Stanley Ngunyi Gathagu

Classification: Decision Trees

One of the questions that we have been focused on in this paper is the question regarding the prediction of possible movements between segments. In the previous section, we clustered all customers into three segments: Good, Better and Best with K-Means and K-Means PCA methods. The next step in our analysis to characterise the type of customers in each of these three groups and to investigate which customers are likely to move from one group to another. For instance, we aim to predict which customers have a large likelihood of moving from Good segment to Better, and so on. So, we need a model that will estimate these transitional likelihoods of each customer from one segment to another. For this purpose, we propose the use of decision tree.

We aim to cluster all customers into three groups as it has been done with K-means, but now we add demographic variables to the classification to analyse whether there are customers in one segment who are more similar to the customers of another segment than to their own (Kotler, 2010). In this way, we will be able to identify customers with the highest potential that are most likely to move from less favorite segment to more favorite one utilizing transition matrix containing transitional probabilities for each customer and each segment. Besides, using the same idea, we can identify the potential droppers who are likely to drop from a more favorite segment to less favorite one.

Haughton, (1993) has used decision tree for analysing customers buying behaviour. Moreover, Duchessi et.al., (2013) have used decision tree to profile the online and mobile technologies and services that ski resorts use for their promotional and advertising strategies for two important segments. Furthermore, Han et.al., (2012) have used decision tree method to extract important parameters related to long-term value, credit, and loyalty. They applied this model to telecom operators in China achieving high prediction accuracy. This unsupervised learning method is very popular because of its higher interpret-ability compared to other classification methods. Another advantage of decision tree is that it can handle easily both missing values in data and irrelevant attributes. Finally the method is very fast and gives compact results in the form of pruned tree. However, decision tree has also disadvantages such as instability of the results once a small change in the input data occurs. Finally, preparing decision tree, especially if they are large with many branches, can be complex and time-consuming challenge.

The idea behind decision tree is in line with the tree analogy where tree starts from the top, and the trees are drawn upside down, moving from the top to the internal nodes and consequently to the terminal nodes, also called leaves. The segments of the trees that connect the nodes are called

branches. The classification tree predicts that each observation belongs to the most commonly occurring class of all observations in the space to which it belongs. The process of building a Classification Tree can be described as follows:

Step 1: Construct the regions R_i,...,R_J such that total classification error E = 1-max$_j(\hat{p}_{mj})$, the fraction of observations in that region that do not belong to the most common class, is minimized:

$$\textbf{arg} \ \ \textbf{min}_{R_j} \ \sum_{j=1}^{J}(1 - \max_j(\hat{p}_{mj})) \tag{0.1}$$

where $\hat{p}_{mj}$ is the proportion of observations in mth that are from the jth class.

Step 2: Divide predictor space, the set of possible values for X_i i = 1,...,p; into R_j j = 1,...,J distinct and non-overlapping spaces(regions).

Step 3: Make the same prediction for each observation from the region R_j, where the prediction is simply the mean of corresponding values of the response variable.

The limitation of this algorithm for constructing the tree is that it is computationally infeasible to consider all possible partitions of feature space into J regions. Moreover, it has been found that classification error is not sufficiently sensitive for tree-growing. This task of growing a classification tree relies on the chosen method for splitting the tree and each of these measures, also called impurity measures, leads to different tree structures. Two widely-known impurity measures are Gini-index and Entropy. Gini-index is a measure of node purity where the small value indicates that the node contains mostly observations from a single class and it is defined as follows:

$$G = \sum_{j=1}^{J} \hat{p}_{mj}(1 - \hat{p}_{mj}) \tag{0.2}$$

Therefore, because of these insights, instead of considering all possible split combinations the recursive binary splitting technique can be used to optimally split the predictor space where each step is via two new nodes on the tree by using Gini-index and above algorithm can be extended as follows:

Step 4: Select the predictor X_j and cut-point s such that the splitting of predictor space into regions R_1(j,s) = {X|X_j<s},the region of predictor space in which X_j takes on a value less than s, and R_2(j,s) = {X|$X_j \geq$s} leads to largest possible information gain which is equivalent to minimizing Gini-index.

Step 5: Repeat Step 4 for finding best predictor j and best cutpoint s to split the tree further until the criterion is reached.

As mentioned earlier, Gini-index is a measure for impurity of the tree, that is when members of one class are in another class where they do not belong. Consequently, we aim to build a tree which has leaves that are as pure as possible and since lower values of Gini-index indicate more homogeneous and purer leaves we aim to minimize it in Step 4. The purpose of using decision tree is in this analysis twofold: building a pure tree and predicting the transitional probabilities of all customers per segment, we use both classification error rate E and Gini-index G for these purposes respectively. So, we use Gini-index for evaluating the quality of each split in the tree, that

is pruning process of the tree, and we use classification error rate in the predicting process applied on the pruned tree which leads to higher prediction accuracy. Coussement, (2014) introduced the connection between decision tree and RFM segmentation. We use the variables from RFM model; Recency, Frequency, and Monetary; as features for decision tree but we also add demographic variables such as Age-category, Gender, Allow-analysis, Opt-in and Loyal-time variables described in section 2.

Using the prediction results from the pruned tree, we calculate the probability's of each cluster in the final nodes, such that each customer has certain probability in being in one of the tree classes and these three probabilities sum up to one. These probabilities can be seen as the transition rates of the customers. We then use algorithm 1 to determine the new clusters of all customers based on the transition rates.

Algorithm 1 Assigning new classes to all customers

Input: transition rates $(p_{i,k})$ for customer $i = 1, ..., N$ and cluster $k = 1, ..., K$; and "real" cluster found by K-means algorithm $(\tilde{\kappa}_i)$
Output: new cluster $(\hat{k}_i)$ for customer $i = 1, ..., N$

- For each $i = 1, ..., N$
  - Get $max.position_i$ that contains all the positions that have $max(p_{i,k})$
  - If $max.position_i$ contains only one position c, then set $\hat{k}_i = c$
  - Else
    - If $max.position_i$ contain $\tilde{\kappa}_i$, then set $\hat{k}_i = \tilde{\kappa}_i$
    - Else, then set $\hat{k}_i = max.position_i$

The idea behind Algorithm 1 is that knowing the actual class labels from K-means we compare this with the class probabilities per customer and if the maximum probability, the probability which is the largest among three probabilities, of being in particular class differs from the actual class from K-means we define that particular class as the new class of that customer. For instance, if the customer has been classified to Class 1(Good segment) in K-means and he has p_1 likelihood for being in Class 1, p_2 likelihood for being in Class 2 and p_3 likelihood for being in Class 3 with p_2 being the largest among all probabilities, then we define the new class of this customer as Class 2 (Better segment).

Example of Decision Tree

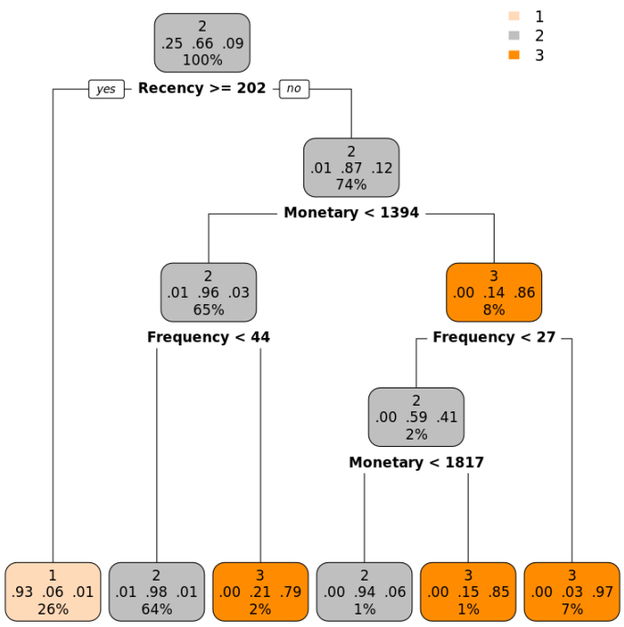

Figure 1: Decision Tree with 3 customer classes

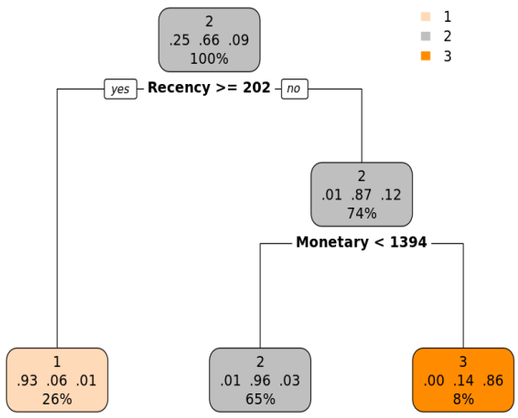

Figure 2: Pruned decision tree