Technical Document

Cardiovascular Risk Prediction

Prepared By

Nitin Solanki
(solankinitin1210@gmail.com)

Contents

1. Objective

Cardiovascular disease is the leading cause of death worldwide and a major public health concern. Therefore, its risk assessment is crucial to many existing treatment guidelines.
Risk estimates are also being used to predict the magnitude of future cardiovascular disease mortality and morbidity at the population level and in specific subgroups to inform policymakers and health authorities about these risks.
We have been given a dataset from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts that provides the patients' information
It includes over 3,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.
The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

2. Introduction

The heart is one of the main organs of the human body. The heart plays the most crucial role in the circulatory system. If the heart does not function properly then it will lead to serious health conditions including death.

Our problem is that we want to predict whether patients have heart disease by giving some features to users. This is important to medical fields. If such a prediction is accurate enough, we can not only avoid wrong diagnoses but also save human resources. When a patient without heart disease is diagnosed with heart disease, he will fall into unnecessary panic and when a patient with heart disease is not diagnosed with heart disease, he will miss the best chance to cure his disease. Such a wrong diagnosis is painful to both patients and hospitals. With accurate predictions, we can solve the unnecessary trouble. Besides, if we can apply our machine learning tool into medical prediction, we will save human resource because we do not need complicated diagnosis process in hospitals. The input to our algorithm is 15 features with number values. We use several algorithms such as Logistic Regression, SVM, Random Forest, Gradient bosting, stacking, KNN to output a binary number 1 or 0. 1 indicates the patient will have the risk of heart disease and vice versa

3. Dataset

We have been given a dataset from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts that provides the patients' information. It includes over 3,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors

Demographic Attributes:
Sex: Categorised as either 0 or 1 such as 0 = female and 1 = male
Age: Age of the patient at the current time of examination
Education: It is an inessential data, because any medical issue doesn't occur as per someone's education level
Behavioural:
Is_smoking: Categorised as either 0 or 1 depending upon whether a patient smoke currently or not i.e., 1 = yes and 0 = no
Cigs Per Day: Depends upon if someone is smoking regularly, then the average no. of cigarettes being smoked by him.
Previous medical history-based information:
• Diabetes: Categorised as either 0 or 1 depending upon whether a patient had diabetes or not, 1 means Yes & 0 means No.
• BP Meds: Categorised as either 0 or 1 depending upon whether a patient is based on medication for blood pressure or not, 1 means having blood pressure medication and 0 means not having blood pressure medication.

• Prevalent Stroke: Categorised as either 0 or 1 depending upon whether the patient had a stroke previously or not, where 1 represents YES & 0represents NO.
• Prevalent Hyp: Categorized as either 0 or 1 depending upon whether the patient was hypertensive (i.e., having abnormally high blood pressure), 1 means Yes & 0 means No.

Current medical condition-based information:
• Heart Rate: heart rate
• Tot Chol: total cholesterol level
• Sys BP: systolic blood pressure
• Dia BP: diastolic blood pressure
• BMI: Body Mass Index
• Glucose: glucose level

Target variable to predict:
• 10-year risk of CHD - (binary: 1 means "Yes" & 0 means "No")

4. Data Modification and Exploration

In this section, Data modification, exploration and pre-processing will be explained, which is required before we input our data set to any machine learning algorithm in order to improve prediction for machine learning models.

Data exploration and pre-processing part will be done using pandas and NumPy packages. All the graphs and figures will use Matplotlib and seaborn package.

Null Values
The first step is to deal with Null values form due to which the dataset becomes redundant and hence leads the models to predict results with poor accuracies. We usually clean the tuples having missing values by either dropping those tuples from dataset or by imputing mean or median values of respective column

in our proposed model we are using mean and median imputation approaches for imputing missing values in data set for attaining its consistency to achieve higher accuracy. Mean/Median imputation is the way of replacing missing values (i.e., 'NA' or 'NULL') data in dataset by mean/median of that parameter.

It can be described as whenever the parameter represents a normal distribution then we can use any one of both mean and median imputation. But if the parameter represents skewed distribution (Just like us continues variable) instead of normal distribution, then the median imputation is preferred over the mean imputation.

Outliers
We know that outliers are one of the main problems when building a predictive model. Indeed, they cause to achieve poorer results than they could. To solve that, we need effective methods deal with that spurious points and remove them.
There are different methods for dealing with outliers, first step is to detect the outliers and to do so we have used IQR method (below steps are performed)

- Sort the dataset in ascending order
- calculate the 1st and 3rd quartiles (Q1, Q3)
- compute IQR=Q3-Q1
- compute lower bound = (Q1–1.5*IQR), upper bound = (Q3+1.5*IQR)
- loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outliers

After finding the outlier to deal with them winsorising technique has been used, where outliers have been replaced with a lower limit or upper limit

Class Imbalance
most machine learning algorithms do not work very well with imbalanced datasets. There are several techniques can help to train a classifier to detect the abnormal class. Like over sampling, Under Sampling, SMOTE
Here we have used the SMOTE, SMOTE Simply adds duplicate records of minority class and often doesn't add any new information to the model. SMOTE has created 2368 new records and updated dataset have 5758 records

5. Implementation and Evaluation Metrics

In this Analysis, the data is split with a ratio of 75:25. All models will be optimized for better accuracy using Grid Search. Grid-search is used to find the model's optimal hyperparameters that provide the most 'accurate' predictions. The function is fed with different hyperparameters, model takes each parameter and calculates accuracy for it and selects that hyper-parameter which gives the best result. To generate the models ScKit-Learn will be used.

Models will be evaluated based on their accuracy using Accuracy, Precision, Recall, F1 score, Roc score

we are working on medical data and predicting the cardiovascular risk before it happens so that it can be treated and prevented to happen so we must predict a higher correct number of positive case

A false positive can lead to unnecessary treatment and a false negative can lead to a false diagnosis, which is very serious since the disease has been ignored. So in our case recall measurement parameter is very important than other parameters.

So, the model which gives us the best recall score will be our best predictor model

6. Different Regression Model

4.1 Logistic Regression

Logistic Regression is a supervised learning that computes the probabilities for classification problems with two outcomes. It can also be extended to predict several classes. In Logistic Regression model, we apply the sigmoid function, which is

$S(x)= 1/1+e^{-x}$

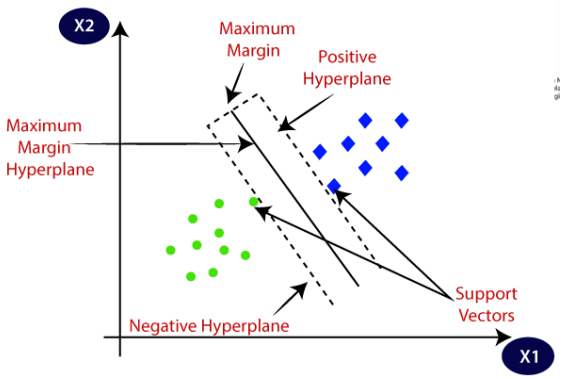
This function successfully maps any number into the value between 0 and 1 and we can regard this value as the probability of predicting classes. For example, we have two classes and they are presence of heart disease and absence of disease. If we set the threshold as 0.5, applying the sigmoid function gives us a value of 0.7, which means the man has the 70% probability of having heart disease so we will predict that he has heart disease.

4.2 K-Nearest Neighbors

K-nearest neighbors (kNN) is a supervised machine learning algorithm that can be used to solve both classification and regression tasks. K-Nearest Neighbors is one of the simplest supervised learning algorithms. KNN 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.
kNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using kNN algorithm.
kNN algorithm can be used for regression as well as for classification problems. kNN 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.
kNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data

4.3 Support Vector classifier

Support Vector classifiers (SVc in short) are machine learning algorithms that are used for classification and regression purposes.
The goal of the SVC algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
SVC chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed a Support Vector Machine/classifier. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



4.4 Decision Tree

Decision trees are statistical models that measure a target value using a collection of binary rules. To make a decision between 2 nodes, decision trees use **Attribute selection measure techniques** to decide to split a node into more sub-nodes.

There are two popular techniques for ASM

- Information Gain
- Gini Index

Both of the above techniques measure the entropy which is nothing but impurity/randomness in a given attribute and based on the result it chooses a split node and builds the decision tree.

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)
Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)
Gini=1−∑ni=1(pi)2

The advantage of this algorithm is that less pre-processing is required as data scaling or data normalization is not necessary. However, this model can be computationally expensive if the data has a lot of features. Decision trees can completely fit the training data but tend to overfit the test data. An alternative to deal with overfitting in decision trees is to use random fforestregressors.

4.5 Random forests

Random forest works by training a large number of decision trees and then calculating the mean prediction of the individual trees. The notion of random implies randomly created decision trees. Random decision trees are created on different subsets of the features and data points. For accurate predictions, random forest regressors can be optimized by hyperparameter tuning to ensure that the model does not depend too heavily on any single feature and that all potentially predictive features are considered equally. Also due to the previously mentioned random creation of decision trees, adding randomness prevents overfitting.

Random forest regression provides is feature importance estimate. Using feature importance, the effort can aid in a deeper understanding of the solved problem and, in some cases, contribute to model improvements.

4.6 Gradient Boosting and XG Boosting

Bosting both are The ensembles is a method used in the machine learning algorithm in this method, multiple models or ‘weak learners’ are trained to rectify the same problem and integrated to gain desired results. Weak models combined rightly give accurate models.

Gradient Boosting Machine (GBM) builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Extreme Gradient Boosting (XG Boost) is just an extension of gradient boosting with the added advantages like Regularization, Parallel Processing, High Flexibility, Handling Missing Values, Tree Pruning, Built-in Cross-Validation, Continuing on Existing Model

4.7 Stacking

Stacking or Stacked Generalization is an ensemble technique.
It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms.
The benefit of stacking is that it can harness the capabilities of a range of well-performing models on a classification or regression task and make predictions that have better performance than any single model in the ensemble.
Unlike boosting, in stacking, a single model is used to learn how to best combine the predictions from the contributing models (e.g. instead of a sequence of models that correct the predictions of prior models). The architecture of a stacking model involves two or more base models, often referred to as level-0 models, and a meta-model that combines the predictions of the base models referred to as a level-1 model.

Level-0 Models (Base-Models): Models fit on the training data and whose predictions are compiled.
Level-1 Model (Meta-Model): Model that learns how to best combine the predictions of the base models.

The meta-model is trained on the predictions made by base models on out-of-sample data. That is, data not used to train the base models is fed to the base models, predictions are made, and these predictions, along with the expected outputs, provide the input and output pairs of the training dataset used to fit the meta-model.
The outputs from the base models used as input to the meta-model may be real values in the case of regression, and probability values, probability like values, or class labels in the case of classification.

7. Model Expalanability

ELI5 was coded to check the model explainability. ELI5 is a python package used to understand and explain the prediction of classifiers such as sklearn regressors and classifiers, XGBoost, CatBoost, LightGBM, and Keras. It offers visualizations and debugging to these processes of these algorithms through its unified API. ELI5 understands text processing and can highlight text data. It can also implement techniques such as LIME and permutation importance.

We have observed, for our selected data point (2nd row) Gender has the most significant effect on the final prediction.

8. Score Board

Model	Accuracy	Precision	F1_Score	Roc_Score	Recall
XG-Boost	0.912	0.918	0.909	0.912	0.901
KNN	0.864	0.835	0.866	0.865	0.899
G-Boost	0.905	0.908	0.902	0.905	0.896
Stacking	0.908	0.933	0.903	0.908	0.875
Random Forest	0.874	0.886	0.868	0.873	0.851
SVC	0.818	0.832	0.809	0.817	0.787
D-tree	0.738	0.755	0.72	0.737	0.688
Logistic	0.794	0.872	0.763	0.791	0.678

9. Conclusion

- Machine Learning Algorithm K-Nearest Neighbour, Gradient boosting and XG Boosting performed the best with a recall of 1.0 on the train data set
- KNN and Gradient boosting performed the best on test data with a recall score of 0.89 and 0.90 respectively
- Logistic Regression is the least accurate as compared to other models performed.
- Model Explanibililty was performed on extrema gradient boosting using eli5 and it's observed that sex(gender) has the most significant impact on the target variable whether the patient will have the risk of CHD or not.
- Model can be improved with more computational resources and with more data.

10. Library and packages are used for analysis

- **libraries for process data**

```
import pandas as pd
import numpy as np
from numpy import math
```

- **libraries for plotting data**

```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

- **To prepare our train and test dataset**

```
from sklearn.model_selection import train_test_split
```

- **For Standardise our dataset**

```
from sklearn.preprocessing import StandardScaler
```

- **For cross validation and hyperperameter tuning**

```
from sklearn.model_selection import GridSearchCV
```

- **For overcome class imballence**

```
from imblearn.over_sampling import SMOTE
```

- **libraries for regression**

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import StackingClassifier
```

- **libraries for measuring performance metrics**

```
from sklearn.metrics import accuracy_score, recall_score, precision_score, f1_score, roc_auc_score, conf
usion_matrix, roc_curve, auc
```