Alec Engel
May 30, 2021

Introduction

The question at hand here is to determine what if any factors play a role in the prediction of diabetes disease progression within an individual. A disease progression from one year ago of over 200 shows a nominal response variable result of “high” for the disease getting worse, and a disease progression from one year ago equal to or lower than 200 has a nominal response variable result of “low” for the disease getting better. We want to run several cross-validation models to see what predictor variables can be strongly correlated with our binary response variable of disease progression. This cross-validation method will allow us to split our dataset into training, validation, and test portions. The split allows for better predictions of future outcomes with reduction in potential overfitting and removing unnecessary model noise amongst highly correlated independent variables.

Predictor variables for diabetes progression consists of ten baseline variables. The most important variables in this data set are of course the nominal response variable of low vs high disease progression; along with key predictor variables of body mass index (BMI), average blood pressure readings (BP), logarithm of triglyceride levels (LTG), along with measures of high-density lipoproteins (HDL) and total cholesterol numbers. As we will elaborate on later, each patient’s BMI, BP and LTG will have highly unique information as it pertains to predictability with diabetes progression.

In order to narrow down our predictor variables, we ran several statistical method models and judged our outcomes. Statistical methods used include cross-validation tactics of Logistic Regression, Elastic Net (Adaptive and Normal), Lasso Regression (Adaptive and Normal). The latter two methods involve assessing penalties to our independent variables in order to place more emphasis on important variables and less emphasis or in some cases eliminating uninformative predictors.

Analysis and Model Comparison

Logistic Regression involved the inclusion of all 10 predictor variables which often leads to overfitting or the misuse of highly correlated variables. While using all variables, there are often noisy and unnecessary data variables that are accounted for. For a clearer picture, we

then move on to penalized regressions of Lasso and Elastic Net models. Lasso will eliminate uninformative or highly correlated variables while Elastic Net may still include such variables. We will also run adaptive versions of each method. Adaptive models will take into account Logistic Regression estimates of all variable importance and implement smaller penalties against our pre-determined most important variables.

In order to begin with these cross-validation techniques, we need to first create a "validation" column in our dataset using JMP's validation column creator. This will split our data into specified training (60%), validation (20%) and testing (20%) sets to allow our model to gain predictive attributes and score our model comparisons on how we believe future data may react to our models. A fixed random seed of 123 is also used during this process so that we are able to replicate the splits. After cross-validation, we created predictor formulas with each of the above five methods described. Each method is then added as a dataset column so that we can easily perform model comparisons.

Model comparison results can be seen below. Here we are not looking at R squared values like in the past because categorical data is being used. We are instead highlighting area under curve (AUC) and misclassification rates. For AUC, we want the highest possible number as this is a percentage measure of how well our model predictions are versus actual outcomes. The opposite goes for misclassification rate as we want a lower number measuring success of our model. First with the "Training" notation, we see all five models having relatively high performance. All five are above 88% AUC have very small misclassification rates as well. These however are our training sub sets of data and we want to focus more towards our testing performance. Moving to "Test" notations, we see the same patterns in regards to relatively strong AUC values across the board with normal distribution models of Adaptive Lasso and Adaptive Elastic Net appearing to be our two strongest models. Although I will dig into each model to see what our strongest prediction profiler variables are, I want to choose the strongest model of Adaptive Elastic Net with a testing set AUC value of 0.8960 and misclassification rate of 0.1910 as my predominant model. We are splitting hairs here between Adaptive Elastic Net and what looks to be a nearly identical model of Adaptive Lasso with the same AUC and misclassification rates, but each model uses identical variables so I will choose Adaptive Elastic Net to narrow down to one model. Adaptive Elastic Net shows nearly 90% accuracy in sorting our unbiased test data.

Model Comparison

Measures of Fit for Y Binary

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Nominal Logistic		0.4012	0.5458	0.3568	0.3387	0.2294	0.1774	265	0.8937
Training	Fit Generalized Lasso		0.3854	0.5289	0.3661	0.3446	0.2460	0.1811	265	0.8871
Training	Fit Generalized Elastic Net		0.3847	0.5281	0.3666	0.3447	0.2470	0.1811	265	0.8880
Training	Fit Generalized Adaptive Lasso		0.3577	0.4984	0.3827	0.3497	0.2635	0.2000	265	0.8809
Training	Fit Generalized Adaptive Elastic Net		0.3575	0.4982	0.3828	0.3498	0.2636	0.2038	265	0.8809
Validation	Fit Nominal Logistic		0.3027	0.4111	0.3423	0.3228	0.2034	0.1023	88	0.8351
Validation	Fit Generalized Lasso		0.3236	0.4352	0.332	0.3179	0.2127	0.1364	88	0.8434
Validation	Fit Generalized Elastic Net		0.3237	0.4354	0.3319	0.3178	0.2136	0.1364	88	0.8434
Validation	Fit Generalized Adaptive Lasso		0.3416	0.4556	0.3232	0.3103	0.2240	0.1136	88	0.8799
Validation	Fit Generalized Adaptive Elastic Net		0.3416	0.4556	0.3232	0.3102	0.2241	0.1136	88	0.8799
Test	Fit Nominal Logistic		0.3307	0.4760	0.4224	0.3749	0.2355	0.2135	89	0.8856
Test	Fit Generalized Lasso		0.3237	0.4679	0.4269	0.3740	0.2512	0.2022	89	0.8747
Test	Fit Generalized Elastic Net		0.3253	0.4697	0.4258	0.3737	0.2522	0.2022	89	0.8753
Test	Fit Generalized Adaptive Lasso		0.3737	0.5245	0.3953	0.3588	0.2601	0.1910	89	0.8960
Test	Fit Generalized Adaptive Elastic Net		0.3737	0.5246	0.3953	0.3588	0.2602	0.1910	89	0.8960

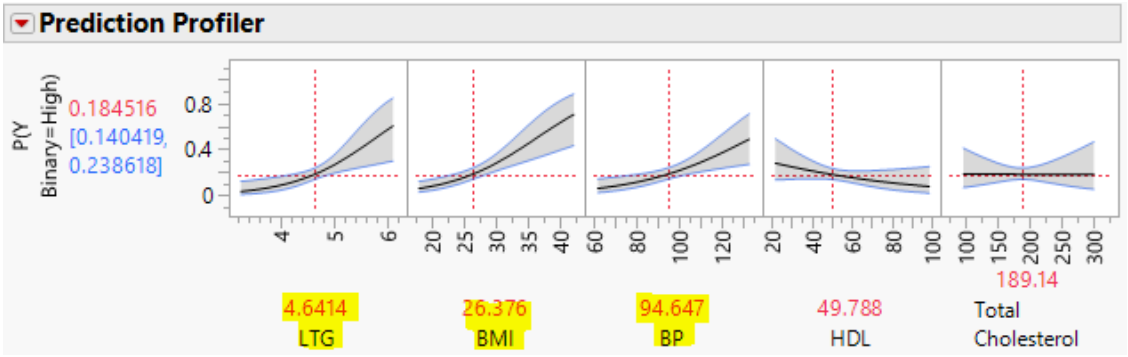
Interpretation

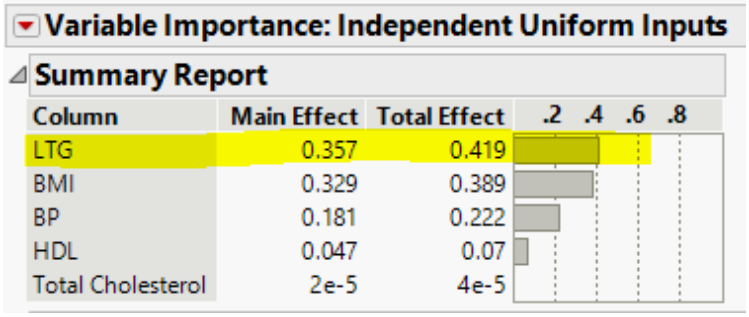
Per below, we can see that five predictor variables were used to maximize our overall model predictions. Uninformative variable selection was used through penalized regression to drop unimportant variables which are denoted by zeroes. The parameter estimates used in our model were BMI, BP, Total Cholesterol, HDL and LTG. We can however see that only the three highlighted predictor variables have statistical significance with p-values less than 5%. We will focus on these three variables of BMI, BP and LTG to further investigate which is best correlated to diabetes progression predictions. With BMI (0.1477), BP (.0374) and LTG (1.296) having positive estimates, we can predict that as each of these variables increases for a patient, his or her diabetes disease has gotten worse "high". This correlation makes complete sense as increasing values in these variables seems unhealthy, thus worsening the diabetes diagnoses.

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-13.94994	2.4453554	32.543217	<.0001*	-18.74275	-9.157128
Age	0	0	0	1.0000	0	0
Gender[1-2]	0	0	0	1.0000	0	0
BMI	0.1477019	0.0372809	15.696354	<.0001*	0.0746326	0.2207712
BP	0.0373983	0.0126585	8.7285363	0.0031*	0.0125882	0.0622084
Total Cholesterol	-6.935e-5	0.005966	0.0001351	0.9907	-0.011762	0.0116238
LDL	0	0	0	1.0000	0	0
HDL	-0.019533	0.0143918	1.8420033	0.1747	-0.04774	0.0086748
TCH	0	0	0	1.0000	0	0
LTG	1.2957469	0.4489226	8.3310059	0.0039*	0.4158748	2.175619
Glucose	0	0	0	1.0000	0	0

When looking at our three statistically important variable profilers ranked by importance, we notice LTG being our most import and the other two previously discussed variables of BMI and BP also having a relatively strong total effect for our model while the other two variables have minimal significance. Logarithm of triglyceride levels (LTG) is our strongest profiler and variable. Triglyceride is a type of fat found in our blood from consumption of calories. When we don't burn off those calories, they cause a spike in our triglyceride which obviously is unhealthy and promotes weight gain that may also correlate to a high BMI. LTG makes a lot of sense here in being our strongest profiler when looking at a patient's diabetes worsening over time. Some consideration can obviously also be given to BMI with nearly 39% variation regarding its positive relationship. But again, LTG is our strongest predictor variable as its positive relationship explains 42% of variations in predicting the worsening of one's diabetes disease.





Finally, aligning with our model preference of Adaptive Elastic Net, we created a new case to investigate whether or not we could predict a new patient's probability of progression as high. We looked at a 47-year-old male (assuming Gender 1 = male) whose BMI is 45, BP is 109, Total Cholesterol is 237, LDL is 100.2, HDL is 70, TCH is 3, LTG is 5.2149 and their Glucose is

107. We came up with an accuracy prediction of 89.53% for this patient's diabetes to be high(worsening).

Probability(Y Binary=High) AElastic Net	Probability(Y Binary=Low) 4	Most Likely Y Binary 5
0.8953443711	0.1046556289	High