Cardiovascular Risk Prediction

Naved mansuri

Data science trainees,

AlmaBetter

Abstract

The WHO estimates that 17.9 million deaths worldwide from heart disease occurred in 2016, accounting for 31% of all fatalities. More than 75% of these fatalities occurred in developing and middle-income nations.

Coronary heart disease, sometimes known as a heart attack, is by far the most prevalent and lethal of all heart conditions. For instance, it is believed that someone in the United States experiences a heart attack every 40 seconds, and that there are roughly 805,000 heart attacks there each year (CDC 2019).

Introduction

Several health conditions, your lifestyle, and your age and family history can increase your risk for heart disease. These are called risk factors. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Some risk factors for heart disease cannot be controlled, such as your age or family history. But you can take steps to lower your risk by changing the factors you can control.

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 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors.

Data Description

Attribute	Description
Sex	male or female("M" or "F")
Age	Age of the patients
Hour	Hour of the day
Education	The level of education of the patient (categorical values - 1,2,3,4)
is_smoking	whether or not the patient is a current smoker ("YES" or "NO")
Cig per day	the number of cigarettes that the person smoked on average in one day.
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
Glucose	glucose level (Continuous) Predict variable (desired target)
TenyerCHD	heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

Exploratory Data Analysis

We were able to understand how various variables in our dataset affect the target variable because to the exploratory data analysis we conducted on our train dataset. To determine if a certain category characteristic was reliant on our binary target variable, we used a chi-square test. In order to determine if the distribution of our continuous variables for the various groups is similar, we also performed a one-way ANOVA test. This assisted us in determining whether or not certain features should be included in our final model.

Checking missing values and treat

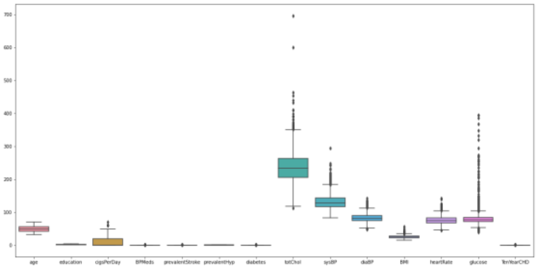
The **glucose** column contains 304 missing values.

Education feature is not a continues variable so we using Mode for filling the missing values.
heartRate only 1 and **BMI** 14 missing values, so we simply drop that nan values.

Checking outliers and treat

Outliers are noticed in:

- totChol
- sysBP
- diaBP
- BMI
- heartRate

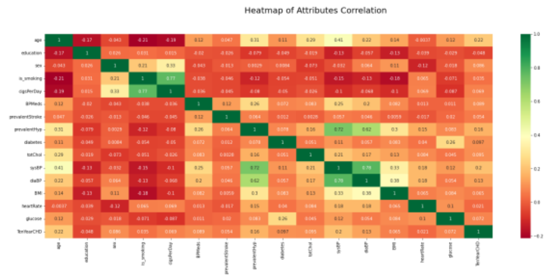


- glucose
- cigsPerDay

totChol, sysBP, diaBP, BMI, heartRate, glucose and cigsPerDay having outliers, so handling those using The interquartile range (IQR).

Label Encoding

In machine learning, we usually deal with datasets that contain multiple labels in one or more than one columns. These labels can be in the form of words or numbers. To make the data understandable or in humanreadable form, the training data is often labelled in words.



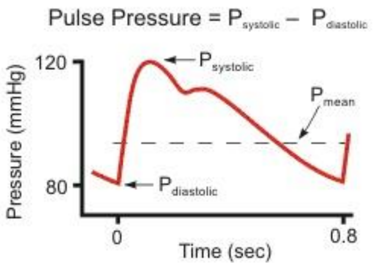
Correlation Analysis

Highest correlation between systolic BP and diastolic BP.
Systolic and Diastolic BP shows a high correlation with hypertension.
Variables such as age, prevalent hypertension, systolic BP, diastolic BP,.influence the risk of heart disease mainly.
cigarette smoking and the number of cigarretes smoked per day.
Systolic BP and age have a positive correleation.
So we would required to select our features for our models performance.

Feature Engineering

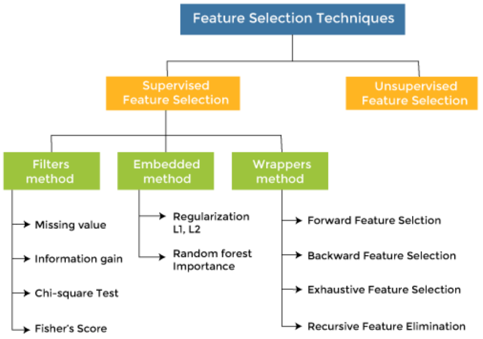
There is a high correlation between **sysBP(Systolic BP)** and **diaBP(Diastolic BP)**, and both of them influence our target variable to a greater extent, so we cannot drop them directly, but rather must find a parameter that can formulate these parameters together in such a way that we can add a single feature without experiencing multicollinearity or **pulse pressure**.

Pulse Pressure = Systolic BP - Diastolic BP



Feature Selection Models:

Filter Method:In this method, features are dropped based on their relation to the output, or how they are correlating to the output. We use correlation to check if the features are positively or negatively correlated to the output labels and drop features accordingly. they models are,



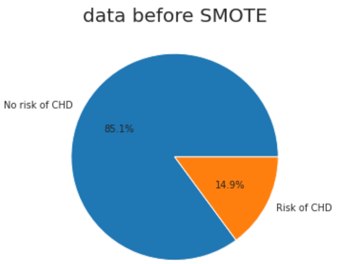
- Information Gain,
- Chi-Square Test,
- Fisher’s Score, etc.

In this model we using **Chi-Square** test for selecting the features that influence the most.

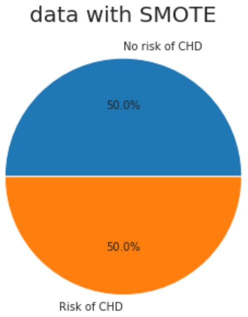
Handling Imbalanced Data

We oversample the train dataset using SMOTE (Synthetic Minority Oversampling

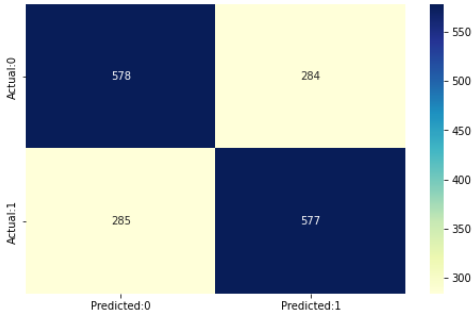
Technique) to ensure that the model is trained equally on all results, and is not biased to one particular result.

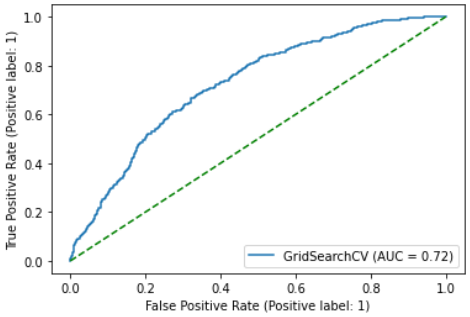


Since our dataset is imbalanced, with more negative cases than positive cases, we may end up with a classifier that is biased towards the negative cases. The classifier may have high accuracy, but poor precision and recall.



We have successfully oversampled the minority class using SMOTE. Now, the model we build will be able to learn from both classes without any bias.





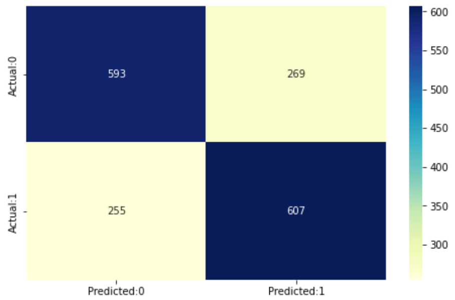
Machine Learning Algorithms

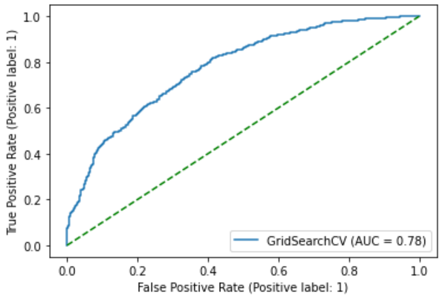
Logistic Regression

Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

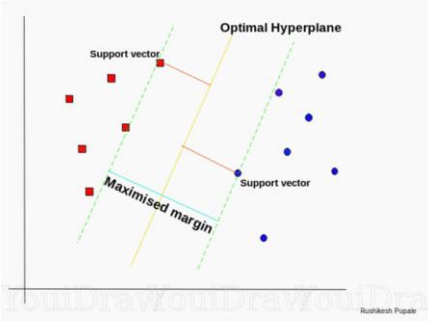
Performance of Logistics regressions
Accuracy : 0.67
Precision : 0.6694
Recall : 0.6702
F1 Score : 0.6698

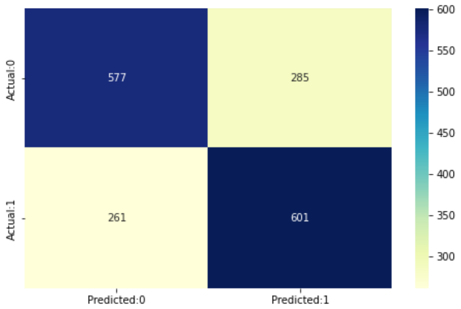
Random Forest





Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

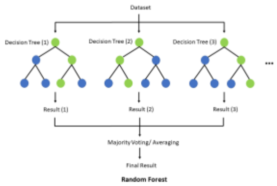


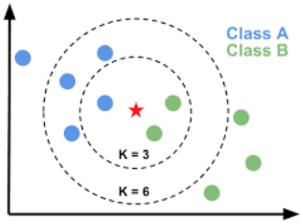


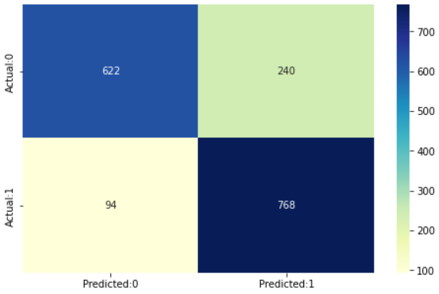
Performance of Random forest classifiers
Accuracy : 0.6961
Precision : 0.7042
Recall : 0.6929
F1 Score : 0.6985

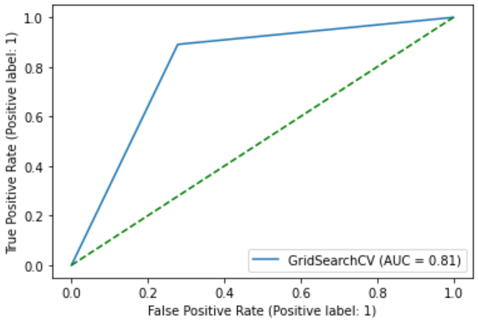
SVM

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.







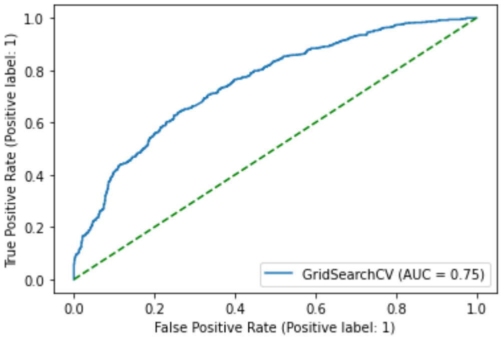


Performance of Support Vector Machine Classifier
Accuracy : 0.6833
Precision : 0.6972
Recall : 0.6783
F1 Score : 0.6876

KNN

The k-nearest neighbors algorithm, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

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



Performance of KNN Classifier
Accuracy : 0.8063
Precision : 0.891
Recall : 0.7619
F1 Score : 0.8214

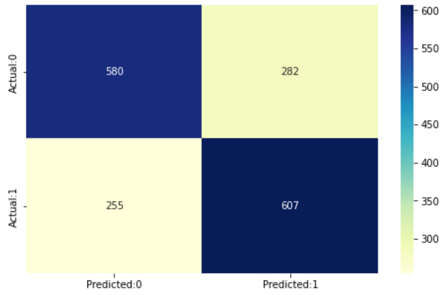
XGBoost

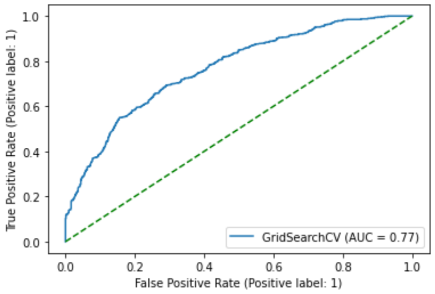
XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values. The most common loss functions in XGBoost for regression problems is linear, and that for binary classification is logistics.

Performance of XGBoost Classifier
Accuracy : 0.6885
Precision : 0.7042
Recall : 0.6828
F1 Score : 0.6933

Conclusion

Cardiovascular heart disease affects a similar number of smokers and non-smokers.





Age, total cholesterol, systolic blood and diastolic blood pressure, BMI, heart rate, and glucose are the main factors in determining a person's 10-year chance of having cardiovasular heart disease.

The K Nearest Neighbour is proved to be best algorithms can be used for the risk prediction of Cardiovasular heart disease.

We chose the oversampling technique because the data provided to us had fewer records. But since there will be a lot of unbalanced and large amounts of health data, we can try to work on cost-sensitive learning, which, rather than changing the data records, only gives more weight to the minority and focuses on the individuals at high risk for heart disease.