Capstone Project

Cardiovascular Risk Prediction

by

Roshan Jamthe

Points for Discussion

- Problem Statement
- Data Summary
- EDA and Data Visualization
- Correlation Heatmap
- ML Models' Description
- Best model
- Performance Comparison
- Data resampling
- Conclusions

Problem Statement

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

The dataset provides the patients' information. It includes over 3,000 records and 15 attributes. Each attribute is a potential risk factor. There are demographic, behavioral, and medical risk factors.

Data Summary

Demographic:

Sex: male or female("M" or "F")

Age: Age of the patient;(Continuous - the concept of age is continuous)

Behavioral:

is_smoking: whether or not the patient is a current smoker ("YES" or "NO")

Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical (history):

BP Meds: whether or not the patient was on blood pressure medication (Nominal)

Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)

Prevalent Hyp: whether or not the patient was hypertensive (Nominal)

Diabetes: whether or not the patient had diabetes (Nominal)

Data Summary (Contd.)

Medical(current):

Tot Chol: total cholesterol level (Continuous)

Sys BP: systolic blood pressure (Continuous)

Dia BP: diastolic blood pressure (Continuous)

BMI: Body Mass Index (Continuous)

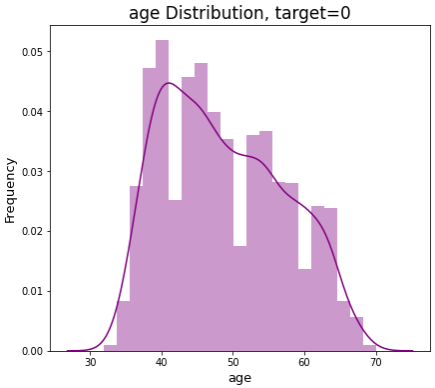
Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

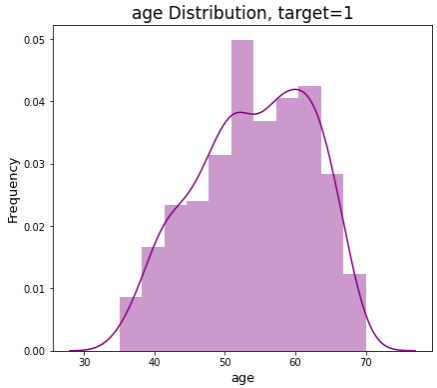
Glucose: glucose level (Continuous)

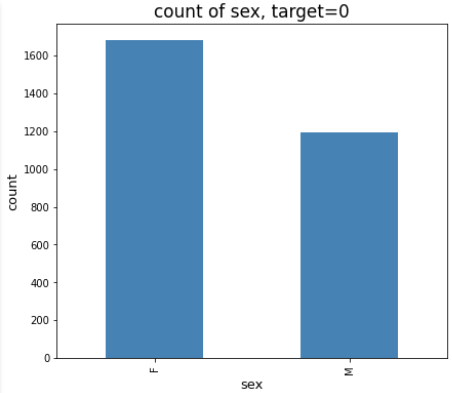
Predict variable (desired target):

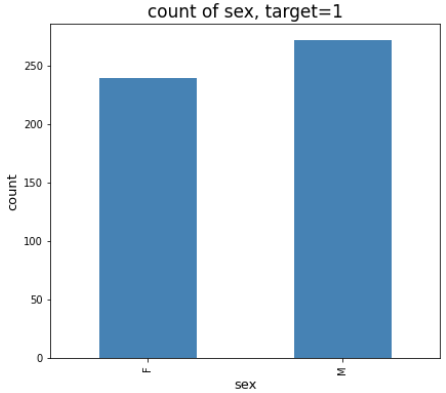
TenYearCHD: 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

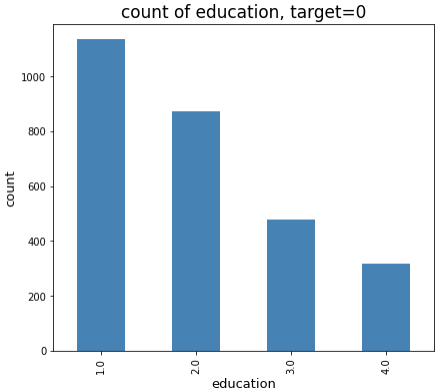
EDA and Data Visualization

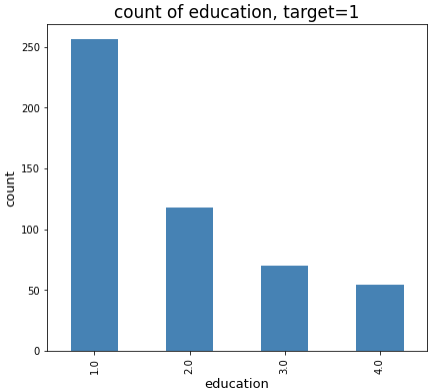


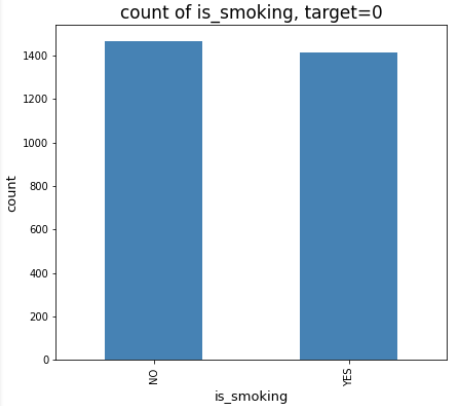


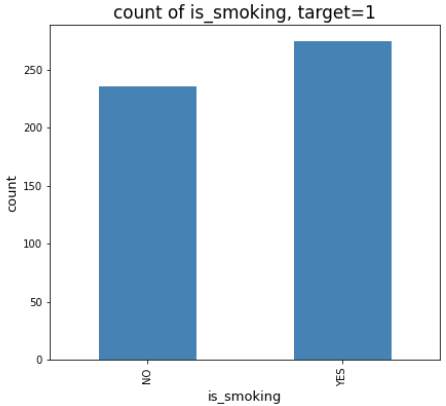




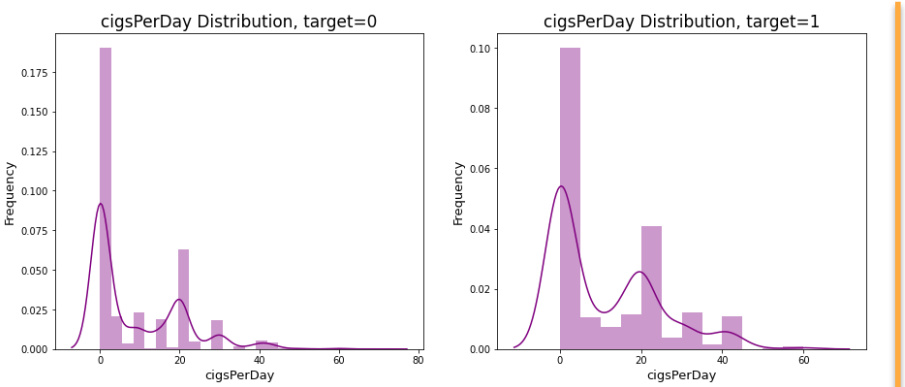


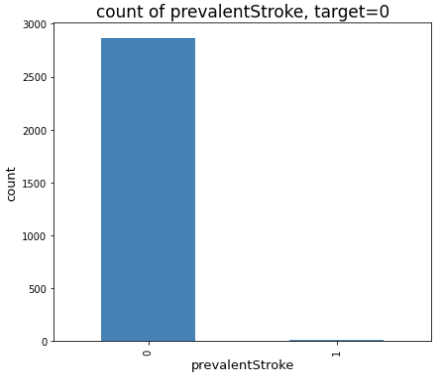


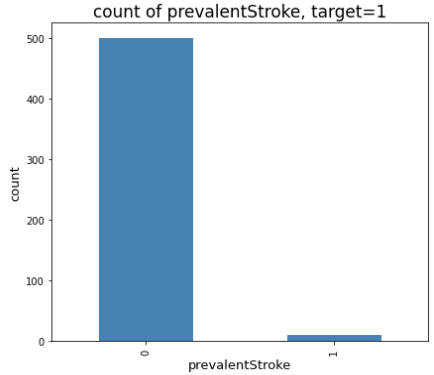


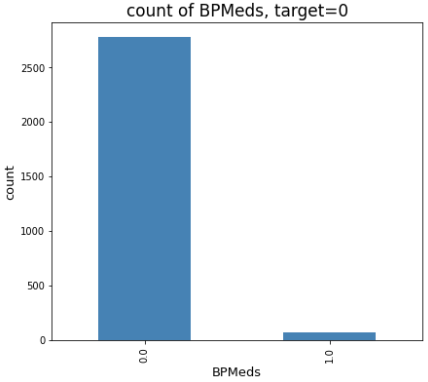


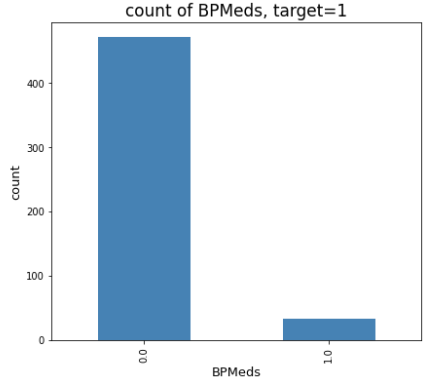
EDA and Data Visualization

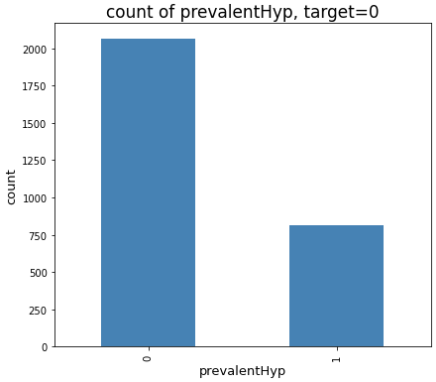


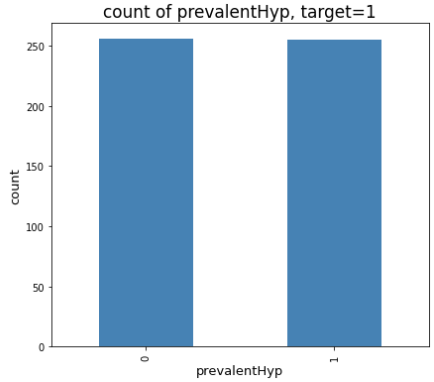




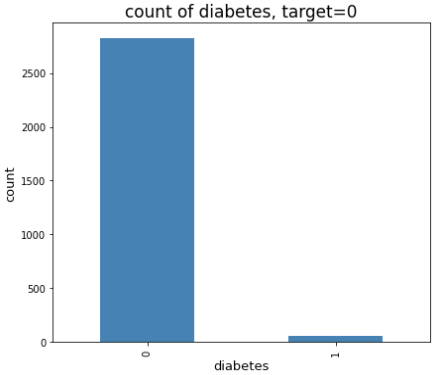


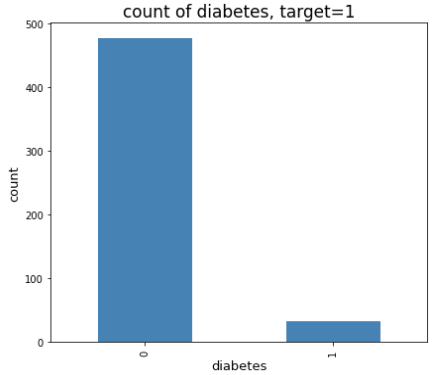


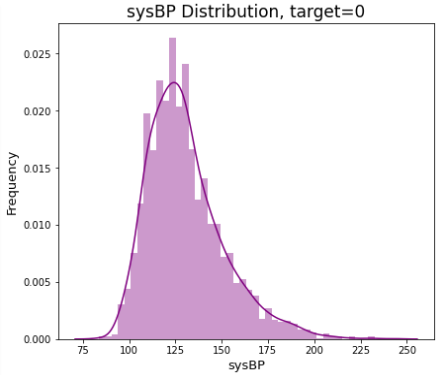


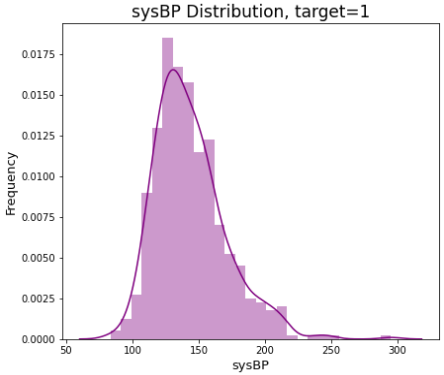


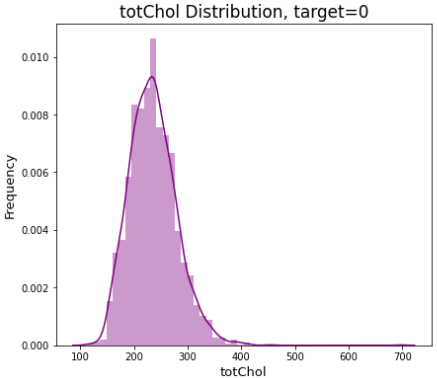
EDA and Data Visualization

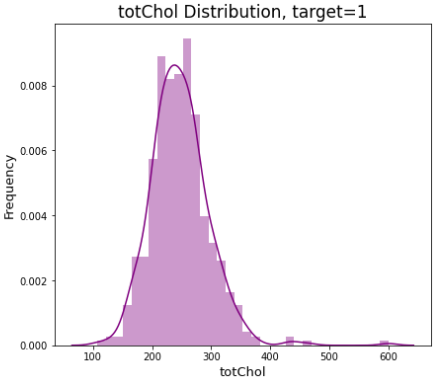


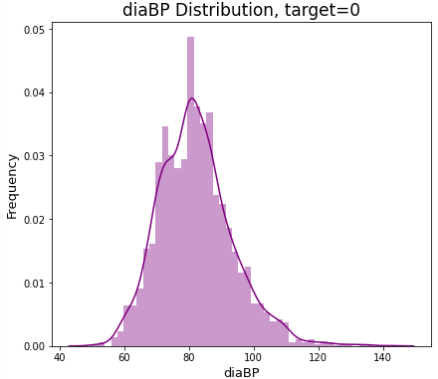


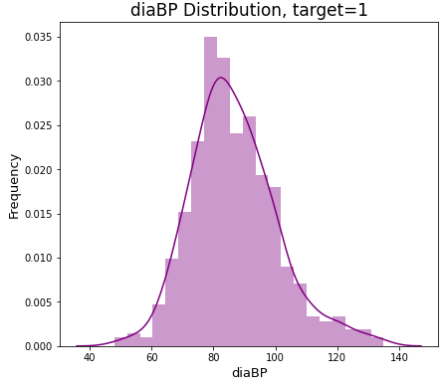




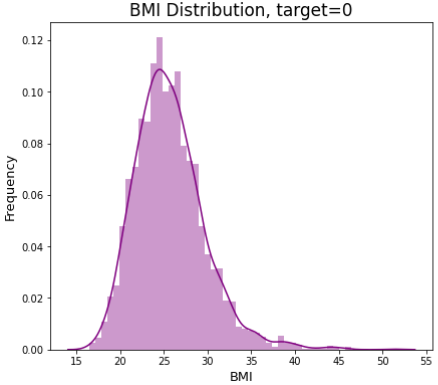


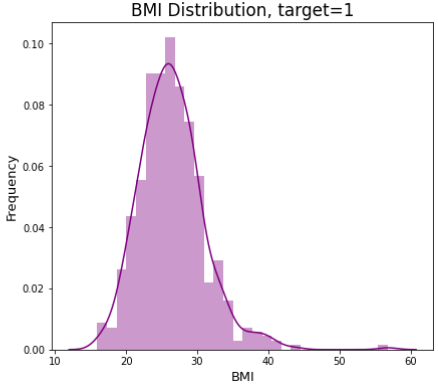


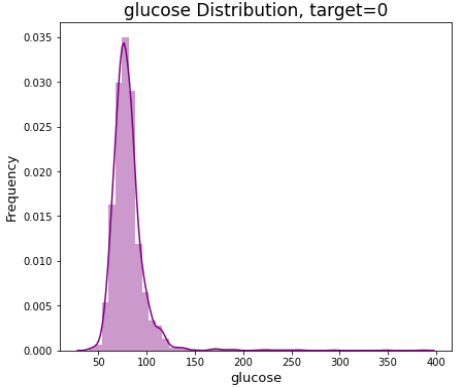


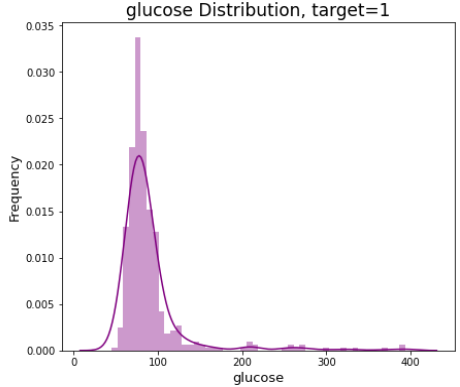


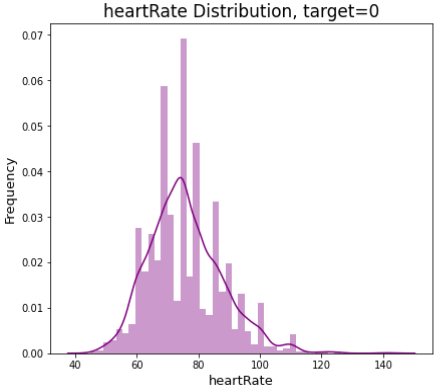
EDA and Data Visualization

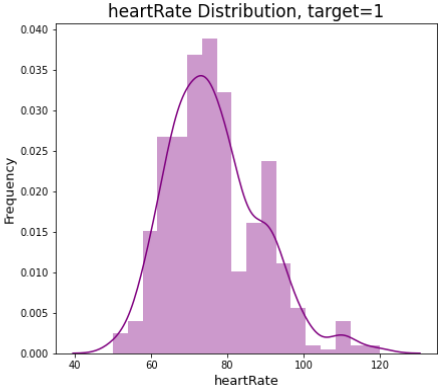


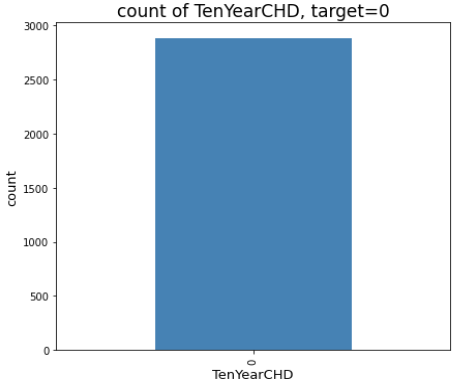


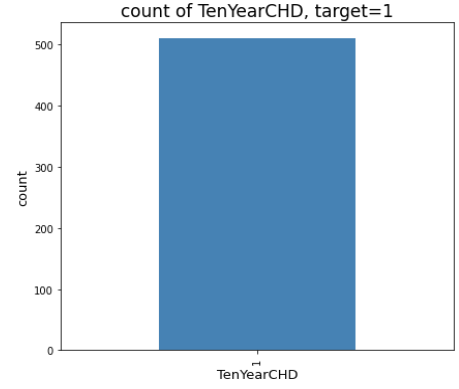












Feature Engineering

Blood Pressures:

Sys_BP: systolic blood pressure (Continuous)

Dia_BP: diastolic blood pressure (Continuous)

New feature,

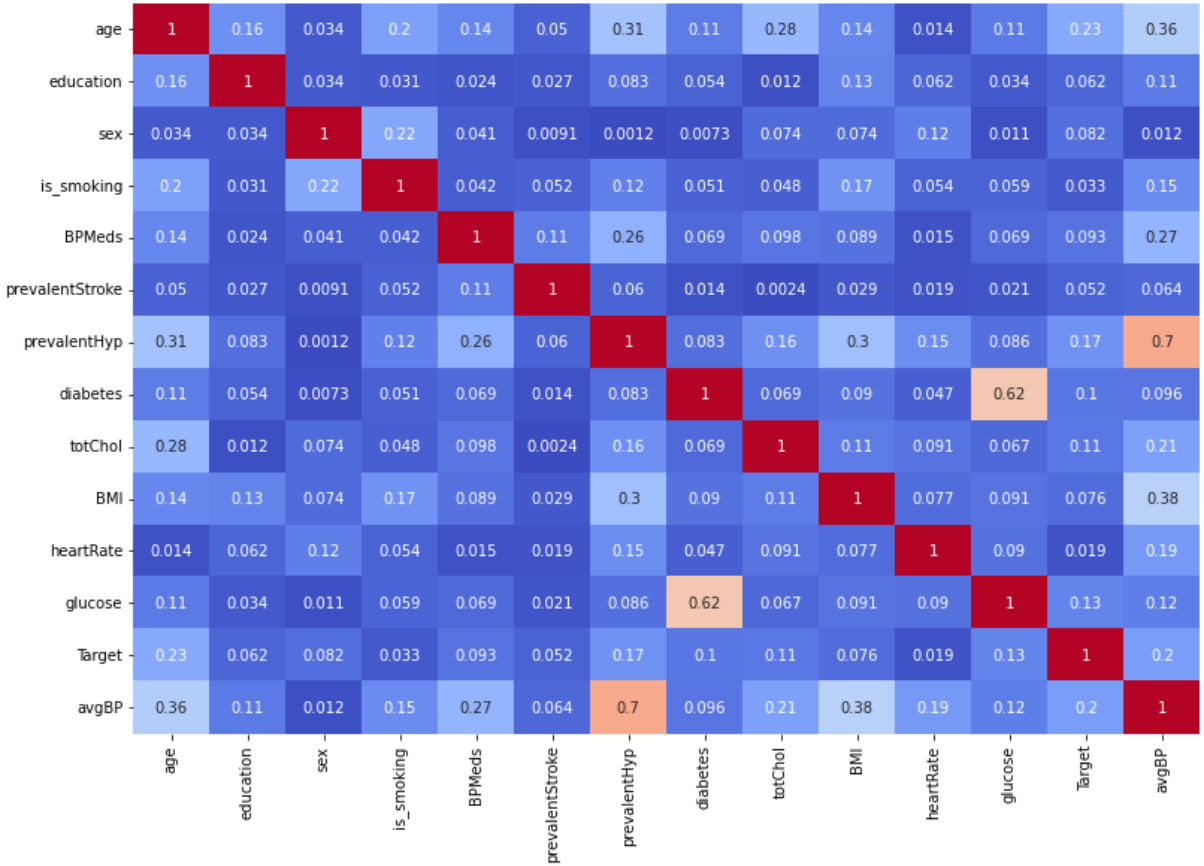
avg_BP = average of sys_BP and dia_BP

Cigarettes count:

Through EDA, I have deduced that whether a person smokes or not, is more important than how many cigarettes a person smoke.

To reduce multicollinearity, I am dropping cigs_Per_Day

Correlation Heatmap



As you can see, there exists a strong correlation between prevalentHyp and avgBP, also between diabetes and glucose.

Giving importance to both history and present features.
So, I am not further reducing the features…

ML Models' Description

Logistic Regression model

Parameters

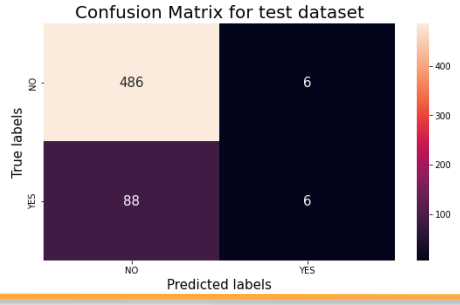
Penalty: l1

Solver: liblinear

Fit_intercept: True

C: 7.995

Output



Random Forest Classifier model

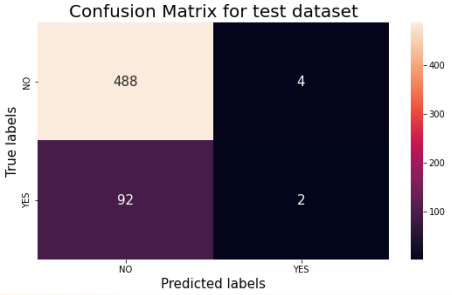
Parameters

Max_depth: 50

Min_samples_leaf: 5

min_samples_split: 5

Output



SVM Classifier model

Parameters

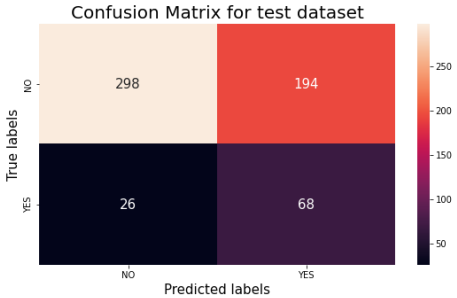
kernel: linear

Class_weight: balanced

gamma: 0.936

C: 6.6729

Output



XGBoost Classifier model

Parameters

Colsample_bytree: 0.4

gamma: 0.8098

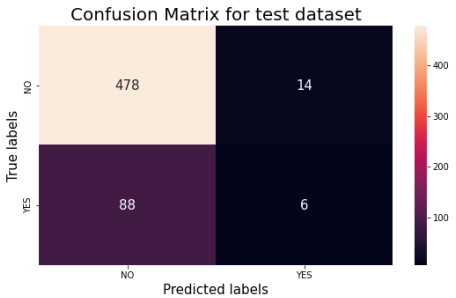
max_depth: 25

Min_samples_weight: 1

Reg_alpha: 0.1

Subsample: 0.9

Output



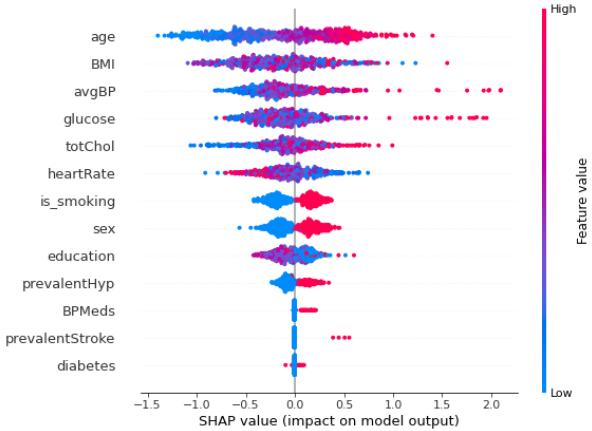
Best model

- In this study, I have done EDA, null values treatment, encoding of categorical columns, feature selection, and then model building.
- Except for the SVM model, all the models performed poorly. Because, the interference between the noise and the signal in the data was too strong.
- The SVM model gave a recall of 72% at the expense of predicting a lot many False Positives'.

Model Name	Roc auc (Train , Test)	Accuracy (Train , Test)	Precision (Train , Test)	Recall (Train , Test)	Log loss (Train , Test)	F1 score (Train, Test)
Logistic Regression	(0.538, 0.526)	(0.859, 0.84)	(0.757, 0.5)	(0.08, 0.064)	(4.884, 5.54)	(0.145, 0.113)
SVM Classifer	(0.679, 0.665)	(0.659, 0.625)	(0.262, 0.26)	(0.709, 0.723)	(11.789, 12.967)	(0.383, 0.382)
Random Forest Classifier	(0.57, 0.507)	(0.871, 0.836)	(1.0, 0.333)	(0.14, 0.021)	(4.441, 5.658)	(0.246, 0.04)
XGBoost Classifier	(0.963, 0.518)	(0.989, 0.826)	(1.0, 0.3)	(0.926, 0.064)	(0.384, 6.012)	(0.961, 0.105)

Model Interpretation : XGBoost, SVM Classifier model

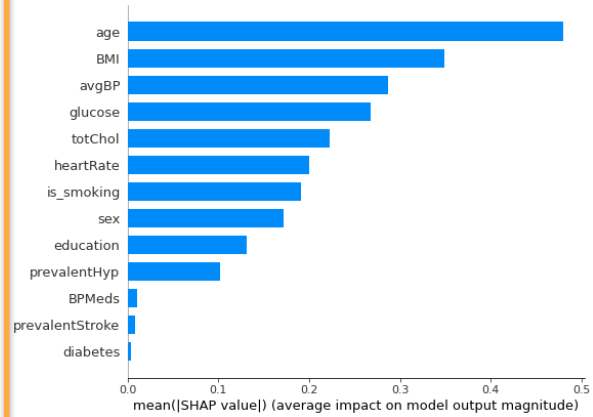
The **noise** in the data can be observed using following shap summary plot of XGBoost classifier model.
Observe the top 6 features.



Feature Impact

-by XGBoost model

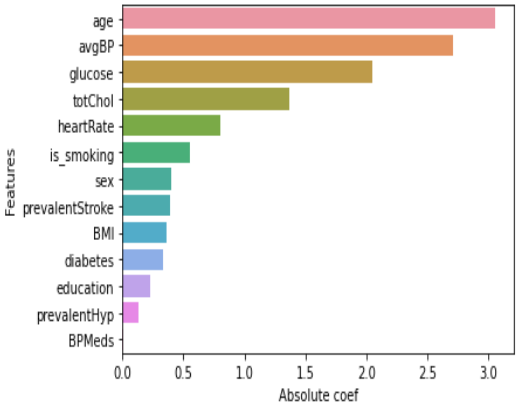
The XGBoost model predicts the results based on impact of age, BMI, avgBP, glucose and totChol as top 5 features.



Feature Impact

-by XGBoost model

The SVM model predicts the results based on importance of age, avgBP, glucose, totChol and is_smoking as top 5 features.



Feature Importance

-by SVM model

Data Resampling

‘Getting a noisy dataset’, treating it as a common problem for a data scientist in an industry. I decided to resample the given dataset using the risky limits of each feature. I tried to balance the dataset by under-sampling the majority class and converting those observations to the minority class. Much like oversampling with ‘Tomek links’. But ‘Tomek links’ failed to give expected results. The queries used for data resampling are,

{

If (**age > 60)**, risk True

If (**prevalent stroke = True)**, risk True

If (**age > 50) & (total cholesterol > 240 or BMI > 30 or heart rate >100 or glucose > 126 or average BP > 120)**, risk True

}

The focus of resampling was to get accurate answer if the patient shows two any strong indications along with age.

ML Models' Description

Logistic Regression model

Parameters

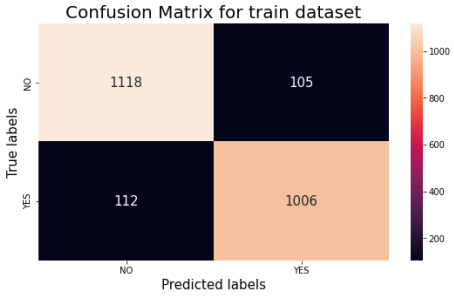
Penalty: l1

Solver: liblinear

Fit_intercept: True

C: 0.85173

Output



Random Forest Classifier model

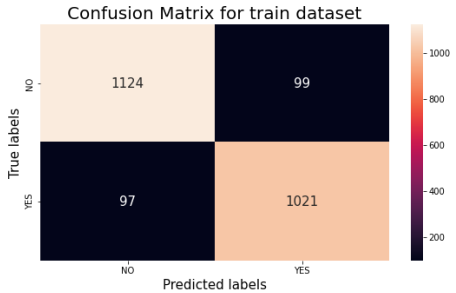
Parameters

Max_depth: 9

Min_samples_leaf: 107

Min_samples_split: 97

Output



SVM Classifier model

Parameters

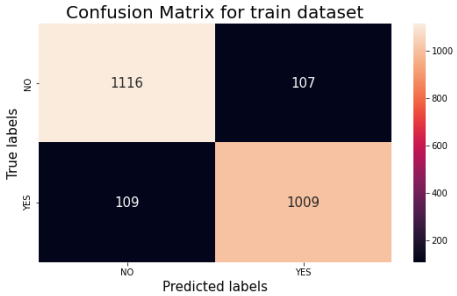
kernel: linear

Class_weight: balanced

gamma: 0.7627

C: 4.2768

Output



XGBoost Classifier model

Parameters

Colsample_bytree: 0.4

gamma: 0.18284

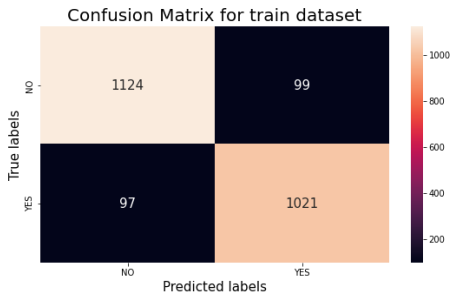
max_depth: 23

Min_samples_weight: 0.97

Reg_alpha: 9.34

Subsample: 0.9

Output



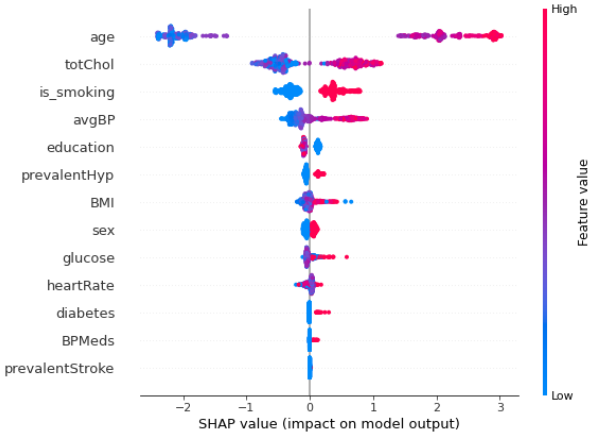
Best model

- After data resampling, I have repeated all the steps from EDA to Model building
- Since, the lot of noise has been cleared all the models performed well
- Judge the best model using following table,

Model Name	Roc auc (Train , Test)	Accuracy (Train , Test)	Precision (Train , Test)	Recall (Train , Test)	Log loss (Train , Test)	F1 score (Train, Test)
Logistic Regression	(0.907, 0.893)	(0.907, 0.894)	(0.905, 0.904)	(0.9, 0.871)	(3.202, 3.654)	(0.903, 0.887)
SVM Classifier	(0.908, 0.9)	(0.908, 0.901)	(0.904, 0.908)	(0.903, 0.882)	(3.187, 3.419)	(0.903, 0.895)
Random Forest Classifier	(0.916, 0.894)	(0.916, 0.894)	(0.912, 0.898)	(0.913, 0.879)	(2.892, 3.654)	(0.912, 0.888)
XGBoost Classifier	(0.943, 0.918)	(0.944, 0.92)	(0.97, 0.95)	(0.911, 0.879)	(1.918, 2.77)	(0.94, 0.913)

Model Interpretation : XGBoost, SVM Classifier model

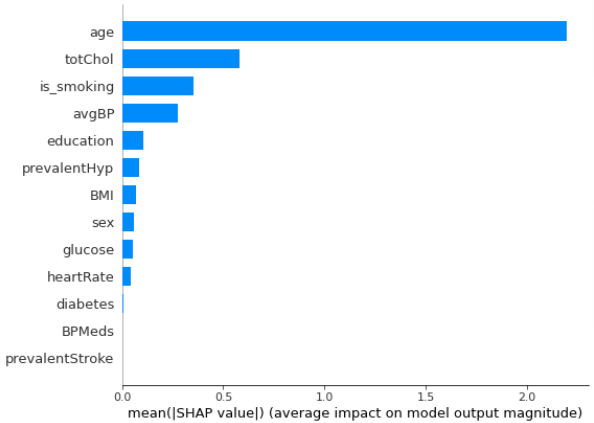
The tree classifier used age feature the most to branch out. There is much reduced noise in the resampled dataset



Feature Impact

-by XGBoost model

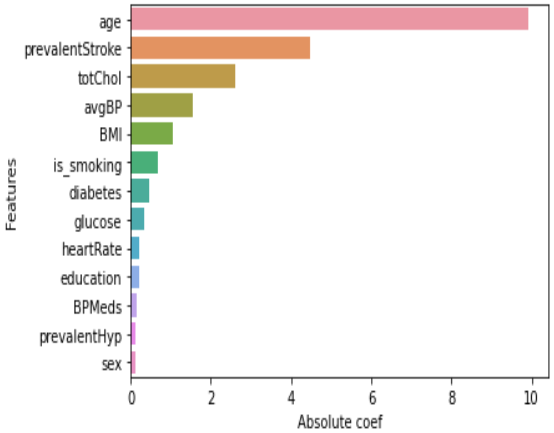
The XGBoost model predicts the results based on impact of age, totChol, is_smoking, and avgBP as top 4 features.



Feature Impact

-by XGBoost model

The SVM model predicts the results based on importance of age, prevalentStroke, totChol, and avgBP as top 4 features.



Feature Importance

-by SVM model

Conclusions

For the given dataset,

* Except for the SVM model, all the models performed poorly.
* The **SVM model gave a recall of 72%** at the expense of predicting a lot many ‘False Positives’.

For the resampled dataset,

* All the models showed high performance on the resampled dataset.
* The **XGBoost classifier** showed the best performance with the least log loss, highest f1 score of 91%, precision score of 95%, accuracy of 92%, and **recall of 88%.**
* All the models reported that the most prominent risk factor is **age**.
* Using feature importance from the Logistic regression model, the SVM model, and feature impacts from the XGBoost classifier model, the top 5 features that increase the risk of CVDs are **age, total cholesterol, and history of the previous stroke, average BP, and BMI**.
* The model has a limitation: It is suitable to predict the risk of cardiovascular diseases when the person is +ve with at least two risk factors.

Thank you!