CAPSTONE PROJECT – Cardiovascular Risk prediction system

Manas Ranjan Behera
Data Science Trainee, Bangalore

## Abstract

Cardiovascular disorders are the leading causes of death in developed countries. The chapter provides an overview of behavioral and psychosocial influences on cardiovascular disorders, with an emphasis on coronary heart disease (CHD) and hypertension. This chapter reviews the pathophysiology of CHD, the role played by standard biological, behavioral, and psychosocial risk factors, including social determinants of health, environmental and psychological stress, individual psychological characteristics, and psychosocial protective factors such as social support. The chapter provides a summary of research examining the utility of interventions targeted at reducing risks of cardiovascular disease associated with psychosocial risk factors.

## 1 . Introduction

The annual mortality of CVD is expected to reach 23.6 million by 2030 (Alissa and Ferns, 2011). CVD is a group of diseases that includes coronary heart disease (CHD), cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, deep vein thrombosis, and pulmonary embolism (WHO, 2012). Sometimes the term circulatory diseases is used to encompass all of these diseases. CHD is a disease of the blood vessels supplying the heart muscle. This condition is often called ischemic heart disease (IHD) and acute myocardial infarction belongs to this entity. Cerebrovascular diseases are subdivided into ischemic and hemorrhagic diseases. Stroke is a widely used unspecific term for a group of cerebrovascular diseases of abrupt onset that cause neurological damage.

Approximately 85% of strokes are caused by inadequate blood flow to the brain, i.e. ischemic stroke.

The most common cause of CVD is atherosclerosis. Atherosclerosis has an inflammatory nature, which was described by the Austrian pathologist Carl von Rokitansky in the 1840s and by Rudolf Virchow somewhat later. Rokitansky believed that inflammation was secondary to other disease processes and Virchow promoted atherosclerosis as a primary inflammatory disease (Frostegård, 2010; Mayerl et al., 2006).

## Project Description

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries.

Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. In the United States, for example, it is estimated that someone has a heart attack every 40 seconds and about 805,000 Americans have a heart attack every year (CDC 2019).

Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

In this project, We will be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease(CHD) using different Machine Learning techniques.

## Project Workflow

1. Splitting to Train,Validation and Test sets to avoid Data bleeding.

2. Simultaneously Data Cleaning of Train and Test sets

3. EDA on features

4. Feature cleaning if needed

5. Solving Class Imbalanced problem

6. Base Model and Candidate Models

7. There Hypertuning

8. Bias-Variance tradeoff

9. Creating Voting Classifier

10. Final Predictions

## Feature Engineering

- From the above EDA we try to establish some patterns which influence the cause of heart disease in that we found people both men and women lying in a particular age group 40-42, 50-51 are more prone to heart disease.
- So what I want to try is to create age buckets of population e.g 18-25 -> 20s, 25-40 -> Mid30s etc in this way we can target the particular age group which have high risk of Heart disease.

## Class Imbalanced issue and Evaluation-metric to be chosen

In this problem we have a dataset of patients where we have to find out whether the given features or symptom a person has he/she has a Cardiovascular disease in future. But here's the catch… the risk rate is relatively rare, only 15% of the people have this disease.

## The Metric Trap

One of the major issues when dealing with unbalanced datasets relates to the metrics used to evaluate their model. Using simpler metrics like accuracy score can be misleading. In a dataset with highly unbalanced classes, the classifier will always "predict" the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one.

### *Synthetic Minority Oversampling Technique*

To solve this problem of imbalance in the dataset is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

An improvement on duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective.

Perhaps the most widely used approach to synthesizing new examples is called the Synthetic Minority Oversampling TEchnique, or SMOTE for short. This technique was described by Nitesh Chawla, et al. in their 2002 paper named for the technique titled "SMOTE: Synthetic Minority Over-sampling Technique." SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

## Different Models

**1 .** Logistic Regression

2. K-Nearest Neighbours

3. Decission Tree

4. Random Forest

5. XGBoosting

6. Naive Bayes

7. Support Vector Machine

After training each model and tuning their hyper-parameters using grid search, I evaluated and compared their performance using the following metrics:

- The accuracy score: which is the ratio of the number of correct predictions to the total number of input samples. It measures the tendency of an algorithm to classify data correctly.
- The F1 Score: Which is defined as the weighted harmonic mean of the test's precision and recall. By using both precision and recall its gives a more realistic measure of a test's performance. (Precision, also called the positive predictive value, is the proportion of positive results that truly are positive. Recall, also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate).
- The Recall: Which provides an aggregate measure of performance across all possible classification thresholds. It gives the probability that the model ranks a random positive example more highly than a random negative example

**1 – Logistic Regression:** Logistic Regression can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. We used GridSearch to tune the hyper parameters of logistic regression to get the best possible test score.

**2 – K-Nearest Neighbours:-** K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

**3 – Decision Tree Classifier** :The decision trees was also built on the training data in order to improve prediction accuracy .Here also we used GridSearch to tune the hyper parameters of Decision Tree to get the best possible test score.

**4 – Random Forest Classifier** : It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

**5 – XGBoosting** : It was used for final prediction of the trip duration in the test dataset. The dataset was very large, as a result for this type of problem XGBoost was applied in which all the attributes were taken and parallel processing of boosting trees executed. Another aspect of XGBoost is that it keeps a nice check between bias and variance which helps in better prediction. The results were interpreted by using GridSearch, the XGBoost hyper parameters.

**6 – Naïve Bayse :-** A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

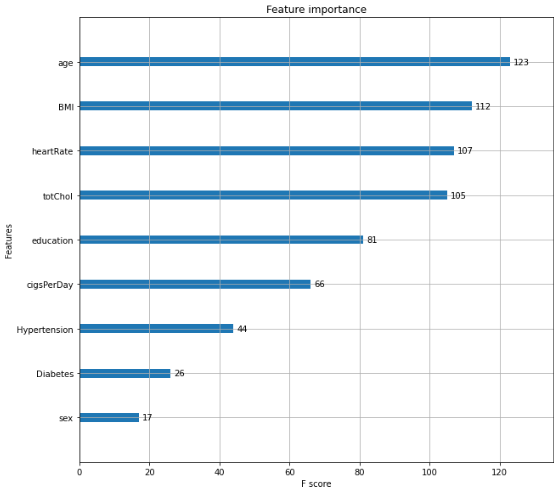
**7 – Support Vector Machine** : We use SVMs is because they can find complex relationships between your data without you needing to do a lot of transformations on your own. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

## Feature Importance

Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

Age, Total Cholesterol level, Heartrate were the important features in the case of XGBoost Classifier Model.



## Conclusion

In this Cardiovascular Risk Prediction Project We tried to fit various models to our dataset to predict the risk of Coronary Heart Disease for next ten years.

Age, Total Cholesterol level, Heartrate were the important features in the case of XGBoost Classifier Model.

Naive Bayes classifier and Logistic Regression were not able to perform well in the prediction of target variable.

XGBoost, Support Vector Machine and Random Forest were better than the rest models.