Cardiovascular Risk Prediction

Ameen Attar, Hrithik Chourasia,
Pradip Solanki, Vridhi Parmar
Data science trainees,
AlmaBetter, Bangalore

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Team Consists of:

1. Ameen Attar
   - Contribution: EDA
   - E-mail: ameenattar92@gmail.com
2. Hrithik Chourasia
   - Contribution: EDA
   - E-mail: hrithik8wel@gmail.com
3. Pradip Solanki
   - Contribution: Baseline model building
   - E-mail: solankipms@gmail.com
4. Vridhi Parmar
   - Contribution: Model building
   - E-mail: vridhiparmar27@gmail.com

Abstract:

In today's world data science plays an important role in Medicine and Healthcare. Traditionally, medicine solely relied on the discretion advised by the doctors. For example, a doctor would have to suggest suitable treatments based on a patient's symptoms.

However, this wasn't always correct and was prone to human errors.

There are several fields in healthcare such as medical imaging, drug discovery, genetics, predictive diagnosis and several others that make use of data science.

Here we are provided with some dataset from an on-going 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

Problem Statement:

The dataset provided is from 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

The description of the data is given below:
Demographic:

- Sex: male or female("M" or "F")
- Age: Age of the patient

Behavioral

- 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

Medical(History)

- BP Meds: whether or not the patient was on blood pressure medication
- Prevalent Stroke: whether or not the patient had previously had a stroke
- Prevalent Hyp: whether or not the patient was hypertensive
- Diabetes: whether or not the patient had diabetes (Nominal) Medical(current)
- Tot Chol: total cholesterol level
- Sys BP: systolic blood pressure
- Dia BP: diastolic blood pressure
- BMI: Body Mass Index
- Heart Rate: heart rate
- Glucose: glucose level
- 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”)

Steps involved :

Exploratory Data Analysis :-

It is a way of visualizing, summarizing and interpreting the information that is hidden in rows and column format. EDA is one of the crucial step in data science that allows us to achieve certain insights and statistical measure. It performs to define and refine our important features variable selection, which will be used in our model.

Null values Treatment and encoding categorical values:-

Our data comprises of huge null values which will disturb our data during modeling and hence we fill the null values with the mode of the data and changing the numerical values like ‘Yes’ and ‘No’ with numerical values like ‘1’ and ‘0’.

Fitting different models

For modeling we tried various classification algorithms like:

1. Naive Bayes Classifier:- 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:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2. KNN:- The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics— calculating the distance between points on a graph.
3. 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:

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

4. Decision Tree:- Decision Trees is where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.
5. Random Forest:- 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.
6. XGBoost:- It 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. 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.

Model Performance:-

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.

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 and false positive rate to build up a summary picture of the classification performance.

Conclusion:-

That's it! We reached here at 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 66 to 91%.
And there is no such improvement in accuracy score even after hyper parameter tuning.
So the accuracy of our best model is 91% 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 but maybe with enough data we can train out model even better.

Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.
Google Drive link: https://drive.google.com/drive/folders/1TxwSQdr_IwV6J_ikWmeVh2_863eJ9BhC?usp=sharing GitHub link: https://github.com/ameenattar92/Cardiovascular-Risk-Prediction