Coronary Artery Disease Risk Prediction

A Logistic Regression Approach

Janco van Niekerk

01-Feb-20

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 of coronary heart disease (CHD) it was decided that these two factors will be used as predictors for a CHD predicting model.

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 due to some patients having developed coronary artery disease before being treated with statins.

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 participant who has 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 and a risk was calculated by counting the number of angina sufferers and dividing this number by the amount of participants in each bin.

The data was then plotted on a 3-dimensional axis with the code provided in 'Data Project\\' as 'ChartPlot'. This was done to get an intuition about how the model should look like - the result is indicated in Figure 1.

From an eyes-glance of the binned data in Figure 1 it became apparent that increasing age was definitely associated with a higher risk.

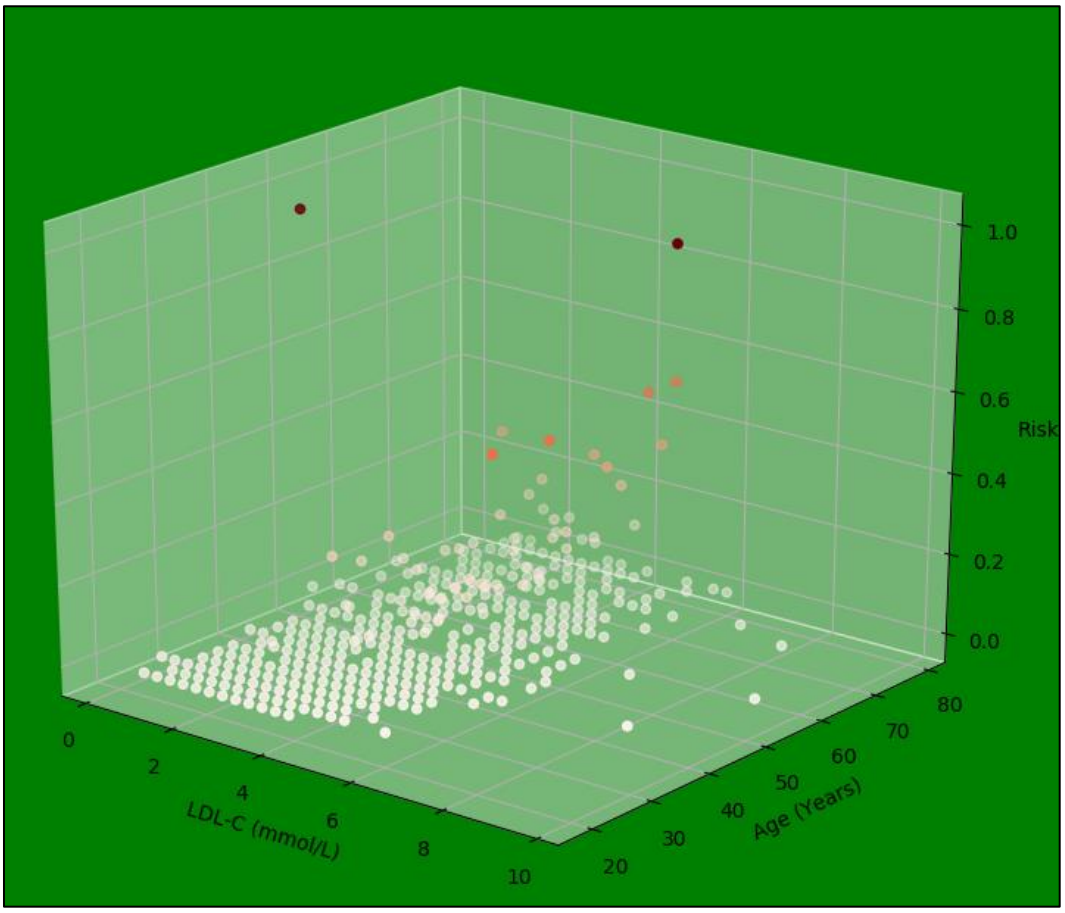

Figure 1 - Risk vs. LDL-C and Age

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 binned data points with risk-values greater than 0.3 were removed from the figure - 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 training data.

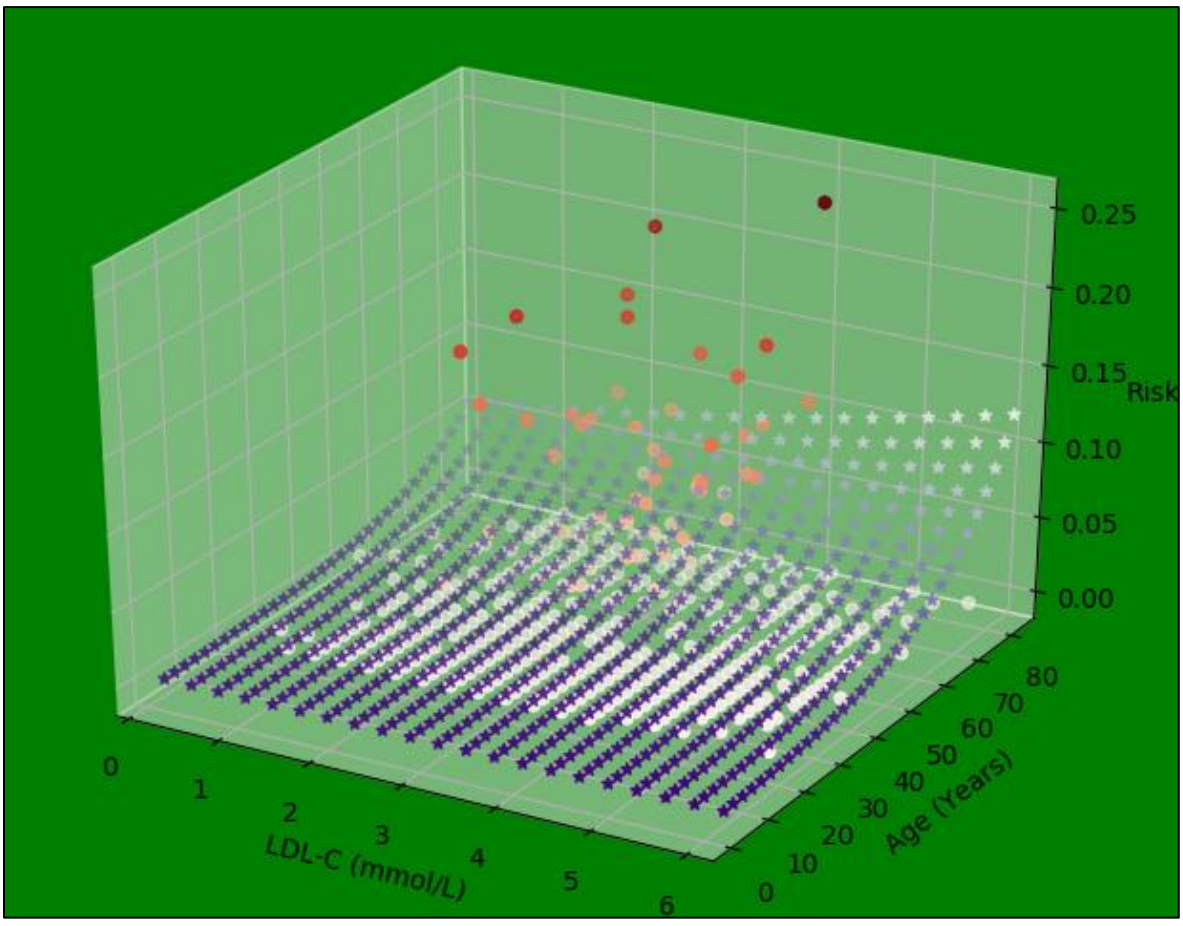

Figure 2 - Logistic Regression model vs. Actual Data

7. Evaluation of Results

The model was evaluated according to a certain set of criteria which will be discussed in the following section. The final output is displayed in Figure 3.

```
ROC AUC Score = 0.7509768314952918
Accuracy at a threshold of 0.5 =98.02361396303901 %
positive cases of angina = 77
total samples = 3896
fraction of negative cases of angina to total samples = 98.023613963 %
Accuracy at threshold of 0.035 = 83.34188911704312 %
Recall at threshold of 0.035 = 42.857142857142854 %
Precision at threshold of 0.035 = 5.172413793103448 %
Accuracy on test data at threshold of 0.035 = 82.42612752721618 %
```

Figure 3 - Evaluation Output

The first step in evaluating the model entailed calculating the ROC AUC score, for a perfect classifier this score will be equal to 1.0 and a random classifier will have a score of 0.5.

The ROC AUC score was calculated and a value of 0.75 was obtained as can be seen from Figure 3. This score was consistent with expectations due to it being larger than the score of a random classifier and smaller than that of a perfect classifier.

The next step taken was to calculate the accuracy of the model. The initial threshold for the model was taken as 0.5.

It was seen that, at this threshold, the model had and accuracy of 98.02 %. This accuracy clearly seems too high and therefore some further investigation was done.

The fraction of negative cases of angina within the total sample was calculated and a value of 98.02 % was obtained.

This means that if a model predicted every single case in the sample as negative then it would have an accuracy of 98.02 %. Such a model, albeit having a high accuracy, would provide a naïve perspective on the relationship between the target and predictor variables.

To classify fewer samples as negative the threshold of the classifier would have to be adjusted. The F-Score was chosen as the measure to be used to select the optimal threshold. The F-score for various thresholds are plotted in Figure 4 – from this plot it could be seen that the maximum F-score is obtained at a threshold of 0.035.

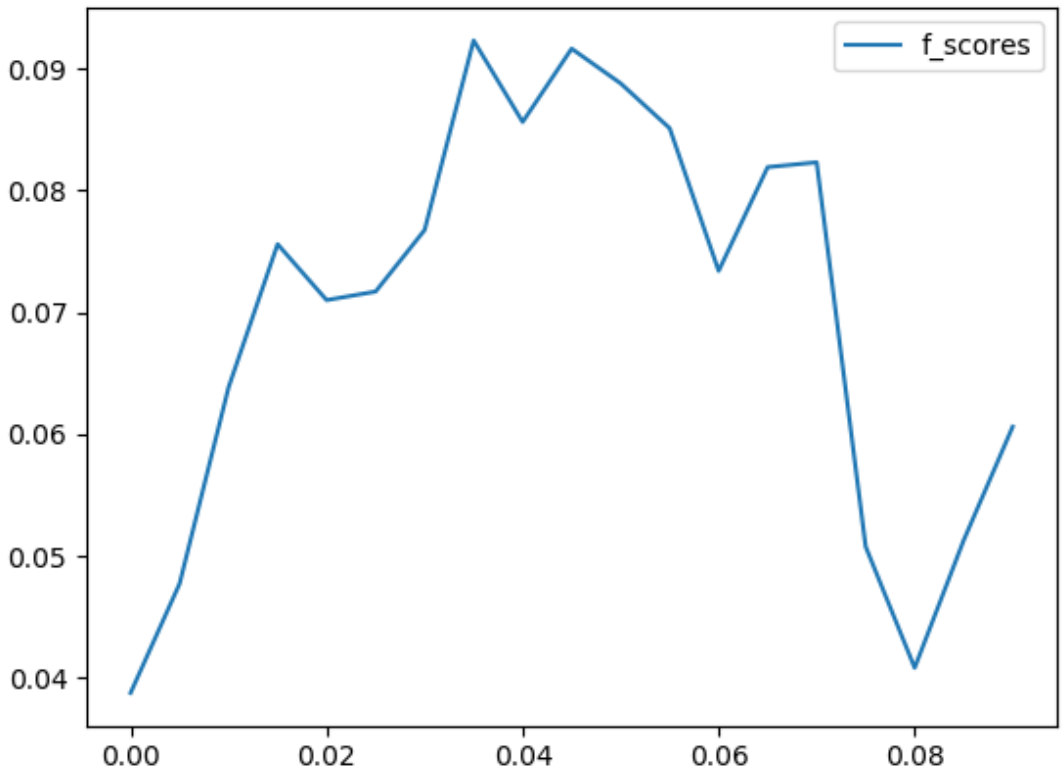

Figure 4 - F-Score vs. Threshold

The accuracy was then calculated at this new threshold and value of 83.34 % was obtained.

As can be seen from Figure 3, the precision and recall were also calculated at this new threshold. A recall of 42.857 % means that the model correctly identified 42% of the all positive cases (individuals with angina) as positive. The precision obtained means that only 5.17 % of the samples that were predicted as positive actually had a case of angina.

Finally, the accuracy at this new threshold was calculated on the test set and found to be 82.426 %.

8. Conclusion

The precision of the model (5.17 %) is quite poor. Which mean the model will label a lot of individuals as having angina when they in fact do not.

On the other hand, the model has a recall of 42.857 %. This means that the model could identify 42.847% of angina sufferers as such.

Essentially, the performance of a classification model should be judged based on the decision that this model will be used for.

For instance, let us propose that our only goal was to decrease mortality rates. Let us also propose that we knew that simply having angina increased mortality risk while some hypothetical drug had no effect on mortality risk alone but could be used to treat angina. We may then conclude that the optimal model will label every single person as having angina. This is due to the fact treating every single individual with the hypothetical drug will ultimately lead to achieving our goal – decreasing mortality rates as much as possible.

It should however be mentioned that such a model would not be an accurate depiction of reality. Such a model would also not be ideal for most use cases. A simple use case which illustrates this point is if we had a drug which increases mortality risk due to side effects – A model and decision strategy which simple prescribes everyone the proposed drug would thus lead to an increase risk in mortality for perfectly healthy individuals.

A model which is the most accurate depiction of reality and applicable on tons of use cases would have the highest possible accuracy, precision and recall.

Such a model would however incorporate all the causal predictors which could possibly be factors such as smoking habits, insulin levels, genetics etc. and would thus be far more complex than a model which simply uses age and LDL-C levels as predictor variables.