CARDIOVASCULAR RISK PREDICTION

Prajwal D U, Data Science Trainee, AlmaBetter, Bangalore.

ABSTRACT :
The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants. Prior to the study almost nothing was known about the epidemiology of hypertensive or arteriosclerotic cardiovascular disease. Much of the now-common knowledge concerning heart disease, such as the effects of diet, exercise, and common medications such as aspirin, is based on this longitudinal study. It is a project of the National Heart, Lung, and Blood Institute, in collaboration with (since 1971) Boston University. Various health professionals from the hospitals and universities of Greater Boston staff the project.

INTRODUCTION :
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.

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 17 attributes.

Data Description :

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)

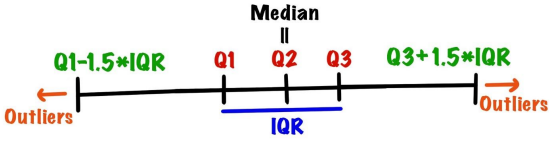
- 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

MISSING VALUE TREATMENT :

Missing data or missing values occur when no data stored null variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Continuing to that we found missing observations in seven columns which we further treated with its median value that corresponds to that column.

HANDLING OUTLIERS :

A data point that varies greatly from other results is referred to as an outlier. An outlier may also be described as an observation in our data that is incorrect or abnormal as compared to other observations.



To find outliers, we can simply plot the box plot for every feature present in the datasets. Outliers are points that are outside of the minimum and maximum values. We have treated the outlier that lies away from the upper boundary and lower boundary with its median values (or 50th percentile).

CLEANING AND MANIPULATING THE DATASET :

Manipulating dataset mean the information to make it more effective and readable,

- Dropping duplicate values in the dataset.
- Checking null values and filling null values.
- Dropping unwanted features like i'd.
- Checking value counts on categorical features to give us some intuition about the feature.
- Defining label encoding in features, it refers to converting the label into numeric form as to convert it into machine readable form.

UNIVARIATE ANALYSIS :

Uni means one so we called it as univariate analysis, This is most basic type of analysis,

In univariate analysis we analyse one variable and find mean and median, here we look at the distribution of features separately.

BIVARIATE ANALYSIS :

In bivariate analysis we find relation between single independent variables and dependent variables, using bivariate analysis association and dissociation between variables at a pre-defined significance level, we can perform bivariate analysis from any combination of categorical or continuous variables.

MULTICOLLINEARITY :

- It tells us how one variable depends on other variables that mean if we change one variable value how it's affect our dataset,
- Multicollinearity is a statistical concept where several independent variables in a model are correlated.
- Two variables are considered to be perfectly collinear if their correlation coefficient is +/- 1.0.
- It is better to use independent variables that are not correlated or repetitive when building multiple regression models that use two or more variables.
- It is better to use independent variables that are not correlated or repetitive when building multiple regression models that use two or more variables.
- Sys_bp, dia_bp, BMI, glucose, heartrate, age, total_chol poses to be highly dependent variables which is not good to perform models on top of it.

VARIANCE INFLATION FACTOR :

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. Generally, we consider a VIF score less than 5.

$$VIF_i = \frac{1}{1 - R_i^2}$$

After considering the features with VIF score less than 10 we are finally left with the following features - age, education, sex, cigsperday, prevalentHYP, BPmeds, diabetes, prevalentstroke.

MODEL BUILDING

PREREQUISITES :

CLASS IMBALANCED ISSUE :

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 our 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. Hence we need to address the imbalance in the classes. There are several ways to tackle this. Lets see how:

RANDOM OVER SAMPLING :

Oversampling can be defined as adding more copies to the minority class. Oversampling can be a good choice when you don't have a ton of data to work with. A con to consider when under-sampling is that it can cause overfitting and poor generalisation to your test set.

SMOTE - TOMEK :

This method combines the SMOTE ability to generate synthetic data for minority class and Tomek Links ability to remove the data that are identified as Tomek links from the majority class,We only oversampled the train data, test data remains untouched form making synthetic duplicates,If you balance the Validation set (test data), your model may work well (may get better score in Val) but in the future after deploying, it may not work better so while training, validate with imbalance data only.

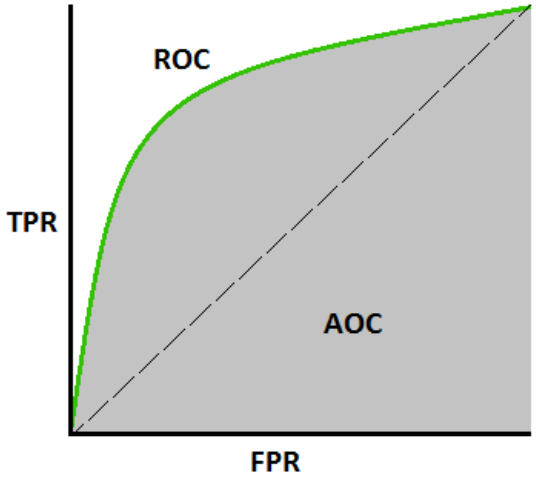
CONFUSION MATRIX :

A Confusion matrix is an N x N matrix used for evaluating the **performance of a classification model**, where **N** is the number of **target classes**. The matrix compares the actual target values with those predicted by the machine learning model.

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

AREA UNDER CURVE - RECEIVER OPERATING CHARACTERISTICS (AUC-ROC) :

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings.



ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

CLASSIFICATION REPORT :

A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report.

METRICS :

- **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.
- **Precision** : Precision is the ratio between the True Positives and all the Positives. For our problem statement, that would be the measure of patients that we correctly identify having a heart disease out of all the patients actually having it.

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

- **Recall** : The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart

disease.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

- **F1 score** : There are also a lot of situations where both precision and recall are equally important. For example, for our model, if the doctor informs us that the patients who were incorrectly classified as suffering from heart disease are equally important since they could be indicative of some other ailment, then we would aim for not only a high recall but a high precision as well. In such cases, we use something called F1-score. F1-score is the Harmonic mean of the Precision and Recall,

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

FEATURE IMPORTANCE

It assigns the score of input features based on their importance to predict the output. More the features will be responsible to predict the output, the more will be their score.We can use it in both classification and regression problems.

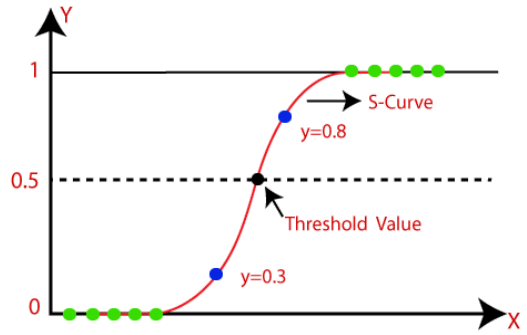
HYPERPARAMETER TUNING

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process.

- n_neighbour - Number of neighbours to use by default 5.
- n-estimator - number of trees in the forest, by default 100.
- max_depth - It governs the maximum height up to which the trees inside the forest can grow. The default is set to None.
- min_samples_split - It specifies the minimum number of samples an internal node must hold in order to split into further nodes. However, the default value is set to 2.
- min_samples_leaf - It specifies the minimum number of samples that a node must hold after getting split. The default value is set to 1.
- learning_rate - The learning rate determines the step size at each iteration while your model optimises toward its objective. A low learning rate makes computation slower, and requires more rounds to achieve the same reduction in residual error as a model with a high learning rate.
- Kernel - Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Some examples are radial basis function, Gaussian, sigmoid and polynomial kernel.

LOGISTIC REGRESSION :

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc.



but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.So, logistic regression cost function is

NAIVE BAYES CLASSIFIER :

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Why called **naive** because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of colour, shape, taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Why called **bayes** because it depends on the principle of Bayes' theorem.

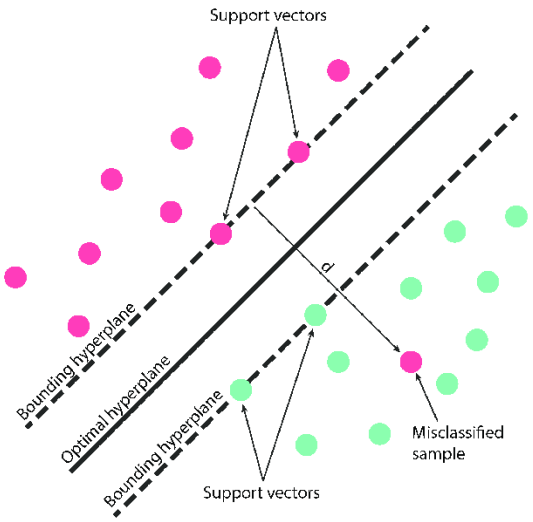
the formula for bayes' theorem is

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

- **P(A|B) is Posterior probability** - Probability of hypothesis A on the observed event B.
- **P(B|A) is Likelihood probability** - Probability of the evidence given that the probability of a hypothesis is true.
- **P(A) is Prior probability** - Probability of hypothesis before observing the evidence.
- **P(B) is marginal probability** - Probability of Evidence.

SUPPORT VECTOR CLASSIFIER (SVM) :

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimised. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

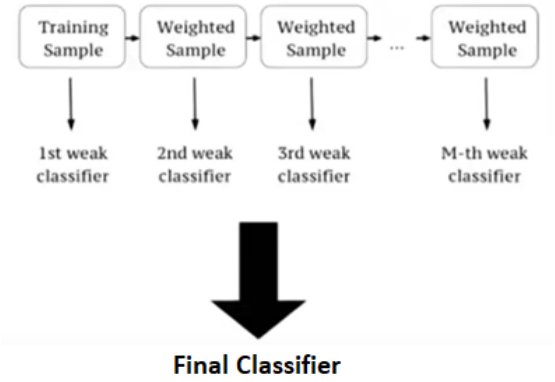


RANDOM FOREST CLASSIFIER :

The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction.

In Layman's term, the training set is given as: [X1, X2, X3, X4] with corresponding labels as [L1, L2, L3, L4], random forest may create three decision trees taking input of a subset of it.

XGBOOST CLASSIFIER :



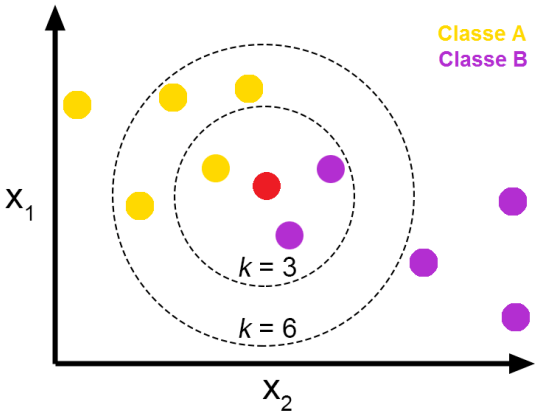
In this algorithm, decision trees are created in sequential form. **Weights** play an important role in **XGBoost**. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model.

It can work on regression, classification ranking and user-defined prediction problems.

K - NEAREST NEIGHBOURS :

- **Instance-based learning** : Here we do not learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data.
- Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance.
- Non -Parametric: In KNN, there is no predefined form of the mapping function.

The distance between the two data points is calculated by the following methods - **Euclidean distance**, **Hamming distance**, **Manhattan distance**, **Minkowski distance**. Euclidean is most popular amongst all.



CONCLUSION :

- If we want to completely avoid any situations where the patient has heart disease, a high recall is desired. Whereas if we want to avoid treating a patient with no heart diseases a high precision is desired.
- Assuming that in our case the patients who were incorrectly classified as suffering from heart disease are equally important since they could be indicative of some other ailment, so we want a balance between precision and recall and a high f1 score is desired.
- Since we have added synthetic data points to handle the huge class imbalance in the training set, the data distribution in train and test are different so the high performance of models in the train set is due to the train-test data distribution mismatch and not due to overfitting.
- Best performance of Models on test data based on evaluation metrics for class 1,
    - Recall - SVC
    - Precision - Naive Bayes Classifier
    - F1 Score - Logistic Regression, XGBoost
    - Accuracy - Naive Bayes Classifier.