HEART DISEASE PREDICTION

Kate Yu | April 2022 | BrainStation Data Science Bootcamp

PROBLEM STATEMENT

In every 36 seconds, one person dies in the United States from cardio vascular disease and that is one in every four deaths [1]. The diagnosis of heart disease has always been a challenge for health care organizations, because it is usually based on obvious signs, symptoms and physical examination. The goal of this project is to help people to better predict the possibility of getting heart disease based on current health status. This can be useful for each of us to improve personal health awareness, and also for health organizations to make proactive treatments and distribute medical resource efficiently.

DATA SOURCE

The data of this project was directly from Kaggle, which is a shortened version of the 2020 annual CDC (Centre for Disease Control and Prevention) survey of 400,000 adults related to their health status. The data consists of numeric, binary, multi-class variables that show the status of health, and also the target variable: if the person has heart disease or not.

DATA PROCESSING

A comprehensive data preprocessing, cleaning and EDA were performed on the dataset. The goal of this step is to check the distribution of each variable, correlation with the target variable, and convert each categorical column into a numeric column. Feature engineering was applied to some variables to make it more sense or prevent multicollinearity. Specifically, binary columns were converted into 1 or 0, and multi-class variables were converted into a series of numbers or using dummy variables. The cleaned version of the dataset was exported as a CSV file after all the preprocessing steps.

MODELING

Based on the newly cleaned dataset, a couple of modeling techniques have been applied. Logistic Regression, Decision Tree, Random Forest, and K Nearest Neighbors are the main models I have used within Pipeline. Since the dataset is heavily imbalanced based on heart disease rate, the up-sampling technique: SMOTE and down-sampling technique: Random Under Sampling were used to deal with the high false-negative problem. The goal of doing modeling is to find the best model based on recall score and F1 score because I hope to capture as many people who might have heart disease as possible. Here is the summary table of the top selected model:

Rank	Model	Recall	F1	Accuracy
1	Logistic Regression + Under Sampling	0.78	0.35	0.75
2	Logistic Regression + SMOTE	0.57	0.32	0.79
3	Decision Tree	0.67	0.26	0.68

Table 1: Top selected models with statistic scores

FINDING AND CONCLUSION

Based on the odds ratio from optimal logistic regression, Figure1 shows the positive or negative relationship between each feature and heart disease..

To summarize, they key findings of this projects are as follows:

- The top 5 indicators of having heart disease are: age, stroke, smoking, have diabetes, and kidney disease.
- The top 2 habits that help to decrease the possibility of getting heart disease are: increase sleeping time and drinking alcohol occasionally.
- Being female and Black or Asian also naturally have less possibility of getting heart disease.

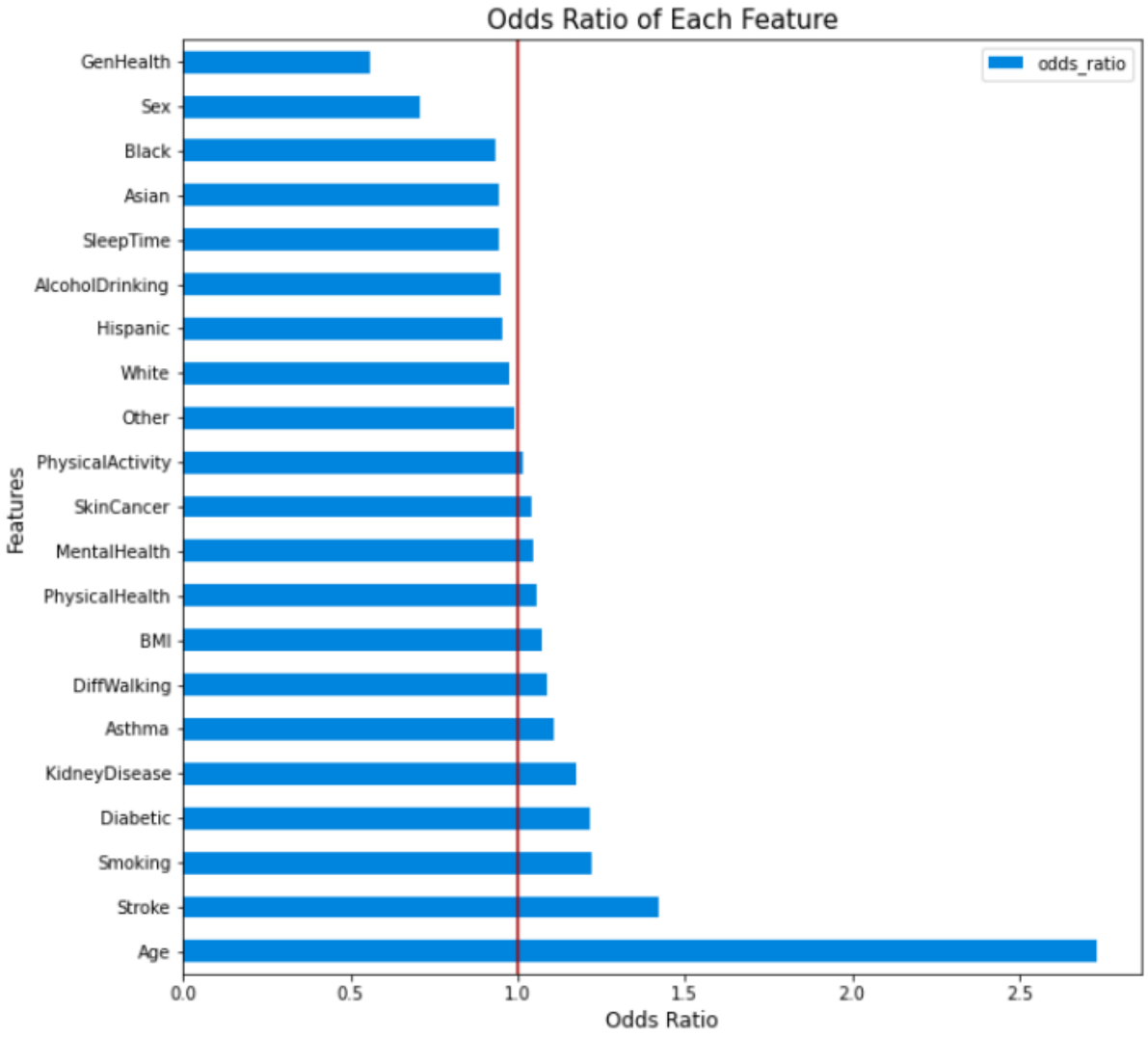

Figure 1: Odds ratio of each independent variable from Logistic Regression Model

NEXT STEPS

As a next step, I would like to look more into the multicollinearity of each feature to get more accurate coefficients. Additionally, checking the original dataset from CDC's official website is another way to add more useful features to increase the prediction power. Finally, since a more balanced dataset will help to increase the precision and recall scores, instead of using the sampling technique, I would like to find a more balanced dataset and refit the models.

[1]: “Heart Disease Fact” , *CDC government*