Kiejstut Bunikiewicz

Professor Brahma

BSAN 6070 – Machine Learning

3/23/21

CA05A – Logistic Regression: Question Responses

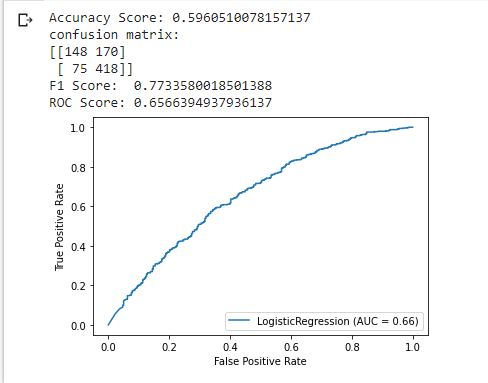
Part 1:

I built a binary classifier model to predict risk of Cardiovascular Disease using a Logistic Regression to estimate risk, using descriptive factors.

Part 2:

Upon running prediction using the model, I created standardized coefficients to estimate the effect of various factors. I took the absolute value of these coefficients and ordered them in descending order.

Part 3: Evaluate Performance Outputs and Draw Conclusions



The accuracy score is the number of correct predictions over the number of total predictions. The accuracy score of 0.59 is not especially troubling because it is on the threshold of 0.6, which is the rule of thumb for a reasonably useful model. However, this model is more useful than previously assumed because it is more likely to capture false positives than false negatives. With medical predictions such as this, raising a false alarm is better than refusing to diagnose a potentially harmful case. The High F1 score of 0.77 means that the model effectively balances out positive (at risk of disease) and negative predictions, rather than opting to prefer positive or negative outcomes. The ROC curve is relatively effective, because it is in the range between 0.6 and 0.7. The score of 0.66 means that the model performs within the acceptable range. This model has a lift of approximately 16% compared to guessing randomly. Therefore, the model is useful when trying to predict the likelihood of being at risk. Overall, the model is useful because it is an effective predictor of risk of cardiovascular disease. Also because this model is being used for medical prediction, the propensity for false positives rather than false negatives is not a problem. The model is satisfying business goals by opting for false positive (needing re-diagnosis), rather than false negatives (which will leave serious cases untreated).