**REPORT BY**

**NAME: KUSHAL SHAH**

**STUDENT ID:A20207420**

**EMAIL ID:kusshah@okstate.edu**

**MANAGER FRIEDNLY SECTION**

**Here we have built a decision tree model to actually predict the customers which are likely to switch to using their 4G network.**

**The data contain many missing values for different variables at different observation. We have fixed this issue by replacing those missing values with corresponding most likely value(mean for continuous and mode for categorical variable). for respective variable at particular observations(see** Figure 1 **for reference.)**

**Data used here is from valid sample of customers.**

**Given data set has been split into two section: one is training data(70% of observations) to train the model and other is validation data(other 30%) to check the validity of the model.(see** Figure 2 **for reference)**

**Classification of people has been done based upon conditional grouping of variables in accordance with the algorithm.(see** Figure 3 **for reference)**

**Our model has accuracy of 73.94% while predicting whether is going to change to 4G or not. (see** Figure 4 **for reference).**

**There are 21 rules each containing different conditions for different variables, based on which classification of people has been done. (see** Figure 5 **for reference).**

**Based upon grouping, we can characterize our model performance as excellent(see** Figure 6 **for reference.)**

**TECHNICAL SUMMARY:**

**A)MISSING DATA:**

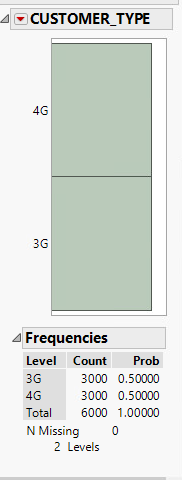
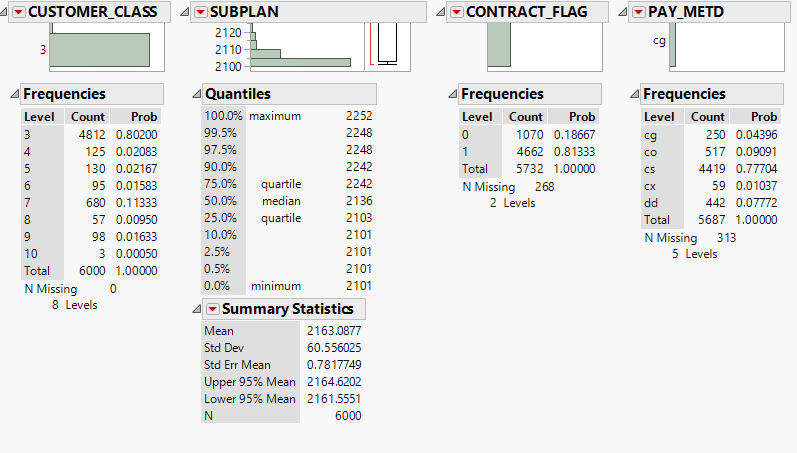
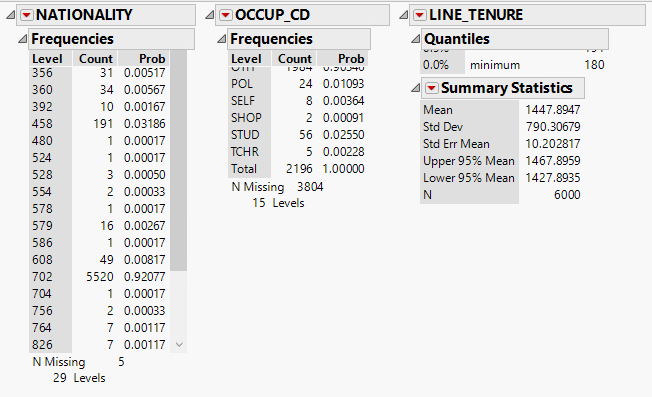
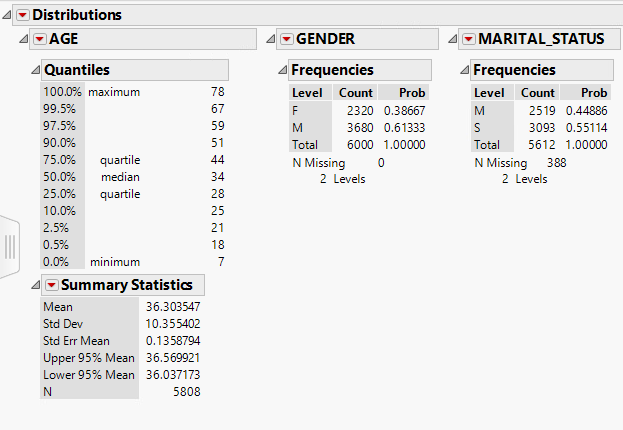
****

Figure 1

Here, if we see in above figures for Marital\_Status, nationality, Occup\_CD, Contract\_flag and PAY\_METD variables, there are 388, 5, 3804, 258, 313 missing values respectively. For other variables, we do not have any missing data.

For predictive modelling, we need to fix those missing data, otherwise we may lose very important information as these data points won’t be included into predictive Analysis. One way to address those issue is to replace those missing values by mean or mode for that particular variable.

Here, for Analysis, I have replaced each missing values by mean of variable when variable is continuous and by mode when variable is qualitative. By this approach We can add sense from missing data as well.

**B)Validity of 70% of training data set:**

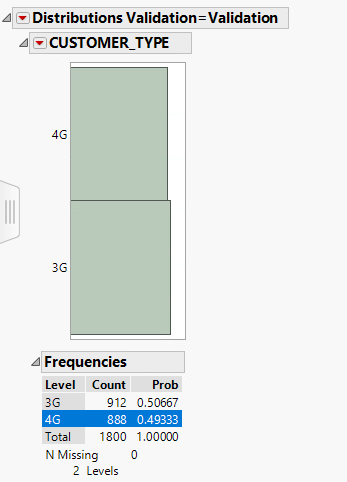
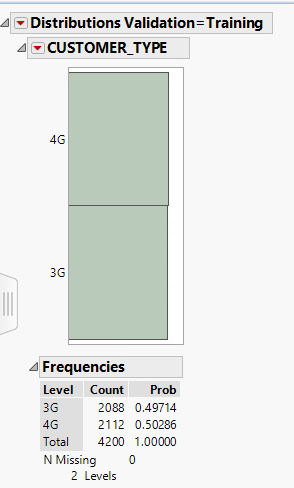
****

Figure 2

I have created a Validation column of training and validation data. Above figure shows distribution of Customer type by Validation. We can see that there are total 4200 values in training data set and 1800 values in validation data set. If we see 4200/(4200+1800)=0.7. That means 70% of total data has been into training data set and other 30% is into validation Data set.

Above figure shows count of 3G and 4G(binary variable) in each training and validation. % of count for both category in both data set almost looks similar(49.17% for 3G and 50.28 for 4G in training, 50.67% for 3G and 49.33% for 4G in validation data set). If we look at distribution for other variables (figures haven’t been shown because it tend to make report very large), we tend to have similar % distribution for each category when compared with respect to training and validation data set. This information ensures that information is from valid sample of their customer.

**C)CLASSIFICATION OF PEOPLE:**

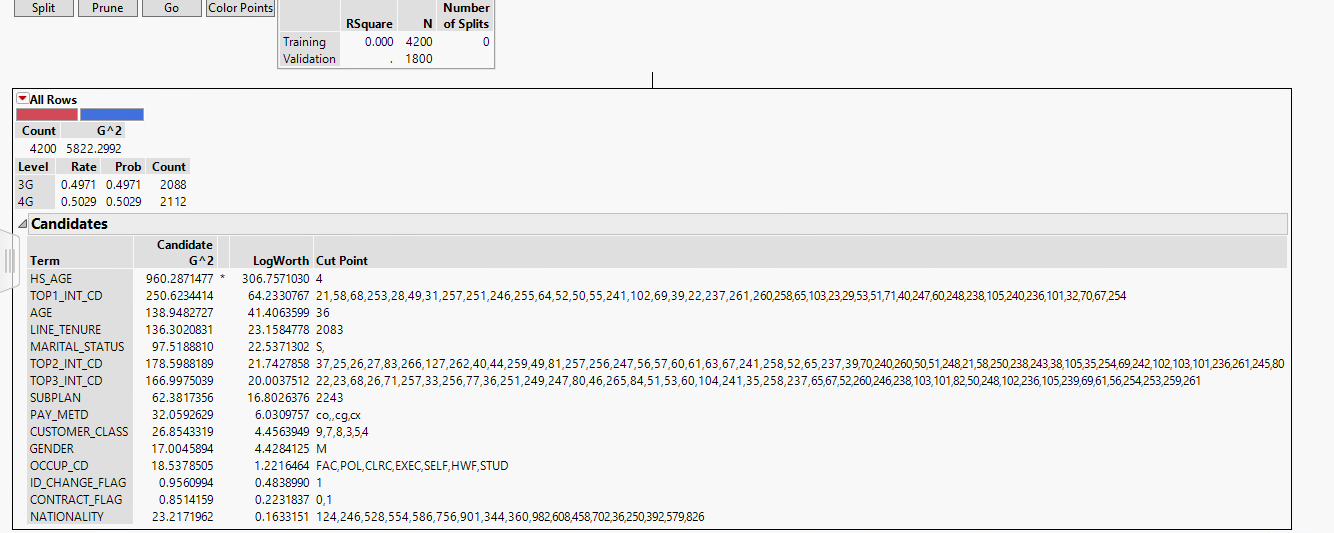


Figure 3

Here, to build a predictive model to decide about whether a person will change to 4G or not, we are using decision tree algorithm. As shown in count box, 4200 observations(70%) have been used to train and build a model. Here, Rate shows the probability in training data and Prob shows probability in validation data.

While building decision tree, classification of people has been done based on logworth of each term. We can see logworth and cut point of each term in the above figure. That means that the tree will classify people at cut points of each variable in sequence of log worth of each term. For an example, HS\_AGE has the highest logworth so people will be classified based on logworth at cut point of 4 in first split. And the split continues according to the log worth. This is how, classification has been done based on logworth of the term.

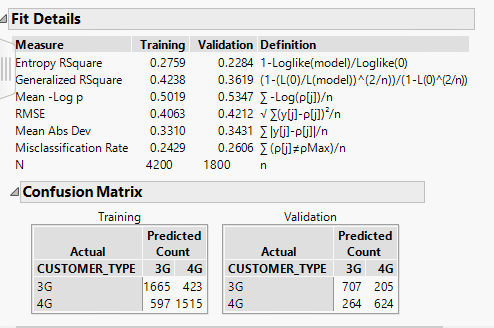


Figure 4

In the above figure, we can see that misclassification rate is 26.06%. Misclassification rate shows about how many observations have been predicted wrongly by model w.r.t actual outcome. Therefore (1-0.2606)\*100 shows us that how well this model can determine those who are really going to change to 4G and those who will not. Therefore we can say that 73.94% times our model determines accurately about those who are really going to change to 4G and those who will not.

Here if we denote 3G as 0 and 4G as 1, we can identify very important matrix similar to sensitivity and specificity.

%of correctly predicted 3G customer= [707/(707+205)]\*100=77.52

% of correctly predicted 4G customers=[624/(624+264)]\*100=70.27%

Those matrix reflects actual % of customers who will change to 4G or not.

**Rules for classification of people.**

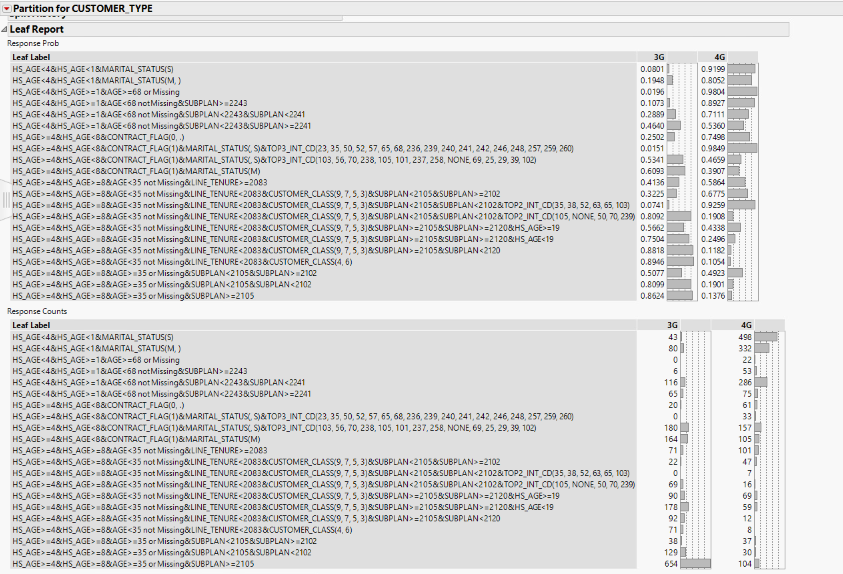
****

Figure 5

Above shown leaf report shows the rules for classification of people and sample size for each rule. To give an example, if we look at the first category of HS\_AGE<4 and HS\_AGE<1 and MARITAL\_STATUS=S, for this category there is 91.99% chance that person will belong to 4G category. In second table we can see total count(43+496=539) for this particular category. This give us the sense of how we can classify people , how to use this classification and how well those rules work.

In this way there are 21 rules where 14 variables has combined in each rule, according to which person can be classified with respect to target variable(customer type).

Second example: if we see last category in first table: category belong to group of HS-AGE>=4&HS\_AGE>=8&AGE>=35 OR MISSING&SUBPLAN>=2015 then we can say that the person belong to this category has probability of 86.24% that person will have 3G. Total sample size for this category is 625+104=729.

**Model Perfomance:**

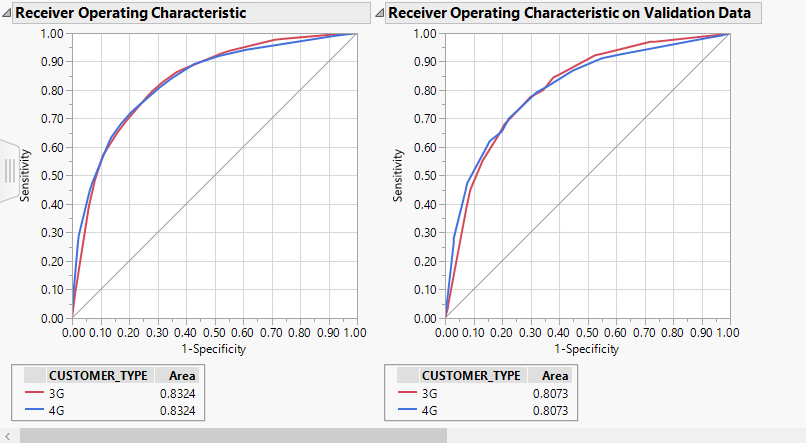


Figure 6

Above shown is the ROC curve(sensitivity vs 1-specificity) for Decision tree model. It poses true positive vs false positive. Straight line in diagram is by random model. Area under curve for random model is 0.5. As true positive is high and false positive is low, area under curve is high and model performance tends to be high as true positive gets high. Here we can see that area under curve is 0.8073. From this area under curve, we can show that the performance of this predictive model for validation data set is excellent(Area>0.8). This 0.8 value has been compared to 0.5(area for baseline model) to determine model performance( as acceptable, excellent or extraordinary).

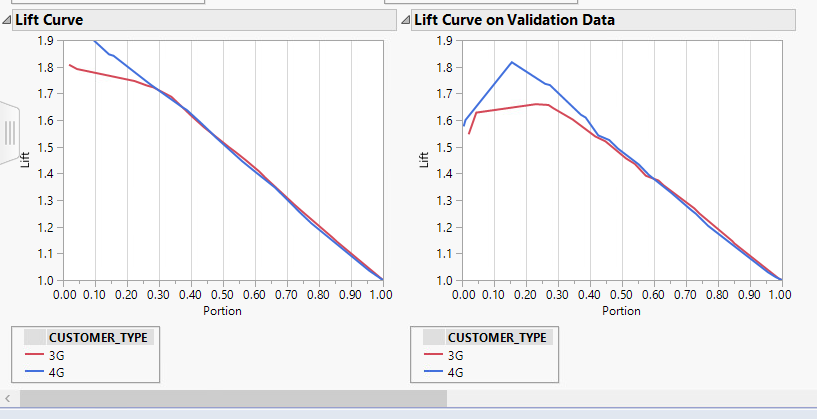


Figure 7

Above shown is the lift curve for decision tree model. Here lift of 1 is for random model. It depicts how better model predicts at given level then the overall probability. It actually displace events from non events. Horizontal axis shows portion of the data. Lift value at any given portion of data shows for any category shows how many times this model shows the value(3G or 4G) for that particular portion then baseline model.

E.g: in validation data, For portion=0.3, lift value is approximately 1.6 for 3G. that means that this model will predict 3G for 60% more than random. Higher the lift value, better is the model performance. We can see that model attains 1.85 of lift value for some particular portion which tends to show better performance of model on validation data set.

**Column Contribution:**

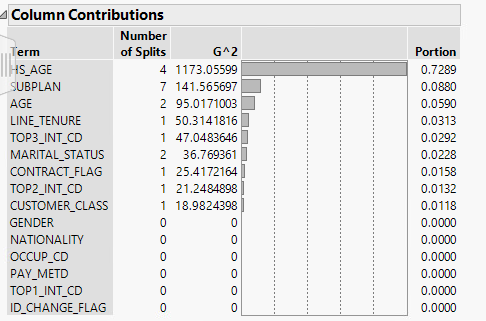
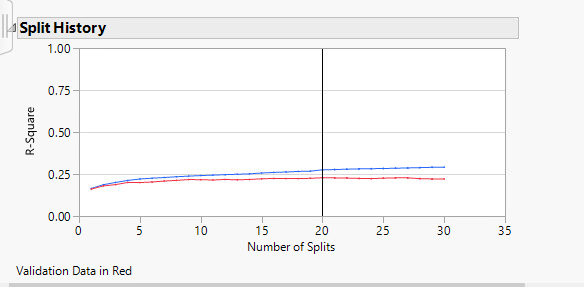


Figure 8

Above shown figure describes importance of every variable measured by G^2 value. Higher the value, higher the contribution of corresponding variable is. We can also check number of splits and

**Decision tree model:**

****

Above figure shows tend of r-square value for training and validation data. As we can see, r-square value increases as number of splits increases for training data set. For validation data set, R-square value first increases and then decreases with increase in number of split.

As R-square value attains the peak, we stop there chose that model as our final model.

This plot give us the sense of how R-square value changes for training and validation data set as complexity of model increases(increase in number of split.)