Heart Disease Prediction Using Logistic Regression

Dominik Szablonski

Abstract

The World Health Organization (WHO) estimates that in 2019, 17.9 million people died from cardiovascular diseases (CVD), representing 32% of all global deaths, while of 17 million premature deaths due to non-communicable diseases in 2019, 38% were caused by CVDs.[1] Further, a WHO report states that the most effective method of preventing, in particular, CHD is through addressing behavioural risk factors, such as diet and smoking.[2] My research aims to be able to predict the development of CHD within a subject through analysing various factors, including behavioural, medical, and demographic, in a logistic regression model.

I was able to streamline the through applying initial data analyses, before going onto applying logistic regression to the data in order to predict a subject's likelihood of developing CHD within the next 10 years. The model, however, was only able to reach 84.8% accuracy, so there is a need to explore changes which we may apply to the model or perhaps to explore a more complex model, perhaps something similar to the model which SANO is currently working on based on Daniel Kahneman's "Thinking, Fast and Slow".

Variables

Below are then listed the variables included within our dataset and the category within which it falls into.

Demographic

- **male:** The sex of the subject. These are binary values, 0 representing female and 1 representing male (Discrete, nominal, binary)
- **age:** The age of the subject, rounded to a whole number. (Continuous, interval)

Behavioural

- **currentSmoker:** If the subject is currently a smoker. (Discrete, nominal, binary)
- **cigsPerDay:** How many cigarettes does the subject consume per day. (Continuous, interval)

Medical (History)

- **BPMeds:** Whether the subject was on blood pressure medication. (Discrete, nominal, binary)
- **prevalentStroke:** Whether or not the subject has experienced stroke before. (Discrete, nominal, binary)
- **prevalentHyp:** Whether or not the subject was previously hypertensive. (Discrete, nominal, binary)

Medical (Current)

- **totChol:** The subject's total cholesterol level. (Continuous, interval)
- **sysBP:** The subject's systolic blood pressure. (Continuous, interval)
- **diaBP:** The subject's diastolic blood pressure. (Continuous, interval)
- **BMI:** The subject's Body Mass index. (Continuous, interval)
- **heartRate:** The subject's heart rate. (Continuous, interval)
- **glucose:** The subject's glucose level. (Continuous, interval)

Predicting Variable

- **TenYearCHD:** 10 year risk of coronary heart disease. This will be the dependant variable within our model. (Discrete, nominal, binary)

Categories of Variables

The dataset provided contains variables of two types: **continuous** and **discrete**. Continuous variable types are those who's value may be attained through measurement. There are then two categories of continuous variable data, interval data and ratio data. In our dataset, the continuous variables are represented through interval data where the data is ranked and categorised through precise and continuous intervals, i.e. temperature in degrees Celsius, cholesterol levels, etc. Our discrete variables then also fall into two categories, nominal data and ordinal data. Our dataset is only concerned with nominal data, in which data is fitted into a group with no scale, i.e. colour or other binary values. Much of our nominal data has binary values, which is important within our logistic regression model, particularly with our predicting variable. This is explained further under the section "Logistic Regression".

Preparing the Data

The first step in preparing our data is to remove any null values. These only accounted for 12% of our data, so we were simply able to remove those columns from the dataset. It was then required of us to check for outliers. We were able to visualise this in Figure 1.

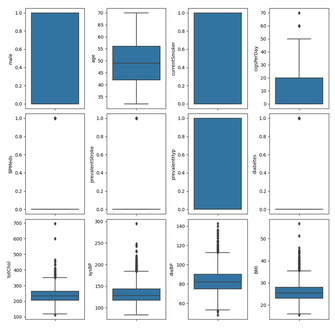


Figure 1. Boxplot to help us visualise potential outliers within the data.

This allows us to see that there are outliers within totChol, sysBP, diaBP, and BMI, so those values must be removed.

Initial analysis

Initial analysis allows us to view the general trends about our data. Figure 1 is contains histograms of various data, and it appears taht our data contains more female subjects than male ones.

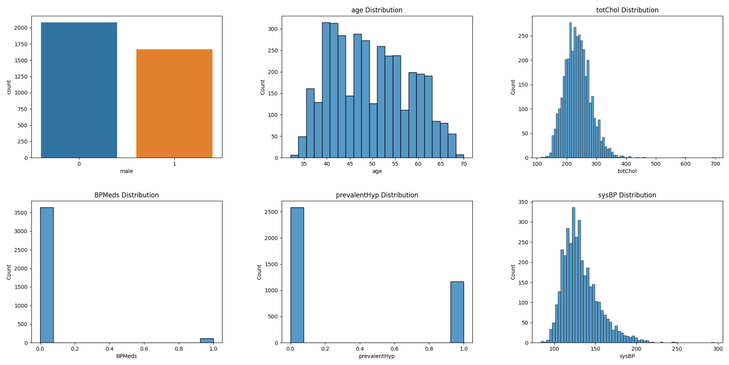


Figure 2. Histograms of variables within the data. (Left to right; Top: Sex, Age Distribution, Total Cholesterol; Bottom: Blood Pressure Medication, Prevalent Hypertension, Systolic Blood Pressure

We may then attempt to further visualise correlation between certain variables through generating a pairplot (Figure 3) and a heatmap (Figure 4).

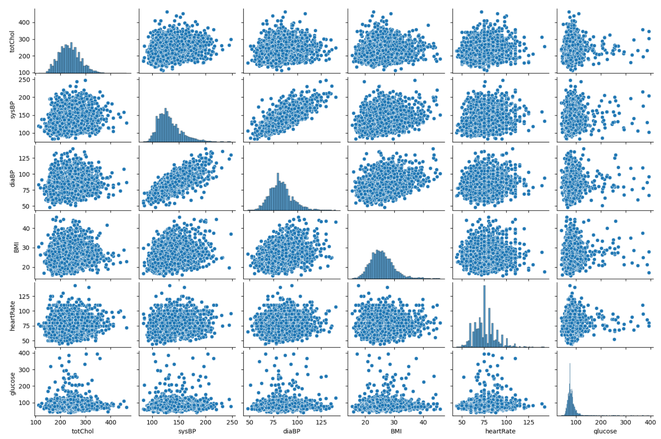


Figure 3. Pair plot to help identify relations between continuous variables.

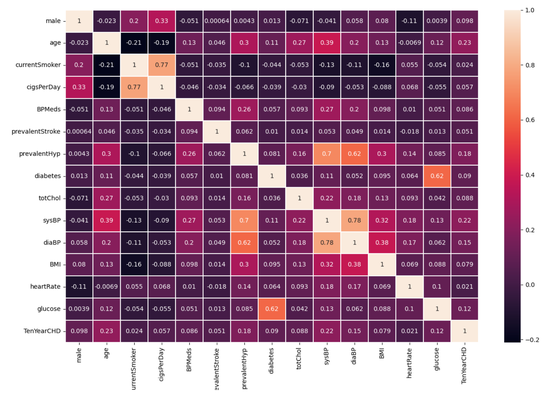


Figure 4. A heatmap graph used to further identify relations between variables.

Highly Correlated Data

From both of these figures, we can identify that diastolic blood pressure ad systolic blood pressure are heavily correlated, along with cigarettes per day and current smoker. These variables are highly correlated and are not necessary within our model, so we may remove cigsPerDay and diaBP from our data.

Logistic Regression

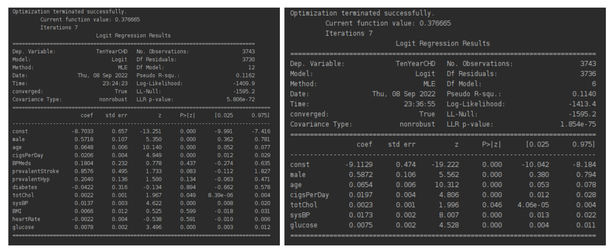
Logistic regression is a form of regression analysis, which is a predictive modelling technique which may be used to identify a relationship between two variables. In the case of logistic regression, this is the prediction of a binary event occurring (in our case, the possibility of developing coronary heart disease within 10 years). The independent variables by which the model is trained on may be either continuous or discrete. Logistic regression follows a rather complex mathematical modal, shown below.

$$p(X) = e^{\beta_0+\beta_1X_1+\beta_2X_2...\beta_pX_p}/(1+e^{\beta_0+\beta_1X_1+\beta_2X_2...\beta_pX_p})$$

This equation follows where X_j is the j^{th} predictor variable (i.e., cigsPerDay, etc.) and β_j is the estimate for the j^{th} coefficient of the predictor variable. $p(X)$ is then our probability, that within whatever threshold will result in either 1 or 0.

Fortunately, libraries like numpy and SKLearn are able to do most of the mathematical work for us. However, before continuing, we must first prepare our data. We do this by generating a constant within the data as well as generating P-Values. The ideal P-Value would be 5% or below, however as seen in Figure 5, initially this is not the case. Therefore, we must remove variables of a P-Value significantly more than 5%, before running regression repeatedly until all P-Values are below 5%.

Logistic Regression (Continued)



```
Optimization terminated successfully.
         Current function value: 0.376665
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              TenYearCHD   No. Observations:                 3743
Model:                          Logit   Df Residuals:                     3730
Method:                           MLE   Df Model:                           12
Date:                Thu, 08 Sep 2022   Pseudo R-squ.:                 0.1162
Time:                        23:24:23   Log-Likelihood:                -1409.9
converged:                      True   LL-Null:                       -1595.2
Covariance Type:            nonrobust   LLR p-value:                 5.806e-72
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -8.7033      0.657    -13.251      0.000      -9.991      -7.416
male           0.5718      0.107      5.360      0.000       0.362       0.781
age            0.0648      0.006     10.140      0.000       0.052       0.077
cigsPerDay     0.0206      0.004      4.949      0.000       0.012       0.029
BPMeds         0.1804      0.232      0.778      0.437      -0.274       0.635
prevalentStroke  0.8576      0.495      1.733      0.083      -0.112       1.827
prevalentHyp   0.2040      0.136      1.500      0.134      -0.063       0.471
diabetes      -0.0422      0.319     -0.134      0.894      -0.667       0.583
totChol        0.0023      0.001      1.987      0.049    8.39e-06       0.004
sysBP          0.0137      0.003      4.622      0.000       0.008       0.020
BMI            0.0066      0.012      0.525      0.599      -0.018       0.031
heartRate     -0.0022      0.004     -0.538      0.591      -0.010       0.006
glucose        0.0078      0.002      3.496      0.000       0.003       0.012
==============================================================================
```



```
Optimization terminated successfully.
         Current function value: 0.376665
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              TenYearCHD   No. Observations:                 3743
Model:                          Logit   Df Residuals:                     3736
Method:                           MLE   Df Model:                            6
Date:                Thu, 08 Sep 2022   Pseudo R-squ.:                 0.1140
Time:                        23:36:55   Log-Likelihood:                -1413.4
converged:                      True   LL-Null:                       -1595.2
Covariance Type:            nonrobust   LLR p-value:                 1.854e-75
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -9.1129      0.474    -19.222      0.000     -10.042      -8.184
male           0.5872      0.106      5.562      0.000       0.380       0.794
age            0.0654      0.006     10.312      0.000       0.053       0.078
cigsPerDay     0.0197      0.004      4.896      0.000       0.012       0.028
totChol        0.0023      0.001      1.996      0.046    4.06e-05       0.004
sysBP          0.0173      0.002      8.007      0.000       0.013       0.022
glucose        0.0075      0.002      4.528      0.000       0.004       0.011
==============================================================================
```

Figure 5. Left: The initial generation of P-Values isn't satisfactory. Right: After removing the variables with highest P-Values and running regression repeatedly we get much more satisfactory results.

Results and Model Evaluation

	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
const	0.000046	0.000602	0.000166	0.000
male	1.436623	2.184140	1.771379	0.000
age	1.053654	1.080377	1.066932	0.000
cigsPerDay	1.012501	1.029132	1.020783	0.000
BPMeds	0.760072	1.887256	1.197686	0.437
prevalentStroke	0.893858	6.218234	2.357587	0.083
prevalentHyp	0.939413	1.600951	1.226358	0.134
diabetes	0.515800	1.781726	0.958653	0.894
totChol	1.000008	1.004482	1.002242	0.049
sysBP	1.007936	1.019738	1.013820	0.000
BMI	0.982229	1.031556	1.006590	0.599
heartRate	0.989672	1.005929	0.997767	0.591
glucose	1.003438	1.012261	1.007840	0.000

Figure 6. Final results for interpretation.

We can draw a lot of conclusions from these results. Males seem to be 77.1% more likely to be diagnosed with coronary heart disease than females, while those with prevalent stroke have a 136% higher chance to be diagnosed with CHD. Surprisingly, we see that glucose and BMI play hardly any role in increasing the chances of potential diagnoses of CHD.

However, we cannot assume that our model is 100% correct. Using the library, we may evaluate our model. Through generating a confusion matrix, we can visualise this. (See Figure 7.)

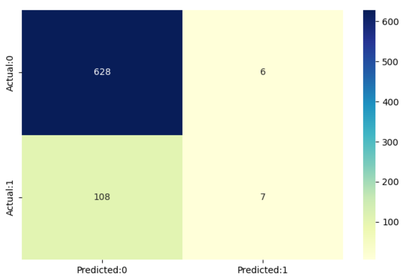


Figure 7. The model made 634 correct predictions and 115 incorrect ones

Running a metric using sklearn has shown us that our model is 84.8% accurate, and in the tests that were run, we received 6 true positives, 628 true negatives, 108 false positives and 7 false negatives. So then, it is clear that our model is flawed in some way, perhaps in the way it identifies positive results. This may be due to our model being highly specific, rather than sensitive, and perhaps also due to the threshold for positive results being too high.

Conclusion

The model outlined in this report has shown the potential for how machine learning and statistical models may allow us to combat non-communicable diseases such as CHD. Our model allowed us to narrow down variables which played significant roles in the development of CHD through analysing their P-values, as well as showing the links between CHD and sex, with males being more likely to be diagnosed with CHD than females, perhaps revealing a bias within our model. It has also led us to a surprising conclusion of how insignificant of a role cholesterol and glucose play in the development of CHD, 0.224% and 0.00784%.

However, our model is not perfect, reaching only 84.8% accuracy in our metric test with many false positives. We may rectify this perhaps by tweaking certain parameters, as mentioned above, or implementing a more complex model, such as the one currently being developed by SANO. (See Figure 8.)

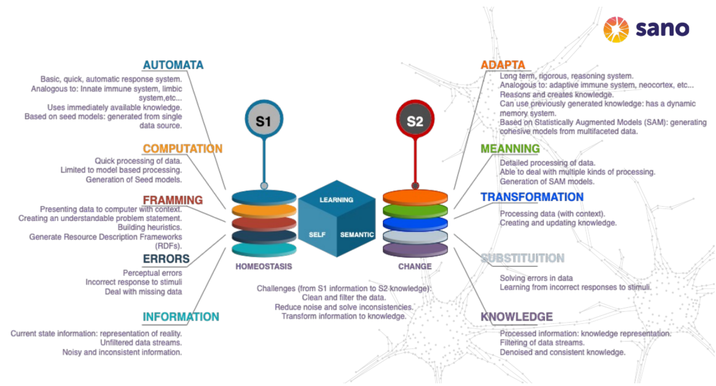


Figure 8. Diagram summarising SANO's approach to AI

References

[1] World Health Organization, Cardiovascular diseases (CVDs), https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] WHO Expert Committee on Prevention of Coronary Heart Disease & World Health Organization. (1982). Prevention of coronary heart disease : report of a WHO expert committee [meeting held in Geneva from 30 November to 8 December 1981]. World Health Organization https://apps.who.int/iris/handle/10665/39293