Cardiovascular Disease Prediction Analysis

Coursera Capstone

by

Chris Buys

July, 2021

Table Of Contents

1. Introduction

1.1 Background

Cardiovascular disease is one of the leading causes of death globally and is the leading cause in the United States. Heart disease costs the United States approximately 219 billion dollars each year which includes the cost of health care services, medicines and lost of productivity due to death (Centers for Disease Control and Prevention, 2020).

1.2 Problem

If people with or without cardiovascular disease could be identified with a certain amount of certainty resources could be better planned for and the target group could be treated more effectively. Especially if we can eliminate false positives (outliers) in treating cardiovascular disease.

1.3 Interest

Addressing this problem would yield benefits for physicians as they get to focus on treating the right people and government who can better plan and allocate resources towards the treatment of cardiovascular disease.

2. Data Requirements & Collection

2.1 Data Description

The data set consists out of twelve total features which are categorised into three types of input features and one target feature (kaggle.com, n.d.). The input features include objective information (factual), examination (result of medical examination) and subjective information (information given by the patient). Table 1 describes the features of the data set:

Table 1 - Data Features

Feature	Category of Feature	Metric	Unit of Measurement
Age	Objective	age	int(days)
Height	Objective	height	int(cm)
Weight	Objective	weight	float(kg)
Gender	Objective	gender	categorical code
Systolic Blood Pressure	Examination	ap_hi	int
Diastolic Blood Pressure	Examination	ap_lo	int
Cholesterol	Examination	cholestorol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective	smoke	binary
Alcohol Intake	Subjective	alco	binary
Physical Activity	Subjetive	active	binary
Presence or Absence of cardiovascular disease	Target Variable	cardio	binary

2.2 Feature Selection

All Features are applicable to include in this analysis however, it would be informative to split the data into features containing the subjective features and not containing to indicate how truthful patients are about this information. Often people either over or under exaggerate and this can introduce issues when trying to treat patients. Depending on the outcome

this could give an indication if whether or not a patients subjective information is needed during this type of analysis.

3. Methodology

The methodology applied in this project will follow the iterative data science methodology as can be seen in Figure 1.

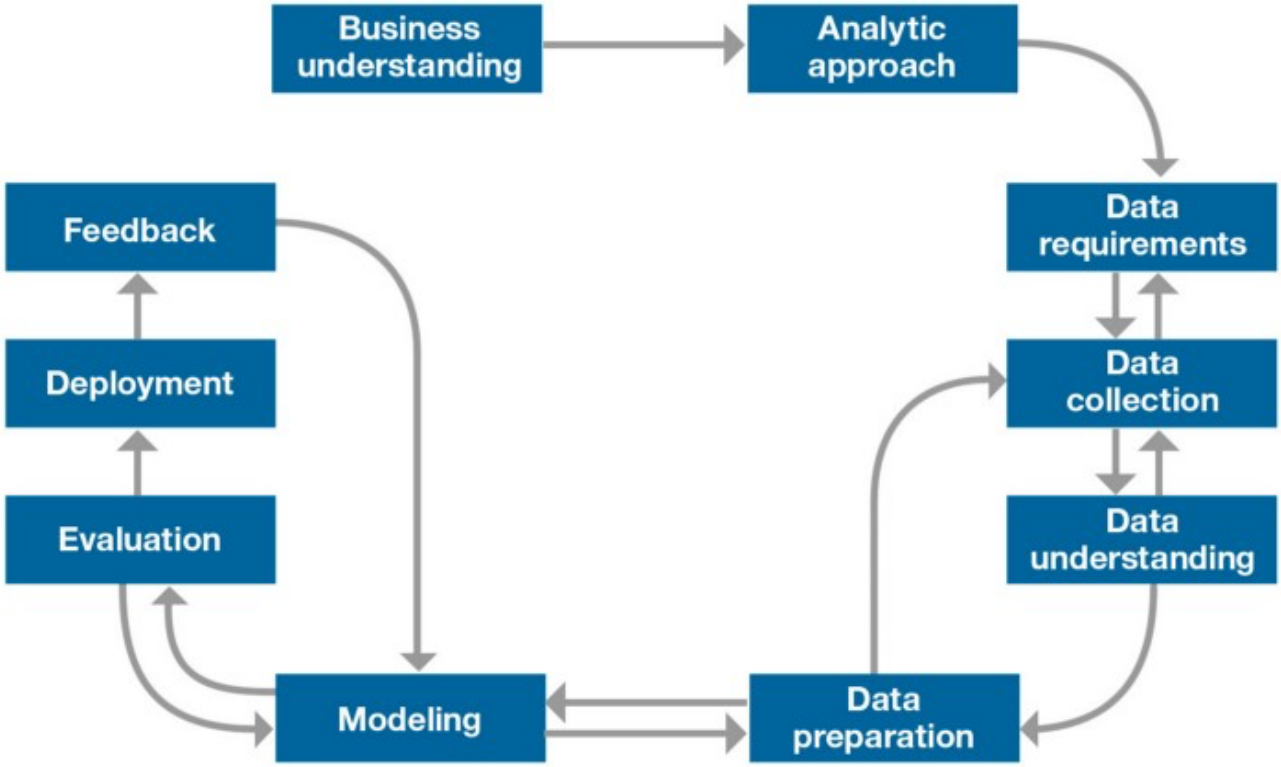

Figure 1 - The Data Science Methodology (Logallo)

Since this project consisted out of trying to identify a business need a relevant problem had to be identified that a business sector could find of interest this was done in Section 1 of this project.

The next step is to identify what types of pattern is needed to to address the question most effectively. In this case the question would require a yes/no answer therefore it would be most effective to use a classification model. The purpose of this study is to classify people has having or not having a cardiovascular disease.

Since the analytic approach to be used requires a classification model a data set needs to be sourced with a binary target variable, this was done in Section 2 of this report. The data that needs to be sourced should also include variables that could have a correlation to the target variable in question.

Once data has been collected it needs to be determined what meaning this data has and if it is representative of the problem that needs to be solved. Descriptive statistic could be used to assist to assist in the understand of the data set.

The next step in the methodology is all about preparing data for analysis. This includes making sure that data is in the correct format, getting rid of any potential outliers and then determining if features need to be engineered. Features are characteristics that might help solve the problem and requires domain knowledge.

Next a data model will be constructed that will be predictive since it tries to yield a yes or no outcome (whether or not a person has a cardiovascular disease). The data will be split into a training and a testing set to construct the model with a 80:20 split.

The model will then be evaluated to measure its performance and whether or not it has potential for further investigation since this is not a ‘real world’ project the model will not be deployed and will not receive feedback for iterative improvements. However, an indication can be given off whether or not this approach can be pursued for this the problem at hand.

At each step in the methodology if anything is unclear an iterative approach will be used where the previous step will be refined to give better input to the current step in the process. Since custom data can’t be sourced for the problem data will need to be found that matches the problem as closely as possible, this limitation will be discussed later in this report.

4. Data Understanding

4.1 Exploratory Data Analysis

Data is checked for any missing or null values whereafter outliers are identified and replaced by the max or minimum threshold of each data set feature depending on what end of threshold the outlier is found. As can be seen in Figure 2 there are abnormal readings for both systolic and diastolic blood pressure, this could be due to data entry errors. It is clear to see that these values are outliers since the normal systolic blood pressure ranges between 90-120 mmHg where as diastolic ranges between 60-80 mmHg (What is blood pressure?, 2021). It is important to get rid of outliers in the data set since it can skew interpretation of the results (Garcia and Garcia, 2021). Figure 3 indicates other outliers in the numerical data of the data set.

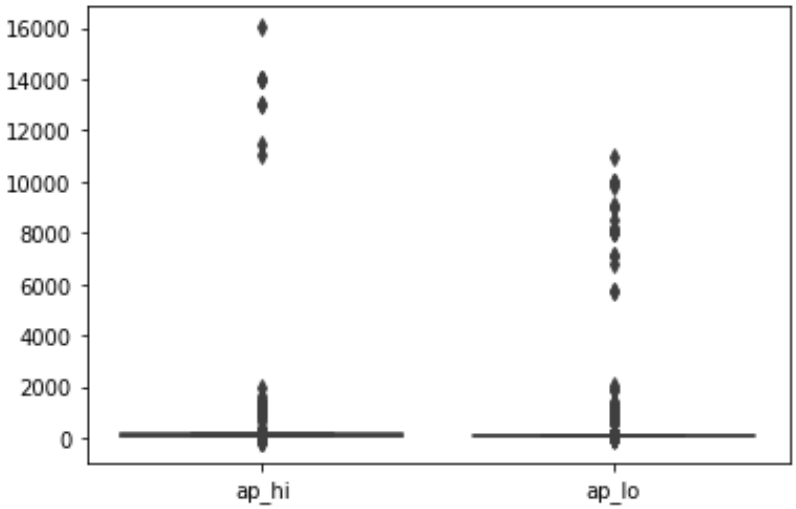

Figure 2 - Outliers Blood Pressure

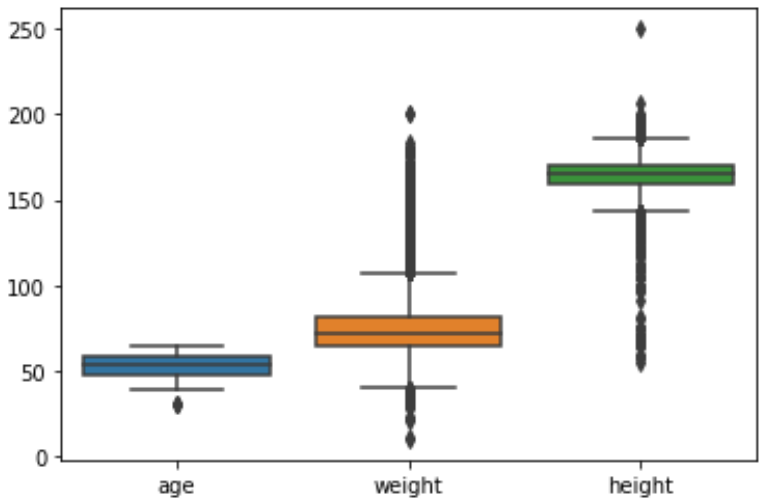

Figure 3 - Other Outliers

Once the outliers are replaced by viewing the summary of the data set is is clear to see that the maximum values for each feature is within realistic ranges as can be seen in Figure 4 where the maximum systolic blood pressure now ranges to a 210.

	age	gender	height	weight	ap_hi	ap_lo
count	70000.000000	70000.000000	70000.0000	70000.000000	70000.000000	70000.000000
mean	53.338686	1.349571	164.3878	74.187047	126.835586	81.841771
std	6.765294	0.476838	7.9838	14.286894	17.604118	10.505388
min	30.000000	1.000000	125.0000	10.000000	50.000000	40.000000
25%	48.000000	1.000000	159.0000	65.000000	120.000000	80.000000
50%	54.000000	1.000000	165.0000	72.000000	120.000000	80.000000
75%	58.000000	2.000000	170.0000	82.000000	140.000000	90.000000
max	65.000000	2.000000	205.0000	145.500000	210.000000	120.000000

Figure 4 - Data Set After Outliers Removed

Dummy variables is also created for the dataset because a machine learning model would treat the order of numbers as an attribute of significance. This means it would read a higher numbers as more important than lower numbers (Data Science in 5 Minutes: What is One Hot Encoding?, 2021).

The data is finally scaled to ensure that there isn't bias in the result and that one feature is not given more weight due its difference in magnitude.

4.2 Inferential Statistical Testing

The correlation between the target variable (cardio) and independent variables are calculated as can be seen in Figure 5. The most significant correlation to cardiovascular disease seems to be blood pressure, weight, cholesterol and age. We would expect to see a positive correlation between smoking and the target variable, yet for this data set there seems to be a weak negative correlation. This is where subjective data of patients is brought in question and it should be asked whether or not people are being honest in this instance. Being active produces a negative correlation as expected while the other correlations seems insignificant.

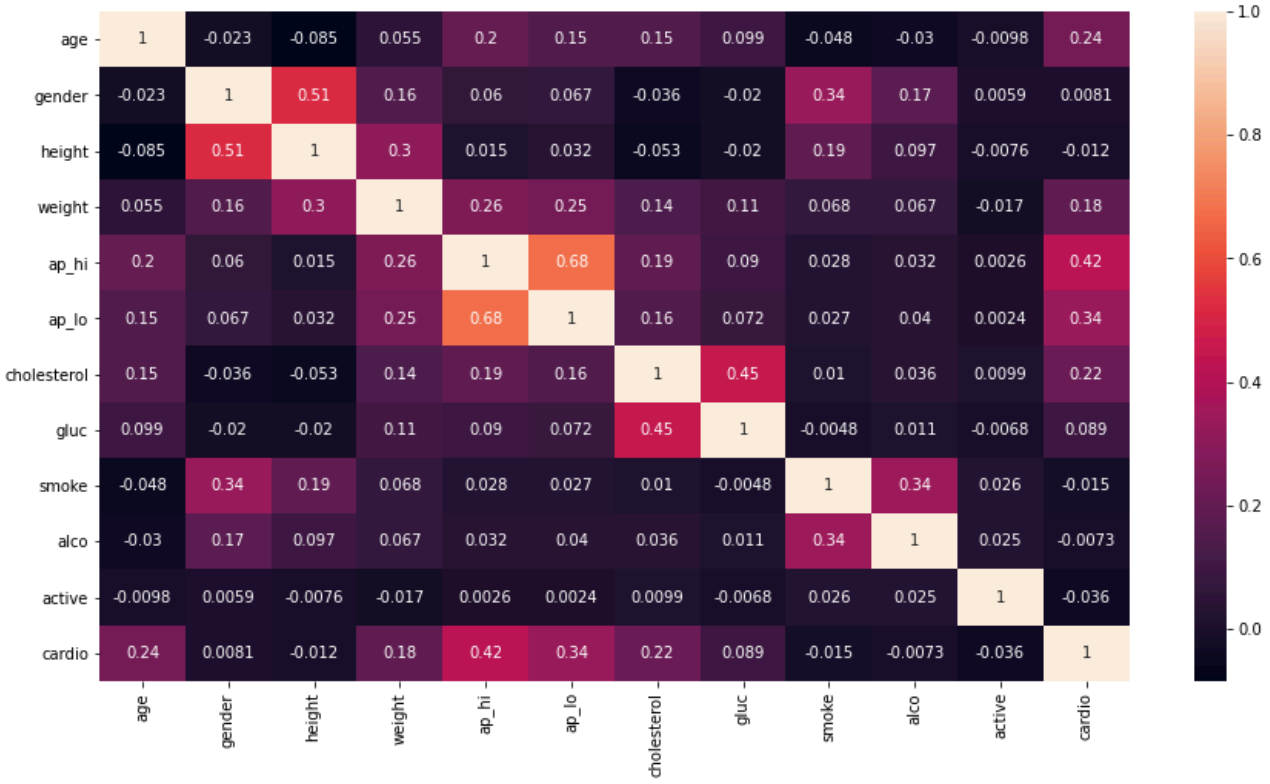

Figure 5 - Correlation Matrix

It must be noted that correlation doesn’t mean causation. This is simply stating that there could be a relationship between these variables.

4.3 Machine Learning Model

The model applied to the data set is a logistic regression model. Since the problem was to categorize whether or not people have a cardiovascular disease along with an amount of certainty that you can make this statement.

Ultimately to help the public health sector to be able to better estimate with a certain confidence level how many people in the general population could potentially have a cardiovascular disease which would allow for more efficient resource planning as mentioned in Section 1.

Although different categorical machine learning models could be applied, this specific problem requires one that will allow a probability value therefore no other classification models are selected.

5. Results

The model built results in the confusion matrix seen in Figure 6. This indicates the models performance on the test data set. The goal of this model is to successfully identify people without cardiovascular disease with a high degree of certainty and to (most critically) not wrongfully classify people as not having a cardiovascular disease when they indeed have (false negatives) as this could potentially lead to loss of life. From Figure 6 it can seen that the false positive rate is 23.94% where as the false negative rate is 29.79%.

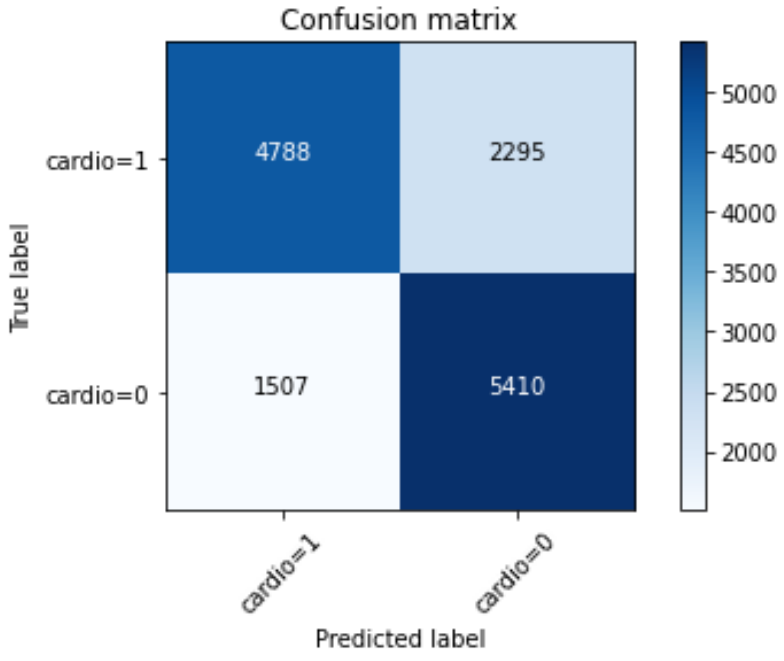

Figure 6 - Confusion Matrix

	precision	recall	f1-score	support
0	0.70	0.78	0.74	6917
1	0.76	0.68	0.72	7083
accuracy			0.73	14000
macro avg	0.73	0.73	0.73	14000
weighted avg	0.73	0.73	0.73	14000

Figure 7 - Model Classification Report

The model produced the classification report as can be seen in Figure 7. This indicated an overall accuracy of 73% and an overall true positive rate of 73%, meaning that is is correctly classifying 73% of the time. The overall weighted average for the model is also 73%, based on the f1-score.

This model produced a log loss value of 0.56, had a training accuracy of 72% and finally a jaccard index of 0.56.

The model is also built omitting the subjective data, a t-test is then done to determine if there is a difference between these models. The p-value is found to be 0.19. Therefore we accept the null hypothesis which states the difference of the means between the two samples is zero. This means that the subjective data makes no difference to the outcome of prediction whether it is included or not.

Features were engineered out of this data set such as bmi, age categories, blood pressure categories etc. These features proofed to have no significance in the performance of the model and were therefore omitted from this report.

6. Discussion

The model that is built has a higher training accuracy than a test accuracy. This can be seen through the jaccard index score of 0.56 meaning that the test data set is not a very good representation of the model and the model won't perform as good as you would expect on the test data set given its 72% training accuracy. There could be various reasons for the this issue however, to overcome this is outside the scope of this project. This can be solved by taking multiple tests sets and getting an average. Another solution would be to get more representative data of the true population.

Some red flags in this data set included the correlation between smoking and cardiovascular disease meaning that either this data set is not representative or that people was not being truthful. This could also then have an impact on the results of the model built.

The model produced a log loss value of 0.56 which we need to be closer to zero to ensure that the probability of outcomes predicted are 100%. When we consider the confusion matrix (Figure 6) what is concerning is the percentage of false negatives (meaning people classified as not having a cardiovascular disease when they in fact do). False positives consequences could lead to over spending by the government sector, which is what this model was aiming to reduce. These false negatives account for 29.79% of the data which puts people at a very high risk if we were to rely on this model to identify people for treatment.

The model is better at identifying people without cardiovascular disease than with. If further refinement is made to the model, by using general population averages we can make an inference on how many people have cardiovascular disease to be able to better allocate resources to treat people more efficiently.

Future work could include determining if the features for this data set is indeed adequate. Just because the feature set is statistically related doesn't mean it is the best feature set to consider. These features can't indicate how far along a person is with their disease and can't act as a pre emptive warning but rather just an indication.

Overall the precision (accuracy) and recall (true positive rate) of the model is fairly good as it is above 70%. There is also no statistical significant difference between the result sets when we consider using people's subjective information or not. Meaning that this has no impact on the accuracy of the results (with a p value of 0.19 therefore we accepted the null hypotheses that the difference between the mean results set is zero). This means that when using this model we could omit people's subjective

information to determine whether or not they have a cardiovascular disease since it might be that they are not truthful in the first instance, this can be illustrated by the correlation between smoking and cardiovascular disease which one would expect to hold a positive relationship yet in this data set there seems to be negative correlation.

7. Conclusion

We can conclude that this could be a good starting point to be able to make an inference on the population and determine the true positive in a population group. However, the inaccuracies in these results could be contributed by an unrepresentative data sample or over fitting. These issues need to be addressed and the best feature set needs to be determined. As mentioned these aren’t necessarily features that are statistically related but biologically rated so that the model could also server as a pre emptive indication (early indicator) of whether or not people will develop a cardiovascular disease.

Future work needs to be done to see which features can best determine cardiovascular and give a pre emptive indication rather than indicating if a person has the disease or not.

This model wouldn't be suited for use but shows promise for this classification problem. Other approaches that could be used are decision trees, k-nearest neighbour and so on.

8. References

Centers for Disease Control and Prevention (2020). *Heart disease facts & statistics*. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/heartdisease/facts.htm.

kaggle.com. (n.d.). *Cardiovascular Disease dataset*. [online] Available at: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

Logallo, Nunzio. “Data Science Methodology 101.” Towards Data Science. [online] Available at: https://towardsdatascience.com/data-science-methodology-101-ce9f0d660336.

nhs.uk. 2021. *What is blood pressure?*. [online] Available at: <https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/> [Accessed 19 July 2021].

Garcia, C. and Garcia, C., 2021. *How to Find Outliers in a Data Set - Atlan | Humans of Data*. [online] Atlan | Humans of Data. Available at: <https://humansofdata.atlan.com/2017/10/how-to-find-outliers-data-set/> [Accessed 19 July 2021].

Educative: Interactive Courses for Software Developers. 2021. *Data Science in 5 Minutes: What is One Hot Encoding?*. [online] Available at: <https://www.educative.io/blog/one-hot-encoding> [Accessed 19 July 2021].