Predicting Risk of Cardiovascular Disease (CVC)

Introduction

Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. This project uses the Framingham Heart Study dataset to create an algorithm which predicts the chance a patient will develop CVD based on the patient's data.

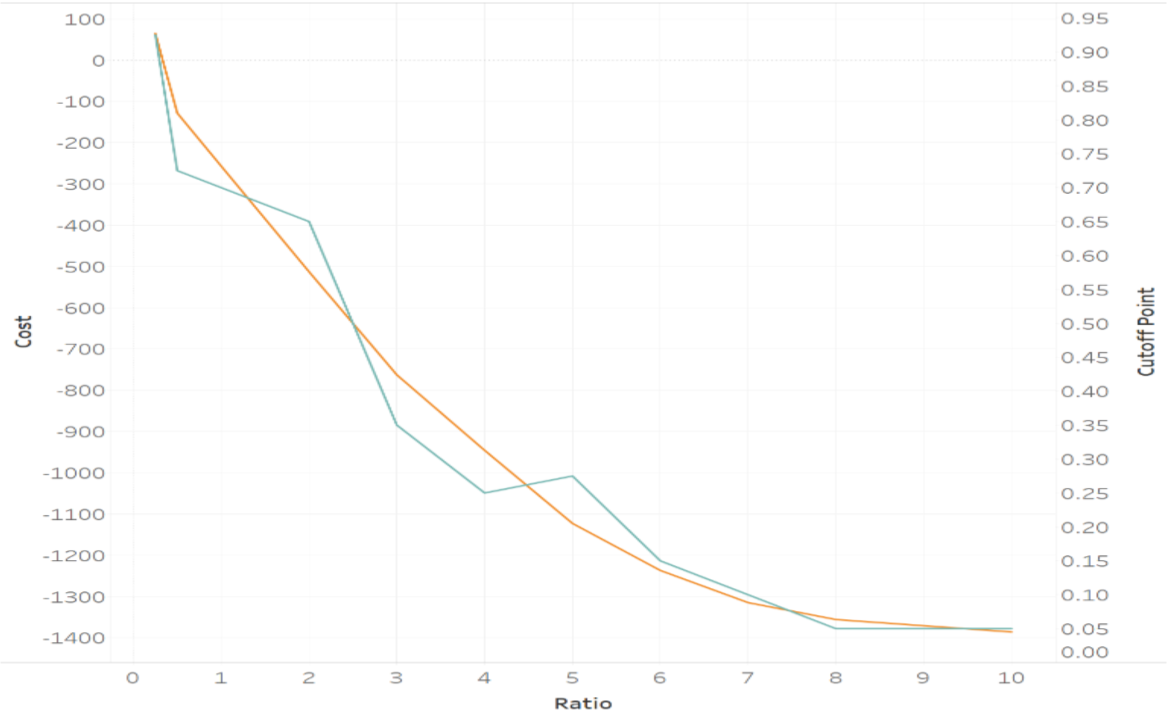
1. Identifying the best Cut-Off Value to use for Logistic Regression based on operation cost

Before evaluating a logistic regression, I must choose what the cut-off point must be. The cut-off point is the cut-off probability that the logistic regression must calculate for a result to be considered positive.

This is important to consider; predicting that a patient doesn't have CVD when they do would be more costly than wrongly predicting the patient does have CVD – this must be taken account of in the model.

How do costs change as the cost ratio changes?

Choosing the best cut-off point depends on the cost ratio. If wrongly predicting costs 5 times more, the ideal cut-off point will be different than if it costs 10 times more. The table below shows the ideal cut-off point and the costs for different possible ratios (maintaining cost of wrong prediction) The blue line corresponds to the cut-off point and the orange line corresponds to the cost.



The trends of Cost and Cutoff Point for Ratio. Color shows details about Cutoff Point. The data is filtered on Ratio as an attribute, which keeps non-Null values only.

.

Costs go down as the ratio decreases (because false negatives become less expensive). However, they change at different rates, with bigger change happening in the lower ratios. The best cut-off value changes drastically at lower ratios. At ratio 1:1, the cut-off point is 0.7. However, at ratio 4:1 the cut-off point is 0.25, and at 7:1 it is 0.1. This means that if the cost of a false negative is relatively very expensive (greater than 7:1 compared to a false positive), a very conservative cut-off point of 0.05 would be justified.

How would these costs be estimated in real life?

I am measuring the risk of someone developing a heart disease in a 10-year span. The costs of not predicting this are potentially huge (death). However, the risks of wrongly predicting it are also large (costly treatment, mental stress). A proper estimation of these costs would involve analysing expenses for patients diagnosed properly and those who weren't. Patients are likely to take multiple tests in a 10-year span, so it should not be necessary to weight all the costs in one test. It could prove very hard to isolate the data properly as it ranges a large timespan.

Does it make sense to use the same cost ratio for each individual?

Some individuals will need more expensive treatment than others. Many factors should be observed (is it lifestyle based? I.e smoking - or based on demographic charactecrics?). Treatments will depend on many factors. How long would they stay in a hospical? How many tests did they use. Did they have to use an ambulance if a misdiagnosed patient had a stroke? If a misdiagnosed party is hospitalised, is someone liable?

How can one use the results from the sensitivity analysis if costs differ among individuals?

If only one model can be chosen, a weighed average cut-off point could be computed, weighted on the number of people that have each cost-ratio. However, if the model cut-off point can be changed, then the sensitivity analysis can dictate what cut-off point should be used for each individual, assuming that reliable information of the costs can be obtained.

2. Modelling

EDA – Multicollinearity with diaBP, and sysBP.

In the model incorporating both sysBP and diaBP, the coefficient for sysBP is positive, but the coefficient for diaBP is negative (with a non-significant p-value). However, diaBP has a positive coefficient when sysBP is not present, with a significant p-value. Diastolic blood pressure is highly correlated (0.78) to systolic blood pressure. This means that running a model with both these variables will lead to multicollinearity issues, which could lead to the model thinking one of the values has a negative coefficient. Diastolic pressure refers to a measurement during the beginning of the cardiac cycle, while systolic pressure occurs when the ventricles contract. Intuitively, it makes sense than an increase in blood pressure in one is highly correlated with the other.

Would Principal Component Analysis (PCA) help improve model accuracy?

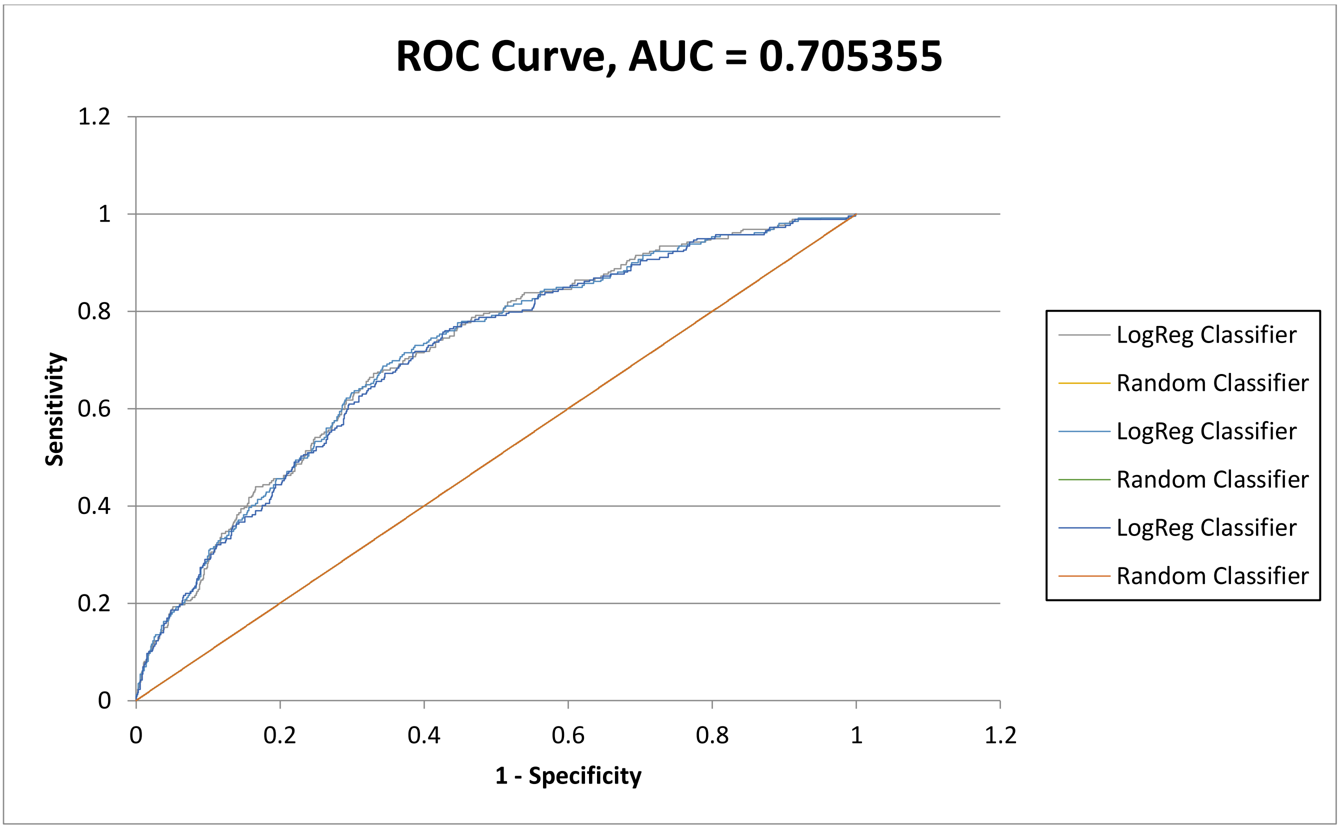
Adding on to the sysBP and diaBP PCA, it may be a good idea to find variables that are highly correlated to sysBP and diaBP. For example, prevalentHyp Is highly correlated with both (0.70 and 0.62), so that would be my first choice. Checking PCA analysis with sysBP, diaBP and prevalentHyp, 80% of the data can be explained with 1 variance, and 93% with 2.

Checking the PCA for all variables, 90% of variance can be explained using 13 of 19 variables (including dummies). Sex has the highest explanatory power, followed by the age and smoking variables.
It could thus be beneficial to implement PCA. Dimensions can be reduced (addressing curse of dimensionality issues), while still keeping a large amount of the explanatory power. Perhaps more

importantly for logistic regression, multicollinearity issues would be dealt with, which would allow usage of all sysBP, diaBP and

Testing Models

I ran three logistic regressions using age, BPMeds, totChol, glucose, and ciggaretes per day, but alternating between sysBP, diaBP and a PCA term of sysBP and diaBP. In terms of AUC, sysBP gives me the best results at 0.7125 R^2 scored. However, the ideal model varies (slightly) depending on which cut-off point is used.



Above are is a comparison of the three models, showing that they regularly overlap.

The ideal cut-off point will depend on the cost ratio, and the ideal model will depend on the cut-off point.