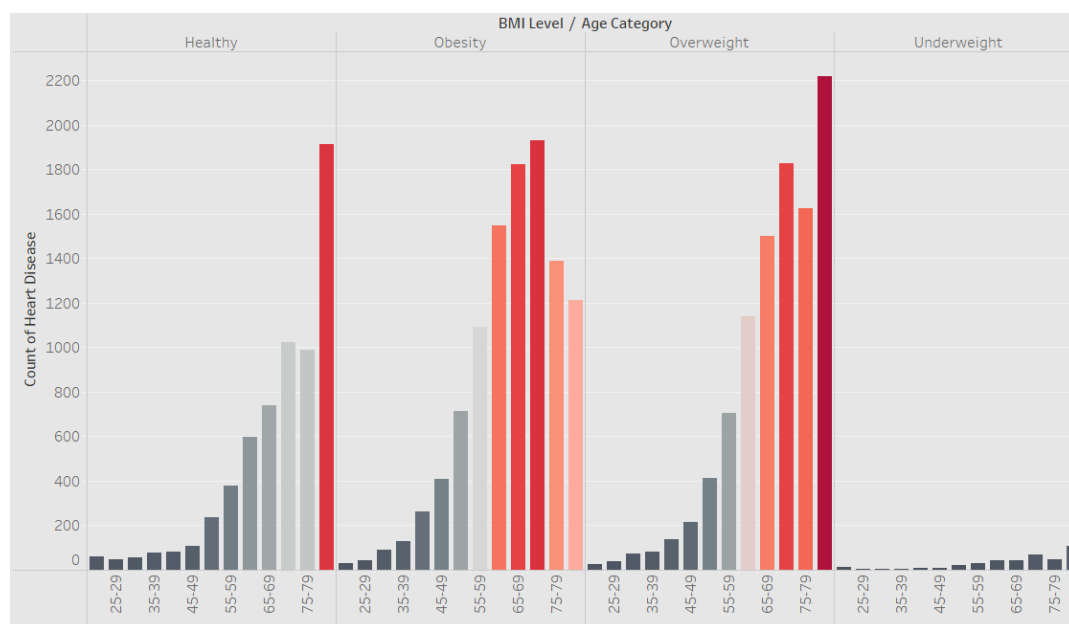


(Figure 7: Snapshot of gender pie in Tableau)



(Figure 8: Snapshot of Race distribution in Tableau)

We also found that based on the BMI index of our respondents, we found a lot of them were obese or overweight. And most of them were of the age 65 and older as well that did have a heart disease. The difficult part in setting up the Tableau visuals was understanding our data and recognizing we had a lot more Dimensional (Categorical) columns than we do Measure (Numerical) columns (Figure 9).



(Feature 9: Snapshot of the Age vs BMI category viz in Tableau)

Website Deployment.

The website was designed as a resource for visitors who are curious to explore the health status of U.S. respondents and those who wish to explore whether our model will predict heart disease (Figure 10). The questions asked in the Prediction page are similar to the questions the BRFSS (Behavioral Risk Factor Surveillance System) uses to conduct telephone surveys and gather data which was used as data for our ML models. However with our website, visitors may answer those questions and find their probability of having heart disease. The questions are split into categories so that it is easier to fill out. Categories for the questions include: About You, Your Daily Habits, and About your Health (Figure 11).

Heart Disease



LEARN MORE ABOUT HEART DISEASE AND ITS RISK FACTORS.
IT'S IMPORTANT FOR EVERYONE TO KNOW THE FACTS ABOUT HEART DISEASE.
TO EXPLORE MORE CLICK ON "GET STARTED".

Get Started →


Tableau Viz

TABLEAU HELPS CREATE INTERACTIVE
GRAPHS AND CHARTS IN THE FORM OF
DASHBOARDS AND WORKSHEETS TO GAIN
INSIGHTS ON HEART DISEASE.

Do you have Heart Disease?

FIND OUT IF OUR MODELS PREDICTS IF YOU HAVE HEART DISEASE.

(Figure 9 : Snapshot of the home page)


Prediction

Please answer questions below to predict if you could have Heart Disease.

About you

| | |
|----------------------------|--------------------------------|
| Choose your Age | How is your general health? |
| 18-24 | Excellent |
| What is your birth gender? | What best describes your Race? |
| Female | White |

Your Daily Habits

| | |
|--|--|
| I am a Heavy Drinker?(14+ drinks men, 7+ drinks women) | # of Days I had physical illness or injury in past 30 days |
| No | 0 |
| Smoked at least 5 packs of cigarettis in my life | # of Days my Mental Health was not good in past 30 days |
| No | 0 |
| Difficulty Walking or Climbing stairs | I have exercised in the past 30 days |
| No | No |

About your Health

| | |
|--|------------------------------------|
| Do you have Diabetes? | Do you suffer from Kidney Disease? |
| No | No |
| Do you have Asthma? | Do you suffer from Skin Cancer? |
| No | No |
| Give an estimate on your BMI (body mass index) | Have you had a Stroke? |
| 28.50 | No |

Make Prediction!

(Figure 11 : Snapshot of the Prediction Page)

Conclusion:

In conclusion, using the random undersample technique to balance the data used to train the LBGT model yielded the best results overall. We would like to know how this model would perform on a similar dataset for a newer year. Also, after exploring our data we found that most of the respondents for our original data considered themselves white, thus the model might be very good at predicting for white respondents rather than another race. .

Limitations and future work

First, we would like to add a predicting percentage for the predicted outcome when a user uses our website. We believe this will give our users great insight as to how close the model predicted they would or would not have heart disease. Additionally, we left out a lot of columns that were used in our prediction model when building the tableau visualizations, and with more time, we would like to further explore how they all correlate to the prediction of our mode. We think this will give users the ability to indulge their curiosity with how all the different questions relate to their health and with other respondents of the survey. Lastly, we would have liked to deploy an LGBM model trained on oversampled data, however due to the sizing limit it was not possible to upload the model to github or Heroku.