**Northeastern University**

**College of Professional Studies**



**ALY-6040 Data Mining Applications**

**TOPIC:** **Model Optimization**

**TO: Prof. Justin Grosz**

**BY:** Snehal Keshav Bende

Khushboo Bharatkumar Oza

Shruti Sawale

Mohit Patel

Rishi Barbhaya

**Quarter 2- WINTER 2020**

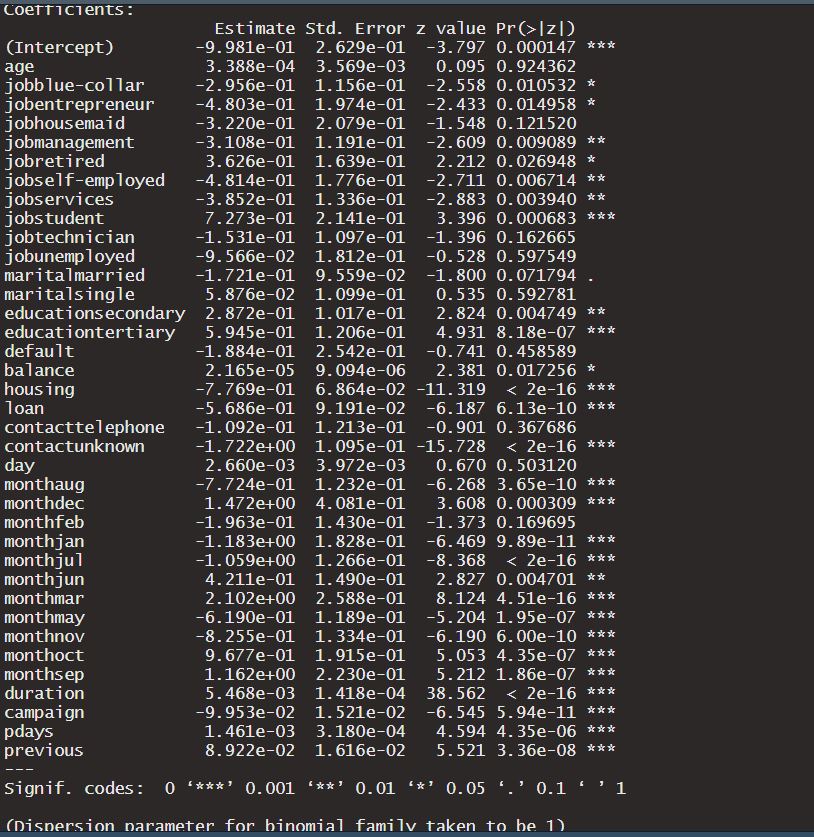
**Introduction:**

After cleaning the data set, we began with implementation of different models and finding the variables that are more significant and the one that are least significant. Moving ahead, its time for model optimization which requires recreation of a dummy variable, which are the one that take value of 0 or 1 in case of categorical data to show the presence or absence of the some categorical effect that might have some impact on the dependent variable.

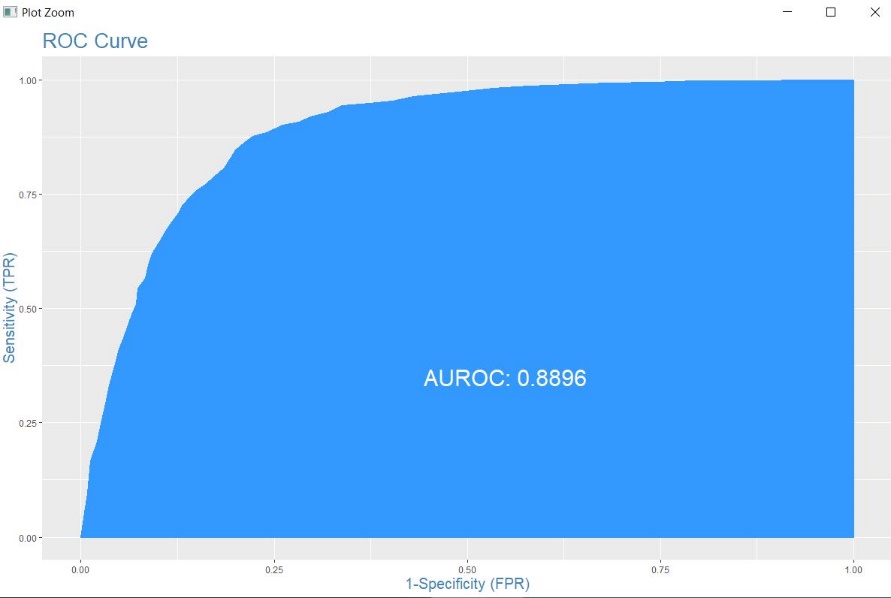
**Analysis:**

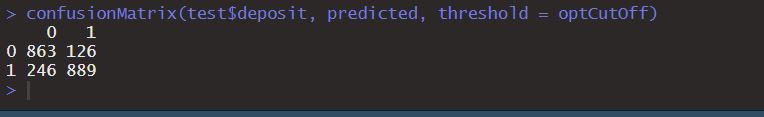
GLM is selected as the base model on which selection optimization model was applied. In case of forward selection, the model will keep on adding up the variables until there is a drop in optimization score where as backward selection considers all the variables at once and continues removing variables one after the other until the optimization score falls. We selected “both” as out selection method which as the term suggests is the combination of both.

Below chart shows the significance of each variable after the implementation of base model which included all the variables after cleaning the data set.

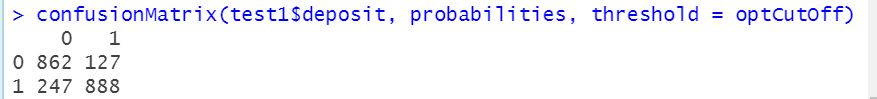


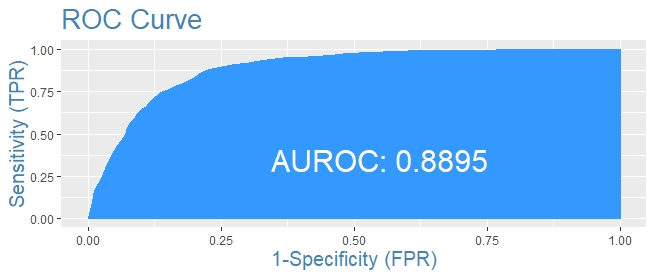
Next, ROC curve is implemented, where the area under the curve shows the forecast accuracy which turned out to be 88.96% in case of GLM and accuracy was observed to be 82.4% as per the confusion matrix and type 2 error is more than type 1 which means that the number of term deposit holders predicted was actually less than the ones that actually do have term deposit and need to be focused.



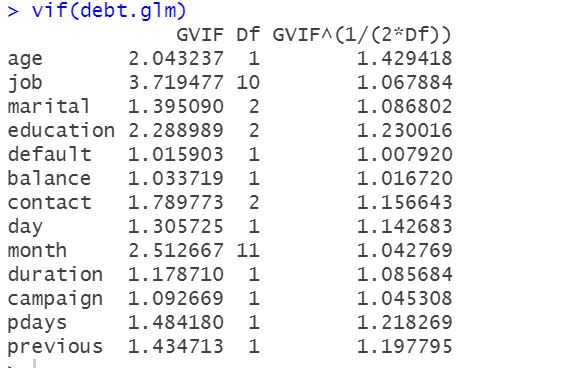


A new dummy variable is created which had information about customers holding any form of loan and is termed as “debt” which was derived by operating logical OR condition on loan and housing attributes. Now the base model was again implemeted on the new data consisting of dummy variable and stepwise selection was applied whose accuracy turned out to be same as earlier i.e. 82.4% and area under the ROC curve was 88.95%.

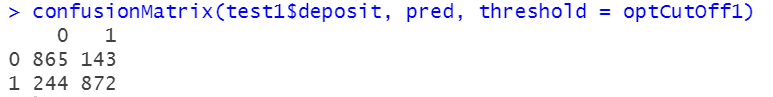


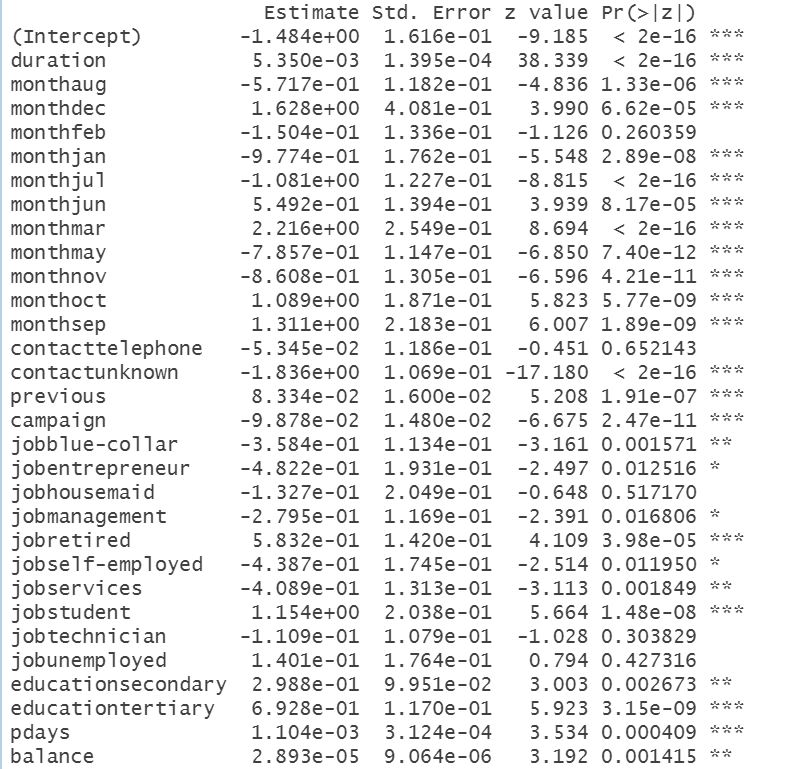


Variance inflation factor (VIF) is determines the correlation between the predictors in the model where higher the value, difficult it gets to access the attribute’s contribution towards the model which is also known as multicollinearity. In our case, all the attributes had values below 5 which categorize them under moderately correlated.

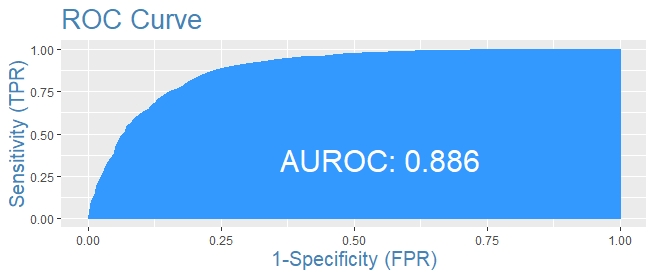


Bayesian Information Criteria (BIC) is used for comparing the model and lesser the value, better the model which was implement in our case having selction criteria as “both”. It was observed after its application which attributes were removed like days, marital, age and default which is shown below. The accuracy was observed to be 81.77% as per the confusion matrix below.

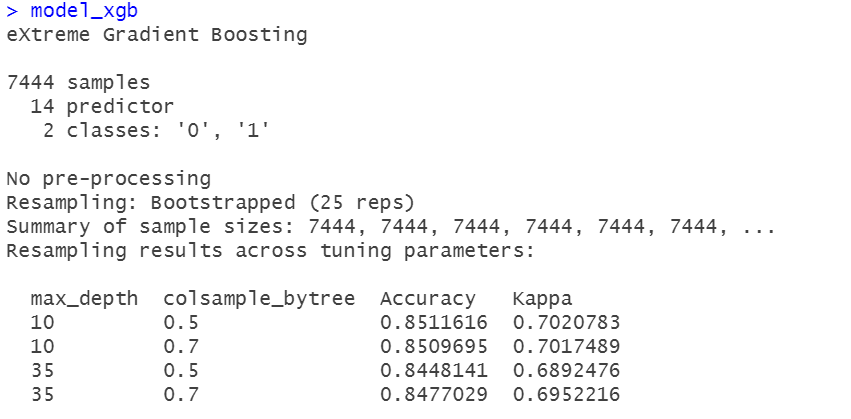




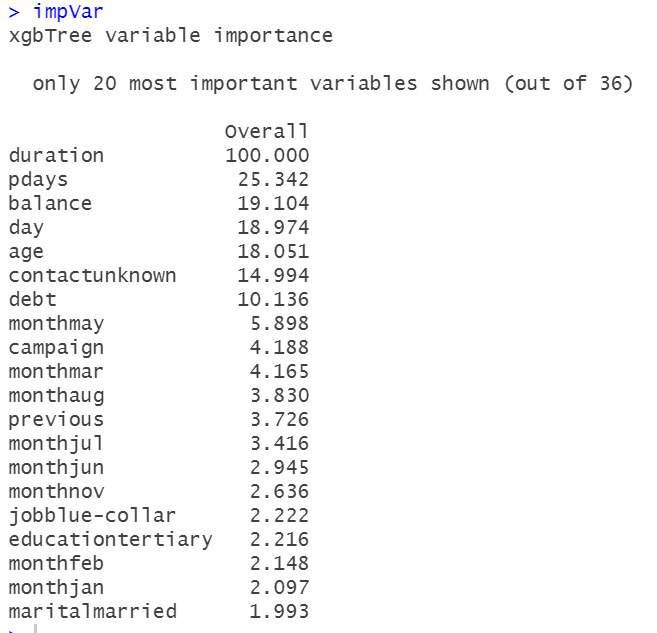
The area under the ROC curve was 88.6% after implementing BIC and stepwise. This model won’t be best for testing as on comparing, no significant changes were observed between base model and selection model.



Moving ahead with optimization with XGBoost, where accuracy was observed to be 78% with the base data set and it turned out to be 85.11% when maximum depth of the tree is 10 and sub sample ratio is 0.5. So, improvement of approximately 7% is observed. Hence, making it more reliable model.



The important variables according to XGBoost are as mentioned below.



**Conclusion:**

The number of contacts with any particular and the duration of the contact matters the most in case of influencing clients for term deposit. There are chances of people ignoring telephonic communication hence, other forms of contact are working out better than those to attract an individual for term deposit. Age of the customer and the amount of balance they have in the account are other factors that can determine their interest in term deposit.

****