**HEART DISEASE PREDICTION**

**ARVIND PAWAR**

**AZAR NAJAFLI**

**DHRUVIN KOTHARI**

**NORTHEASTERN UNIVERSITY**

**ALY6015 INTERMEDIATE ANALYTICS**

**INSTRUCTOR NAME- TENGLONG LI**

**DATE- 03/28/2019**



**INTRODUCTION**

Cardiology is one of the most important and yet very difficult field of health care. Heart diseases can prove lethal if not detected in early stages. Every year around 610000 people die from heart diseases in the United States of America, which is 1 in every 4 deaths, according to the data provided by CDC (CDC, 2017). The same report says that the disease is main cause of death for both men and women, however more than half of similar diseases have been observed in men in 2009. Every year more than 700000 Americans have heart attacks in which more than 500000 have it for the first time and more than 200000 have already have attack before. Because of these statistics, every year researches are done by different people including doctors all around the world and data analysts/scientists. The role of data science in those researches is that doc

With all the information we had above, we decided to choose this topic as part of our final project in which our main objective is to construct a predictive model that will help us to determine the disease beforehand any anomalies in patient’s health.

Our team used the dataset called heart.csv received from Kaggle.com which is an open source web site for various real-life datasets. Our dataset contains 14 variables (columns) and 304 observations (rows). We decided to define the variable called “target” as dependent one which means, our model will predict that variable using others (independent). As the variable “target” is binomial which is either yes or no (1 – if the patient has heart disease or 0 – the patient does not have a heart disease), the best model to use should be logistic regression.

RStudio is used for all our analysis and visualizations.

We installed all the required packages.

install.packages("DataExplorer") library(ggplot2

install.packages("Hmisc")

install.packages("data.table")

install.packages("caret")

install.packages("extrafont")

install.packages("ggthemes")

install.packages("caret",

repos = "http://cran.r-project.org",

dependencies = c("Depends", "Imports", "Suggests"))

> library(caret)

> library(data.table)

> library(Hmisc)

> library(DataExplorer)

> library(ggplot2)

> library(carData)

> library(car)

> library(dplyr)

> library(lattice)

> library(tidyr)

> library(caret)

> library(MASS)

> library(broom)

> library(ROCR)

> library(corrplot)

> setwd("C:/Users/Arvind/Desktop/Projects/Heart disease/")

> heart\_data= read.csv("heart.csv")

> head(heart\_data)

ï..age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target

1 63 1 3 145 233 1 0 150 0 2.3 0 0 1 1

2 37 1 2 130 250 0 1 187 0 3.5 0 0 2 1

3 41 0 1 130 204 0 0 172 0 1.4 2 0 2 1

4 56 1 1 120 236 0 1 178 0 0.8 2 0 2 1

5 57 0 0 120 354 0 1 163 1 0.6 2 0 2 1

6 57 1 0 140 192 0 1 148 0 0.4 1 0 1 1

> names(heart\_data)

[1] "ï..age" "sex" "cp" "trestbps" "chol" "fbs" "restecg" "thalach" "exang"

[10] "oldpeak" "slope" "ca" "thal" "target"

> colnames(heart\_data)[colnames(heart\_data)=="ï..age"]<- "age"

Renamed the age column name

**Some of the variables of our data has different categories for example chest pain has type 0,1,2,3, ECG level 0,1,3 and so on. To determine the effect of each variable and better results we decided to perform factorization.**

> factor\_heart\_data<- copy(heart\_data)

> colnames(factor\_heart\_data)[colnames(factor\_heart\_data)=="ï..age"]<- "age"

> head(factor\_heart\_data)

age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target

1 63 1 3 145 233 1 0 150 0 2.3 0 0 1 1

2 37 1 2 130 250 0 1 187 0 3.5 0 0 2 1

3 41 0 1 130 204 0 0 172 0 1.4 2 0 2 1

4 56 1 1 120 236 0 1 178 0 0.8 2 0 2 1

5 57 0 0 120 354 0 1 163 1 0.6 2 0 2 1

6 57 1 0 140 192 0 1 148 0 0.4 1 0 1 1

>col\_names<-c("age","sex","chest\_pain","rest\_bp","chol","fasting\_bloodsugar","rest\_ecg","max\_heartrate",

+ "exercise\_angina","ST\_depression","slope","n\_major\_vasel","thal","target")

> names(factor\_heart\_data)<- col\_names

> names(factor\_heart\_data)

[1] "age" "sex" "chest\_pain" "rest\_bp" "chol"

[6] "fasting\_bloodsugar" "rest\_ecg" "max\_heartrate" "exercise\_angina" "ST\_depression"

[11] "slope" "n\_major\_vasel" "thal" "target"

> factor\_heart\_data$sex <- as.character(heart\_data$sex)

> factor\_heart\_data$sex <- ifelse(heart\_data$sex=="0", 'female', 'male')

> factor\_heart\_data$chest\_pain<-as.factor(heart\_data$cp)

> factor\_heart\_data$fasting\_bloodsugar[heart\_data$fasting\_bloodsugar == 1]= "Diabetic"

> factor\_heart\_data$fasting\_bloodsugar[heart\_data$fasting\_bloodsugar == 0] = "Normal"

> factor\_heart\_data$rest\_ecg[heart\_data$rest\_ecg == 0] = "Normal"

> factor\_heart\_data$rest\_ecg[heart\_data$rest\_ecg == 1] = "Abnormality"

> factor\_heart\_data$rest\_ecg[heart\_data$rest\_ecg == 2] = "Probable or definite"

> factor\_heart\_data$exercise\_angina[heart\_data$exercise\_angina == "1"]= "yes"

> factor\_heart\_data$exercise\_angina[heart\_data$exercise\_angina == "0"] = "no"

> factor\_heart\_data$slope=as.factor(heart\_data$slope)

> factor\_heart\_data$thal=as.factor(heart\_data$thal)

> factor\_heart\_data$target=as.factor(heart\_data$target)

> factor\_heart\_data$sex=as.factor(heart\_data$sex)

> factor\_heart\_data$fasting\_bloodsugar=as.factor(heart\_data$fbs)

> factor\_heart\_data$exercise\_angina=as.factor(heart\_data$exang)

> #View(factor\_heart\_data)

> head(factor\_heart\_data)

age sex chest\_pain rest\_bp chol fasting\_bloodsugar rest\_ecg max\_heartrate exercise\_angina ST\_depression slope

1 63 1 3 145 233 1 0 150 0 2.3 0

2 37 1 2 130 250 0 1 187 0 3.5 0

3 41 0 1 130 204 0 0 172 0 1.4 2

4 56 1 1 120 236 0 1 178 0 0.8 2

5 57 0 0 120 354 0 1 163 1 0.6 2

6 57 1 0 140 192 0 1 148 0 0.4 1

n\_major\_vasel thal target

1 0 1 1

2 0 2 1

3 0 2 1

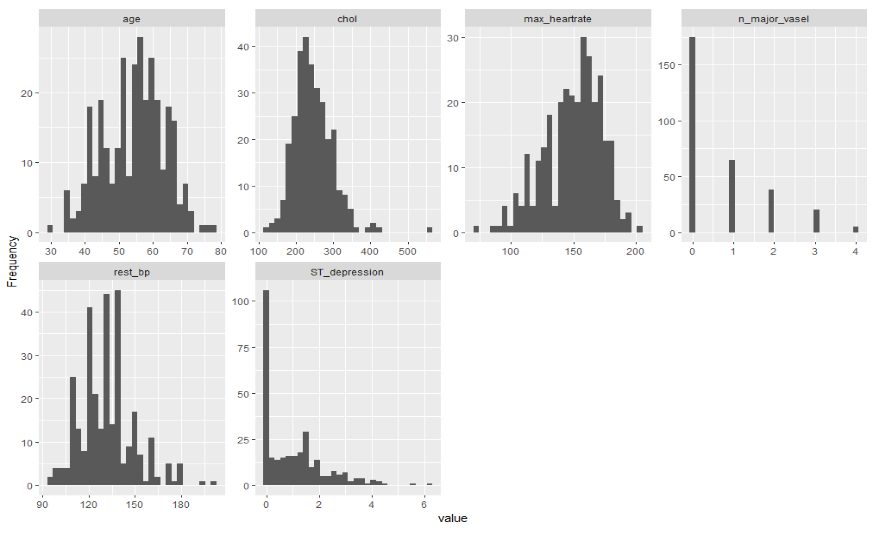
4 0 2 1

5 0 2 1

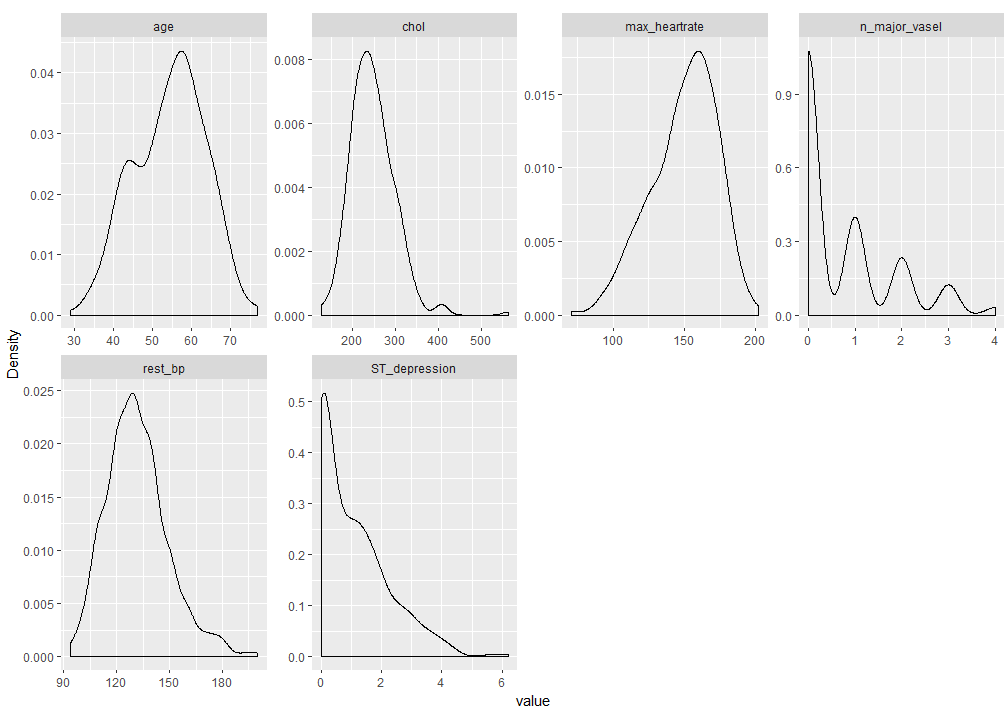
6 0 1 1

Our response variable is “target” that has binary data, 0 and 1. 0 means a person don’t have heart disease and 1 means a person has heart disease.

> plot\_histogram(factor\_heart\_data)



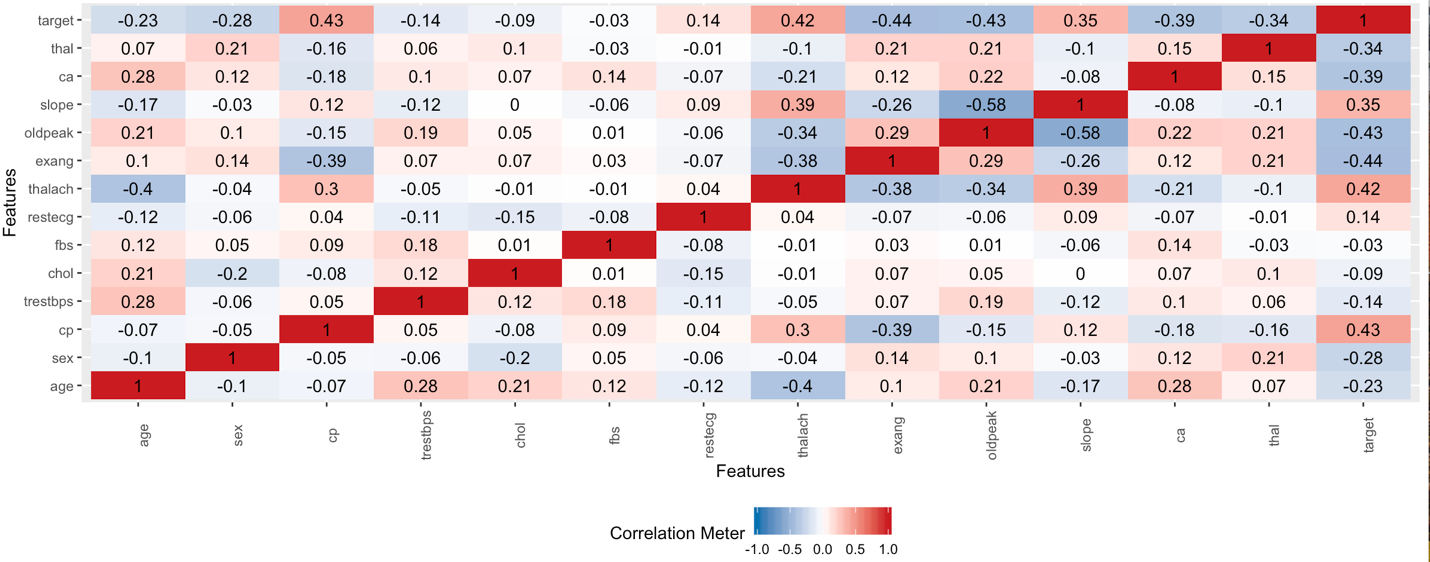
> plot\_density(factor\_heart\_data)



> plot\_correlation(heart\_data)

From the above results we can say that Age, Cholesterol, maximum heart rate and blood pressure follows normal distribution.

We started to plot the correlation matrix to see the relationship between “target” and all other variables visually. As we had, relatively large number of independent variables, we decided to focus on 4 (cholesterol, heart rate, chest pain, and rest ECG test (this is the test advised by doctors if the patient has those three symptoms)) which doctors pay attention more when they try to determine the illness. Below you can see the correlation matrix:



Surprisingly, we can see from the matrix that, although the doctors have more focus on those 4 variables we mentioned above, cholesterol, heart rate, and rest ECG test do not have strong correlation with dependent variable.

**VISUALIZATION**

Age Distribution:

> ggplot(factor\_heart\_data,aes(age, fill=target)) +

+ geom\_histogram(aes(y=..density..),breaks=seq(0, 80, by=1), color="grey17") +

+ geom\_density(alpha=.1, fill="black")+

+ facet\_wrap(~target, ncol=1,scale="fixed") +

+ scale\_fill\_manual(values=c("blue","red"))+

+ xlab("Age") +

+ ylab("Density / Count") +

+ ggtitle("Age Distribution")

Output:

A picture containing screenshot

Description automatically generated

**Conclusion:**

From the above graph we can conclude that, heart disease is uniformly spread out across Age.

Gender count for heart disease:

colnames(heart\_data)[colnames(heart\_data)=="ï..age"]<-"Age"

head(heart\_data)

ggplot(factor\_heart\_data,aes(target, fill=target)) +

geom\_bar(stat = "count") + facet\_wrap(sex~.) + scale\_fill\_manual(values=c("Blue","red")) +

ggtitle("Gender count for Heart Disease")

Output:

A picture containing screenshot

Description automatically generated

**Conclusion:**

From the above graph we can conclude that, more females have Heart Disease as compared to male patients.

2) To construct a predictive model, we focused performing analysis on the following parameters.

* **Chest Pain (Diabetic Patient):** As the chest pain increases there is higher probability that the patient will have a heart problem.

> ggplot(factor\_heart\_data,aes(target,fill=target)) +

+ geom\_bar(stat = "count") + facet\_wrap(chest\_pain~.) + ggtitle("Count of Heart Patients having different types of chest Pains") + theme\_bw() +

+ scale\_fill\_manual(values=c("Blue","red"))+

+ xlab("Target")

Output:

A picture containing screenshot

Description automatically generated

**Conclusion**: More Heart Disease patients have chest pain type 1 or 2

* **Cholesterol Level**: As the cholesterol level increases chances of having stroke, diabetes or high blood pressure also increases which may lead to severe heart problems.

> ggplot(factor\_heart\_data,aes(factor\_heart\_data$chol, fill=target)) +

+ geom\_histogram(aes(y=..density..),breaks=seq(90, 550, by=25), color="grey17") +

+ geom\_density(alpha=.1, fill="black")+

+ facet\_wrap(~target, ncol=1,scale="fixed") +

+ theme\_economist() +

+ scale\_fill\_manual(values=c("blue","red")) +

+ xlab("Cholestoral levels") +

+ ylab("Density / Count") +

+ ggtitle("Cholestrol level test")

Output:

A close up of a map

Description automatically generated

**Conclusion:** More Heart Disease patients seem to have between 200 and 250 mg/dl.

* **Heart Rate Monitoring (Blood Pressure**): For the patients having blood high blood pressure are often considered as higher probability of having heart problems.

> ggplot(factor\_heart\_data,aes(factor\_heart\_data$max\_heartrate, fill=target)) +

+ geom\_histogram(aes(y=..density..),breaks=seq(70, 205, by=10), color="grey17") +

+ geom\_density(alpha=.1, fill="black")+

+ facet\_wrap(~target, ncol=1,scale="fixed") +

+ theme\_economist() +

+ scale\_fill\_manual(values=c("blue","red"))+

+ xlab("Maximum Heart Rate Achieved") +

+ ylab("Density / Count") +

+ ggtitle("Max Heart Rate Histogram")

Output:

A close up of a map

Description automatically generated

**Conclusion:** Heart Disease patients have higher maximum heart rate than healthy patients

* **Rest ECG Tests**: ECG test are carried out to monitor heart conditions.

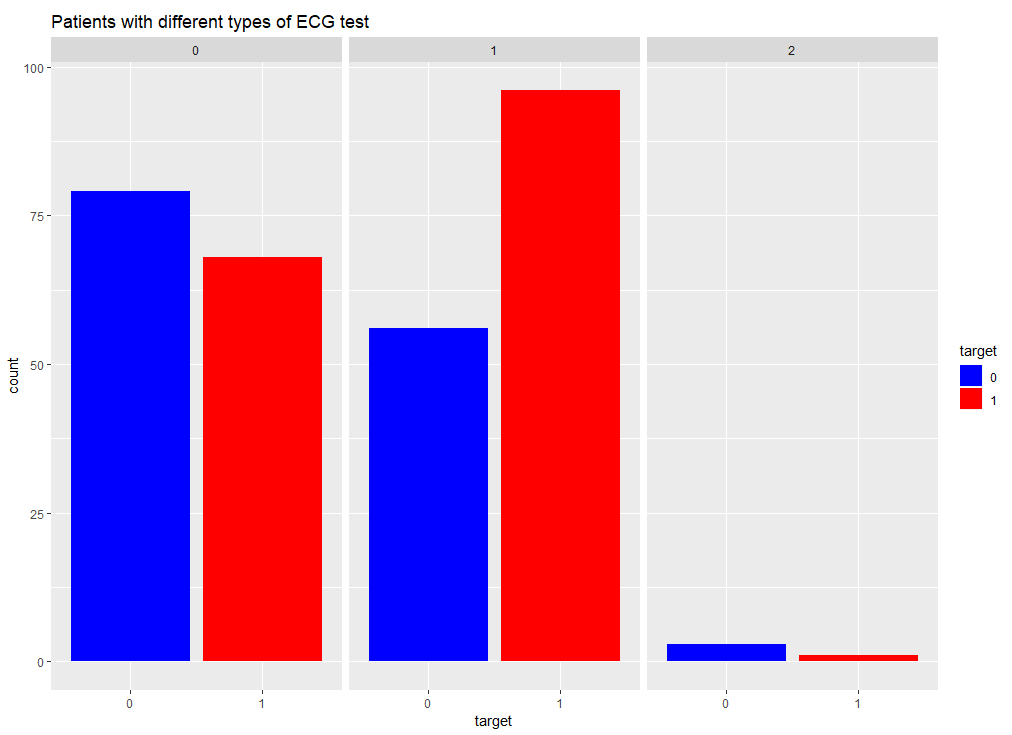
> ggplot(factor\_heart\_data,aes(target,fill=target)) +

+ geom\_bar(stat = "count") + facet\_wrap(rest\_ecg~.)+

+ ggtitle("Patients with different types of ECG test")+

+ scale\_fill\_manual(values=c("Blue","red"))

Output:



**Conclusion:** From the above graph we can observe that Patients with Rest ECG 1 have more probability of having Heart Diseases.

**PREDICTION**

**We are going to examine the relationship between one categorical response variable and various categorical independent variables, so we chose Logistic Regression for prediction.**

**We divided the dataset into training and testing data in 75% and 25% respectively.**

> set.seed(12345)

> train <- floor(0.75\*nrow(factor\_heart\_data))

> train\_ind <-sample(seq\_len(nrow(factor\_heart\_data)),size = train)

> trainset <- factor\_heart\_data[train\_ind, ]

> testset <- factor\_heart\_data[-train\_ind, ]

> dim(trainset)

[1] 227 14

> dim(testset)

[1] 76 14

**Above are the dimensions of our training and testing dataset.**

**First, we applied the logistic regression model on all the independent variables using glm() function.**

> logit<-glm(target~., data=trainset, family = binomial)

> summary(logit)

Call:

glm(formula = target ~ ., family = binomial, data = trainset)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.6911 -0.3425 0.1598 0.6103 2.6969

Coefficients:

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

(Intercept) 1.258e+01 1.455e+03 0.009 0.993106

age 6.877e-03 2.703e-02 0.254 0.799172

sex1 -1.677e+00 5.846e-01 -2.868 0.004133 \*\*

chest\_pain1 8.921e-01 5.946e-01 1.500 0.133532

chest\_pain2 2.034e+00 5.486e-01 3.708 0.000209 \*\*\*

chest\_pain3 1.947e+00 7.560e-01 2.575 0.010016 \*

rest\_bp -1.469e-02 1.315e-02 -1.117 0.263945

chol -5.670e-03 4.236e-03 -1.339 0.180657

fasting\_bloodsugar1 -2.108e-01 6.912e-01 -0.305 0.760370

rest\_ecg1 4.284e-01 4.259e-01 1.006 0.314494

rest\_ecg2 -4.029e-01 2.361e+00 -0.171 0.864495

max\_heartrate 2.541e-02 1.220e-02 2.083 0.037226 \*

exercise\_angina1 -6.633e-01 4.795e-01 -1.383 0.166528

ST\_depression -4.378e-01 2.672e-01 -1.638 0.101338

slope1 -1.322e-01 8.605e-01 -0.154 0.877894

slope2 7.741e-01 9.387e-01 0.825 0.409615

n\_major\_vasel -8.687e-01 2.348e-01 -3.700 0.000216 \*\*\*

thal1 -1.124e+01 1.455e+03 -0.008 0.993840

thal2 -1.176e+01 1.455e+03 -0.008 0.993551

thal3 -1.302e+01 1.455e+03 -0.009 0.992861

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 312.74 on 226 degrees of freedom

Residual deviance: 160.10 on 207 degrees of freedom

AIC: 200.1

Number of Fisher Scoring iterations: 14

**From above output, we observed that sex, chest pain, number of major vessels observed, max heart rate has effect on heart disease.**

**Then, we decided to create another dataset that has significant variables only.**

> cor\_data<-trainset[,c(2,3,9,10,12,14)]

> summary(cor\_data)

sex chest\_pain exercise\_angina ST\_depression n\_major\_vasel target

0: 96 0:143 0:204 Min. :0.00 Min. :0.0000 0:138

1:207 1: 50 1: 99 1st Qu.:0.00 1st Qu.:0.0000 1:165

2: 87 Median :0.80 Median :0.0000

3: 23 Mean :1.04 Mean :0.7294

3rd Qu.:1.60 3rd Qu.:1.0000

Max. :6.20 Max. :4.0000

**However, before applying the model let’s see the variance inflation factor to check whether there is multicollinearity or not.**

> #variance inflation factor

> vif(glm(target ~ ., data=cor\_data, family="binomial"))

GVIF Df GVIF^(1/(2\*Df))

sex 1.056310 1 1.027770

chest\_pain 1.311732 3 1.046263

exercise\_angina 1.093148 1 1.045537

ST\_depression 1.152195 1 1.073403

n\_major\_vasel 1.045553 1 1.022523

**VIF seems to be low, and there is no multicollinearity in data. So, we can move forward to apply the logit model on significant variables data.**

> logit1<-glm(target~., data=cor\_data, family=binomial)

> summary(logit1)

Call:

glm(formula = target ~ ., family = binomial, data = cor\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.3277 -0.5202 0.2011 0.5714 2.5038

Coefficients:

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

(Intercept) 1.9614 0.4348 4.511 6.44e-06 \*\*\*

sex1 -1.4117 0.3894 -3.625 0.000289 \*\*\*

chest\_pain1 1.3498 0.4868 2.773 0.005560 \*\*

chest\_pain2 2.0905 0.4192 4.987 6.12e-07 \*\*\*

chest\_pain3 2.0161 0.6086 3 .313 0.000924 \*\*\*

exercise\_angina1 -1.2217 0.3721 -3.283 0.001028 \*\*

ST\_depression -0.8060 0.1810 -4.454 8.42e-06 \*\*\*

n\_major\_vasel -0.7635 0.1662 -4.595 4.34e-06 \*\*\*

---

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: 238.32 on 295 degrees of freedom

AIC: 254.32

Number of Fisher Scoring iterations: 5

**From the above results we can see that all the variables are significant, which is good for prediction. Let’s visualize the correlation.**

> cor\_data$chest\_pain<-as.factor(cor\_data$chest\_pain)

> factor\_heart\_data$chest\_pain<-as.factor(heart\_data$cp)

> logit1.df<-tidy(logit1)

> library(ggthemes)

> library(extrafont)

> logit1.df %>%

+ mutate(term=reorder(term,estimate)) %>%

+ ggplot(aes(term,estimate, fill=estimate))+

+ geom\_bar(stat="identity")+

+ ggtitle("Effect of variables resulting Heart Disease")+

+ scale\_fill\_gradient(low="blue", high="red")+

+ theme\_economist()+

+ geom\_hline(yintercept=0)+

+ coord\_flip()

Output:



**From the above output, we observed that chest pain type 2 has the most impact on heart disease. Person having any type of chest pain is likely to have heart disease. Increase in number of major vessels and exercise angina reduces the chances of heart disease.**

**Our varianvce inflation factor result was good but still to avoid the problem of overfitting we performed cross validation using trainControl() function.**

> fitControl <- trainControl(method = "repeatedcv",

+ number = 10,

+ repeats = 10,

+ classProbs = TRUE,

+ summaryFunction = twoClassSummary)

> trainset$target<-make.names(trainset$target)

> set.seed(142)

> trainset$target<-as.factor(trainset$target)

> generalized\_model <- caret::train(target ~ .,

+ data = trainset ,

+ method = "glm",

+ trControl = fitControl,

+ metric="ROC")

There were 20 warnings (use warnings() to see them)

> generalized\_model

Generalized Linear Model

227 samples

13 predictor

2 classes: 'X0', 'X1'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 10 times)

Summary of sample sizes: 205, 204, 205, 205, 205, 203, ...

Resampling results:

ROC Sens Spec

0.8685239 0.7367273 0.8478205

> pred <- predict(generalized\_model, testset,type='raw')

> summary(pred)

X0 X1

33 43

**Here we passed the test dataset for prediction using predict() function.**

> pred

[1] X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X1 X0 X0 X1 X1 X1 X1 X1 X1 X1 X0 X1

[37] X1 X1 X1 X1 X1 X0 X0 X0 X0 X0 X0 X1 X0 X0 X0 X0 X0 X0 X0 X0 X0 X0 X0 X0 X0 X1 X0 X0 X0 X0 X0 X0 X0 X1 X1 X0

[73] X0 X1 X0 X0

Levels: X0 X1

> testset$target

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0

[55] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Levels: 0 1

We can compare the predicted values and observed values.

Here, we created a confusion matrix

> Confusion\_Matrix<-table(testset$target, pred)

> Confusion\_Matrix

pred

X0 X1

0 30 5

1. 3 38

In the end we determined the accuracy of our model.

> accuracy<-sum(diag(Confusion\_Matrix))/sum(Confusion\_Matrix)

> accuracy

[1] 0.8947368

We got 89.47% accuracy.

**CONCLUSION**

We analyzed various parameters that risk in causing heart disease, and we created a predictive model with 89.47% accuracy that can help us determine beforehand whether preventive measure needs to be taken so that we can avoid heart disease and heart strokes. We examined the relationship between one categorical response variable and various categorical independent variables, so we chose Logistic Regression for prediction. To get better results, we can keep on increasing various parameters and increase the size of the data.

TEAM CONTRIBUTION:

Everyone contributed equally as we helped each other when anyone of us was stuck on certain point and that was the key highlight of our project. Our team-work led to a successful completion of our project.

REFERENCES

* Heart Disease UCI. (2019). Kaggle.com. Retrieved 29 March 2019, from <https://www.kaggle.com/ronitf/heart-disease-uci/kernels>
* ggplot2 histogram : Easy histogram graph with ggplot2 R package - Easy Guides - Wiki - STHDA. (2019). Sthda.com. Retrieved 29 March 2019, from <http://www.sthda.com/english/wiki/ggplot2-histogram-easy-histogram-graph-with-ggplot2-r-package>
* A Short Introduction to the caret Package. (2019). Cran.r-project.org. Retrieved 29 March 2019, from <https://cran.r-project.org/web/packages/caret/vignettes/caret.html>
* Cui, B. (2019). Introduction to DataExplorer. Cran.r-project.org. Retrieved 29 March 2019, from <https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html>
* Machine Learning Part 3 : Logistic Regression. (2017). Towards Data Science. Retrieved 29 March 2019, from <https://towardsdatascience.com/machine-learning-part-3-logistics-regression-9d890928680f>
* Evaluating Logistic Regression Models. (2015). R-bloggers. Retrieved 29 March 2019, from <https://www.r-bloggers.com/evaluating-logistic-regression-models/>
* package, F., docs, R., & browser, R. (2019). theme\_economist: ggplot color theme based on the Economist in ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. Rdrr.io. Retrieved 29 March 2019, from <https://rdrr.io/cran/ggthemes/man/theme_economist.html>