# Capstone Project

## Cardiovascular Risk Prediction

by

Roshan Jamthe

# Points for Discussion

- Problem Statement
- Data Summary
- Data Visualization
- Feature engineering
- Feature selection
- Train Test datasets
- ML Models' performance
- Best model
- Feature Importance
- 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)

## 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)

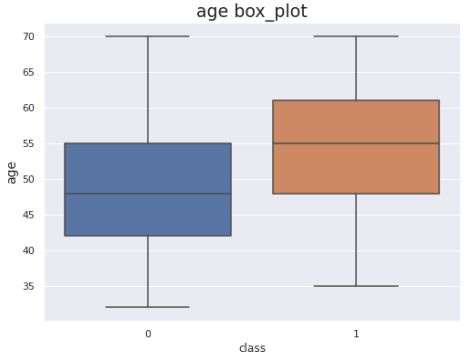
**Heart Rate:** heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

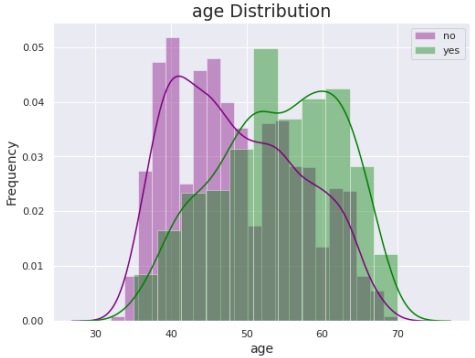
**Glucose:** glucose level (Continuous)

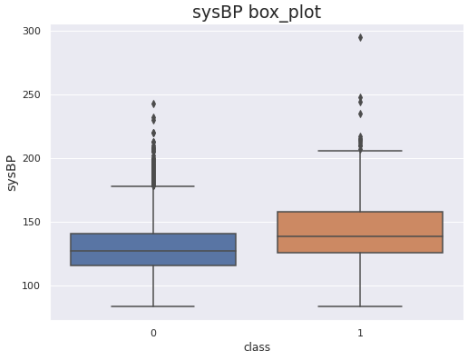
## Predict variable (desired target):

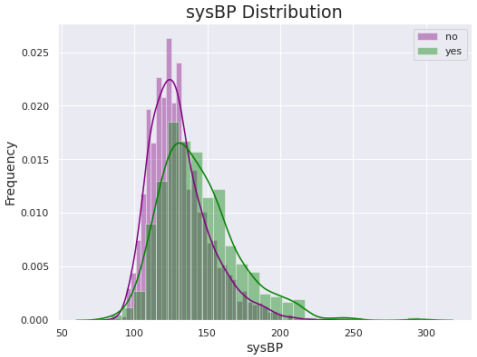
**TenYearCHD**: 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

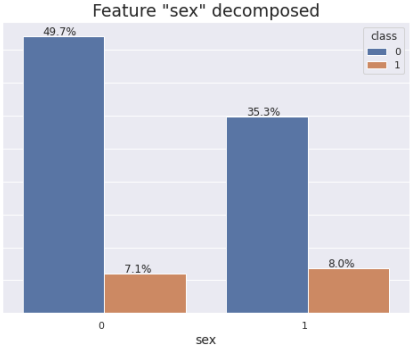
# Data Visualization

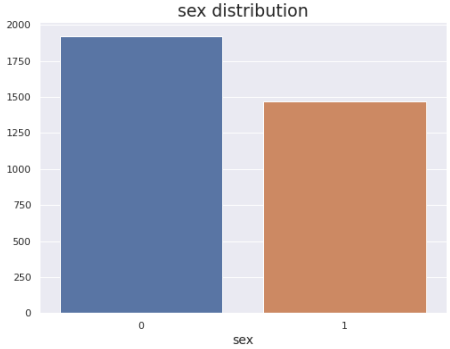



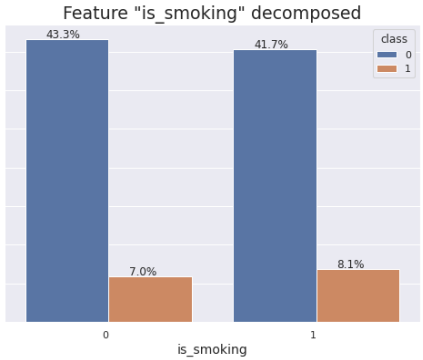



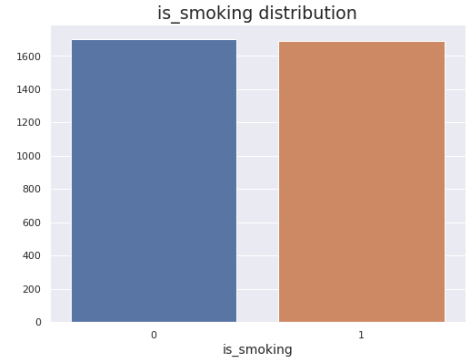



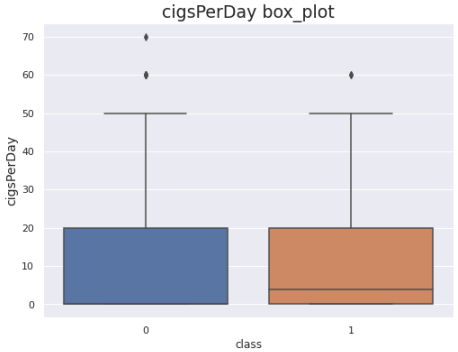



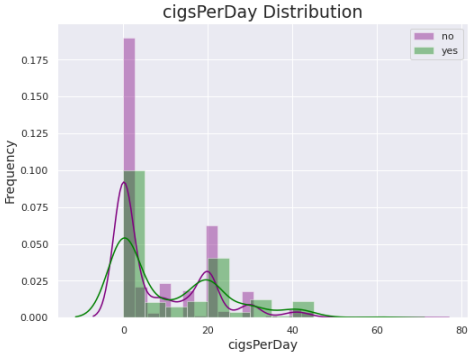

Feature "sex" decomposed
class
0
1
49.7%
7.1%
35.3%
8.0%
0
1
sex

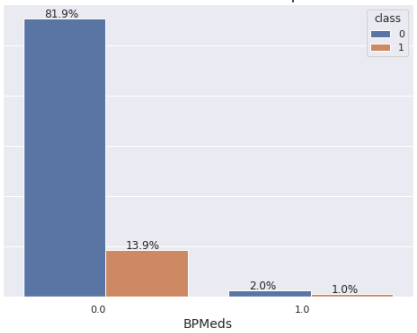

sex distribution
2000
1750
1500
1250
1000
750
500
250
0
0
1
sex

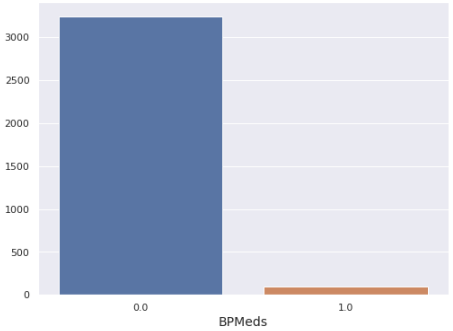

Feature "is_smoking" decomposed
class
0
1
43.3%
7.0%
41.7%
8.1%
0
1
is_smoking

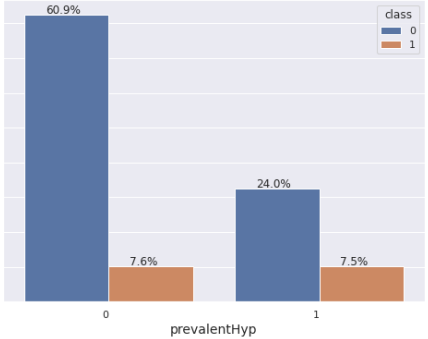

is_smoking distribution
1600
1400
1200
1000
800
600
400
200
0
0
1
is_smoking

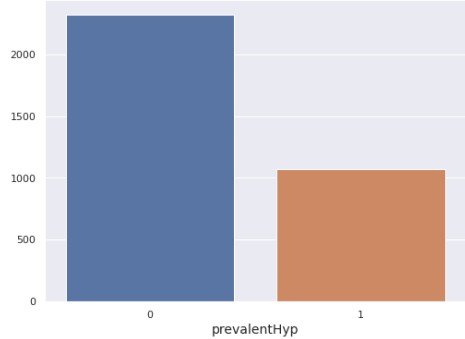

cigsPerDay box_plot
70
60
50
40
30
20
10
0
cigsPerDay
0
1
class

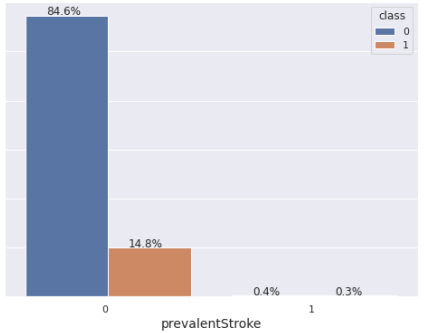

cigsPerDay Distribution
no
yes
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Frequency
0
20
40
60
80
cigsPerDay

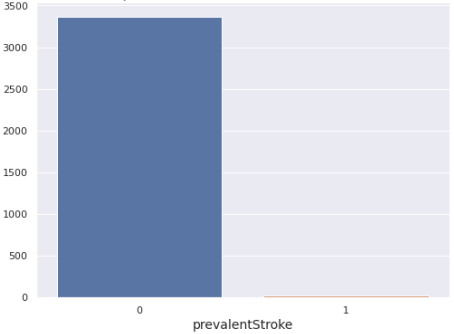

Feature "BPMeds" decomposed
class
0
1
81.9%
13.9%
2.0%
1.0%
0.0
1.0
BPMeds

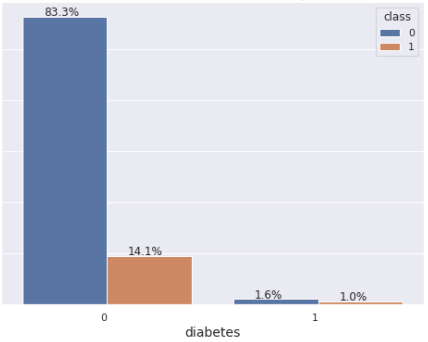

BPMeds distribution
3000
2500
2000
1500
1000
500
0
0.0
1.0
BPMeds

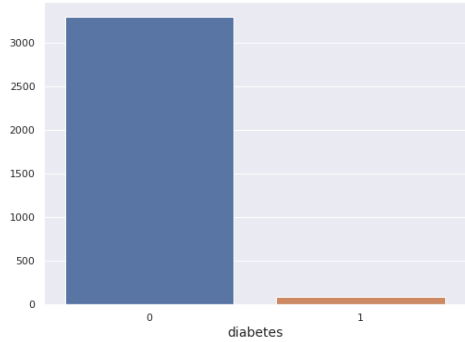

Feature "prevalentHyp" decomposed
class
0
1
60.9%
7.6%
24.0%
7.5%
0
1
prevalentHyp

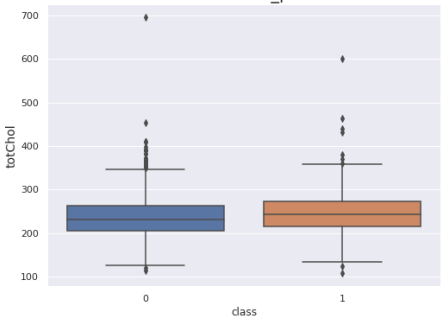

prevalentHyp distribution
2000
1500
1000
500
0
0
1
prevalentHyp

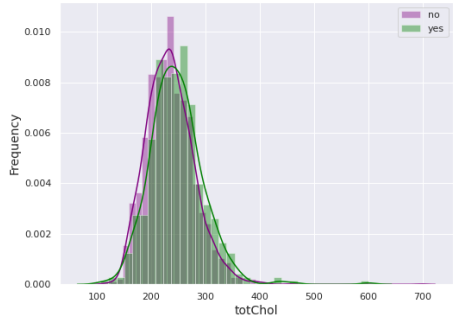

Feature "prevalentStroke" decomposed
class
0
1
84.6%
14.8%
0.4%
0.3%
0
1
prevalentStroke

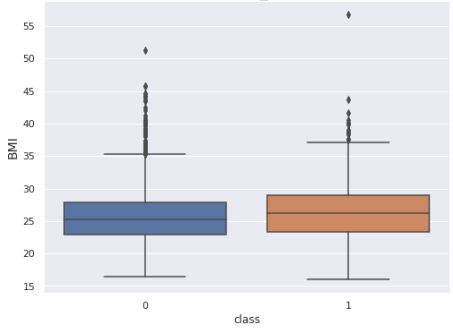

prevalentStroke distribution
3500
3000
2500
2000
1500
1000
500
0
0
1
prevalentStroke

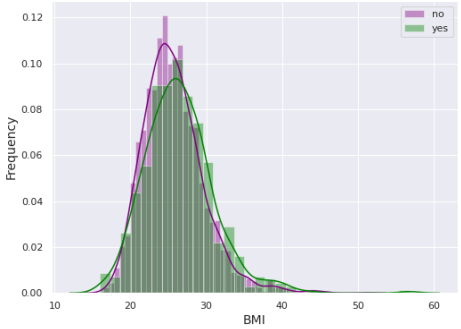

Feature "diabetes" decomposed
class
0
1
83.3%
14.1%
1.6%
1.0%
0
1
diabetes

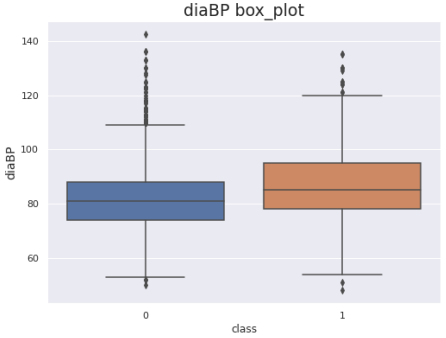

diabetes distribution
3000
2500
2000
1500
1000
500
0
0
1
diabetes

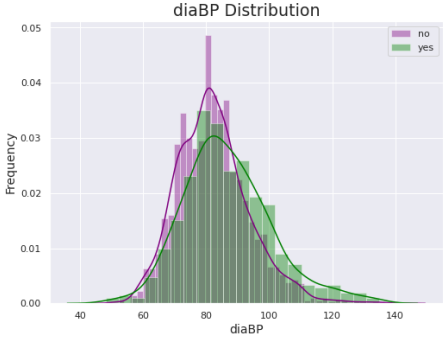

totChol box_plot
totChol
class
0
1

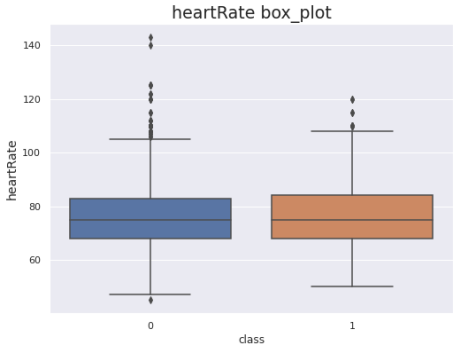

totChol Distribution
Frequency
totChol
no
yes

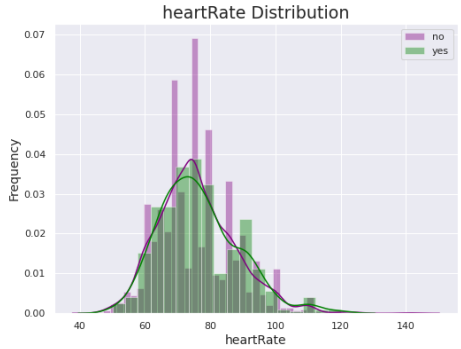

BMI box_plot
BMI
class


BMI Distribution
Frequency
BMI
no
yes


diaBP box_plot
diaBP
class


diaBP Distribution
Frequency
diaBP
no
yes


heartRate box_plot
heartRate
class


heartRate Distribution
Frequency
heartRate
no
yes

# Feature Engineering

## Blood Pressures:

**Sys_BP:** systolic blood pressure (Continuous)

**Dia_BP:** diastolic blood pressure (Continuous)

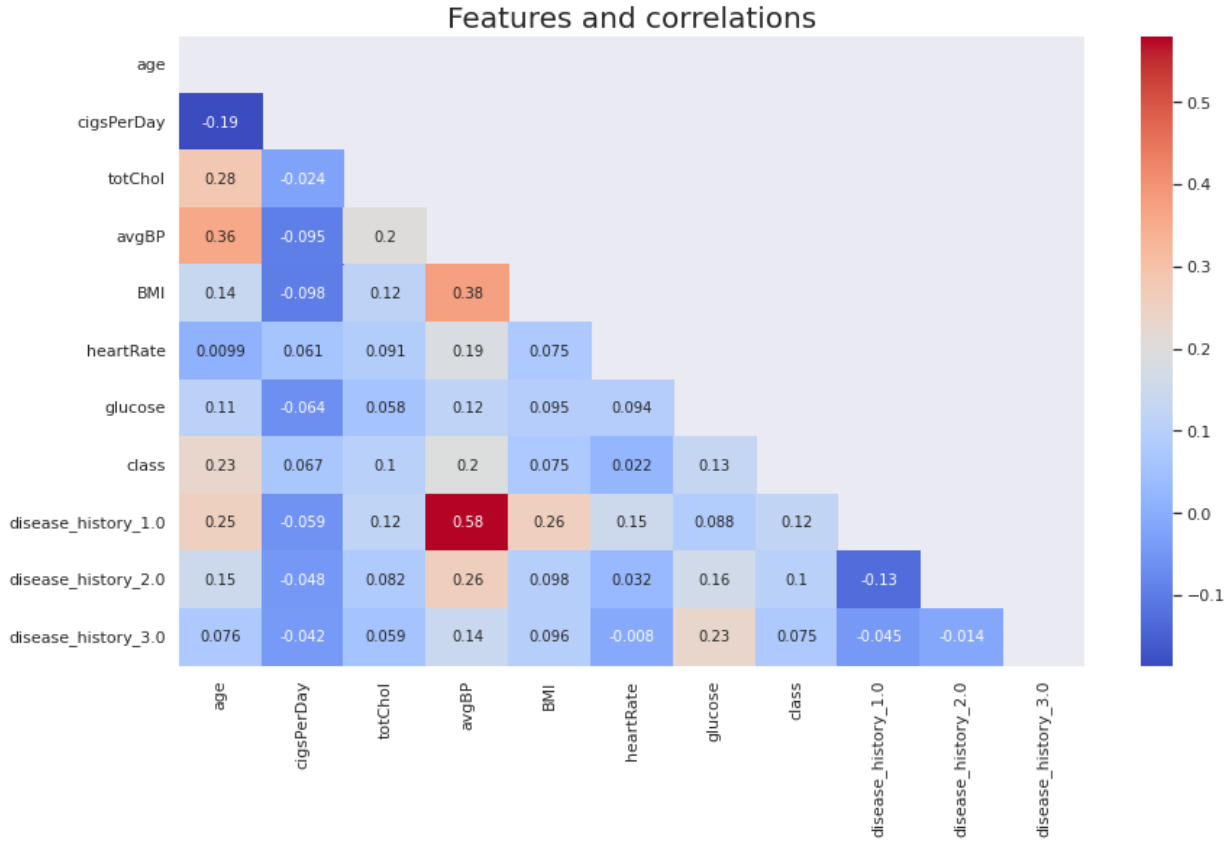
New feature,

avg_BP = average of sys_BP and dia_BP

## Historical features:

disease_history = BPMeds + prevalentStroke + prevalentHyp + diabetes

# Feature Selection




## Method

**For feature selection we used Boruta algorithm which is build on top of Random Forest**

# Train Test datasets


## Datasets Explanation

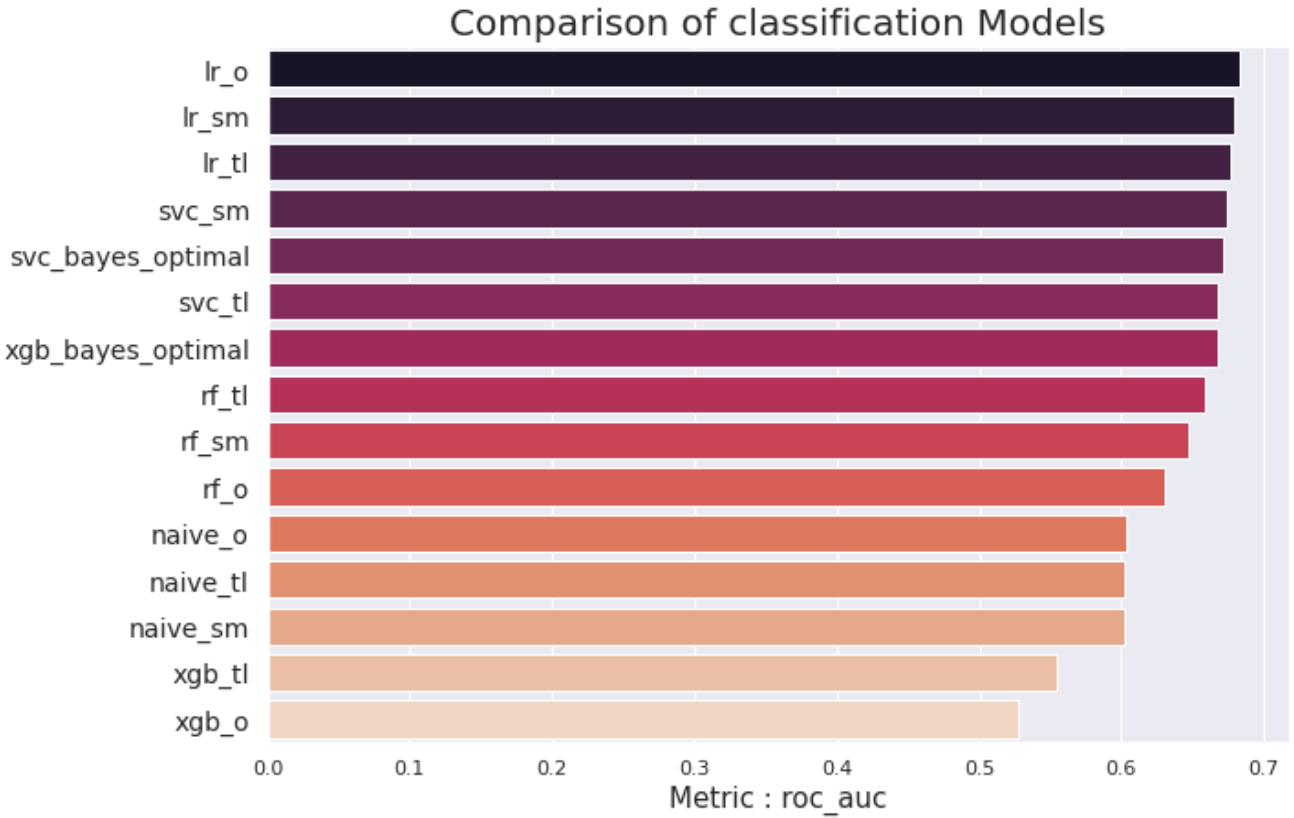
**y_train :**
**Untreated imbalanced train dataset**

**y_tl:**
**Undersampled with tomeklinks but still imbalanced train dataset**

**y_smote :**
**Oversampled with SMOTE, so balanced train dataset**

**y_test:**
**Imbalanced test dataset**

# ML Models' performance




## Abbrevations

### Models

**lr :**
**Logistic Regression**

**svc:**
**Support vector machine**

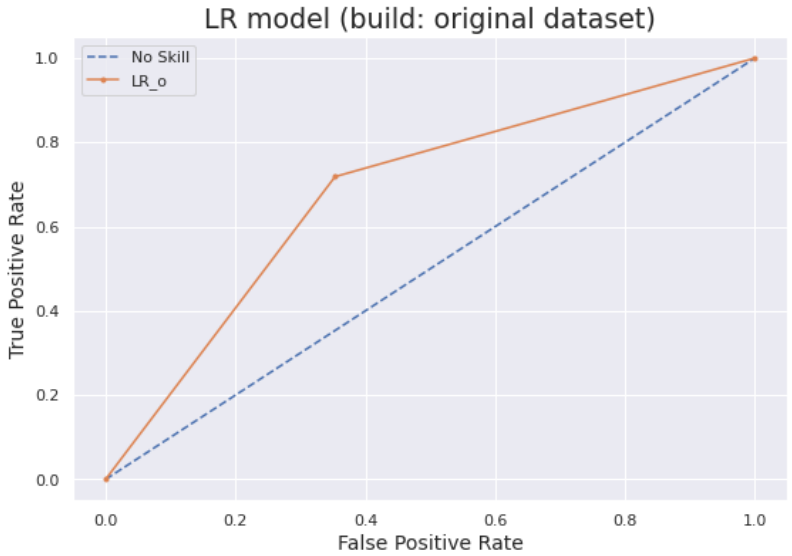
**rf :**
**Random forest**

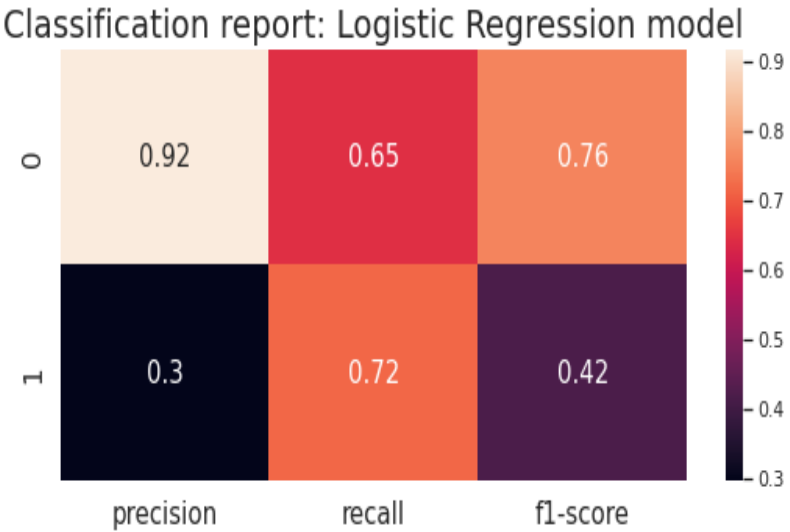
**xgb :**
**XGBoost**

### Dataset

**o : original**
**tl : under-sampled**
**sm : over-sampled**

# Best model : Logistic Regression model



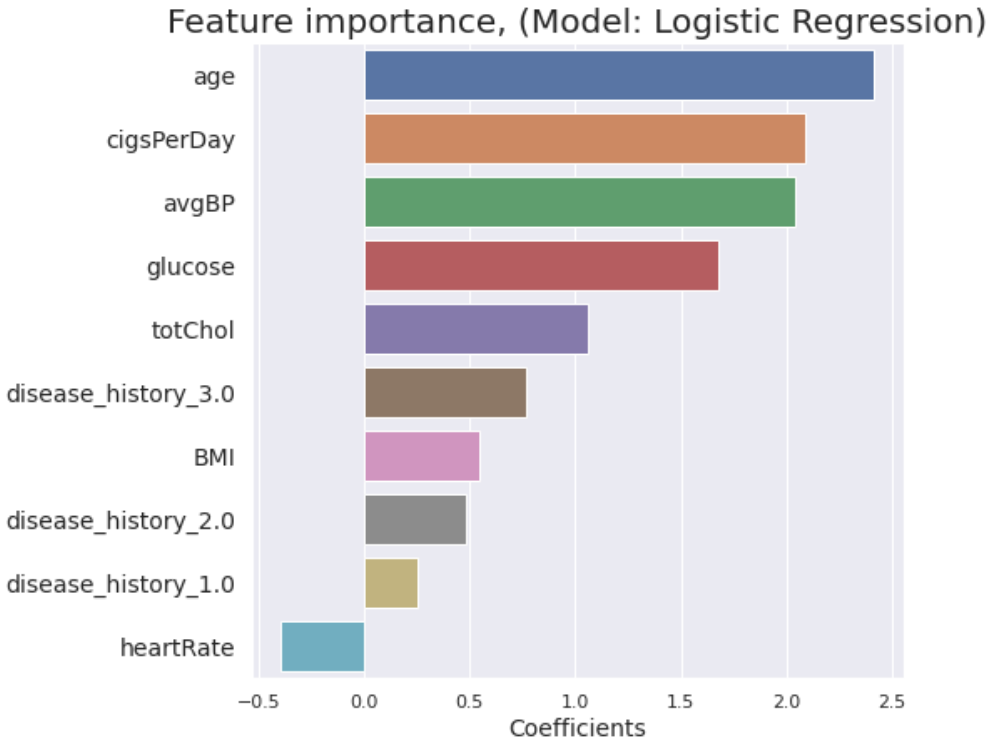




**Properties of the model:**

The model has the best precision and the best auc_roc score.

The model is really fast to react, takes less than 1 second to give out the results.

# Feature Importance




## Top 3 features

**The top 3 features are +vely correlated with the dependent variable.**

**The top risk factor is age followed by cigarettes per day and blood pressure.**

# Conclusions

- Selected features for model building : age, cigsPerDay, totChol, avgBP, BMI, heartRate, glucose, disease_history.
- Three train datasets were used for the analysis.
- Built naive bayes, logistic regression, support vector machine, random forest and XGBoost model using the 3 train datasets.
- On comparing all the models, Logistic regression model trained with original dataset gave best performance.
- We found out that, Logistic Regression model is performing the best with the minimum roc_auc score of 0.68, accuracy of 0.66, recall of 0.69 and precision of 0.263.
- Top 3 features that are helpful in predicting the class in decreasing order of their importance are age, cigs_per_day, avgBP.

# Thank you!