Cardiovascular Risk Prediction

Dharmeshbhai Patel

(PRO-Program)

Abstract:

The independent variables such as age, education, is_smoking etc.… are the determinants of the dependent variable 'TenYearCHD'. I was provided with already classified labels in our data set. Our experiment can help understand what could be the reason for the classification of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the correct classification.

Keywords: machine learning, TenYearCHD, support vector machines, classified labels

1. 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. Variables Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

2. Introduction

The TenCHD is predicted using the following Independent Variables:

- **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 has 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):

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

Our goal here is to build a predictive model, which could help the hospitals in predicting Chronic Heart Disease proactively.

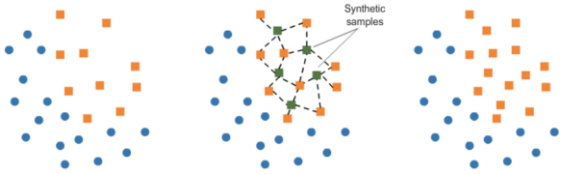
3. Steps Involved

- Exploratory Data Analysis
- Null Value Treatment
- Encoding of Categorical columns
- Feature Selection
- Standardization of features
- Fitting Different Models
  For modelling I tried various models:
  1. Logistic Regression
  2. Random Forest Classifier
  3. XG-Boost Classifier
  4. Support Vector Machine

4. Application of SMOTE to the Dataset

This technique generates synthetic data for the minority class.

SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbours for this point. The synthetic points are added between the chosen point and its neighbours.



SMOTE algorithm works in four steps:

- Choose a minority class as the input vector

- Find its k nearest neighbors (k_neighbors is specified as an argument in the SMOTE() function)
- Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor
- Repeat the steps until data is balanced

5. Algorithms

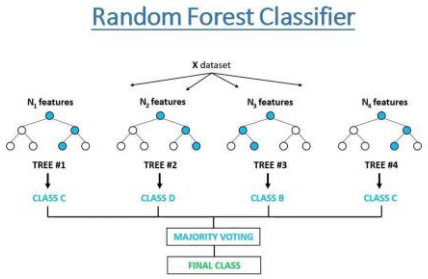
1. Logistic Regression

Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by: **f(x)= 1/1+e ^(-x)**

2. Random Forest Classifier

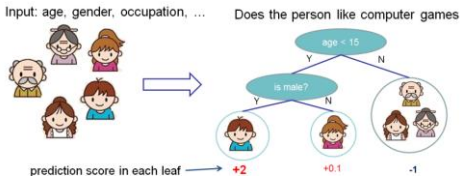
Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.



3. XGBoost

To understand XGBoost we have to know gradient boosting beforehand.

- Gradient Boosting- Gradient boosted trees consider the special case where the simple model is a decision tree



In this case, there are going to be 2 kinds of parameters P: the weights at each leaf, w, and the number of leaves T in each tree (so that in the above example, T=3 and w=[2, 0.1, -1]).

When building a decision tree, a challenge is to decide how to split a current leaf. For instance, in the above image, how

could I add another layer to the (age > 15) leaf? A 'greedy' way to do this is to consider every possible split on the remaining features (so, gender and occupation), and calculate the new loss for each split; you could then pick the tree which most reduces your loss.

XGBoost is one of the fastest implementations of gradient boosting. trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits

4. **Support Vector Machine Classifier**

SVM is used mostly when the data cannot be linearly separated by logistic regression and the data has noise. This can be done by separating the data with a hyperplane at a higher order dimension. In SVM I use the optimization algorithm as:

Consider the general SVM optimization problem:

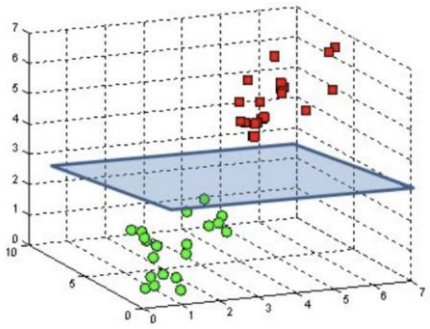
$$\begin{aligned} \min_{\xi,w,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t} \quad & y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0;\ i = 1, \ldots, m. \end{aligned}$$

where C is a cost parameter and ξ_i's are slack variables.

(a) Write the Primal and Dual problems.

(b) Derive the equation of the Dual problem and apply the KKT conditions to identify the support vectors.

(c) (optional) Write an R program to solve the Dual optimization problem numerically using the SMO Algorithm.



6. Model Performance

Model can be evaluated by various
metrics such as:0

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)0**

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.

7. Conclusion:

Starting with loading the data so far, I 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 65 to 90%.

So, the accuracy of our best model is 89% 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.