Project 2 – CHD

Author(s):

Gabriel Jackson

Naad Kundu

Thao Nguyen

Kayla Nguyen

Halbert Nguyen

Ben Willoughby

Omar Zeineddine

DS 3001: Foundations of Machine Learning

April 16, 2024

Summary

Our main focus for this project was to build predictive algorithms in order to find out the likelihood of a person developing coronary heart disease. There were many different variables we could use, such as sex, age, education, currentSmoker, cigsPerDay, etc. We used provided data to build our training model, as well as to test our model. Firstly, we cleaned our data by updating any NAN values. We had to clean the education variable, the glucose variable, the BPMeds variable, the totChol variable, the BMI variable, the cigsPerDay variable, and the heartRate variable in terms of NANs. We then saved these cleaned datasets into updated csv files. After cleaning the data, we used linear regression and decision trees in order to predict the likelihood of a person developing coronary heart disease. For our linear models, we used the sklearn library. We first read in the csv files, including the test data. Our linear regression models results were interesting, due to our dummy education variable. We decided to make two linear models to compare to each other- one with all variables, and one with all variables except the education variable. We found that the adjusted R^2 values for the model with all variables was higher than the model without the education variable, showing that the model with all variables had more explanatory power for predicting CHD. For our decision tree model, the results were also interesting- our model had an 85% accuracy rating. It did end up having a lot of false positives though- it incorrectly predicted yes to CHD for 143 cases where the person didn't have CHD.

Data

In terms of the data, two csv files were provided- one for training our model and one to test the model. The variable we are trying to predict is 10YearCHD, which is the 10 year risk of coronary heart disease. Firstly, we had to clean the training dataset. For our education variable, we replaced all NAs with 0s, in order to keep its value numerical. For our glucose variable, we filled all NAs with 85, which is the average. This was done to keep the data from being skewed for our models. For our BPMeds variable, we replaced all NAs with 0- to represent that they are not being taken. For our totChol (total cholesterol), BMI (body mass index), cigsPerDay, and heartRate variables, we decided to drop all the NANs. We then saved these cleaned datasets into their respective updated csv files, in order to prepare them to be used by botb our linear regression model and our decision tree. To begin our linear regression models, we had to update our education variable first. One big challenge we encountered was with our education variable. The education variable was categorical, but it appeared numerical. We proceeded to change the entries into text, so that we could easily create dummy variables. We replaced its values from 0-5 to ‘Unknown education’, ‘Some high school’, ‘High school/GED’, ‘Some college’, ‘College’, respectively. In our dataset overall, there are multiple dummy variables. For a lot of the variables, it is intuitive, as 1 is true and 0 is false. But for our sex variable, 0 represents female and 1 represents a male. We proceeded to construct a linear model with no intercept, using the variables sex, age, our education dummy variable, and currentSmoker, cigsPerDay, BPMeds, prevalentStroke, and prevalentHyp. Calculating R^2 and adjusted R^2, we got 0.0998 and 0.0943 respectively. We proceeded to make another linear model, this time without the education variable, and our R^2 was 0.0832 and our adjusted R^2 was 0.0790. It didn’t seem like education played much of a factor in coronary heart disease, which is why we decided to create two models- one with the education variable and one without it. Looking at the R^2 values, it seems like the model with all available variables, including education, had more explanatory value due to the higher adjusted R^2 value. We encountered some challenges when trying to make our predictive model. Firstly, the variable we were trying to predict, 10YearCHD, can only take a binary value of 0 or 1, meaning we have a linear probability model. This implies that the predicted values are understood as probabilities of the event occurring. Also, under linear probability models, the regression line will never fit the data perfectly if the dependent variable is binary and the regressors are continuous.This means the R^2 value, our primary tool for measuring the explanatory power or 'usefulness' of a model, loses its interpretation. Since linear models predict probabilities, we decided to try and use a decision tree as well, to compare our results.

To make a decision tree, it was fairly simple to write the code. We also used the sklearn library for the decision trees. We used the same cleaned data files as we did for our linear models. A couple of challenges did arise though- the biggest one was attempting to limit the amount of false positives the decision tree kept giving us. The decision tree incorrectly predicted yes for 143 cases when they had no coronary heart disease. The accuracy was 85%.

Results

Conclusion

Appendix