

**MGS 616–Predictive Analytics**

Assignment 2

Karan Ashok

**Questions – Conducting an Association Analysis**

c. What is the highest lift value for the resulting rules? Which rule has this value?

Display a screen shot of the Statistics Line Plot.

Highest Lift value for the resulting rules is: 3.60

Rule: Perfume =>ToothBrush and Toothbrush => Perfume is having this value.

It means the person who purchases perfume is 3.6 times more likely to purchase

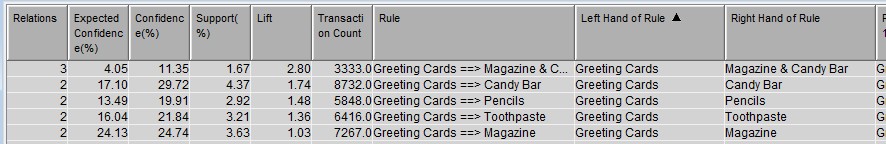
Toothbrush and vice-versa. Screenshot for the same is –



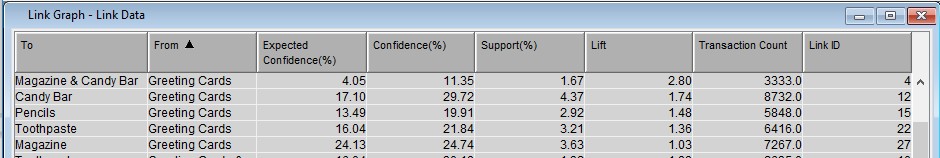
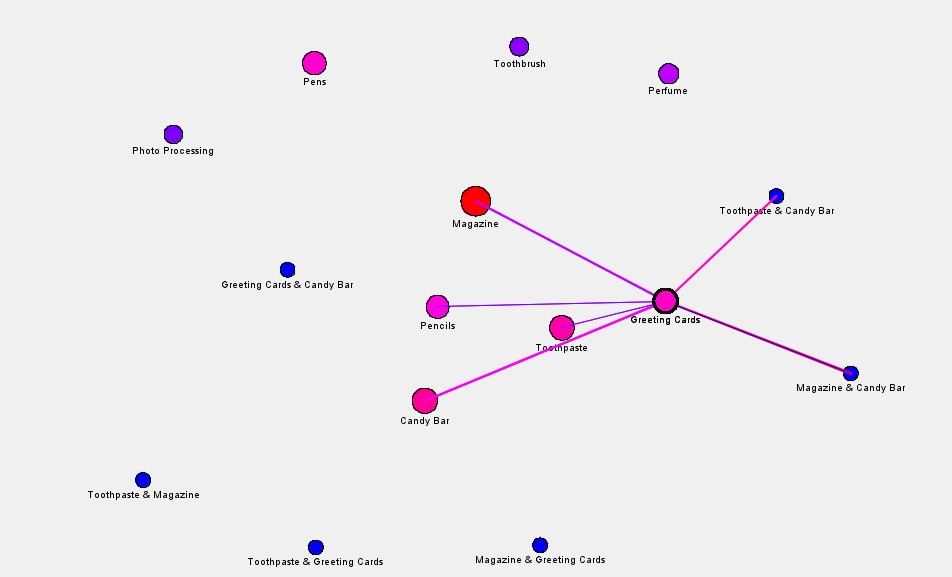
d. Based on the rules, what are the other products these individuals are most likely to purchase? Display a screen shot of the link graph showing only the links for greeting cards.

The other products individuals are likely to purchase when they purchase greeting card are

Magazine & Candy Bar, Candy Bar, Pencils, Toothpaste, Magazine



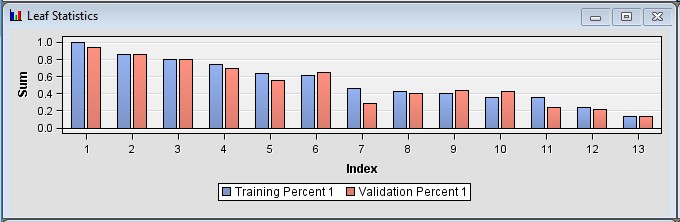
Link graph for the same is attached below –



**Questions – Predictive Modeling using decision trees**

c. How many leaves are in the tree?

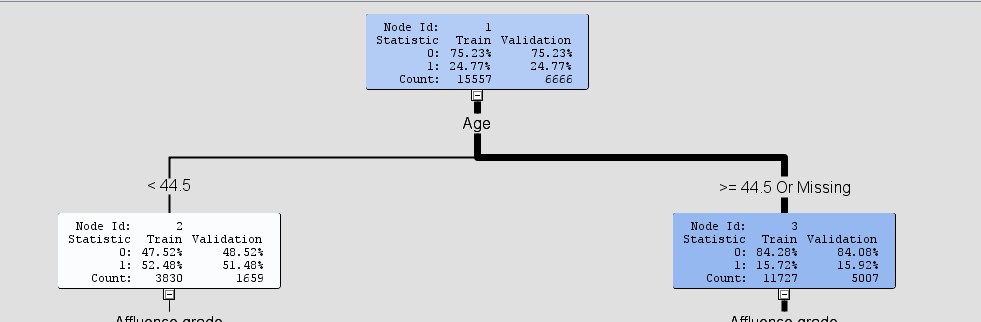
There are 13 leaves in the tree. Screenshot of Leaf statistics is attached below –



d. Which variable was used for the first split? Display a screenshot of the decision

tree.

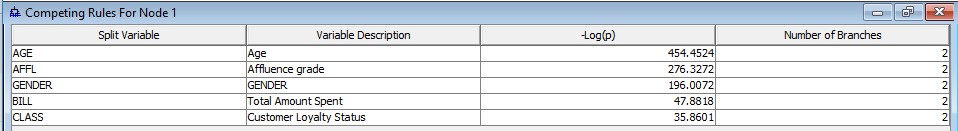
Variable ‘Age’ is used for the first split. Screenshot of decision tree is attached below –



e. In the decision tree, what were the competing splits for this first split? Display a screenshot of the Competing Rules for the first split.

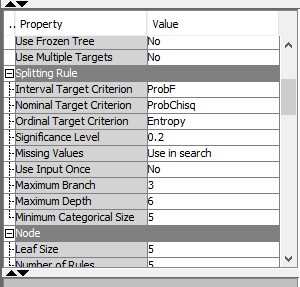
The competing splits for the first split were Age, Affluence Grade(AFFL), GENDER, BILL and

CLASS



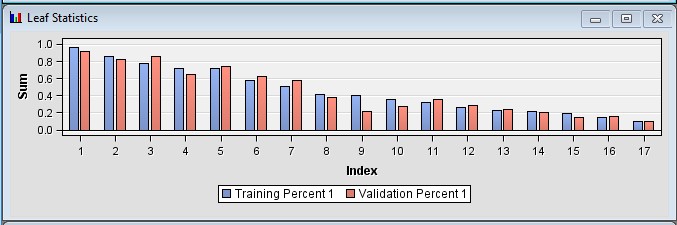
g. Display a screenshot of the properties panel for this step.

Properties panel screenshot for second decision tree –



h. How many leaves are in the tree?

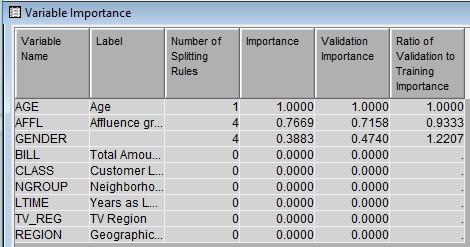
There are 17 leaves in second decision tree –



i. Which variables were important in this tree? Display a screenshot of the Variable

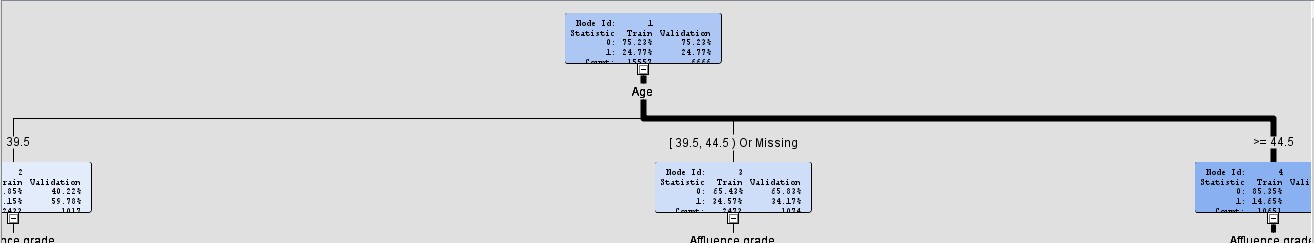
Importance section from the Results of the model.

Important variables in this tree are: AGE, AFFL and GENDER ( with positive importance). Other variables with zero importance are: BILL, CLASS, NGROUP, LTIME, TV\_REG, REGION Screenshot for the same -



j. Which variable was used for the first split?

Variable ‘Age’ is used for the first split.

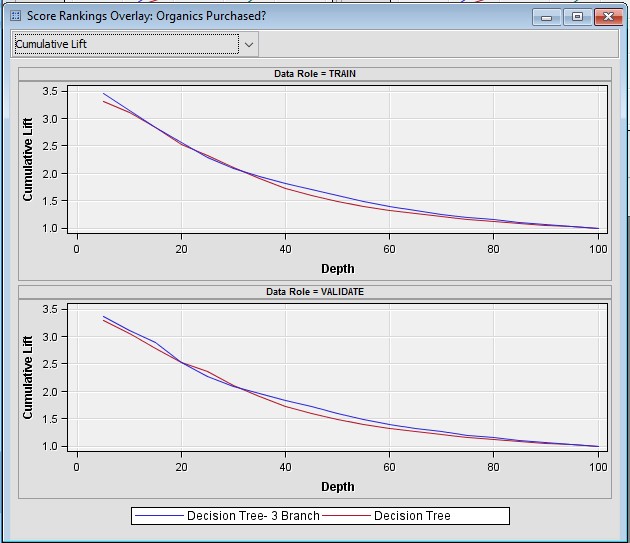


k. Which of the decision tree models appears to be better and how did you make that determination?

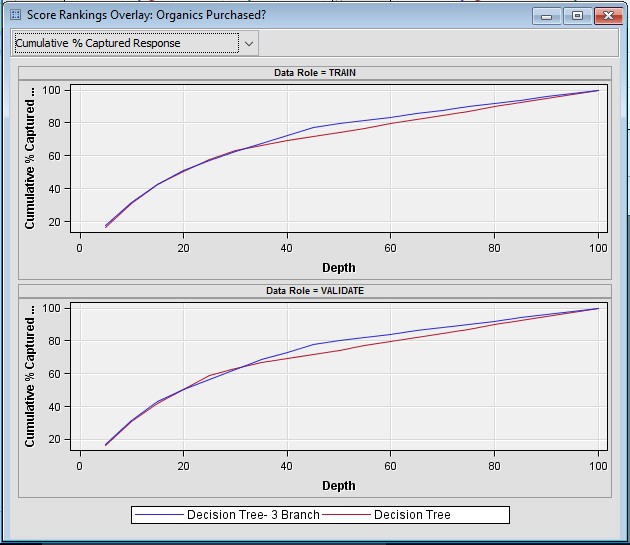
Second Decision tree model “Decision tree with maximum branch as 3” is better as compared to the default decision tree model because of the following reasons -

1. In cumulative lift chart, the cumulative lift for ‘Decision tree with maximum 3 branches’ is higher for lower percentiles till 20th percentile and then again from 40th percentile to 60th percentile.

At higher percentiles more than 70th percentile, the performance of two tree models is not appreciably different.



2. From Cumulative % captured response chart, we can check that decision with maximum branches as 3 is performing well because if 50% of data is selected, then this model has identified over 80% of the people who would have used ORGANICS while default decisionn tree has identified less than 80% of the people.



3. Also, if we check the fit statistics of comparision of both models, we can check that Decision tree with maximum of 3 branches is better because –

Misclassification rate is lower

Average squared error is less

Root average squared error is less

