CAPSTONE PROJECT REPORT CORONARY HEART DISEASE RISK PREDICTION

SANJU HYACINTH C

CAPSTONE PROJECT NOTES ON CORONARY HEART DISEASE RISK STUDY

1. Introduction:

This study on the coronary heart disease risk is an ongoing one, called the ***Framingham Heart Study***. This was conducted mainly after the demise of *President Franklin D. Roosevelt* who died of a heart disease and stroke in ***1945***. Owing to the President's premature demise, the fact that most of the Americans were convinced in the 1940s that heart diseases are unavoidable and one of three deaths happen to be of heart diseases, and to help themselves to understand the risk factors that may lead to CHD; the Framingham Heart study was conducted on the participants from ***Framingham, Massachusetts*** starting from its first participant in 1948.

I. Problem Statement:

The project requires us to understand the given *demographic and behavioral variables* of approximately **4200** patients, analyze the details we are provided with (the personal and medical such as glucose level details of a patient) to meaningfully assess their ***health status*** and ***predict*** if they are close to having a risk of heart disease in the near future.

II. Need for the Study:

The ultimate goal or objective of the study would be to predict, with the given information, if the patient ***will have a risk of coronary heart disease risk in the next 10 years.*** Our need for the study expands to also

- Identify the crucial and deciding parameters (or variables) that increase the risk of patients contracting the diease
- Find meaningful, actionable insights from the analysis
- Provide valuable recommendation for the business and individuals

III. Scope:

The scope of the study includes most of the objectives, that is,

- The analysis of details given by the various demographic, medical and behavioral parameters of the approximately 4200 patients in the Massachusetts region
- Explanation of the important variables that has been identified from the data, and their part in causing or controlling the risk
- Actionable insights for the business growth.

2. Exploratory Data Analysis:

The visual inspection of the dataset on coronary heart risk study gives the below details.

- The coronary heart risk study data comprises of **4240** entries (rows) spread across **16** variables (columns).
- The details include that of the demographic and medical details and figures of the patient (independent variables), and also the target variable (dependent variable) which has to be predicted using the various machine learning techniques.
- Upon visually inspecting the dataset, we find that there are quite an amount of **missing values** (NAs)
- The data consists of 8 nominal fields (including the target variable) and 8 continuous variables.

I. Uni-variate Analysis:

The uni-variate analysis of the dataset contains the summary of all the variables, and the individual analysis of each variable. We have visualizations for a few important variables alone The summaries of continuous variables and categorical variables is given below (Fig 1.a and Fig 1.b)

Fig 1.a - SUMMARY OF CONTINUOUS VARIABLES IN THE DATASET

Variable	Minimum	Maximum	Mean	Median	Outliers	NA
Age	32	70	49.58	49	-	-
Cigsperday	0	70	9	0	yes	29
TotChol	107	696	236.7	234	yes	50
SysBP	83.5	295	132.4	128	yes	-
DiaBP	48	142.5	82.9	82	yes	-
BMI	15.54	56.8	25.8	25.4	yes	19
Heartrate	44	143	75.88	75	yes	1
Glucose	40	394	81.96	78	yes	388

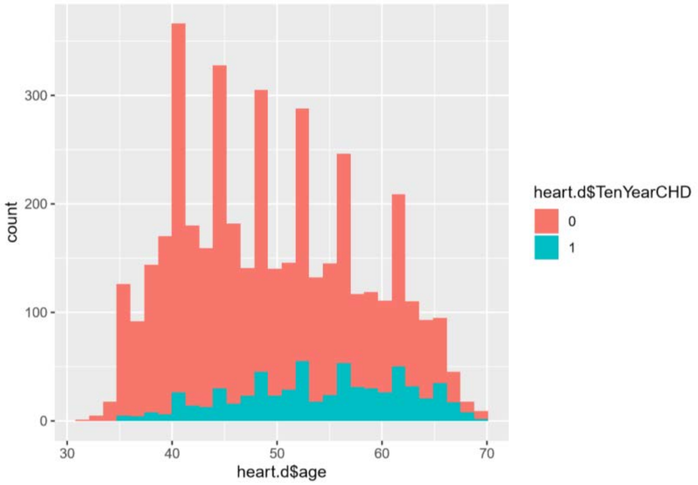
Fig 1.b - SUMMARY OF CATEGORICAL VARIABLES

Variable	No	Yes	1	2	3	4	NA
Male	2420	1820	-	-	-	-	-
CurSmoke	2145	2095	-	-	-	-	-
BPMeds	4063	124	-	-	-	-	53
PrevStroke	4215	25	-	-	-	-	-
PrevHype	2923	1317	-	-	-	-	-
Diabetes	4131	109	-	-	-	-	-
TenYearCHD	3596	644	-	-	-	-	-
# Education	-	-	1720	1253	689	473	105

# contains levels

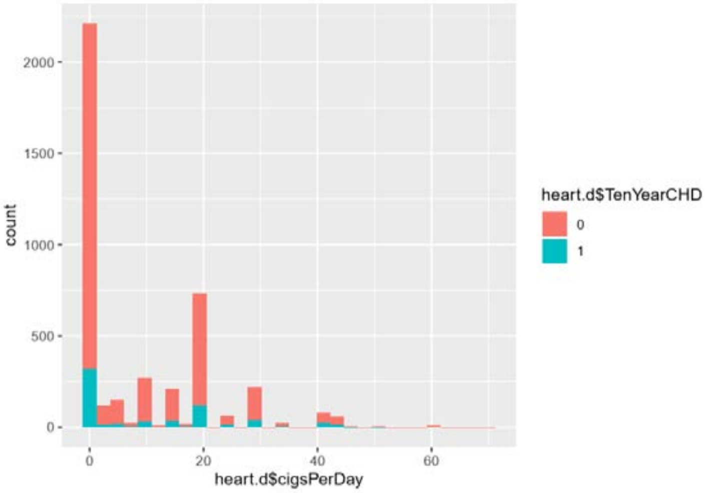
Histogram Visualizations (Fig 2.a – Fig 2.c)

Fig 2.a: Age Histogram:



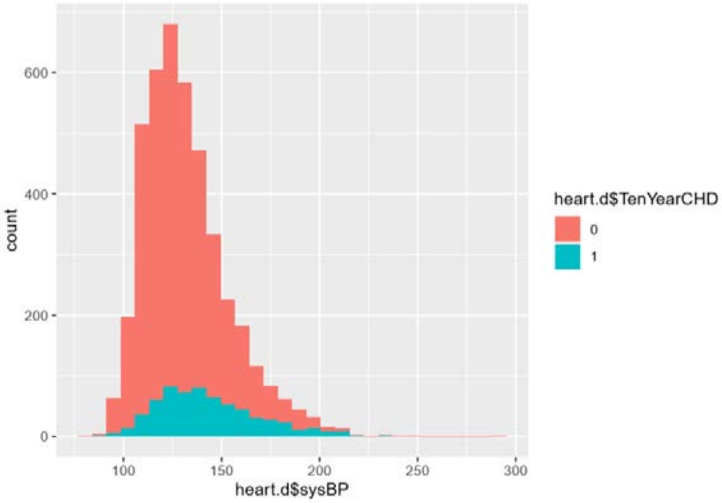
Most of the patients are between the age group 35-65. And though most of them do not suffer a risk of heart disease, quite a lot of the risk bearers are between the range of 48-65 years of age.

Fig 2.b: Cigarettes per Day Histogram:



A large peak at 0 explaining most are non-smokers. The fact is we also find people who have a 0 cigarette count bearing a risk. We see a small area covered at the range of 60 cigs where all are classified to not bear a risk in the next 10 years.

Fig 2.c: Systolic BP Histogram:



It is a left skewed histogram with most of the systolic pressures between 100 and 150. The highest peak shows at 125. The risk is between 120-140

II. Bi-variate Analysis:

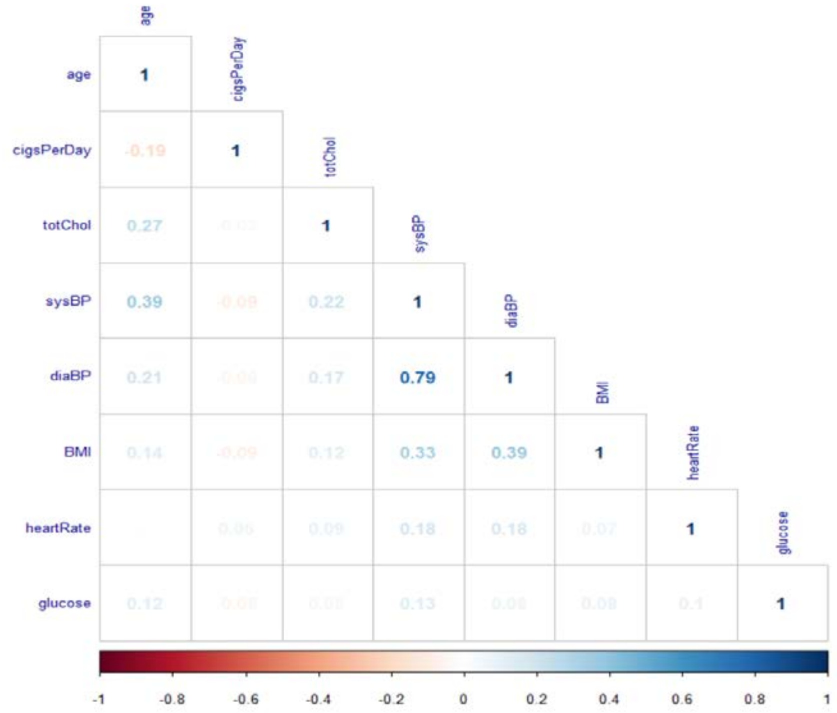
The bi-variate analysis of variables includes ***correlation plots, relationship graphs (scatter plots),*** *and* ***boxplots*** as well. They are used to describe how a change in one factor can influence the other.

a. CORRELATION:

Looking at the correlation between the numeric variables (Fig 3), we find that

- The systolic and the diastolic blood pressures show a high amount of correlation (79%).
- Also the other independent variables that correlate between themselves are Systolic BP and Age (39%) and also Diastolic BP and BMI (39%).
- The others could be BMI and Systolic BP (33%), age and cholesterol which is about 27% and, age with cigarettes per day showing an obvious negative correlation of 19%, and the heart rate the blood pressures (SysBP and DiaBP) show some correlation of 18%

Fig 3: Correlation between numeric variables

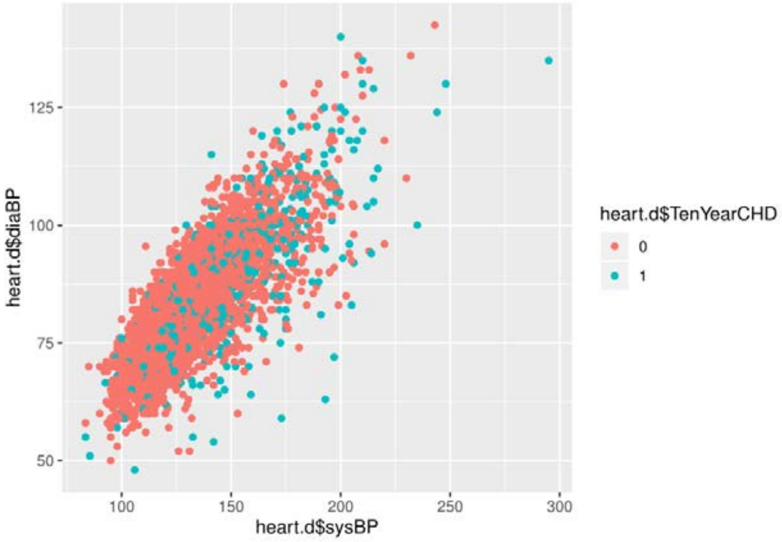


b. RELATIONSHIP BETWEEN VARIABLES (w.r.t Dependent variable)

The analysis of different variables with respect to the dependent variable is given below. We are suing some scatter plots and box plots to find the relationship between many independent variables, with respect to the dependent variable (TenYearCHD). The charts (Fig 4.a – Fig 4.c) are given below with explanation.

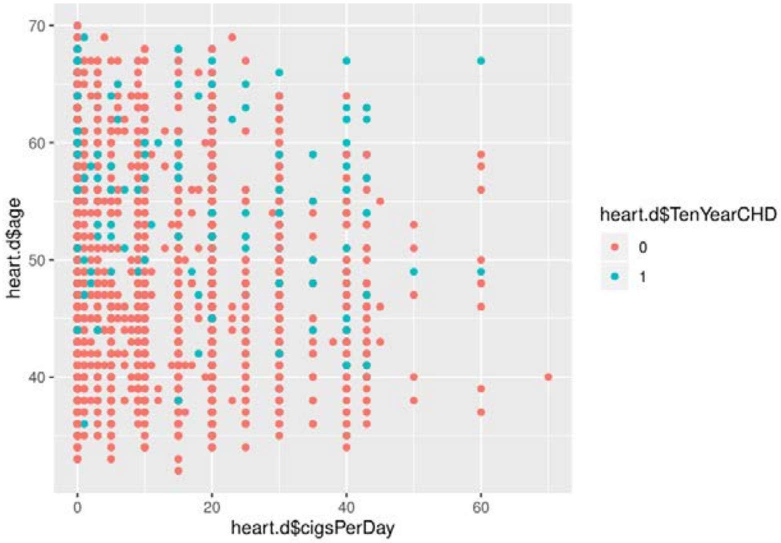
- ***Relation between Systolic and Diastolic BP:*** In this plot, we see a direct relation between the BPs. Patients with a steady and balanced range of both the pressures are at the risk free side and the patients who deviate largely are at the risk side, for example, a patient with 175/65 (systole/diastole) is said to have elevated blood pressure and likewise while a person who has them in control (120/80) is less likely to face the risk.

Fig 4.a – Systolic BP Vs Diastolic BP



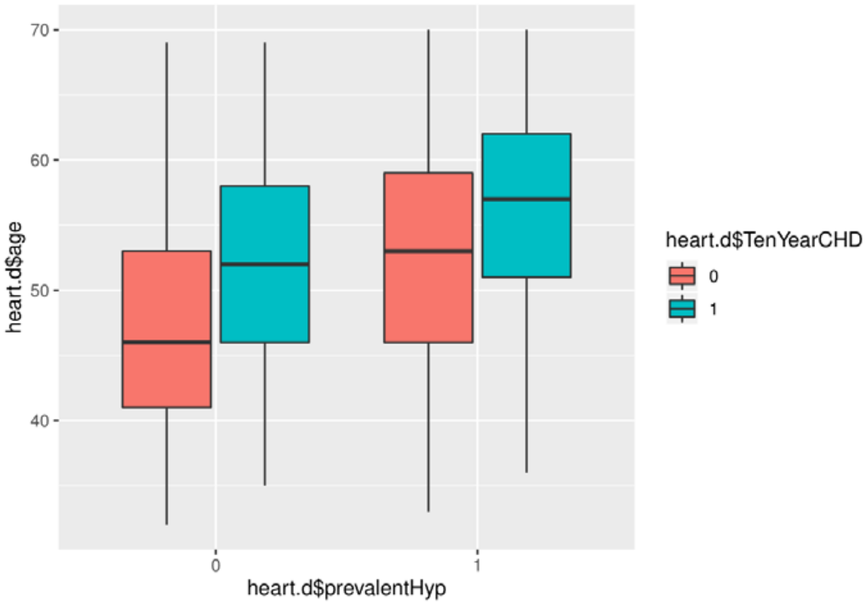
- ***Relation between Cigar count and Age:*** From this graph we see that, though cigar count is an important criteria, we still find that aged people are more susceptible to bearing the risk. We also can confirm this by stating an example from the graph. Two patients smoking 60 cigarettes per day are classified risky when compared to 10 other patients who are classified non risky. So age and cigar count are hand in hand

Fig 4.b – Cigarettes Per Day Vs Age



- ***Relation between Age and Hypertension:*** We find that in both the cases, the patients over the age of 50 have high chances of contracting the heart disease. And we do see an increasing pattern. Aged patients have a higher risk with hypertension, as it increases with age.

Fig 4.c – Prevalent Hypertension Vs Age



3. DATA CLEANING AND PRE-PROCESSING:

Data pre-processing is the most important and perhaps, the longest process that requires more time. Only a clean, well groomed data can be used to build predictive analytical models. And needless to mention, consists of many methods or treatments involved in it. Let us go through the various steps that make the data clean and usable for high prediction reliability.

A. OUTLIER TREATMENT:

By using the **box plots** during the individual (uni-variate analysis) variable analysis, we have found that all our numeric variables have outliers present in them except the ***age*** variable. Hence we will apply a ***function*** to bring them under control limits.

This function helps to locate the data points below and above the ***interquartile range*** of the boxplot and ***replace*** the lower and upper inter quartile values with the ***5th and the 95th*** percentile values. This method, although may alter the provided values, might not greatly vary the outlier values as they are brought to the corresponding ***max or min*** values. The below two box plots (Fig 5.a and Fig 5.b) is the example of the Cholesterol distribution before and after doing the outlier treatment.

Fig 5.a – Before Outlier Treatment

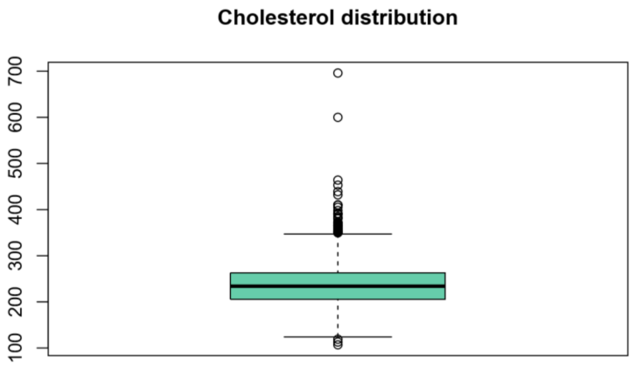
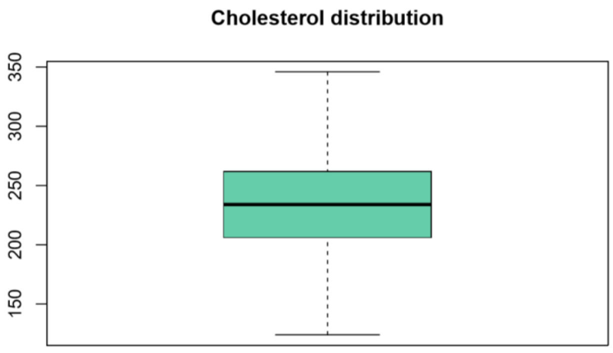


Fig 5.b – After Outlier Treatment



B. MISSING VALUE TREATMENT:

From the summaries (uni-variate analysis) of the individual variables, we have the exact number of missing values under each variable. Without treating the missing values in a data, it is nearly impossible to carry out the analysis. Hence missing value treatment is an important step towards deriving the best model.
Our data has missing values as given below in the tabular column (Fig 6)

Fig 6 - MISSING VALUES COUNT

Variables	NA (count)
Glucose	388
Education	105
BPMeds	53
TotChol	50
Cigsperday	29
BMI	19
Heartrate	1

We have found that about 13.73% of the data is missing in the dataset. We could just remove the missing values (NA) or treat them by ***imputation.*** Imputation could be by mean, median or other prediction methods. But since ours is a healthcare dataset, comprising of data for risk prediction (such as a coronary heart risk) on patients, we rather not ***manipulate or pollute*** the data with compromised values. This will have an effect on our models, which may probably provide a biased result.

Hence we are removing all the missing values (NA) from the dataset and consider only the complete cases for the model building process.

C. VARIABLE TRANSFORMATION:

Variable transformations in R are done to prevent any misleading results when it comes to model building, in other terms, to build an effective model. This step includes Outlier treatment, Missing value treatment, Removing unwanted variables, removing multicollinearity amongst variables, etc. We have thus far, treated outliers and missing values in the data.

We assume that the data does not need scaling as most of them are on the same scale. Scaling may affect the authenticity of the data provided as the cholesterol values are supposed to be high and scaling may affect the same as well as some other variables like age which may get tampered.

D. REMOVAL/ ADDITION OF VARIABLES:

On inspecting the dataset, we have identified the variables that would be of use and some that might not be that insightful. In that case we have a few variables of conflict, but need not be removed from the dataset. One such variable is from the demographic information of the patient, that is, ***education (has 4 levels)*** We have found the variable to be of not use because of the anonymity and uncertainty of the variable's use in the data. The other variable, being ***current smoker,*** which conflicts with the ***cigarettes per day*** variable. Both these variables somewhat explain the same information. We could not have both these variables in the same model. As per the clarity in information, we have cigarettes per day giving enough details to proceed and exclude current smoker from our models.

We consider the variables education and current smoker to be of least importance in building models on the dataset.

As for adding any new variable, we so far consider not creating any as model would already have sufficient parameters to work with. We do not want to create a new variable, that may result in multicollinearity and have an effect on the model.

4. MODEL BUILDING:

Our next step is the most important one in the study. In order to bring to effect, our analysis and findings so far, we need to build models on our data to predict the efficiency of the same so as to provide a usable model for future business needs.

In order to do so, we first need to ***split our dataset into train and test dataset.*** This is done, so that we can develop a model on the train dataset, which helps the algorithm to familiarize with the situations of how each variable influences in the decision making of whether something is there or not. Further we have the test dataset, as the name suggests, the dataset on which the developed model is ***tested or validated on.*** This helps us to ***confirm if our model is working at its best or not.*** As this is a healthcare data, we need to have great predictions for the same to be reliable and useful.
We do our split on the main dataset using the ***caTools*** package in R, in the ration of ***3:1*** for the train and test data respectively. This means that the algorithm gets

trained on the randomly selected 75% of the data and will be validated on the remaining 25% of the data allotted for testing.

Now that we have our data split, we shall now go to the models used for prediction, why they are used and the interpretation of the same.

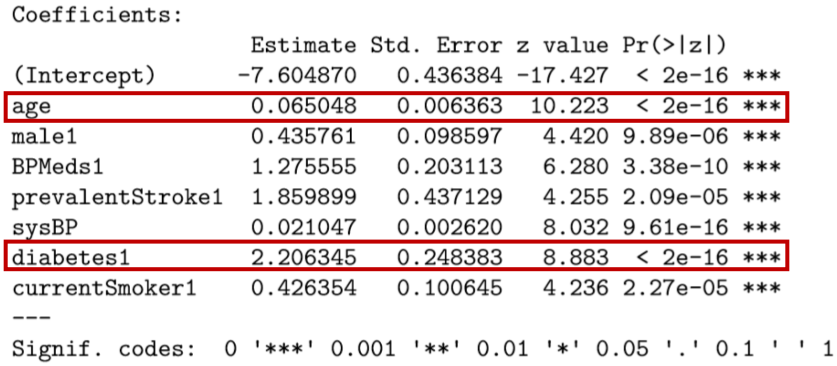
4.1 LOGISTIC REGRESSION:

Logistic regression is one of the most preferred algorithms when the target is binomial (0 or 1) It is therefore an ideal technique for this dataset. We have conducted the model on the dataset that does not have imputed values. But we are working on the dataset that has been treated for **imbalance (SMOTE).**

SMOTE helps to balance the data points by synthetically generating some values in the data, shuffling it between the majority and the minority class. Therefore it helps to balance out the majority (minimizing the entries) and minority class (maximizing the entries) This makes the dataset more balanced for training the model. ***This is done only on the train dataset that is used for the model to get developed, but not on the test dataset as this will affect the results of the model.***

From our model 1, that was conducted on all the variables, we have had a good result, for the 86% of data available. The model identified the important variables as ***age, male, BP meds, prevStroke, diabetes, sysBP.*** From model 2 conducted on our important variables identified, we have got a better result with all the used variables as highly significant. The image of the variables is shown below for reference.

Fig 7 – Significant variables list.



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.604870   0.436384 -17.427  < 2e-16 ***
age               0.065048   0.006363  10.223  < 2e-16 ***
male1             0.435761   0.098597   4.420 9.89e-06 ***
BPMeds1           1.275555   0.203113   6.280 3.38e-10 ***
prevalentStroke1  1.859899   0.437129   4.255 2.09e-05 ***
sysBP             0.021047   0.002620   8.032 9.61e-16 ***
diabetes1         2.206345   0.248383   8.883  < 2e-16 ***
currentSmoker1    0.426354   0.100645   4.236 2.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The highlighted variables are found to be the most significant values from the model as per the p value identified.

4.2 NAÏVE BAYES:

Our next model conducted on the smote dataset is the ***Naïve bayes model.*** Naive bayes is a common technique used in the field of medical science and is especially used for cancer detection. This is a probabilistic approach generally preferred for large datasets, but we are using the model to identify if the model can well detect or identify the people who will have a heart disease in 10 years tenure. And as it is a model that works well on ***categorical variables,*** we are using the model on non-numeric data only

One reason for using this technique comes from the principle itself. It is that, when we have narrowed down our important factors to an extent, the probability of the something happening provided the knowledge of a factor makes it better for prediction. That is the importance of naïve bayes. As well as it has predicted the non risk bearers, we have very little accuracy with the risk bearers (~32%) which would not give reliable results for a business model.

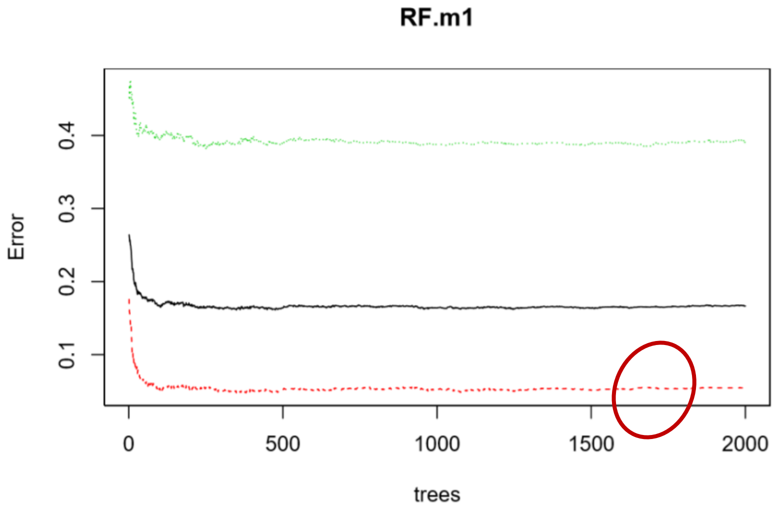
4.3 RANDOM FOREST:

Our next model is one of the most used models, called as ***Random Forest.*** As we know that RF is an ensemble technique that builds multiple models from subsets of original data (bootstrap aggregating) to give us a more robust unbiased model. The left out data points from each of the subsets is called ***OOB dataset*** (Out of Bag) on which the next model will be created and so on. This is used simply for the fact that ***the result is given on the basis of majority predictions*** given by an ensemble of models. This makes the model a robust and working model. Until a desired model is chosen out of all the models based on performance.

First we assign the ***mtry*** value, which is usually the square root of the number of columns available to us. This is important as, this assigns the number of variables available for splitting at each tree node. Our mtry value is 4. After this the random forest model is built with the variables with respect to the dependent variable. The parameters such as the number of trees (***ntree***) is given as per requirement, after which we print the model to observe the OOB error rate, the confusion matrix, the class error, etc. This is sort of our base model.

Upon plotting the same model, we get a graph that helps us to ***prune the tree to a good number of trees that is sufficient.*** This step is essential as it helps us to bring a better model than a random model, while assigning random numbers. In our model, we have found the correct number of trees after which there is not much variability captured by the model. This value is then replaced in our formula for a better model. The graph is given below.

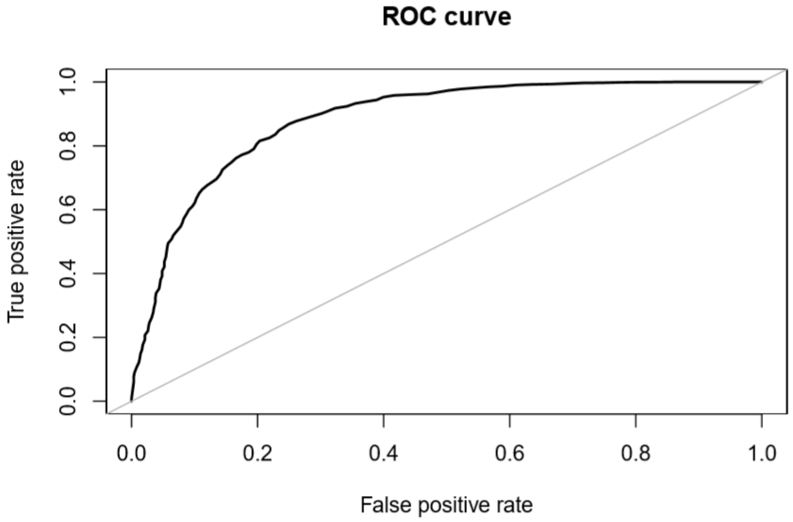
Fig 8 – Random Forest plot with the tree number



The circled portion is where the model needs to be tuned to. The number of trees to which we have tuned in our model us 1700. This is the step where we find the most significant variables. From the ***importance*** command, we have found the important variables to be ***age, cigsperday*** and ***sysBP.*** The mean decrease in accuracy is the factor which gives in number, how much accuracy is lost when the variable is not present. The tab is given below

Fig 9 – Variable importance of RF model:

Variable	MeanDecreaseAccuracy
age	129.92255
cigsperday	111.51891
sysBP	100.47808



4.4 EXTREME GRADIENT BOOSTING (XGB):

The last one is that we tried a XGBoost (Extreme Gradient Boosting) model on the dataset. XGB is a very powerful tool that boosts the model performance like anything. The one main criteria is that the dataset used contains all numeric parameters. Hence only numeric variables have been used.

The accuracy in predicting the target class is okay, and we have not tweaked the control parameters (eta…) much, as it has also reduced our overall accuracy and specificity. Hence this model remains with eta at 0.2 (lower the value, more robust the model is which tends to be overfitting)

This was done to improve the accuracy given from any of our other models but as we observed a decrease in overall accuracy and specificity, we have not gone further with model tuning.

5. MODEL VALIDATION:

We have evaluated the models mainly by the accuracies from the confusion matrix and most importantly based on the specificity only. The best model here is evaluated as the logistic regression model (mod3) It has given a good identification and prediction of the patients will be at a risk of disease.

Fig 10 – Confusion matrix of all the model accuracies.

NO.	**Algorithm Used**	**Specificity (%)**	**Sensitivity (%)**	**Overall Accuracy (%)**	**Model**
1	Logistic Regression	70.68	63.43	69.61	Mod3
2	Logistic Regression	72.09	61.94	70.60	Mod4
3	Naïve Bayes	86.62	31.75	82.84	NB2
4	Random Forest	94.38	61.11	83.29	Rf.m1
5	XGB	64.27	56.72	63.17	xgb.fit1

6. INSIGHTS AND RECOMMENDATION TO BUSINESS:

We have identified something that maybe a good factor in analysis from the dataset itself that did not work on the models. That is the effect of ***Hypertension*** on the people. It is proved in studies that ***hypertension paired with BP can be a crucial factor*** or even a decisive combination is causing a heart disease.

- We have observed a progressive growth of hypertension with age. About 60% of patients aged above 50 were hypertensive, which explains that hypertension has a parity with age.
- Almost 55% of our data comprises of people between ages 30 and 50, who practically would be yielding more to hypertension (stress) and blood pressure. They will be our people of high importance, as they will go through difficult time in the future.
- There has been a finding that smoking can to an extent handle hypertension, as we have the data describing about ***79% of women who are smokers and hypertensive, having only 25% of risk in 10 years.***
- Stroke paired with prevalent hypertension can yield more risk to the patients.

Some recommendations to businesses:

- The healthcare industry has the power to make and break the society. ***Conducting medical camps for patients*** and ***awareness programs targeting patients with specific needs*** can help the patients identify their status of well being.
- Another important development is making medical diagnostics ***accessible to most patients.***
- ***Digitalisation is the future.*** While everything becomes computerized and while artificially intelligence is at its peak, we need to make sure that most of our data is stored safely, making it as much computerized without manual interruptions. ***A tracker to the patient data*** with all the influential parameters identified, can help safe guard the lives of many patients. This will have many positive outcomes such as big data developments, confidential data handling reducing data mishandling to a large extent, cost effectiveness, etc.

IMPORTANCE OF THIS STUDY: The importance of this study is seen in many industries unlike healthcare industries. The insurance companies can formulate better plans for people who are at risk (high premium charges), the Rehabilitation centres (mentally counseling people in need), Fitness and Hospitality firms (better well being programs and food awareness)

NOTE: The R file is submitted separately with the codes.