

CARDIOVASCULAR RISK PREDICTION

Deepak Kumar Dubey

Data Science Trainee ,
Almabetter,Bangalore

1.Introduction:

Cardiovascular disease is the leading cause of death worldwide and a major public health concern. Therefore, its risk assessment is crucial to many existing treatment guidelines. Risk estimates are also being used to predict the magnitude of future cardiovascular disease mortality and morbidity at the population level and in specific subgroups to inform policymakers and health authorities about these risks. Additionally, risk prediction inspires individuals to change their lifestyle and behaviour and to adhere to medications.

Although several risk prediction models of cardiovascular disease have been developed for different populations in the past decade, the validity of these models is a cause of concern. Most data for model formation and validation have been provided from a small set of populations, mostly from developed countries. Therefore, using this set for the classification of individuals from different risk groups of other populations might lead to risk overestimation. This, in turn, can result in increased costs of guidelines and health interventions. These models might also cause risk underestimation, which can lead to missing vulnerable cases. Consequently, providing a valid model for cardiovascular disease risk classification of each population has become a high priority for scientists and organisations working in this field.

2.Problem Statement:

- Currently cardiovascular diseases (CVDs) are the main cause of death worldwide. Disease risk estimates can be used as prognostic information and support for treating CVDs.
- The commonly used Framingham risk score (FRS) for CVD prediction is outdated for the modern population, so FRS may not be accurate enough.
- In this project ,a CVD prediction system based on machine learning is proposed.
- 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 both demographic, behavioural, and medical risk factors.

3.Data Summary:

Demographic:

- Sex: male or female("M" or "F")
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioural

- 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 a large number of possible values.)
 - Glucose: glucose level (Continuous)
- Predict variable (desired target)

Dependent variable

- 10-year risk of coronary heart disease CHD(binary: "1", means "Yes", "0" means "No") -

CONTINUOUS	NOMINAL
AGE	IS_SMOKING
EDUCATION	SEX
CIGS_PER_DAY	DIABETES
TOT_CHOL	BP_MEDS
SYS_BP	PREVALENT_STROKE
DIA_BP	PREVALENT_HYP
BMI	DIABETES
HEART_RATE	
GLUCOSE	

Observations:3390 Features:16

4.Approach:

- Data understanding
- Handling null values
- Exploratory data analysis
- Model fitting
- Analysing evaluation metrics

4.1.Data Understanding:

Age-we have data for people having age between 32 to 70 with mean age being around 49 years.

Cigarettes per day- most people don't smoke any cigarette in a day but those who smoke smoke from 1 to 70 cigarettes with mean smoked cigarettes being around 9.

Totchol- Total cholesterol value has mean value of 237 with minimum being 107 and max 690.

Sys BP-Normal: Below 120

> Elevated: 120-129

> Stage 1 high blood pressure (also called hypertension): 130-139

> Stage 2 hypertension: 140 or more

> Hypertensive crisis: 180 or more

> In our data ,we have systolic bp distributed between 83.5 to 295 with mean being 132.

dia BP- This is what your diastolic blood pressure number means:

> Normal: Lower than 80

> Stage 1 hypertension: 80-89

> Stage 2 hypertension: 90 or more

> Hypertensive crisis: 120 or more

> In our data,We have diaBP values distributed between 48 to 142.5 with mean value being 83

BMI- Category BMI range - kg/m²

> Severe Thinness < 16

> Moderate Thinness 16 - 17

> Mild Thinness 17 - 18.5

> Normal 18.5 - 25

> Overweight 25 - 30

> Obese Class I 30 - 35

> Obese Class II 35 - 40

> Obese Class III > 40


We have BMI index values ranging between 15.96 to 57 with mean value being 25.79.

heart rate- A normal resting heart rate for adults ranges from 60 to 100 beats per minute.

> In our data we have heart rate values ranging from 45 to 143 with mean value being 76.

Glucose-

In our data, we have glucose ranging from 40 to 394 with mean value being 82.

BLOOD GLUCOSE CHART				
Mg/DL	Fasting	After Eating	2-3 hours After Eating	
Normal	80-100	170-200	120-140	
Impaired Glucose	101-125	190-230	140-160	
Diabetic	126+	220-300	200 plus	

4.2.Handling missing data:

Missing Data Count

```
glucose      304
education    87
BPMeds       44
totChol      38
cigsPerDay   22
BMI          14
heartRate    1
dtype: int64
```

Missing Data Percentage

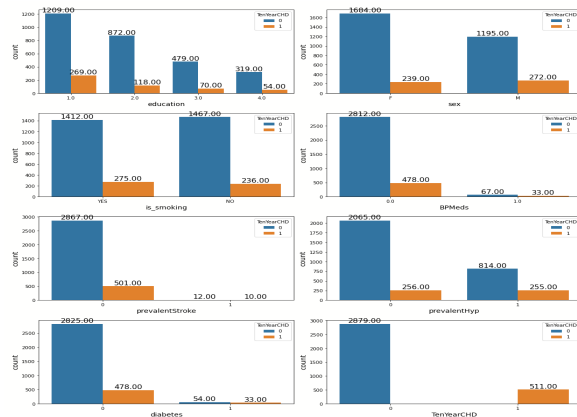
```
glucose      8.97
education    2.57
BPMeds       1.30
totChol      1.12
cigsPerDay   0.65
BMI          0.41
heartRate    0.03
dtype: float64
```

As we can see from the above table that none of the features have more than 10 % missing values.So we can fill up the null

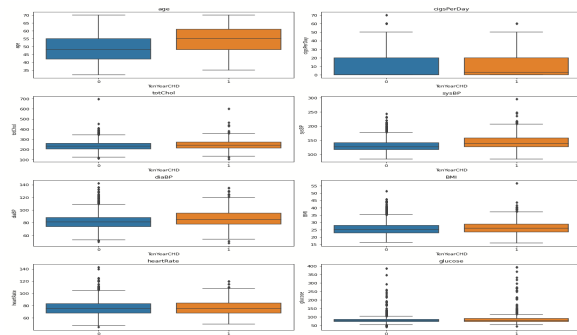
values and it won't create bias. For continuous features which have a nearly normal distribution, null values have been replaced with median values while for categorical values null values have been replaced with mode values.

4.3. EDA

Count plots were drawn for nominal/ordinal features divided w.r.t whether ten year CHD or not.

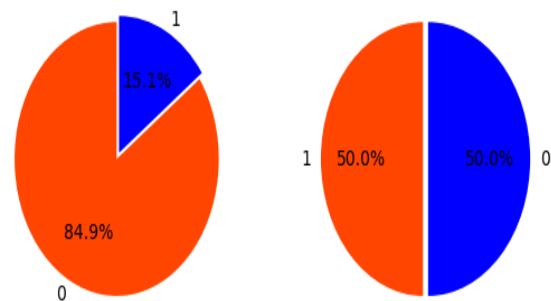
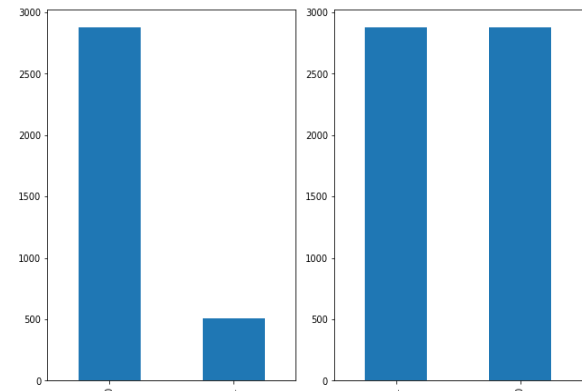


Boxplots were drawn for continuous/discrete features divided w.r.t whether ten year CHD or not.



4.4. Target variable data balancing of target variable using SMOTE

There was quite a high data imbalance in the target variable with no ten year CHD(0) being around 2700 observations and ten year CHD(1) being around 600. We can balance this imbalance using SMOTE technique.



4.5. Model fitting

The following classification models have been used :

- Logistic Regression
- K nearest neighbours Classifiers
- Support vector Classifier
- Decision Tree
- Random Forest Classifier
- Bagging Classifier
- Adaboost Classifier
- XG boost Classifier
- Catboost Classifier
- Stacking Classifier

4.6.Evaluation metrics analysis:

index	model	Accuracy	Recall	Precision	ROC-AUC	F1 Score
1	Logistic	0.61	0.95	0.57	0.6	0.71
2	KNN	0.82	0.92	0.77	0.81	0.84
3	SVC	0.79	0.82	0.79	0.79	0.81
4	Decision tree	0.75	0.83	0.71	0.74	0.77
5	Random forest	0.83	0.94	0.78	0.83	0.85
6	Bagging	0.83	0.94	0.78	0.83	0.85
7	Adaboost	0.88	0.95	0.85	0.88	0.89
8	XG boost	0.87	0.89	0.87	0.87	0.88
9	Catboost	0.85	0.88	0.83	0.85	0.86
10	Stacking	0.87	0.9	0.86	0.87	0.88

5.CONCLUSIONS:

Conclusions based on model performances-

1)Ensemble models are performing better as compared to other models with adaboost being the best performing model followed by stacking and bagging.

2)We can see from the evaluation metric table that the best performing model is adaboost classifier because it has the best accuracy and has the highest recall as well.

3)In this type of problem our priority should be to reduce the number of False Negatives or find maximum Recall score. If we misclassify someone as having no risk of heart disease, it can be highly detrimental, it can lead to loss of life. Adaboost gives us an excellent Recall and at the same time doesn't compromise on Precision. If we require a model with more strict Recall values we can opt for KNN.

Conclusions based on EDA-

1)Males are more prone to CHD as compared to females.

2)Smokers are more prone to CHD as compared to non smokers though it makes only a little difference.

3)People on BP medications are more prone to CHD.

4)People who had any previous prevalent stroke are more prone to having CHD although people who had prevalent stroke are very less.

5)Those who have prev hyp are more prone to having CHD.

6)Those who have diabetes are also more prone to having CHD although the number of people who have diabetes are very less.

7)Median age of people who have CHD is higher as compared to people who don't have CHD.This implies that people having a

higher age are more prone to having CHD.

8) People who had CHD smoke more cigarettes as compared to people who had no CHD.

9) People who have CHD have more total chol, sysBP, diaBP, BMI.

6. References

- Analytics vidhya
- Towards Data Science
- Stack overflow
- Geek for geeks