STA 141A – Final Project

Predicting Heart Disease

Department of Statistics
University of California, Davis
Dec. 10, 2021

Eric Chagon | echagnon@ucdavis.edu
Meng-Tien Tsai | mtts@ucdavis.edu
Wun-Syuan Wu | wswu@ucdavis.edu
Sanskruti More | smore@ucdavis.edu

Introduction

According to the Center for Disease Control, heart disease is the leading cause of death for humans beating out even all forms of cancer combined ("Leading Causes of Death"). Early detection of heart disease would allow for faster treatments before the disease reaches a critical stage, and as a result improve the lives of patients worldwide. The aim of this paper is to attempt to diagnose a patient with heart disease based on a collection of biographical data on the patient such as Cholesterol Levels, Resting Heart Rate, Age, Sex, etc. The relationship between the risk factors of heart disease and the actual presence of heart disease is complex, and the aim of this paper is to create a model that can learn this relationship and determine if an individual is affected by heart disease (Cohn et al). Some key questions that this paper intends to answer are:

- If a model can be fit to accurately predict the presence or absence of heart disease, then what predictors are the most influential in this determination?
- Are all the predictors necessary to predict the presence of heart disease?

As mentioned above, our goal is to identify the features that increase the risk of heart failure. The University of California, Irvine maintains a repository of datasets, several of which revolve around heart disease. The dataset being analyzed was published on kaggle.com, and is the union of 5 datasets containing similar data. The dataset contains 918 observations with 12 attributes. The response variable, HeartDisease*,* is a binary classifier for heart disease where a 1 indicates that the individual had some form of heart disease, and a 0 indicates an individual with no heart disease. Information on the predictor variables can be seen below in Table 1.

Table 1. Description of Predictors

Variable	Description	Type
Age	Age of the patient (years)	Numeric
Sex	Sex of patient (M: Male, F: Female)	Categorical
ChestPainType	chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)	Categorical
RestingBP	Resting blood pressure (mm Hg)	Numeric
Cholesterol	Serum cholesterol (mm/dl)	Numeric
FastingBS	Fasting blood sugar (1: if FastingBS > 120 mg/dl, 0: otherwise)	Categorical
RestingECG	Resting electrocardiogram results (Normal: Normal, ST: having ST-T wave abnormality.	Categorical

	LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria)	
MaxHR	Maximum heart rate achieved	Numeric
ExerciseAngina	Does the patient have exercise induced angina, yes or no	Categorical
Oldpeak	ST depression induced by exercise relative to rest	Numeric
ST_slope	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]	Categorical

Methods and Results

Exploratory Data Analysis (EDA):

The goal of this paper is to examine which attribute has the potential to be used to predict future heart disease. Therefore, the exploratory data analysis objectives here are first to investigate features of attributes that are able to predict the binary response variable HeartDisease, and second, to look into outliers and anomalies that may have an influence on further modeling and analyzing.

A high level view of the data shows us that the data is observed over ages 30 to 80, made up of 193 females (21%) and 725 males (79%). Nearly, 55% of the observations report heart disease and 45% report no heart disease.

Since, there are more patients (79%) with heart disease above the age of 55, it might be interesting to study the effects of Age with Cholesterol, MaxHeart Rate and other predictors.

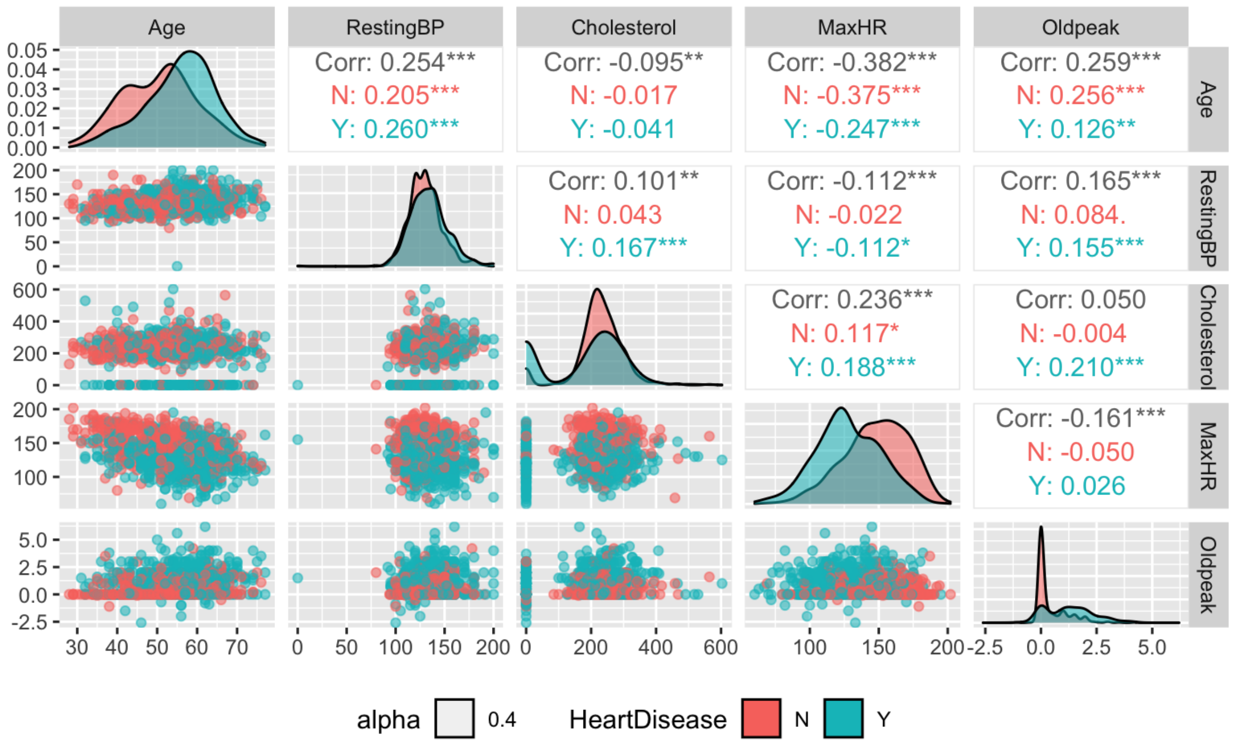


Figure 1: Multicollinearity Matrix

From Figure 1 (see above) it is clear that overall there are no strong linear relationships between age and other numeric predictors, so we can assume no multicollinearity within any of these predictors and age. Additionally, observing the correlation matrix we see that there are no other strong linear relationships among all numeric predictors.

Observing the scatterplots (Fig 1) with respect to Age yields the following observations:

Age vs RestingBP - No strong correlation or pattern is observed. One potential outlier/mistaken data point with restingBP 0 at Age 55.

Age vs Cholesterol shows no linear relationship. Between the range 400-600 mg/dl there are more patients with heart disease specifically below the age of 55 as compared to without heart disease. One possibility is that younger individuals with high cholesterol are at high risk of heart disease as compared to other younger individuals with controlled cholesterol levels.. There are some data points at 0 mg/dl cholesterol. Also most of these 0 mg/dl observations have heart disease. Cholesterol levels cannot be 0 mg/dl. Therefore, these data points are either missing values filled with 0s or some other error, we will further examine this in the outlier analysis section.

Age vs MaxHR has a weak negative correlation as observed by the downward trend. As age increases the MaxHeart Rate decreases. Moreover, overall the MaxHR for those with heart disease is lower than those without heart disease.

Age vs Oldpeak - No clear linear relation

Now, we will take a look at the categorical predictors proportions by heart disease to further understand the data.

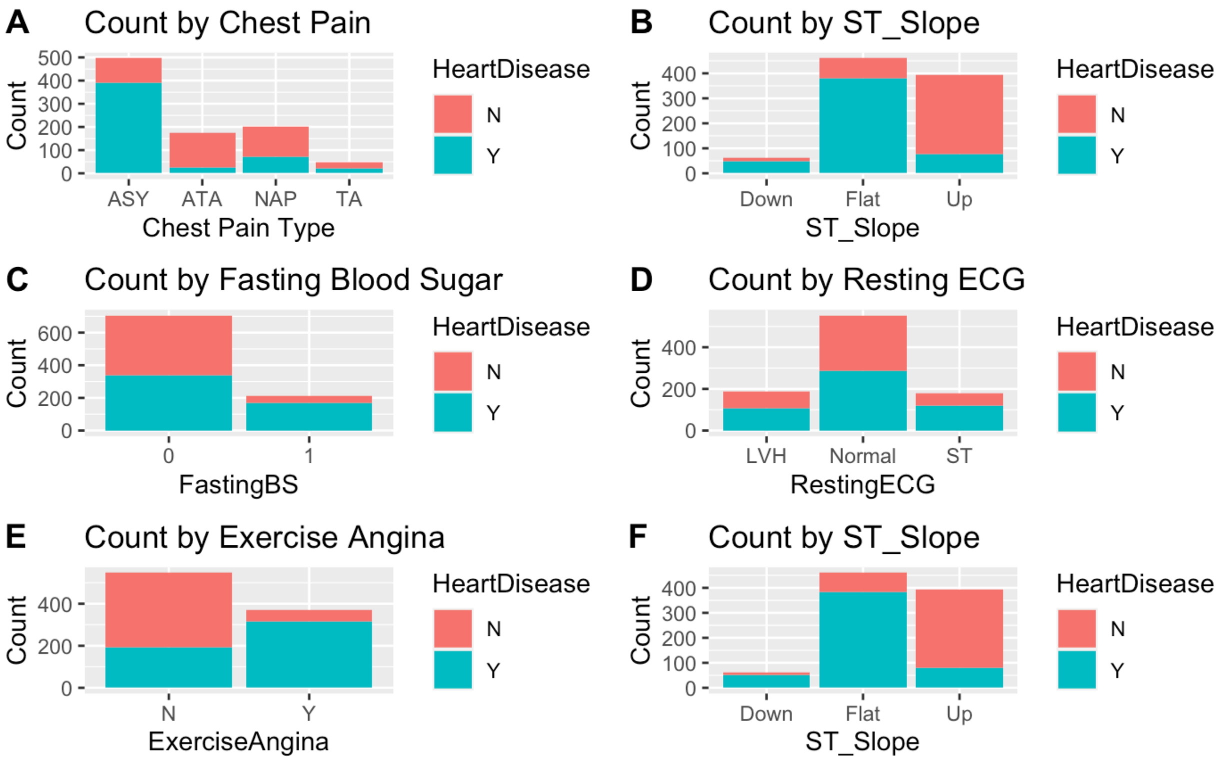


Figure 3: Stacked Scatter Plot of all Categorical Predictors

A. Asymptomatic chest pain is the most frequent chest pain type, and the majority of the Heart Disease patients experience this Asymptomatic chest pain.
B. The proportion of females with heart disease is smaller than that of Males. While this could lead to a pattern later it is important to note that the sample size of females is a lot smaller than Males and collecting more data points for females could lead to a different result.
C. Patients with Fasting Blood Sugar over 120 mg/dl (1) are at a significantly higher risk of heart disease as observed nearly all the patients in this category have been identified with a heart disease. This could be interpreted as a strong association between heart disease and blood sugar levels.
D. Overall, most people who had heart disease had a normal ECG. Moreover, within each type there is an almost equal proportion of Heart Disease and normal.
E. Individuals who reported Exercise Induced Angina (pain in chest, left shoulder region) are more likely to have a heart condition as observed here majority of individuals who have a heart condition reported Exercise Induced Angina.

F. The Down ST_Slope almost always indicates the presence of heart disease. While a flat and upward slope are more common in the ECG graph. A large proportion of the flat slope patients also have heart disease.

Outlier/Extreme Value Analysis:

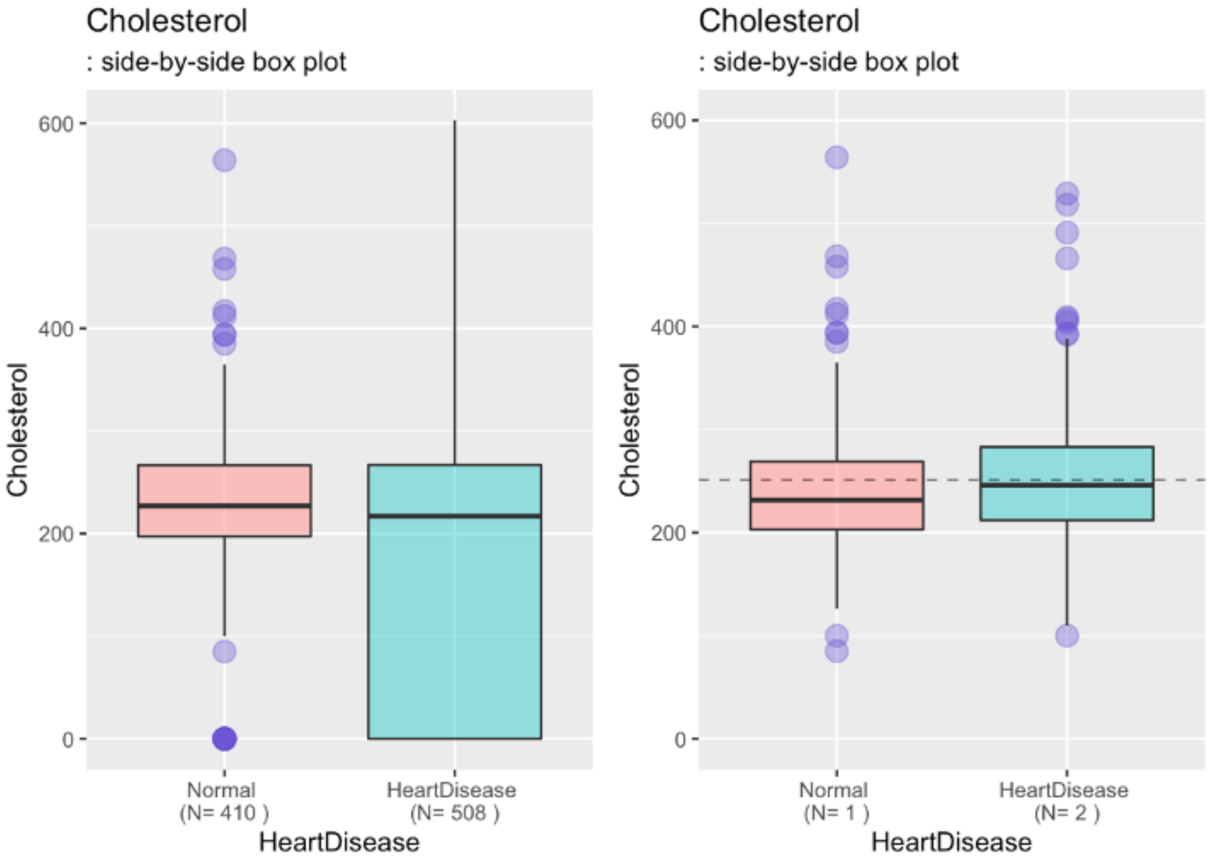


Figure 4: Side-by-Side boxplot of Cholesterol levels with 0 mg/dl (L) and without 0 mg/dl (R)

We note the effect of removing the 150, 0 mg/dl values from the dataset. As we know, most of these values come from the subset of people with heart disease. The plot on the right shows the IQR for patients with heart disease changed significantly, exposing other potential outliers in this group that were previously masked by the 0 values. The median values have moved very slightly. Given the size of the subset with 0 mg/dl cholesterol, we believe that other attributes of these data points will be significant to our analysis and therefore choose to not remove them. However, we will exercise caution when we make inferences about the same.

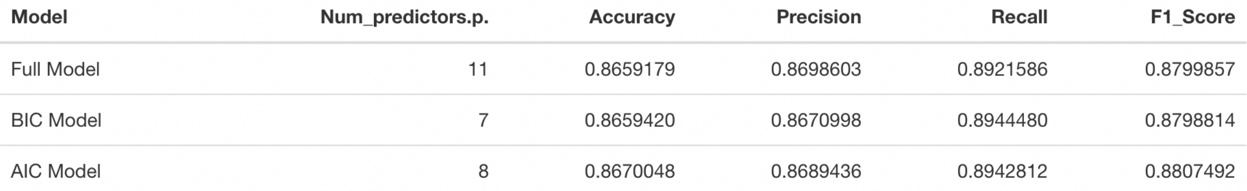
Model Building

Initially a model with all predictors present was created in order to have the full model needed for variable selection. Since the number of total possible predictors is relatively small (P = 11) the Best Possible Subset selection is a valid first choice for variable selection methods. In order to keep the model as simple as possible, the BIC criteria was used to judge the models' output from the best possible subset. The results can be seen here *Table/Figure*. From the output, the model with the lowest BIC is at model size equal to 7, with the predictors ST_Slope, ChestPainType, Sex, FastingBS, ExerciseAngina, Cholesterol, and Oldpeak. Next, the Forward Stepwise Procedure was used in order to be exhaustive. When BIC was the criteria to be minimized, the model output from the Forward Stepwise procedure was the same model obtained from the Best Possible Subset. Finally, the Forward Stepwise Procedure was used with AIC as the criteria to be minimized. In this case the model output was slightly different from the previous model. It contained all the same variables, with the addition of the Age variable. These two models were then compared against the full initial model in order to determine which model was the best at predicting the presence of heart disease.

Model Validation

To validate the models a function simulating k-fold cross validation was created (see attached). The three models were validated using this method with $k = 10$. Their average accuracy, precision, recall, and F1 scores were recorded and can be seen in the following table.

Table 2. Validation Criteria by model

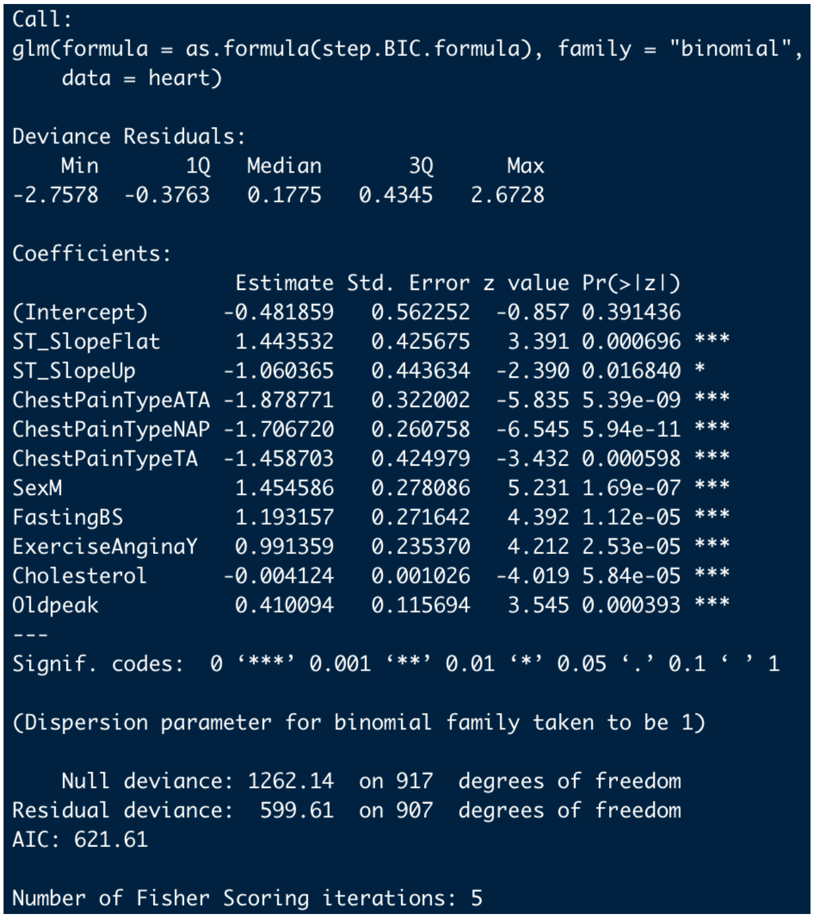


Model	Num_predictors.p.	Accuracy	Precision	Recall	F1_Score
Full Model	11	0.8659179	0.8698603	0.8921586	0.8799857
BIC Model	7	0.8659420	0.8670998	0.8944480	0.8798814
AIC Model	8	0.8670048	0.8689436	0.8942812	0.8807492

The models have very similar performance metrics, with every metric being with 1% of each other. It is important to note that in the context of this problem recall is the most important criteria. When dealing with healthcare, and especially a diagnosis problem, failing to identify a case of someone who actually has heart disease would have catastrophic effects for the patient. Looking back at Table 2 it is evident that the BIC model has the highest recall. On top of this the BIC model is the least complex model due to it having the least number of variables. As a result the BIC model was selected to be the final model.

Conclusions

After the final model was selected, it was refitted on the entire dataset. The final model can be seen in Figure 5 below.



```
Call:
glm(formula = as.formula(step.BIC.formula), family = "binomial",
    data = heart)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7578  -0.3763   0.1775   0.4345   2.6728  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -0.481859   0.562252  -0.857 0.391436    
ST_SlopeFlat      1.443532   0.425675   3.391 0.000696 ***
ST_SlopeUp       -1.060365   0.443634  -2.390 0.016840 *  
ChestPainTypeATA -1.878771   0.322002  -5.835 5.39e-09 ***
ChestPainTypeNAP -1.706720   0.260758  -6.545 5.94e-11 ***
ChestPainTypeTA  -1.458703   0.424979  -3.432 0.000598 ***
SexM              1.454586   0.278086   5.231 1.69e-07 ***
FastingBS         1.193157   0.271642   4.392 1.12e-05 ***
ExerciseAnginaY   0.991359   0.235370   4.212 2.53e-05 ***
Cholesterol      -0.004124   0.001026  -4.019 5.84e-05 ***
Oldpeak           0.410094   0.115694   3.545 0.000393 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  599.61  on 907  degrees of freedom
AIC: 621.61

Number of Fisher Scoring iterations: 5
```

Figure 5. Final model summary

This model has some interesting features that aren't exactly intuitive. Any level of chest pain that isn’t type ASY (asymptomatic) actually decreases the odds of heart disease being present according to this model. Furthermore, a 1 unit increase in Cholesterol raises the odds of heart disease being present, albeit by a very minute amount. High cholesterol is usually a key indicator of heart disease. However, not every coefficient in this model goes against traditional thinking, for example being a male increases the odds of heart disease by 1.45 compared to females, which is supported by males being more likely to have heart disease than females (Weidner). Similarly, the flat ST slope is usually indicative of heart disease, while an upward slope is not indicative of heart disease (Burns; Misumida).

From this output the predictors with the largest effect on the presence of heart disease are ChestPainType, ST_Slop, and Sex, as seen from the values of their estimated coefficients. While the full model performed well, this models success shows that not all the variables are required to model HeartDisease. Figure 6 below shows the resulting predictions on the full dataset.

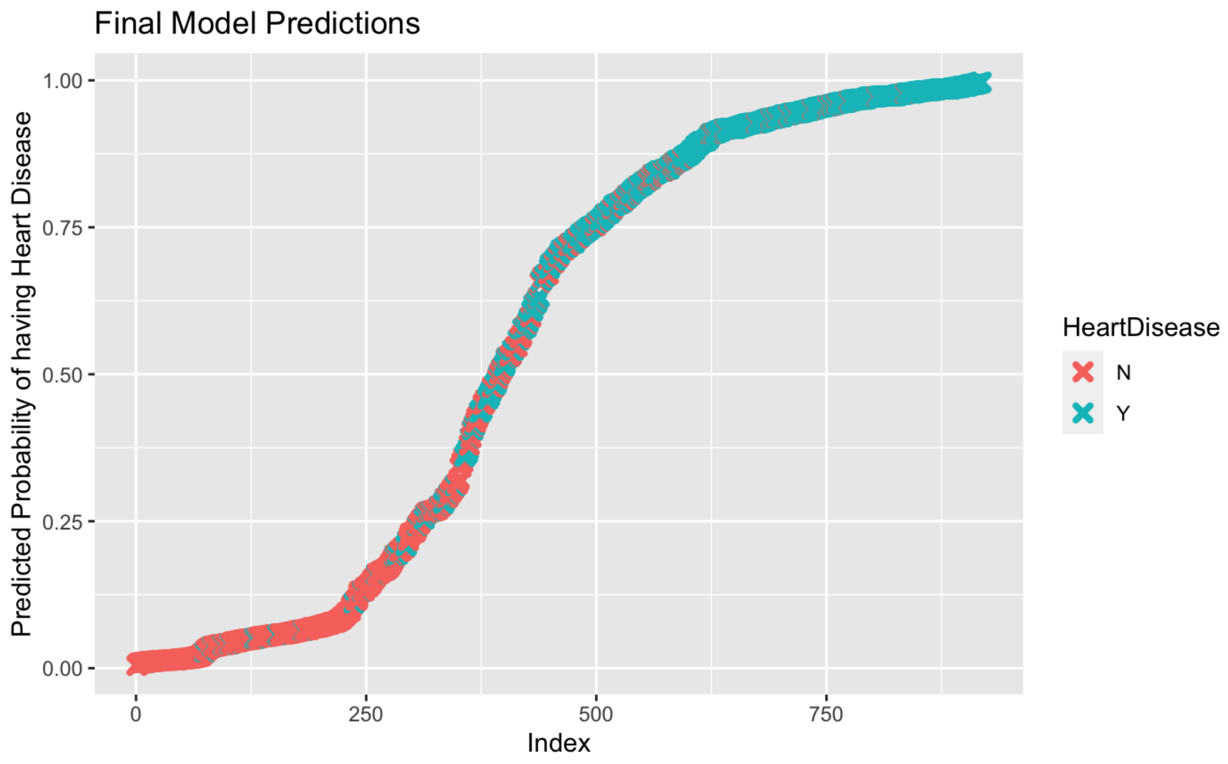


Figure 6. Final model predictions on whole dataset

While the model does an overall good job of predicting the presence of heart disease there is still room for improvement. This is best showcased in the tails of the plot, where the model was predicting incorrectly with a high probability. In the way the model was trained there is no difference in being wrong with a prediction of 51% or 99%. When in reality the model predicting incorrectly at greater than 90% is much worse. In future work either better metrics, or training methods should be used to weigh far misses to be worse than near misses.

In the context of the problem, diagnosing a patient with a deadly disease, these results likely wouldn't be sufficient enough to warrant any practical use. The model has a recall of 0.8944, therefore for every 100 people who have heart disease, this model would only be able to successfully diagnose 11 of them. These 11 missed people's conditions would worsen as a result. Future work should make attempts to increase the model's recall, and accuracy as a whole.

After reviewing the feedback from the proposal, the idea of using unsupervised clustering and knn was discarded and instead this project focused solely on logistic regression.

Works Cited

Burns, Ed. “The ST Segment • LITFL • ECG Library Basics.” *LITFL.com*, 24 March 2021, https://litfl.com/st-segment-ecg-library/.

“Leading Causes of Death.” *CDC*, Center for Disease Control, 2021, https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

Misumida, Naoki. “Prevalence and Clinical Significance of Up-Sloping ST-Segment Depression in Patients With Non-ST-Segment Elevation Myocardial Infarction.” *NCBI*, 25 October 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5295568/.

Weidner, G. “Why do men get more heart disease than women? An international perspective.” *PubMed*, 2000, https://pubmed.ncbi.nlm.nih.gov/10863872/.

Appendix

Model Building/Analysis

```
require(ggplot2)
require(class)
require(caret)
require(MASS)

heart=read.csv("heart.csv")

set.seed(999)
idx <- sample(1:nrow(heart), nrow(heart)*.9, replace = FALSE)
heart_train <- heart[idx, ]
heart_test <- heart[-idx, ]


########## Model Selection ##########
full.model = glm_fit <- glm(HeartDisease~., data = heart_train,
family="binomial")
null.model = glm_fit <- glm(HeartDisease~1, data = heart_train,
family="binomial")

n = nrow(heart_train)
## Bidirectional Stepwise
# BIC as criteria
step.fit = stepAIC(null.model, scope = list(upper = full.model, lower
= ~1), trace=F, direction="both", k=log(n))

step.fit.summary <- step.fit$anova
step.fit.summary

# AIC as criteria
step.fit2 = stepAIC(null.model, scope = list(upper = full.model,
lower = ~1), trace=F, direction="both", k=2)

step.fit.summary2 <- step.fit2$anova
step.fit.summary2
```

```
## Best Possible Subset
require(bestglm)

qwer = heart_train
#can only take in numeric and factors

qwer$Age = as.numeric(qwer$Age)
qwer$Sex = as.factor(qwer$Sex)
qwer$ChestPainType = as.factor(qwer$ChestPainType)
qwer$RestingBP = as.numeric(qwer$RestingBP)
qwer$Cholesterol = as.numeric(qwer$Cholesterol)
qwer$FastingBS = as.numeric(qwer$FastingBS)
qwer$RestingECG = as.factor(qwer$RestingECG)
qwer$MaxHR = as.numeric(qwer$MaxHR)
qwer$ExerciseAngina = as.factor(qwer$ExerciseAngina)
qwer$Oldpeak = as.numeric(qwer$Oldpeak)
qwer$ST_Slope = as.factor(qwer$ST_Slope)
qwer$HeartDisease = as.numeric(qwer$HeartDisease)


test = bestglm(qwer, family = binomial, IC = "BIC")
test$BestModel
# This is the same model as stepBIC


########## Model Validation ##########

k_fold <- function(k, glm_formula, data, seed){
  set.seed(seed)
  size_of_test = ceiling(nrow(data) / k) #size of each testing
dataset
  test_indexes = matrix(nrow=k-1, ncol=size_of_test) #the indexes
that will make up the test dataset for each fold
  total_rows = c(1:nrow(heart))
  for(i in c(1:k)){
    if(length(total_rows) < size_of_test){
      leftovers = total_rows #when the dataset isnt divisible by
size_of_test there will be leftovers
```

```
    }
    else{
      idx = sample(total_rows, size_of_test, replace = FALSE)
      total_rows = setdiff(total_rows, idx)
      test_indexes[i,] = idx
    }
  }
  accuracy <- c()
  precision <- c()
  recall <- c()
  f_measure <- c()
  for(i in c(1:k)){
    if (i == k){ #after all the full folds are done we have to run
the leftovers as test data so every observation is able to be part of
the test data
      test_data = data[leftovers, ]
      train_data = data[-leftovers,]
    }
    else{
      test_data = data[test_indexes[i,], ]
      train_data = data[-test_indexes[i,],]
    }

    glm_fit <- glm(as.formula(glm_formula), data = train_data,
family="binomial") #fit model with formula that was passed in
    response_v <-
as.character(attributes(glm_fit$terms)$variables[[2]])

    predicted <- predict(glm_fit, test_data, type="response")
#predictions of the model on the test data this fold
    confusion <- table(ifelse(predicted > 0.5, 1, 0),
                       test_data[[response_v]],
                       dnn = c("Predicted","True"))
    #performance metrics for the current fold
    accuracy[i] <- sum(diag(confusion)) / sum(confusion)
    precision[i] <- confusion[2,2] / sum(confusion[2,])
    recall[i] <- confusion[2,2] / sum(confusion[,2])
    f_measure[i] <- (2 *(precision[i] * recall[i])) / (precision[i] +
recall[i])
```

```
  }
  result <- setNames(c(mean(accuracy), mean(precision), mean(recall),
mean(f_measure)), c("Accuracy", "Precision", "Recall", "F-measure"))
  return(result)

}

step.BIC.formula = HeartDisease ~ ST_Slope + ChestPainType + Sex +
FastingBS + ExerciseAngina +
  Cholesterol + Oldpeak

step.AIC.formula = HeartDisease ~ ST_Slope + ChestPainType + Sex +
FastingBS + ExerciseAngina +
  Cholesterol + Oldpeak + Age

#best.subset.formula = HeartDisease ~ ST_Slope + ChestPainType + Sex
+ FastingBS + ExerciseAngina +  Cholesterol + Oldpeak

# k-fold cross validation on the models
folds = 10

step.bic.kfold = k_fold(folds, step.BIC.formula, heart, 12082021)
step.aic.kfold = k_fold(folds, step.AIC.formula, heart, 12082021)
full.model.kfold = k_fold(folds, HeartDisease~., heart, 12082021)

valid.df = data.frame(Model = c("Full Model", "BIC Model", "AIC
Model"),
                      "Num_predictors(p)" = c(11, 7, 8),
                      Accuracy = c(full.model.kfold[1],
step.bic.kfold[1], step.aic.kfold[1]),
                      Precision = c(full.model.kfold[2],
step.bic.kfold[2], step.aic.kfold[2]),
                      Recall = c(full.model.kfold[3],
step.bic.kfold[3], step.aic.kfold[3]), "F1_Score" =
c(full.model.kfold[4], step.bic.kfold[4], step.aic.kfold[4]))

require(kableExtra)

valid.df %>%
  kbl() %>%
  kable_styling()

final.model = glm(as.formula(step.BIC.formula), data = heart, family
= "binomial")
summary(final.model)

predicted.data <- data.frame(prob.of.HeartDisease =
final.model$fitted.values, HeartDisease = heart$HeartDisease)

predicted.data <-
predicted.data[order(predicted.data$prob.of.HeartDisease, decreasing
= FALSE), ]

predicted.data$index <- c(1:nrow(heart))

final.pred = ggplot(data=predicted.data, aes(x=index,
y=prob.of.HeartDisease)) + geom_point(aes(color=HeartDisease), alpha
= 1, shape = 4, stroke = 2) + xlab("Index") + ylab("Predicted
Probability of having Heart Disease") + labs(title = "Final Model
Predictions")

ggsave("Final Model Predictions.pdf", final.pred)


########## EDA and Data Visualization ##########

require(ggplot2)
library(GGally)
library(ggpubr)
library(tidyverse)

#prepare the data
df <- read.csv('heart.csv')
df[sapply(df, is.character )] <- lapply(df[sapply(df, is.character)],
as.factor)
df$FastingBS <- as.factor(df$FastingBS)
df$HeartDisease <- ifelse(df$HeartDisease == 1, "Y", "N")
df$HeartDisease <- as.factor(df$HeartDisease)
```

```
#data overview
summary(df)
prop.table(table(df[(df$Age >= 55),]$HeartDisease))

#scatter plot matrix (Fig 1)
num_df <- df[, c(1,4,5,8,10)]
ggpairs(num_df, legend = 1, aes(color = df$HeartDisease, alpha =
0.4)) +
  theme(legend.position = "bottom") +
  labs(fill = "HeartDisease")

#stacked scatter plots
bp_ct <- ggplot(df, aes(ChestPainType, fill = HeartDisease)) +
geom_bar() +  labs(title = "Count by Chest Pain ", x = "Chest Pain
Type", y = "Count")
bp_s <- ggplot(df, aes(Sex, fill = HeartDisease)) + geom_bar() +
labs(title = "Count by Sex ", x = "Sex", y = "Count")
bp_f <- ggplot(df, aes(FastingBS, fill = HeartDisease)) + geom_bar()
+  labs(title = "Count by Fasting Blood Sugar", x = "FastingBS", y =
"Count")
bp_r <- ggplot(df, aes(RestingECG, fill = HeartDisease)) + geom_bar()
+  labs(title = "Count by Resting ECG", x = "RestingECG", y =
"Count")
bp_e <- ggplot(df, aes(ExerciseAngina, fill = HeartDisease)) +
geom_bar() +  labs(title = "Count by Exercise Angina", x =
"ExerciseAngina", y = "Count")
bp_s <- ggplot(df, aes(ST_Slope, fill = HeartDisease)) + geom_bar() +
labs(title = "Count by ST_Slope", x = "ST_Slope", y = "Count")
ggarrange(bp_ct, bp_s, bp_f, bp_r, bp_e, bp_s,
             ncol = 2, nrow = 3, labels = c('A','B', 'C', 'D', 'E',
'F'))

#side-by-side boxplot for cholesterol
p3 <- ggplot(data = heart, mapping = aes(x = HeartDisease, y =
heart[,quant_name_index[3]], fill = HeartDisease)) +
    geom_boxplot (alpha = 0.4, outlier.color = "slateblue",
outlier.size = 4 ) +
    theme (legend.position = "none" ) +
    scale_x_discrete (labels = paste(c("Normal", "HeartDisease"),
"\n(N=", table(heart[,quant_name_index[3]]), ")", sep=" ") ) +
    ggtitle(paste(quant_heart_name[3]), ": side-by-side box plot") +
    xlab("HeartDisease") + ylab(quant_heart_name[3])

p6 <- ggplot(data = heart_rm, mapping = aes(x = HeartDisease, y =
heart_rm[,quant_name_index[3]], fill = HeartDisease)) +
    geom_boxplot (alpha = 0.4,
                  outlier.color = "slateblue",
                  outlier.size = 4 ) +
    scale_x_discrete (labels = paste(c("Normal", "HeartDisease"),
"\n(N=", table(heart_rm[,quant_name_index[3]]), ")", sep=" ")) +
    ylim(0,600) +
    geom_hline(yintercept =
mean(heart_rm$Cholesterol[which(heart_rm$HeartDisease == 1)]),
               linetype = "dashed", alpha = 0.5) +
    ggtitle(paste(quant_heart_name[3]), ": side-by-side box plot") +
    xlab("HeartDisease") + ylab(quant_heart_name[3]) +
    theme (legend.position = "none" )
grid.arrange(p3, p6, nrow = 1, ncol = 2)

##Frequency plot
heart$HeartDisease <- as.factor(heart$HeartDisease)
for(i in quant_name_index){
    label <- paste(colnames(heart)[i], ":", "frequency plot")
    p <- ggplot(data = heart, mapping = aes(x = heart[,i], color =
HeartDisease))+
    geom_freqpoly()
    print(p + ggtitle(label = label) + xlab(colnames(heart)[i]) +
ylab("Count"))
}
```