

Daniel Sanchez

BAN 525

Dr. Cetin Ciner

Module 3: Assignment 1 - Understanding Diabetes Progression

May 30<sup>th</sup>, 2021

### **Introduction**

This Memorial Day weekend will mark the unofficial start of Summer in the United States. Eager to celebrate, families are sure to flock to area beaches to enjoy some fun in the sun. To help beat the heat, beachside vendors will surely be offering a wide array of icy cold sweet treats and beverages. Lemonade, popsicles, ice cream, mixed drinks, just to name a few. These seasonal indulgences, however, hold a hidden risk for certain individuals. Indeed, for some, the threat of Diabetes is ever-present. This chronic condition affects many underlying bodily functions, and it can be particularly difficult to manage. To better understand the factors that contribute to Diabetes progression, we may be able to analyze physical characteristics of the afflicted.

The dataset for this analysis contains ten baseline measurements from a Diabetes study involving 442 participants. These measurements consist of age, gender, body mass index, average blood pressure, and six blood serum measurements. After one year, researchers measured the Diabetes progression of the study participants. The result of the disease progression is the response variable in our analysis. This variable of interest is nominal, with two factor levels ("High" for worsening, and "Low" for improving). We are hoping to determine the predictors of Diabetes worsening in patients.

The analysis in this report has been performed in JMP Pro 16, using Ordinary Logistic Regression and the following penalized logistic regression methods:

- 1.) Lasso
- 2.) Adaptive Lasso
- 3.) Elastic Net
- 4.) Adaptive Elastic Net

## Analysis and Model Comparison

Ordinary Logistic Regression: This is being used as a benchmark for the analysis. It is not expected to outperform the other models in the group. By retaining all of the data set variables in its results, this method tends to overfit models to the training set.

Lasso: Using the absolute value of a determined penalty, this method shrinks the variables that contain little information, so that those are not included in the final model. This method also tends to eliminate redundant variables that are highly correlated with others.

Adaptive Lasso: This method is similar to normal Lasso. However, it first takes into consideration the results of Ordinary Logistic Regression before applying its penalization factor to unimportant variables.

Elastic Net: This method applies both an absolute (Lasso Regression) and squared (Ridge Regression) penalization value to uninformative variables. While unimportant variables are eliminated in the resulting models, highly correlated variables are normally retained (unlike Lasso).






Adaptive Elastic Net: This method considers the results of Ordinary Logistic Regression before applying its penalized regression approach.

Since our Diabetes data set is cross-sectional in nature (as opposed to being time-series), the cross-validation for this analysis was performed using randomization. The data set has been split into three sections. JMP Pro 16's Make Validation Column option was selected, utilizing a random seed of 123. Implementing a 60-20-20 split, 265 rows were used to train the models, 88 rows were tied to validation, and 89 rows were set aside for testing. It is worth acknowledging that this generalized cross-validation did not account for any stratification of the data set. With classification problems, this approach may lead to an underrepresentation of minority factor levels.

Hoping to determine the predictors of worsening Diabetes progression, Y Binary was selected as the response variable, with a Target Level of High. All of the other variables in the data set were included as predictor variables, aside from the Y variable (from which Y Binary was derived). Using JMP Pro 16, this analysis was conducted with all five methods. As

predictions from each model were generated, they were saved and stored in new columns in the data set. The figure below shows the comparison of how all five models performed with the test data.

#### Model Comparison (Relative Performance on Test Data)

Creator		Entropy RSquare	Generalized RSquare	RASE	Misclassification Rate	AUC
Fit Nominal Logistic		0.3307	0.4760	0.3749	0.2135	0.8856
Fit Generalized Lasso		0.3237	0.4679	0.3740	0.2022	0.8747
Fit Generalized Adaptive Lasso		0.3737	0.5245	0.3588	0.1910	0.8960
Fit Generalized Elastic Net		0.3253	0.4697	0.3737	0.2022	0.8753
Fit Generalized <b>Adaptive Elastic Net</b>		0.3737	0.5246	0.3588	<b>0.1910</b>	<b>0.8960</b>

The ideal model in this comparison has the lowest Misclassification Rate and highest AUC (Area Under the ROC Curve). As the name implies, the Misclassification Rate is a measurement of how often the model provides an incorrect classification, either predicting a False Positive or False Negative. The AUC represents overall success of the model, by plotting Sensitivity with 1-minus Specificity, and by measuring the area under the curve of this relationship across a 0-to-1 range of probability thresholds. Adaptive Lasso and Adaptive Elastic Net equally outperformed the other models, in terms of Misclassification Rate and AUC. We will review and interpret the results of the Adaptive Elastic Net model for the rest of this analysis.

#### Interpretation

The Adaptive Elastic Net model shows 3 out of 10 variables having a significant impact on the worsening progression of Diabetes in study participants. Ordered in terms of importance, these variables are LTG (Triglyceride Levels), BMI (Body-Mass Index), and BP (Blood Pressure). Each of these features has a positive correlation with worsening Diabetes. The coefficients of the selected model effects were 1.2957469 for LTG, 0.1477019 for BMI, and 0.0373983 for BP. Although other features in the data set (Total Cholesterol and HDL) have negative correlation coefficients, these relationships did not carry a significant p-value. Still, their inclusion in the model shows that they have more impact than the variables that were excluded.

Adaptive Elastic Net indeed reduced the coefficients of Age, Gender, LDL, TCH, and Glucose to zero. This was executed by applying a Lambda Penalty of 8.3091001. It is

particularly interesting that glucose levels were not made part of the final model. This points to the chronic aspect of Diabetes as an internal, systemic affliction. The Parameter Estimates for the data set are reflected in the table below:

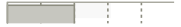

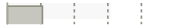
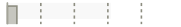
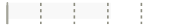
**Parameter Estimates - Adaptive Elastic Net**

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

This table also shows that Age and Gender do not appear to have an effect on the worsening of Diabetes. A missing component in this analysis, along similar demographic lines, is race. It may be worth collecting this information in future studies, since Diabetes has been shown to disproportionately affect minority populations. In terms of functional treatment, however, this analysis seems to indicate that Diabetes patients should aim to reduce these measurements: Triglyceride Levels, Body-Mass Index, and Blood Pressure. In addition, ensuring that a patient's Total Cholesterol incorporates more High-Density Lipoproteins appears to be potentially beneficial.

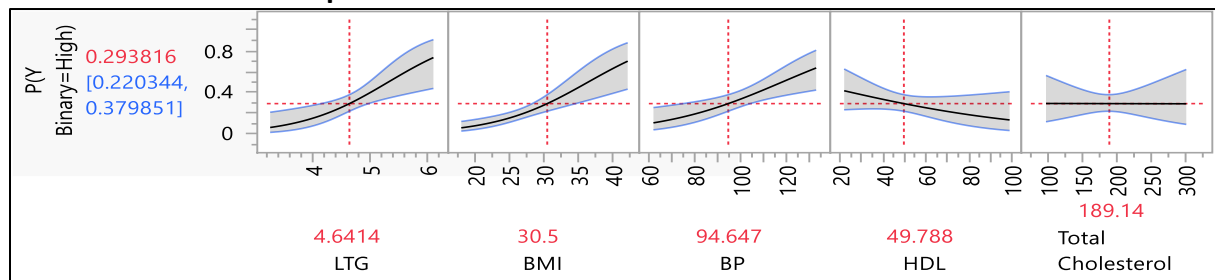
Through variable importance analysis, we can see that LTG and BMI account for most of the worsening Diabetes progression in patients (40.9% and 39.1%, respectively). BP comes next, having a total effect of 21.3%. HDL has a minimal impact of 6.6%, and Total Cholesterol's total effect is very small.

**Variable Importance: Independent Uniform Inputs**

Column	Main Effect	Total Effect	
LTG	0.357	0.409	
BMI	0.341	0.391	
BP	0.174	0.213	
HDL	0.047	0.066	
Total Cholesterol	4e-5	7e-5	

The prediction profiler for the Adaptive Elastic Net model is shown below, featuring a visual representation of the correlation coefficients of the model. The similar effects of LTG and BMI can be seen here, with their positive correlations with the response variable. Making an adjustment to LTG, BMI, or BP appears to have a substantial impact on the other model variables. Modifying HDL and/or Total Cholesterol does not seem to have as large of an impact on the other selected features.

#### Prediction Profiler - Adaptive Elastic Net



The finalized Adaptive Elastic Net model provides the opportunity to predict a probability of Diabetes worsening in a patient with a series of biometrics. An example set of biometrics is listed below:

Age	Gender	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose
47	1	45	109	237	100.2	70	3	5.2149	107

The final model predicts that this individual would have a .895 probability of their Diabetes worsening. This result was found by adding a row to the data set in JMP Pro 16, and having the program calculate a value in the Adaptive Elastic Net prediction formula column. Interestingly, the same result is achieved by simply adjusting values in the Prediction Profiler for the selected model features. This example is presented below:

#### Prediction Profiler - Adaptive Elastic Net (Example Biometrics Series)

