Determining an Ideal Model for Predicting Diabetes Amongst the Pima Indian Heritage Group of Arizona

Authors: Noah Lindley (nwl268), Chandershekhar Shori (cs56788)

Introduction:

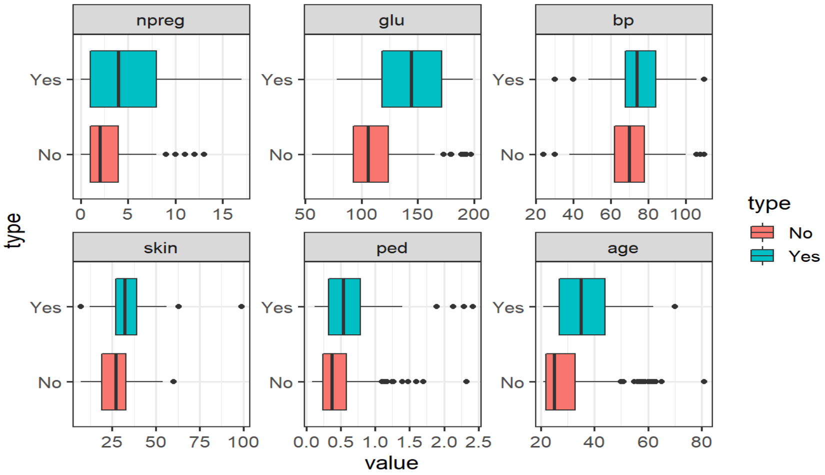
This investigation will be studying diabetes amongst individuals belonging to the Pima Indian Heritage group living near Phoenix, Arizona. The reason this study will be significant is because previous studies have shown that minority groups typically tend to have worse health outcomes than other groups with illnesses such as diabetes, but whether these outcomes are a result of health maladies or social factors remains unknown (Riley, 2012.; Schulz *et al.*, 2015) . Thus, our question that we will be investigating will involve comparing members of the Pima group that have diabetes to members that do not in order to determine if the presence of diabetes is correlated with differences in certain health measures in the Pima group or if no differences in health measures are found amongst diabetic and non-diabetic Pima group members and produce a model to help in diabetes diagnosis. Doing so, we will also identify specific health measures that differ in our dataset. Note that if there is no significant difference found amongst diabetic and non-diabetic Pima group members, it is suggested that the presence of diabetes does not have a significant effect on health measures for the Pima group. From this we can state our null hypothesis to be that there is no difference amongst diabetes patients and normal individuals in the Pima group in Phoenix, Arizona.

Now we will introduce our dataset, which is found in the "MASS" library. We will construct our dataset by conducting a row bind on Pima.tr and Pima.te, which are predefined training and test sets respectively in the "MASS" library. This will give us 532 individuals of the Pima-group we will be investigating for this study. The variables we will be studying are the following with their dataset abbreviations in parentheses: number of pregnancies (npreg), glucose levels (glu), blood pressure (bp), triceps skin fold thickness (skin), diabetes pedigree function (ped), and age of the individual (age). Note that the diabetes pedigree function determines the influences of heredity through genetics on diabetes onset. The column called "type" will determine if the individuals have diabetes or not by denoting it as "Yes" or "No." Additionally, we will have the body mass index column (bmi) serve as a class to assist in classifying the individuals in the study.

Before proceeding further, we would like to note we do expect to see variations between diabetic and non-diabetic patients regarding glucose levels and blood pressure as these two factors are well-known to be a defining factor of this malady. However, this study will prove or disprove this expectation, while also bringing light to any other variations that may exist.

Preprocessing and Descriptives:

For the preprocessing of the data, we would like to begin by producing boxplots to compare the distribution of values amongst diabetic and non-diabetic patients for the following variables as shown below: number of pregnancies (npreg), glucose levels (glu), blood pressure (bp), triceps skin fold thickness (skin), diabetes pedigree function (ped), and age of the individual (age):



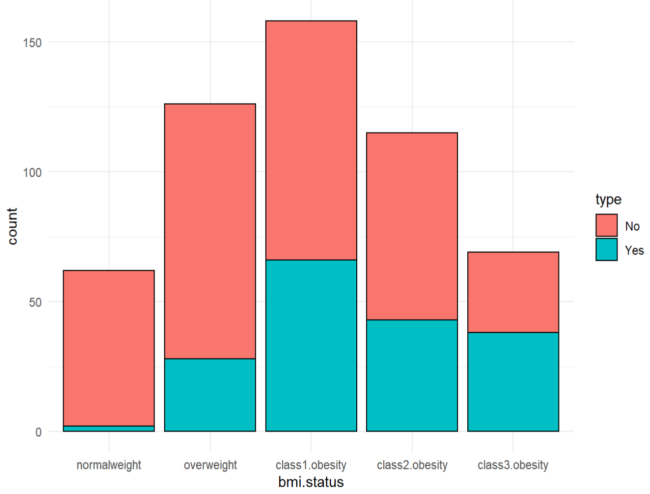
Observe that the distributions of all the variables for diabetic and non-diabetic patients do somewhat overlap to a degree, which implies that the distributions for the variables are somewhat similar amongst the two groups studies. It is worth noting that the distributions of the glucose levels (glu) variable, number of pregnancies (npreg) variable, and age of the individual (age) variable have distributions that overlap to a lesser degree in comparison to the other variables. This may imply that the glucose levels (glu) variable, number of pregnancies (npreg) variable, and age of the individual (age) variable vary amongst diabetic and non-diabetic patients. Also note that some outliers are seen in the box plot above in all of the distributions. These outliers are expected as the data we have is from human patients, and some anomalies do occur in the real-world situations.

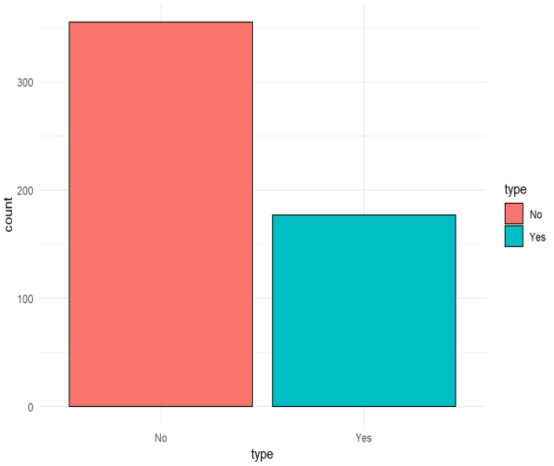
Next, let us look at a summary of all the predictors below for our reference to understand the distribution of the numerical values being studied:

Metric	**npreg**	**glu**	**bp**	**skin**	**ped**	**age**
Min.	0.000	56.0	24.00	7.00	0.0850	21.00
1st Qu.	1.000	99.0	64.00	22.00	0.2590	23.00
Median	2.000	115.5	72.00	29.00	0.4180	28.00
Mean	3.526	121.1	71.52	29.23	0.5040	31.65
3rd Qu.	5.000	141.8	80.00	36.00	0.6595	38.00
Max	17.000	199.0	110.00	99.00	2.4200	81.00

Observe that the range of values for the glucose levels (glu) variable is the largest amongst the variables, while the range of values for the diabetes pedigree function (ped)) variable is the smallest amongst the variables. We can also see that the age of the patients ranges from 21 to 81 years old while the average patient is about 31.65 years old.

Now we will move onto looking at a histogram of BMI classifications of the patients studied in the investigation, which is classified by the presence of diabetes. Note, before we begin, it is worth noting that we removed the underweight group in our investigation as there were only two individuals in this group. Also, note that BMI classifications are a class of categorical variables that can be view in the left graph below:





Notice that most of the individuals studied seem to fall under the class 1 obesity category, while the least number of individuals fall under the underweight category. Additionally, observe that the diabetes patients slightly tend to be skewed towards having obesity outcomes as more diabetic patients fall under the obesity category. On the other hand, the normal patients seem to be distributed more evenly throughout the histogram. Once again, this implies at first glance that diabetic patients have worse health outcomes regarding body mass index. Although this is an interesting observation, it can be attributed to the fact that we have more normal individuals in the dataset, which is demonstrated in the right histogram displayed above.

Methods:

Logistic Regression

We performed a logistic regression using the diabetes status as the response with all other variables as predictors. Our control group of reference in our model is a BMI status of normal. A logistic regression was used since our response variable is binary (diabetic or non-diabetic) and it will allow us to view how the presence of our predictors affects the probability of a patient being diabetic or non-diabetic.

Analyzing Logistic Regression

To interpret and analyze our logistic regression model, we used log-odds scaled coefficients, 10-fold cross validation, a confusion matrix, and an ROC curve to quantify how well we were fitting and predicting overall.

Regularization

We then used lasso regression and ridge regression on our current predictors to see if our model is overfitting on the current data. Overfitting is important to check for because it could cause the model to conform to just the training set meaning it would fail to generalize on cases in the testing set of data. Thus, allowing us to make more accurate predictions on diabetic status.

10-Fold Cross-Validation

Cross validation was used to compare the classification diagnostics such as the accuracy, sensitivity, specificity, positive predictive value, and the area under the curve between the full model, lasso model, and the ridge model. By making this comparison between the models we will be able to identify if our current model is overfitting or not.

Results:

By performing a logistic regression, we obtained the following coefficients for our model in terms of just odds:

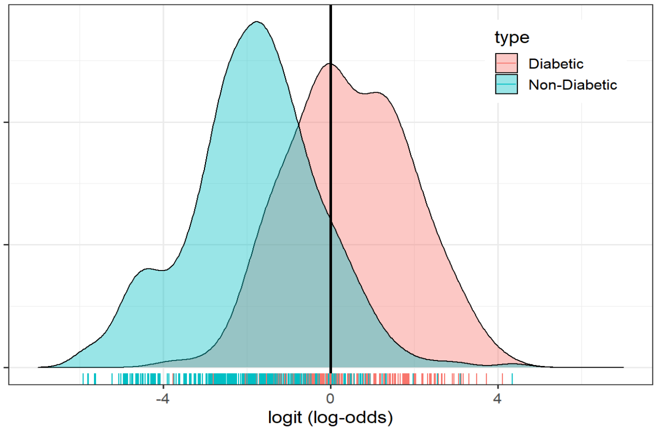
npreg	**glu**	**bp**	**skin**	**ped**	**age**	**overweight**	**class1.obes**	**class2.obes**	**class3.obes**
1.131	1.034	0.997	1.013	3.882	1.024	5.824	12.003	9.09	16.08

Observe in the coefficients above that most of our variables, besides BMI, result in an increased odds likelihood of an individual having been diagnosed with diabetes. This means that all of these health measures do increase an individual's chances, which confirms that variations in these health measures does have an effect on diabetes onset in the Pima group. Something interesting to note is that it seems that the odds of being diabetic decreased from class 1 obesity to class 2 obesity when we expected it to increase. Next, we produced a confusion matrix for our logistic model as shown below.

	Truth		
Pred	0	1	Sum
0	313	70	383
1	40	107	147
Sum	353	177	530

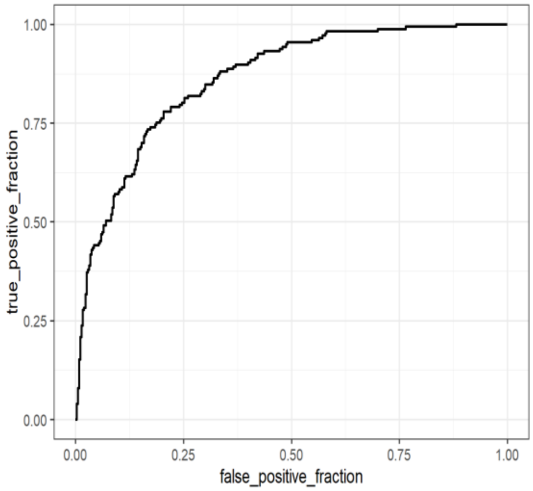
Observe that we have a true positive rate of about 60.45% and a true negative rate of 88.73%. Although the true negative rate is very high, this measure is not as beneficial as knowing the true positive rate. The reason being is that if a patient who is diabetic is diagnosed as not having diabetes, this can lead to detrimental outcomes for the patient due to delayed treatment. On the other hand, if a patient who is not diabetic, but is diagnosed as having diabetes will not have detrimental outcomes in the long run as they do not possess diabetes to begin with.

Density plot for the diabetic vs non-diabetic patients below:



Observe that based on our predicted probabilities, if we have a predicted probability above .5, we predict the status of an individual to be diabetic. On the other hand, if we have a probability below .5, we will then predict the individual's status to be non-diabetic. The overlapping region in the density plot above represents misclassifications. The right of 0, is the proportion of non-diabetic individuals that we predict to be diabetic (false positives). The left of 0, is the proportion of diabetic individuals that we predict to be non-diabetic (false negatives).

Looking at the ROC plot below and the AUC value that is attained below:



Observe that our AUC value for the ROC plot is about 86.3%. Since this AUC value is not great, but it is good. This means that our model does an overall good job in helping predict diabetic and non-diabetic patients based on the predictors used in our investigation.

Now, we are going to compare the full model to other logistic models generated from Lasso, and Ridge regression to see how well the models perform. We chose to compare our full model (uses all predictors) to a Lasso and Ridge model because they are both good methods to help avoid overfitting models and to reduce the feature selection so the model can be more interpretable.

Lasso regression selected predictors, 1 meaning selected & 0 meaning unselected:

npreg	glu	bp	skin	ped	age	overweight	class1.obes	class2.obes	class3.obes
1	1	0	1	1	1	0	0	0	1

Some pros to the Lasso variable selection are that it reduces the number of variables selected. In doing so it removes some variables that we believe to be important to our study which is the BMI status of some patients (overweight, class 1 obesity, class 2 obesity). Previous literature has shown that variables such as being overweight, class 1 obesity, class 2 obesity, and blood have been correlated with having diabetes in patients (Schulz *et al.*, 2015). Although the Lasso regression does remove these variables, these variables are essential in assisting to provide a complete picture of a patient that is needed for proper diabetes diagnosis.

Ridge regression selected only one predictor, that predictor was blood glucose levels. This would make our model very interpretable, but this removes all the other predictors from our study.

Now, the models are going to be compared to see which one performs the best via 10fold cross validation.

	acc	sens	spec	ppv	auc
full	0.7754717	0.5960919	0.8670542	0.7017305	0.8493765
lasso	0.7773585	0.5892664	0.8732238	0.7062496	0.8525276
ridge	0.7660377	0.5211559	0.8930983	0.7217623	0.7911217

Observe that all three models perform at very similar accuracy. It is worth noting that although the ridge model has a slightly lower AUC level in comparison to the models, the fact that it performs almost on par with the other models indicates that the sole variable of the ridge model, which is glucose, does play a significant role in assisting in diabetes diagnosis. Additionally, it is also worth noting that the Lasso model performs slightly better than the full model due higher accuracy, but the Lasso model removes variables that are known to be important in diabetes diagnosis as shown in previous literature (Schulz *et al.*, 2015). This possibly suggests that the training set used in this study is an anomaly such that the diabetes onset in the Pima group may not be influenced on variables traditionally thought to be important to diabetes. Rather, it is possible that certain unknown social and psychological factors are at play leading to diabetes onset.

Conclusion:

In this investigation, we found the glucose variable to be of importance in predicting diabetic status, we determined this because the ridge model performed almost as well as the Lasso and full model indicating glucose is a significant predictor of diabetic status. Observe that this contradicts our null hypothesis which states that there is no difference amongst diabetes patients and normal individuals in the Pima group in Phoenix, Arizona. This means we reject our null hypothesis as there do indeed exist variations in health measures in between diabetic and non-diabetic patients in the Pima group. Additionally, we found that the Lasso model and the full model perform very similarly, with the Lasso model performing slightly better. This implies that the Lasso models predictors are good predictors for the Pima group, but may not necessarily hold for larger populations as some essential variables are excluded that are found in the full model. Such as blood pressure (bp), which has been known throughout previous literature to be important in predicting diabetes status. In the case of the Pima sampling, it seems that blood pressure isn’t an essential variable to predict diabetic status.

In conclusion, we have found that within the Pima population glucose is a significant predictor of diabetes as expected, but blood pressure seems to not be a significant predictor within the sampling which contradicts previous literature findings (Schulz *et al.*, 2015). In future investigations, we hope to study other minority groups to help in generalizing the results along with studying underweight patients that were excluded in this study due to low sample size. Ultimately, the results of this study and future investigations will culminate into a model that can help in diagnosing diabetic patients to the best degree to help in preventive measures and improve healthcare outcomes.

Works Cited

Riley, Wayne J. “Health Disparities: Gaps in Access, Quality and Affordability of Medical Care.” *Transactions of the American Clinical and Climatological Association*, American Clinical and Climatological Association, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3540621/.

Schulz, Leslie O, and Lisa S Chaudhari. “High-Risk Populations: The Pimas of Arizona and Mexico.” *Current Obesity Reports*, U.S. National Library of Medicine, Mar. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/.