Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

author: **Jinhui HU**

Table of contents

1 Introduction: Cardiovascular disease

Cardiovascular diseases are the leading cause of death in all areas of the world except Africa. Together CVD resulted in 17.9 million deaths (32.1%) in 2015, up from 12.3 million (25.8%) in 1990.[1]

In this project we will try to predict whether a person have cardiovascular disease or not. Specifically, this report will be targeted to doctors interested in using machine learning to predict a patient have ardiovascular disease or not based on some information.

We will use two machine learning algorithms including logistic model and decision tree to build models and use them to predict.

2 Data

Most data can be found in this Kaggle datasets here ('https://www.kaggle.com/sulianova/cardiovascular-disease-dataset').[2] These data include 3 different types:

- Objective: factual information;
- Examination: results of medical examination;
- Subjective: information given by the patient.

And there are some descriptions of the features:

Features	Type	Names	Type of data
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective Feature	smoke	binary
Alcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular disease	Target Variable	cardio	binary

All of the dataset values were collected at the moment of medical examination.

There are 68692 records in our data set.

3 Methodology

There are two machine learning algorithms I will use in this report: logistic model and decision tree. The logistic model is a regression model and we can use it to accomplish our goal. The other one is classfication model which is different from regression model but still can handle our task.

3.1 Logistic Model

The mean squared error is 0.272176, and the residual sum of squares is 0.272176. There are the coefficients of the logistic model:

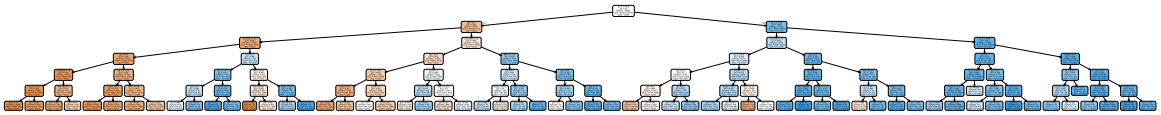
age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke
0.04979299	-0.01919647	-0.03781909	0.04606342	0.05269945	0.01782537	0.48804562	-0.11105226	-0.1305706

3.2 Dicision Tree

We use the grid search to bulid the tree. Here is the visualization of the tree:

In [35]:

Out[35]:



4 Results and Discussion

We bulit two models to predict and the logistic model seems better. And we can draw some conclusion from these two models. There are the coefficients of the logistic model:

age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke
0.04979299	-0.01919647	-0.03781909	0.04606342	0.05269945	0.01782537	0.48804562	-0.11105226	-0.1305706

From the coefficients of the logistic model we can find out that there are several features which are positively correlated with CVD: age, weight, systolic blood pressure, diastolic blood pressure and cholesterol. These information tell us that we should pay attention to our health. If you have hypertension and you are no longer young, and maybe found yourself with high cholesterol from your medical report, you should really care about your heart and pay attention to CVD.

And similarly, we can also find some interesting conclusion from the decision tree model. Young people with high cholesterol seem have greater chance to have CVD, which may be the new research topic.

5 Conclusion

We bulid two models to predict if a person have CVD, including logistic model and decision tree model. And doctors our yourself can use this model to predict whether you have CVD or not. **BUT THE RESULTS CANNOT REPLACE THE DOCTOR'S ADVICE!** These models only tell you may have more chance to have CVD and if you are predicted as no CVD, there is also a chance that you have CVD.

And we also draw some interestiong conclusions. If you have hypertension and you are no longer young, and maybe found yourself with high cholesterol from your medical report, you should really care about your heart and pay attention to CVD. And young people with high cholesterol seem have greater chance to have CVD. We may find more information from these models.

6 Reference

[1] https://en.wikipedia.org/wiki/Cardiovascular_disease (https://en.wikipedia.org/wiki/Cardiovascular_disease)

[2] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset)