Math 218: Final Report

Team-CN

Christina Chen, Siyuan Niu

12/16/2022

Introduction

Heart disease is one of the leading causes of death in the United States. About 697,000 people in the United States died from heart disease in 2020 — that is one in every five deaths. In order to better inform the public about the risk factors for heart disease, the CDC named high blood pressure, high cholesterol, and smoking as three critical factors. However, there are other medical conditions and lifestyle choices, including diabetes, physical inactivity, excessive alcohol use, etc., that can put people at a higher risk for heart disease. Therefore, physical health indicators and demographic information besides the three key factors are often equally imperative in predicting whether an individual has heart disease. This final project aims to answer two research questions: 1) What combination of demographic and physical health indicators yields the least misclassification rate in lasso logistic regression? 2) Among the best lasso logistic regression, a pruned decision tree, and a Naive Bayes mode, which model performs best in predicting heart disease?

Data used to answer the two research questions came from the CDC's annual telephone survey, which surveyed U.S. residents' health status. The observation unit in this 2020 CDC survey dataset is individual, with a total of 319,795 observations. Although the original dataset contained nearly 300 variables, we reduced the dataset to only 17 variables for analysis and simplicity purposes. Some health indicators included body mass index, alcohol consumption status, physical activity, and only one of the three key risk factors - smoking. Other categorical variables that remained in this dataset are gender, age, and race. Additionally, the response variable is a binary variable that indicates whether a person has heart disease.

EDA

The unbalanced nature of this dataset is shown in *Table 1*, as only 8.65% of the total individuals reported having heart disease. Body Mass Index (BMI), Sleep Time, Mental Health, and Physical Health are the four continuous variables in this dataset, reflecting each individual's health status.

	No-Heart Disease	Yes-Heart Disease
Proportion	91.44%	8.65%

Table 1. **The proportion of Respondents w/o Heart Disease.**

On the one hand, some health and demographic predictors have already demonstrated a significant variation between people wsho claimed to have heart disease and those do did not have it through EDA. In particular, *Figure 2* showed that people who reported having heart disease exhibited a more significant number of days to have not good physical health during the past 30 days than their counterparts who did not have heart disease. In addition, a higher proportion of people claimed to have heart disease among people whose ages fall into the following three ranges: 70-74, 75-79, and 80 or older, as shown in *Figure 3*. Smoking is another effective indicator of predicting heart disease, underscored in *Figure 3*, as among the 8.65% of people who had heart disease, a higher proportion of them claimed to smoke at least 100 cigarettes (5 packs) in their

entire life. On the other hand, summary statistics and EDA failed to provide any more helpful information regarding what are the most reliable health factors that can be used to predict heart disease. Hence, further research is necessary to address these research questions.

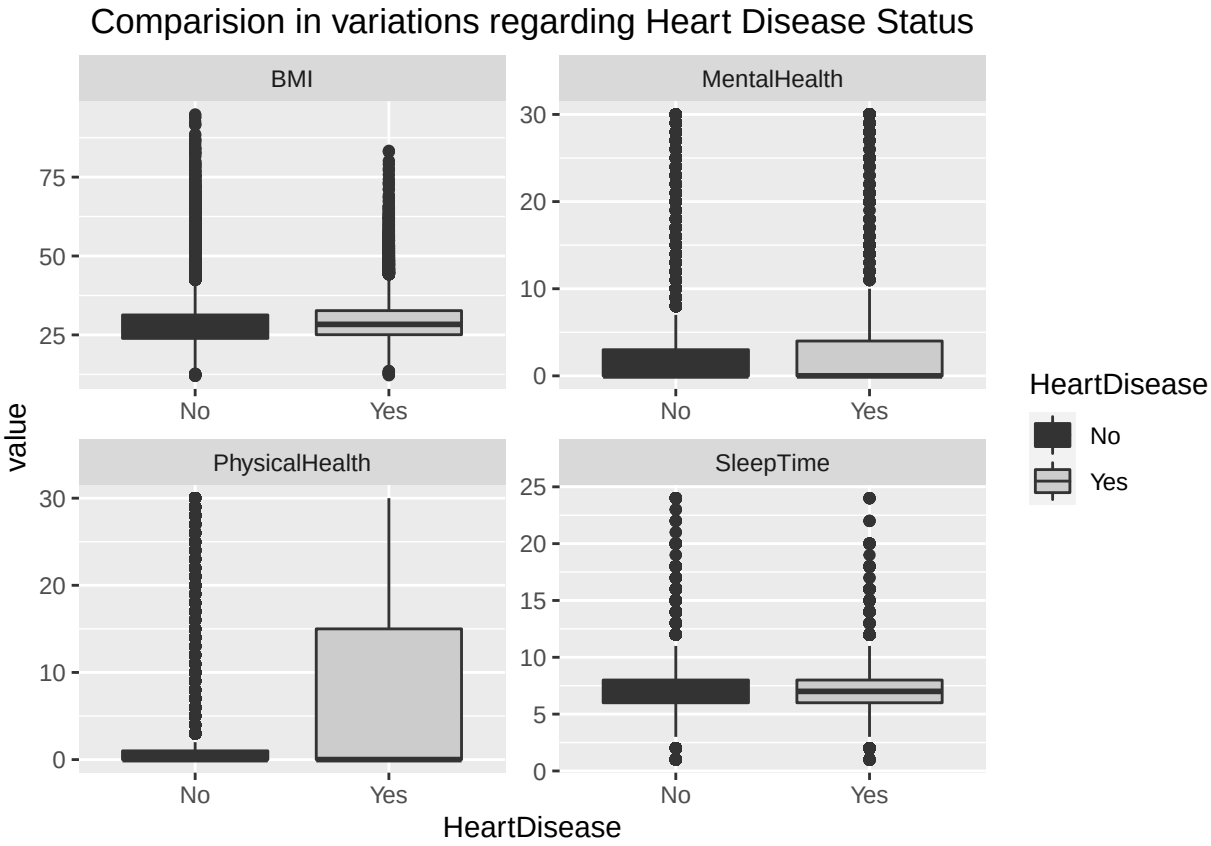

Figure 1. A comparison of the variations in Heart Disease Status across four continuous variables. No obvious difference was observed between people who had heart disease and people who did not for BMI and Sleep Time. Nevertheless, a higher proportion of people with heart disease reported having more days of bad physical health conditions.

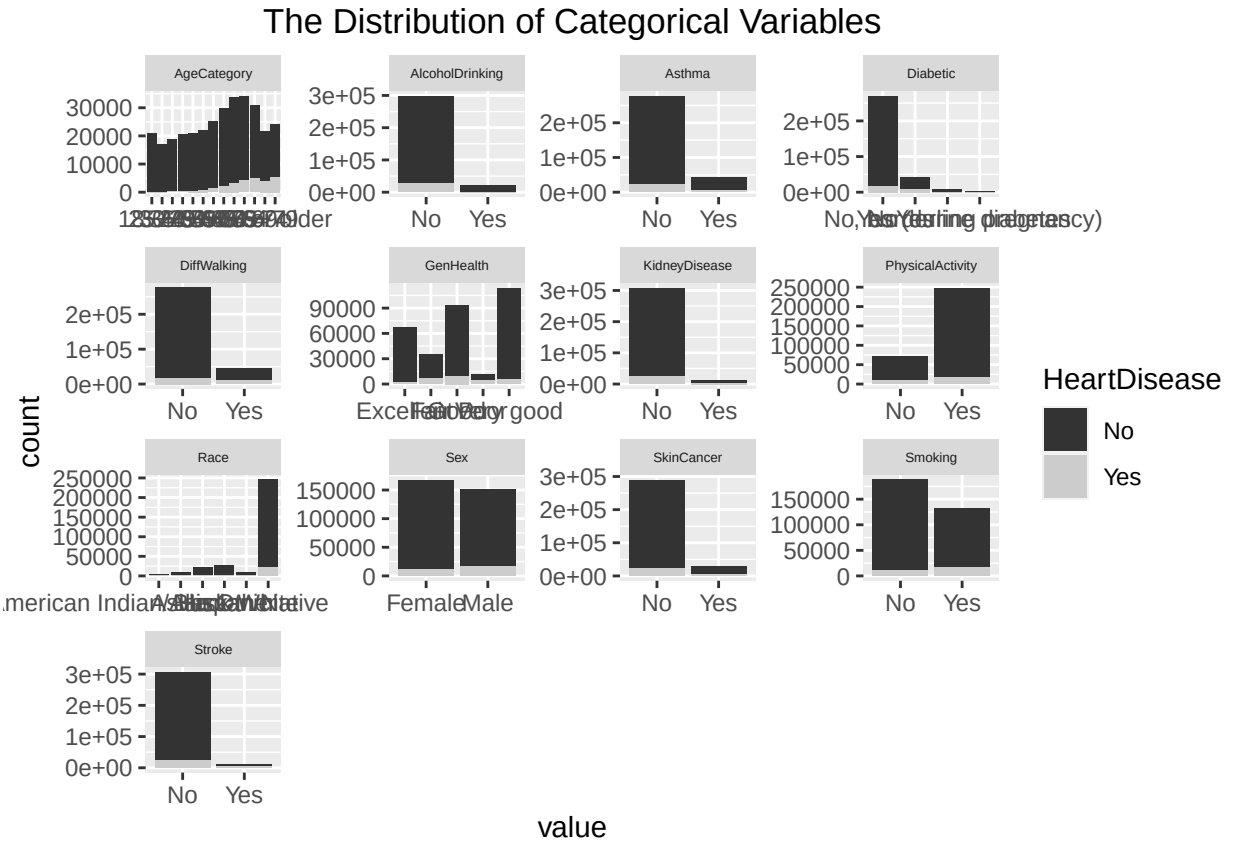

Figure 2. A comparison of the variations in Heart Disease Status across thirteen categorical variables. Age Category and Smoking are effective predictors demonstrating evident variations in the Heart Disease Status.

Methodology

The purpose of this section is to develop methods for obtaining the optimal model for the binary classification task of heart disease. The logistic is to divide the dataset into train/test splits, perform model hyperparameter tuning and obtain the optimal model for lasso logistic regression, decision tree, and Naive Bayes, and finally select the best model out of the three models based on yielded error rate.

Given the imbalanced nature of the dataset, we first obtain the "ids" of observations with and without heart disease separately and perform a random 60-40 train-test split on the full dataset. Then we combine the train "ids" of the two sets with 60 percent of the whole dataset and the rest test "ids" as the remaining 40 percent.

We then perform parameter tuning of the models. The optimal lambda parameter of the lasso logistic regression is selected by running k-fold cross-validation using the training dataset. Specifically, the k is set to 10, and the random seed is set to 10. To figure out whether there exists an optimal level of tree complexity, we first fit a classification decision tree on the training set. Then, we run another 10-fold cross-validation using this trained model to find the optimal. Using the misclassification rate of pruned and unpruned trees, we can determine which was the best among the two. Lastly, no parameter tuning is done for the Naive Bayes model.

After training each model on the same training set, we compare the accuracy of the three models based on their misclassification rate on the test dataset. In particular, we will calculate the misclassification rate by using the confusion matrix of each model.

One advantage of obtaining the lasso regression model is that it can give us insights into the significance of each indicator. By running lasso logistic regression on the entire dataset, we can obtain the set of significant and irrelevant indicators by looking at estimated regression coefficients. Specifically, a zero coefficient would indicate the corresponding indicator is relatively irrelevant to consider.

Results

The optimal lasso logistic regression with its best lambda value chosen by 10-fold cross-validation contained twelve of the seventeen predictors. The combination of predictors that produce the least misclassification and their corresponding coefficients are shown in *Table 2*.

According to *Table 7*, the lasso logistics regression has the lowest test misclassification rate, the Naive Bayes model has the highest test misclassification rate but the lowest test false negative rate, and the decision tree model yields the same result for test misclassification rate and test false negative *Table 5* and *Table 7*.

Notice that we use the misclassification rate of an unpruned 5-node tree instead of a pruned 1-node tree, because both has the same cross-validation error, as shown in *Table 4*. Therefore, pruning in this case fails to improve the tree's performance in prediction.

```
##                     (Intercept)                             BMI
##                    -4.647462818                     0.006522824
##                      SmokingYes               AlcoholDrinkingYes
##                     0.369356622                    -0.101397476
##                       StrokeYes                  PhysicalHealth
##                     1.033942587                     0.004827961
##                   DiffWalkingYes                         SexMale
##                     0.261754787                     0.610985457
##                 AgeCategory25-29                 AgeCategory30-34
##                    -0.752465693                    -0.444243631
##                 AgeCategory35-39                 AgeCategory40-44
##                    -0.483980090                    -0.192960885
##                 AgeCategory45-49                 AgeCategory50-54
##                    -0.035726088                     0.088145343
##                 AgeCategory55-59                 AgeCategory60-64
##                     0.366257934                     0.652459064
##                 AgeCategory65-69                 AgeCategory70-74
##                     0.895652446                     1.188640667
##                 AgeCategory75-79           AgeCategory80 or older
##                     1.353412907                     1.622888935
##                       RaceBlack                     RaceHispanic
##                    -0.024676177                    -0.006693980
##                       RaceWhite DiabeticNo, borderline diabetes
##                     0.163556702                     0.050415468
##                     DiabeticYes                    GenHealthFair
##                     0.544231736                     1.109599223
##                    GenHealthGood                    GenHealthPoor
##                     0.649311969                     1.492194177
##                GenHealthVery good                       SleepTime
##                     0.048616207                    -0.010689491
##                       AsthmaYes                  KidneyDiseaseYes
##                     0.188131718                     0.588829339
##                    SkinCancerYes
##                     0.131704377
```

Table 2. Coefficients of all the predictors that were remained in the model after performing the best lasso logistic regression.

```
##      true
## preds     No    Yes
##   No  125016  10746
##   Yes    892   1118
```

Table 3. Confusion Matrix for best lasso logistic regression.

```
## $size
## [1] 5 1
##
## $dev
## [1] 16410 16410
##
## $k
## [1] -Inf    0
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

Table 4. Result of CV and pruning for the tree model

```
##      true
## preds     No    Yes
##   No  125908  11864
##   Yes      0      0
```

Table 5. Confusion Matrix for the unpruned tree model.

```
##      true
## preds     No    Yes
##   No  114293   7223
##   Yes  11615   4641
```

Table 6. Confusion Matrix for the Naive Bayes model.

```
##         model missclassification false_negative
## 1 Lasso Logit         0.08447290     0.07915322
## 2        Tree         0.08611329     0.08611329
## 3 Naive Bayes         0.13673315     0.05944073
```

Table 7. Summary of misclassification rate and false negative rate across three models.

Discussion

Why are all misclassifications of the tree model false negatives?

There exists a bias toward the majority class because the majority vote performed at each node in a tree ignores numeric differences. But Naive Bayes and logistics regression are more sensitive to the numeric differences by accounting for probabilities of each category. As a result, tree-based methods fail to perform well for highly imbalanced data.

Which model is perferred in practice?

When predicting heart disease, we consider the false negative rate more critical than the overall misclassification rate, since it is worse to predict someone who does not have heart disease but in fact has one.

Based on *Table 3* and *Table 5*, all misclassifications in tree models are false negatives, as are most misclassifications in lasso logistic regression. Therefore, we recommend using the Naive Bayes model in practice, given its lowest test false negative rate (*Table 6*).

How different seed might affect error rates?

We are also aware that changing the seeds for the train-test split may affect test misclassification rates. By trying different seeds, we may be able to see what effect they have.

Future work - an alternative comparing method

When comparing different models, cross-validation is an alternative to the validation set approach. We obtain and compare the candidate with the lowest misclassification rate across each model type, with each candidate derived from a different fold or test data set.

We adjusted the random sampling of the train-test split of the imbalanced data but not that for cross-validation when obtaining the best lambda. Thus, we may implement a lasso logistic regression with balanced folds to obtain a better lambda.

Reference

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease