Cardiovascular Risk Prediction

Sumit Kumar Dhibar
Data Science Trainee
AlmaBetter, Bangalore

Abstract:

Coronary heart disease is a type of heart disease where the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. Coronary artery disease affects the larger coronary arteries on the surface of the heart. Another type of heart disease, called coronary microvascular disease, affects the tiny arteries in the heart muscle. Coronary microvascular disease is more common in women. Symptoms of coronary heart disease may be different from person to person, even if they have the same type of coronary heart disease. However, because many people have no symptoms, they do not know they have coronary heart disease until they have chest pain, blood flow to the heart is blocked, causing a heart attack, or the heart suddenly stops working, also known as cardiac arrest.

This study is based on an ongoing cardiovascular study of residents of the town of Framingham, Massachusetts. And our goal is to predict whether the patient has a 10-year risk of future coronary heart disease (**CHD**). We were provided with a dataset having information about people like age, gender, medication history, and whether that person was prone to heart disease or not in the coming ten years. We did **Univariate Analysis**, **Bivariate Analysis**, Exploratory Data Analysis (**EDA**) and **Data Wrangling** on the given data. We observed that we had imbalanced target labelled data, so we picked recall, precision, and f1 score as our metrics to compare the performance of classification algorithms. We also resampled the data to get a balanced dataset, then again applied machine learning algorithms to find the improvement in predicting the target variable for unseen data. We realized that the resample technique gave good results.

Different classification algorithms were used for machine learning models. They were **Logistic Regression, Naive Bayes Classifier**, **Decision Tree Classifier**, **Random Forest Classifier**, **KNN Classifier**, and **XGBoost Classifier**.

Keywords: EDA, CHD, Logistic Regression, Naïve Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, KNN Classifier, XGBoost Classifier.

1. Problem Statement:

Look at the given dataset and find the distribution of each feature. Look for null values and outliers and replace or remove them if it is meaningful for our study. 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 patient's information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic and behavioral and medical risk factors. Use appropriate metrics to compare the performance of machine learning algorithms.

2. Data Description:

2.1 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)

2.2 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.)

2.3 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)
• SysBP: systolic blood pressure (Continuous)
• DiaBP: 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)
• **Target variable: t**he patient has a 10-year risk of future coronary heart disease (CHD).

3. EDA and Data Wrangling:

3.1 Null values treatment:

Null values are present in features named glucose, education, BP medication, total cholesterol, cigarettes per day, BMI and heart rate. The null values in above features replaced with respective feature values median.

3.2 Univariate Analysis:

A histogram plot is used to determine the distribution of numerical features and categorical features. We also used a count plot too. The ages range from 30 to 70. Cigarettes per day range from 0 to 70. Total cholesterol ranges from 100 to 700. Systolic blood pressure ranges from 100 to 300. Diastolic blood pressure ranges from 40 to 140. The BMI ranges from 15 to 55. The heart rate ranges from 40 to 140 beats per minute. The glucose level is from 40 to 400. We observed outliers in our dataset; outliers in total cholesterol, systolic blood pressure, diastolic blood pressure, body mass index, heart rate, and glucose level. If we remove all outliers without considering their importance, we may lose important pieces of our study.

The people in our dataset are divided into four groups. As education levels increase, the number of people in that group also decreases. The percentage of females is higher compared to males. Half of people are smokers. Most of the people are not on BP medication. A few people (around 60 people among 3390) had prevalent strokes. Among 3390 people, around 1050 had hypertension issues. Less than 100 people in our data set had diabetes. Around 500 people are prone to heart disease in the coming ten years and others are free.

3.3 Bivariate Analysis:

The number of people with heart disease decreases with education level because most people are from level 1 and fewer are from level 4. Even though females are high in our data set, the males are contributing more positive cases of heart disease compared to their female counterparts. Cigarette smokers are more prone to heart disease compared to non-smokers. The percentage of people having heart disease is higher for prevalent stroke people compared to non-prevalent stroke people. Same with people on BP medication.

3.4 New Features Creation:

Two feature systolic blood pressure and diastolic blood pressure were used to create new feature called blood pressure ratio which is equal to ratio of systolic and diastolic blood pressure. Cigarettes per day was created using is smoker and cigarettes per day.

3.5 Dropping Features:

Now we already had information about systolic and diastolic blood pressures in blood pressure ratio. We removed these both features systolic and diastolic blood pressures. Also, we removed features named is smoking and cigarettes per day.

3.6 Resampling:

We had imbalance data as heart disease positive cases were around 500 and negative cases are more than 2700. So, we up sampled the positive cases.

4. Steps involved in Model Building:

4.1 Scaling the data: Because the units and quantities of different feature different. So, the importance may go to unimportant feature. To overcome this difficulty we choose scaling, it can be done by using minmax scaler and standardised scaler.

4.2 Fitting different models: For modelling we tried various classification algorithms. They are

- Logistic Regression
- Naive Bayes Classifier
- Decision Tree Classifier
- KNN Classifier

- XGBoost Classifier
- Random Forest Classifier

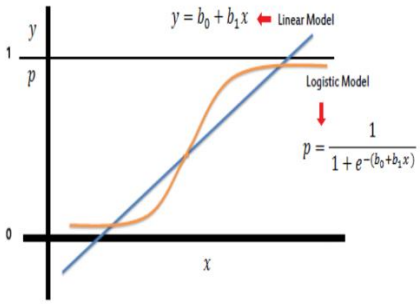
4.3 Tuning the hyperparameters for better accuracy:

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree-based models like Decision Classifier, Random Forest Classifier.

5. Algorithms:

5.1 Logistic Regression:

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. The function used in Logistic Regression is sigmoid function or the logistic function shown in below figure along with linear regression.

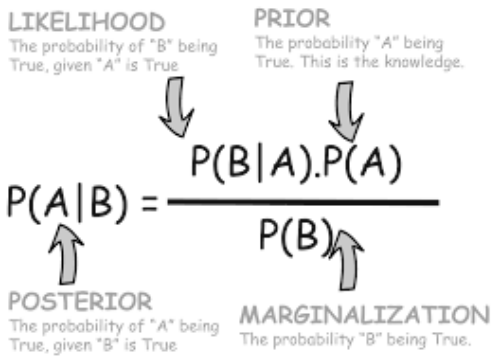


The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:

$$\ln L(\mathbf{y},\beta) = \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

5.2 Naive Bayes Classifier:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:



where A and B are events and P(B) ≠ 0. With regards to our dataset, we can apply Bayes' theorem in following way:

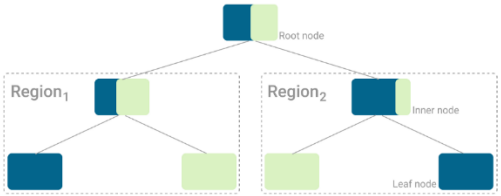
$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

We know that features independent to each other. If A and B are independent to each other. Then P (A, B) = P(A)*P(B). Then we reach to the result as shown below:

$$P(Y|X) = \frac{P(X_1|Y).P(X_2|Y) \ldots \ldots P(X_n|Y).P(Y)}{P(X_1).P(X_2) \ldots \ldots P(X_n)}$$

5.3 Decision Tree Classifier:

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



The algorithm tries to completely separate the dataset such that all leaf nodes, i.e., the nodes that don't split the data further, belong to a single class. These are called pure leaf nodes. On every split, the algorithm tries to divide the dataset into the smallest subset possible. So, like any other Machine Learning algorithm, the goal is to minimize the loss function as much as possible. A loss function that compares the class distribution before and after the split, like Gini Impurity and Entropy. Gini Impurity is formulated as below:

$$\mathrm{G(node)} = \sum_{k=1}^{c} p_k(1 - p_k)$$

Probability of *not* picking a data point from class k

Probability of picking a data point from class k

$p_k = \frac{\text{number of observations with class k}}{\text{all observations in node}}$

Similarly, to Gini Impurity, Entropy is a measure of chaos within the node. And chaos, in the context of decision trees, is having a node where all classes are equally present in the data.

$$\mathrm{Entropy(node)} = -\sum_{i=1}^{c} p_k \log(p_k)$$

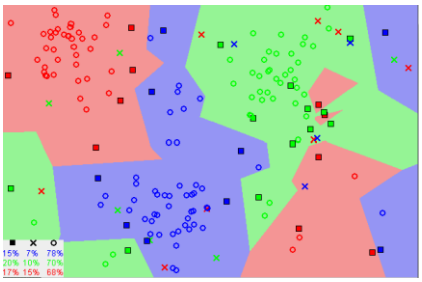
Probability of picking a data point from class k

$p_k = \frac{\text{number of observations with class k}}{\text{all observations in node}}$

Using Entropy as loss function, a split is only performed if the Entropy of each the resulting nodes is lower than the Entropy of the parent node. Otherwise, the split is not locally optimal.

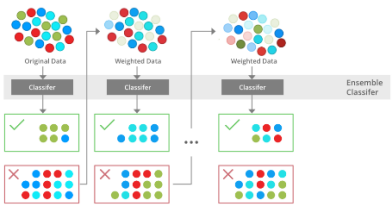
5.4 KNN Classifier:

The k-nearest neighbours (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). Result of KNN Classification on a dataset looks like as shown below:



5.5 XG Boost Classifier:

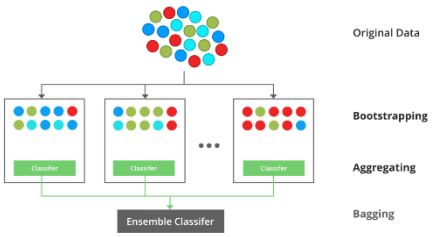
XG Boost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. XG Boost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees. In contrast to bagging techniques like Random Forest, in which trees are grown to their maximum extent, boosting makes use of trees with fewer splits. Such small trees, which are not very deep, are highly interpretable. The boosting technique is shown in below figure:



Parameters like the number of trees or iterations, the rate at which the gradient boosting learns, and the depth of the tree, could be optimally selected through validation techniques like k-fold cross validation. Having a large number of trees might lead to overfitting. So, it is necessary to carefully choose the stopping criteria for boosting.

5.6 Random Forest Classifier:

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Bagging reduces overfitting (variance) by averaging or voting. The bagging classifier shown in below figure:



Random Forest is a bagging type of decision tree algorithm that builds several decision trees from a randomly chosen subset of the training set, gathers the labels from these subsets, and then averages the final prediction based on the most instances of a label that have been did predict out of all of them.

6. Model Performance:

6.1 Confusion Matrix:

Confusion Matrix is the visual representation of the Actual VS Predicted values. It measures the performance of our Machine Learning classification model and looks like a table-like structure. In below figure you can see that confusion matrix table.

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

Elements of Confusion Matrix:

TP: True Positive: The values which were actually positive and were predicted positive.
FP: False Positive: The values which were actually negative but falsely predicted as positive. Also known as Type I Error.
FN: False Negative: The values which were actually positive but falsely predicted as negative. Also known as Type II Error.

TN: True Negative: The values which were actually negative and were predicted negative.

6.2 Accuracy:

It is calculated by dividing the total number of correct predictions by all the predictions. Its formula is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

6.3 Recall:

In an imbalanced classification problem with two classes, recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

6.4 Precision:

Precision checks how many outcomes are actually positive outcomes out of the total positively predicted outcomes. Formula for precision is:

$$Precision = \frac{TP}{TP + FP}$$

6.5 F1 Score:

F1 score is the harmonic mean of Precision and Recall and it captures the contribution of both of them. Formula for F1 score is:

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

6.6 Area under 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. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

7. Conclusions:

Predicting the risk of coronary heart disease is critical for reducing fatalities caused by this disease, we can avert deaths by taking required medications and precautions if we can foresee the danger of this illness ahead of time.

- It is important to have a high recall score in this scenario because It is okay if the model incorrectly identifies a healthy person as a high risk patient, because it will not result in death, but if a high risk patient incorrectly identified as healthy, it may result in fatality. Support Vector Machine with rbf kernel is the best model with recall score of 0.88.
- There may be a case where the patients who are incorrectly classified as suffering from heart disease is equally important as patients who are correctly classified as suffering from heart disease, because patients who are incorrectly classified they may have some other illness, so in that case high f1 score is desired. Logistic Regression, XGBoost, K-NN these are the model with most F1 score.
- From our analysis, it is found that the 'Age' of the patient is the most important feature in determining the risk of coronary heart disease, middle and older age people are more prone to coronary heart disease than younger people followed by 'cigarettes per day', 'BP Meds', 'Prevalent Hypertension' are also very important feature in determining risk of heart disease.

- Future developments must include a strategy to improve models scores with the help of more data from people with different medical history.

References:

• Geeks For Geeks
• Analytics Vidhya
• Wikipedia