Heart Attack Prediction Capstone Project

John Sowers

9/22/2020

INTRODUCTION

This project report is submitted to fulfill part of the requirement for the Edx Data Science: Capstone Course. The program code for this project was compiled utilizing R version 4.02 / R Studio version 1.3.1073. It was run on a Dell m15 R3 laptop computer with an i7 Central Processing Unit (CPU) with 16 Gigabytes of Random Access Memory (RAM) utilizing Microsoft Windows 10.

According to the Center for Disease Control (CDC) website of https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm, the leading cause of death in the United States is Heart Disease, claiming over 500,000 lives each year. Early detection of this disease in critical in lowering these numbers; and according to an article listed in the Artificial Intelligence News at the website of https://artificialintelligence-news.com/2019/05/14/ml-algorithm-predicts-heart-attacks, machine learning could be utilized to predict heart attacks with a 90% accuracy.

The purpose of this project is to validate this claim, utilizing a public domain data set that was previously cleansed of all personal private information. This data set can be found on Kaggle at the following website https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility; and it was unzipped, cleansed, and downloaded to the GitHub Project website of https://github.com/johndsowers/HeartAttack/raw/master/heart.csv for ease of downloading.

The data structure contains 303 unknown patient medical conditions with 13 medical tests and the heart disease condition determination. This downloaded data set was further split into separate training and testing data sets with the training data containing 90% of the cases and the testing data containing 10%. The structure of the combined data set is shown below:

```
##  [1] "'data.frame':\t303 obs. of  14 variables:"
##  [2] " $ age     : num  63 37 41 56 57 57 56 44 52 57 ..."
##  [3] " $ sex     : num  1 1 0 1 0 1 0 1 1 1 ..."
##  [4] " $ cp      : num  3 2 1 1 0 0 1 1 2 2 ..."
##  [5] " $ trestbps: num  145 130 130 120 120 140 140 120 172 150 ..."
##  [6] " $ chol    : num  233 250 204 236 354 192 294 263 199 168 ..."
##  [7] " $ fbs     : num  1 0 0 0 0 0 0 0 1 0 ..."
##  [8] " $ restecg : num  0 1 0 1 1 1 0 1 1 1 ..."
##  [9] " $ thalach : num  150 187 172 178 163 148 153 173 162 174 ..."
## [10] " $ exang   : num  0 0 0 0 1 0 0 0 0 0 ..."
## [11] " $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ..."
## [12] " $ slope   : num  0 0 2 2 2 1 1 2 2 2 ..."
## [13] " $ ca      : num  0 0 0 0 0 0 0 0 0 0 ..."
## [14] " $ thal    : num  1 2 2 2 2 1 2 3 3 2 ..."
## [15] " $ target  : num  1 1 1 1 1 1 1 1 1 1 ..."
```

METHODOLOGY

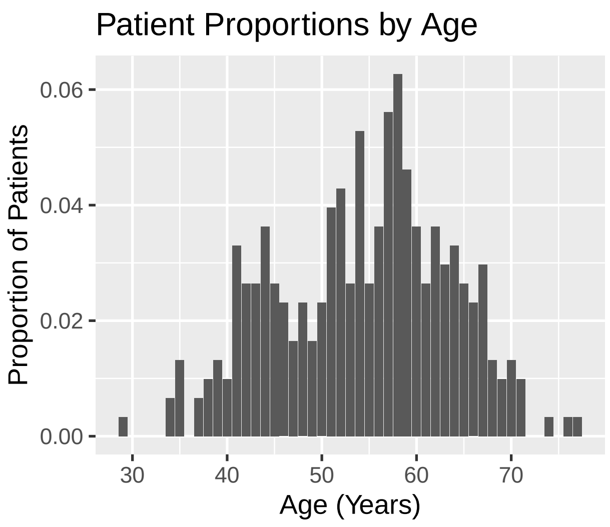
The data set was created from the records of patients' who had been tested for the presence of heart disease. These patients may have previous indications of a heart condition that initiated the scheduling of the appointments; and this may skew some resulting distributions. However, the predictors contained in the data and the resulting heart disease condition remain valid. Medical nomenclature and abbreviations were utilized for the predictors and heart disease determination in the data set and are used throughout this report. These factors include:

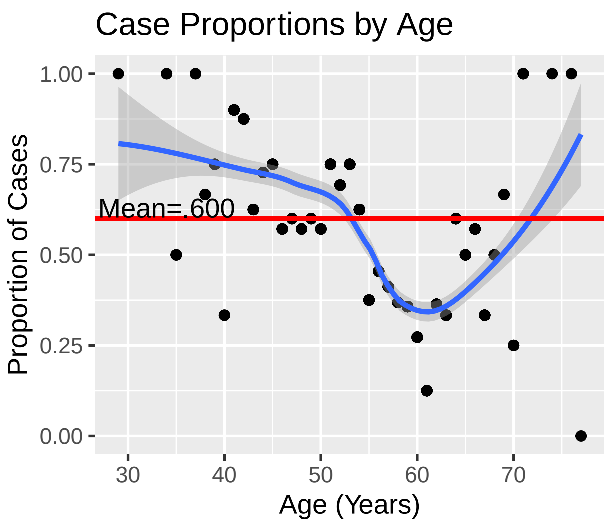
1. Age
2. Gender (sex)
3. Chest Pain Type (cp)
4. Resting Blood Pressure (trestbps)
5. Cholesterol (chol)
6. Fasting Blood Sugar (fbs)
7. Resting Electrocardiographic Results (restecg)
8. Maximum Heart Rate Achieved (thalach)
9. Exercise Induced Angina (exang)
10. ST Depression Induced by Exercise Relative to Rest (oldpeak)
11. Slope of the Peak Exercise ST Segment
12. Number of Major Vessels (ca)
13. Thalassemia Blood Disorder (thal)
14. Presence of Heart Disease (target)

Age Factor

The first factor to consider is that of age. The age of the patients who were tested for heart disease ranged from 29 years through 77 years with the mean age of patients of 54.4 years. The age groups that contained the minimum number of patients were those that were at the extreme left and right of this range at 29 years as well as 74 and 77 years of 1 patient. The age group that contained the maximum number of patients was 58 years with 19 patients. The proportion of patients follow these numbers, as are shown in the first figure below.

The proportion of patients by age who were tested and then diagnosed with heart disease differ from the proportions of patients. The minimum is 0% for the one patient of 77 years with the maximum of 100% for patients who are 29, 34, 37, 71, 74, and 76 years with an overall mean heart disease rate of 60%. The data indicates that the proportion of patients who tested positive for an age group decreased with increasing number of patients as can be seen around the age group of 60 years. The proportion of patients diagnosed with heart disease is shown in the second figure below.

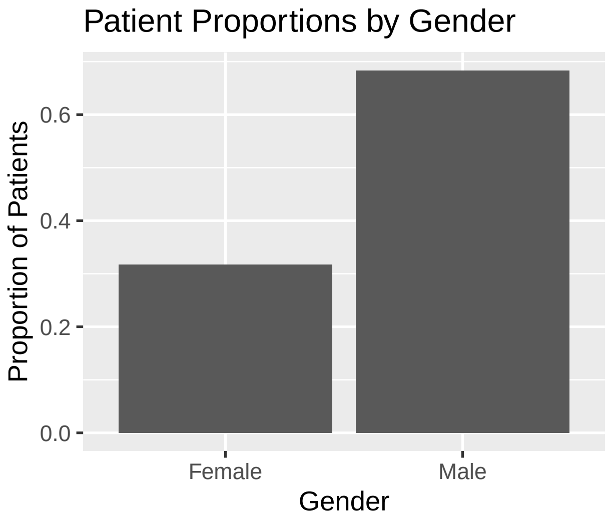


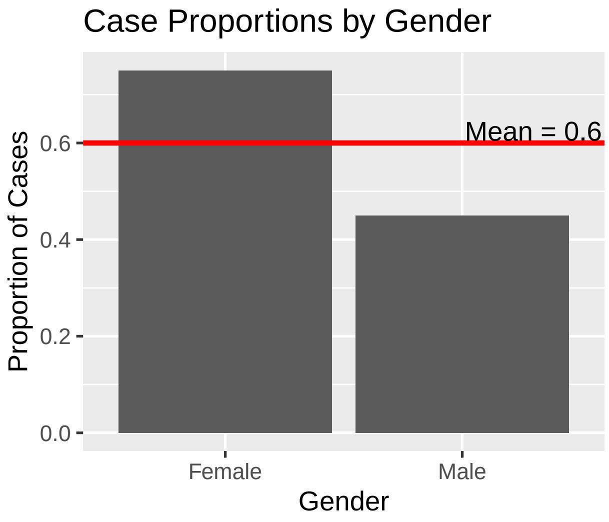


Gender Factor

The next factor to consider is that of gender. From the data set, patients were listed as a “0” for female or “1” for male. Of the patients tested, males made up the majority with 68.3% male versus 31.7% female. The proportion of patients who were tested for heart disease is shown in the first figure below.

The proportion of patients by gender who were tested and then diagnosed with heart disease differ from the proportions of the number of patients. Of the patients tested for heart disease, 44.9% of males tested positive for the disease while 75.1% of females tested positive. The proportion of patients diagnosed with heart disease is shown in the second figure below.

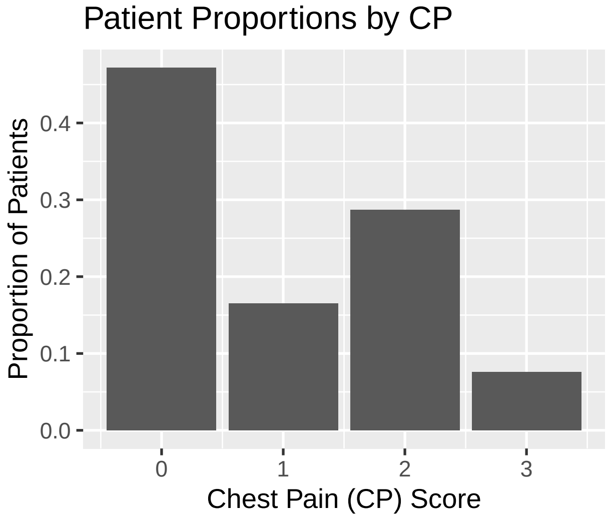


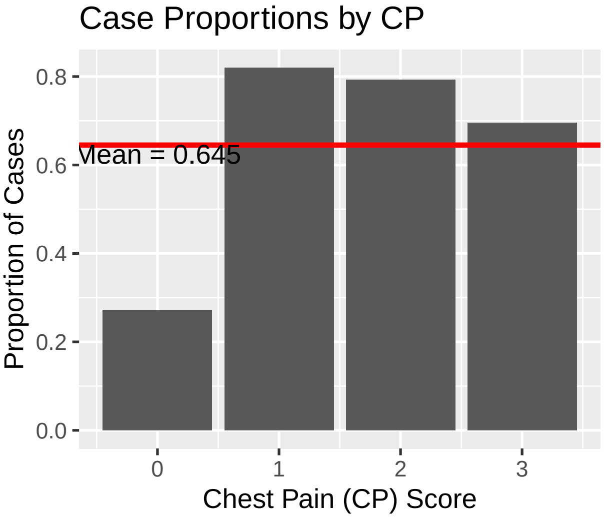


Chest Pain Factor

The next factor to consider is that of chest pain. From the data set, patients were listed with chest pain types of “0” through “3”. Of the four types of chest pain, 47.2% of the patients reported a pain type of “0”, 16.5% reported a pain type of “1”, 28.7% reported a pain type of “2”, and 7.59% reported a pain type of “3”. The proportion of patients who were tested for heart disease is shown in the first figure below.

The proportion of patients by pain types indicate those with with pain types of higher numbers having higher risks of heart disease. Of those testing positive for the disease, 27.3% reported pain type of “0”, 82.0% reported a pain type of “1”, 79.3% reported a pain type of “2”, and 69.6% reported a pain type of “3”. The proportion of patients diagnosed with heart disease is shown in the second figure below.

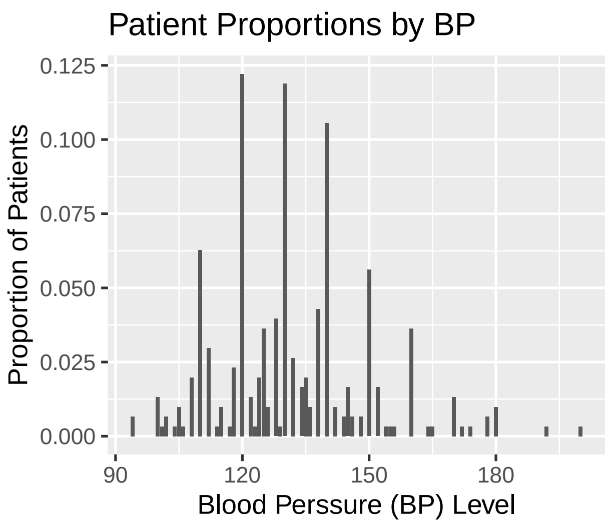


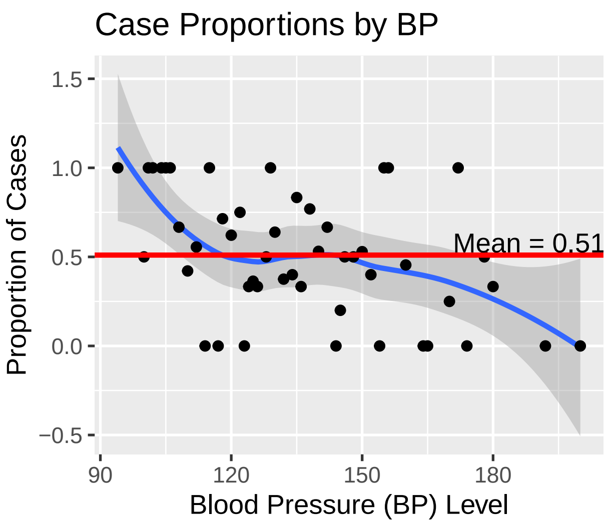


Blood Pressure Factor

The next factor to consider is that of resting blood pressure levels, as measured in mm Hg. From the data set, patients’ blood pressure results ranged from a minimum of 94 to a maximum of 200. From this range the minimum proportion of patients showed that .330% of the patients had a blood pressure of 101 with the maximum proportion of patients showed that 12.2% had a blood pressure of 120, which is considered in the normal range of blood pressure. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, the presence of the disease varied among patients with positive cases having both low and high blood pressures but generally trending downward with increasing blood pressure levels. The proportion of patients diagnosed with heart disease is shown in the second figure below.

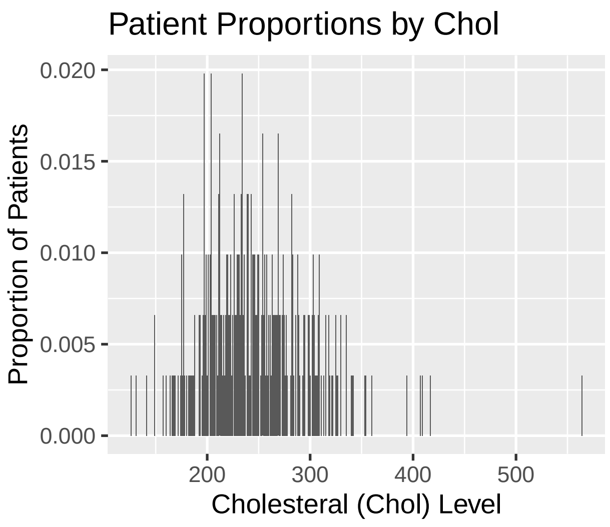


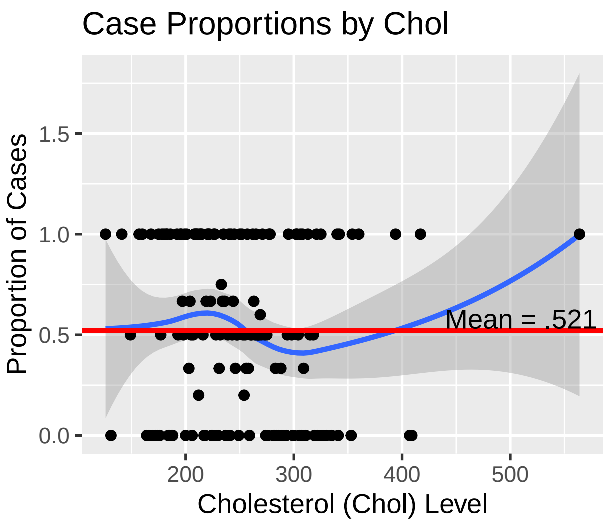


Cholesterol Factor

The next factor to consider is that of resting cholesterol levels, as measured in mg / dl. From the data set, patients' cholesterol results ranged from a minimum of 126 to a maximum of 564. From this range .330% of the patients had a cholesterol levels of 126 with 1.98% of patients having a cholesterol level of 198. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, the presence of the disease varied among patients with positive cases having both low and high cholesterol levels but generally trending upward with increasing cholesterol levels. The proportion of patients diagnosed with heart disease is shown in the second figure below.

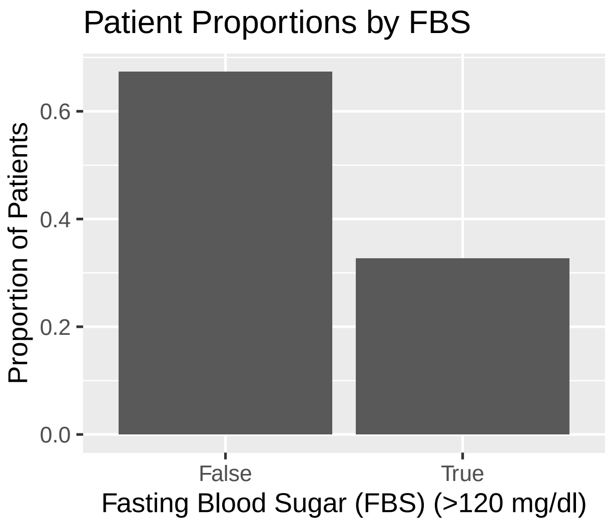


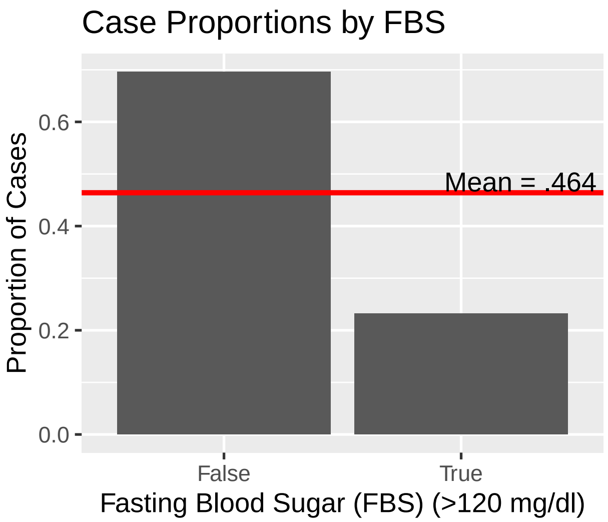


Blood Sugar Factor

The next factor to consider is that of fasting blood sugar levels, as measured in mg / dl. The data indicates returns this value as either a "0" or "1" with "0" indicating blood sugar below or equal to 120 mg/dl while a "1" indicating blood sugar above 120 mg/dl. From the data set, 67.3% of the patients tested had blood sugar levels less than or equal to 120 mg/dl while 32.7% of the patients had blood sugar levels exceeding 120 mg/dl. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, heart disease was present in patients with lower fasting blood sugar than those with higher blood sugar levels with 69.6% of the patients with heart disease having lower blood sugar levels and 23.2% of the patients having higher blood sugar levels. The proportion of patients diagnosed with heart disease is shown in the second figure below.

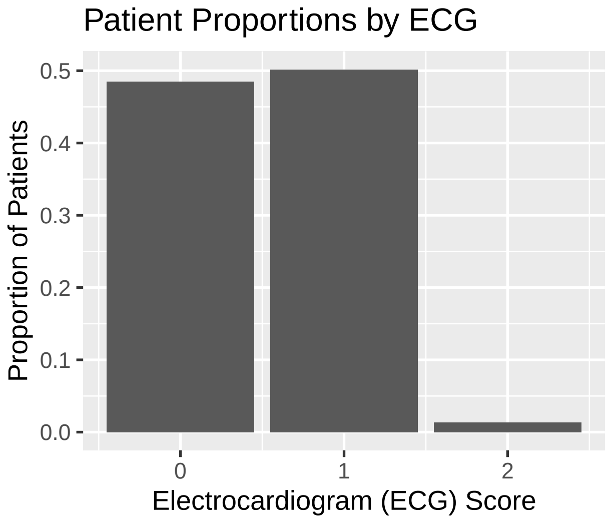


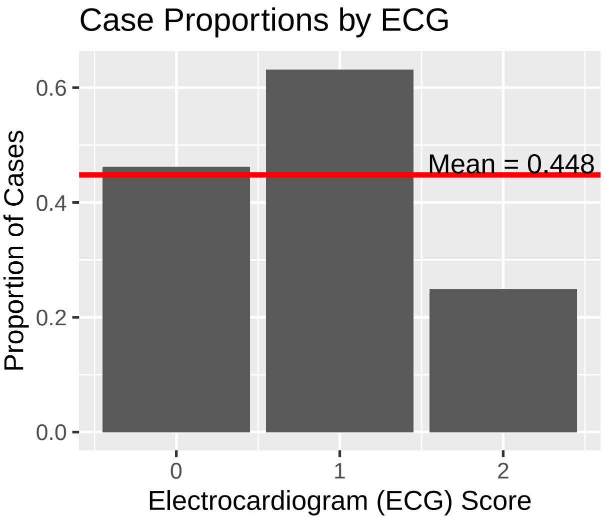


Electrocardiographic (ECG) Factor

The next factor to consider is that of resting electrocardiographic (ECG) results. These results are categorized by values of “0”, “1”, and “2.” Of the patients who were tested, 48.5% returned ECG results of “0”, 50.2% returned results of “1”, and 1.32% returned results of “2.” The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, an ECG result of “0” indicated the presence of heart disease 46.3% of the time, a result of “1” indicated the presence of the disease 63.2% of the time, and a result of a “2” indicated the presence of the disease 25% of the time. The proportion of patients diagnosed with heart disease is shown in the second figure below.

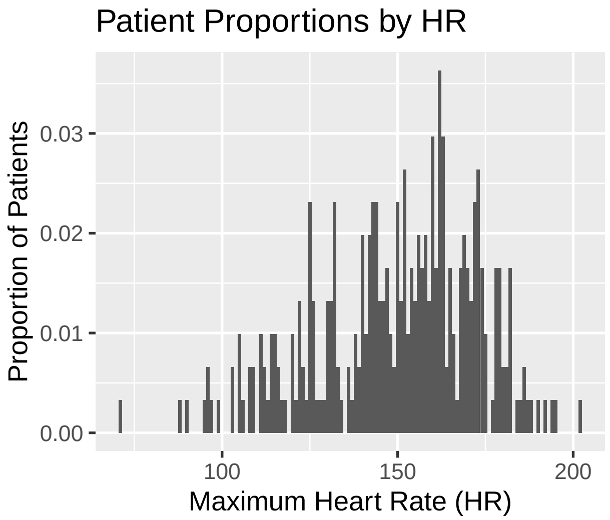


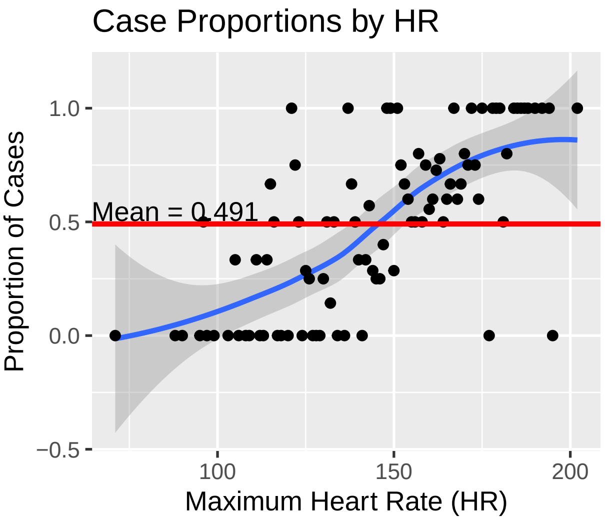


Maximum Heart Rate Factor

The next factor to consider is that of maximum achieved heart rate. From the data set, patients' heart rate results ranged from a minimum of 71 to a maximum of 202. From this range the minimum proportion of patients showed that .330% of the patients had a heart rate level of 71 with the 3.63% of the patients having a maximum heart rate level of 162 with the overall mean of 145 beats per minute. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, the presence of heart disease results varied but generally increased with patients with higher maximum heart rates.

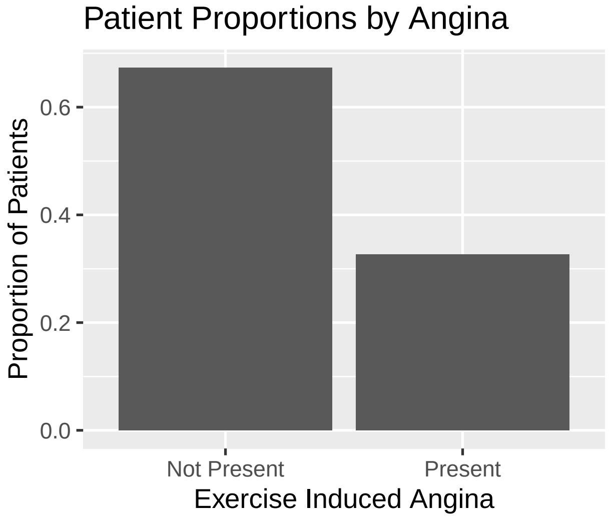


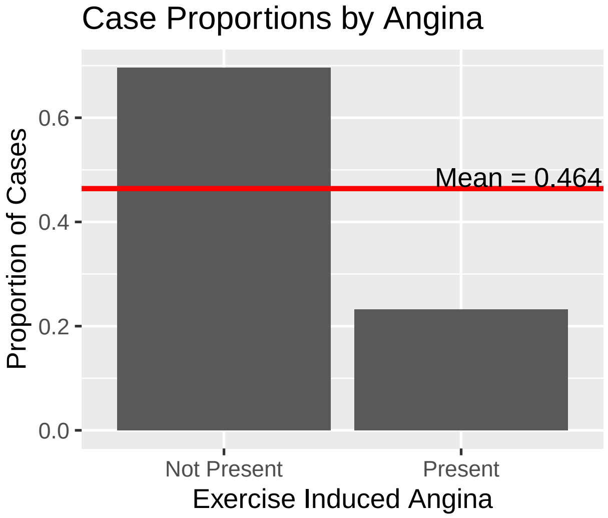


Angina Factor

The next factor to consider is that of exercised induced angina. The data returns this value as either a "0" or "1" with "0" showing the absence of angina indications and "1" showing the presence of angina indications. From the data set, 67.3% of the patients tested did not have indications of exercise induced angina with 32.7% of the patients having indications. The proportion of patients who were tested for heart disease is shown in the first figure below.

Likewise, of the patients who tested positive for heart disease, 69.6% of the time patients without angina tested positive for the disease while 23.2% of the time patients with angina tested positive for the disease. The proportion of patients diagnosed with heart disease is shown in the second figure below.

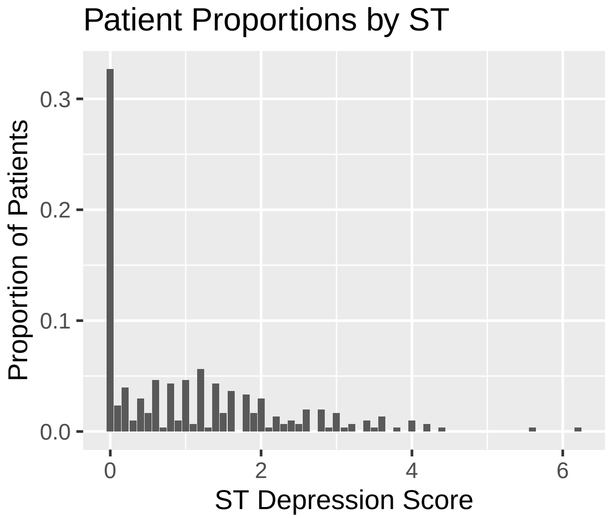


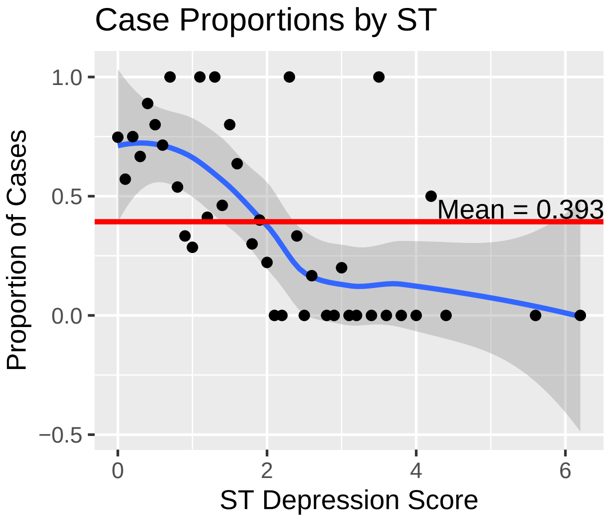


ST depression Factor

The next factor to consider is that of the ST depression on the ECG results that is induced by exercise relative to rest, also known as "oldpeak." In the data, this value ranges from 0 to 6.2 with proportions ranging from a minimum of .330% for the value of 0.7 to a maximum of 32.7% for the value of 0.0. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, the presence of heart disease results varied but probability for the disease generally decreased with increasing oldpeak values. The proportion of patients diagnosed with heart disease is shown in the second figure below.

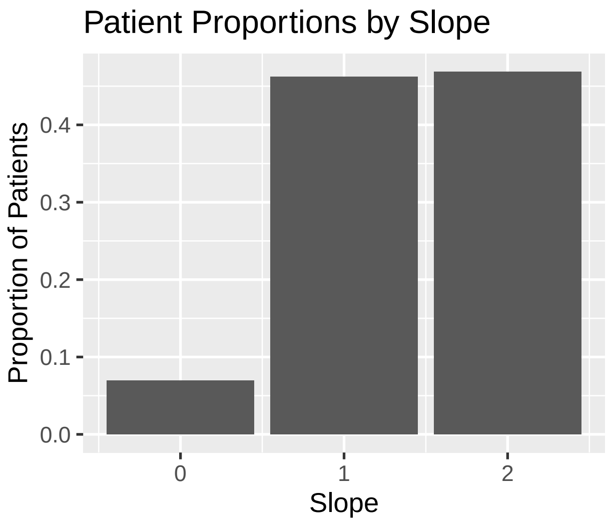


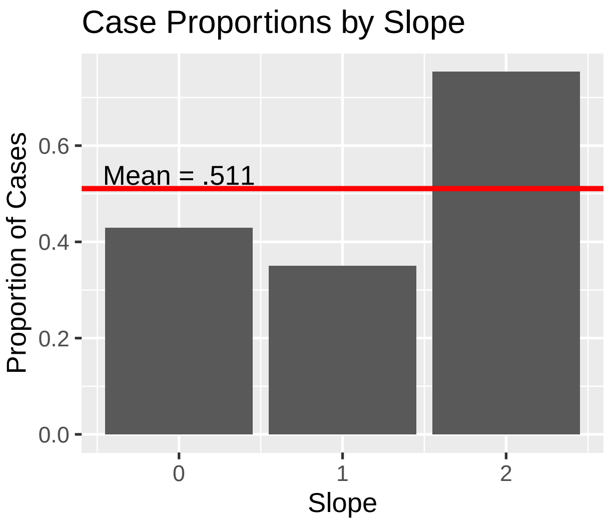


Slope Factor

The next factor to consider is that of the slope of the peak exercise ST segment, taken from the ECG results. These results are categorized by values of “0”, “1”, and “2”, where “0 indicated a decreasing or negative slope,”1" indicates a flat line, and “2” indicates an increasing or positive slope. Of the patients who were tested, 6.93% showed a negative slope, 46.2% had a flat line, and 46.9% had a positive slope. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, patients with a negative slope tested positive for the disease 42.9% of the time, patients with a zero slope or flat line tested positive 35% of the time, and patients with a positive slope tested positive 75.4% of the time. The proportion of patients diagnosed with heart disease is shown in the second figure below.

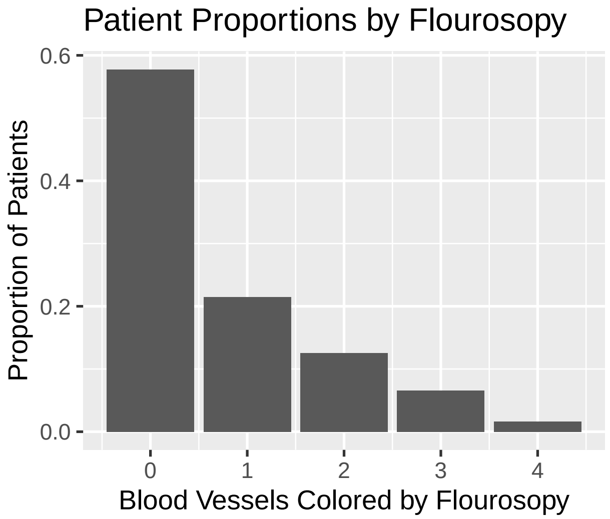


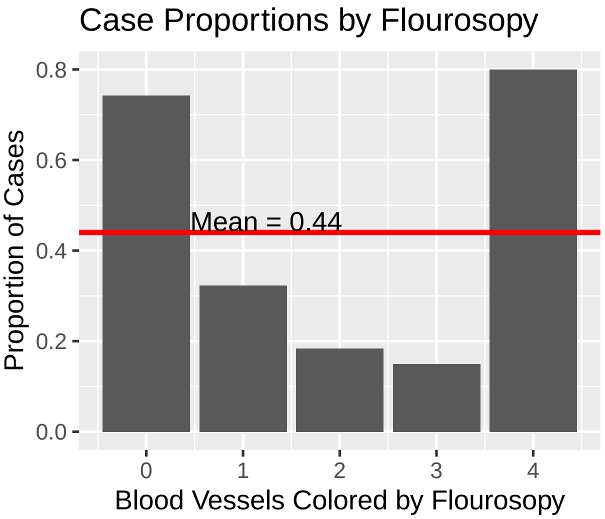


Visible Major Blood Vessels During Flourosopy

The next factor to consider is that of the number of major blood vessels shown during flourosopy. These results range from 0 through 4 blood vessels visible. Of the patients who were tested, 57.8% had 0 vessels visible, 21.5% had 1 vessel visible, 12.5% had 2 vessels visible, 6.6% had 3 vessels visible, and 16.5% had 4 vessels visible. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, those patients with 0 vessels visible had a 74.3% chance of having heart disease. Those patients with 1 vessel visible had a 32.3% chance of having the disease. Those patients with 3 vessels visible had a 15.0% chance of having the disease; and those patients with 4 vessels visible had an 80% chance of having the disease. The proportion of patients diagnosed with heart disease is shown in the second figure below.

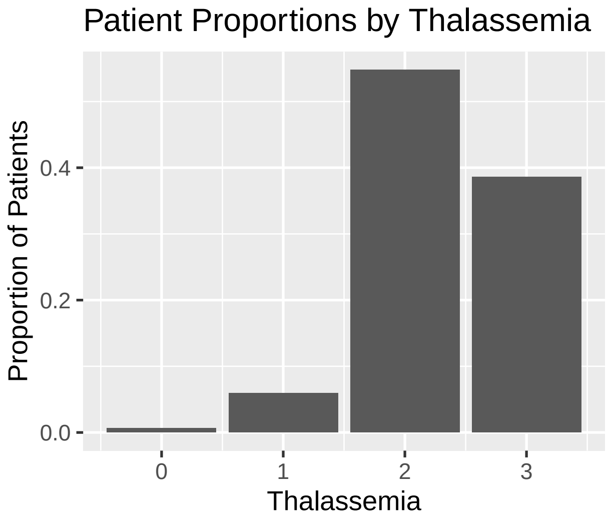


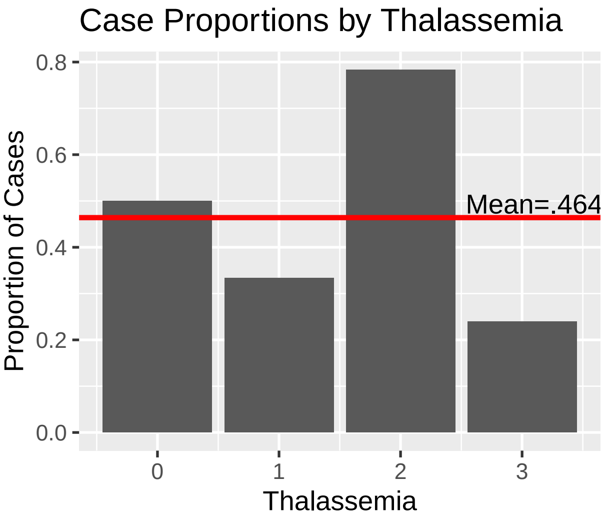


Thal Factor

The final factor to consider is that of a blood flow disorder called "thalassemia." These results range from 0 through 3. Of these values, .660% of the patients returned a value of "0", 5.94%% of the patients had a value of "1", 54.8% of the patients had a value of "2", and 38.6% of the patients had a value of 38.6%. The proportion of patients who were tested for heart disease is shown in the first figure below.

Of the patients who tested positive for heart disease, those patients with a rating of "0" had a 50% chance of having heart disease, patients with a rating of "1" had a 38.3% chance of having the disease, patients with a rating of "2" had a 78.3% chance of having the disease, and patients with ratings of "3" had a 23.9% chance of having the disease. The proportion of patients diagnosed with heart disease is shown in the second figure below.





RESULTS

The 13 factors, as described in the previous section are used to predict the presence of heart disease and regression models compared to determine the best fit. These models include:
1. Linear Regression (LN)
2. Logistic Regression (GLM)

3. Local Regression (Loess)
4. K-Nearest Neighbors (KNN)
5. Random Forest (RF)
6. Ensemble

Linear Regression (LM)

Utilizing linear regression provides promising results with the model returning 81.8% accuracy in diagnosing patients with heart disease and 95.0% accuracy in diagnosing patients without the disease for an overall accuracy of 90.3%

```
##   method accuracy
## 1     lm    0.903
```

The variables, utilized by this function are listed below in order of decreasing importance with the variable "ca" (i.e. the number of visible blood vessels) and "cp" (i.e. chest pain) taking prominent roles in the calculations.

```
##    predictor     lm
## 1         ca 100.00
## 2         cp  96.42
## 3        sex  78.87
## 4      exang  64.10
## 5    oldpeak  57.13
## 6       thal  52.73
## 7    thalach  47.82
## 8    restecg  31.16
## 9   trestbps  28.54
## 10     slope  15.64
## 11      chol   9.48
## 12       fbs   6.86
## 13       age   0.00
```

Logistics Regression (GLM)

Utilizing logistics regression indicates similar results to that of the linear regression model, returning 81.8% accuracy in diagnosing patients with heart disease and a 95.0% accuracy in diagnosing patients without the disease for an overall accuracy of 90.3%.

```
##   method accuracy
## 2    glm    0.903
```

The variables, utilized by this function are listed below in order of decreasing importance with "ca" (i.e. number of visible blood vessels) and "cp" (i.e. chest pain) taking prominent roles in the calculations.

```
##    predictor    glm
## 1         ca 100.00
## 2         cp  96.42
## 3        sex  78.87
## 4      exang  64.10
## 5    oldpeak  57.13
## 6       thal  52.73
## 7    thalach  47.82
## 8    restecg  31.16
## 9   trestbps  28.54
## 10     slope  15.64
## 11      chol   9.48
## 12       fbs   6.86
## 13       age   0.00
```

Loess

Utilizing local regression returns a lower accuracy than that of linear and logistics regression models, returning 81.8% accuracy in diagnosing patients with heart disease and 81.8% accuracy in diagnosing patients without the disease for an overall accuracy of 87.1%.

```
##   method accuracy
## 3  loess    0.871
```

The variables, utilized by this function are listed below in order of decreasing importance with the variable "exang" (i.e. exercise angina) taking the prominent role in the calculations.

```
##    predictor  loess
## 1      exang 100.00
## 2         ca  49.05
## 3         cp  48.85
## 4        sex  47.52
## 5      slope  30.07
## 6       thal  22.73
## 7    restecg  14.85
## 8        age   6.02
## 9   trestbps   2.57
## 10   oldpeak   1.93
## 11      chol   1.52
## 12   thalach   1.15
## 13       fbs   0.00
```

K Nearest Neighbors (KNN)

Utilizing the K-Nearest Neighbors Model with the best tuned "k" of 6 results a lower accuracy than that of linear regression, logistics regression, and local regression models, returning 81.8% accuracy in diagnosing patients with heart disease and 50.0% accuracy in diagnosing patients without the disease for an overall accuracy of 61.3%.

```
##   method accuracy
## 4    knn    0.613
```

The variables, utilized by this function are listed below in order of decreasing importance with the variables "exang" (i.e. exercise angina), "oldpeak" (i.e. ST depression induced by exercise), "cp" (i.e. chest pain), "thalach" (i.e. maximum heart rate achieved), and "ca" (i.e. number of visible blood vessels) taking prominent roles in the calculations.

```
##    predictor   knn
## 1      exang 100.0
## 2    oldpeak  94.2
## 3         cp  90.8
## 4    thalach  90.0
## 5         ca  81.5
## 6      slope  51.7
## 7        thal  50.9
## 8         sex  34.2
## 9         age  29.8
## 10       chol  18.1
## 11    restecg  14.8
## 12   trestbps  14.3
## 13        fbs   0.0
```

Random Forest (RF)

Utilizing the Random Forest Model with the best tuned “mtry” of 2 results in a similar accuracy compared to that of linear regression and logistics regression models, returning 90.9% accuracy in diagnosing patients with heart disease and 90.0% accuracy in diagnosing patients without the disease for an overall accuracy of 90.3%.

```
##   method accuracy
## 5     rf    0.903
```

The variables, utilized by this function are listed below in order of decreasing importance with the variables “cp” (i.e. chest pain), “thalach” (i.e. maximum heart rate achieved), “ca” (i.e. number of visible major blood vessels) and “oldpeak” (i.e. ST depression induced by exercise relative to rest) taking prominent roles in the calculations.

```
##    predictor    rf
## 1         cp 100.0
## 2    oldpeak  97.7
## 3    thalach  95.8
## 4         ca  94.5
## 5        age  68.2
## 6       thal  67.2
## 7      exang  55.4
## 8       chol  52.1
## 9   trestbps  51.2
## 10     slope  29.2
## 11       sex  19.5
## 12   restecg  11.0
## 13       fbs   0.0
```

Ensemble

The final model is the ensemble model, utilizing the results of the most accurate models of linear regression, logistics regression, and random forest. This model returns 81.8% accuracy in diagnosing patients with heart disease and 95.0% accuracy diagnosing patients without the disease disease for an overall accuracy of 90.3%.

```
##     method accuracy
## 6 ensemble    0.903
```

The variables, utilized by this function are listed below in order of decreasing importance with the variables “ca” (i.e. number of visible blood vessels) and “cp” (i.e. chest pain) taking prominent roles in the calculations.

```
##    predictor ensemble
## 1         ca    98.15
## 2         cp    97.62
## 3    oldpeak    70.65
## 4    thalach    63.81
## 5      exang    61.20
## 6        sex    59.09
## 7       thal    57.54
## 8   trestbps    36.11
## 9    restecg    24.44
## 10      chol    23.68
## 11       age    22.74
## 12     slope    20.17
## 13       fbs     4.57
```

CONCLUSION

While a member of the medical field, through the analysis of present data on as few as 303 patient cases using 13 parameters, an overall accuracy of 90.3% is possible, resulting in early detection and saved lives.

Additional studies are warranted as improvents in accuracy may be possible by increasing the number of relevant predictors (e.g. smoking habits, genetic predisposition). Additionally, increasing the number of patient cases will help train the models and increase accuracy.

This form of analysis is not just limited to the medical profession and the use of parameters to make predictions may be utilized in any industry to make data informed decisions and understand the importance of the parameters driving the models and demonstrates the value of data science to an organization.