Coronary Heart Study – Capstone Project

Predicting Ten-year Coronary Heart Disease

Table of Contents

I. Part 1

1. Project brief

The dataset provides the risk factors associated with heart disease for ~4200 patients and whether they have a risk of coronary heart disease in the next 10 years.Based on the dataset provided:

- Create a segmentation of the patients based on the demographic, behavioural and health data and analyse the risk propensity of heart disease for each segment
- Predict the probability of a patient suffering a coronary heart disease in the next 10 years
- Identify the most important factors that influence heart disease
- Come up with recommendations for
    - Preventing / reducing chances of getting a heart disease
    - Extrapolated applications of the model you build and its findings

2. Project Background

The origin of the study is closely linked to the cardiovascular health of President Franklin D. Roosevelt and his premature death from hypertensive heart disease and stroke in 1945. 17.5 million people die each year from cardiovascular diseases (CVDs), an estimated 31% of all deaths worldwide. Of these deaths, estimated 7.4 million are due to coronary heart disease. Epidemiologic studies have played an important role in elucidating the factors that predispose to CVD and highlighting opportunities for prevention. Most CVDs can be prevented by addressing behavioural risk factors

The Framingham Heart Study (FHS), the most influential investigation in the history of modern medicine is a long-term, ongoing cardiovascular study on residents of the town of Framingham, Massachusetts, USA. The study began in 1948 with 5209 adult subjects from Framingham and is now on its third generation of participants. Much of the pathophysiology of heart disease came from the results of studies from the FHS. It established the traditional risk factors, such as high blood pressure, diabetes, and cigarette smoking for coronary heart disease. Framingham also spearheaded the study of chronic non-infectious diseases in the USA and introduced preventive medicine

3. Project Objective

Problem statement: to identify which variables in the data set are able to predict coronary heart disease in individuals so that recommendations can be generated to maintain the parameters within range and prevention of CHD can happen at societal levels

- Build appropriate models to predict CHD by choosing right variables and interpret the output
- Build appropriate models on both the test and train data . Interpret all the model outputs and do the necessary modifications wherever eligible (such as pruning)
- Check the performance of all the models built (test and train). Using performance measures.

4. Assumptions

Below are the key main assumptions

- Required sample size is provided
- No imbalance in the data

5. Project Implications:

The Coronary heart Study tried to define variables that can predict heart disease in individuals.

- The understanding of the most relevant variables is critical to provide recommendations to maintain the parameters within range so that prevention of CHD can happen in society
- The use of this data is for international health organizations like WHO, World heart and diabetes associations that provide a range of health indicators so that doctors can monitor patients as per specified health parameters
- The use of this data can also be for wearables and fitness devices which can help common people keep a daily check on health vitals so that CHD can be avoided
- This data can also be used by nutritionists and nutrition-based food companies to create diet plans and appropriate marketing claims respectively
- Wellness and Fitness organization can use this data to market their programs so that fitness enthusiasts can stay in the healthy ranges
- Governments can use this data to create policy decisions for their citizens and education programs
- Pharmaceutical companies can use this data to create drugs that can be measured to reach the range values and also to generate awareness programs to generate patient flow for treatment
- Diagnosis Laboratories use this data to benchmark their patient's health data within the benchmark range data
- Tobacco avoidance products can be developed by research companies

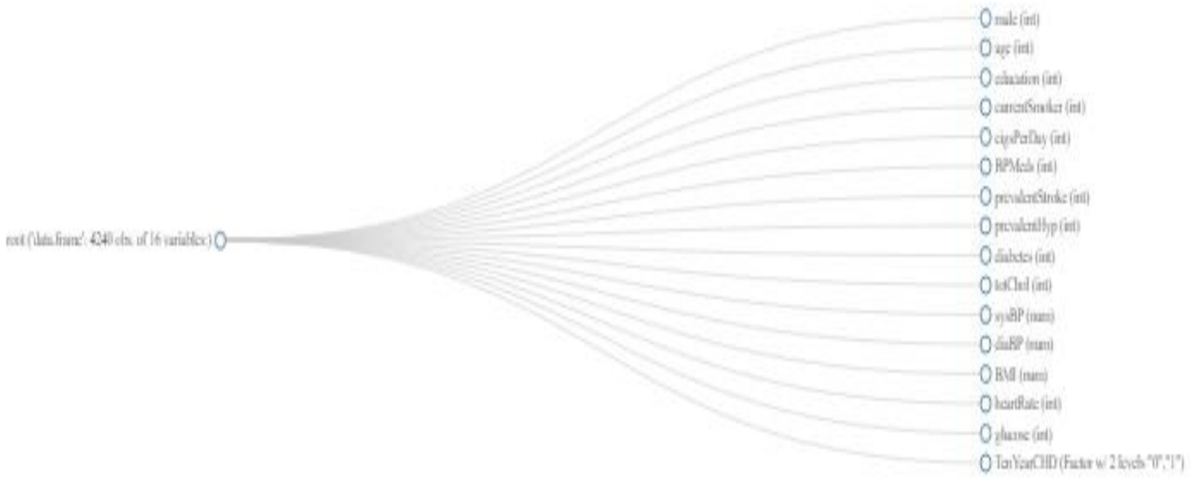
6. Exploratory data analysis

PART1: observations:

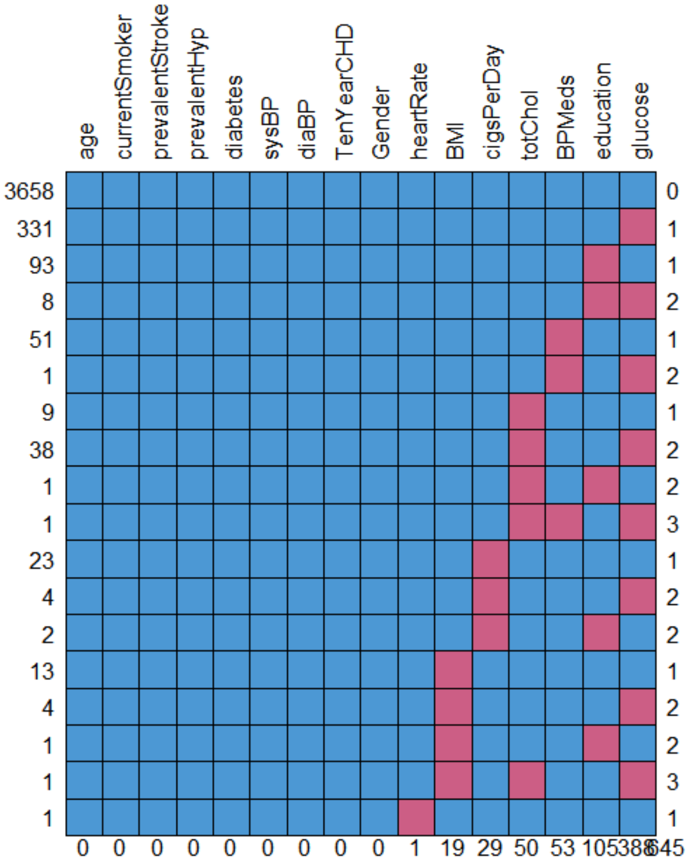
1. The data frame has 16 variables with 4240 observations
2. 8 of these variables are actually nominal or factors. The reclassification was done initially
3. Ten-year CHD is the dependant variable. 15% of the dependant variable has CHD. Data is not imbalanced
4. Missing values exist in the data set, mostly in glucose values (388). <10% of the data is incomplete and hence only complete rows can be taken for analysis
5. Outliers exist in all continuous variables on mostly upper quartiles
6. The Median age of the group is 49

	Data type	n	missing	mean	sd	median	min	max	range
age	integer	4240	0	49.58	8.57	49	32	70	38
Gender*	factor	4240	0	1.43	0.5	1	1	2	1
education*	factor	4135	105	1.98	1.02	2	1	4	3
currentSmoker*	factor	4240	0	1.49	0.5	1	1	2	1
cigsPerDay	integer	4211	29	9.01	11.92	0	0	70	70
BPMeds*	factor	4187	53	1.03	0.17	1	1	2	1
prevalentStroke*	factor	4240	0	1.01	0.08	1	1	2	1
prevalentHyp*	factor	4240	0	1.31	0.46	1	1	2	1
diabetes*	factor	4240	0	1.03	0.16	1	1	2	1
totChol	numeric	4190	50	236.7	44.59	234	107	696	589
sysBP	numeric	4240	0	132.35	22.03	128	83.5	295	211.5
diaBP	numeric	4240	0	82.9	11.91	82	48	142.5	94.5
BMI	numeric	4221	19	25.8	4.08	25.4	15.54	56.8	41.26
heartRate	numeric	4239	1	75.88	12.03	75	44	143	99
glucose	numeric	3852	388	81.96	23.95	78	40	394	354
TenYearCHD*	factor	4240	0	1.15	0.36	1	1	2	1

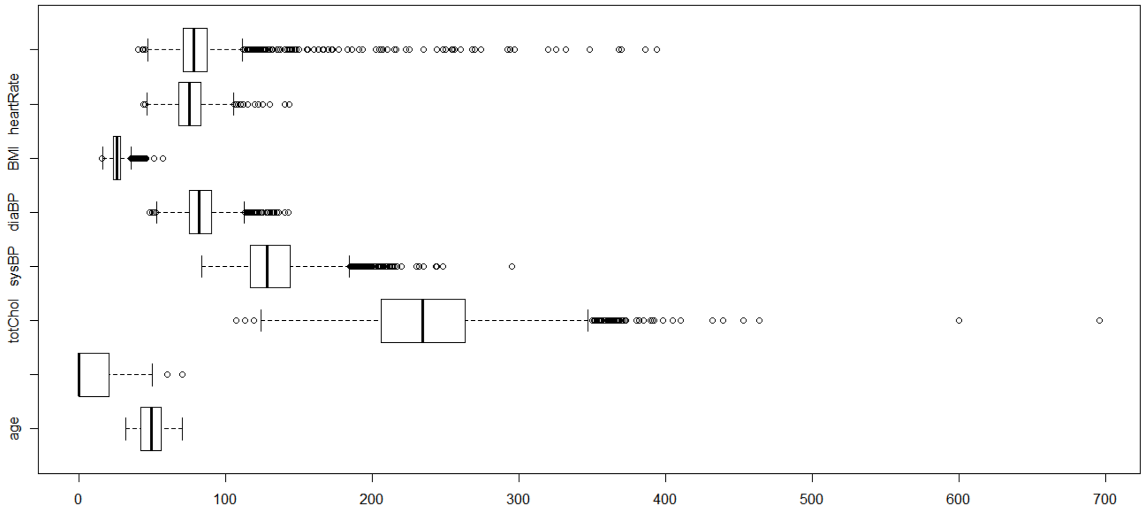
Data Structure



- Missing values: 331 values in Glucose were the highest NAs.
- Dealing with Missing Values: The highest number of Missing values were in Glucose column. Since glucose values can fluctuate within the same individual during the day, it was **not considered prudent to impute** the values. Rather since >10% of the values were NAs, they were filtered out using Complete.cases function



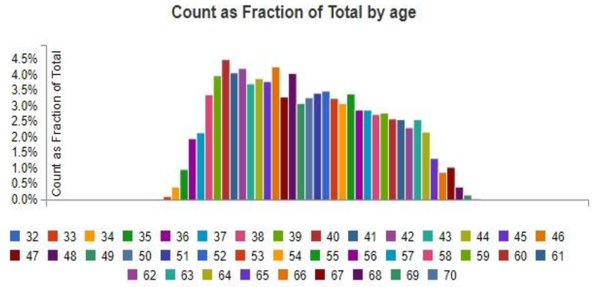
Outliers: Since all numerical variables had extreme outliers as expected in a healthcare dataset, a separate dataset was also created where all variables were capped to the 95th percentile of values

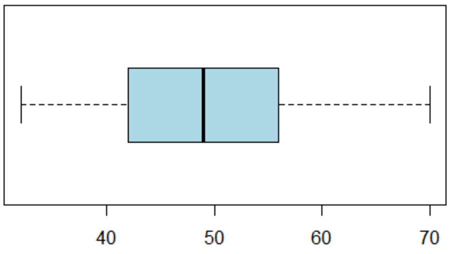


Univariate Analysis:

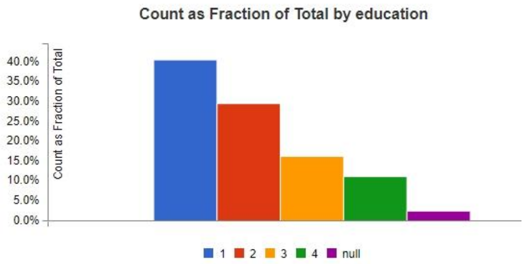
A pivot package was used to look at the distribution of variables along with boxplots for continuous variables

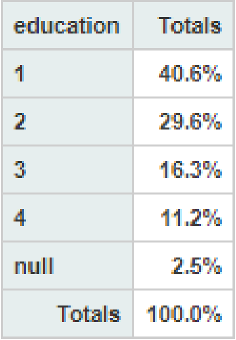
1. Age: This variable describes the age at the enrolment. the distribution is right skewed. Median value is 49 but highest no of individuals is of age 40





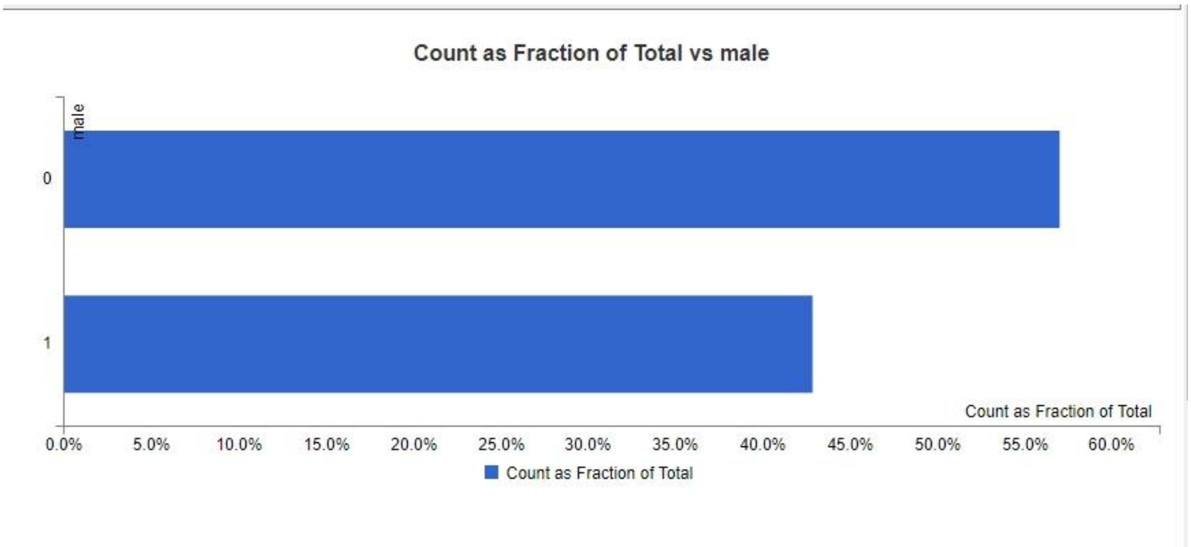
2. Education: Education is a 4-level factor variable where the level 1 is assumed to be least educated and Level 4 is the highest educated. The bar chart shows that 40% of the population is level 1, 29.6% is level2, 16.3% is level 3 and level 4 is 11.2%.



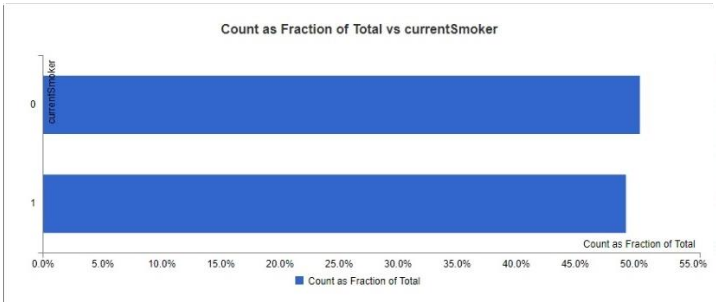


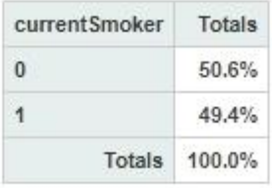
education	Totals
1	40.6%
2	29.6%
3	16.3%
4	11.2%
null	2.5%
Totals	100.0%

3. Gender: original data set had the name of the column as Male. It was changed to factorial variable as gender where 0 was female, 1 was for male. Females forms 57% of the data set.



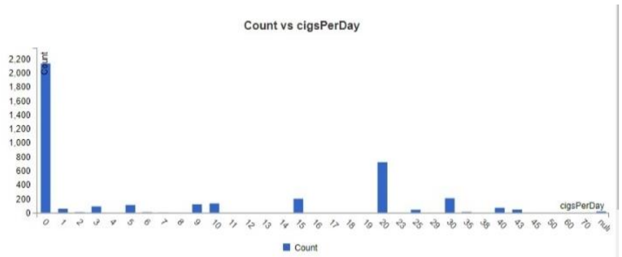
4. Current Smoker: this is a factor variable with 2 levels indicating smoking behaviour. 50.6% of the population was non-smoker currently which shows equal distribution of the risk variable

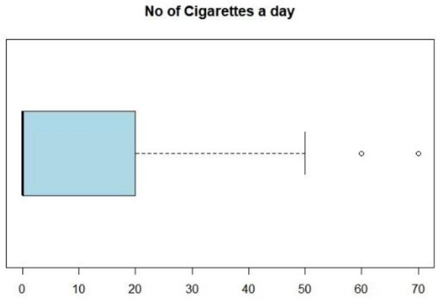

Count as Fraction of Total vs currentSmoker



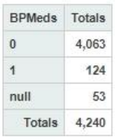
currentSmoker	Totals
0	50.6%
1	49.4%
Totals	100.0%

5. Cigs per day: it is a continuous variable with a range of 0 to 70 cigarettes smoked per day by the individual. 50% of the population was non-smoker. Outliers clearly exist in this data variable for 0.3% of the people smoke >60 cigs/day

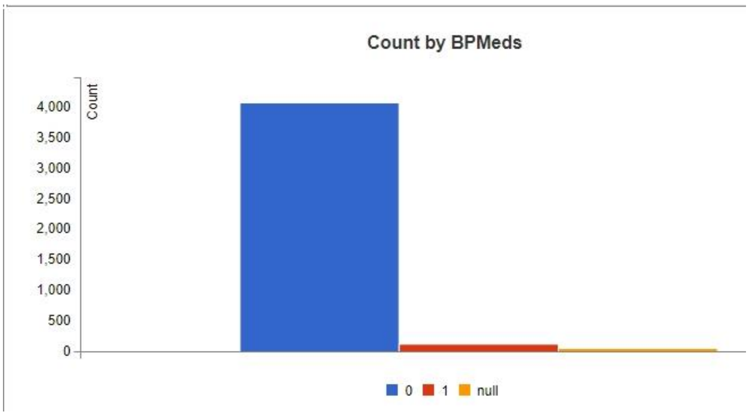

Count vs cigsPerDay


No of Cigarettes a day

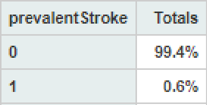
6. Taking BP Meds: this is a factor variable with 2 levels. The data seems to be imbalanced as just 3% of the people take BP meds. Missing values exist for 1.5% of data. Since it is a small proportion of the data, we can ignore these missing rows.



BPMeds	Totals
0	4,063
1	124
null	53
Totals	4,240

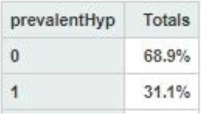

Count by BPMeds

7. Prevalent stroke: this is a factor variable with 2 levels. This is a highly imbalanced data variable as 99.4% of the population didn't have a stroke while enrolling in the study. This corelates with real world evidence as a Stroke is a cardiovascular/Brain event that happens on having extremely high blood pressure combined with a clot in blood vessels. it will be interesting to analyse only the subset having stroke on what their characteristics are for CHD



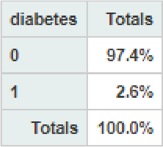
prevalentStroke	Totals
0	99.4%
1	0.6%

8. Prevalent hypertension: this is a factor variable with 2 levels. Nearly 70% of the population doesn't have hypertension.



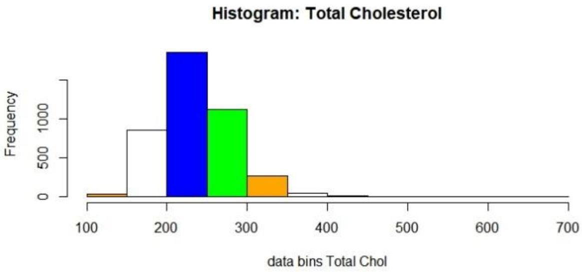
prevalentHyp	Totals
0	68.9%
1	31.1%

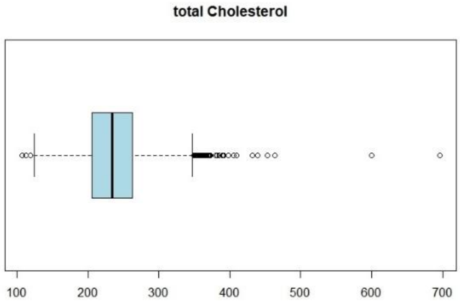
9. Diabetes: this is a factor variable with 2 levels. Nearly 97.4% of the population didn’t have diabetes at time of enrolling in study. It is an imbalanced data



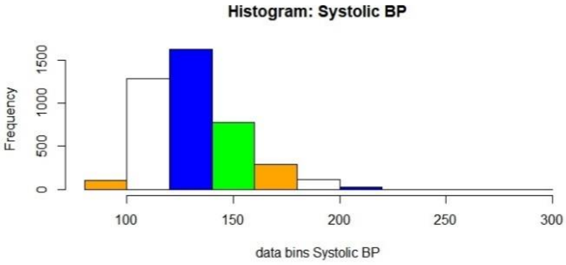
diabetes	Totals
0	97.4%
1	2.6%
Totals	100.0%

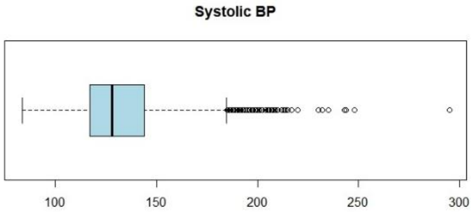
10. Total Cholesterol: this factor describing level of total cholesterol (or fat levels in blood) is a continuous variable that shows normal distribution. There are few outliers on both quartiles. 1.5% of data exist as missing values. Median value is 234 while the range is from 107 to 696



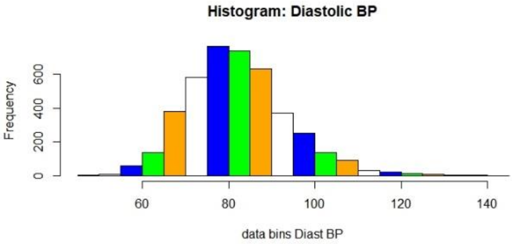


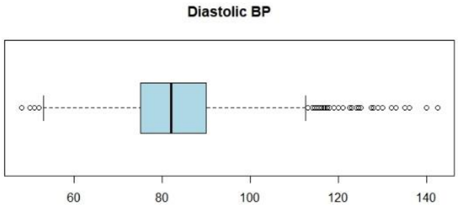
11. Systolic Blood pressure: it is the blood pressure on the walls of blood vessels when the heart pumps out the blood. This is a continuous variable with normal distribution . Outliers exist only on higher side. Median value is 128 which is higher than the standard BP of 120.





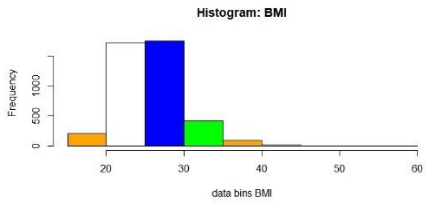
12. Diastolic Blood pressure: it is the blood pressure on the walls of blood vessels when the heart is at rest. This is a continuous variable that shows normal distribution. There are few outliers on both quartiles. Outliers exist outside both quartiles indicating very low Blood pressure which could be error. Median value is 82 which is close to the standard BP 80.

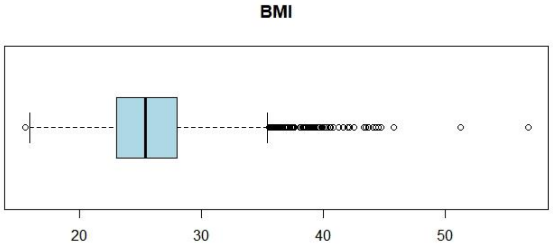




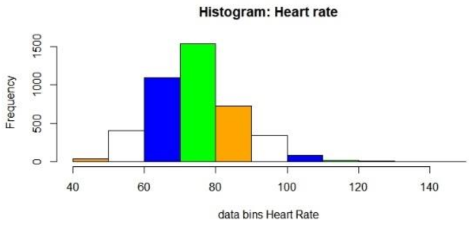
13. BMI: Body Mass Index describes the ration of height to weight and is an indicator of obesity This is a continuous variable that shows right skewed distribution. There are few outliers on

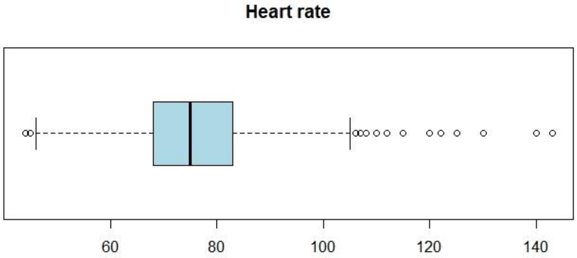
higher quartiles. Median value is 25.4 which is higher than the standard BMI 22.



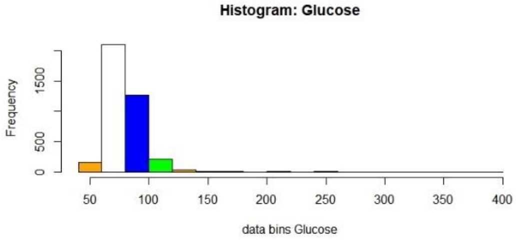


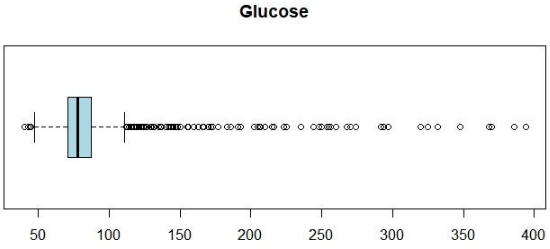
14. Heart rate: it is the per min rate at which the heart is beating and a lower heart rate indicates efficiency. this is a continuous variable that shows normal distribution. There are few outliers on higher quartiles. Heart rate can vary between individuals as well as intra-individual at different points of time so cannot be standardized.



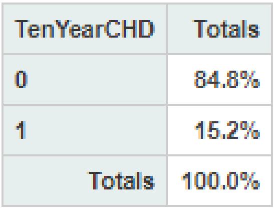


15. Glucose: this is a continuous variable that shows left skewed distribution. There are many outliers on higher quartiles. The blood values can be corelated with diabetes .

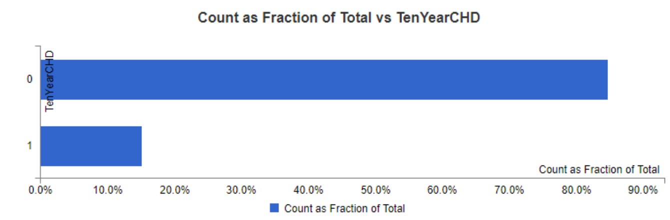




16. Ten year CHD: this is the y variable as a factor. 15% of the population have experienced a coronary heart disease



TenYearCHD	Totals
0	84.8%
1	15.2%
Totals	100.0%

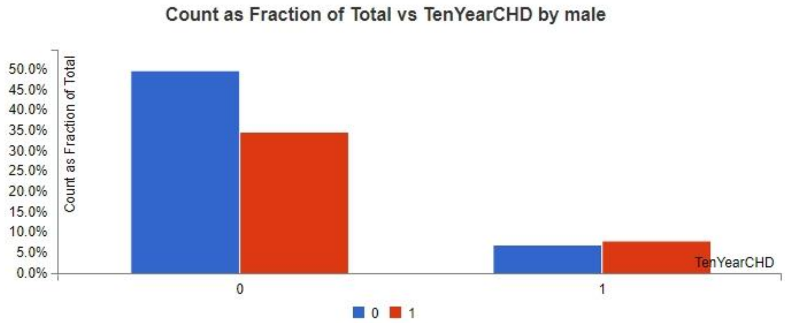


Bivariate Analysis:

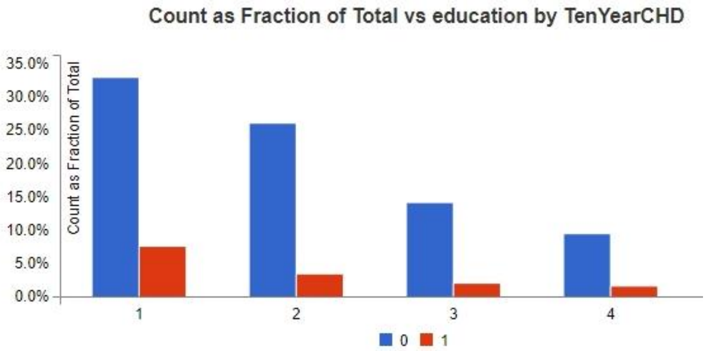
Bivariate analysis can be run with respect to comparing impact of each variable on Ten Year CHD as well as correlation between the variables.

The Impact on Ten year CHD is explored through barchart, histogram , box plot, pair plot as below

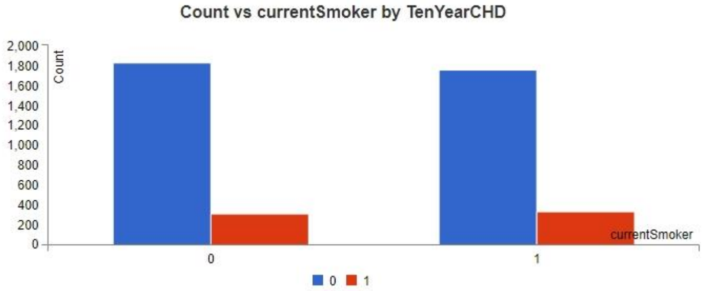
1. Gender: men seem to be more a risk of CHD compared to women in a 10 year period



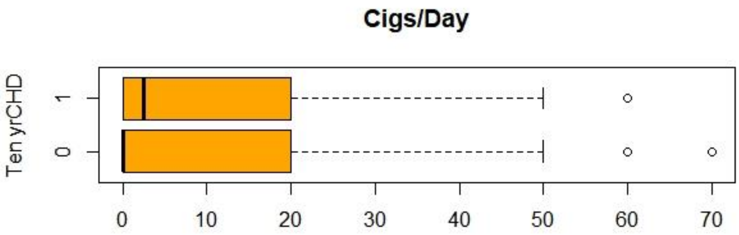
2. Education: incidence of CHD events seem higher in lower levels of education



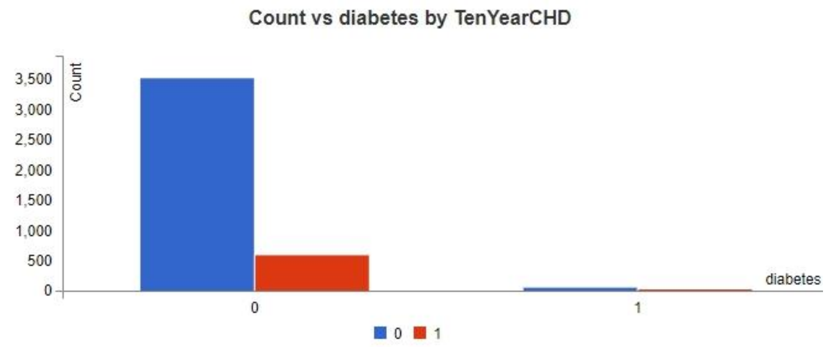
3. Current Smoker incidence of CHD does not seem to be influenced by current smoking status



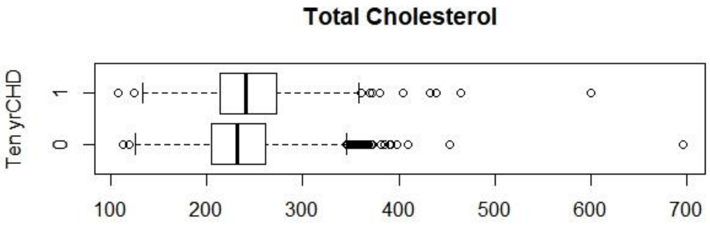
4. Cigs per day: patients with ten yr CHD had clear evidence of slightly higher no of cigs/day



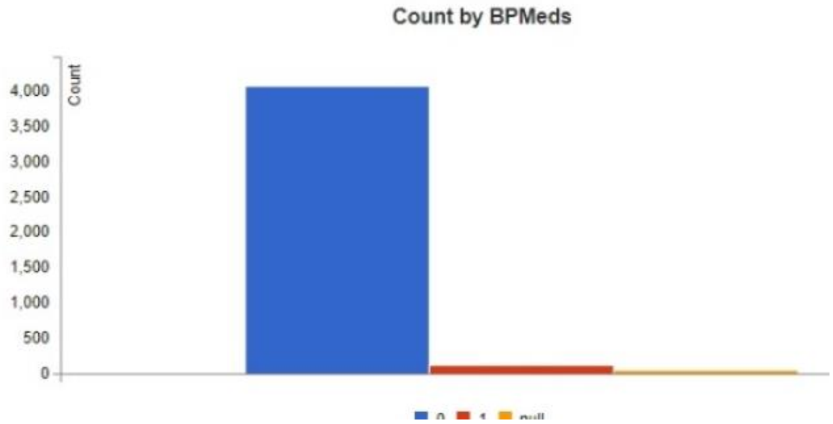
5. Diabetes: it did not seem to influence CHD but data was too small for comparison



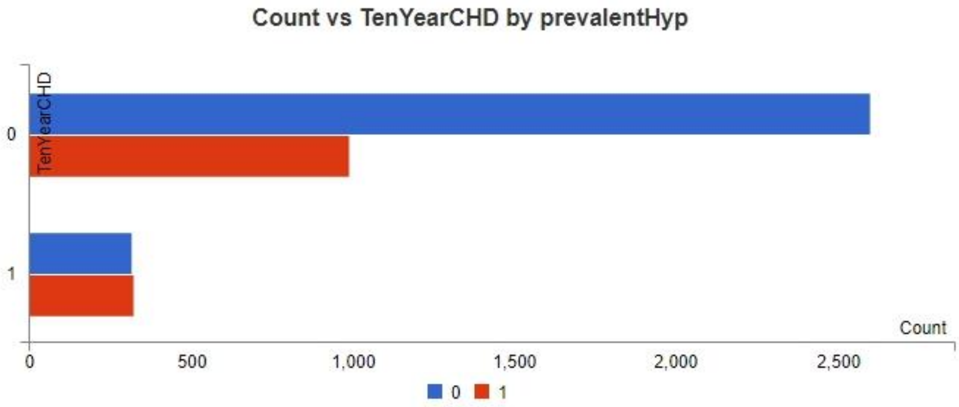
6. Total Cholesterol: patients with ten yr CHD had evidence of slightly higher total Cholesterol



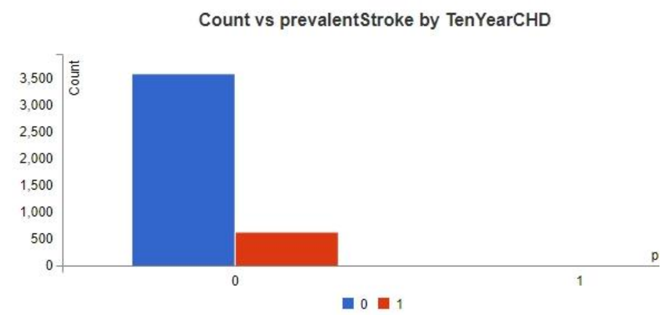
7. BP meds: individuals taking BP meds were too few to give an assessment



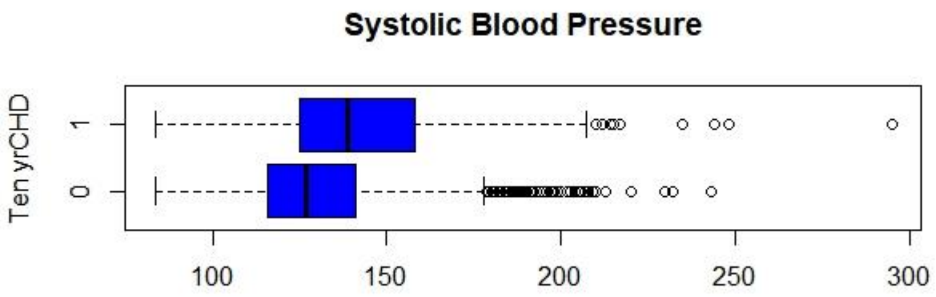
8. Prevalent Hypertension: individuals with hypertension showed higher risk of CHD



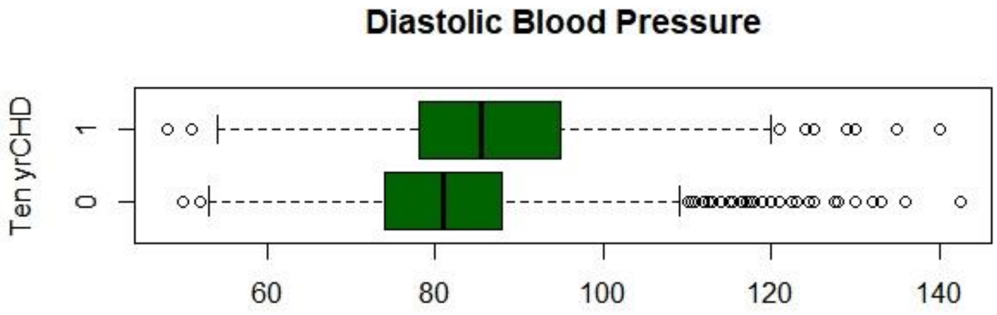
9. Prevalent stroke: didn’t seem to impact CHD as the data was too small of stroke patients



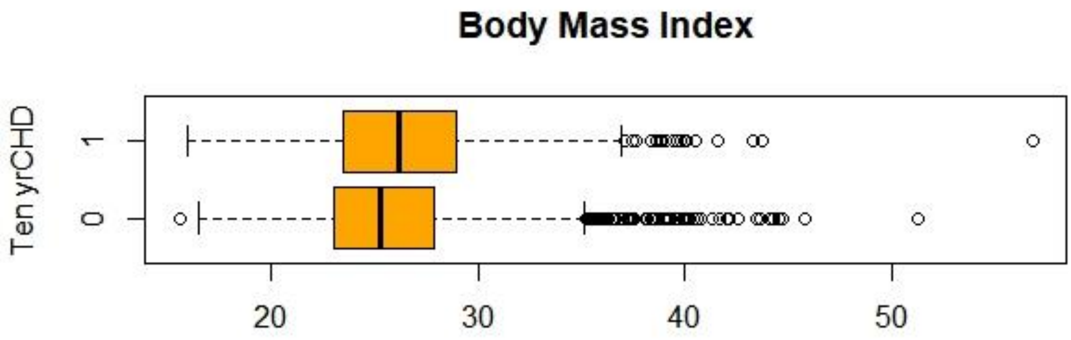
10. Systolic BP: patients with ten yr CHD had clearly evidence of higher Systolic BP



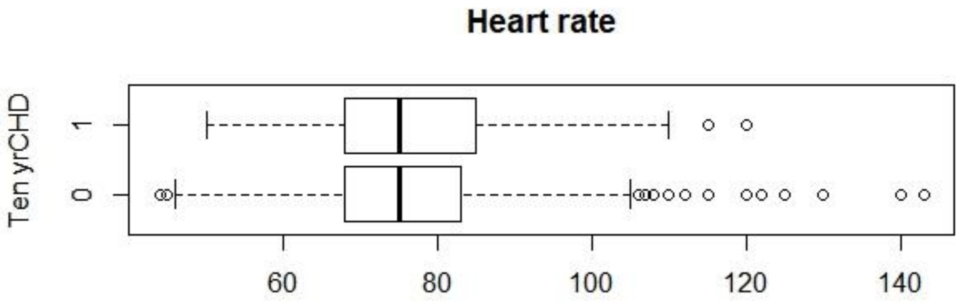
11. Diastolic BP: patients with ten year CHD had clearly evidence of higher Diastolic BP



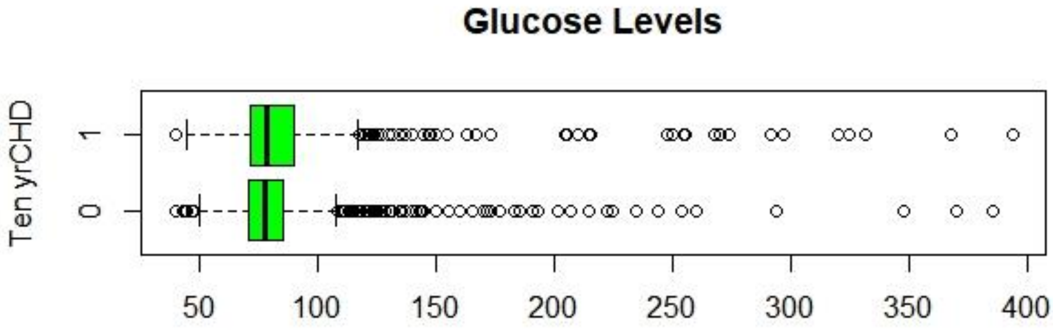
12. BMI: patients with ten-year CHD had evidence of slightly higher BMI



13. Heart rate: patients with ten-year CHD had NO evidence of higher heart rate



14. Glucose: patients with ten-year CHD had NO evidence of higher glucose levels



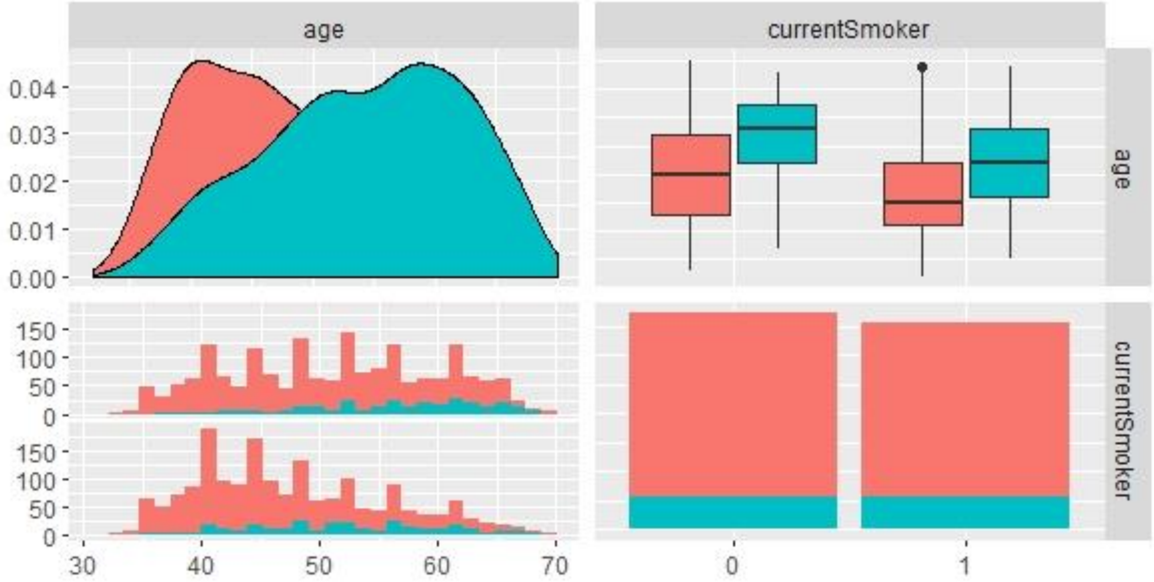
<u>INFERENCE</u>

- Based on observations, current smoker status seems to have less impact on ten-year CHD
- Ten-year CHD seems to be impacted with lower education level, Male gender, no of cigs/day, Systolic and Diastolic Blood pressure

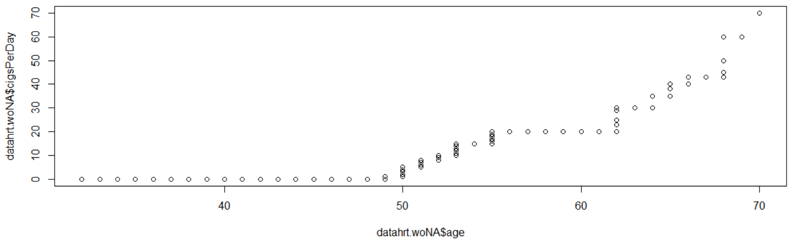
Multivariate analysis through QQPlots, Pairplots

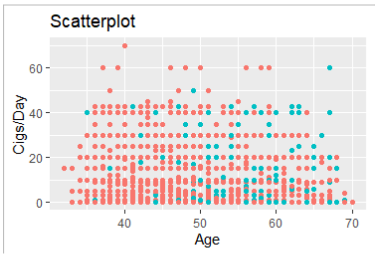
- **Age Vs Current Smoker**

Legend: Color of the plots- Red is 0 for no risk in Ten year CHD, Green is 1 for risk in Ten year CHD

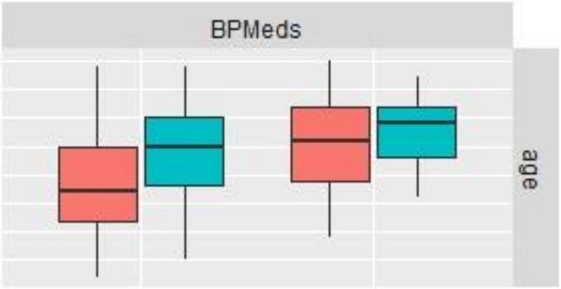


- ten-year CHD is more in People at higher Age
- Current Smoker does not seem to increase risk of Ten year CHD

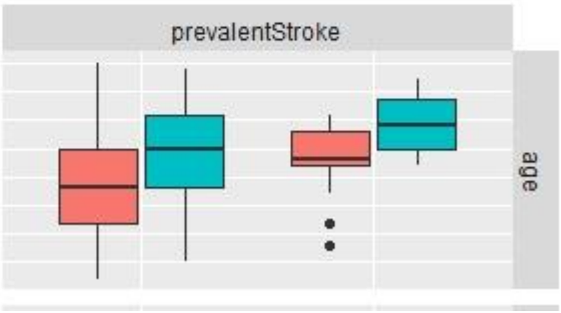




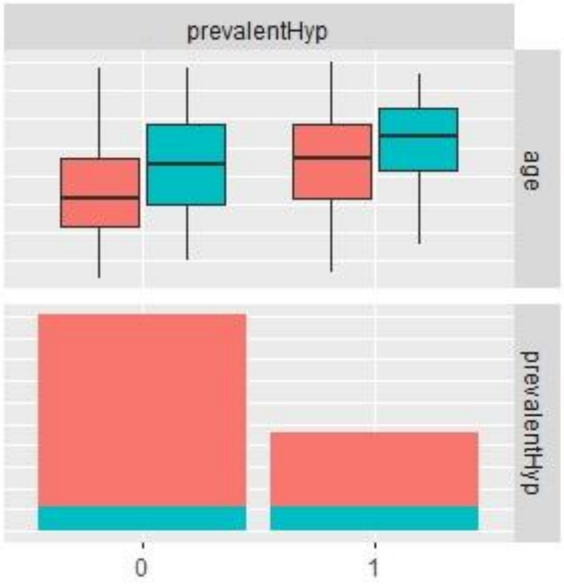
- **Age Vs Cigs/Day** showed a higher tendency to smoke more no of cig/day at higher Age that reflects addiction to smoking that becomes a habit difficult to quit with increasing age.
  Higher no of Cigs/day combined with increasing age is higher risk for ten year CHD



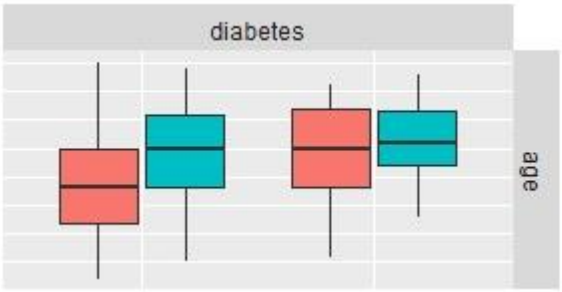
Age Vs BP Meds showed that events occurred at higher age but when BP Meds were consumed, the median age at which the events occur increase slightly



Age Vs Prevalent Stroke: Possibility of a patient having CHD and a stroke were lower and also both existed in very high age group

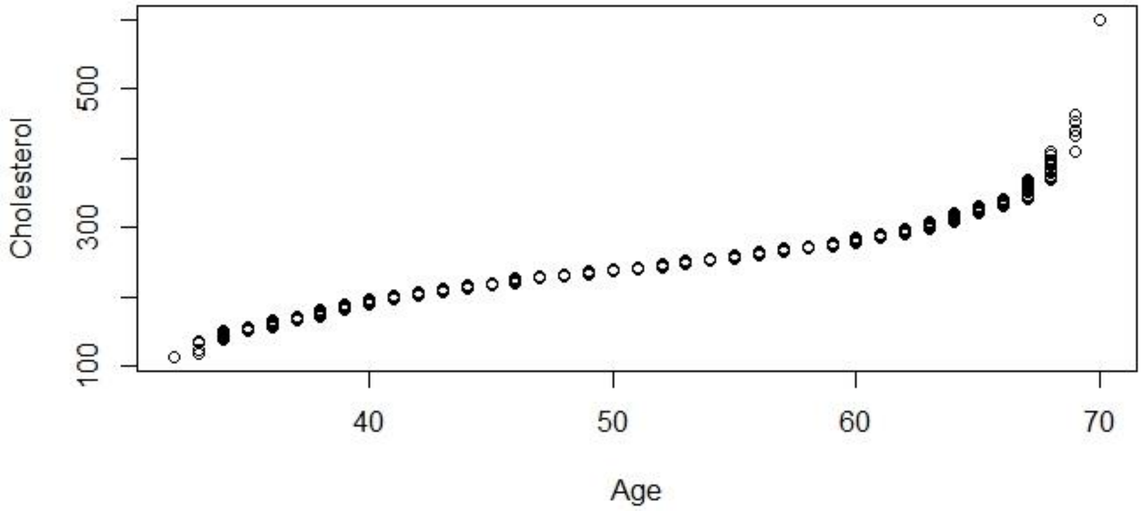


Age Vs Prevalent Hypertension: the % of patients with hypertension and CHD is higher compared to % of patients without hypertension but with CHD

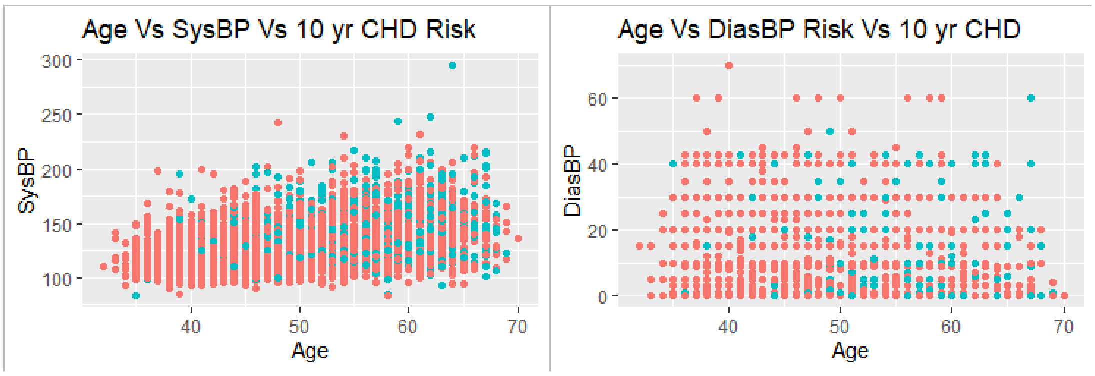


Age Vs Diabetes: the % of patients with diabetes and CHD is low while there is no difference between patients in higher age having CHD and diabetes

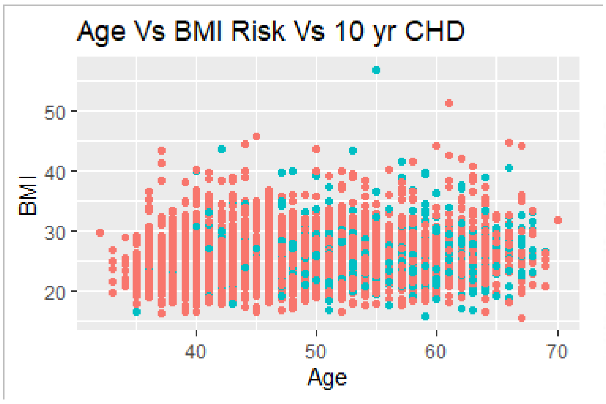
- **Age Vs Cholesterol** showed a slow increase in cholesterol till 60 post which it increased

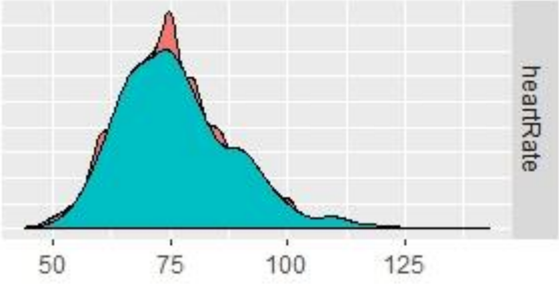


- **Age Vs Blood Pressure Vitals**
  Age is better corelated with systolic BP and CHD events increase with increasing BP

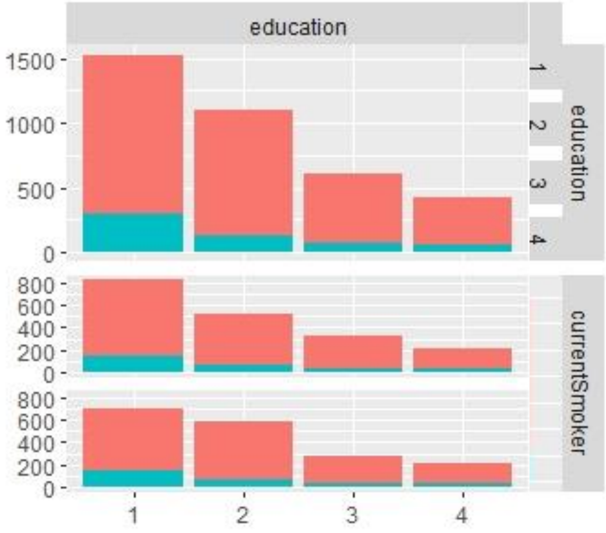


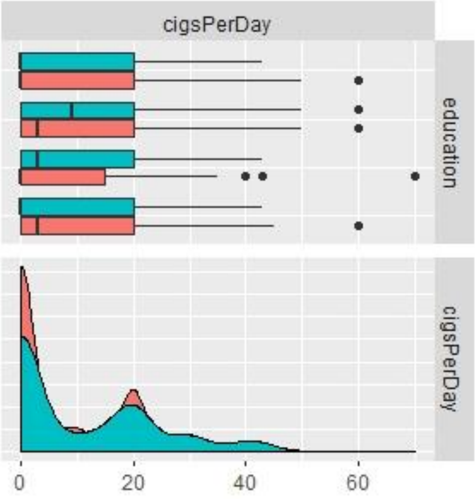
- **Age Vs BMI and Heart rate**
  There seems to be not much difference between CHD and increasing BMI or heart rate





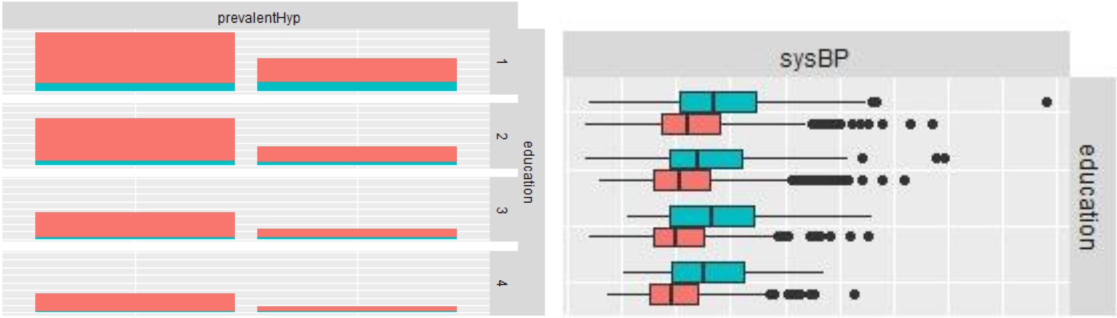
- **Education Vs Current Smoker Vs Cigarettes/day**





Being a Current smoker seems to have no impact on CHD at any education level, but cigarettes per day seems to be influencing CHD as median levels are higher in smokers even at higher education levels. Smokers with CHD also seem to have more cigarettes/day

Education Vs Prevalent hypertension

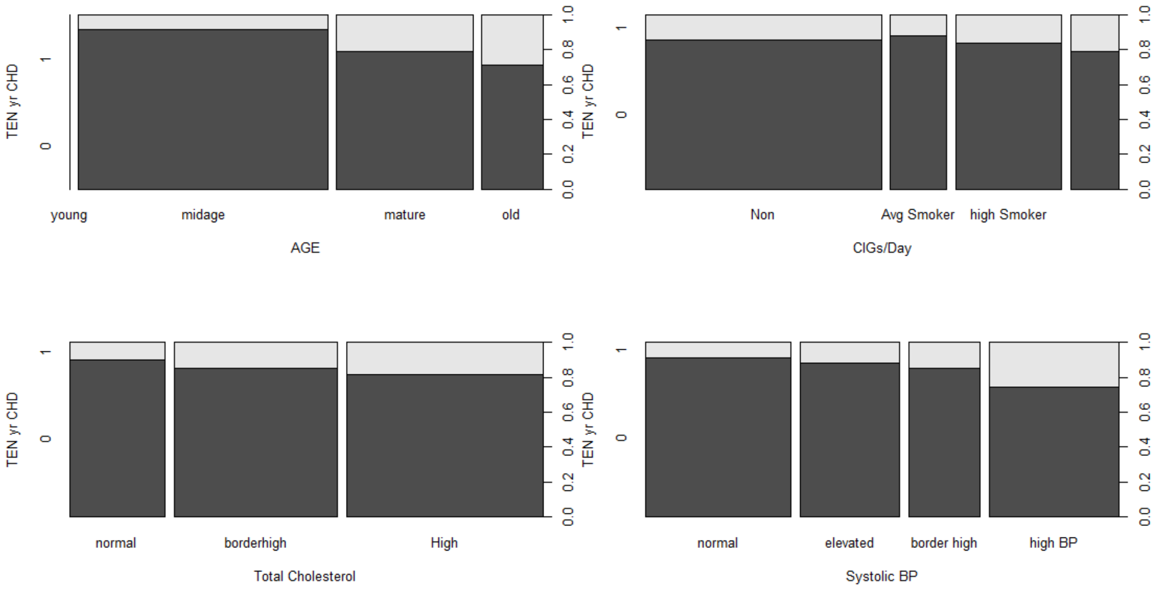


% Patients with CHD seems to be Higher in patients with higher education &prevalent hypertension. Across all education levels, Systolic BP seems to be linked to higher CHD events. Impact of Education and any other variable was not found to be significant on CHD

EDA with Feature Transformation

Numerical Variables were binned as per health guidelines and 4 variables showed higher risk of CHD

Age	<30	30-50	50-60	>60
Category	Young	Mid age	Mature	Old
Cigs/Day	0	1-10	10-20	>20
Category	Non smoker	Average	High smoker	Chain Smoker
Tot. Cholesterol	<200	200-240	>240	
Category	Normal	Border High	High	
Systolic BP	<120	120-130	130-140	>140
Category	Normal	Elevated	Border High	High
BMI	<18.5	18.5-25	25	
Category	Underweight	Healthy	Overweight	
Heart rate	<60	60-80	>80	
Category	Athletic	Regular	Elevated	
Glucose	<120	120-140	140-150	>150
Category	Normal	Prediabetic	Diabetic	High Glucose

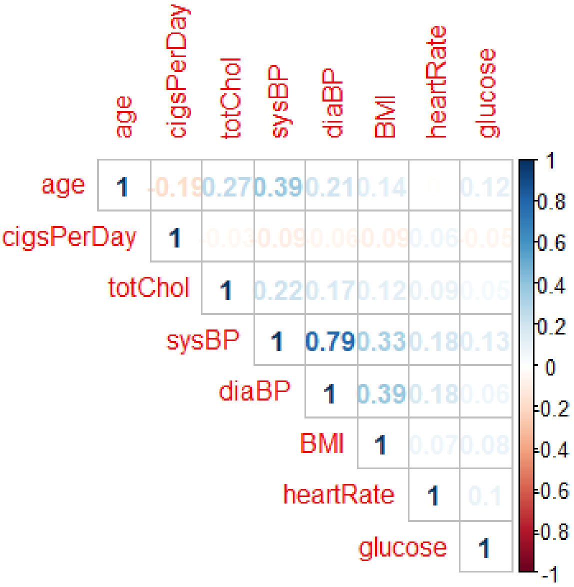


Part2

7. Correlation between variables

- Once Missing values are treated, we run a Correlation plot to check the relationship between the variables. Blood pressure variables of Systolic and Diastolic were found to be strongly corelated at 0.79. Age had weak but positive correlations with cholesterol and blood pressure variables BMI corelated with Blood pressure values too

Cor-relation



- **Chi-Square Test results**

A Chi Sq test run on the factor-based variables compared to the dependent variable (ten-year CHD) gives below results for p value. **Being a Current Smoker** did not seem to be significant impact on ten-year CHD whereas other Variables had low p value & can influence Ten Year CHD.

```
           ROW           Column    ChiSq        p.value

2: TenYearCHD        education    31.19     7.716846e-07
3: TenYearCHD    currentSmoker     1.23     2.656692e-01
4: TenYearCHD       TenYearCHD  3650.25     0.000000e+00
5: TenYearCHD           Gender    30.24     3.816766e-08
6: TenYearCHD           BPMeds    27.64     1.457333e-07
7: TenYearCHD  prevalentStroke     6.86     8.775829e-03
8: TenYearCHD     prevalentHyp   119.26     9.161520e-28
9: TenYearCHD         diabetes    30.34     3.608263e-08
```

8. Data pre-processing:

Checking for Multicollinearity:

A numeric data frame was created and a linear regression model was run. Variance Inflation Factor function was run to check for Multicollinearity. The variables did not show VIF >4. So none of the numeric variables are considered to be dropped for multicollinearity

age	cigsPerDay	sysBP	diaBP
1.32	1.05	3.11	2.86
heartRate	glucose	BMI	totChol
1.06	1.04	1.19	1.11

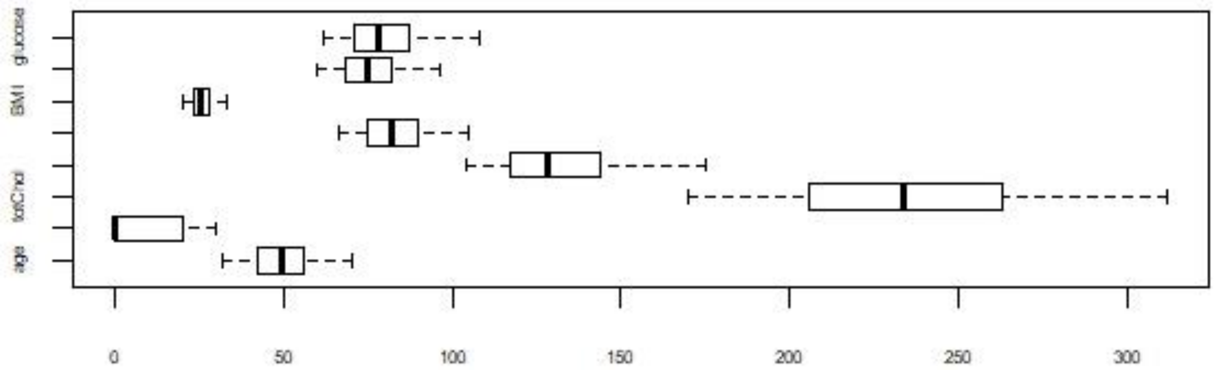
Eliminating variables through Chi sq test:

Chi Sq test was run between factor variables to see if there was low significance value due to correlation. No evidence was found to drop factorial variables (except current smoker) due to correlation between factors & ten-year CHD due to p value being >0.05

Data Pre-processing

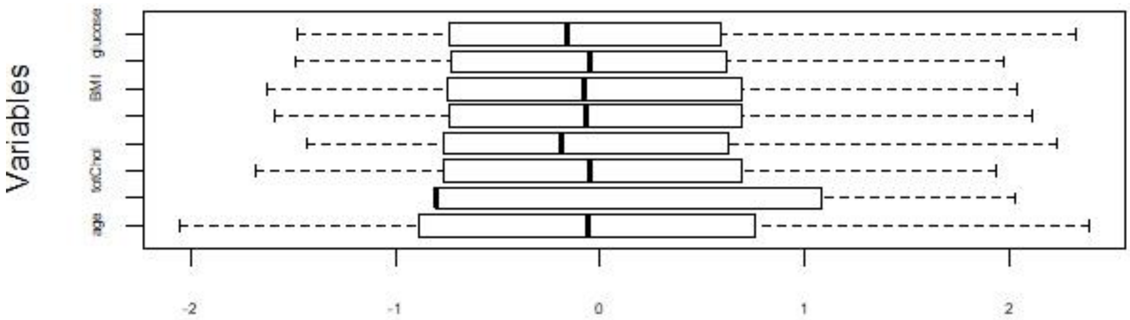
Capping technique for removing outliers:

Since most of the numeric variables have multiple outliers on both sides, we considered doing an outlier treatment with capping technique on 7 numeric variables. Since most numeric variables in health like blood pressure, cholesterol, BMI, heart rate, glucose are range bound, capping will translate outlier data at the extreme limit of the Ranges (=0.05 or =0.95). Any outlier will be capped to the outside limit of the range instead of using median values.



Variable transformation: Scaling the data

The above graph also shows that the data between variable is at different levels and hence we need to scale the data so that a comparison can be done. Since the objective is to do a supervised learning by checking influence of variables on ten-year CHD as a factorial variable, and not as an equation, then the scaling will help in comparison of the variable.



Data Pre-processing

Variable transformation: eliminating variables

Among numeric variables,

- heart rate does not seem to be influencing the Ten-year CHD. Also, since heart rate is a measurement at a single point of time, and may tend to fluctuate within individuals during a single day, the significance of this variable is considered low for influencing ten-year CHD
- Variables which will be evaluated to be eliminated after model creation will be diabetes glucose as both variables capture similar data points

Among factor variables,

- Current Smoker: the median values in people who are current smoker with CHD are same to the median values of non-smokers with CHD. Also, there is another numeric variable, Cigarettes per day which may be capturing similar information. Hence current smoker can be eliminated
- Education: among all the variables, this variable stands out distinctly as a measure of demographic of the population rather than measuring a risk factor. Also, there was limited difference between CHD at different education levels. Hence this variable will be initially eliminated and then later considered to see if it improves any score
- Prevalent Stroke: the data set seems to be imbalanced for this variable as only 0.6% of the population (25/4125 observations) seems to have prevalent stroke. While splitting in a ratio, the values in train model will be very low to capture significant information. We can consider eliminating this variable for now and then check later if by adding it, the score improves
- Total cholesterol and BMI also have similar medical implications and hence will be checked

Creation of test and train data set

The data is split into test and train data sets. Since more than 3600 observations are available, a ratio between 80: 20 is considered fair for the train: test ratio as significant amount of data is available to train the model as well as validate the data from test dataset..

The elimination of the variables was done on train model first and then we attempted various models on the data to improve the performance score by adding some of the variables to check if scores improve. The final data set such created would be validated with the test data set.

Checking for proportion of separation

The split data was checked for proportion of separation of the y variable ten-year CHD. The values were comparable

final	**0**	**1**	**total**
values	3101	557	3658
%	84.8	15.2	
train			
values	2315	428	2743
%	84.4	15.6	
test			
values	786	129	915
%	85.9	14.1	

9. Analytical Approach:

Checking for various Models:

The problem statement entails a supervised learning technique. Since this data is looking to classify risk factor of ten-year CHD, we first attempted a **logistic regression** model.

Other Models which were considered are

1. **CART** as classification can be attempted based on various rules based on variables suggested by the model

2. **Random Forest** as classification will be given based on importance of different variables created by generating multiple trees

Base Model 1: Running a logistic regression on factorial variables:

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -2.23720    0.10726 -20.858  < 2e-16 ***
education2       -0.46265    0.11674  -3.963 7.40e-05 ***
education3       -0.36844    0.14352  -2.567   0.0103 *  
education4       -0.32670    0.15857  -2.060   0.0394 *  
currentSmoker1    0.16485    0.09870   1.670   0.0949 .  
Gender1           0.49358    0.09883   4.994 5.91e-07 ***
BPMeds1           0.51785    0.22139   2.339   0.0193 *  
prevalentStroke1  0.77123    0.47938   1.609   0.1077    
prevalentHyp1     0.91120    0.09938   9.169  < 2e-16 ***
diabetes1         0.93004    0.22380   4.156 3.24e-05 ***
```

Base Model2: running a logistic regression on numerical variables:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -8.610615   0.595463 -14.460  < 2e-16 ***
age          0.067336   0.006608  10.190  < 2e-16 ***
cigsPerDay   0.027681   0.003947   7.014 2.32e-12 ***
totChol      0.001552   0.001114   1.393    0.164    
sysBP        0.015826   0.003437   4.605 4.13e-06 ***
diaBP        0.001174   0.006321   0.186    0.853    
BMI          0.011888   0.012277   0.968    0.333    
heartRate   -0.005571   0.004161  -1.339    0.181    
glucose      0.007579   0.001677   4.518 6.23e-06 ***
```

Univariate Logistic regression on these doubtful variables shows diastolic BP, BMI and Total Cholesterol to be significant but **current Smoker** & **heart rate is still found insignificant. Hence these variables will be dropped for initial model**

	Estimate	Std. Error	z value	Pr(>|z|)	
heartRate	0.0047	0.0037	1.240	0.215	
diaBP	0.0326	0.0036	8.902	<2e-16	***
BMI	0.052	0.010	4.926	8.37e-07	***
currentSmoker1	0.106	0.092	1.159	0.247	
prevalentStroke1	1.241	0.451	2.749	0.00598	**

Analytical Approach

A Full logistic regression model was built on the remaining 13 variables. The model **has AIC of 2132**

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -2.21857    0.12677 -17.500  < 2e-16 ***
age               0.51445    0.06529   7.880 3.28e-15 ***
cigsPerDay        0.23565    0.05972   3.946 7.95e-05 ***
totChol           0.07500    0.05858   1.280 0.200393    
sysBP             0.44583    0.10224   4.360 1.30e-05 ***
diaBP            -0.13435    0.08899  -1.510 0.131123    
BMI               0.01719    0.06111   0.281 0.778499    
glucose           0.03607    0.05700   0.633 0.526846    
education2       -0.12464    0.14077  -0.885 0.375941    
education3       -0.05365    0.17060  -0.315 0.753134    
education4       -0.08756    0.18700  -0.468 0.639622    
Gender1           0.48063    0.12453   3.860 0.000114 ***
BPMeds1           0.10234    0.27396   0.374 0.708719    
prevalentStroke1  1.09408    0.54869   1.994 0.046155 *  
prevalentHyp1     0.28731    0.16660   1.725 0.084612 .  
diabetes1         0.42044    0.29403   1.430 0.152743    
```

An Anova test on the model results show diastolic BP, glucose, BMI , BPMeds education insignificant.

```
                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                             2742     2375.6              
age              1  153.414      2741     2222.2 < 2.2e-16 ***
cigsPerDay       1   25.133      2740     2197.1 5.352e-07 ***
totChol          1    2.842      2739     2194.2   0.09181 .  
sysBP            1   64.506      2738     2129.7 9.624e-16 ***
diaBP            1    0.600      2737     2129.1   0.43852    
BMI              1    1.231      2736     2127.9   0.26731    
glucose          1    1.381      2735     2126.5   0.23986    
education        3    1.640      2732     2124.9   0.65036    
Gender           1   15.333      2731     2109.5 9.014e-05 ***
BPMeds           1    0.548      2730     2109.0   0.45917    
prevalentStroke  1    4.006      2729     2105.0   0.04533 *  
prevalentHyp     1    2.891      2728     2102.1   0.08909 .  
diabetes         1    1.983      2727     2100.1   0.15912    
```

The accuracy of the model is about 84.8% on the train model and the test model is also 85.7%

Train

	FALSE	TRUE
0	2300	15
1	401	27

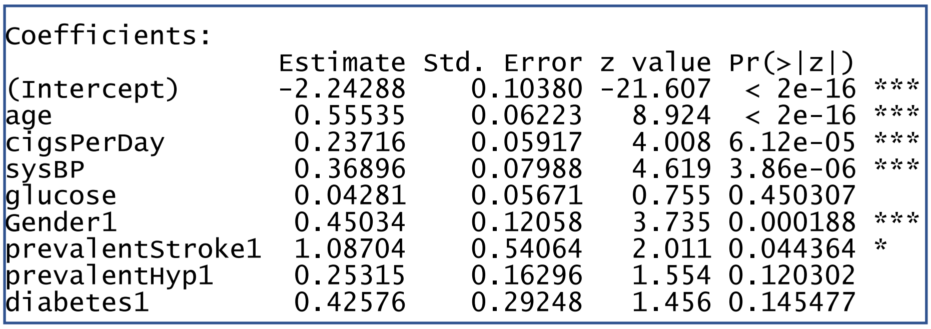
test

	FALSE	TRUE
0	777	9
1	122	7

An iteration was done on the model to improve

- AIC score
- Confusion matrix

The variables finally selected for the model were as below. The AIC score reduced to 2123 from 2132



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -2.24288    0.10380 -21.607  < 2e-16 ***
age                0.55535    0.06223   8.924  < 2e-16 ***
cigsPerDay         0.23716    0.05917   4.008 6.12e-05 ***
sysBP              0.36896    0.07988   4.619 3.86e-06 ***
glucose            0.04281    0.05671   0.755 0.450307    
Gender1            0.45034    0.12058   3.735 0.000188 ***
prevalentStroke1   1.08704    0.54064   2.011 0.044364 *  
prevalentHyp1      0.25315    0.16296   1.554 0.120302    
diabetes1          0.42576    0.29248   1.456 0.145477    
```

the improvement in the confusion matrix was an accuracy of 85% in train model and 86% in test model

	FALSE	TRUE
0	2301	14
1	398	30

	FALSE	TRUE
0	779	7
1	121	8

Analytical Approach

Part3

10. Modelling data

- **Logistic regression**

Reason to choose: A binary classification problem is most suited to be solved by a logistic regression algorithm as it not only gives the most important variables among the available variables but also gives a **formula** expressing the weightage of multiple variables and the interaction between them.

- Option1: We had run a log regression model in the analytical approach with scaling and capping
- Option2: We attempted to run the logistic model without scaling and capping. The hypothesis was, higher values of observations among variables are more likely to impact the result of ten-year CHD.

Variables which had shown no impact earlier like Glucose, Prevalent Stroke, Education & total cholesterol started showing significance. This was revaluated by anova test and observation was confirmed.

The variables dropped for log modelling were, diastolic BP, Heart rate, BMI

Further dropping variables was tested with Blorr

Model-
Logit

BLORR

Using Blorr Library, we found out which variables were most significant

Variables Entered/Removed:

- ✓ **age**
- ✓ **sysBP**
- ✓ **cigsPerDay**
- ✓ **Gender**
- ✓ **glucose**
- ✓ **totChol**
- ✓ **prevalentHyp**

No more variables to be added or removed.

```
                       Stepwise Summary
-------------------------------------------------------------
Variable          Method        AIC         BIC        Deviance
-------------------------------------------------------------
age              addition    2333.78     2345.75     2329.78
sysBP            addition    2289.78     2307.73     2283.78
cigsPerDay       addition    2259.51     2283.43     2251.51
Gender           addition    2245.65     2275.56     2235.65
glucose          addition    2232.38     2268.27     2220.38
totChol          addition    2229.47     2271.34     2215.47
prevalentHyp     addition    2228.87     2276.72     2212.87
-------------------------------------------------------------
```

These are final variables chosen.

AIC of earlier model was 2233.9

Revised model with few variables, shows AIC of 2228.9 which is an improvement

Results of train model with logistic regression at threshold of 0.5

```
          Reference
Prediction    0    1
         0 2461   16
         1  414   36
```

```
               Accuracy : 0.8531
                 95% CI : (0.8397, 0.8657)
    No Information Rate : 0.9822
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1152

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8560
            Specificity : 0.6923
         Pos Pred Value : 0.9935
         Neg Pred Value : 0.0800
```

Results of test model with logistic regression

```
          Reference
Prediction   0   1
         0 622   2
         1 101   6
```

```
               Accuracy : 0.8591
                 95% CI : (0.8317, 0.8835)
    No Information Rate : 0.9891
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0857

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.86030
            Specificity : 0.75000
         Pos Pred Value : 0.99679
         Neg Pred Value : 0.05607
```

Model-Logit

Observations, conclusions & Model accuracy measures for Logistic Regression model

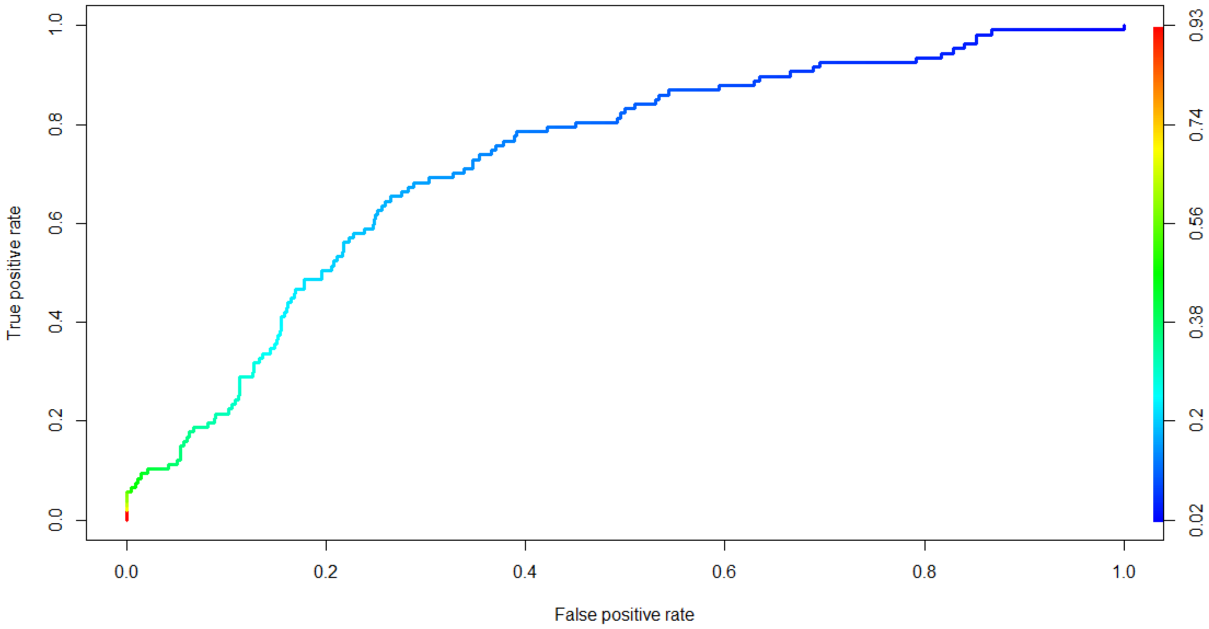
- Accuracy of train model was 86% tested by confusion matrix
- Accuracy of the model holds at same levels in the test data as well.
- Specificity which aims at predicting people with **Ten-year CHD is at 75% in test** data. This is critical portion as the objective was to classify no of people with risk of CHD
- Attempt was made to increase or reduce the threshold from 0.5 to 0.6 or 0.4
- But the results were found best at 0.5 threshold and hence retained.
- Mcfadden Rsq through blorr gives a **value of 0.119** for the current model (likelihood of full model vs intercept model). 0.1 to 0.3 is considered to be a **moderately good model**
- The risk of getting CHD from the LOG model was described by below exponential of coefficient of the model

Log odds of tenyear CHD= 0.0012 +1.073 x age +1.019x cigs/day +1.275 x prevalent Hypertension

+1.013 x Systolic BP +1.0071 x glucose +1.631 x Male gender

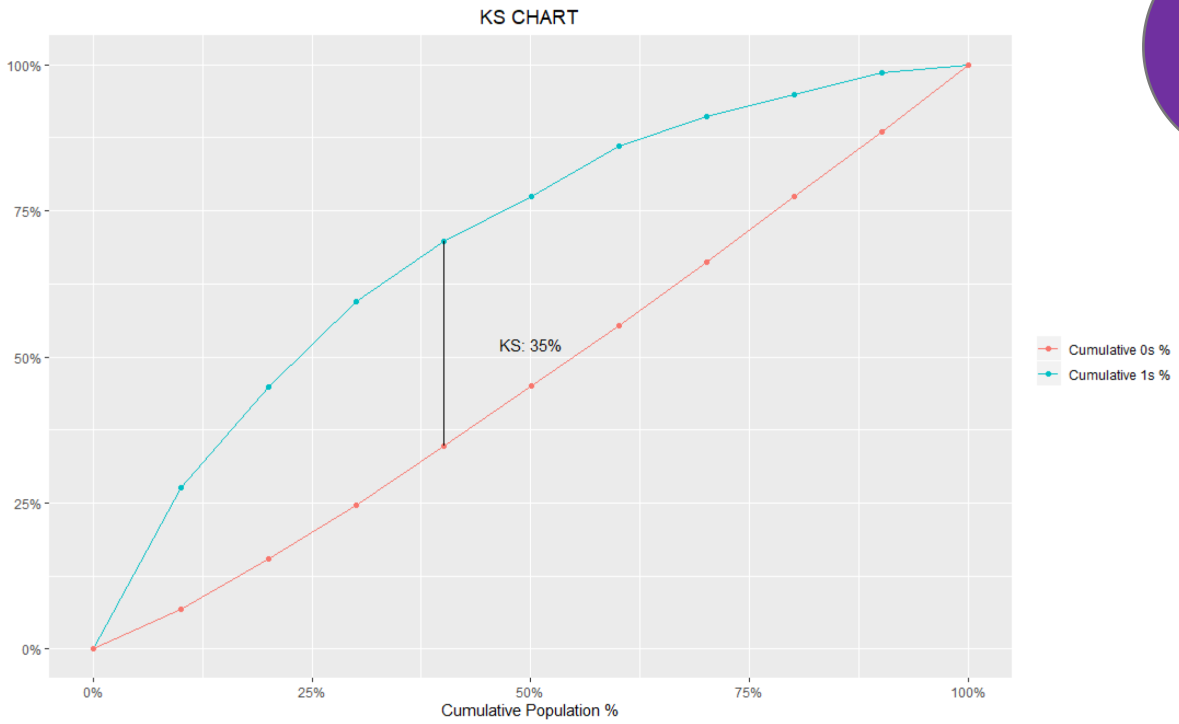
Example: a one unit increase in age leads to a 7.3% increase in odds likelihood to get CHD

- An ROCR was conducted to compare TPR vs FPR



The cumulative response achieves maximum separation at 35%. This is the decile till which we should be looking at for maximum CHD

Model-Log

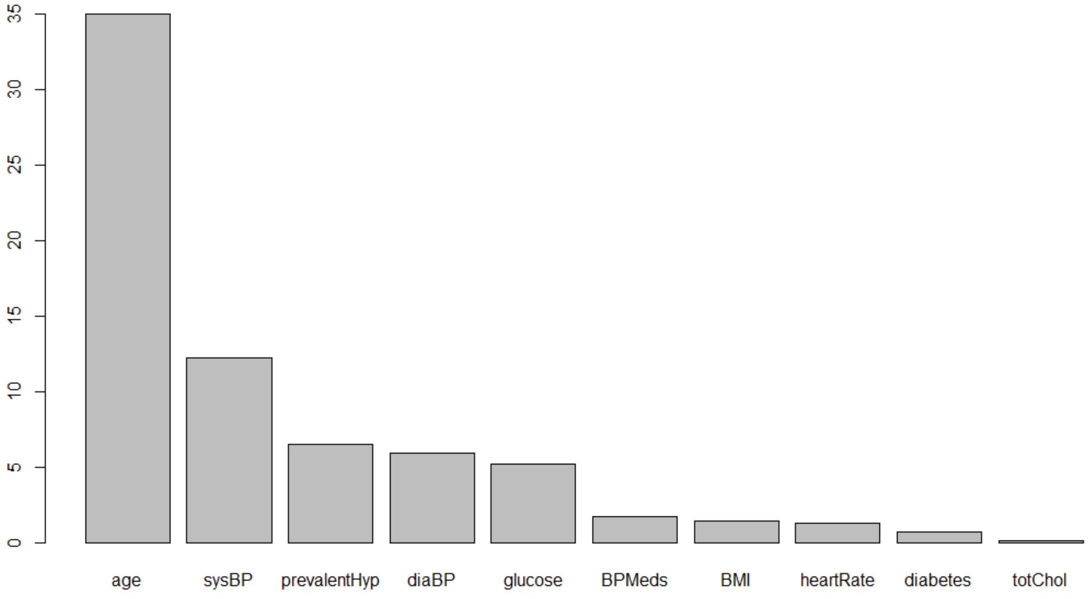


- **<u>CART Model:</u>**

Reason to choose: Since decision trees seem less sensitive to outliers, a CART model was attempted as it will also tell the importance of various variables while classifying

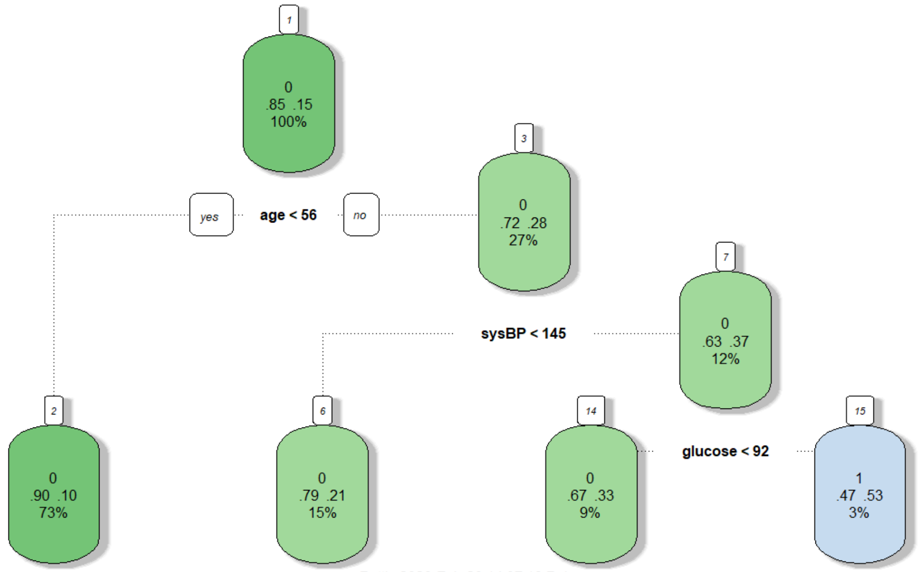
We ran a CART model in 2 ways. We ran all the variable through it and then choose certain important variables only (identified in log regression model). parallelly we also checked if outliers or scaling changed the outcomes in any manner.

Part 1: In full Model, the below variables were considered by the algorithm as critical to form the tree



CART tree gives the following rules to classify the data with age, Systolic BP and Glucose levels being the most significant

```
TenYearCHD                                                      cover
       0.10 when age <  56                                        73%
       0.21 when age >= 56 & sysBP <  145                         15%
       0.33 when age >= 56 & sysBP >= 145 & glucose <  92          9%
       0.53 when age >= 56 & sysBP >= 145 & glucose >= 92          3%
```



Rattle 2020-Feb-20 14:07:13 Rohan

Model-CART

But, treating for outliers, improved the prediction score of the model. Scaling did not improve it

Specificity improved **from 53% to nearly 56%** but is still less than 75% achieved by Log model

Confusion matrix of untreated model

```
          Reference
Prediction    0    1
         0 2289   35
         1  379   40
```

```
               Accuracy : 0.8491
                 95% CI : (0.8351, 0.8623)
    No Information Rate : 0.9727
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1212
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.85795
            Specificity : 0.53333
```

Confusion matrix of Outlier treated model

```
          Reference
Prediction    0    1
         0 2272   41
         1  380   52
```

```
               Accuracy : 0.8466
                 95% CI : (0.8326, 0.8599)
    No Information Rate : 0.9661
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1507
Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8567
            Specificity : 0.559
```

Model-
CART

Part 2: we tried an ensemble method, where variables shown important in logistic regression were used for CART model.

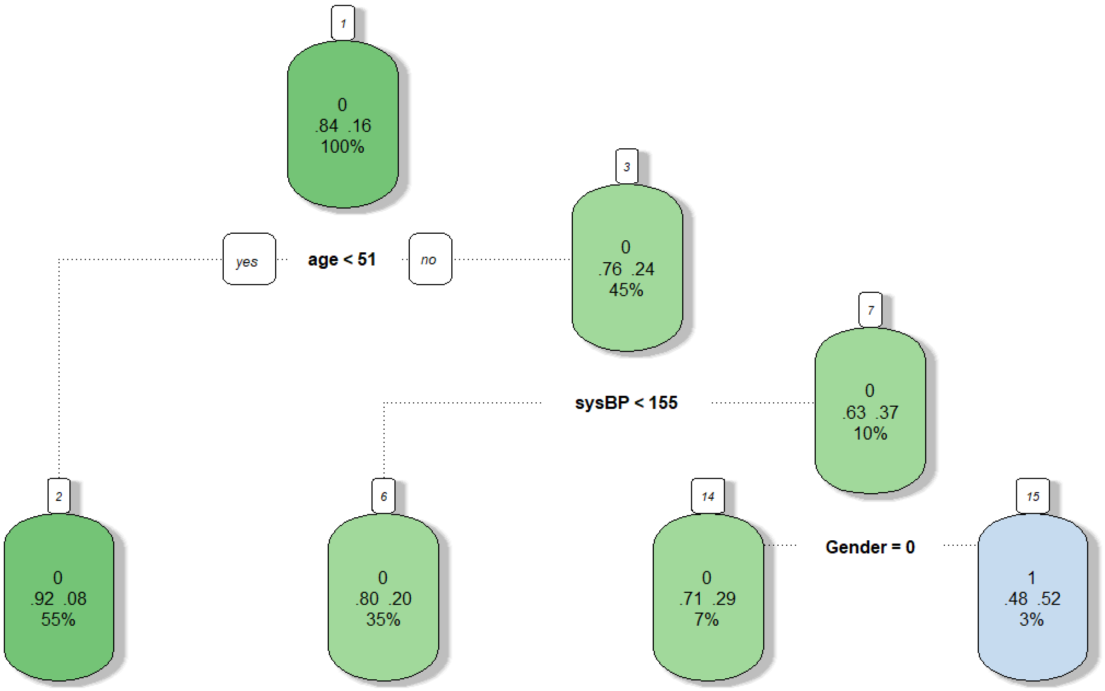
CONFUSION MATRIX for ensemble model did not show improvement over previous model

```
          Reference
Prediction    0    1
         0 2293   29
         1  387   34
```

```
               Accuracy : 0.8483
                 95% CI : (0.8344, 0.8616)
    No Information Rate : 0.977
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1047

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.85560
            Specificity : 0.53968
```

The highest probability of ten year CHD was in age group >51 , followed by sys BP>155 and male

```
TenYearCHD                                                    cover
       0.08 when age <  51                                      55%
       0.20 when age >= 51 & sysBP <  155                       35%
       0.29 when age >= 51 & sysBP >= 155 & Gender is 0          7%
       0.52 when age >= 51 & sysBP >= 155 & Gender is 1          3%
```



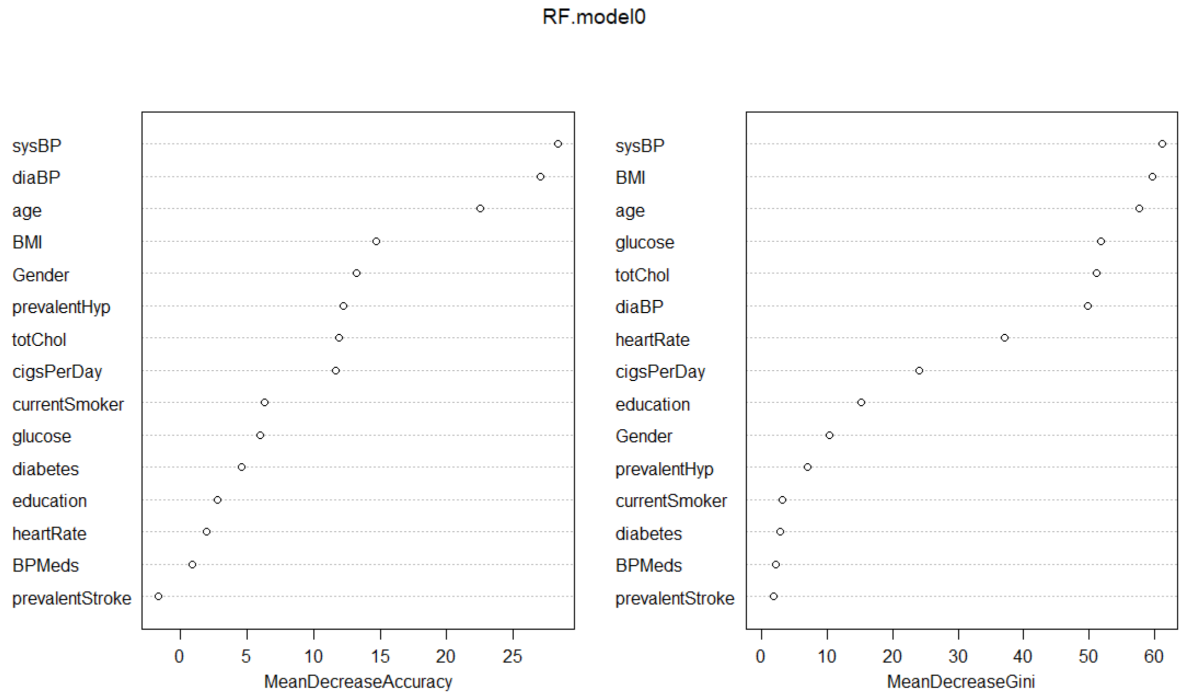
Rattle 2020-Feb-20 17:06:08 Rohan

Model-
CART

RANDOM FOREST

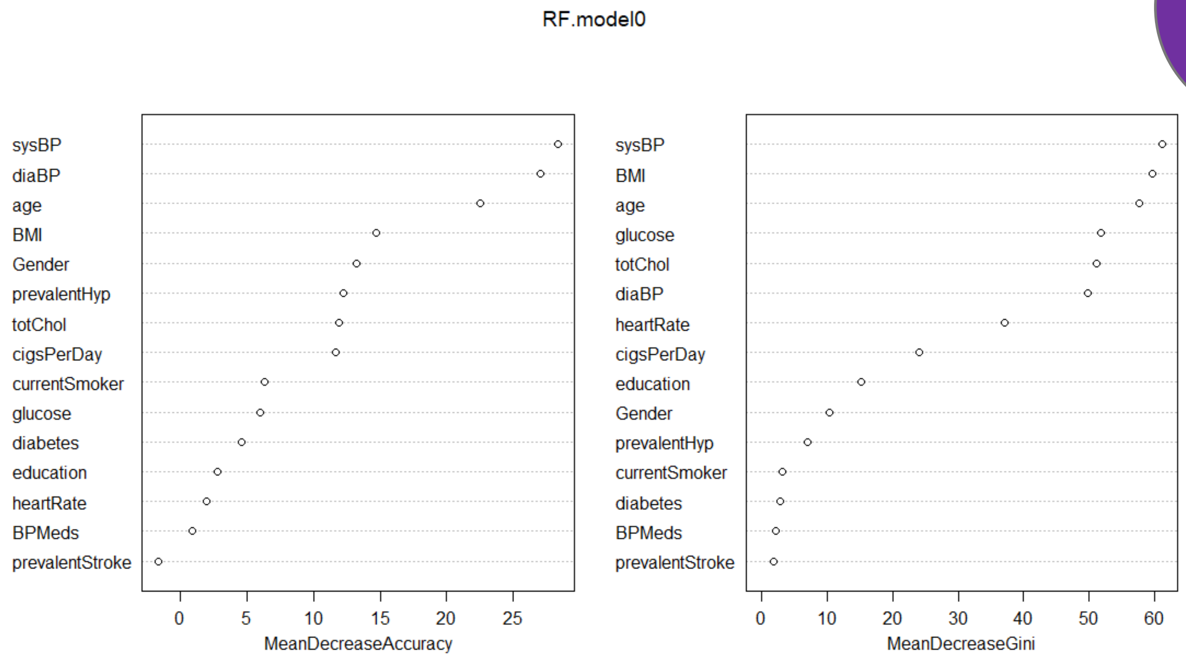
Reason to choose: Random forest has ability to bag no of variables as well as rows and hence was attempted as an ensemble algorithm. Also it is less sensitive to outliers

A general random forest model was run with all variables. Based on Mean decrease of accuracy and GINI. Few variables were dropped. (totChol, currentSmoker, BPMeds, prevalentStroke)

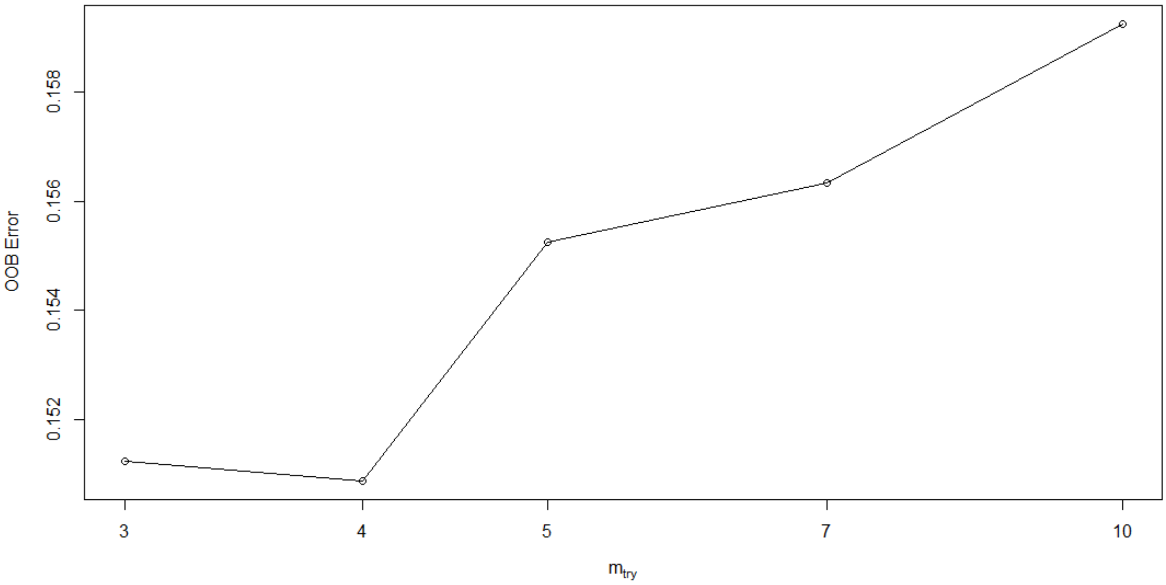


After dropping the variables, we tuned the model & a new model was done of 1001 trees and mtry of 4 (4 variables at a time) and the improvement was seen in the variables

Model- Random Forest



The Out of bag error (OOB) improved from 15.93% to 15.16% at mtry of 4



On the training model, the confusion matrix improved from full model .

Caret package was used to predict the Model scores on train and test

Train scores

Model-
Random
Forest

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2331    0
         1  230  183

               Accuracy : 0.9162
                 95% CI : (0.9052, 0.9263)
    No Information Rate : 0.9333
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.5748

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9102
            Specificity : 1.0000
```

<u>Conclusion:</u> Random forest models showed a **lot of overfitting** in the train data where specificity of over 99% is achieved but in most test data the specificity **<u>drops to near 89%. This is a significant improvement from log model</u>**

```
          Reference
Prediction   0   1
         0 764   6
         1  98  48
               Accuracy : 0.8865
                 95% CI : (0.8641, 0.9063)
    No Information Rate : 0.941
    P-Value [Acc > NIR] : 1

            Sensitivity : 0.8863
            Specificity : 0.8889
```

KNN

Reason to choose: KNN is a simple algorithm for clustering. KNN being a distance-based clustering algorithm it is usually insensitive to outliers. Since people susceptible to CHD will demonstrate more outliers, knn can be considered as an algorithm to predict CHD. But KNN is a classification algorithm that cannot show probabilities and hence cannot be tuned further

- Step 1: data is scaled so as to remove influence of differential units in data variables
- Step 2: Scaled data is presented to KNN

The Confusion matrix of KNN is as follows

```
predknn   0   1
      0 716 129
      1  49  21
```

Accuracy of KNN model: 80%

Specificity of KNN Model: 14%

Conclusion:

KNN did not show significant accuracy or improvement in specificity and hence may not be recommended

Model-
Random
Forest

Naïve Bayes

Reason to choose: Naïve bayes is a probability-based clustering algorithm it is usually insensitive to outliers as data is always binned and in categorical format.

- Step 1: Numerical data is converted to Categorical data. The categories are chosen to reflect the medical accepted guidelines in blood pressure, cholesterol, BMI, heart rate& glucose. Age and cigs/day are categorized as per the histogram bins

Age	<30	30-50	50-60	>60
Category	Young	Mid age	Mature	Old
Cigs/Day	0	1-10	10-20	>20
Category	Non smoker	Average	High smoker	Chain Smoker
Tot. Cholesterol	<200	200-240	>240	
Category	Normal	Border High	High	
Systolic BP	<120	120-130	130-140	>140
Category	Normal	Elevated	Border High	High
Diastolic BP	<80	80-85	85-90	>90
Category	Normal	Elevated	Border High	High
BMI	<18.5	18.5-25	25	
Category	Underweight	Healthy	Overweight	
Heart rate	<60	60-80	>80	
Category	Athletic	Regular	Elevated	
Glucose	<120	120-140	140-150	>150
Category	Normal	Prediabetic	Diabetic	High Glucose

- Step 2: the categorical data is presented to Naïve Bayes model
- Step3: alternatively R was allowed to do random binning at 4 levels in all numerical variables

Conclusion: NAÏVE bayes showed 79% accuracy on train and 33% on Specificity parameters

```
Confusion Matrix and Statistics
          Reference
Prediction    0    1
         0 2153  324
         1  287  163
                                          
               Accuracy : 0.7913          
                 95% CI : (0.7761, 0.8059)
    No Information Rate : 0.8336          
    P-Value [Acc > NIR] : 1.0000          
                                          
                  Kappa : 0.2239          
                                          
 Mcnemar's Test P-Value : 0.1453          
                                          
            Sensitivity : 0.8824          
            Specificity : 0.3347          
```

Model-
Naive
Bayes

11. Model Comparison

- Not just accuracy but capturing the sensitivity and Specificity are important metrics to capture model performance as the objective was to capture True Positive and True Negative
- Random Forest & Logistic regression models were most consistent models in predicting outcomes based on ROC and Specificity

	ROC	Sens	Spec
GLM	0.734	0.856	0.6923
CART	0.640	0.866	0.5385
GLM+CART (Ensemble)	0.642	0.8576	0.5208
Random Forest	0.899	0.8863	0.8889

Discussion

- 5 models were tested to predict ten-year CHD- Logistic regression, CART, Random forest, KNN, Naïve Bayes
- 2 models have showed good performance in prediction.
- Logistic regression had 86% accuracy and 75% prediction accuracy for ten-year CHD events (specificity). It even gives a formula with scores given to predict log odds and influence of each variable. It however has to drop a few variables for prediction. The Mcfadden r Square score shows it to be a moderate model
- Random Forest had 88.6% accuracy and 88.9% prediction accuracy for ten-year CHD events (specificity). Random forest gives probabilities for the data and hence can be further tuned based on more data but has a challenge of overfitting the data
- Naïve Bayes showed the lowest performance with a 33% specificity on train scores. Since this is a parametric algorithm, we expect the results to hold

12. Final Conclusion

The factors/ patient segments most at risk for ten-year CHD are as follows

- Demographic wise:
  - Age>49, every one-unit increase, increases the odds likelihood by 7.3%
  - Gender- Male, odds likelihood of 63%
- Behavioural
  - Cigarettes /Day: Among Smokers every one unit increase, increases odds likelihood by 1.9%
- Medical History
  - Prevalent Hypertension: increases odds likelihood of 27.5%
- Medical (current value)
  - Systolic BP: Among hypertensives every TEN unit increase, increases odds likelihood by 13%

Conclusion

  - Glucose: Among pre/diabetics every TEN unit increase, increases odds likelihood by 7%

Business insight:

- Some variables were shown to be very high predictors of CHD in all models. Notably, systolic BP, Age above 49, Gender- male came out strongly
- Key recommendations would be to drop the Systolic Blood pressure to normal levels, reduce glucose levels in blood with regular screening, reduce Smoking and screen regularly for these risk factors especially in men
- This insight an be used to make policy decision for health-related outcomes by various organization including government to do continuous monitoring of these risk groups and reduce blood pressure or glucose parameters.
- Since the model is used for long time horizon to predict ten-year CHD, the 3 models can have different utility in different scenarios
- If the business objective is to give out a prediction score based on the variables presented by a patient, a logistic model will be fast as it is a simple equation-based model. The use of this can be in initial Risk score assessment on 1st time cases
- Fitness Apps can use the Random forest algorithm which is compute resource intensive to do calculation on cloud and keep sensitizing patients on real time updates on health parameters
- Health organization and pharma companies can work towards keeping drugs to maintain health vitals below threshold decided

13. Recommendations and Business Applications

- Men>50 should be regularly screened for health parameters to be kept under check
- The benchmark vitals for measurement of 10 yr CHD risk are systolic BP>150, Age>50 , high glucose and BMI
- Programs concerning Adherence to BP medications, smoking cessation, weight reduction can reduce risk of CHD in risk groups
- Fitness apps should provide real time feedback to patients with high risk of CHD
- Increased awareness is required for patients with lower education demographic
- A Disease risk score should be created at every possible health screening as initial health assessment for filtering individuals for further test

- Applications
- For a young nation like India, the model can be used to make health policies that will lower Cardiovascular risk factors particularly in lower educated demographic
- CV Score can be assessed by insurance and Corporate sector to help individuals have early warning signals and prevent CHD
- Fitness apps can create an ecosystem of Drs, caregivers, wearable devices to monitor high risk profile individuals and provide emergency systems based on real time dat
- Pathology Labs, Pharmaceutical drug companies and Health organization can create another ecosystem to measure drug performance and efficacy on high risk individuals and check if early intervention can help prevent CHD

Conclusion