Cardiovascular Study and Heart Disease Prediction

Nirav Desai SID: 490540262

Subject Code	COMP5310, Semester 2, 2019
Subject	Principles of Data Science
Purpose	Assignment 2 report
Name	Nirav Desai
Student ID	490540262
Uni Key	ndes8735

Research Problem

As per the World Health Organization report, an estimated 18 million people died from Cardiovascular Disease in 2016. This figure represents 31% of all global deaths. Also, the WHO studies state that cardiovascular diseases can be detected based on other risk factors such as hypertension, diabetes, irregular blood pressure etc. One of the leading ongoing research in this area is conducted by "Framingham Heart Study" since 1948.

Framingham dataset contains number of features such as Diabetes, BP Medications, Glucose levels, and age etc. It also has a Ten Year Cardiovascular Heart Disease score (CHD) for selected patients. Our study focuses on a question:

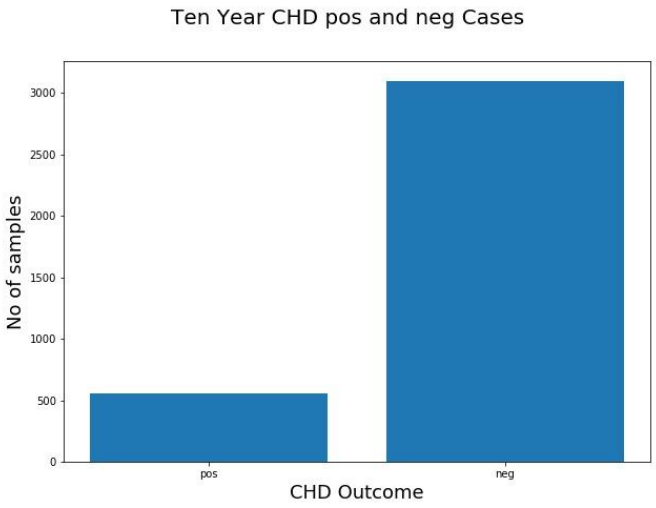
1. Can we successfully predict a patients risk for contracting cardio vascular disease based on their vitals such as Diabetes condition, BP Medications, Glucose levels, and age etc?
2. Is age a major factor to the risk of developing heart disease?

Null Hypothesis:

1. The given set of vitals (features) are not sufficient in predicting heart disease risk.
2. A risk of developing heart disease does not increase with a person's age.

Evaluation

As per the data exploration in phase 1, there has been an imbalance in our dataset.



The number of positive cases of risk score are far lower than negative cases.

Due to class imbalance, we cannot rely on a generic accuracy reading to measure model effectiveness. The negative cases are 85% of samples, and thus a simply assigning negative value to all test sample can give us 85% accuracy. To avoid this fallacy, we will use f1-score, precision and recall measurement to evaluate our model.

Why F1-score?

A successful model should predict a patient's CHD risk score accurately. This means that it should be able to identify both positive and negative cases in effective manner. A wrongly tagging either of them have serious consequence on patient's physical health.

Instead of just any particular case, we would like to identify:

i) True Positive – A patient with no CHD risk tagged with score 0.
ii) True Negative – A patient with CHD risk tagged with score 1.
iii) False Positive/False negative – A patient misclassified with CHD score.

F1-Score, with precision and recall measurement, represents our desired outcome to find True Positives and True Negatives and thus we would use it with accuracy measurement.

Experiment and Approach

Predictor imbalance

Following 2 techniques are used to handle to imbalance in our dataset:

1. Under sampling: This method will make sure than we have equal samples of positive and Negative cases to train and test from actual dataset.
2. Over sampling: This method creates synthetic positive cases to equalize training samples for modeling. We will use SMOTE python library in our modeling to model using over sampling. It defers from under sampling in two ways:
   i) The actual training samples are much more in case of synthetic sampling.
   ii) The synthetic sampling "generates" training data from existing samples. It doesn't represent actual samples and may not give desired results.

Feature Selection

The dataset contains 15 features. As per the correlation heatmap in phase 1, features such as cigarettes per day are not relevant to our research. A simple feature selection was performed using ANOVA F-value measurements. The results are:

1. In case of under sampling, following 8 feature are highly relevant:
   'age', 'sysBP', 'diaBP', 'glucose','BMI','BPMeds','prevalentHyp','diabetes'
2. In case of over sampling, following 8 features are highly relevant:
   'age', 'sysBP', 'diaBP', 'glucose','BMI', 'totChol', 'sex_male','prevalentHyp'

Benchmarking

A benchmarking was performed using simple logistic regression model with following results:

Model	Runtime	F1-Score	Accuracy	Precision	Recall
Logistic Regression	100 ms	0.63	0.6285	0.63	0.63

Modelling Approach

The following models will be implemented and cross checked against the benchmark:

1. Classification using Random Forest
2. Decision Trees with Extreme Gradient Boosting (XGBoost)
3. Support Vector Machines

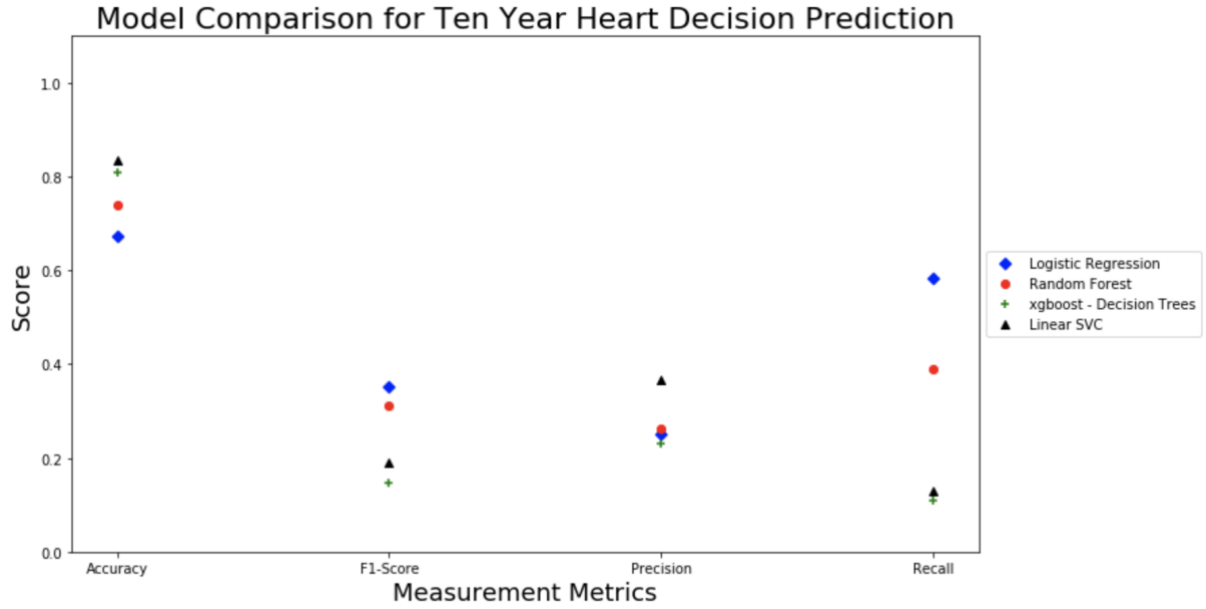
The following approach was taken during the study:

1. After feature selection, each feature was pre-processed either with one hot encoding or standardization technique based on type of data.
2. The selected models were hyper tuned using 10 fold cross-validation.
3. Each model was tested based on best hyper parameters and verified as per the evaluation measure described.
4. Both oversampled and under sampling tests were performed using train/test split of 75% and 25% respectively.

Results Analysis

Over Sampling Results

The following chart contains results for models using over sampling. The results are underperforming benchmark. Also, while accuracy is high, the F1, precision and recall scores clearly point to unstable modeling.



Under Sampling Results

On the other hand, under sampling gives reliable results with all models. The XGBoost based decision tree performed better than our benchmark logistic regression model. The F1-score on XGBoost was best (66%). The detailed results are described in below tables for better comparison.

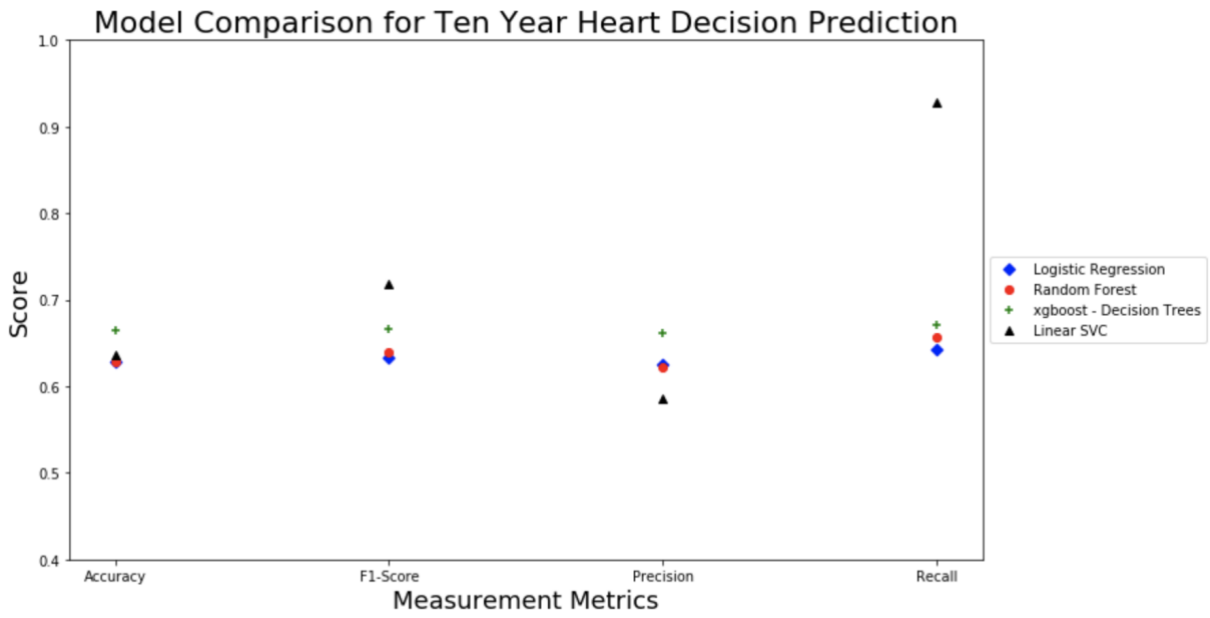
The confusion matrix:

Out of 140 Positives (CHD score 1) and 140 Negatives (CHD score 0)

Classifier	True Positive	True Negative	False Positive	False Negative
Logistic Regression	86	90	54	50
Random Forest	84	92	56	48
XGBoost Decision Trees	92	94	48	46
Support Vector Machines	48	130	92	10

The evaluation matrix:

Classifier	Runtime	F1-Score	Accuracy	Precision	Recall
Logistic Regression	2 secs	0.63	0.6285	0.63	0.63
Random Forest	163 secs	0.63	0.63	0.63	0.63
XGBoost Decision Trees	12 Secs	0.66	0.66	0.66	0.66
Support Vector Machines	5 Secs	0.60	0.61	0.71	0.64



Conclusion

From the results above, we can conclude our original null hypothesis can be rejected. Our model can clearly predict a patient's CHD risk score based on given set of vitals with 66% accuracy for both positive and negative cases.

We can also reject our null hypothesis related to patient's age. As per the ANOVA tests in feature selection, we confirm that age is highly related to cardiovascular heart disease risk.