HW5 - Model Comparison

Kayla Kippes

Task 1: Conceptual Questions

- What is the purpose of using cross-validation when fitting a random forest model?

Cross-validation is used when fitting a random forest model in order to split the data into multiple folds and only train the data on some of those folds while testing it on others. This helps pick the best model by seeing how the model will do on unseen data and it will help prevent overfitting.

- Describe the bagged tree algorithm.

Create many bootstrap samples of the data and then fit trees to each of those samples of the training data. For each tree, find a prediction value and take all predictions and either average (for regression) or chose the most common (for classification) to create one final prediction.

- What is meant by a general linear model?

General linear model is a group of models that allows for the response variable to taken on different distributions so the predictors could be continuous or categorical. One example would be modeling a response variable that is binary.

- When fitting a multiple linear regression model, what does adding an interaction term do? That is, what does it allow the model to do differently as compared to when it is not included in the model?

Adding an interaction term allows for the value of one predictor be dependent on the value of another predictor. Without it, the model assumes that each predictor affects the response independently.

- Why do we split our data into a training and test set?

We split our data into a training and test set so that the model can be fit on data it has never seen before. It can learn from the training data and from there we can evaluate it's performance on the test data.

Task 2: Data Prep

Packages and Data

```r
## load the necessary packages
library(tidyverse)
library(tidymodels)
library(caret)
library(yardstick)

## load the data and save it as a tibble
heart <- read_csv('heart.csv') |>
  as.tibble()
```

Question 1

```r
## summary of the heart data set
summary(heart)
```

```
      Age            Sex            ChestPainType        RestingBP
 Min.   :28.00   Length:918         Length:918         Min.   :  0.0
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   Mode  :character   Median :130.0
 Mean   :53.51                                         Mean   :132.4
 3rd Qu.:60.00                                         3rd Qu.:140.0
 Max.   :77.00                                         Max.   :200.0
  Cholesterol      FastingBS       RestingECG            MaxHR
 Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
 1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
 Median :223.0   Median :0.0000   Mode  :character   Median :138.0
 Mean   :198.8   Mean   :0.2331                      Mean   :136.8
 3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
 Max.   :603.0   Max.   :1.0000                      Max.   :202.0
 ExerciseAngina        Oldpeak          ST_Slope            HeartDisease
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
                    Mean   : 0.8874                      Mean   :0.5534
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000
                    Max.   : 6.2000                      Max.   :1.0000
```

a. Heart Disease is currently a quantitative variable.

b. This does not make sense because it is a binary classification (0 or 1) so it should be a factor.

Question 2

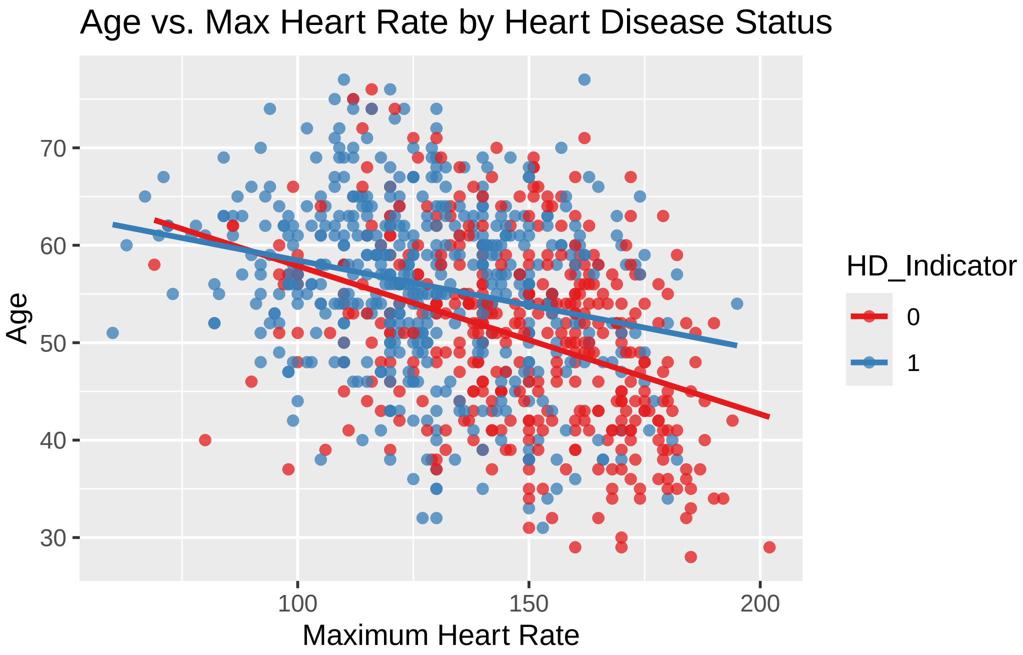
```r
## fix heart disease variable type and select all but two columns
new_heart <- heart |>
  mutate(HD_Indicator = as.factor(HeartDisease)) |>
  select(-ST_Slope, -HeartDisease)
```

Task 3: EDA

Question 1

```r
## load ggplot
library(ggplot2)

## plot visual
ggplot(new_heart, aes(x = MaxHR, y = Age, color = HD_Indicator)) +
  geom_point(alpha = 0.75) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_brewer(palette = "Set1") +
  labs(
  title = "Age vs. Max Heart Rate by Heart Disease Status",
  x = "Maximum Heart Rate",
  y = "Age")
```



Question 2

Based on the plot above, I think an interaction model is more appropriate here due to the differing slopes for each factor of the heart disease indicator.

Task 4: Testing and Training

```
## set seed
set.seed(101)

## split the data into train and test sets
heart_split <- initial_split(new_heart, prop = 0.8)
train <- training(heart_split)
test <- testing(heart_split)
```