DATA MINING APPLICATIONS  
  
  
ALY6040, WINTER 2020  
MODULE 3 PROJECT ASSIGNMENT

MODEL OPTIMIZATION

SUBMITTED BY: KEYUR SHAH, ANUPREETA MISHRA, SHIVANI ADSAR,

NIKHIL SAKINAL, SHRADDHA GOPALKRISHNAN, JAYETA BISWAS

NUID: 001089242, 001050752, 001399374, 001069354, 001376444, 001068448

SUBMITTED TO: PROF. JUSTIN GROSZ

DATE: 03/07/2020

**Introduction**

In this assignment, we work on categorizing the raw data that had been cleaned earlier in order to reduce our “Healthcare Stroke” dataset to more convenient chunks for processing. This alluring method is termed as feature engineering. We use feature extraction to reduce the elements of our faceted data containing more relevant information. The advantage here is it does not negatively affect our machine learning model’s performance. [1] Moreover, the process increases the accuracy of the predictive model. Further, stepwise regression is being used only to lessen the number of errors within our model. The analysis has helped us make better judgements of the factors responsible for patients to suffer a heart stroke.

**Analysis**

Here, we follow according to the pre-model steps for the analysis as seen below:

1. Created dummy variables for smoking\_status, work\_type and Residence\_type

using dummyVars function.

1. We then stored these values as a data.frame named “trsf”.
2. We have appended the created data frame with the original dataset of “Healthcare Stroke”.
3. We used the leaps and MASS libraries which helps in feature extraction and statistics for predicting the regressions.
4. Further we have divided the updated dataset into training and testing sets similar to how we did in Week 2.
5. We have used logistic regression on the training set for all the predictors.
6. Since we are using one of our selection criteria as Stepwise Selection, we have performed regressions using the Stepwise function “stepAIC” on our logistic model. Here we give direction as “both” which compares Forward and Backward Selection models and chooses the best out of the two.
7. We have performed prediction of the logistic model’s Testing set. In order to build our confusion matrix, we have used a threshold value of 0.02 to obtain a binary classification.

Last Week’s Update:

On the basis of our analysis we found that Age, heart disease, and glucose level are highly statistically significant factors which affect the stroke levels of patients. And the accuracy of the model was 75.72% which summarized the performance of the classification algorithm. After building the confusion matrix, the following results were obtained:

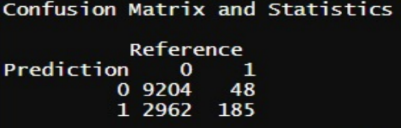


Figure 1: Previous Confusion Matrix

Observations: From the above table we can see that age, heart disease, and glucose level are highly statistically significant factors which affect the stroke levels of patients. However, even though hypertension is significant but relatively less dominant than the other factors mentioned above. The remaining factors do not significantly affect stroke as their z-score values are very small.

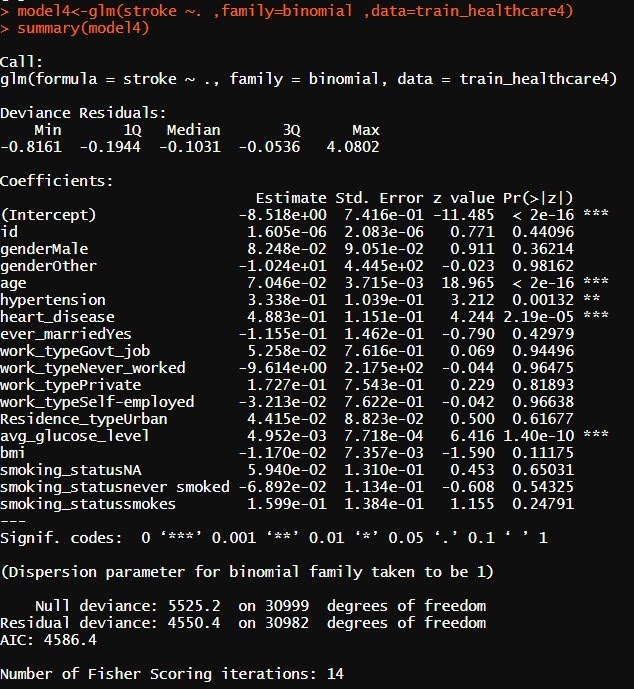
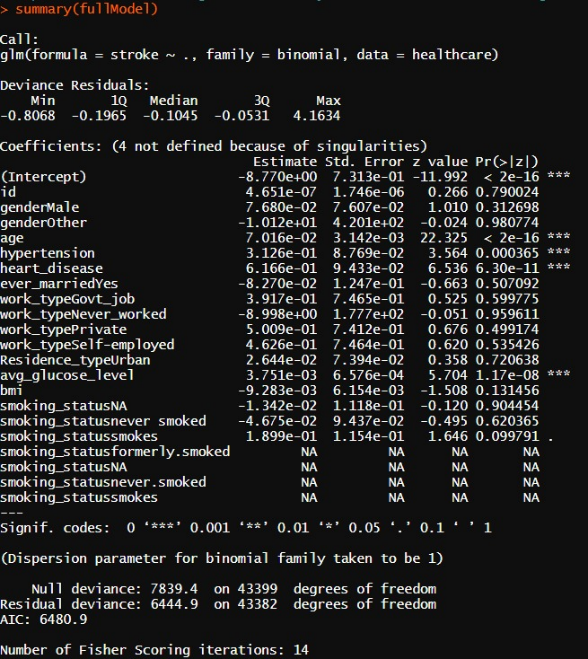


Figure 2: Summary of Week 2 Logistic Model

Comparison between the 2 Week’s Models-

Observations: As we can see, the results obtained from logistic regression and Confusion Matrix, show that age, hypertension, heart\_disease, avg\_glucose\_level are significant factors that contribute to a patient suffering a heart stroke.

Figure 3: Summary of Week 2 Model

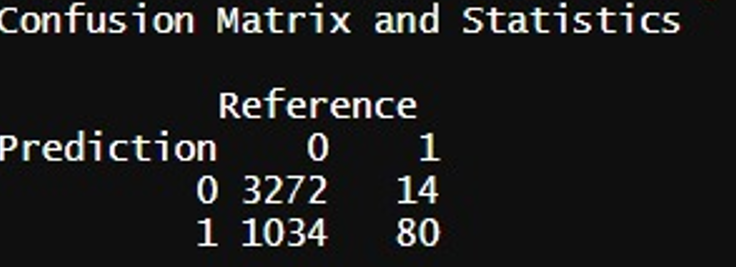


Figure 4: Updated Confusion Matrix

Observations: Post Stepwise regression, we can see that the parameters age, hypertension, heart\_disease, avg\_glucose\_level and smoking\_status for the people who smoked are statistically significant.

In contrast to the results of previous model, the number of iterations have reduced from 14 to 8. This shows us that Stepwise selection method has been effective in faster computations of prediction models.

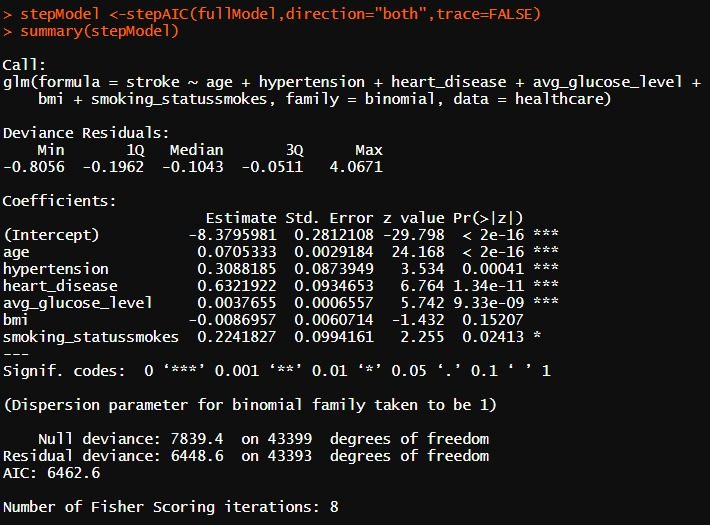


Figure 5: Code for Stepwise Regression

**Conclusion**

* Age, heart disease, average glucose level and smoking status for people who smoked are highly statistically significant factors which affect the stroke levels of patients.
* We understand that the Dummy variables allow us to perform easier categorization of variables and they act like 'switches' that turn various parameters on and off in an equation.
* Stepwise Regression is an advanced method for both forward and backward selections where the added variables are checked to note if the significance has reduced considerably to a specific tolerance level.

**References**

1. Feature Extraction. (2019, May 17). Retrieved from https://deepai.org/machine-learning-glossary-and-terms/feature-extraction
2. Eckerson, W. (n.d.). Secrets of Analytical Leaders. Retrieved from https://learning.oreilly.com/library/view/secrets-of-analytical/9781935504344/chap13.xhtml