Zackery Bradley

BAN 525

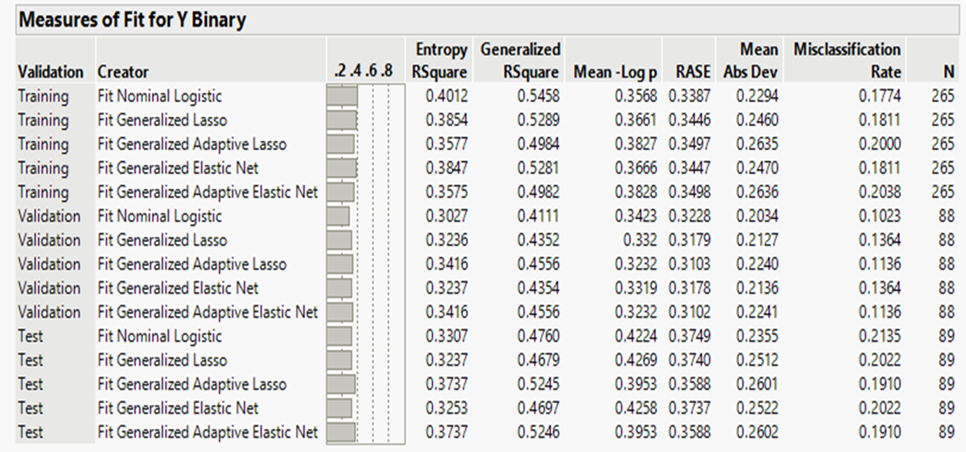
20220206

Model 3 Assignment

In this week’s assignment, we have been tasked to provide an analysis to the factors that contribute to diabetes. Our data is comprised of ten baseline variables (age, gender, body mass, index, average blood pressure, and six different blood serum measurements) on behalf of 442 diabetes patients. We will be using the ‘Y Binary’ variable as our response, which indicates whether the disease gas gotten better or worse (low, high) in regard to the particular patient. On the topic of our variables of interest, we will be focusing on the blood serum measurements specifically, as they are hypothesized to have the biggest impact on model predictions. Particularly, we are predominantly interested on the effects of Total Cholesterol, and low-density lipoprotein (LDL), as they tend to be significant factors in regard to diabetes. Throughout our analysis, we will be using various methods to ensure we get the best model for the data in question. These methods will include our standard method, Ordinary Logistic Regression (OLS) and various penalized logistic methods such as Lasso, Adaptive Lasso, Elastic Net, and Adaptive Elastic Net respectfully. All of these methods have distinct capabilities that can be highly desirable based on the type of analysis being performed. Lasso has the capability of shrinking uninformative coefficient all the way to zero, essentially eliminating their respective influences. Lasso typically bodes well in an environment where prediction accuracy is of the upmost importance. Elastic Net is a combination of the Ridge and Lasso methods in which a mixing function (usually fixed) is implied. Like Lasso, Elastic Net has also shown great prediction accuracy. We will also be utilization adaptation in regard to these two methods, as this strategy will give us an innovative idea about which variables are of the most importance. Employing adaptive strategies would allow us to penalize important variables at a much lesser rate, thus creating a more accurate model.

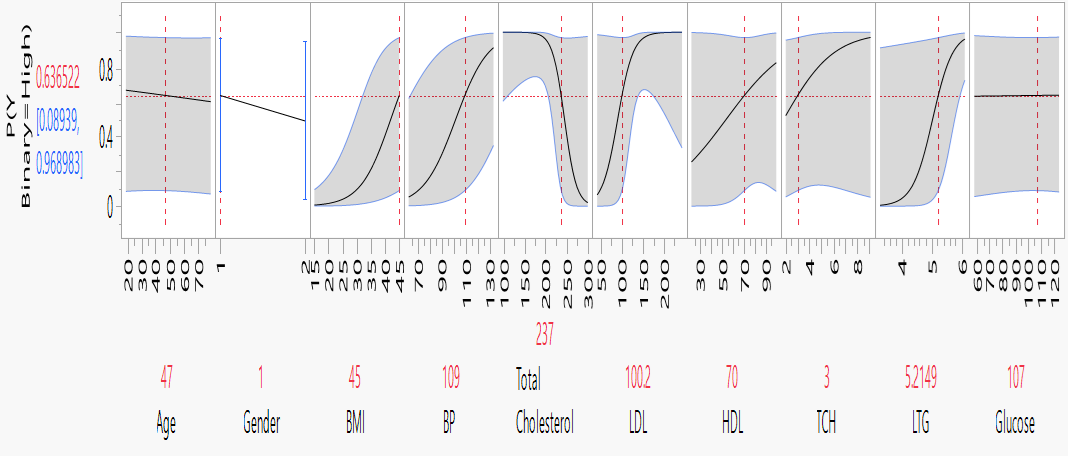
While all of our methods have been proven to be dependable in the right conditions, each method does not come complete without its own pros and cons. Lasso, for example, typically avoids over fitting, however it also tends to be highly biased based on selections. Elastic Net can be especially valuable when more than one predictor is utilized but can also be detrimentally more expensive to compute than Lasso. Given the constraints of each method, the adaptive and Cauchy strategies are implemented as to enhance our overall capability. Furthermore, penalized regression methods are typically considered modern approaches in statistical methodology. These methods work by shrinking the size of coefficients and retaining those with coefficients greater than zero, usually producing a slightly better result than traditional methods. Procedurally, we will be employing a cross-validation analysis. Cross-validation techniques are built on the concept of leaving out a part of the data out of the estimation process as a buffer. As the predictions of the model stop improving as the data holdout process occurs, model growth then stops, and estimates are obtained. This creates a ‘split’ in the data. These splits are then broken down into three parts in the terms of our model. Those parts include training, validation, and testing splits respectfully. Training data is the portion that is used to estimate our model. Validation data is not used in estimate directly, but instead works in the background to determine the optimal point in which the model stops. Lastly, the test data is never actually used in the model estimation. Test data is used to represent “new observations” and assists in providing and unbiased analysis of the predictive ability of the model. With the validation method only being utilized indirectly, we choose our model that displays the best performance on the test data.

Upon examination of our data, we can see that while all methods were similar in analysis, two models in particular proved to be remarkably close together. For our Adaptive Lasso method, we see a RSquare of 0.5245, followed by a RASE of 0.3588. In comparison, our Adaptive Elastic Net model produced an RSquare of 0.5246 and a RASE of 0.3588. While these models are within a fraction of each other, our analysis results in the Adaptive Elastic Net method being the most efficient model. Given the outcome of this model, our data performance has shown that utilizing several different statistical methods in our models is of the upmost importance. Those small degrees of performance enhancement can cost companies millions if the wrong decision is made based on inferior data.



After thoroughly sifting through our data, our chosen model displayed which variables were of most importance. Shown in the visual below, we can see that the variables responsible for the biggest impact of whether patients had made positive progression of not were Total Cholesterol, LDL, and LTG respectfully. Total Cholesterol, accounted for an astounding total of 0.567 total overall effect when examining patient data. This is followed by LDL with a 0.372 total effect and LTG with a total effect of 0.234. Given these three variables’ effects, it can be strongly determined that Total Cholesterol, along with LDL and LTG levels play a role of great significance in diabetes progression. 

Upon further analysis, we were asked to determine the predicted probability of a certain individual given the data we had found. This individual was that of a 47-year-old with a BMI of 45, BP of 109, Total Cholesterol of 237, LDL of 100.2, HDL of 70, TCH of 3, LTG of 5.2149, and a Glucose of 107. Given the numbers for the requested information, we came of with a prediction profiler that measured a Y Binary of 0.636522 for this particular individual. The visual of this individual is displayed below.



In hindsight of our analysis, we uncovered a plethora of useful information. For parameter estimate, our selected model indicated that there were a number of insignificant variables in correlation to our data analysis. Shown below, we uncovered that Age, Gender, LDL, TCH, and Glucose levels had an insignificant impact on whether our patients made progress or declined in their health.



In closing of our analysis, we can say with certain confidence that Adaptive Elastic Net performs adequately as opposed to other methods in regard to this particular data. This can be seen on our overall analysis of the other performance methods as Adaptive Elastic Net has proven to be superior statistically speaking. Given the detrimental impact one variable can have on decision making, it is imperative that all companies study model performance in order to generate the best outcome for their futures.