Coronary Artery Disease Risk Prediction

A Logistic Regression Approach
(a project in progress)

Janco van Niekerk

14-Apr-18

Contents

List of Figures

1. Introduction

The goal of following project was to determine the risk of developing coronary artery disease.Python 2.7 was used in conjunction with Pandas and Scikit-learn to analyze the available data.

This document contains information from NHANES 2011-2012 and NHANES 2013-2014, which is made available here:

- https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013
- https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011

This project is currently still a work in progress – plans include the evaluation of the predictive model with cross validation, the incorporation of more independent variables which may influence the risk of coronary artery disease as well as more comprehensive data exploration and visualization.

2. Background

Coronary heart disease is a condition where blood flow to the heart is reduced due to arterial plaque buildup. The condition typically occurs in older individuals.

Heart attacks and angina (chest pain) is a common effect of coronary artery disease. The Rose Angina Questionnaire is commonly used to diagnose Angina.

High levels of Low-density lipoprotein cholesterol (LDL-C), commonly referred to as 'bad' cholesterol, is associated with increased risk of developing coronary artery disease.

Due to the increased risk of high levels of LDL-C, it is common for high risk individuals to be prescribed cholesterol lowering drugs (statins).

3. Problem Statement

It was required to determine the risk that an individual has coronary artery disease. Due to the common consensus that increasing age and LDL-C levels increase the risk it was decided that the risk will be calculated with respect to these two factors.

The data-set included the following relevant information:

- Whether a participant is taking Statins (LDL-C lowering medication).
- A participant's age.
- Whether a participant has previously been diagnosed with coronary artery disease.
- Whether a participant has had a heart attack in the past.
- Answers given by the participant to the Rose Angina Questionnaire.

Due to the fact that we would like to establish the relationship between LDL-C and coronary artery disease we would have to exclude all the participants on LDL-C lowering drugs –otherwise a participant could have low levels of LDL-C while showing symptoms of coronary artery disease. This may lead one to incorrect conclusions about the relationship between LDL-C and coronary artery disease.

Furthermore, three possible effects of coronary artery disease are:

- Heart attacks.
- Testing positive for Angina according to the Rose Questionnaire.
- Being diagnosed with coronary artery disease.

So, at first glance, we could take the any of the three events described above as a positive indicator of coronary heart disease. However, it is known that not all heart attacks are due to coronary artery disease; therefore individuals who have suffered a heart attack which was not due to coronary artery disease may not necessarily be prescribed statin drugs.

Testing positive for Angina as well as being diagnosed with coronary artery disease may not necessarily mean a participant has coronary artery disease – but let's assume, for this investigation at least, that they do imply that a participant has coronary artery disease. We will thus use a positive test for Angina according to the Rose Questionnaire and coronary artery disease diagnoses as conditions which are equivalent to having coronary artery disease. As mentioned earlier, not all heart attacks are due to coronary heart disease; however a heart attack will likely cause some chest pain for a few individuals and will thus influence the answers on the Rose Questionnaire. Due to this fact we will exclude from the investigation any participants who have suffered a heart attack.

4. Data Pre-processing

It was decided that two datasets would be used:

- NHANES 2011-2012
- NHANES 2013-2014

These two datasets were both individually cleaned in the following manner:

- All entries of participants who use cholesterol lowering medication were removed.
- All entries of participants who had suffered heart attacks were removed.

An additional column called 'Angina' was then added and the data was then screened to produce either a 'Yes' or 'No' value for this column according to criteria specified in the Rose Questionnaire.

The code for the pre-processing of both datasets is included in 'Data Project\\Wrangling Code\\' as:

- Code 2011 to 2012
- Code 2013 to 2014

5. Data Analysis

The two datasets mentioned in the previous section were merged for further analysis.

The data was then binned according to LDL-C levels as well as age. The data was then plotted on a 3-dimensional axis with the code provided in 'Data Project\\' as 'ChartPlot'. The result is indicated in Figure 1.

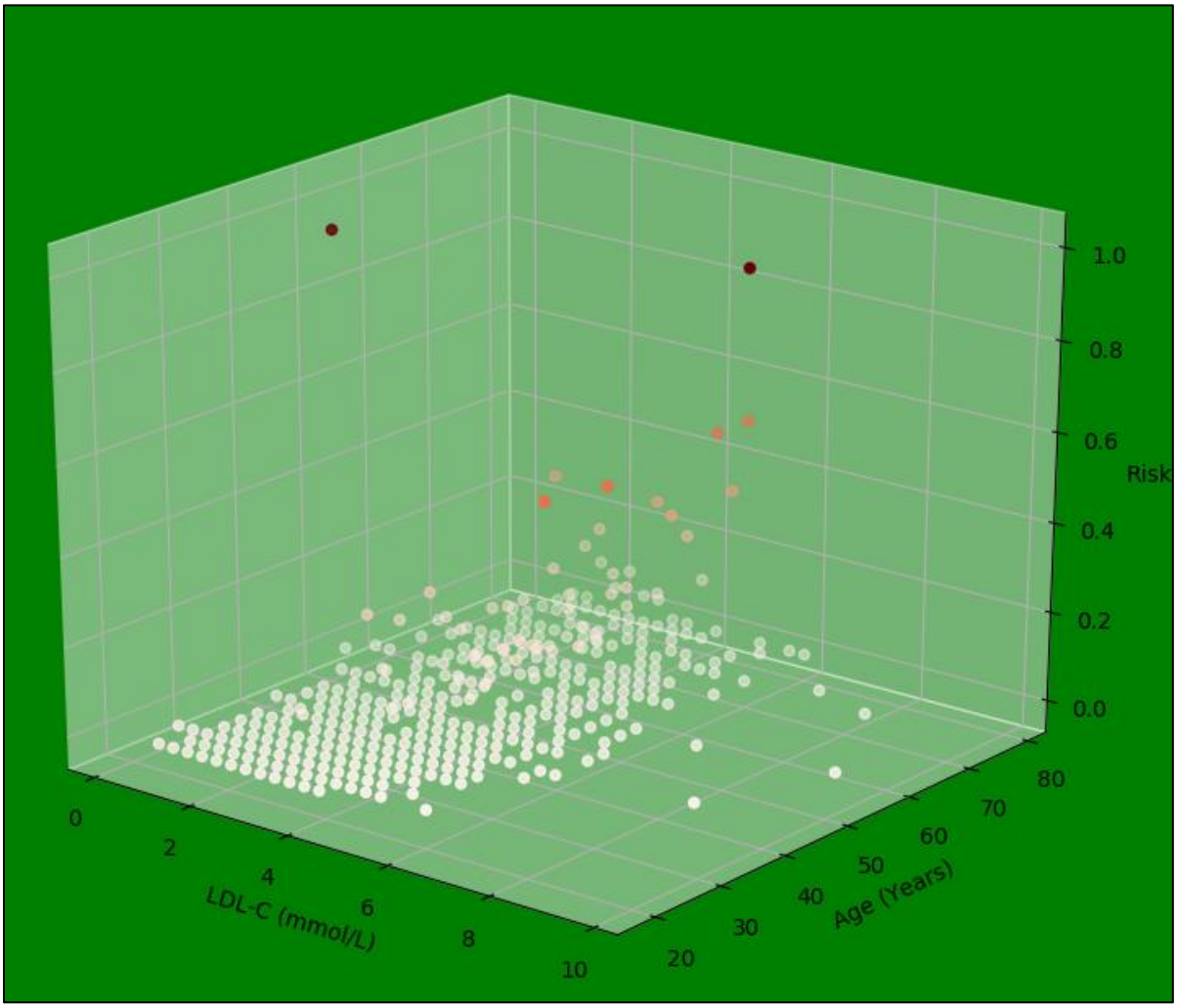

Figure 1 - Risk vs. LDL-C and Age

6. Predictive Modeling using Logistic Regression

Due to the fact that we are essentially faced with a classification problem-i.e. does a given participant have coronary artery disease or not- it was decided that a Logistic Regression model will be used to estimate the data.

After the model was trained it was plotted against the data points from the previous section as per the code included in 'Data Project\\' under the name 'Logistic Regression and comparison'.

The results of the model are shown in purple (decreasing in darkness as the risk **increases**) and the actual (binned) data points are displayed in red (decreasing in darkness as the risk **decreases**) on Figure 2. Note that the entries with risk values greater than 0.3 were removed from the actual data- this was done to accentuate the colour change for increased risk and was thus done for display purposes only, the predictive model was still trained on all available data.

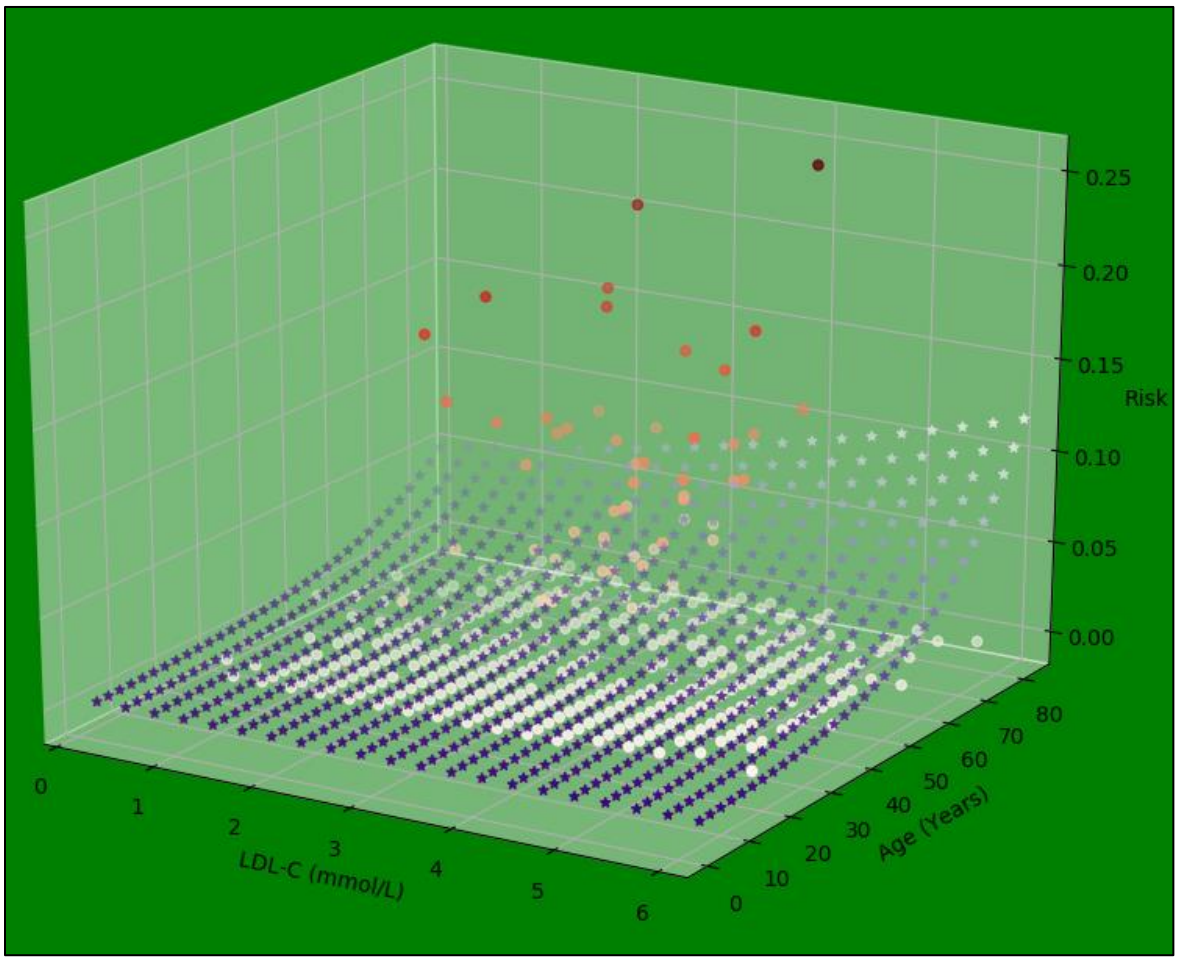

Figure 2 - Logistic Regression model vs. Actual Data

7. Evaluation of Results

The predictive model clearly shows that as age and LDL-C levels increase the risk of having coronary artery disease also increases.

From the actual data points it can be seen that there does exist participants of old age with high LDL-C levels who show no symptoms of coronary artery disease. This may be all due to misdiagnoses but a far more plausible answer is that the process of developing coronary artery disease depends on more variables than just age and LDL-C alone.