Capstone Project

Cardiovascular Risk Prediction

Vijay N

Steps Performed

1. Defining the problem statement
2. Data Summary
3. EDA and Preparation of dataset
4. Applying the Model
5. Model Evaluation and Selection
6. Conclusion

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).
Let's see how this can be accomplished in the coming sections.

Data Summary

- **Demographic**
  - Sex
  - Age
  - Education

- **Medical (history)**
  - BP Meds
  - Prevalent Stroke
  - Prevalent Hyp
  - Diabetes

- **Dependent or Predicted variable**
  - TenYearCHD

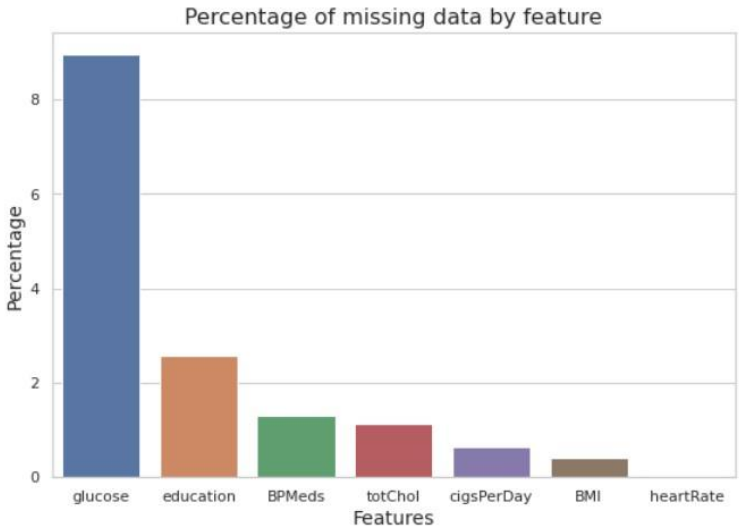
- **Behavioral**
  - Is_smoking
  - Cigs per day

- **Medical (current)**
  - Tot Chol
  - Sys BP
  - Dia BP
  - BMI
  - Heart rate
  - Glucose

Our dataset has 3390 rows and 17 columns to begin with.

Spread of Missing values

	Total	Percentage
glucose	304	8.967552
education	87	2.566372
BPMeds	44	1.297935
totChol	38	1.120944
cigsPerDay	22	0.648968
BMI	14	0.412979
heartRate	1	0.029499

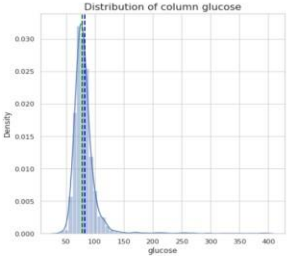


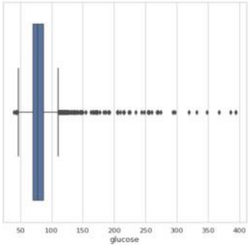
We have around **15% of missing values**.

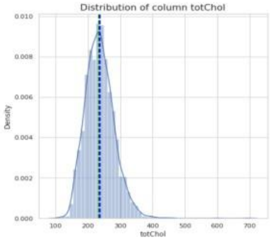
- Since **Education** qualification of person won't be having any dependency in heart disease, dropped them
- Imputed missing value of **Glucose** with a median glucose value based on the record that has diabetes or not.
- Imputed missing **BPMeds** with a prevantHyp value. Because, if the person is suffering from hypertension, he/she will be under medication for the same.
- Missing value of **cigsPerDay** will be imputed with mean cigsPerDay.
- Since the distribution is close to normal imputing missing value of **totChol** with median totChol, **BMI** with median BMI and **heartrate** with median heartrate.

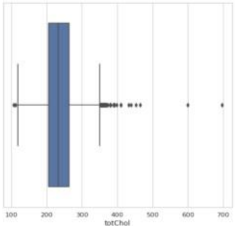
Outlier Treatment

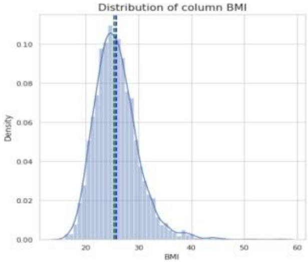
Observation of outliers before treatment

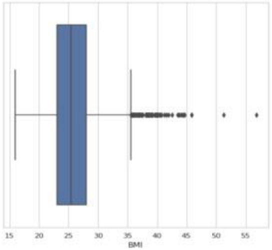


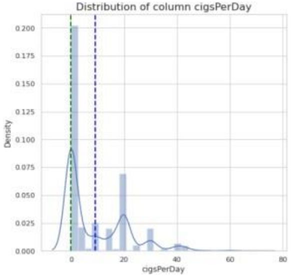


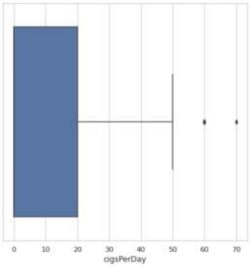


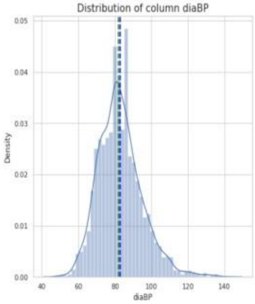


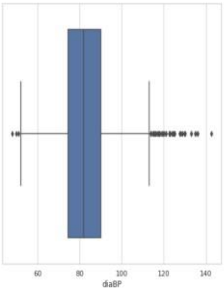


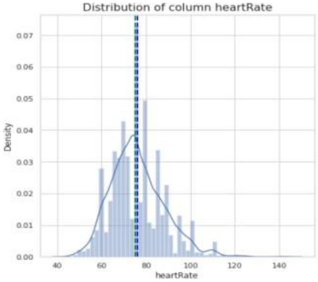


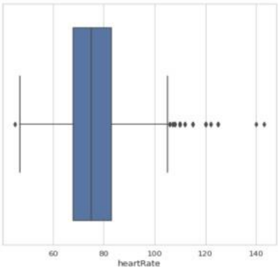






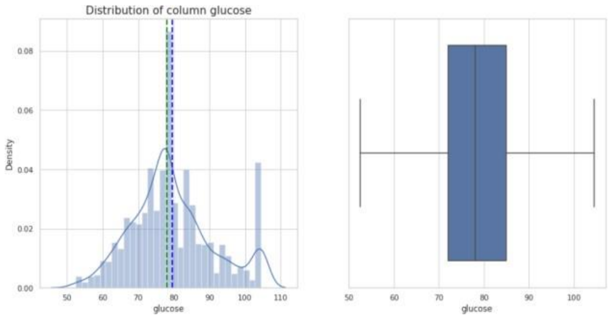


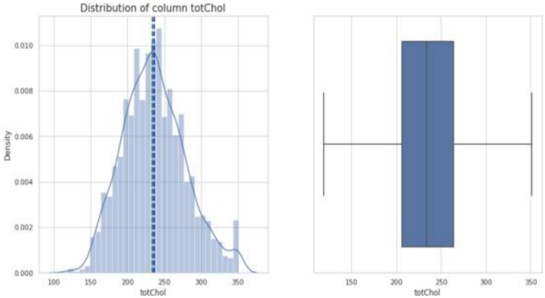


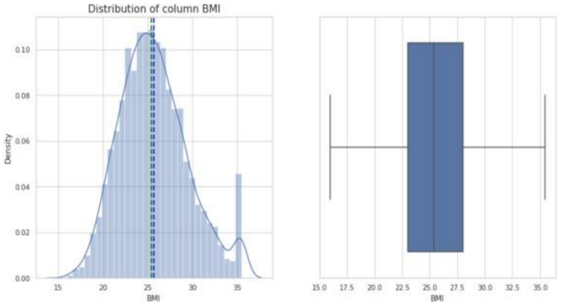


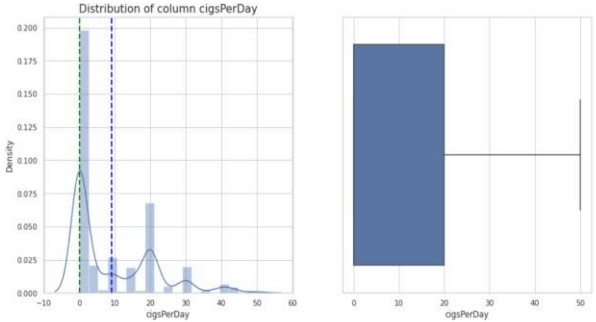
Outlier Treatment (contd.)

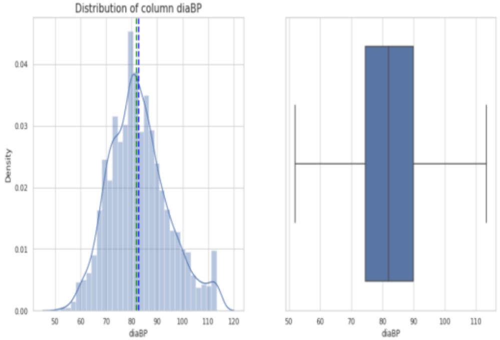
Outliers handled

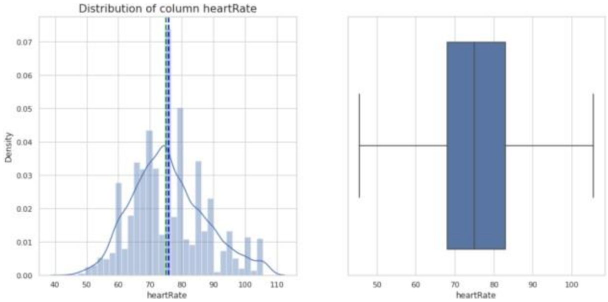




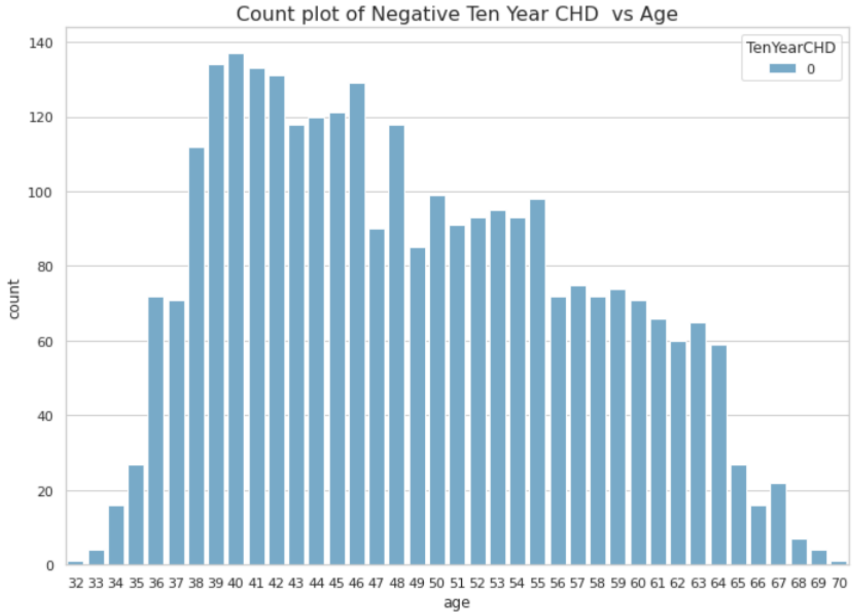


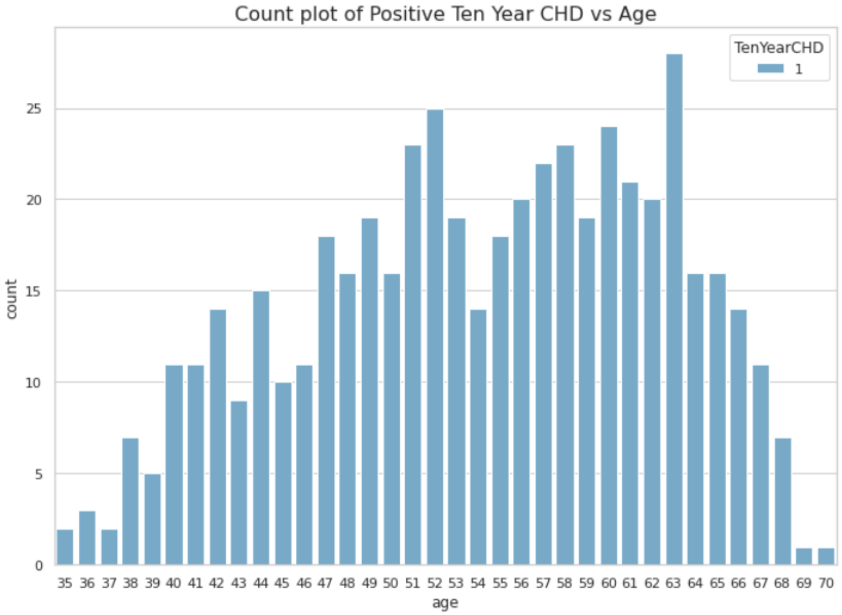






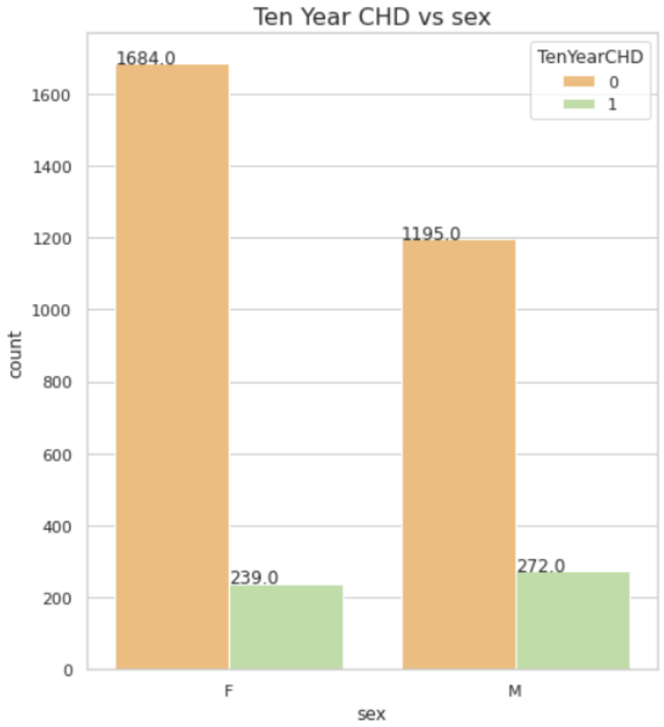
Exploratory Data Analysis



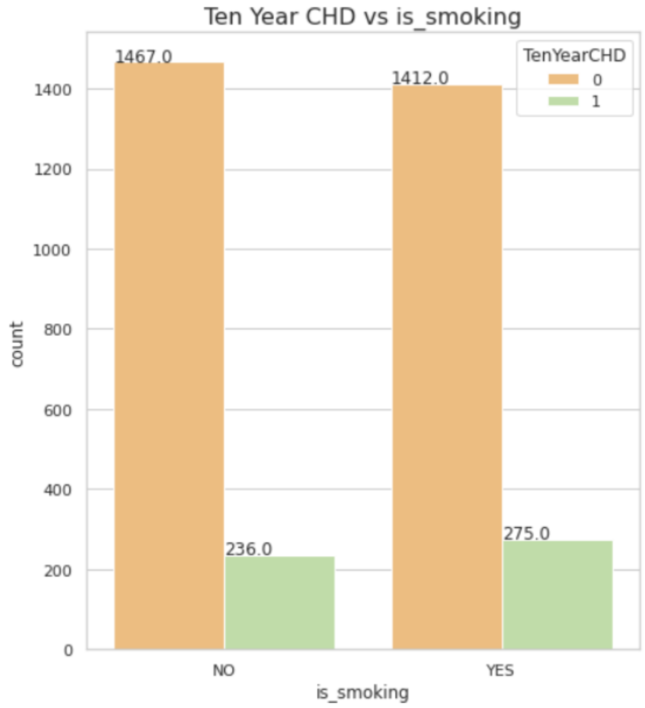


The chances of Getting Coronary Heart Disease is less for the lower age groups.

Exploratory Data Analysis (contd.)



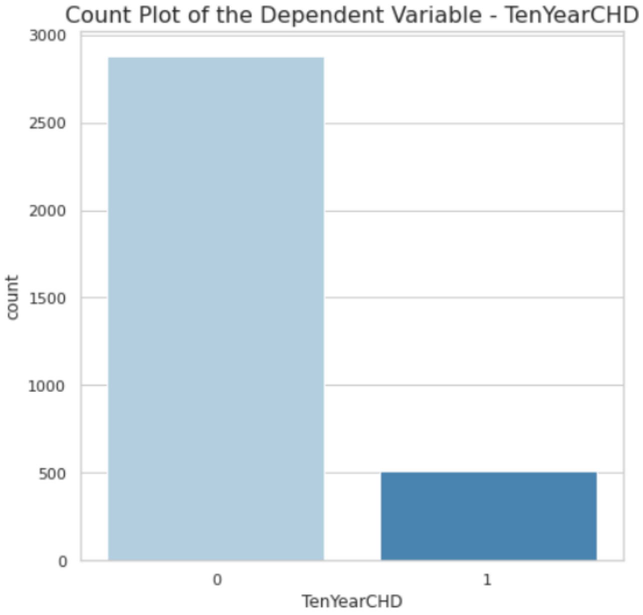
Chances of CHD in 10 years is more among Males.



Chances of CHD in 10 years is more among Smokers.

Dependent Variable Analysis

TenYearCHD is our dependent variable. This gives us the information whether that person will have a risk of getting coronary heat disease (CHD) in 10years. It is a categorical variable.



We can observe a huge imbalance in the dependent variable. So, we will be using SMOTE technique to solve this imbalance issue.

Multivariate Analysis

Preparation of Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3390 entries, 0 to 3389
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   age         3390 non-null   int64  
 1   cigsPerDay  3390 non-null   float64
 2   sysBP       3390 non-null   float64
 3   glucose     3390 non-null   float64
 4   sex         3390 non-null   int64  
dtypes: float64(3), int64(2)
memory usage: 132.5 KB
```

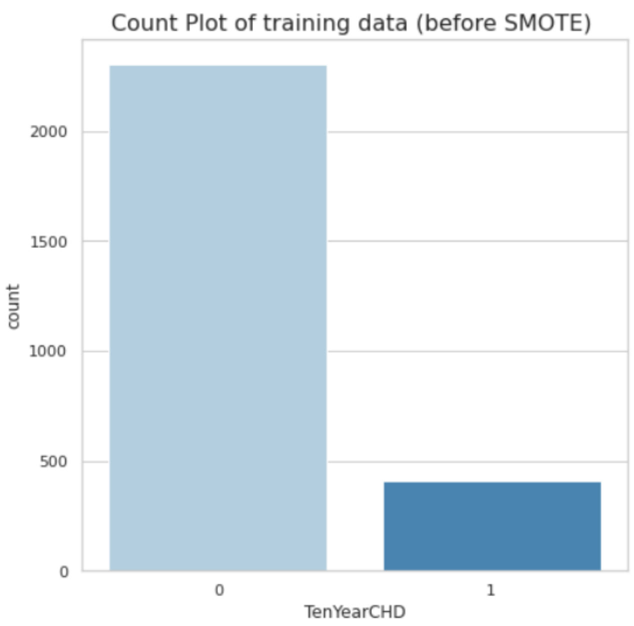
Task – Classification

Train dataset – (2712, 5)

Test dataset – (678, 5)

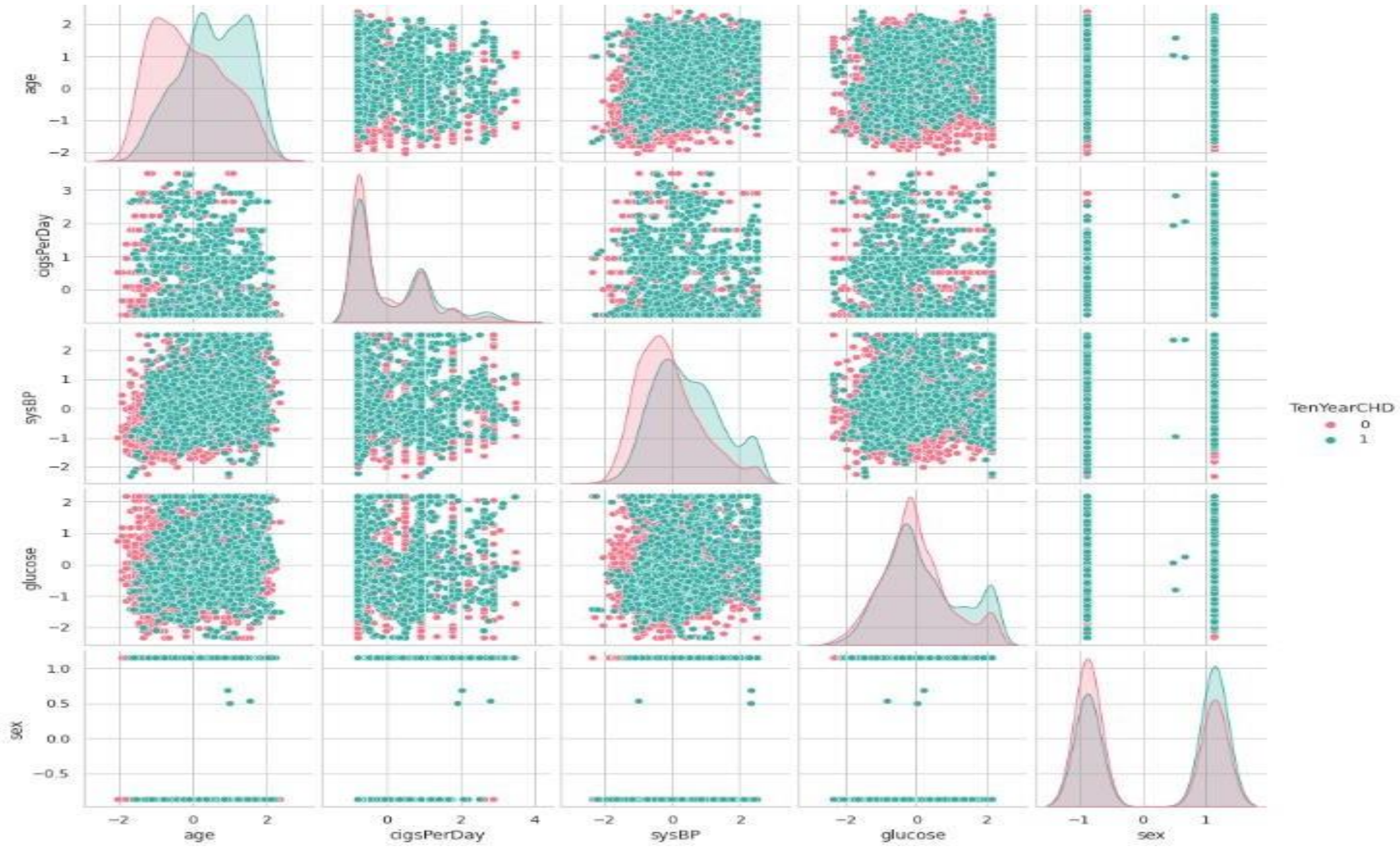
Response – Categorical variable (prediction of 10 year risk of CHD)

Handling Class Imbalance



Apply SMOTE TOMEK for training dataset

Pair plot of features after SMOTE

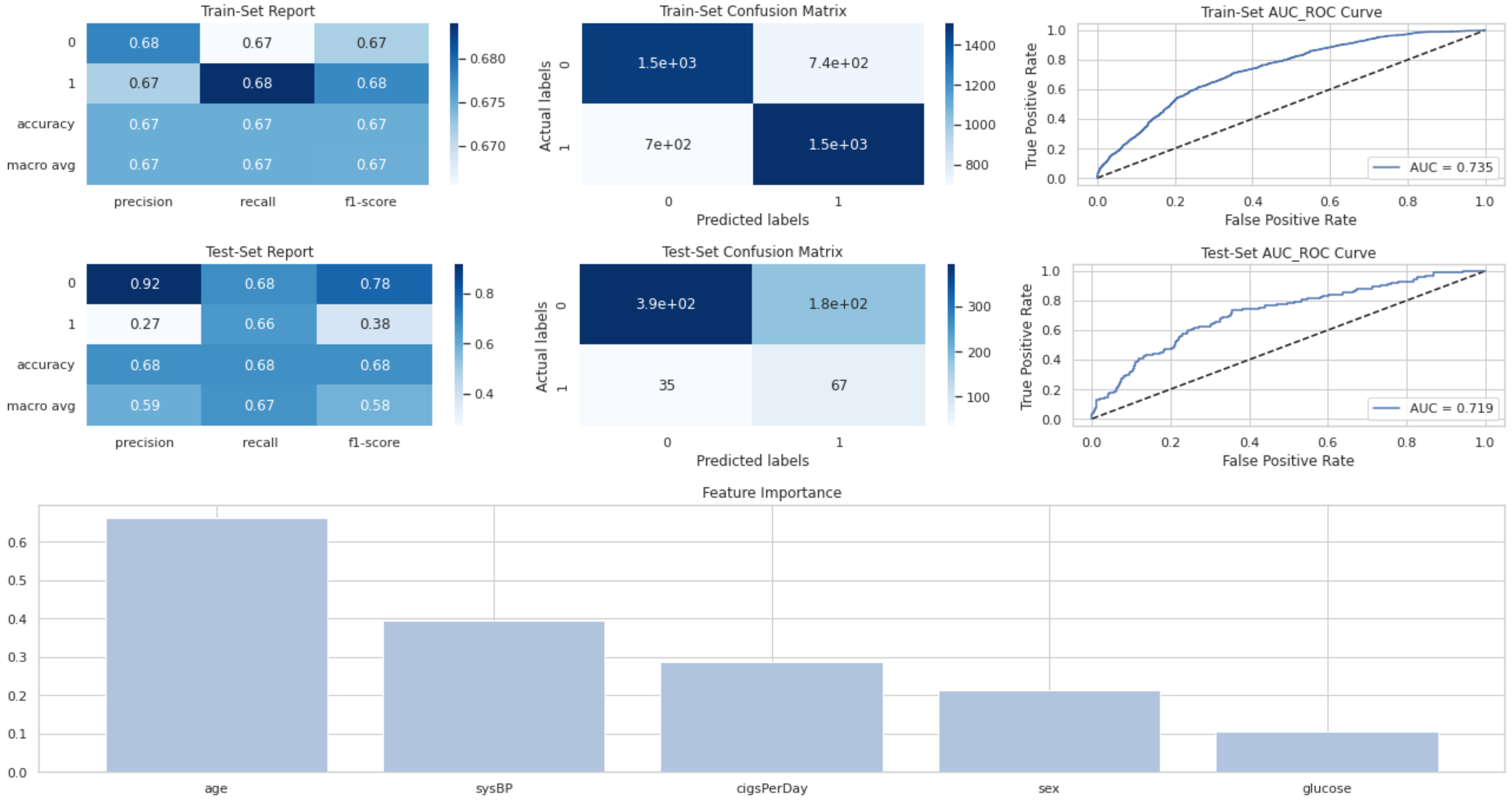


Evaluation Metrics

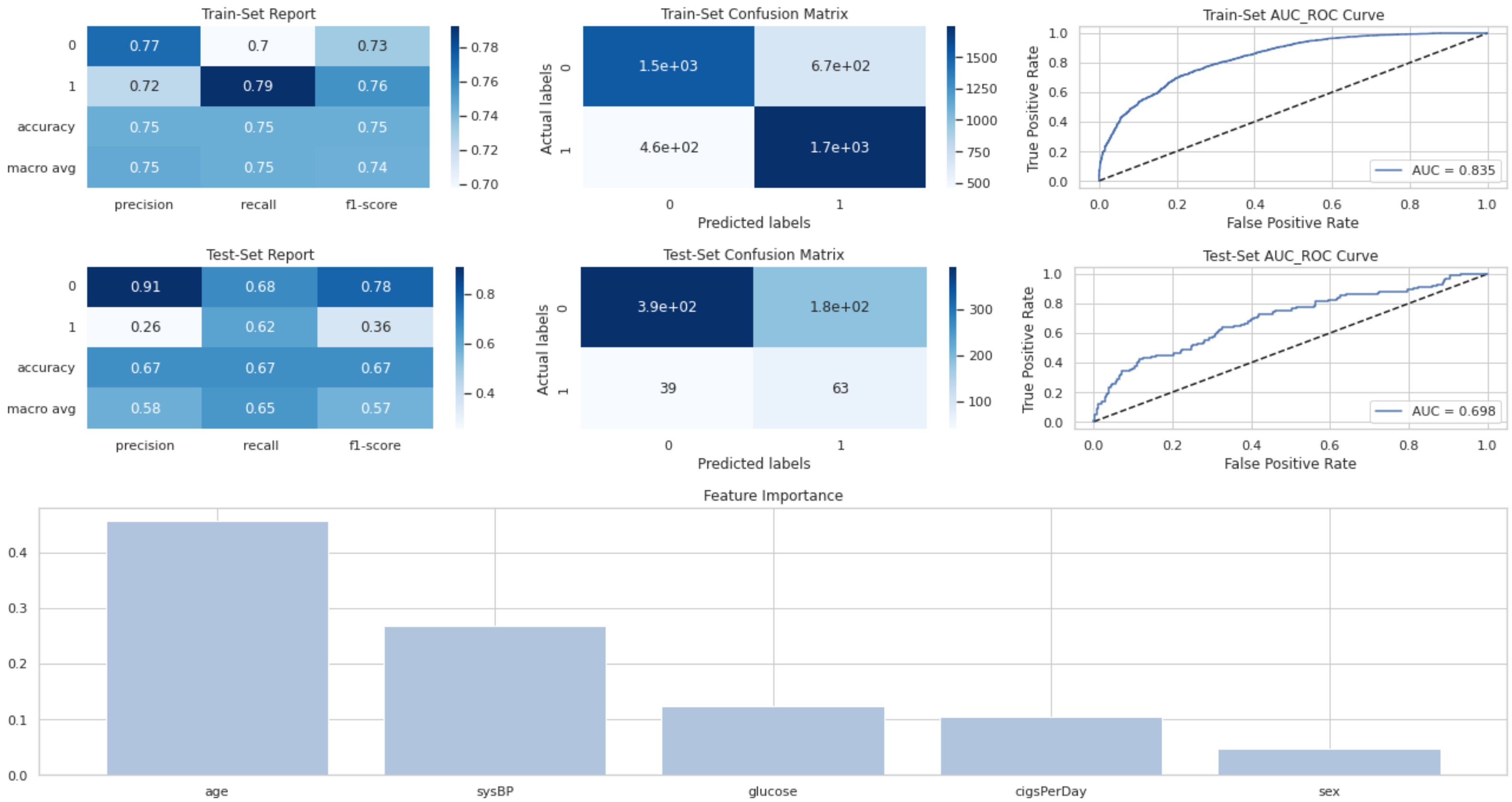
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\
specificity &= \frac{TN}{TN + FP}
\end{aligned}
$$

Logistic Regression



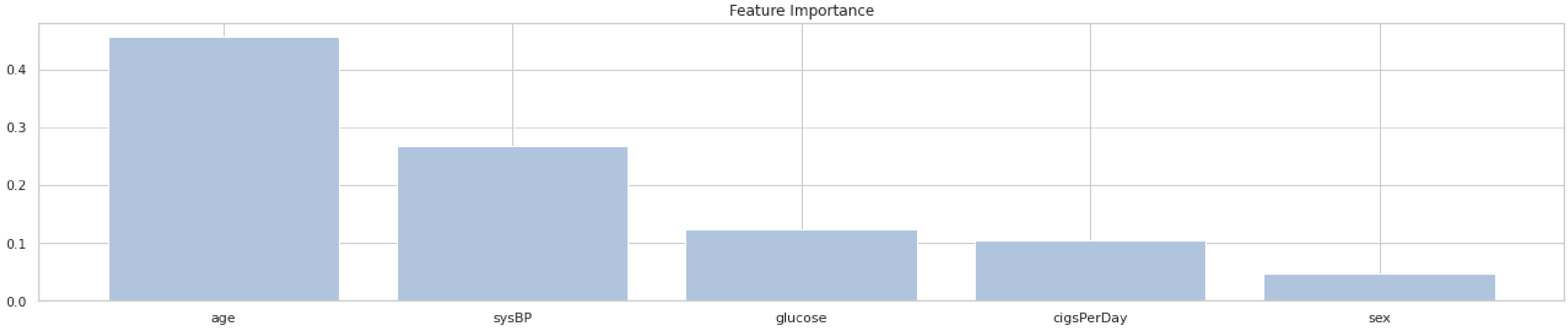
Random Forest Classifier



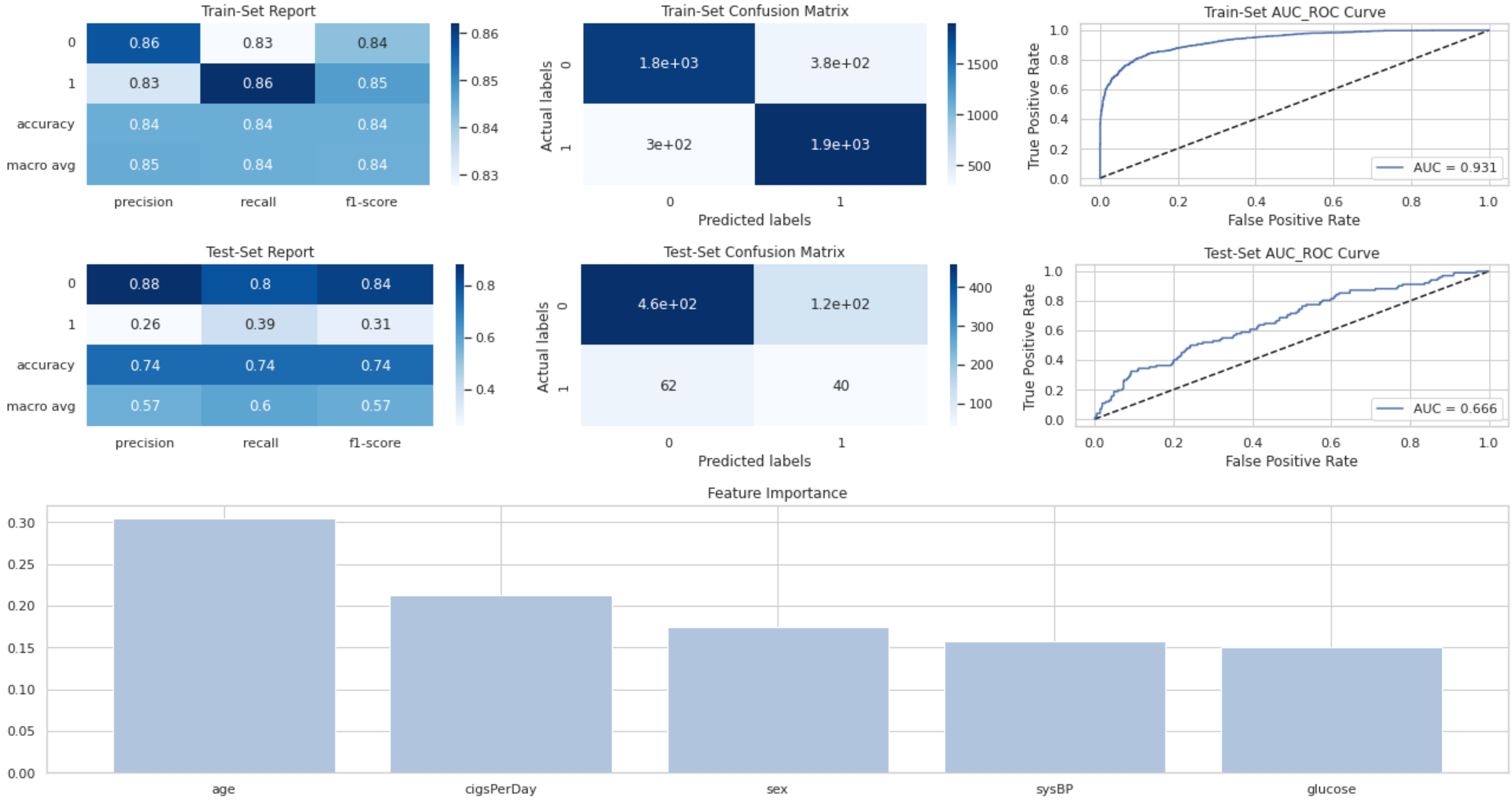
Random Forest Classifier (contd.)

BEST FIT PARAMETERS:

Max_depth – 8
Min_samples_leaf – 40
Min_samples_split – 100
N_estimators - 50



Extreme Gradient Boost (XGB)

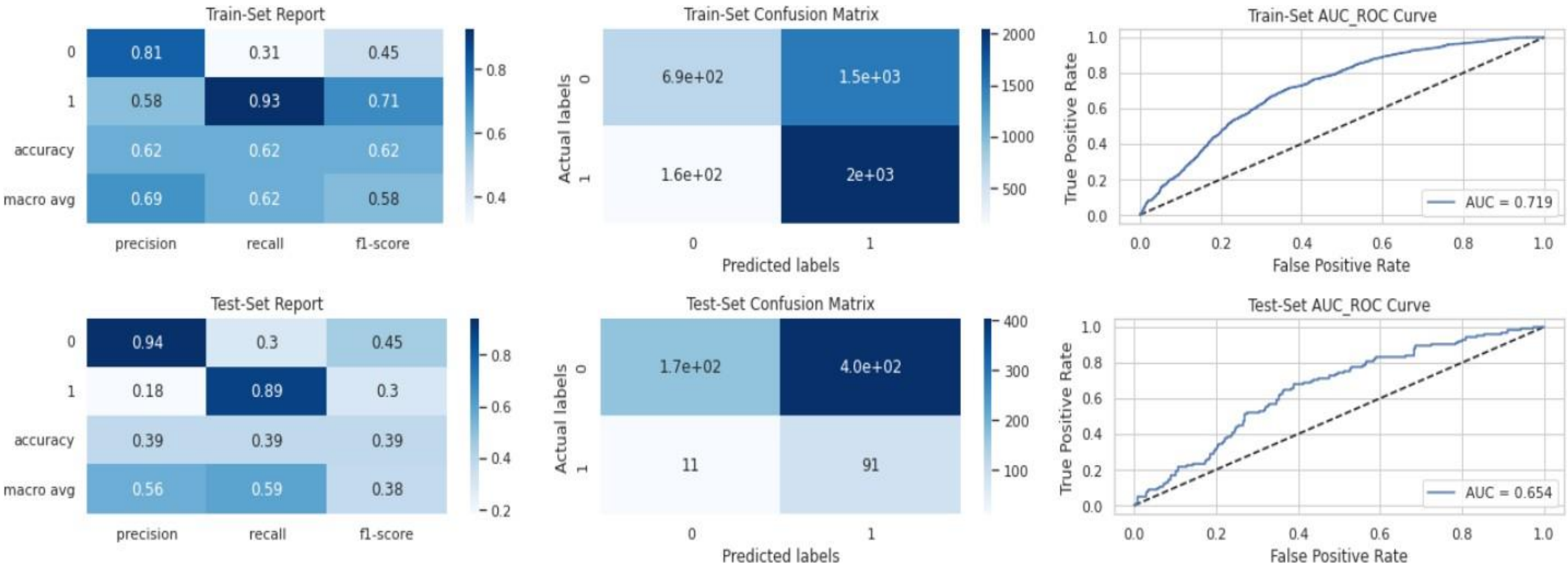


Extreme Gradient Boost (XGB) (contd.)

BEST FIT PARAMETERS:

Learning_rate – 0.1
Min_samples_leaf – 30
Min_samples_split – 20
N_estimators - 140

Support Vector Machine (SVM)



Conclusion

- **We have successfully built predictive model that can predict a patients risk for CHD based on their demography, lifestyle and medical history.**
- **Considered Recall score has the best metric to measure.**
- **Logistic Regression and other tree based algorithms were not quite good in classifying our data with accuracy.**
- **SVM worked has best classification model with recall score of 93% in training data and 89% in test data.**

Challenges

- **Computation time**

  Multiple iterations are run on a single model to tune the hyperparameters.

- **Less amount of data**

  Efforts must be put in gathering more data so that we can improve the model and can save more lives.

Q & A

Thank You