CAPSTONE PROJECT 3

Classification Model

Cardiovascular Risk Prediction

Prepared by:

PRAVEEN .S

(Cohort Everest)

Data Science Trainee

Almabetter

Abstract:

Coronary heart disease **(CHD)** is the most common type of heart disease. It is also called coronary artery disease (CAD). CHD is plaque buildup in your arteries. It's known as hardening of the arteries, too. Arteries carry blood and oxygen to your heart.

Coronary heart disease is now the leading cause of death worldwide. An estimated 3.8 million men and 3.4 million women die each year from CHD1. In developed countries heart disease is the leading cause of death in men and women.

Our experiment can help understand what could be the chance of diagnosing for CHD using logistic regression of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the current trend.

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 **4,000** records and **15** attributes.

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

Introduction:
I had presented my exploratory analysis,,visualization,Correlation plots and developed three machine learning models and lot of other interesting insights into the given dataset.I choose this particular dataset as this is the domain I have expertise in my bachelors.My goal here is to build a predictive model, which could help in reducing death rate as well as it will create an awareness among others in their future endeavors.

Steps involved:

Exploratory Data Analysis

After loading the dataset we performed this method by comparing our target variable that is “TenYearCHD” with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

Null values Treatment

Our dataset consist of approximately 12% of null values.Owing to this fact, we have dropped all the null value entries or rows.We made this decision as 12 % of null values will create huge shift in accuracy and providing mean values will change the originality of the dataset.

Data Visualization:

Using Data visualization,we found that the features(sex,is_smoking,BPMeds,prevalentStroke,prevalentHyp ,diabetes) are to be nominal as they consists of binary values.And features(age,cigsperDay,totChol,sysBP,diaBP,BMI,heartrate,glucose) are to be continous.The target variable"TenYearCHD" consists of two outputs which are nominal where there is a signifiant difference in between the two values.Since it is classification dataset,it is hard to find variables which are correlated to each other

Data Preparation:

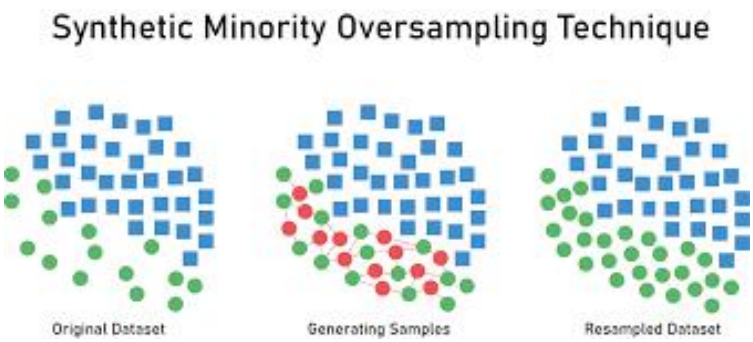
The features "sex" and "is_smoking" are found to have two values which are nominal.Thus we have converted the values "M" and "F" to binary values 1 and 0 and similarly for "Yes" and "No" to 1 and 0.We have also implemented two techniques for feature selection and Handling class imbalances.

ChiSquare test:

Using Chisquare test,As per the above logistic regression results, P >= 0.05 shows a low statistically significant relationship with the probability of heart disease. Hence, a backward elimination approach has been used to remove the attributes with the highest P values. The process will be continued until all the attributes of P values less than 0.05.thus we get the following features to be significant than others(sex,age,cigsPerDay,totalChol,sysBP,glucose)

SMOTE:

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling.. It potentially performs better than simple oversampling and it is widely used.



Fitting different models

For modelling we tried various classification algorithms like:

- Logistic Regression
- K Nearest Neighbour
- Support Vector Machine

Tuning the hyperparameters for better accuracy

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting and underfitting

Algorithms:

Logistic Regression:

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

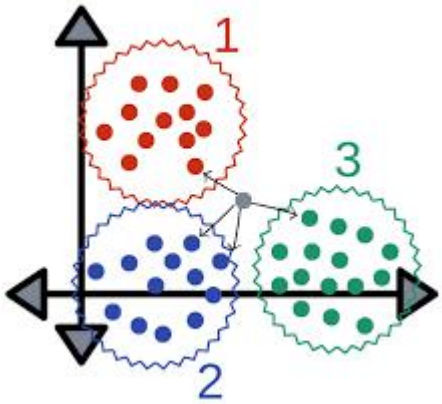
$$\beta_0 + \beta_1 * Sex + \beta_2 * age + \beta_3 * cigsPerDay + \beta_4 * totChol + \beta_5 * sysBP + \beta_6 * glucose$$

K Nearest Neighbour:

The k-nearest-neighbors is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it.

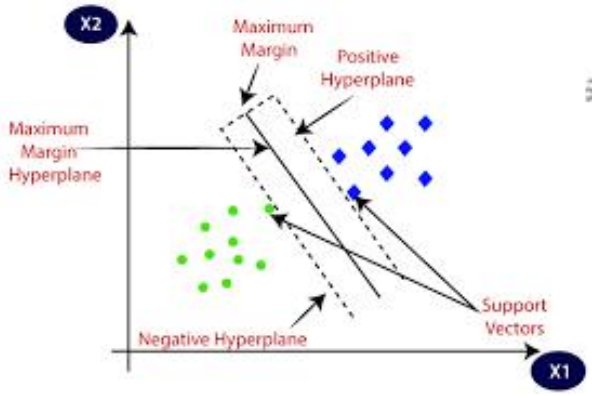
An algorithm, looking at one point on a grid, trying to determine if a point is in group A or B, looks at the states of the points that are near it. The range is arbitrarily determined, but the point is to take a sample of the data. If the majority of the points are in group A, then it is likely that the data point in question will be A rather than B, and vice versa.

The k-nearest-neighbor is an example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand. The only calculations it makes are when it is asked to poll the data point's neighbors. This makes k-nn very easy to implement for data mining.



Support Vector Machines:

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.



Model performance:

Classification model can be evaluated by various metrics such as:

1.Confusion Matrix:

The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

2.Precision/Recall:

- Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP
- Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

3.Accuracy:

Accuracy is given by the number of correctly classified examples divided by the total number
of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

4.Area under ROC Curve(AUC):

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance

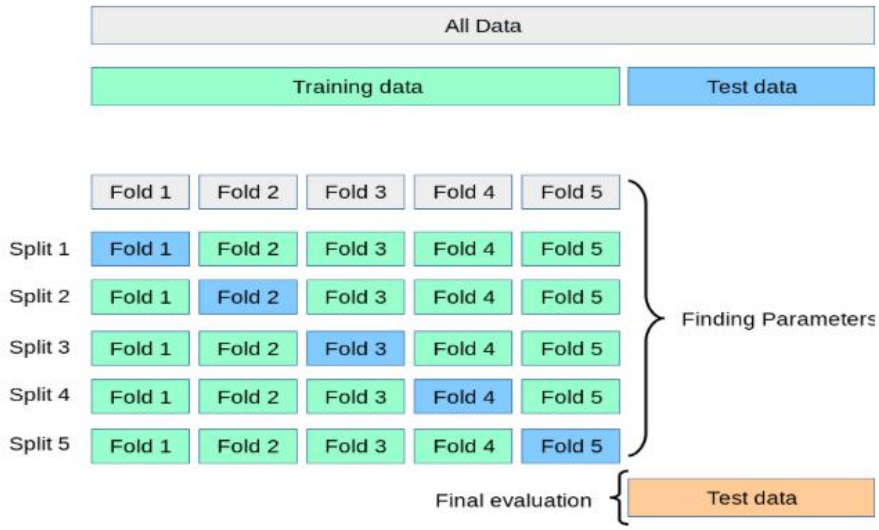
Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

There are three types namely Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Grid Search CV.

Grid Search CV:

Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.



8. Conclusion:

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.In all of these models our accuracy revolves in the range of 80 to 85%.And there is no such improvement in accuracy score even after hyperparameter tuning.So the accuracy of our best model is 83% which can be said to be good for this large dataset. This performance could be due to various reasons like: no proper pattern of data, too much data, not enough relevant features etc.

References-

1. **MachineLearningMastery**
2. **GeeksforGeeks**
3. **Analytics Vidhya**