KOGOD SCHOOL *of* BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics Project

"Predicting the Likelihood of Heart Attack "

Team 1: ITEC 621 -001
Victor Stephen

May 4, 2023

Final Report

1. Business Question and Case

1.1. Business Question

What are the key risk factors contributing to heart disease and what targeted intervention programs can reduce healthcare costs?

1.2. Business Case

According to the Centers for Disease Control and Prevention (CDC), heart disease is a global health concern, accounting for approximately 31% of all deaths worldwide. In the United States alone, over 600,000 people die from heart disease each year, making it the leading cause of death for both men and women[1]. Despite significant advancements in medical technology, many heart attacks are still unpredictable, leading to delayed treatment and increased mortality rates. By identifying the key risk factors associated with heart disease, healthcare providers and insurance companies can design targeted intervention programs to mitigate risks and reduce costs. In fact, Early detection and treatment of heart disease have been demonstrated to decrease the need for expensive interventions like bypass surgery, resulting in significant cost savings of up to $1.5 billion annually, according to a report by the American College of Cardiology in 2017[2].

The value proposition of this project is to provide actionable insights to healthcare providers and insurance companies, enabling them to make informed decisions and optimize their resources.

2. Analytics Question

What is the effect of patient characteristics, such as age, chest pain type (cp), cholesterol levels (chol), maximum heart rate achieved (thalach), and the number of major vessels colored by fluoroscopy (ca) on the likelihood of heart disease? And to what extent do these factors influence the presence of heart disease?

Our analytical question aims to identify the most influential factors contributing to heart disease by examining the relationships between patient characteristics, listed in the analytical question. Key predictors include the aforementioned patient characteristics, which are a mix of numeric, categorical, and binary variables.

Our primary goal for the project is predictive accuracy. This enables informed decisions on prevention, risk management, and personalized treatments while ensuring a comprehensive understanding of factors associated with heart disease.

3. Data Set Description

The "Heart Disease UCI dataset" is from Kaggle.com and contains information on 303 patients and their characteristics associated with heart disease. The dataset includes 14 quantitative, binary, continuous, and categorical predictors, capturing essential patient characteristics such as age (in years), resting blood pressure, serum cholesterol (in mg/dL), maximum heart rate achieved,thalach (in bpm), and oldpeak (ST depression induced by exercise relative to rest). The categorical variables included sex, chest pain type (typical angina, atypical angina, non-anginal pain, asymptomatic), fasting blood sugar (greater than or equal to 120 mg/dL or less than 120 mg/dL), exercise-induced angina (yes or no), and resting electrocardiographic results show whether the patient had normal or abnormal results, number of major vessels colored by fluoroscopy, and thalassemia represents the blood disorder. The outcome variable of interest is a binary classification variable indicating whether the patient was diagnosed with heart disease. Data source: https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility)

4. Descriptive Analytics

4.1 Descriptive Statistics of Key Variables

Visual and Classification Analytics: Again, our outcome variable for this heart disease analysis is the presence or absence of heart disease, a binary classification. The distribution of the outcome variable is relatively balanced, with 55.5% of individuals having heart disease and 45.5% not having heart disease.

Initial descriptive analysis focused on important predictors highlighted above in the "Analytics Question" and "Dataset Description. The dataset contains 207 males (68.3%) and 96 females (31.7%). For chest pain type, 47.2% of the cases are typical angina, while 29.0% are non-anginal pain, 16.5% are atypical angina, and 7.3% are asymptomatic. Regarding resting ECG results, 50.5% are normal, 49.2% have ST-T wave abnormality, and 0.3% show probable or definite left ventricular hypertrophy.

Fig.2, there is a moderate positive correlation between age and resting blood pressure (0.279), indicating that older individuals tend to have higher resting blood pressure. Furthermore, there is a negative correlation between the maximum heart rate achieved and age (-0.399), suggesting that younger individuals typically have a higher maximum heart rate. Also, the outcome variable (presence of heart disease) is negatively correlated with maximum heart rate achieved (r = -0.4) This suggests that higher maximum heart rates tend to be associated with a lower likelihood of heart disease and positively correlated with age (r = 0.2), resting blood pressure (r = 0.15), and oldpeak (r = 0.43).

The ANOVA test results show significant differences in the presence of heart disease across sex (F value: 25.79, P-value: < 0.001), chest pain type (F value: 69.77, P-value: < 0.001), and exercise-induced angina (F value: 70.95, P-value: < 0.001). However, the fasting blood sugar level (P-value: 0.627) and resting electrocardiographic results (P-value: 0.0168) show less significance in differentiating heart disease presence.

Box Plot and Chi-square tests indicate that the presence of heart disease varies by chest pain type (highest in atypical angina), sex (higher in males), and resting electrocardiographic results (higher in those with ST-T wave abnormality). Other categorical variables, like fasting blood sugar (fbs) and exercise-induced angina (exang), do not show significant associations with the presence of heart disease.

Finally, an initial observation that older individuals seem to have a higher prevalence of heart disease led to a chi-square test between age groups (younger than 50, 50-60, 61-70, and older than 70) and heart disease presence. A significant chi-square test result confirms dependence between age group and heart disease presence.

4.2 Data Pre-Processing and Transformations:

The data sourced has been preprocessed and there were no omitted values.

4.3 Initial Set of Predictors:

The initial set of predictors for the heart disease classification model was chosen based on domain knowledge and factors generally considered to influence heart disease risk. These predictors include age(in years), as older individuals are generally more prone to cardiovascular issues; sex, as males and females may have different heart disease risks; chest pain type, which can indicate varying levels of heart disease risk; resting blood pressure (in mmHg), a known risk factor for heart diseases; serum cholesterol, which can contribute to the development of arterial plaque; fasting blood sugar, which may be associated with increased heart disease risk; resting

electrocardiographic results, which can indicate potential heart issues; maximum heart rate achieved, as a lower rate might suggest decreased heart adaptability to physical stress; and exercise-induced angina, which could be a sign of underlying heart issues due to insufficient oxygen-rich blood during physical activity.

5. Modeling Methods and Model Specifications

5.1. Initial Logistic Regression Modeling

A logistic regression model fit using the 13 predictors for heart disease produced significant estimates, indicating that the model is better at predicting heart disease than the null model. Significant predictors at the .05 significance level include sex, chest pain type (cp), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), number of major vessels colored by fluoroscopy (ca), and thalassemia type (thal).

Initial Logit Model Results

The logistic regression analysis found that the coefficients for sex, cp, thalach, exang, oldpeak, ca, and thal are statistically significant at the 5% level, indicating that these variables are significantly associated with the presence of heart disease. The accuracy of the model is reported as 0.836, indicating that the model can correctly predict the presence or absence of heart disease in approximately 85% of cases. Additionally, there is a notable reduction between the null and residual deviance (49.3%), suggesting that the model is a good fit for the data. Overall, these results suggest that the logistic regression model can identify significant predictors of heart disease and has a reasonable level of predictive accuracy.

5.2. Logistic Regression Assumptions (Second Set of Predictors)

The stepwise(backward) variable selection method was used to select only the significant predictors. The logistic regression was then fitted with a smaller set of significant predictors, leading to improved model performance. With the smaller set of predictors, the condition index has dropped to 47.33 (pg.16), indicating acceptable levels of multicollinearity in this specification, indicating the predictors are independent (XI). Additionally, the highest Variance Inflation Factor (VIF) is 1.49, which is well below the threshold of concern, suggesting that multicollinearity is not a problem for this specification (XI). The accuracy of the stepwise model is 0.852, indicating that it performs well in predicting the presence of heart disease.

5.3. Model Specification Candidates and Rationale

The first model specification used in this exercise was the initial set of 13 predictors selected based on domain knowledge and clinical understanding. The second model specification was selected using stepwise variable selection at p-value threshold of 0.05 to include only the most significant predictors. The full model for this variable selection exercise included all predictors. The lower bound was the null model plus “age,” a variable whose inclusion was based on intuition and the positive relationship between it and the outcome variable, regardless of its significance. Stepwise variable selection refined our second specification down to 11 predictors: age, sex, chest pain type (cp), resting blood pressure (trestbps), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and thalassemia type (thal). All predictors are significant at the .05 significance level except for age

and restecg. Due to the variable selection resulting in a subset of predictors, we will refer to this specification as the "small" subset of predictors or specifications.

5.4. Methods Evaluated:

First, we began our analysis by examining a logistic regression model for our classification, then examined the problem using the stepwise set of predictors. This initial model served as a baseline to compare the performance of more advanced methods that followed.
Next, we examined both random forests and boosted tree models for our classification problem using the same full and stepwise set of predictors. These ensemble methods were chosen to improve predictive accuracy, reduce overfitting, and extract a clearer understanding of which predictors are most "important" in predicting the presence of heart disease.

5.5. Cross-Validation Testing and Final Model Selection

In all six combinations of model and specification, we evaluated six different combinations of models and specifications. For each combination, we computed the confusion matrix and assessed the model performance using Accuracy, Sensitivity, ROC, and AUC.
For the logistic regression models, we used the glm function with type = "response" to predict the outcomes and calculate the confusion matrix. We then employed the ROCR package to analyze the AUC and ROC. Interestingly, both the full and reduced logistic models displayed the same sensitivity of 0.97, AUC of 92.7%, and an accuracy of 83.6%.
We also fitted a random forest model for the full dataset, which yielded the following results: an accuracy of 82%, AUC of 94.3%, and sensitivity of 76% for the reduced model. The random forest model with fewer predictors had an accuracy of 83.8%, AUC of 92.9%, and sensitivity of 80%.
Lastly, we evaluated the boosted tree models. The reduced model demonstrated an accuracy of 81.9%, the sensitivity of 97.2%, and AUC of 95%. The full model yielded similar results, with an accuracy of 80.3%, sensitivity of 97.2%, and AUC of 95%.

5.6 Final Method/Specification Selected:

Based on the parameters we used for the comparison of the three models, the small model of the boosted trees yielded the best results with an accuracy of .819, sensitivity of 0.972 and area under the curve of 0.95. From the plot, we also see that the graph hugs the top left corner and tends towards the value of 1.
It is worth noting that We adjusted the tuning parameters for the confusion matrix to 0.3 to encourage the model to classify more positives. Additionally, we set the number of trees to 500, considering the small size of our dataset. To address degrees of freedom issues arising from the limited data points, we allocated 80% of the data for the training set during the partitioning process.

6. Analysis of Results

The model demonstrates a strong ability to accurately predict patients with a likelihood of heart disease, achieving an accuracy of 82%, which surpasses the threshold for a good model. Moreover, the model exhibits high sensitivity, indicating its effectiveness in classifying patients at risk of heart attacks, which is our primary objective.
Interestingly, the random forest and boosted tree models identified chest pain (cp), the number of major vessels colored by fluoroscopy (ca), Thalium stress test results (thal), and ST depression induced by exercise relative to rest (old peak) as the most significant contributors to reducing the %MSE. These variables play a crucial role in enhancing the predictive performance of the model for heart disease risk.

Important Predictors:
The output (pg.32) provided allows us to make several interpretations about predicting heart disease. Chest Pain Type (cp) is the most important variable with a relative importance of 19.84%, indicating that it has the strongest influence on predicting heart disease among all variables considered. The Number of Major Vessels (ca) and Thalassemia (thal) are the second and third most important variables with relative importance of 15.83% and 14.25%, respectively. ST Depression Induced by Exercise (oldpeak) and Maximum Heart Rate Achieved (thalach) follow with relative importance values of 12.24% and 9.77%. Other variables, such as age, exang, trestbps, slope, sex, and restecg, have lower relative importance ranging from 8.17% to 1.88%. Although their influence is less substantial compared to the top five variables, it is crucial to consider all these factors in the model to gain a comprehensive understanding of the elements that influence heart disease prediction.

7. Conclusions and Lessons Learned

7.1. Conclusions from the Analysis

Based on our model, the most significant factors contributing to heart attacks include chest pain (cp), the number of major vessels colored by fluoroscopy (ca), Thalium stress test results (thal), ST depression induced by exercise relative to rest (old peak), and maximum heart rate achieved (thalach). Interestingly, age, which was not initially considered significant but was included based on domain knowledge, emerged as the 6th most influential contributor to the reduction of %MSE. Factors such as fasting blood sugar (FBS), sex, and cholesterol trailed far behind in importance.

The variable importance plots from both the boosted trees and random forest models support the conclusion that these predictors are the primary factors influencing the likelihood of heart attacks. The model's purpose is to assist medical institutions in addressing the increasing number of deaths resulting from heart attacks and to save billions in medical expenses. With access to more data, we anticipate even further improvement in the model's predictive power.

7.2 Project Issues, Challenges and Lessons Learned

A major challenge that the team faced was the initial worry of data quality. Although the dataset was sourced from Kaggle and after careful investigation, we discovered that the dataset originally had 76 attributes. Additionally, the initial sets of predictors were chosen from domain knowledge, this could lead to our results being insufficient since there is no data to support that decision.

Model complexity and overfitting was a concern considering the number of predictors in the initial model feature; however, we employed the stepwise technique which was essential in reducing overfitting.

Throughout this project, the team were made aware of the importance of choosing the right evaluation metrics which were critical for our model assessment and validation.

Appendix Contents

Contents

A. Data Information

Provides statistical information of each variable in the dataset

describe(Heart.Att)

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	1	303	54.36633663	9.08210099	55	54.54320988	10.3782	29	77	48	-0.20046319	-0.56912374	0.521753069
sex	2	303	0.683168317	0.466010823	1	0.728395062	0	0	1	1	-0.78351743	-1.39065743	0.026771622
cp	3	303	0.9669967	1.032052489	1	0.864197531	1.4826	0	3	3	0.479943598	-1.20511726	0.059289866
trestbps	4	303	131.6237624	17.53814281	130	130.436214	14.826	94	200	106	0.706716973	0.868396024	1.007539979
chol	5	303	246.2640264	51.83075099	240	243.4855967	47.4432	126	564	438	1.132104929	4.362840855	2.977598844
fbs	6	303	0.148514851	0.356197875	0	0.061728395	0	0	1	1	1.967025388	1.875410992	0.020463033
restecg	7	303	0.528052805	0.525859596	1	0.518518519	0	0	2	2	0.160916654	-1.37083448	0.030209844
thalach	8	303	149.6468647	22.90516111	153	150.9753086	22.239	71	202	131	-0.53210047	-0.09992646	1.315867125
exang	9	303	0.326732673	0.469794465	0	0.283950617	0	0	1	1	0.735195923	-1.46428695	0.026988987
oldpeak	10	303	1.03960396	1.161075022	0.8	0.855555556	1.18608	0	6.2	6.2	1.257176106	1.500339664	0.066702017
slope	11	303	1.399339934	0.616226145	1	1.46090535	1.4826	0	2	2	-0.50329386	-0.65252214	0.035401267
ca	12	303	0.729372937	1.022606365	0	0.53909465	0	0	4	4	1.297476206	0.780652237	0.058747201
thal	13	303	2.313531353	0.612276507	2	2.358024691	0	0	3	3	-0.47201256	0.2517144	0.035174366
HeartDisease	14	303	0.544554455	0.498834784	1	0.555555556	0	0	1	1	-0.17804456	-1.97478493	0.02865731

Fig 1: Dataset Statistics

B. Visuals, Graphs, and Plots

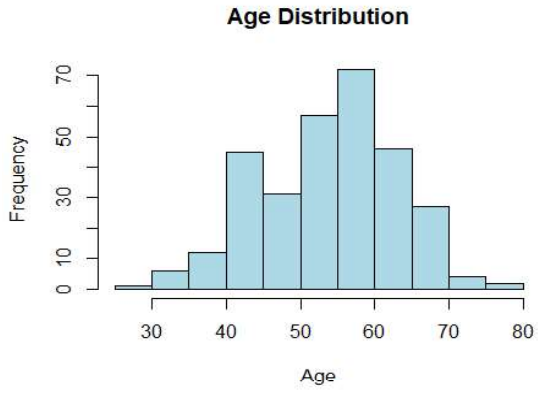


Fig 2: Age Distribution

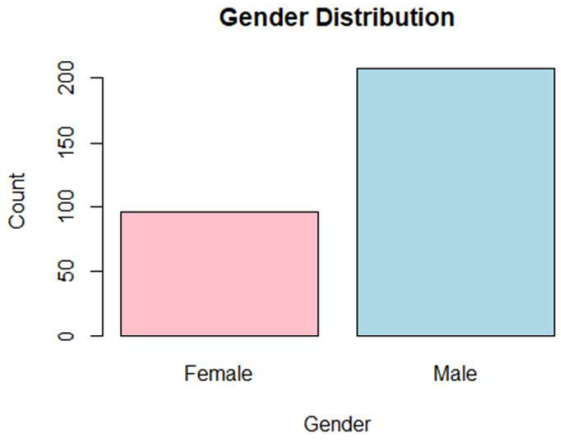


Fig 3: Gender Distribution

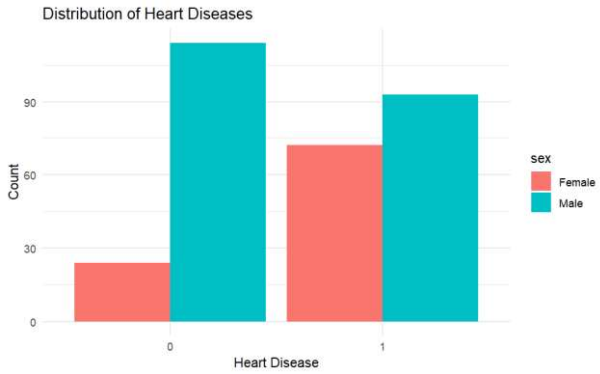


Fig 4. Heart Disease Distribution (Gender)

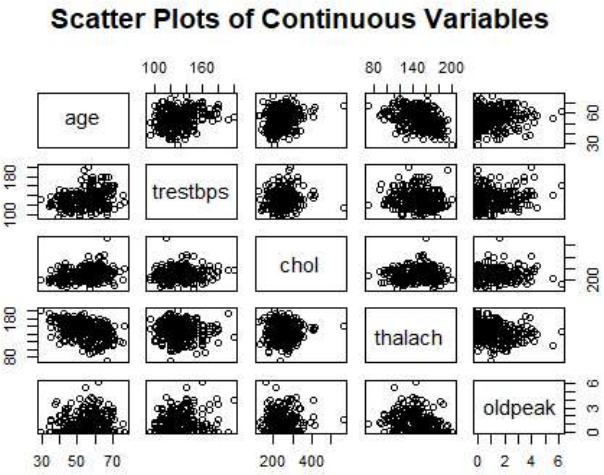


Fig 5: Scatter Plots of Continuous Variables

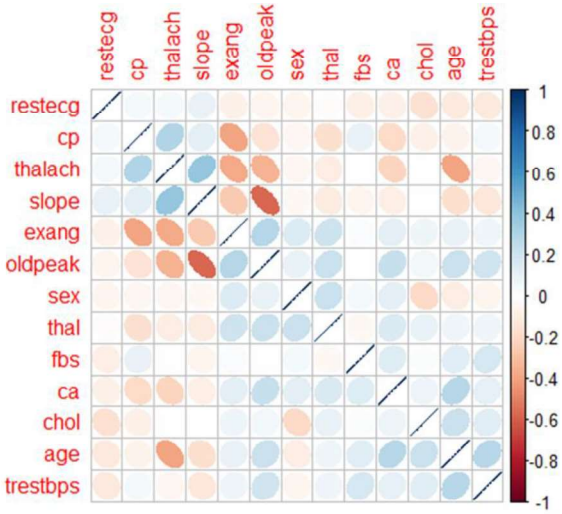


Fig 6: Predictors Correlation

Boxplots:

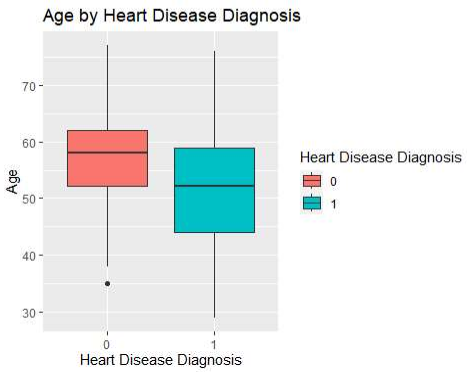


Fig 7: Age by Heart Disease Diagnosis

Fig 8: Distribution of Resting Blood Pressure

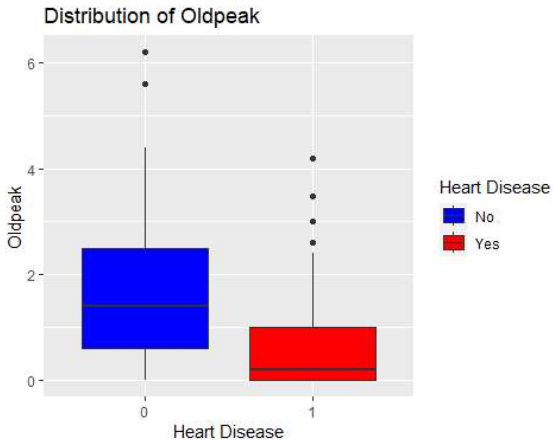


Fig 9: Distribution of Oldpeak

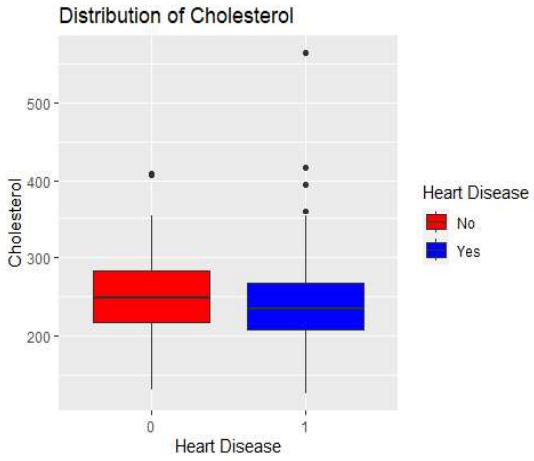


Fig 10: Distribution of Cholestrol

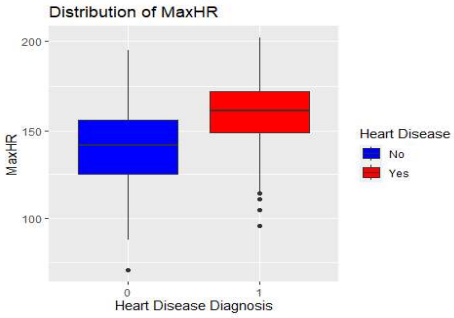


Fig 11: Distribution of Max Heart Rate

C. Hypothesis Testing (ANOVA &Chi-Square)

```
anova_sex <- aov(HeartDisease ~ sex, data = Heart.Att)
summary(anova_sex)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## sex           1   5.93   5.931   25.79 6.68e-07 ***
## Residuals   301  69.22   0.230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Convert binary variable to factor
heart  <- factor(Heart.Att$HeartDisease)

# Run ANOVA for HeartDisease and sex
anova_age <- aov(HeartDisease ~ age, data = Heart.Att)
summary(anova_age)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## age           1   3.82   3.819   16.12 7.52e-05 ***
## Residuals   301  71.33   0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Convert HeartDisease to numeric
Heart.Att$HeartDisease <- as.numeric(Heart.Att$HeartDisease) - 1

# Run ANOVA for HeartDisease and age
anova_hd_age <- aov(HeartDisease ~ age, data = Heart.Att)
summary(anova_hd_age)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## age           1   3.82   3.819   16.12 7.52e-05 ***
## Residuals   301  71.33   0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_hd_sex <- aov(HeartDisease ~ sex, data = Heart.Att)
summary(anova_hd_sex)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## sex           1   5.93   5.931   25.79 6.68e-07 ***
## Residuals   301  69.22   0.230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_hd_sex <- aov(HeartDisease ~ sex, data = Heart.Att)
summary(anova_hd_sex)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## sex           1   5.93   5.931   25.79 6.68e-07 ***
## Residuals   301  69.22   0.230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_hd_cp <- aov(HeartDisease ~ cp, data = Heart.Att)
summary(anova_hd_cp)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## cp            1  14.14  14.142   69.77 2.47e-15 ***
## Residuals   301  61.01   0.203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_hd_exang <- aov(HeartDisease ~ exang, data = Heart.Att)
summary(anova_hd_exang)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## exang         1  14.34  14.335   70.95 1.52e-15 ***
## Residuals   301  60.81   0.202
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova_hd_fbs <- aov(HeartDisease ~ fbs, data = Heart.Att)
summary(anova_hd_fbs)

##              Df Sum Sq Mean Sq F value Pr(>F)
## fbs           1   0.06 0.05911   0.237  0.627
## Residuals   301  75.09 0.24947

anova_hd_restecg <- aov(HeartDisease ~ restecg, data = Heart.Att)
summary(anova_hd_restecg)

##              Df Sum Sq Mean Sq F value Pr(>F)
## restecg       1   1.42   1.415   5.777 0.0168 *
## Residuals   301  73.73   0.245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#ChiSq

```
  # Sex and HeartDisease
table_hd_sex <- table(Heart.Att$HeartDisease, Heart.Att$sex)
chisq.test(table_hd_sex)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_hd_sex
## X-squared = 22.717, df = 1, p-value = 1.877e-06

# ChestPainType and HeartDisease
table_hd_cp <- table(Heart.Att$cp, Heart.Att$HeartDisease)
chisq.test(table_hd_cp)

##
##  Pearson's Chi-squared test
##
## data:  table_hd_cp
## X-squared = 81.686, df = 3, p-value < 2.2e-16

# FastingBS and HeartDisease
fbs_hd_table <- table(Heart.Att$fbs, Heart.Att$HeartDisease)
chisq.test(fbs_hd_table)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  fbs_hd_table
## X-squared = 0.10627, df = 1, p-value = 0.7444
```

```
# RestingECG and HeartDisease
restecg_hd_table <- table(Heart.Att$restecg, Heart.Att$HeartDisease)
chisq.test(restecg_hd_table)
```

```
## Warning in chisq.test(restecg_hd_table): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  restecg_hd_table
## X-squared = 10.023, df = 2, p-value = 0.006661
```

```
# ExerciseAngina and HeartDisease
exang_hd_table <- table(Heart.Att$exang, Heart.Att$HeartDisease)
chisq.test(exang_hd_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  exang_hd_table
## X-squared = 55.945, df = 1, p-value = 7.454e-14
```

```
# ST_Slope and HeartDisease
st_slope_table <- table(Heart.Att$slope, Heart.Att$HeartDisease)
chisq.test(st_slope_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  st_slope_table
## X-squared = 47.507, df = 2, p-value = 4.831e-11
```

```
age_groups <- cut(Heart.Att$age, breaks = c(0, 30, 40, 50, 60, 70, 120), labe
ls = c("0-30", "31-40", "41-50", "51-60", "61-70", "71+"))

# Perform chi-squared test between age groups and HeartDisease
age_table <-table(age_groups, Heart.Att$HeartDisease)
chisq.test(age_table)
```

```
## Warning in chisq.test(age_table): Chi-squared approximation may be incorre
ct
```

```
##
##  Pearson's Chi-squared test
##
## data:  age_table
## X-squared = 17.565, df = 5, p-value = 0.003544
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

D. Classification R Output

Fitting the OLS Regression model (initial set of predictors)

Initial predictors selected using business rationale:

```
Heart.Att <- read.table("Hearts.csv",header = T,sep = ",")

heart.logit.full <- glm(HeartDisease ~ ., family = binomial(link = "logit"),
data = Heart.Att)

summary(heart.logit.full)

##

## Call:

## glm(formula = HeartDisease ~ ., family = binomial(link = "logit"),

##     data = Heart.Att)

##

## Deviance Residuals:

##     Min       1Q   Median       3Q      Max

## -2.5849  -0.3872   0.1551   0.5863   2.6249

##

## Coefficients:

##              Estimate Std. Error z value Pr(>|z|)

## (Intercept)  3.450472   2.571479   1.342 0.179653

## age         -0.004908   0.023175  -0.212 0.832266

## sex         -1.758181   0.468774  -3.751 0.000176 ***

## cp           0.859851   0.185397   4.638 3.52e-06 ***

## trestbps    -0.019477   0.010339  -1.884 0.059582 .

## chol        -0.004630   0.003782  -1.224 0.220873

## fbs          0.034888   0.529465   0.066 0.947464

## restecg      0.466282   0.348269   1.339 0.180618

## thalach      0.023211   0.010460   2.219 0.026485 *

## exang       -0.979981   0.409784  -2.391 0.016782 *

## oldpeak     -0.540274   0.213849  -2.526 0.011523 *
```

```
## slope        0.579288   0.349807   1.656 0.097717 .
## ca          -0.773349   0.190885  -4.051 5.09e-05 ***
## thal        -0.900432   0.290098  -3.104 0.001910 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 211.44  on 289  degrees of freedom
## AIC: 239.44
##
## Number of Fisher Scoring iterations: 6
nulldev.logit <- heart.logit.full$null.deviance
resdev.logit <- heart.logit.full$deviance
Dev.Rsq <- (nulldev.logit - resdev.logit) /nulldev.logit
Dev.Rsq
## [1] 0.4937339
set.seed(1)
train <- sample(1:nrow(Heart.Att),0.7*nrow(Heart.Att))
train.heart <- Heart.Att[train,]
test.heart <- Heart.Att[-train,]`
glm.heart <- glm(HeartDisease~.,family = binomial(link = "logit"),data = trai
n.heart)
glm.prob.heart <- predict(glm.heart,test.heart,type = "response")
thres <- 0.5
glm.predict.heart <- ifelse(glm.prob.heart>0.5,1,0)
conf.mat <- table("Predicted"=glm.predict.heart,"Actual"=test.heart$HeartDise
ase)
colnames(conf.mat) <- c("No","Yes")
rownames(conf.mat) <- c("No","Yes")
conf.mat
##          Actual
## Predicted No Yes
```

```
##       No  30   4
##       Yes 10  47

accuracy <- mean(glm.predict.heart == test.heart$HeartDisease)
print(paste("Accuracy: ", round(accuracy, 3)))

## [1] "Accuracy:  0.846"
```

E. XI (Predictors are Independent):

Inspect for Multicollinearity(Full Dataset)

```
cond.index(heart.logit.full,data = Heart.Att)

##  [1]  1.000000  3.034843  3.345410  4.120956  4.443615  4.637964  5.768480
##  [8]  5.858653 11.007922 14.120515 18.317894 19.528439 28.746040 50.785451
```

Condition Index /VIF after variable selection

```
cond.index(heart.logit.small,data = Heart.Att)

##  [1]  1.000000  2.859495  3.797908  4.173373  4.344945  5.385541  5.709484
##  [8] 10.514332 13.431967 18.376342 26.768035 47.333980

vif(heart.logit.small)

##      age      sex       cp trestbps  restecg  thalach    exang  oldpeak
## 1.332740 1.192676 1.249748 1.125914 1.052987 1.410850 1.131778 1.407201
##    slope       ca     thal
## 1.489843 1.092859 1.038936
```

Variable Selection for Second Specification:

A brief snapshot of the setup for stepwise variable selection:

Age is included based on business knowledge

```
heart.full <- glm(HeartDisease ~ ., family = binomial(link = "logit"), data =
Heart.Att)
heart.null <- glm(HeartDisease ~ age, family = binomial(link = "logit"), data
= Heart.Att)

#To run stepwise variable selection
heart.step.backward <- step(heart.full,scope=list(lower = heart.null, upper =
heart.full), direction = "both",test="F")

## Start:  AIC=239.44
## HeartDisease ~ age + sex + cp + trestbps + chol + fbs + restecg +
##     thalach + exang + oldpeak + slope + ca + thal
```

```
## Warning in drop1.glm(fit, scope$drop, scale = scale, trace = trace, k = k,
: F
## test assumes 'quasibinomial' family
```

```
##            Df Deviance    AIC F value    Pr(>F)
## - fbs       1   211.44 237.44  0.0059 0.9386128
## - chol      1   212.91 238.91  2.0090 0.1574484
## - restecg   1   213.24 239.24  2.4629 0.1176549
## <none>          211.44 239.44
## - slope     1   214.13 240.13  3.6768 0.0561618 .
## - trestbps  1   215.08 241.08  4.9862 0.0263152 *
## - thalach   1   216.64 242.64  7.1177 0.0080633 **
## - exang     1   217.09 243.09  7.7345 0.0057737 **
## - oldpeak   1   218.24 244.24  9.3053 0.0024969 **
## - thal      1   221.36 247.36 13.5660 0.0002749 ***
## - sex       1   227.77 253.77 22.3261 3.596e-06 ***
## - ca        1   228.84 254.84 23.7960 1.772e-06 ***
## - cp        1   236.18 262.18 33.8247 1.596e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=237.44
```

Logistic Regression Testing for the smaller set of predictors (second specification):

```
heart.logit.small <- glm(HeartDisease ~age+sex + cp + trestbps + restecg +
    thalach + exang + oldpeak + slope + ca + thal, family = binomial(link = "
logit"), data = Heart.Att)
summary(heart.logit.small)
```

```
##
## Call:
## glm(formula = HeartDisease ~ age + sex + cp + trestbps + restecg +
##     thalach + exang + oldpeak + slope + ca + thal, family = binomial(link
= "logit"),
##     data = Heart.Att)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6327  -0.4002   0.1573   0.5904   2.5452
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.796966   2.502636   1.118 0.263734
## age         -0.009209   0.022534  -0.409 0.682787
## sex         -1.589990   0.437347  -3.636 0.000277 ***
## cp           0.873115   0.182389   4.787 1.69e-06 ***
## trestbps    -0.019420   0.010258  -1.893 0.058346 .
```

```
## restecg      0.537152   0.340871   1.576 0.115067
## thalach      0.021162   0.010208   2.073 0.038166 *
## exang       -0.979848   0.404411  -2.423 0.015397 *
## oldpeak     -0.561492   0.212197  -2.646 0.008143 **
## slope        0.564993   0.347014   1.628 0.103492
## ca          -0.753383   0.187182  -4.025 5.70e-05 ***
## thal        -0.936476   0.283632  -3.302 0.000961 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 212.91  on 291  degrees of freedom
## AIC: 236.91
##
## Number of Fisher Scoring iterations: 6
```

```
nulldev.logits <- heart.logit.small$null.deviance
resdev.logits <- heart.logit.small$deviance
Dev.Rsq.s <- (nulldev.logits - resdev.logits) /nulldev.logits
Dev.Rsq.s
```

```
## [1] 0.4902054
```

```
cond.index(heart.logit.small,data = Heart.Att)
```

```
##  [1]  1.000000  2.859495  3.797908  4.173373  4.344945  5.385541  5.709484
##  [8] 10.514332 13.431967 18.376342 26.768035 47.333980
```

```
vif(heart.logit.small)
```

```
##      age      sex       cp trestbps  restecg  thalach    exang  oldpeak
## 1.332740 1.192676 1.249748 1.125914 1.052987 1.410850 1.131778 1.407201
##    slope       ca     thal
## 1.489843 1.092859 1.038936
```

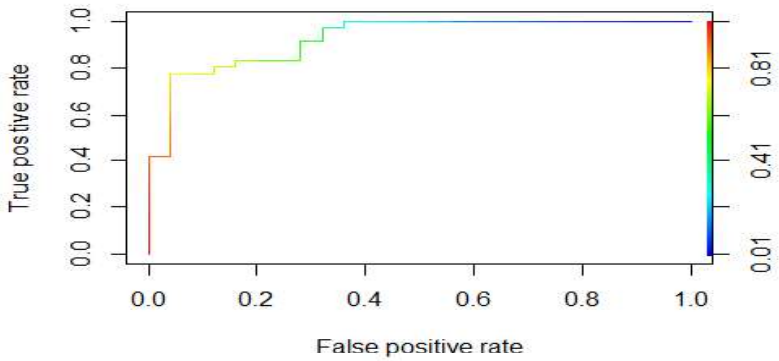
```
#Confusion Matrix (full dataset)

accuracy.f <- accuracy <- mean(logit.predict.heart.f==test.heart$HeartDisease
)

TruN <- conf.mat[1, 1]
TruP <- conf.mat[2, 2]
FalN <- conf.mat[1, 2]
FalP <- conf.mat[2, 1]
TotN <- TruN + FalP
TotP <- TruP + FalN
Tot <- TotN + TotP
Accuracy.Rate <- (TruN + TruP) / Tot
```

```
Error.Rate <- (FalN + FalP) / Tot
Sensitivity <- TruP / TotP
Specificity <- TruN / TotN
FalseP.Rate <-1- Specificity
logit.rates.30 <- c(Accuracy.Rate, Error.Rate, Sensitivity, Specificity, Fals
eP.Rate)
names(logit.rates.30) <- c("Accuracy Rate", "Error Rate", "Sensitivity", "Spe
cificity", "False Positives")
print(logit.rates.30, digits = 2)
```

```
##   Accuracy Rate      Error Rate     Sensitivity     Specificity False Posi
tives
##            0.84            0.16            0.97            0.64
0.36
```

#AUC (Full Dataset)

```
rocr.pred.logit <- prediction(logit.prob.heart.f, test.heart$HeartDisease)

rocr.perf.logit <- performance(rocr.pred.logit, measure = "tpr", x.measure =
"fpr")
plot(rocr.perf.logit, colorize = TRUE)
```



```
#AUC
auc_roc <- performance(rocr.pred.logit, measure = "auc")
auc_roc_value <- as.numeric(auc_roc@y.values)
print(paste("AUC-ROC =", round(auc_roc_value, 3)))
```

```
## [1] "AUC-ROC = 0.927"
```

```
#Confusion Matrix(Small Dataset)

set.seed(1)
thresh <- 0.80
train <- sample(1:nrow(Heart.Att),thresh*nrow(Heart.Att))
train.heart <- Heart.Att[train,]
test.heart <- Heart.Att[-train,]
heart.logit.small <- glm(HeartDisease~age+sex + cp + trestbps + restecg +
    thalach + exang + oldpeak + slope + ca + thal,family = binomial(link = "l
ogit"),data = train.heart)
logit.heart.predict.s <- predict(heart.logit.small,test.heart)
logit.prob.heart.s <- predict(heart.logit.small,test.heart,type = "response")
thres <- 0.3
logit.predict.heart.s <- ifelse(logit.prob.heart.s>0.3,1,0)
conf.mat <- table("Predicted"=logit.predict.heart.s,"Actual"=test.heart$Heart
Disease)
colnames(conf.mat) <- c("No","Yes")
rownames(conf.mat) <- c("No","Yes")
conf.mat
```

```
##          Actual
## Predicted No Yes
##       No  17   1
##       Yes  8  35
```

```
accuracy.s <- mean(logit.predict.heart.s == test.heart$HeartDisease)
print(paste("Accuracy: ", round(accuracy, 3)))
```

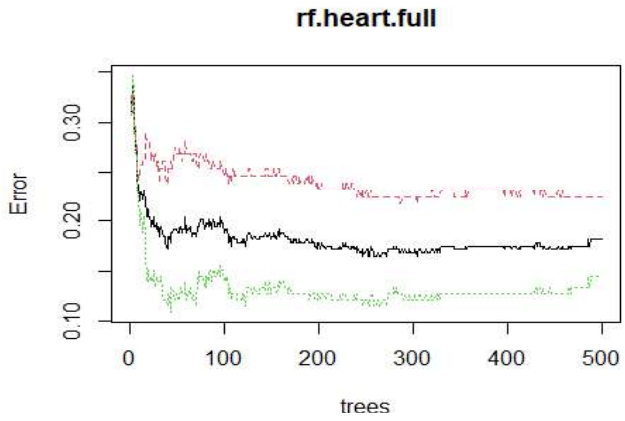
```
## [1] "Accuracy:  0.836"
```

Random Forest(full)

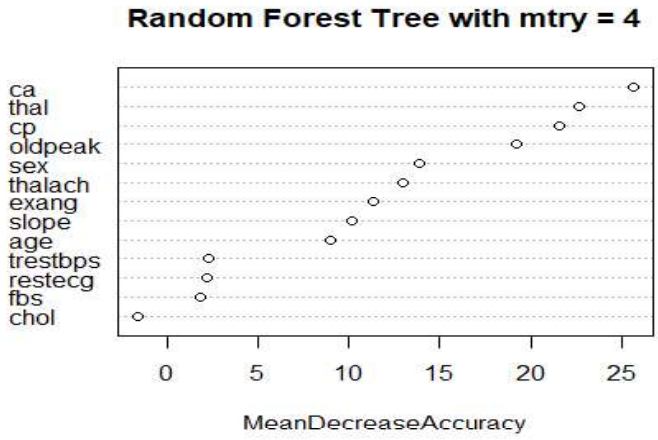
```
Heart.Att$HeartDisease<- as.factor(Heart.Att$HeartDisease)
rf.heart.full <- randomForest(HeartDisease ~ ., data = Heart.Att, mtry = 4, i
mportance = T)
rf.heart.full
```

```
##
## Call:
##  randomForest(formula = HeartDisease ~ ., data = Heart.Att, mtry = 4,
importance = T)
```

```
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 18.15%
## Confusion matrix:
##     0   1 class.error
## 0 107  31   0.2246377
## 1  24 141   0.1454545

plot(rf.heart.full)
```




```
m <- 4
varImpPlot(rf.heart.full, type=1, main = paste("Random Forest Tree with mtry
=", m))
```




```
importance(rf.heart.full)

##                    0          1 MeanDecreaseAccuracy MeanDecreaseGini
## age       4.1212540  7.8476046             8.978033        12.475781
## sex       7.0565246 12.6416010            13.850461         5.037384
```

```
## cp       17.6602735 13.9987585          21.617562        21.813120
## trestbps  0.6691086  2.1963669           2.227821        10.358314
## chol     -3.5113215  0.4670988          -1.675368        11.269069
## fbs      -0.8331074  2.6616233           1.789777         1.393731
## restecg   2.2713676  0.6196615           2.138430         2.730737
## thalach   5.7166501 11.6678083          12.996518        17.067945
## exang    10.3657471  5.3466938          11.322344         8.725838
## oldpeak  15.0830690 13.2962063          19.269077        17.081487
## slope     8.4991163  5.4953347          10.104248         6.294197
## ca       17.8096583 22.0153440          25.687544        18.744350
## thal     13.2818145 19.3349991          22.671327        16.398559
```

10FCV-Random Forest(full)

```
rf.full.caret.10FCV <- train(HeartDisease ~ ., data = Heart.Att, method = "rf
",trControl=trainControl(method = "cv", number = 10))
rf.full.caret.10FCV
```

```
## Random Forest
##
## 303 samples
##  13 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 272, 273, 272, 273, 273, 272, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8218984  0.6378505
##    7    0.8086726  0.6121335
##   13    0.8052243  0.6038263
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

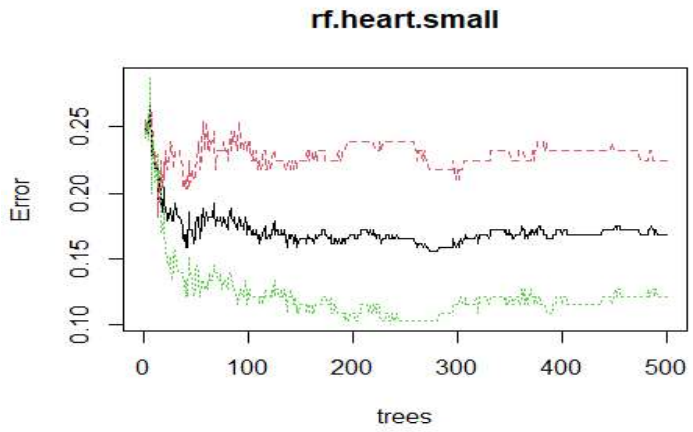
Random Forest (Small)

```
set.seed(1)
Heart.Att$HeartDisease<- as.factor(Heart.Att$HeartDisease)
rf.heart.small <- randomForest(HeartDisease ~  age+sex + cp + trestbps + rest
ecg + thalach + exang + oldpeak + slope + ca + thal,data = Heart.Att, mtry =
4, importance = T)
rf.heart.small
```

```
##
## Call:
##  randomForest(formula = HeartDisease ~ age + sex + cp + trestbps +      re
stecg + thalach + exang + oldpeak + slope + ca + thal,      data = Heart.Att,
mtry = 4, importance = T)
```

```
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 16.83%
## Confusion matrix:
##     0   1 class.error
## 0 107  31   0.2246377
## 1  20 145   0.1212121
```

```
plot(rf.heart.small)
```



10FCV-Random Forest(Small)

```
rf.small.caret.10FCV <- train(HeartDisease ~ age+sex + cp + trestbps +
restecg + thalach + exang + oldpeak + slope + ca + thal, data = Heart.Att,
method = "rf",trControl=trainControl(method = "cv", number = 10))
rf.small.caret.10FCV$results$Accuracy
```

```
## [1] 0.8383871 0.8086022 0.8052688
```

```
importance(rf.heart.small)
```

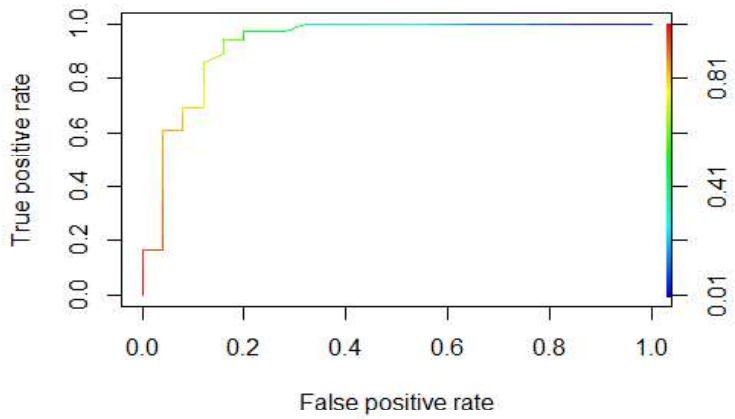
```
##                   0         1 MeanDecreaseAccuracy MeanDecreaseGini
## age       3.9048044  9.499234             9.963012        14.304423
## sex       8.4607570 10.849387            13.662894         5.125847
## cp       20.4306471 15.681169            24.762305        24.478321
## trestbps  0.8113599  3.062232             2.798764        12.336002
## restecg   4.4841251  1.708364             4.286152         3.471794
## thalach   3.5029409 10.558383            10.434142        18.968627
## exang    11.6133258  4.192153            10.580438         8.259809
## oldpeak  14.3721144 12.856467            18.847557        18.406634
## slope     8.7148702  3.591583             9.086645         6.908307
## ca       19.7334440 21.640006            26.994532        19.453096
## thal     14.2668222 17.133564            20.930589        17.578785
```

```
rf.heart.small$confusion
```

```
##     0   1 class.error
## 0 107  31   0.2246377
## 1  20 145   0.1212121
```

Sensitivity and AUC (RF Small)

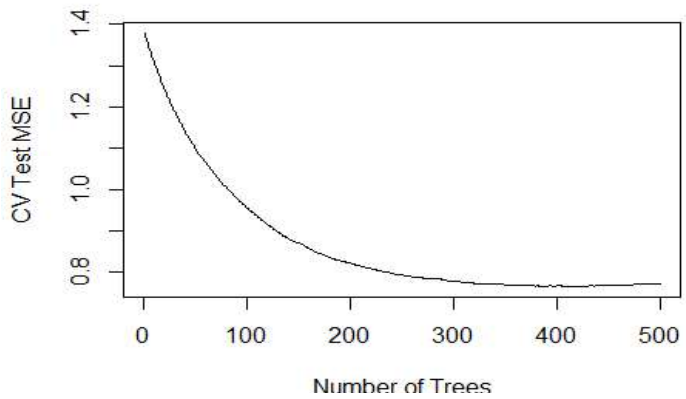
```
set.seed(1)
rf.model.small <- randomForest(HeartDisease ~ age+sex + cp + trestbps + reste
cg + thalach + exang + oldpeak + slope + ca + thal, data = train)
predicted.probs.small <- predict(rf.model.small, newdata = test, type = "prob
")
pred <- prediction(predicted.probs.small[, 2], test$HeartDisease)
perf <- performance(pred, "tpr", "fpr")
plot(perf, colorize = T)
```



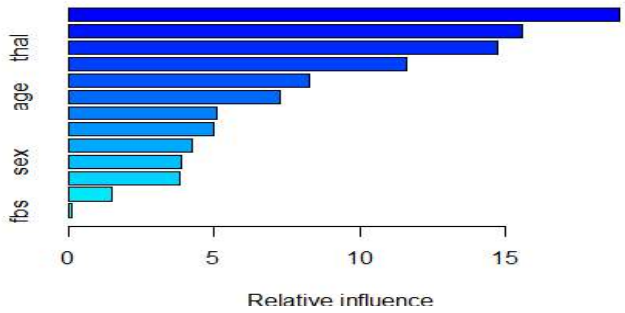

```
auc <- performance(pred, "auc")
auc.name <- auc@y.name[[1]]
auc.value <- round(auc@y.values[[1]], digits = 3)
paste(auc.name, "is", auc.value)

## [1] "Area under the ROC curve is 0.929"

predicted_class <- predict(rf.model.small, newdata = test)
conf.matrix.f <- confusionMatrix(predicted_class,test$HeartDisease)
sensitivity <- conf.matrix.f$byClass["Sensitivity"]
sensitivity

## Sensitivity
##         0.8
```

Boosted Trees

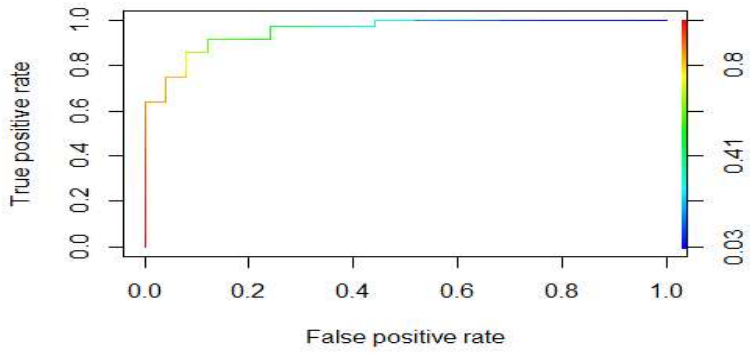
```
Heart.Att <- read.table("Hearts.csv",header = T,sep =",")
library(gbm)

## Loaded gbm 2.1.8.1

library(ROCR)
boost.heart <- gbm(HeartDisease ~ ., data = Heart.Att, distribution = "bernou
lli", shrinkage = 0.01,cv.folds = 10, n.trees = 500, interaction.depth = 4)
boost.heart

## gbm(formula = HeartDisease ~ ., distribution = "bernoulli", data = Heart.A
tt,
##     n.trees = 500, interaction.depth = 4, shrinkage = 0.01, cv.folds = 10)
## A gradient boosted model with bernoulli loss function.
## 500 iterations were performed.
## The best cross-validation iteration was 426.
## There were 13 predictors of which 13 had non-zero influence.

best.num.trees <- which.min(boost.heart$cv.error)
min.10FCV.error <- round(min(boost.heart$cv.error),digits = 4)

paste("Min 10FCV Test Error =", min.10FCV.error,"at", best.num.trees, "trees"
)

## [1] "Min 10FCV Test Error = 0.7662 at 426 trees"

plot(boost.heart$cv.error, type = "l",xlab = "Number of Trees", ylab = "CV Te
st MSE")
```




```
summary(boost.heart)
```




```
##                var     rel.inf
## cp             cp 18.91012924
## ca             ca 15.58203887
## thal         thal 14.72645405
## oldpeak   oldpeak 11.61839176
## thalach   thalach  8.27284839
## age           age  7.28348159
## exang       exang  5.09172860
## chol         chol  4.97615601
## slope       slope  4.25823097
## sex           sex  3.87872569
## trestbps trestbps  3.82899893
## restecg   restecg  1.48188317
## fbs           fbs  0.09093272

set.seed(1)
split.boost <- sample(1:nrow(Heart.Att), round(0.8 * nrow(Heart.Att)))
train <- Heart.Att[split.boost, ]
test <- Heart.Att[-split.boost, ]


boosted.model <- gbm(HeartDisease ~ ., data = train, distribution = "bernoull
i", n.trees = 500, interaction.depth = 3, shrinkage = 0.01)

predicted.probs.boosted <- predict(boosted.model, newdata = test, type = "res
ponse", n.trees = 500)

pred <- prediction(predicted.probs.boosted, test$HeartDisease)
perf <- performance(pred, "tpr", "fpr")
plot(perf, colorize = T)
```




```
auc <- performance(pred, "auc")
auc.name <- auc@y.name[[1]]
auc.value <- round(auc@y.values[[1]], digits = 3)
paste(auc.name, "is", auc.value)
```

```
## [1] "Area under the ROC curve is 0.954"
```

```
predicted.probs.f <- predict(boosted.model, newdata = test, type = "response"
)
```

```
## Using 500 trees...
```

```
threshold <- 0.3
predicted.class <- ifelse(predicted.probs.f > threshold, 1, 0)
conf.matrix <- table(Predicted = predicted.class, Actual =test$HeartDisease)
print(conf.matrix)
```

```
##           Actual
## Predicted  0  1
##         0 14  1
##         1 11 35
```

```
cm.boost <- table(Predicted = predicted.class, Actual = test$HeartDisease)
sensitivity <- cm.boost[2, 2] / (cm.boost[2, 2] + cm.boost[1, 2])
specificity <- cm.boost[1, 1] / (cm.boost[1, 1] + cm.boost[2, 1])
FPR <- cm.boost[2, 1] / (cm.boost[2, 1] + cm.boost[1, 1])
accuracy <- sum(diag(cm.boost)) / sum(cm.boost)
cat("Sensitivity:", sensitivity, "\n")
```

```
## Sensitivity: 0.9722222
```

```
cat("Specificity:", specificity, "\n")
```

```
## Specificity: 0.56
```

```
cat("False Positive Rate (FPR):", FPR, "\n")
```

```
## False Positive Rate (FPR): 0.44
```

```
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.8032787
```

Boosted Trees (Small)

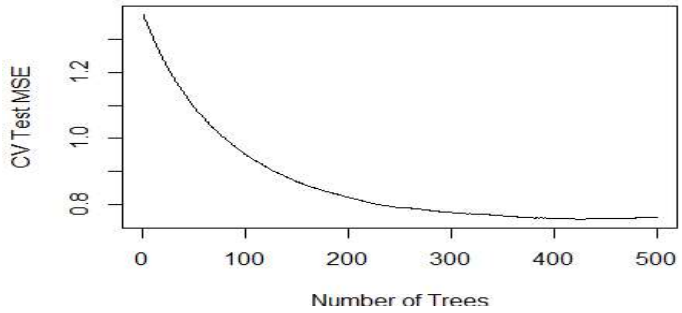
```
boost.heart.small <- gbm(HeartDisease ~ age+sex + cp + trestbps + restecg + t
halach + exang + oldpeak + slope + ca + thal, data = Heart.Att, distribution
= "bernoulli", shrinkage = 0.01,cv.folds = 10, n.trees = 500, interaction.dep
th = 4)
boost.heart.small
```

```
## gbm(formula = HeartDisease ~ age + sex + cp + trestbps + restecg +
##     thalach + exang + oldpeak + slope + ca + thal, distribution = "bernoul
li",
##     data = Heart.Att, n.trees = 500, interaction.depth = 4, shrinkage = 0.
01,
##     cv.folds = 10)
## A gradient boosted model with bernoulli loss function.
## 500 iterations were performed.
## The best cross-validation iteration was 425.
## There were 11 predictors of which 11 had non-zero influence.
```

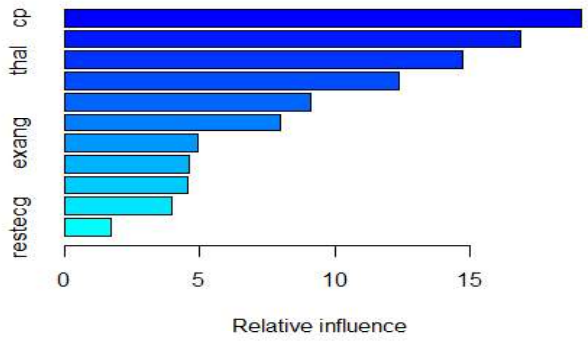
```
best.num.trees.s <- which.min(boost.heart.small$cv.error)
min.10FCV.error.s <- round(min(boost.heart.small$cv.error),digits = 4)

paste("Min 10FCV Test Error =", min.10FCV.error.s,"at", best.num.trees.s, "tr
ees")
```

```
## [1] "Min 10FCV Test Error = 0.7544 at 425 trees"
```

```
plot(boost.heart.small$cv.error, type = "l",xlab = "Number of Trees", ylab =
"CV Test MSE")
```

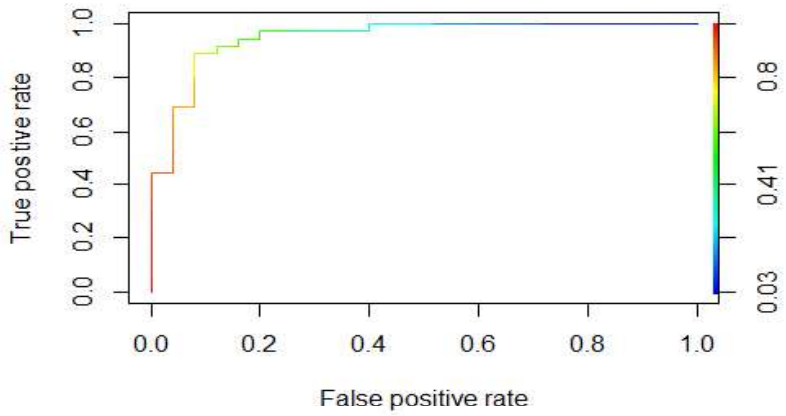



```
summary(boost.heart.small)
```




```
##                var   rel.inf
## cp             cp 19.125469
## ca             ca 16.884845
## thal         thal 14.705087
## oldpeak   oldpeak 12.385715
## thalach   thalach  9.112681
## age           age  7.981683
## exang       exang  4.928493
## slope       slope  4.605910
## trestbps trestbps  4.580999
## sex           sex  3.956486
## restecg   restecg  1.732631
```

```
set.seed(1)
split.boost <- sample(1:nrow(Heart.Att), round(0.8 * nrow(Heart.Att)))
train <- Heart.Att[split.boost, ]
test <- Heart.Att[-split.boost, ]


boosted.model.small <- gbm(HeartDisease ~ age+sex + cp + trestbps + restecg +
thalach + exang + oldpeak + slope + ca + thal, data = train, distribution = "
bernoulli", n.trees = 500, interaction.depth = 3, shrinkage = 0.01)

predicted.probs.small <- predict(boosted.model.small, newdata = test, type =
"response", n.trees = 500)

pred <- prediction(predicted.probs.small, test$HeartDisease)
perf <- performance(pred, "tpr", "fpr")
plot(perf, colorize = T)
```




```
auc <- performance(pred, "auc")
auc.name <- auc@y.name[[1]]
auc.value <- round(auc@y.values[[1]], digits = 3)
paste(auc.name, "is", auc.value)
```

```
## [1] "Area under the ROC curve is 0.95"
```

```
predicted.probs <- predict(boosted.model.small, newdata = test, type = "respo
nse")
```

```
## Using 500 trees...
```

```
threshold <- 0.3
predicted.class <- ifelse(predicted.probs > threshold, 1, 0)
conf.matrix <- table(Predicted = predicted.class, Actual =test$HeartDisease)
print(conf.matrix)
```

```
##           Actual
## Predicted  0  1
##         0 15  1
##         1 10 35
```

```
cm.boost.s <- table(Predicted = predicted.class, Actual = test$HeartDisease)
sensitivity <- cm.boost.s[2, 2] / (cm.boost.s[2, 2] + cm.boost.s[1, 2])
specificity <- cm.boost.s[1, 1] / (cm.boost.s[1, 1] + cm.boost.s[2, 1])
FPR <- cm.boost.s[2, 1] / (cm.boost.s[2, 1] + cm.boost.s[1, 1])
accuracy <- sum(diag(cm.boost.s)) / sum(cm.boost.s)
cat("Sensitivity:", sensitivity, "\n")
```

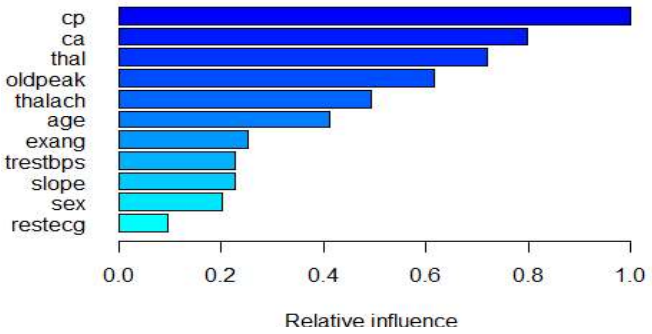
```
## Sensitivity: 0.9722222
```

```
cat("Specificity:", specificity, "\n")
```

```
## Specificity: 0.6
```

```
cat("False Positive Rate (FPR):", FPR, "\n")
```

```
## False Positive Rate (FPR): 0.4

cat("Accuracy:", accuracy, "\n")

## Accuracy: 0.8196721


#fit with entire dataset
library(gbm)
boosted.model.all <- gbm(HeartDisease ~ age+sex + cp + trestbps + restecg + t
halach + exang + oldpeak + slope + ca + thal, data = Heart.Att, distribution
= "bernoulli", n.trees = 500, interaction.depth = 4, shrinkage = 0.01)

summary.gbm(boosted.model.all,)


A gradient boosted model with bernoulli loss function.

500 iterations were performed.

There were 11 predictors of which 11 had non-zero influence.
```




```
##                var   rel.inf
## cp             cp 19.836137
## ca             ca 15.830102
## thal         thal 14.252398
## oldpeak   oldpeak 12.243008
## thalach   thalach  9.765153
## age           age  8.170182
## exang       exang  5.027810
## trestbps trestbps  4.496952
## slope       slope  4.494526
## sex           sex  4.004988
## restecg   restecg  1.878743
```

References

1. CDC-Heart Disease Facts: https://www.cdc.gov/heartdisease/facts.htm