**Boosted Neural Network and Diabetes**

**Susan Wiggins**

**Cases in Business Analytics BAN 525**

**Professor: Dr. Cetin Ciner**



**Table of Content:**

Introduction…………………………………………………………………………………………………………………………………….3-4

Analysis and Model Comparison………………………………………………………………………………………………………4-6

Interpretation………………………………………………………………………………………………………………………………….6-8

References……………………………………………………………………………………………………………………………………….9

Appendix A **(Nominal Logistic Regression)**………………………………………………………………………………………10

Appendix B **(Neural Network Boosted with Squared Penalty)**………………………………………………………..11

Appendix C **(Neural Network** **Boosted with Absolute Penalty)**……………………………………………………….12

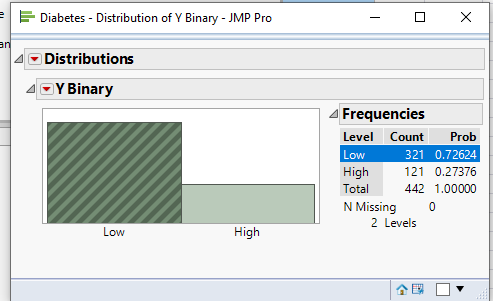
Appendix D **(Model Comparison )**……………………………………………………………………………………………………13-14

Appendix E **(Neural Network Final Model including Profilers and Variable Importance Summary)**..14-16

**Introduction**

Well who doesn’t like the occasional sugar rush of the four main food groups candy, ice cream, cake, and pie? I must say I know I do! However, even though this is not the prescribed main four groups that the population is advertised to eat they are an indulgence of many. Diabetes is a disease which your blood sugar is to high. The population needs to be mindful of the choices that we eat based on the factor that glucose (sugar) appears in the foods that we consume, and glucose provides the energy trigger to produce insulin. Insulin gives you energy. The reduction of insulin leads to the possibilities for diabetes to occur. In the United States alone Diabetes has been reported to exist in more than 100 million Americans. The Central for Disease Control and Prevention stated in a report in 2015 that, “30.3 million Americans an estimation of 9.4 percent of the U.S. population have diabetes. Another 84.1 million have prediabetes, a condition that if not treated often leads to type 2 diabetes within five years” (1). In addition, a listing on the web notes that, “the population should be aware that normal blood sugar levels are less than 100 mg/dL after eating (fasting) for at least eight hours. As well as a reported 140 mg/dL two hours after eating” (1). This case study will investigate the factors which contribute to diabetes. The dataset consists of 442 records in the study and variables of interest are age, gender, body mass index, average blood pressure, total cholesterol, LDL, HDL, TCH, LTG, and glucose. The dependent variable being measured in this study is Y Binary (low and high) a classification variable and the predicator candidates are age, gender, body mass index, average blood pressure, total cholesterol, LDL, HDL, TCH, LTG, and glucose. Three methods will be addressed in this case nominal logistic regression, Neural Network boosted with squared penalty, and Neural Network boosted with absolute penalty. The newest technique deployed in this study boosted neural network is recent to the machine learning methods. Dr. Cetin Ciner points out that, “Boosting is a learning process built sequentially and this technique evolves by learning from its previous mistakes. A predication is set forth based on the model being averaged and are not unstable” (6-7).

Logistic Regression is used in this case based on the categorical variable (Y Binary) being used to determine the outcome of the analysis. The distribution of Y Binary already informs the reader that there are a great many more of low than high response for diabetes as seen below.



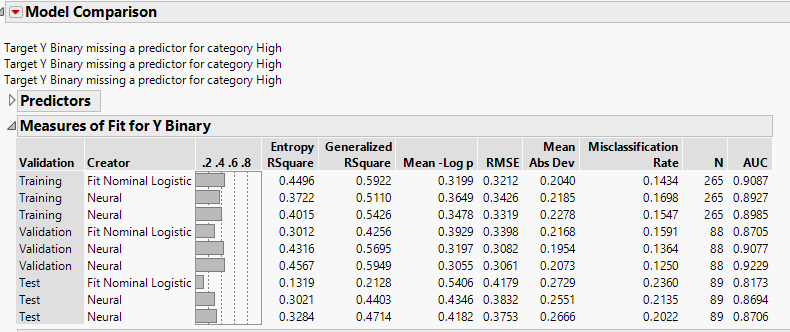
The probability of low is reported at 73 % whereas the probability of high only reports probability of 27%. Logistic Regression a power house for categorical data in essence will report the probability of an event through predictive factors used in the equation.

**Analysis and Model Comparison**

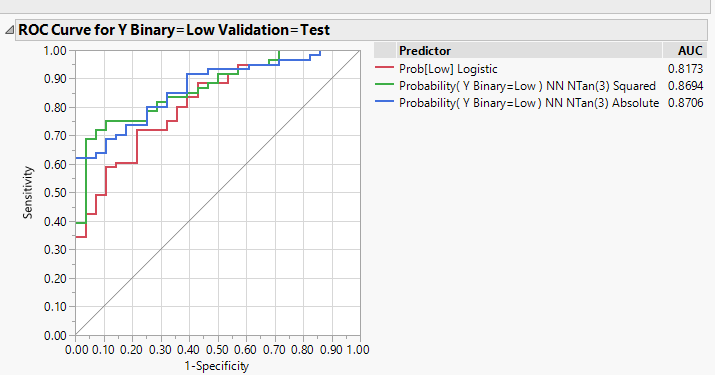
Ordinary Logistic Regression a modeling technique most highly used in prediction methods in categorical data can cause problems when Big data is being used. For instance, Ordinary Logistic Regression modeling can cause large variances in the data, models random noise, and poor forecasting can exists. Therefore, other machine learning techniques need to be implemented such as Neural Networks with Boosting. Neural Networks are highly complex modeling techniques and Neural Networks have ability to model very complex relations. However, Neural Networks tend to be difficult to understand when taking a first glance.

In order to reach the best results in the analysis one must depend on cross validation which will decrease the ability to interpret random noise. In cross validation, data is held out and the predication criterion is based on this interpretation. Estimates are then built on the data after the completion of each modeling affect occurs. In the case predication of Y binary (low), the cross validation is established on 60/20/20 split of the data with a random seed of 123. The final interpretation of the model will be based on the results of the test data due to the basis of new observations.

Next, the following modeling methods were performed nominal logistic regression, Neural Network boosted with squared penalty, and Neural Network boosted with absolute penalty. In both neural network assuming the default layer with one layer and three nodes with TanH as the activation function is deployed as well as the 40 models are used for boosting and conducted with 20 tours. The results of each one of these models are listed in the appendix A, B, and C. A model comparison was run for analysis of best results of modeling predications and chart is listed below.

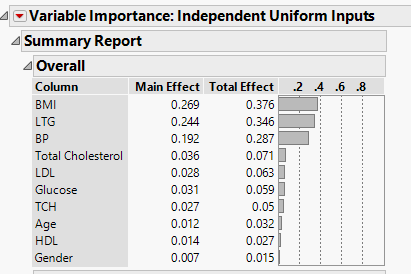


The model comparison revealed that Neural Network II with absolute penalty was the best modeling technique based on the lowest Misclassification Rate of .2022. In addition, the Area Under the Curve (AUC) reveals how well the data is being sorted in the modeling technique and an indication of a higher value towards the upper level of the curve at 1.0 informs the reader that the model sorting is better. The AUC for the Neural Network II with absolute penalty is the highest towards value in the reported model comparison at .8706. Therefore, the Neural Network II method accounts for 87.06 % of the testing data for Y Binary (low). The ROC curve below is further evidence of comparison for AUC values. The ROC curve is a graphical representation of true positive (sensitivity) verses false positive rate (specifity). In this case the sensitivity is the classifying correctness of Y Binary “Low”. Specifity deals with when the model classifies as Y Binary but when the Y Binary is “High” in fact. The perfect test for a ROC curve hugs the upper quadrant of the graph and the gray vertical line indicates that the modeling technique would be useless below this reference line. Truly, the Neural Network II Boosting with absolute penalty is the model comparison of choice for the case study of Diabetes.



**Interpretation**

The variable importance is included in the predication of Y Binary (Low) for Diabetes and the following table gives us the results below.



The total effect takes into account the variations in the data and the Body Mass Index (BMI) explains 38% of the variations in the data associated with diabetes “low”. The most important variable to look at is BMI. The second variable to study is concentration of lamorigine (LTG) listed at percentage rate of 35% followed by the third most important variable listed as Blood Pressure (BP) at 29%. Concentration of lamotrigine (LTG) is listed to be between the range of 1.5 to 10 g/mL. BMI and LTG have about the same impact on diabetes “low” around 30 percentiles. The model prediction profiler reveals that BMI, LTG, and BP has a negative relationship towards Y Binary “Low” meaning that as these variables increase in value there is a decrease in Y Binary “Low”. Total Cholesterol which is listed as the fourth variable of importance shows that there is a slight increase when studying marginal model association with Y binary “Low”. Therefore, this case reveals that in order for a patient to have lower chance of decreasing diabetic tendency that the individual should watch their BMI, LTG, and BP.

In conclusion, the selected model to predict Y Binary “Low” was Neural Network II Boosting with absolute penalty. The highest variable for fluctuations in Y Binary “Low” is related to the BMI at 38 percent and has a negative relationship. The profiler reveals that if the BMI increases to 35 there is an increased likelihood for the patient to have diabetes. The second most important variable associated with Y Binary “Low” is the LTG and has a negative relationship on Y Binary. The profiler determines as LTG increases for example 5.0 than a decrease for Y binary “Low” decreases whereby making a deduction for the patient to have a better likelihood to have diabetes also. The final variable studied in the Neural Network II is the Blood Pressure (BP) reported percentage at 29% and has a negative relationship in account that when the Y Binary “Low” decrease as BP decreases. Blood pressure weighs into the study based on increasing level of 130 would decrease the low response factor on Y binary “Low” enforcing the likelihood of diabetes within the patient. Therefore, one could associate through actively staying within the BMI parameters there is a lower chance of diabetes. In addition, when having the specified desire range of LTG concentration of lamotrigine one can also derive that there is a lower possibility for one to be diagnosed with diabetes. Blood Pressure is also a factor to limit the degree of diabetes in this case study. Overall, the variable which makes the greatest impact on Y binary “Low” is Body Mass Index (BMI). Also, AUC for modeling selection was the highest for Neural Network II absolute penalty at .8706. The Misclassification rate is also the lowest in the modeling comparison at .2022. The confusion matrix reveals that there is a rate of “low” of 50 times which is a fairly good rate. The predication for the model is good for the testing data at 50 times.

**Reference**

Ciner, Cetin. October 2019. Module 6. “Boosted Neural Networks and Fuel Efficiency.” Slides 1-25. Retrieved from: <https://uncw.instructure.com/courses/24252/pages/module-6-lecture-materials-boosted-neural-networks-and-fuel-efficiency?module_item_id=524522>

Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R. (2004). Least Angle Regression. Annals of Statistics (with discussion), 32, 407-499.

Nath Ras, Rabindra. “Diabetes and Obesity Determinants based on Blood Serum.” Endocrinology and Diabetes Research. 19 Jan. 2017. Web. 4 Oct. 2019. <https://www.scitechnol.com/peer-review/diabetes-and-obesity-determinants-based-on-blood-serum-CRno.php?article_id=6015>

“New CDC Report: More than 100 Million Americans have Diabetes or Prediabetes.” Center of Disease Control and Prevention. CDC Newsroom. 18 July 2017. Web. 3 Oct. 2019. <https://www.cdc.gov/media/releases/2017/p0718-diabetes-report.html>

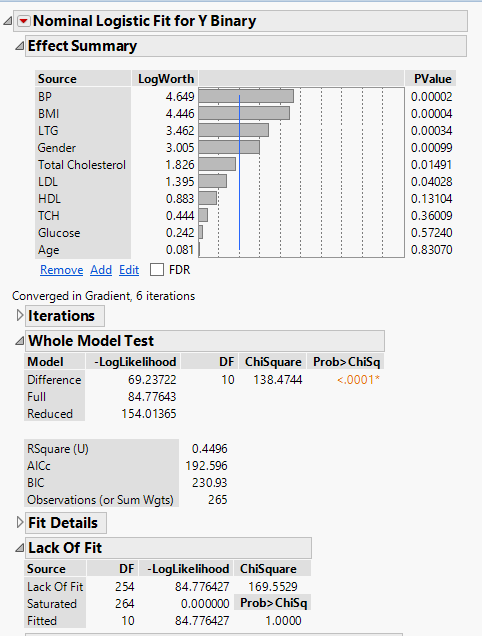
Website. Blood Sugar Levels. <https://www.google.com/search?source=hp&ei=AAqWXbP8Mo6KtQX3_7mABA&q=blood+sugar+normal&oq=blood+sugar+normal&gs_l=psy-ab.3..0l10.904.8212..8732...2.0..0.62.1067.20......0....1..gws-wiz.......0i131.ehfQYuaoWiM&ved=0ahUKEwiz_IzRqoDlAhUORa0KHfd_DkAQ4dUDCAg&uact=5>

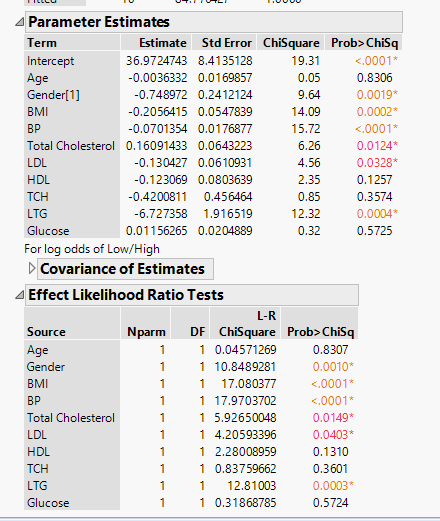
Website.Diabetes.<https://ww.google.com/search?source=hp&ei=uQiWXcrnNYiStQWQtZ9g&q=diaebetes&oq=diaebetes&gs_l=psy-ab.3..0i10l10.1968.5446..6162...0.0..0.152.824.8j2......0....1..gws-wiz.....6..0i362i308i154i357j0j0i131.e9C3pucF5M4&ved=0ahUKEwiKqJm1qYDlAhUISa0KHZDaBwwQ4dUDCAg&uact=5>

Website. lamorigine (LTG). <https://www.google.com/search?ei=JnOXXdFBhaT9BtDxmIgO&q=what+is+serum+concentration+of+lamorigine+%28LTG%29&oq=what+is+serum+concentration+of+lamorigine+%28LTG%29&gs_l=psy-ab.3...8345.9408..10433...0.4..0.91.649.8......0....1..gws-wiz.......0i71j33i10.lmd2q8oBlbk&ved=0ahUKEwjR7JqGg4PlAhUFUt8KHdA4BuEQ4dUDCAs&uact=5>

**Appendix A**

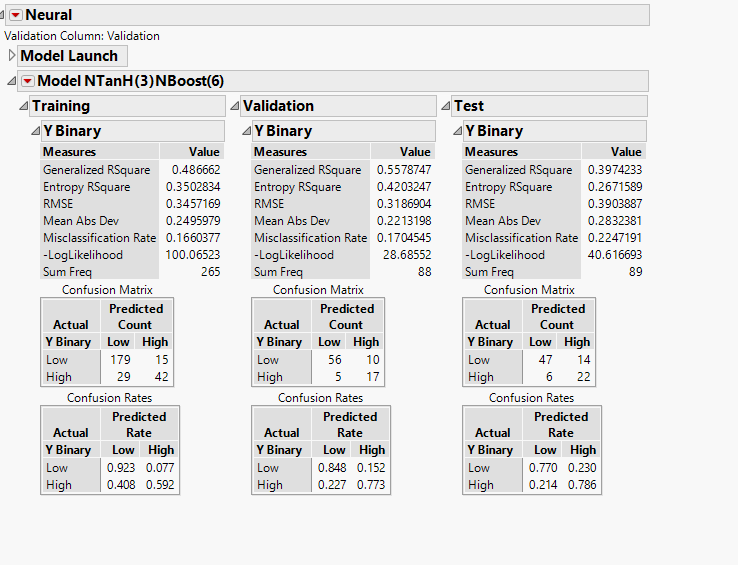
**Nominal Logistic Regression**





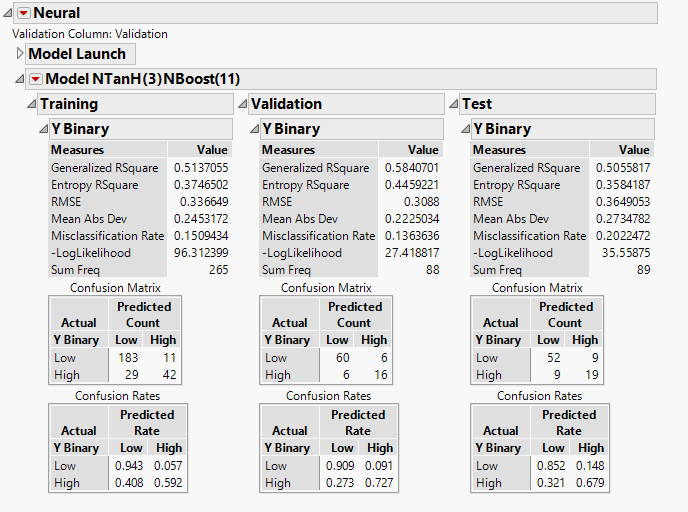
**Appendix B**

**Neural Network Boosted with Squared Penalty**



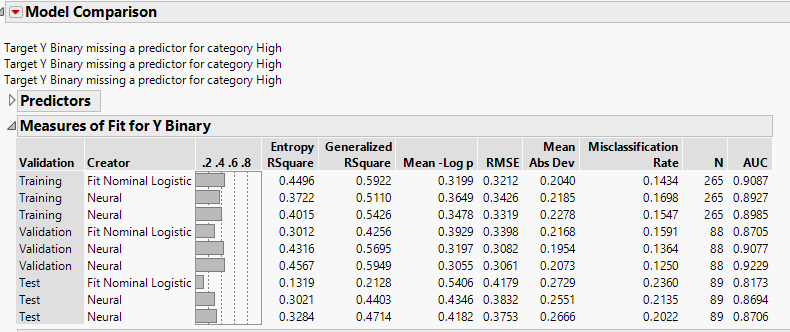
**Appendix C**

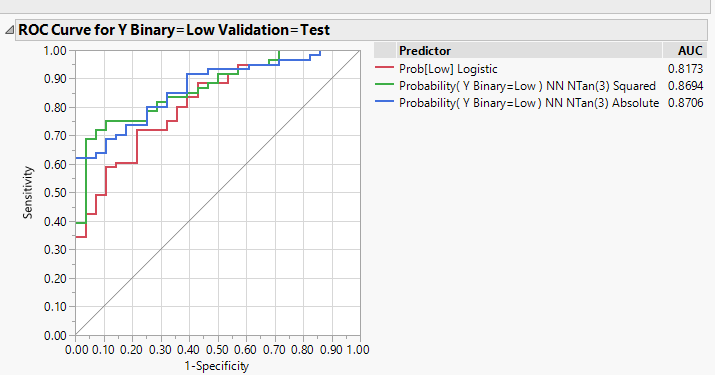
**Neural Network** **Boosted with Absolute Penalty**

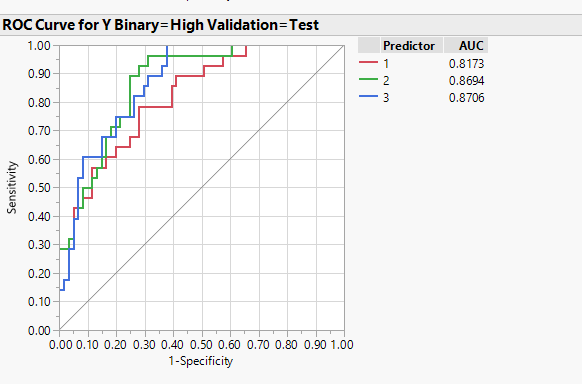


**Appendix D**

**Model Comparison**







**Appendix E**

**Neural Network Final Model including Profilers and Variable Importance Summary**

