CAPSTONE PROJECT

CORONARY HEART RISK STUDY

GREAT LEARNING – 2019-2020 BATCH III

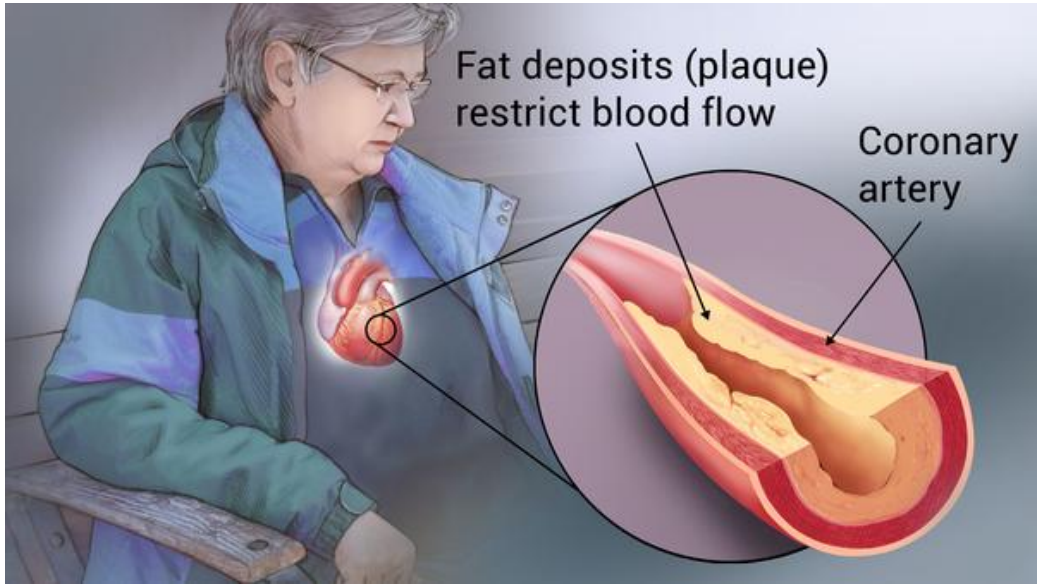
BY SWARUP KUMAR

BOSCH

Capstone : CORONARY HEART RISK STUDY

Business Problem Understanding

- **What : Coronary** artery **disease** develops when the major blood vessels that supply your **heart** with blood, oxygen and nutrients (**coronary** arteries) become damaged or diseased. In today's world of changing lifestyle, Coronary heart disease has become one of the key diseases to tackle.
- With the dataset provided, need to predict factors that causes people to develop Coronary heart disease in the next 10 years
- **Why:** Better treatment to patients can be provided by doctors, based on their habits & medical condition



- **Why: There is no cure for CHD**. More lives can be saved if preventive actions are taken at early stage or corrective actions if at extreme level.

Capstone : CORONARY HEART RISK STUDY

Business Problem Understanding

- Constraints
  - Dataset provided is not big, only **4240** rows of which around **645** cells with null or NA values
    - **40** rows have NA values in both glucose and cholesterol field
  - Close to **50**% of the total variables are Factorial in nature (sparse of 0,1)
    - previous Hypertension, gender, current Smoker, previous Stroke, BP Medication, diabetic
  - 85%-15% is the distribution of Yes and No in Dependent variable (TenYearCHD)

- Assumptions
  - Education field mapping is 1->Illiterate,2 ->High School, 3 -> University, 4 -> Post graduate
  - All the data is collected over a period of time from 1 common region
  - Insights to be provided based on the factors in the dataset ONLY (no extrapolation)

Capstone : CORONARY HEART RISK STUDY

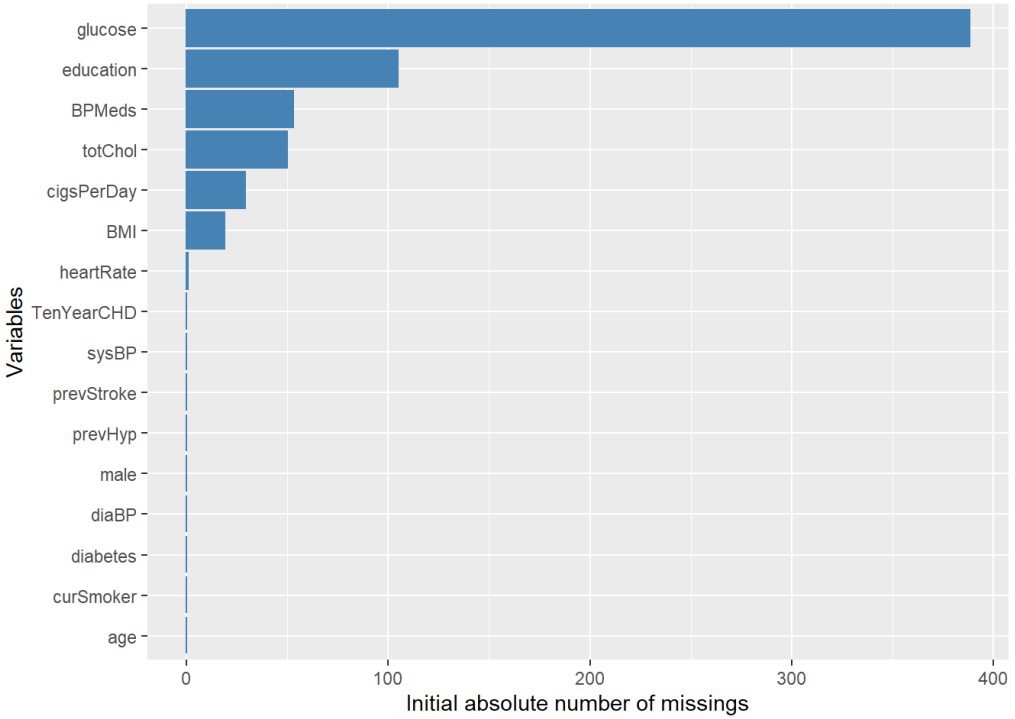
Business Problem Understanding

- Objectives
  - Minimize the fatalities due to CHD i.e.
    - Correctly Predicting + ve cases of CHD (True Positive)
    - Correctly Predicting – ve cases of CHD (True Negative)
    - **[Type II Error]** Cost of incorrect predicting of + ve cases of CHD is Higher than cost of incorrect predicting of – ve cases**[Type I Error]**

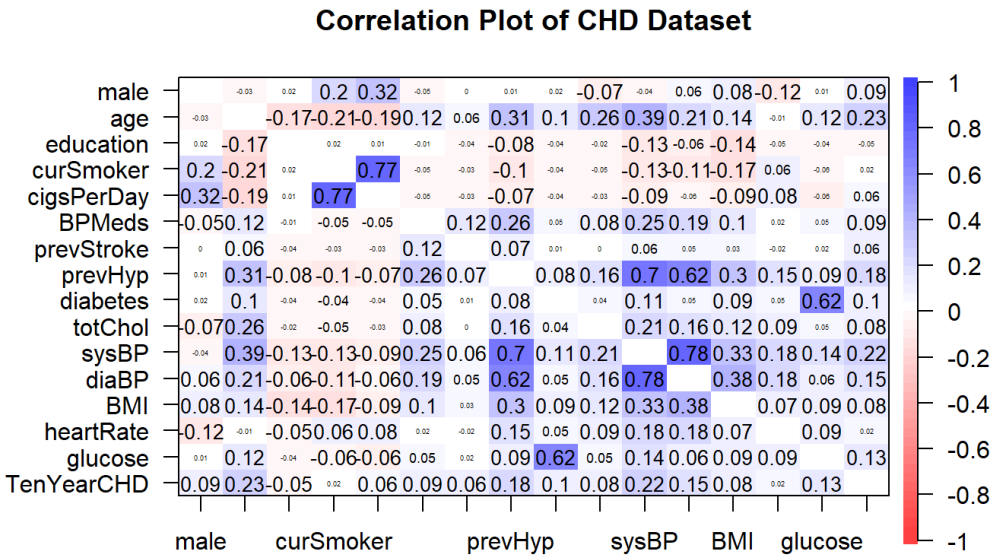
- Deliverable
  - Exploratory data analysis of dataset
  - Come up with best predictive classification model on all parameters like
    - **Sensitivity(Type II)**, Precision(True +ve/Total +ve), Kappa(chance agreement), F1(FN,FP) values

Capstone : CORONARY HEART RISK STUDY

EDA (1/2)



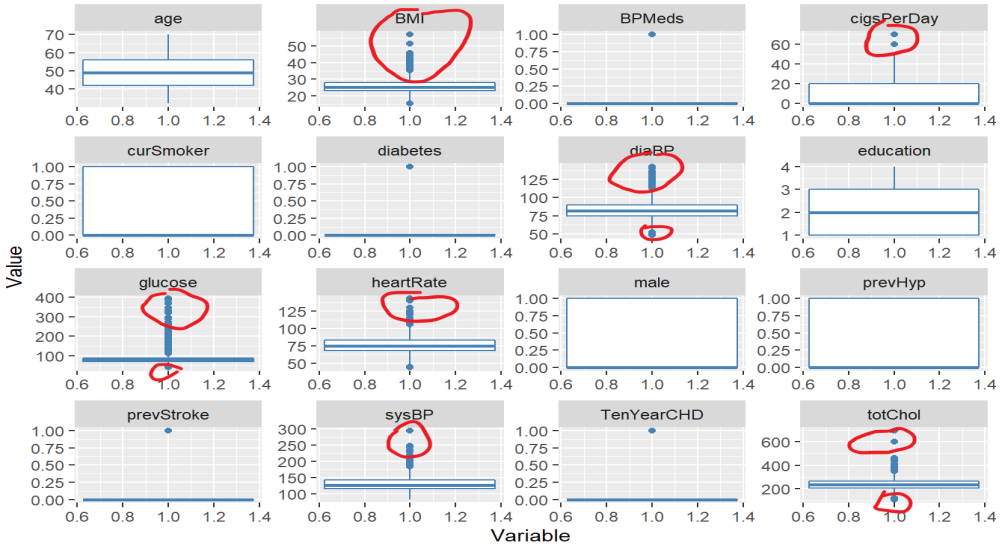
glucose	education	BPMeds	totChol	BMI	heartrate	cigsPerDay	Total
388	105	53	50	19	1	29	645

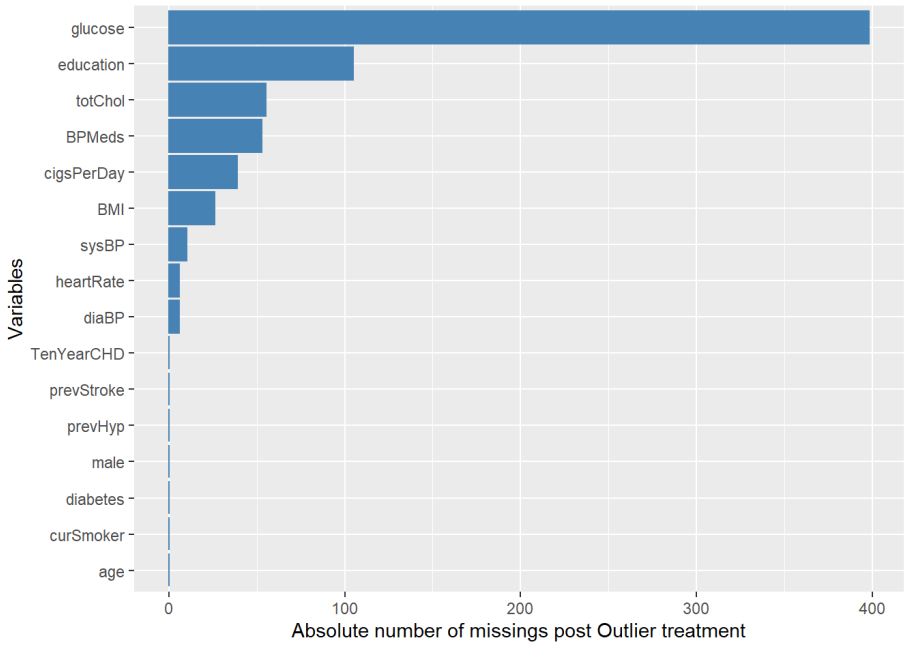


1. No correlation found > 92%
2. Multicollinearity doesn't exist and hence PCA and other treatment not required

Capstone : CORONARY HEART RISK STUDY

EDA (2/2)





- Outliers exist for numerical variables like:

  BMI, cigsPerDay, diaBP, glucose, heartrate, sysBP, totChol

- Find outliers using cooks distance and then replacing them with NA

glucose	sysBP	diaBP	totChol	BMI	heartrate	cigsPerDay	Total
398	10	6	55	26	1	39	698

Capstone : CORONARY HEART RISK STUDY

Modelling approach used and Why (1/2)

Outlier Treatment
- Cooks distance method
- Replace outlier with NA

⇒

NA Value treatment
- NAs are missing at Random, hence MICE package will be used

⇒

Binning
- Body Mass Index and Age are generally mentioned in Ranges
- Binning for age, BMI fields

⇒

Test, train split
- 70-30 split keeping proportion of Dependent variable in mind

⇒

Train Control
- Resample = 10 fold, 3 repeat cross validation to reduce Overfit
- Additional sampling 'smote'

⇒

Hyper parameter tuning
- Metric = ROC
- tuneLength = 10

⇒

Model Execution + Predict
- Predict on test
- Print Metrics
- varImp for top variables
- confusionMatrix for TP,FP,FN,TN

⇩

4 datasets created:
1. **Outlier capping(5%-95%) + replace NA using dlookr**
2. **Dropping all NA values and no outlier treatment**
3. **Cooks distance + replace NA using mForest**
4. **Cooks distance + replace NA using mice**

All 4 datasets shall be passed to various algorithms and the best model will be selected

⇩

Criteria for model selection:
1. **Highest Sensitivity value +**
2. **Higher Precision +**
3. **Kappa value > 0.2 +**
4. **Higher F1 +**
5. **Higher Accuracy**

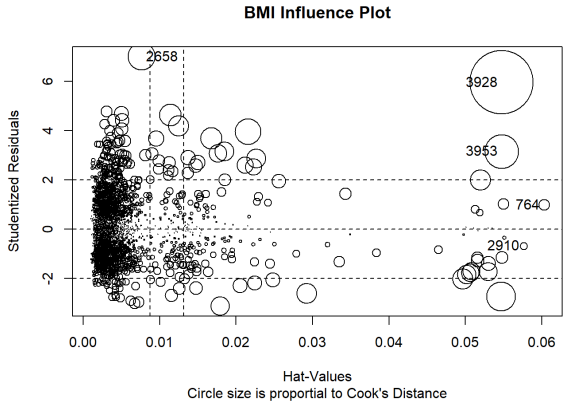
Capstone : CORONARY HEART RISK STUDY

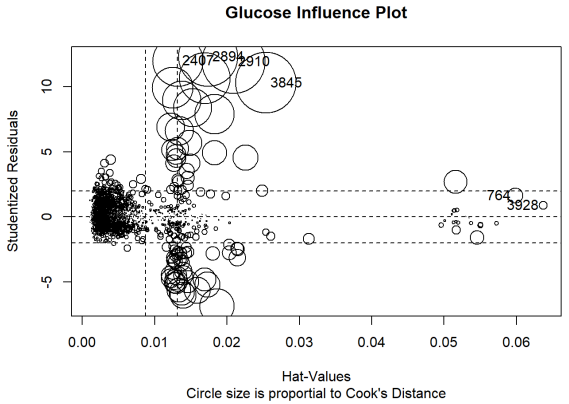
Modelling approach used and Why(2/2)

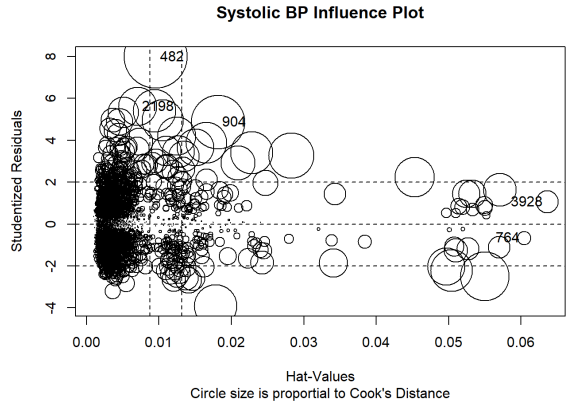
Sl No	Modelling Technique	Why	Remarks
1	**Logarithm Regression (with boost)**	Binary classification	
2	**Naïve Bayes**	Binary classification, more factorial variables, faster	Learns over time
3	**KNN**	K Neighbours (for comparison purpose)	Since outliers are treated can be used or else performance reduces
4	**Random Forest**	To get more insights like variable importance	Can handle missing values automatically and
5	**SVM Classification**	Works well with small data and for imbalanced data	
6	**eXtreme Gradient Boosting**	Ensemble method for quick implementation, extreme computation limits for scalability	Trees are built in series and compared on weighted leaf scores
7	**Gradient Boosting**	Ensemble method	
8	**CTREE**	To get decision nodes	Similar insights like CART
9	**CART**	To get decision nodes	Comparatively better performance than CTREE

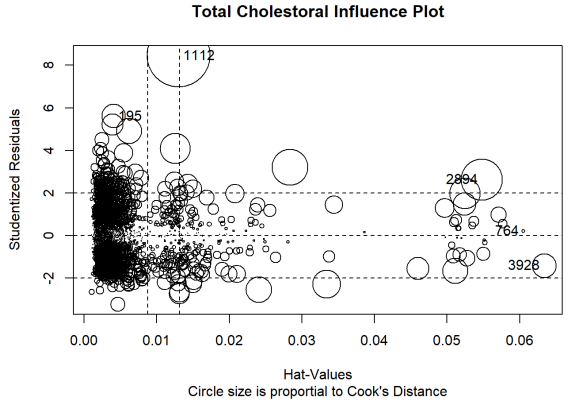
Capstone : CORONARY HEART RISK STUDY

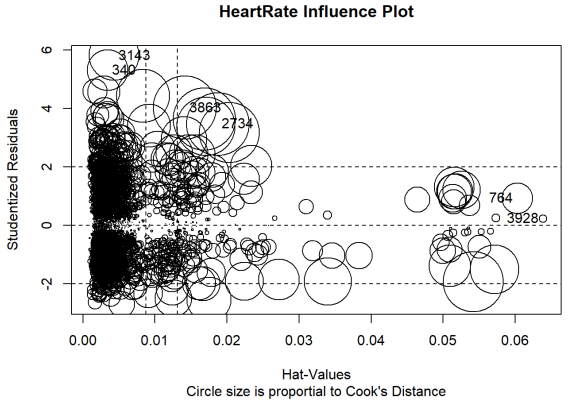
Cook's distance Outlier + NA treatment (mice)

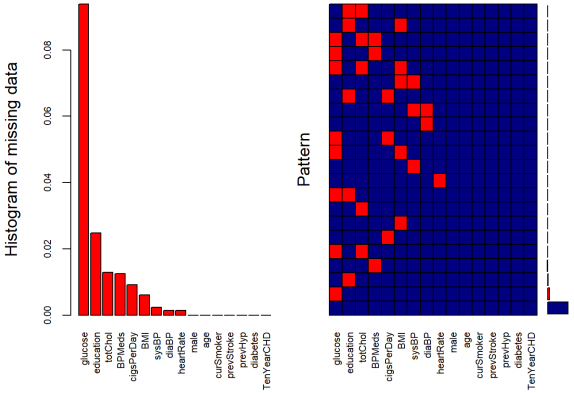










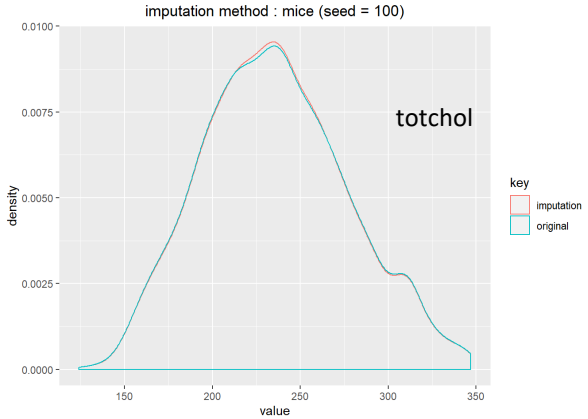


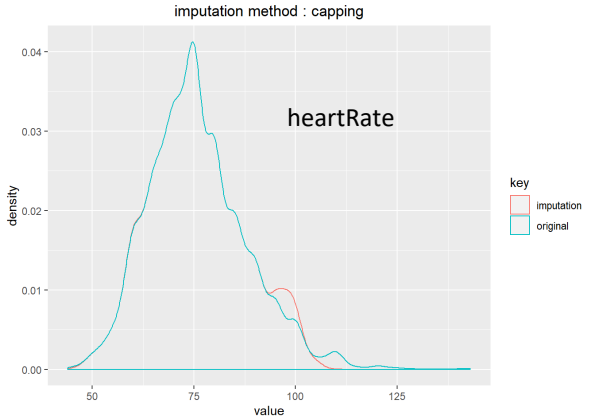
BOSCH

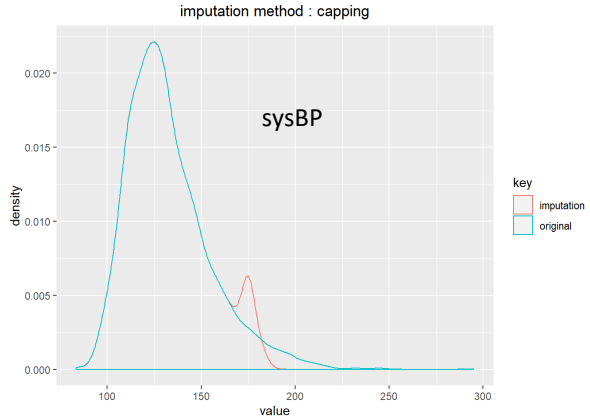
Capstone : CORONARY HEART RISK STUDY

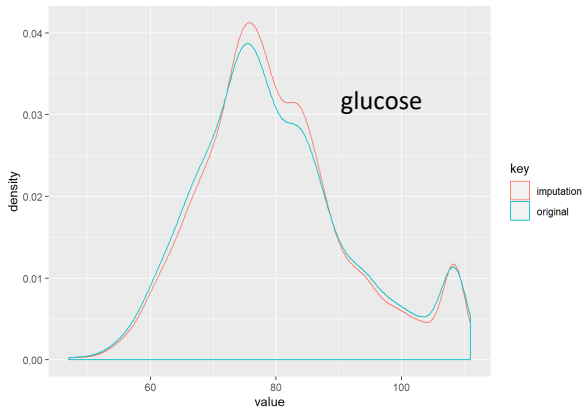
Jan-2020

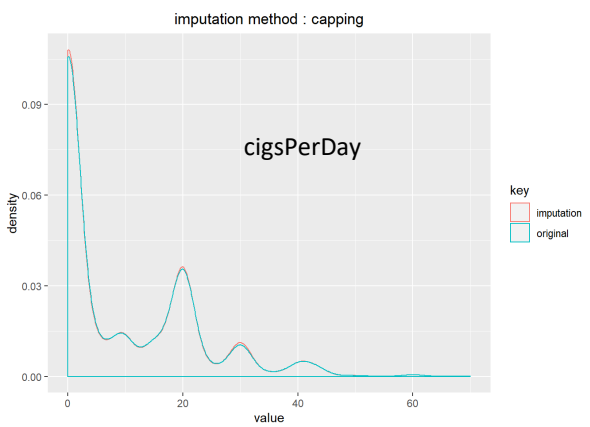
Outlier capping + mice imputation of NA

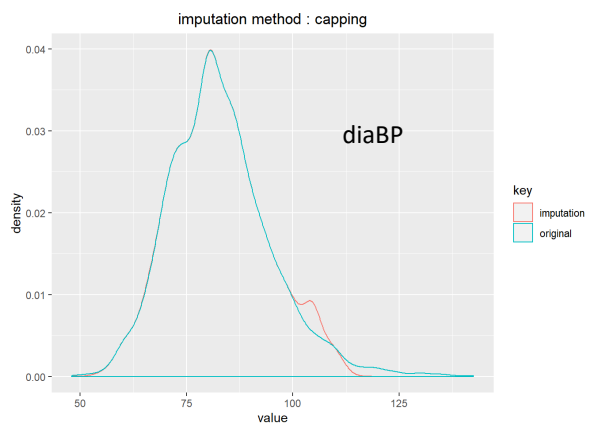












BOSCH

Capstone : CORONARY HEART RISK STUDY

Observations from Model execution

- No models were overfitting when outlier capping was done, CV of 10 fold helped
- Random forest tends to overfit when dropped NAs dataset is passed
- KNN and SVM were overfitting on dataset with cooks distance treated outliers
- Gradient Boosting was the best performing model on dataset with dropped NA values and dataset with capped outliers
- CTREE model gave highest sensitivity while using mForest method but Kappa value < 0.2
- xgBoost was overfitting maybe due to insufficient tuning of hyper parameters and scaling of variables
- Gradient boosting took comparatively less time for model execution on capped outliers dataset
- Naïve Bayes performed comparatively well on all datasets except on dropped NAs

Note: Refer to Appendix for Source code and HTML output

BOSCH

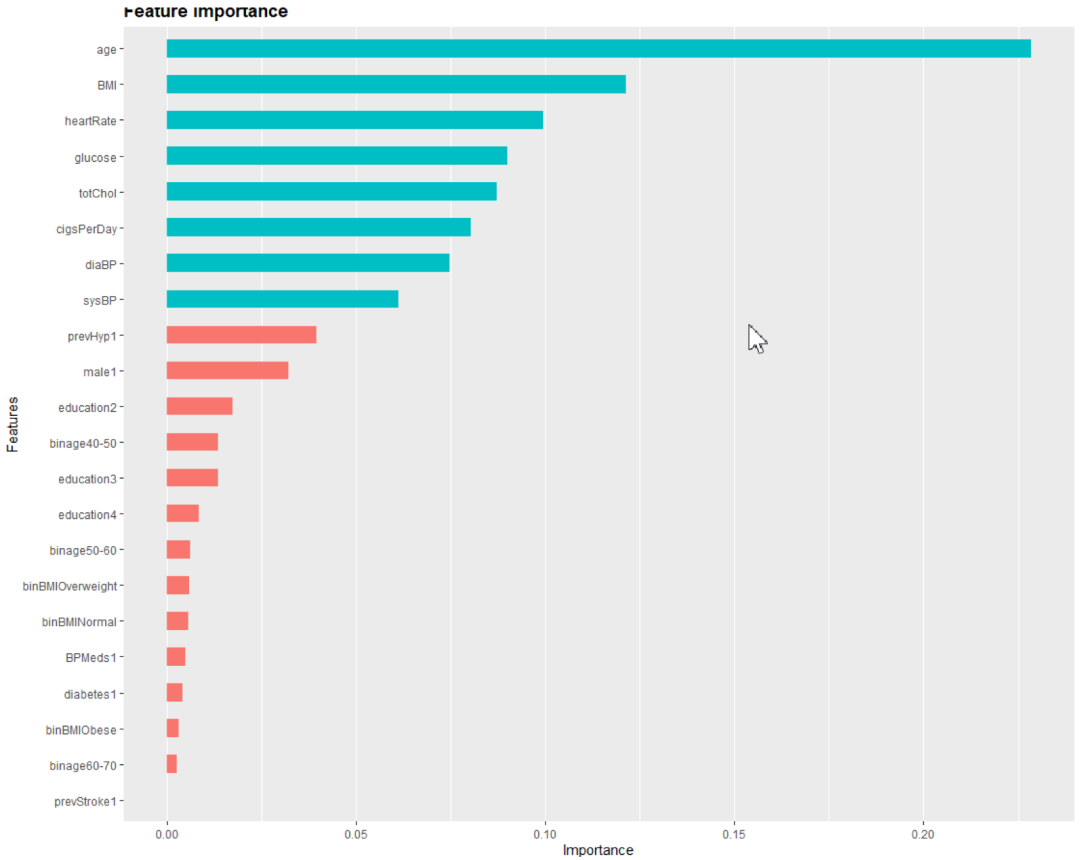
Capstone : CORONARY HEART RISK STUDY

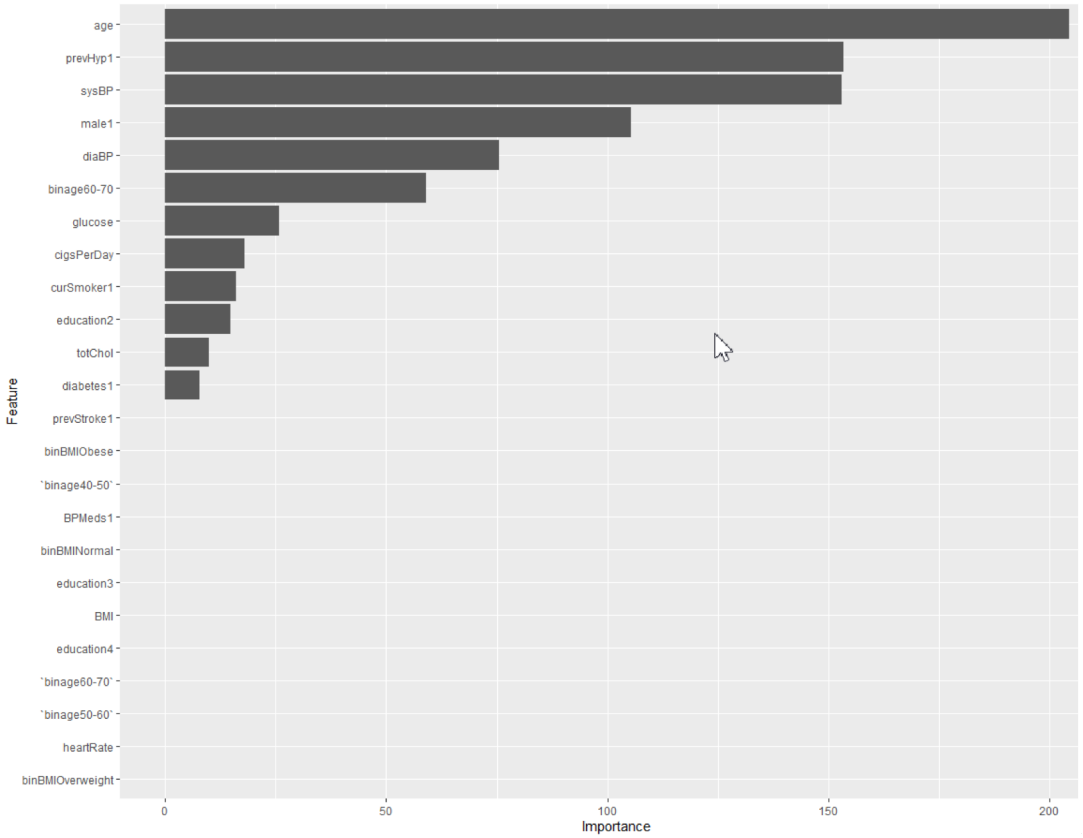
Confusion Matrix

		0	1
CTREE	0	811	95
	1	267	98
Naïve Bayes	0	832	90
	1	246	103
KNN	0	738	90
	1	340	103
CART	0	830	103
	1	248	90
GBM	**0**	**872**	**89**
	1	**206**	**104**
XGBoost	0	955	123
	1	123	70

Capstone : CORONARY HEART RISK STUDY

varImp Plots from various Modelling Technique





Capstone : CORONARY HEART RISK STUDY

Model Metrics

Technique	Sensitivity	Precision	Kappa	F1	Specificity	Accuracy	Exec Time (min)
Logit Boost	0.31	0.24	0.12	0.27	0.82	0.74	1.13
SVM	0.52	0.19	0.07	0.28	0.6	0.59	4.97
Naïve Bayes	0.53	0.28	0.21	0.36	0.75	0.72	0.42
KNN	0.57	0.25	0.17	0.34	0.69	0.67	2.60
xgBoost	0.4	0.34	0.24	0.37	0.85	0.78	0.42
Random Forest	0.4	0.31	0.22	0.35	0.84	0.77	2.13
CTREE	0.51	0.27	0.19	0.35	0.75	0.71	2.70
CART	0.46	0.26	0.19	0.34	0.77	0.72	0.25
GBM	**0.54**	**0.33**	**0.28**	**0.41**	**0.8**	**0.76**	**9.14**

Note: Refer to (Annex 1)ModelMetrics_alldatasets.xlsx for consolidated values from all datasets

BOSCH

Capstone : CORONARY HEART RISK STUDY

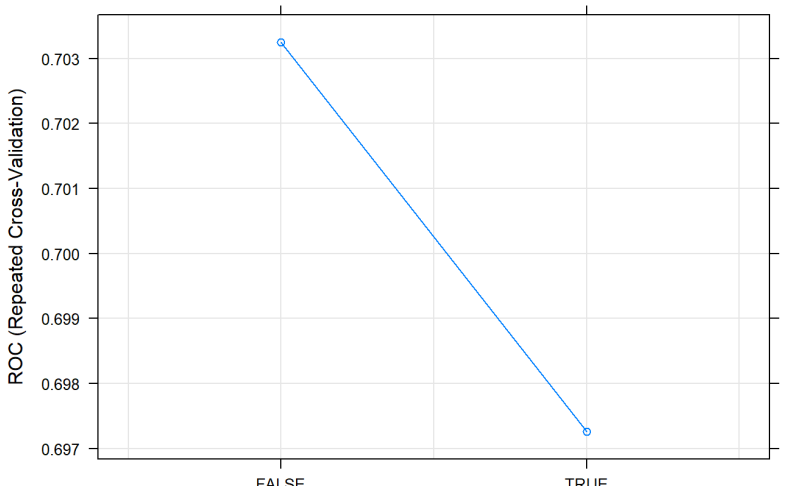
Best Performing model and metrics

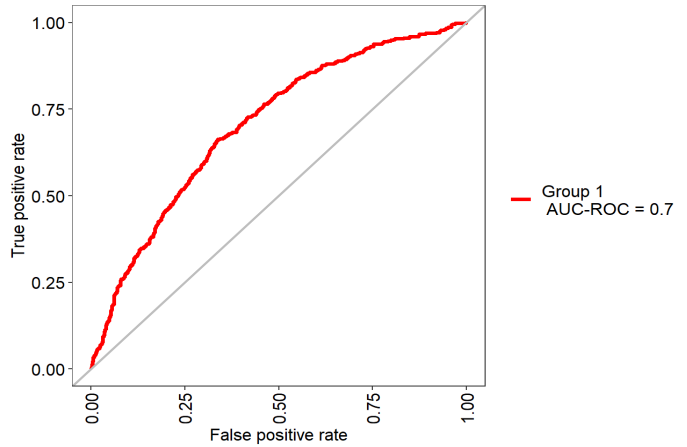
- Naïve Bayes is the preferred model on most datasets
- Confusion Matrix

	0(-ve)	1(+ve)
0(-ve)	832(TN)	90(FP)
1(+ve)	246(FN)	103(TP)

Other Metrics	Value
Sensitivity/Recall	**0.53**
Precision	0.29
Kappa	0.23
F1	0.38
Specificity	0.77
Accuracy	0.73
Model Execution time	0.62 min

- ROC





Capstone : CORONARY HEART RISK STUDY

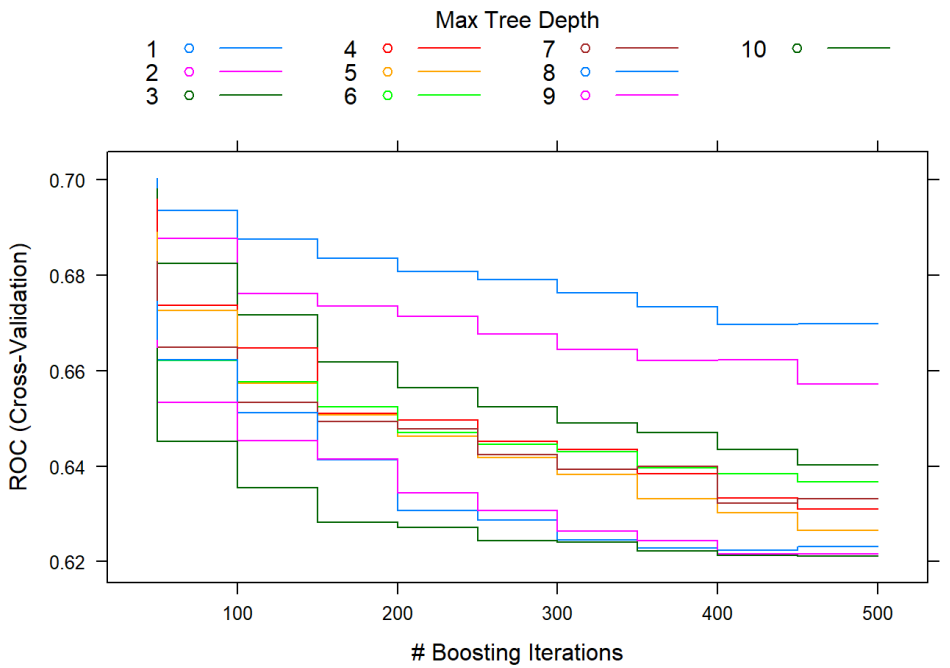
Best Performing model and metrics

- **Gradient Boosting** was the best performing model
- Confusion Matrix

	0(-ve)	1(+ve)
0(-ve)	872(TN)	89(FP)
1(+ve)	206(FN)	104(TP)

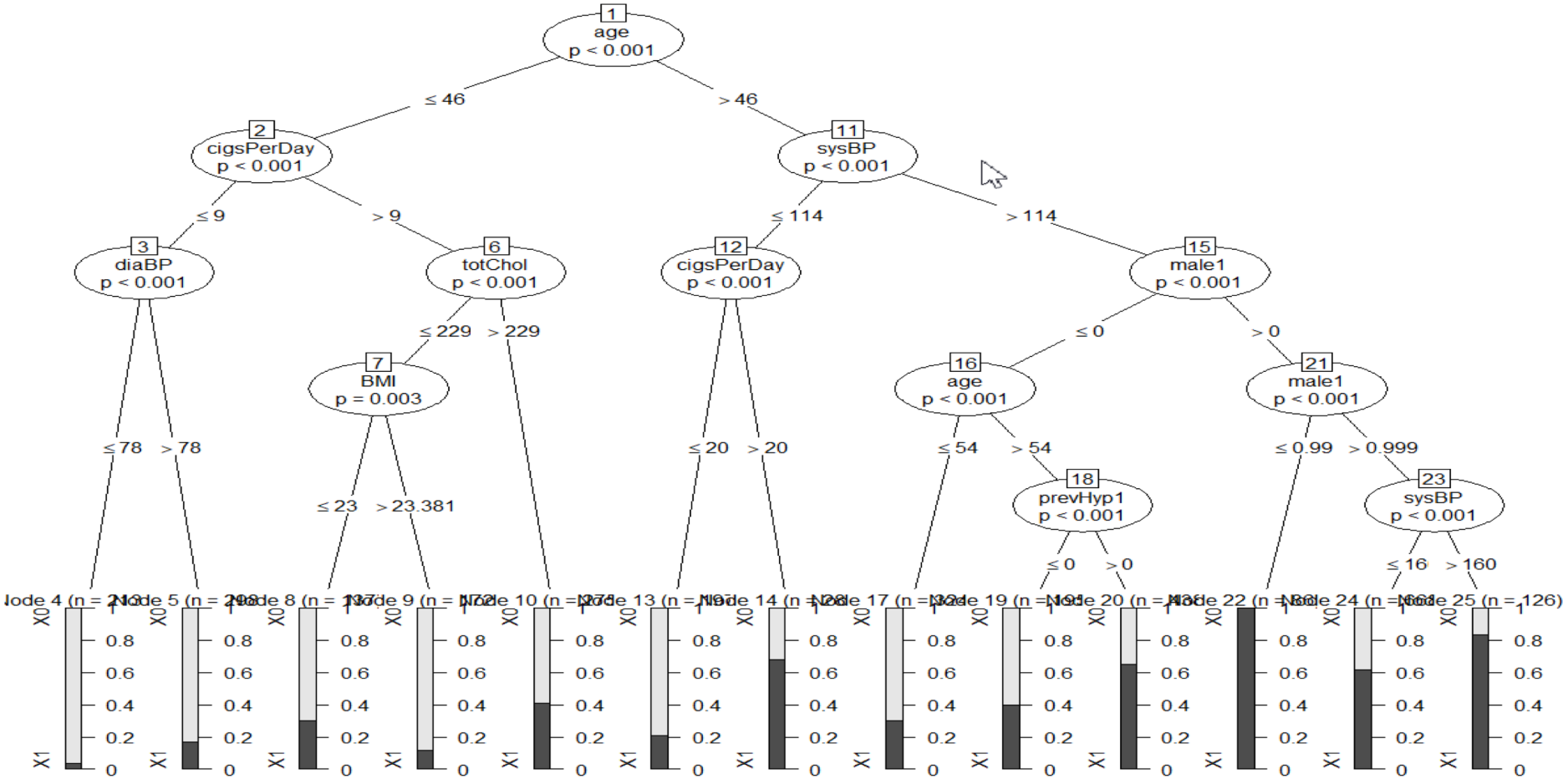
Other Metrics	Value
Sensitivity/Recall	**0.54**
Precision	0.33
Kappa	0.28
F1	0.41
Specificity	0.8
Accuracy	0.76
Model Execution time	9 min

- ROC

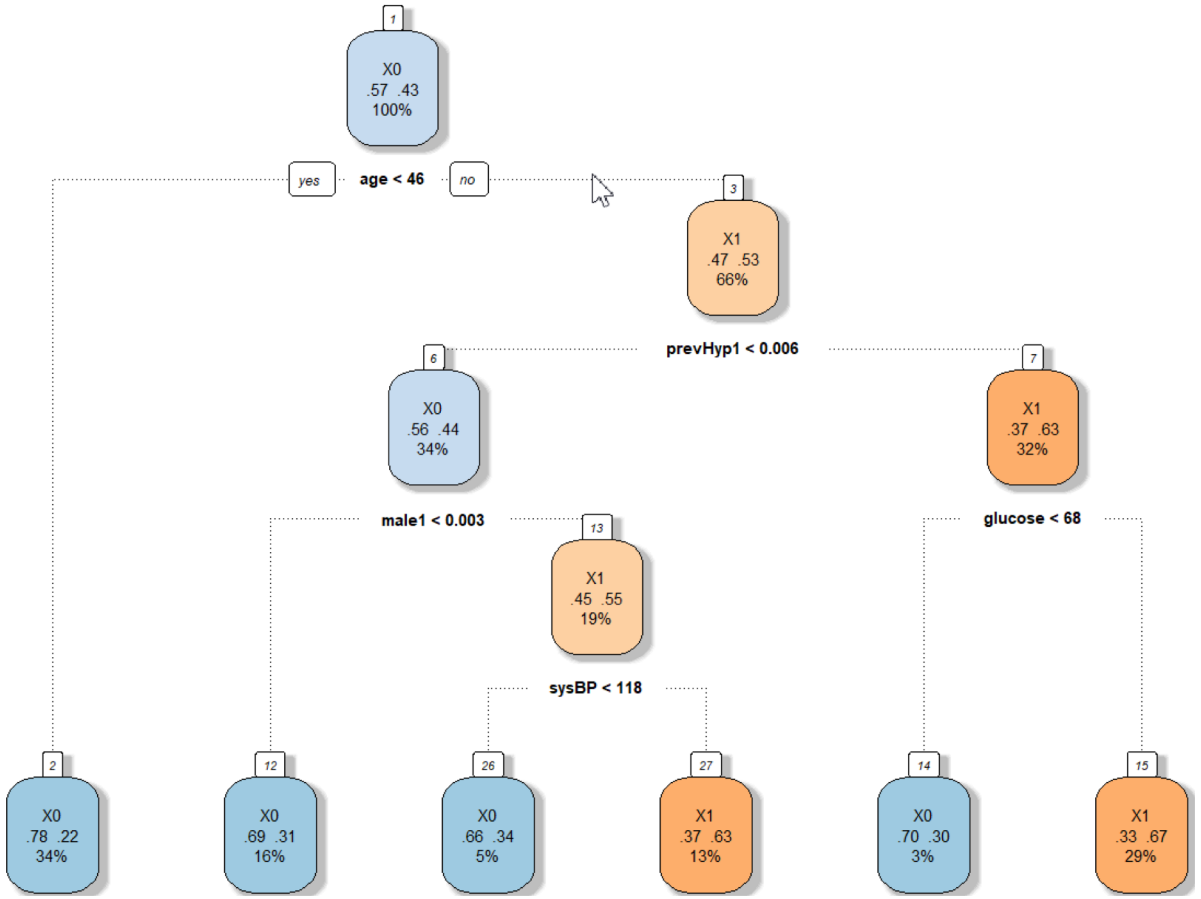


BOSCH

CTREE model



CART model



Capstone : CORONARY HEART RISK STUDY

Insights from Analysis (Top 10 causes of CHD)

- 1. **Age**: As we age, plaque builds up. In our case, age group of 40-50 followed by 60-70 is found to be having higher chances of CHD

- 2. **Sex**: Male / Female.

Men were found to have higher chances of CHD among >46 age group (19%) while Women >54 had higher chances if CHD

- 3. **Systolic blood pressure**: Amount of pressure in the arteries during the contraction of the heart muscle.

Higher systolic BP (>114), higher the value higher chances of CHD

- 4. **Total Cholesterol:** It is the waxy substance found in the blood.

Higher the cholesterol (>229) levels higher the fatty deposits in the blood

- 5. **Glucose**: Is simple sugar, a sub category of Carbohydrates. Over time, higher glucose levels (>68)leads to greater accumulation of fat ,which raises risk of CHD

- 6. **cigsPerDay**: Carbon monoxide, nicotine and other substance in tobacco smoke can promote clumping platelets and then block coronary arteries.

Higher the cigarettes (>9 per day) smoked higher the chances of CHD

- 7. **Diastolic blood pressure**: The pressure in the arteries, when the heart rests and beats. This is the time when the heart fills with blood and gets oxygen.

Higher the value (>90), the person has high blood pressure

- 8. **prevHyp**: Past history of having high blood pressure. People who had past history were found to have higher chances (32%) of CHD

- 9. **Heartrate**: Speed of the heartbeat measured by the number of contractions per minute. Heartrate > 100 higher chances of CHD

- 10. **BMI**: Higher BMI (>24) meant higher cholesterol and high blood pressure. People who were overweight had higher chances of CHD

Capstone : CORONARY HEART RISK STUDY

Jan-2020

Recommendations

- Men above 46 yrs, have to maintain sysBP<160 or else chance of CHD is 80%
- Female above 54 years to maintain sysBP < 114 + hypertension or will have 60% chance of CHD
- People older than 46 yrs with previous history of Hypertension to maintain glucose level under 68 or else chance of CHD is 29%
- People above 46 yrs with sysBP <114 to either quit smoking or reduce (with 20 cigs / day to have 70% chance of CHD)
- People below 46 years of age to maintain cholesterol level under 229 mg/dL & quit or reduce smoking (<9 cigs/day)
  - In addition, people who were Underweight to absolutely quit smoking (30% chance to have CHD)

- Factors that did not have major impact on CHD were:
  - Undergoing BP medication (Yes/No)
  - Has a history of stroke (Y/N)
  - Education field

- Suggestions for data collection
  - Values of important fields like **Cholesterol, glucose** cannot be left blank
  - NA values in education, BP Medication history(Y/N) is of little significance (can be removed totally)
  - Other factors causing CHD to be captured instead for ex: lifestyle, alcohol consumption

Capstone : CORONARY HEART RISK STUDY

Appendix

Sl No	Description	File name
1	Summary of all model metrics	ModelMetrics_alldatasets.xlsx
2	Source code of cookdistance + mice NA	Capstone_HeartDiseaseRisk_cMice.Rmd
3	Html Knit of cookdistance + mice NA	Capstone_HeartDiseaseRisk_cMice.html
4	Source code of capping + mice NA	Capstone_HeartDiseaseRisk_cdlookr.Rmd
5	Html Knit of capping + mice NA	Capstone_HeartDiseaseRisk_cdlookr.html
6	Source code of missForest +cooks distance	Capstone_HeartDiseaseRisk_mForest.Rmd
7	Html Knit of missForest +cooks distance	Capstone_HeartDiseaseRisk_mForest.html
8	Html Knit of dropped NA values	Capstone_HeartDiseaseRisk_dropNA.html