Understanding Diabetes Progression

BAN 525

Khayrayyah Haamid-Day

In this assignment the factors that contribute to diabetes will be analyzed. The dataset to be used contains various collected metrics that show the one-year progression of patients with diabetes. The response variable will be Y Binary which is classified as ‘Low’ or ‘High’ and corresponds to the variable Y; this variable contains numerical values of measurement. The predictor variables range from age and gender to total cholesterol and glucose, essentially variables deemed to be of importance to the health of a diabetic patient. Preliminary analysis indicate there are strong relationships between Y Binary and many of the predictor variables as well as significance. This analysis will be using a variety of statistical methods to discover the best predictive model. These methods include ordinary logistic regression and the penalized logistic regression, that is, lasso, adaptive lasso, elastic net, and adaptive elastic net.

Each method employed have their strengths and weakness that are essential to finding the most optimal predictive model. The first method to be created is the ordinary logistic regression and this due to the response variable being binary. The results from this estimation method are the same as the maximum likelihood results. Like the other baseline methods, logistic regression will fit the entire model and no variable selection will be performed. The next estimation method that will be used is lasso, this is a penalized regression that will apply a penalty that will reduce redundant coefficients to zero. All coefficients are equally penalized and will perform variable selection simultaneously. The modified version of lasso, adaptive lasso, will also be used and will employ weighted penalties. The main disadvantage to using the lasso method is that it may ignore insignificant variables that could still be of interest. The final estimation method is elastic net, and this method uses the and penalty. These penalties used together ensures variable selection and improvement of the predictive model by shrinking the coefficients. The adaptive elastic net method will also be used and works similarly to the adaptive lasso method. Elastic net typically provides more accurate predictions than lasso when there is high correlation among the variables. This analysis will also be utilizing the random validation column feature in JMP with a 60/20/20 split on training, validation, and test set. This validation will also be given a random seed of 123 and will help reduce random noise.

Text

Description automatically generated with medium confidence



A quick analysis of these models shows very similar RSquare values as well as each showing a significant relationship with variables BMI, BP, LTG. The penalized regression models also returned lambda penalties, with the adaptive methods returning higher penalties than their counterparts. Due to the fac that this is a binomial distribution, the RSquare values are not the only aspect to be considered and, as a result, the ROC curve will be considered when determining the best predictive model. These current results indicate that the test set will likely be very close amongst each model and that the best predictive model will only be the best by a small margin.

Table

Description automatically generated

This model comparison looks at model with the predictor variable of Y Binary (High) and as expected, each test set share similar RSquare values, both entropy and generalized. Entropy RSquare is simply comparing the log-likelihoods from the fitted model and probability model, this is only used with categorical variables. Regardless of this, the same rules apply, values closer to 1 indicate a better model. The next important column to consider is AUC, area under the ROC curve, and this is also only used with categorical variables. A strong indication of a well fit model is an AUC value closer to 1. In this comparison, both adaptive lasso and elastic net share the same AUC of 89% but adaptive elastic net has the higher RSquare by 1 point, 0.5246 versus 0.5245. Due to this, adaptive elastic net will be considered the best predictive model.

Text

Description automatically generated with medium confidence

The parameter estimates of the adaptive elastic net show that Y Binary = High has a significant relationship with predictor variables BMI, BP, and LTG. When considering the overall impact of one’s health and the effect this may have on preexisting conditions, such as diabetes, these are unsurprising results. There is a known relationship between a person having a high body mass index and high glucose levels. This is because a high BMI suggest obesity which can increase the risk of becoming diabetic as well as exacerbate symptoms if a person already diabetic. The same is true for a person with high blood pressure, those with high blood pressure have significantly higher concentrations of glucose. LTG refers to triglyceride levels, higher amounts of this are an indication of insulin resistance which keep cells from absorbing glucose intake. This will result in the need for higher levels of insulin that, in turn, raises a person glucose levels. It should be noted that this model removed the variables of Age, Gender, LDL, TCH, and Glucose.

Table

Description automatically generated

The variable importance summary reports that the most important variable in this model is LTG, the measurement of triglyceride levels, but this is only by a small margin. The small difference between a persons BMI and LTG, in terms of which is the most important predictor, indicates that the two might be closely related when it comes to a person having high glucose levels. One predictor variable that was not found to be significant in the parameter estimates, but is for variable importance, is HDL. This is high-density lipoprotein that is considered to be the “good” type of cholesterol wherein high levels of this actually lower risk of heart disease and stroke. A diabetic will typically have high levels of triglycerides and low HDL levels and the parameter estimates does show that HDL has a negative estimate, even if it is not considered significant.

Chart, line chart

Description automatically generated

The prediction profiler further elaborates on these variables of importance, it is shown that as LTG, BMI, and BP so does the response variables. This is reinforcing what was stated previously, as is the relationship between HDL and the response variable. HDL is the only variable that has an inverse relationship with Y Binary = High. Total cholesterol does not show a significant relationship, the impact it has on the response variable is not strong. To further test out the results of this model, a new case was added. This case is a patient Age = 47, Gender = 1, BMI = 45, BP = 109, Total Cholesterol = 237, LDL = 100.2, HDL = 70, TCH = 3, LTG = 5.2149, and Glucose = 107. The adaptive elastic net model already states that a high LTG, BMI, and BP will also increase the probability of Y Binary = High; it also states that lower HDL will increase the response variable. The added case reiterates this, it shows a probability of 0.6365 or 64%.

In conclusion, the predictions in this model reflect the reality of diabetic patients and highlights the way high glucose levels impact overall health.