Global Journal of Enterprise Information System

DOI: 10.18311/gjeis/2016/15771

Knowledge and Choice based Decision Trees for Manufactures

Abhisek Mukherjee*

Analytics and Insights, Tata Consultancy Services, Bangalore, Karnataka, India; abhi.mukherjee1@tcs.com

Abstract

With recent advances in data-driven analytics, and the resultant improved capabilities in working with huge datasets, strategic planning has become more complex for business units, and subsequently for the retail domain. The purpose of this study is to understand and define the hierarchy of decisions that leads to the purchase of consumer packaged goods. Consumers are driven by a much complex decision making process inside a store that finally leads to the purchase of a product. In this paper we are trying to emphasize on the consumer decision process and how it could be utilized by the retailers as well as the manufacturer to meet the needs of the customers. In a way it will help to make more successful sales conversion which in turn helps the manufactures of different products. The focus here is to capture the decision pyramid by using core statistical techniques.

Keywords: Agglomerative Clustering, Decision Trees, K-Means Clustering, Switching Matrix

1. Introduction

In today's world before we make any purchase, search is an activity which most of us engage regularly to extract up-to-date information which enable us to make the right purchase. However uncertainty is the driving force for consumer search. It should be mentioned that when an individual enter a store in his/her sub conscious mind they will already have an idea what they want to purchase and also a grief idea of the attributes they want in their product. If consumers had a perfect knowledge about their preferences and the market offerings are in line then life for the consumers will be much easier. This is an optimum condition which is very difficult to achieve. Generally our preference starts flickering moment we are given multiple choice.

If the decision trees created are clear and simple then depictions of product sets considered in a hierarchical order will help towards understanding products which can substitute each other. This representation of category attributes is extremely helpful in making assortment recommendations and in evaluating product development or brand positioning opportunities.

The decision trees based on consumer data should not be confused with the planograms[1]. However it can influence the shelf space in the context which items to include or exclude. So it helps in understanding groups of products which are forming the clusters. This cluster directly or indirectly influence the buyers while they are shopping.

These kind of product clustering has benefits the consumers in more than one ways like you get what you require more easily which in turn saves an individual time and also itrequires less effort. All these add to customer satisfaction and a happier customer tend to spend more money in his/her stipulated time. This will help in the revenue generation for the manufacturer.

The manufacturer historically spend a lot time to understand the market requirement and the requirement of their potential customers. These knowledge and choice based decision trees provides a great boon to the manufacturer. They always be a leg up in our competitive world. These models help in building customer loyalty within the brand and retailer.

A decision tree primarily focuses on how an individually within a store behaves and finally ends up purchasing the product[2].

There is very few researches have been cited for particular analysis, algorithm, or software program that can be used for generating the perfect CDT (consumer decision tree (CDT)). Therefore, we have elaborated the idea to create a CDT which will provide insights on planogram arrangement. The manufacturer can make the decisions for future manufacturing schedule on the basis of provided planogram arrangement.

*Author for correspondence

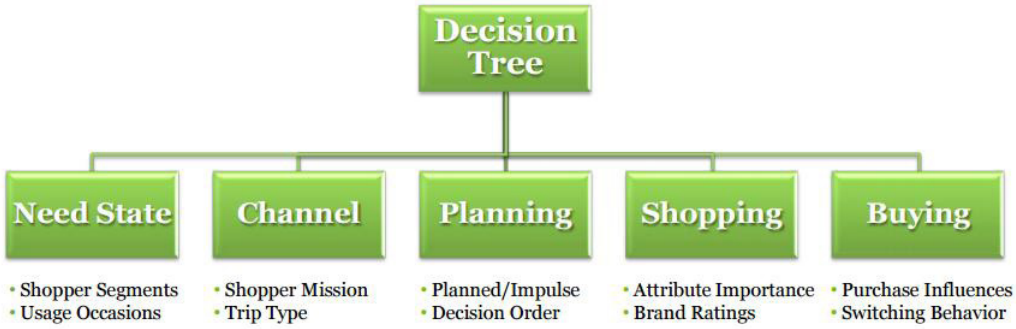

Figure 1. In store Path to Purchase[2].

In this paper, we have proposed an approach that will help the manufacturer with insight into which product and attributes are more important to the customers. Hence this will lead to the revenue realization for the manufacturer. Here we concentrate on a particular category of our choice mostly in the retail consumer packaged goods sector. Within that category we intend to identify the key players (brands) and the hierarchy of the choices that people are making. In order to achieve that we will primarily use a combination of supervised and unsupervised clustering techniques. Also highlighting how to resolve some of the challenges that we will face during the analysis phase.

The paper is structured in the following manner we first look into some of the existing methods and the challenges we face, following which we will discuss a solution framework and methodology. Finally we conclude by mentioning the advantages of the solution and the future scope of improvement.

2. Challenges with Current Solutions

The decision making process is a very complex process which an individual takes before finally purchasing the product. Some of the advanced statistical techniques used to create decision trees are as follows[2]:

- Logistic regression
- Statistical classification trees
- Hierarchical cluster analysis

However the decision trees which are widely used has its own pitfalls:

- Stability Issue
- Complexity
- Cost
- Biased Nature
- Too much information

3. Stability Issue

Decision trees are affected by addition or deletion of observation. The tree structure will change to a great extent. Also if we add or delete any variables from our models this will impact our analysis.

4. Complexity

This is one of the major problem of decision tree making. Though decisions trees are easy to comprehend than other modeling techniques. However is the size of the decision tree increases it tends to become much more complex and a time consuming affair.

Computing probabilities of all possible branches, determining the best split of each node, and selecting optimal combining weights to prune algorithms contained in the decision tree are complicated tasks that require much expertise and experience[3].

5. Costs

The cost component of decision trees are an indirect effect of the complexity of the decision tree. We can very well understand that from the fact that well trained analyst would be required in order to build decision trees which are big. The time involved to train individual the concepts of different statistical methods to take care of the decision trees. Also one need to have in-depth business knowledge to make correct inferences. This makes decision trees an unlikely choice to many analysts as it tends to become more expensive[3].

6. Biased Nature

In our dataset if we have too many categorical variables then in that case the decision becomes more biased towards those categorical variables which has many levels.

7. Too Much Information

Another major detractor for decision trees is that it is resource intensive. In this case we get bombarded with information which at some point causes "paralysis of analysis ". The time taken to process this magnitude of data is long which in turn results delay in decision making process. In today's cut throat competition every minute lost may result in our revenue loss[3].

When we compare decision trees primarily classification trees with logistic regression has its own merits and demerits.

Logistic regression results are easier to interpret compared to decision trees especially in cases when we have too many attributes or features to be incorporated. Logistic regression is also better than decision trees in the context of over fitting. Also it is observed that the computational time required by logistic regression is faster. Logistic regression much more reliable and provides much flexibility when we are modeling. In logistic regression we have different methods for variable selection like subset selection, forward, backward, stepwise etc. We can also apply LASSO with logistic regression. The result interpretation for logistic regression is also easier[4,5].

However both Logistic and decision trees are plagued with the problem of instability. Though logistic regression provides a probabilistic framework but it gets impacted badly by any change in the variable list. In the following section we share how we can tackle these issues.

8. Analytic Solution Approach

8.1 Data Requirement

The data used to feed the tree is a critical part of the analysis. The data can be from panel data based on a statistically valid sample and that includes the offline population, cell phone-only households, and key ethnic groups. The data should have details of the each item transaction and key attributes information over the last couple of years.

8.2 Methodology

We assume that we have a heterogeneous set of population of consumers choosing among various brands.

It is well known that in consecutive purchases, many customers will re-purchase the same brand they bought last time, while many others will try out different brands. Marketers are interested in both the aspects of this phenomena. Repurchasing is measured as a behavioral loyalty metric, while brand switching is one among the number of methods used to identify competitive market structure. Examination of competitive market structure provides an understanding of the intensity of competition between particular brands or product variants which also gives fair amount information about the different attributes of the competitive products[6].

For example, the marketing manager for company A is interested to know to which Brands Company A loses sales to when a current company buyer buys a competitor brand on their next purchase. Likewise, which brands does Company A take sales from. If there is a particular brand that company A competes against very intensely, then company A can try to determine the cause of this, because the other brand represents a threat[6].

A measure of substitutability between competing brands in market research is to build a brand switching matrix. A brand switching matrix can be constructed either by cross elasticity or by brand switching probabilities. Brand switching probabilities are estimated from panel or survey data as cross classification probability (proportion of times Brand i and j are purchased on two adjacent occasions).

Table 1. Sample Brand Switching Matrix[6]

Second purchase									
First purchase	1	2	3	4	5	6	7	8	9
1 Heinz Tom	391	158	90	22	25	9	13	1	0
2 Heinz Chick	148	136	60	5	15	1	1	1	0
3 Heinz Veg	105	64	122	2	5	5	1	2	2
4 Campbell Tom	29	10	8	103	85	19	2	2	0
5 Campbell Chick	27	13	9	92	128	17	4	3	1
6 Campbell Veg	5	2	4	31	24	27	0	1	0
7 WW Tom	7	0	0	1	5	0	42	18	15
8 WW Chick	4	0	2	2	3	0	17	15	3
9 WW Veg	2	0	1	0	0	0	10	6	8

In the switching matrix between brand i and j, a higher number will indicate that these products are substitutable and similarly a lower number indicates that there is hardly any switching. This forms the basis of our distance matrix that we will use for clustering.

Once we have the distance matrix we will provide that as an input to k-means clustering. The k-means cluster will provide us the initial set of clusters. The optimal number of cluster is selected through an iterative process where we basically plot the between cluster variance against the number of cluster and find the point where percent difference from the previous cluster converges. Also we will be using Hubert Index and other statistical measure to justify the optimal number of cluster selected based on majority rule[7,8].

In k-means clustering it is observed that some of the cluster will have observation whose profile might not match with the dominant features of the cluster. In those kind of scenario we have correlation analysis to find the best match for that particular observation and reallocate them. Also we will calculate the gini index or any other heterogeneity index. To validate that the reallocation is working properly. Based on the latest k-means cluster result with the reclassified observation we will perform the Agglomerative Hierarchical Cluster which will provide us the tree structure[9].

From the optimal set of cluster we need to profile each segment. Agglomerative Hierarchical Cluster on the data will provide us the tree structure. A closer inspection of the tree will provide us the levels to identify the customer priorities.

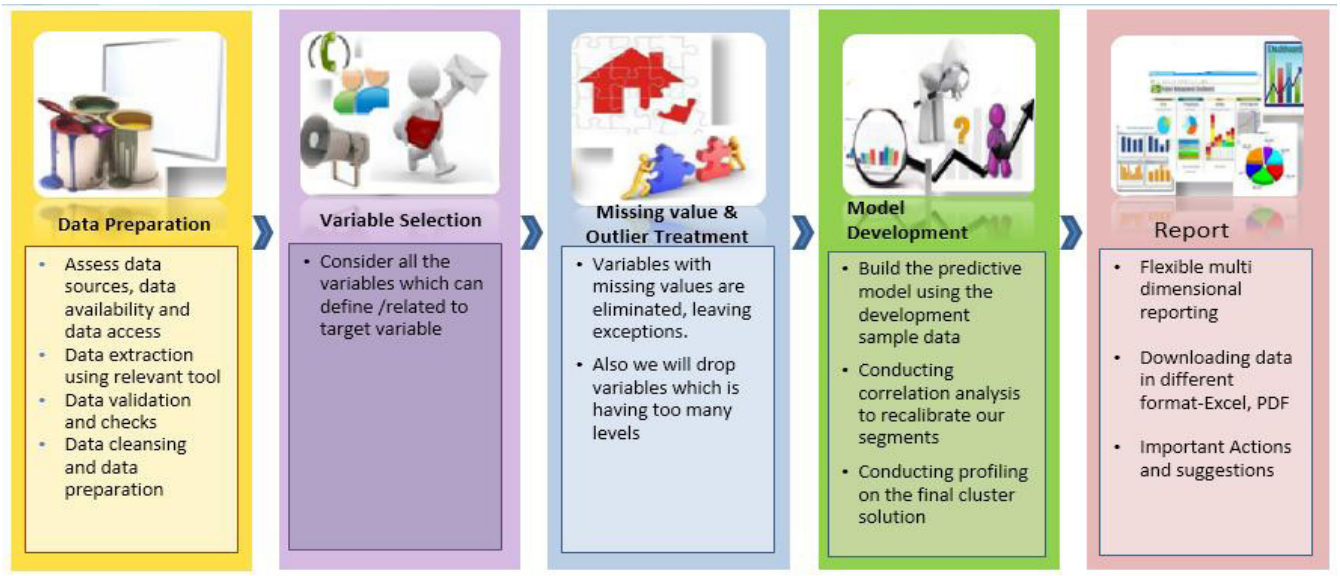

Figure 2. Modeling Framework.

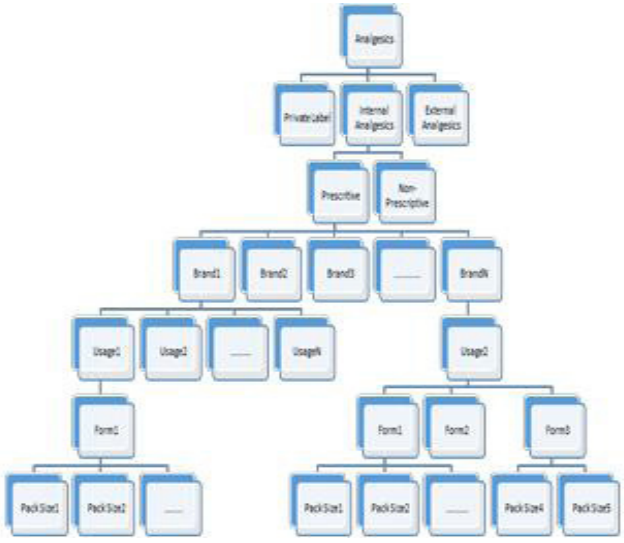

Figure 3. Illustrative Output.

9. Managerial Discussion

In this paper we are basically using an indirect approach to track customer preference. Traditionally we generally use propensity to buy models based on various attributes of the data along with customer information. However here we are looking at the purchase pattern of customers over a period of time to understand that within a category the competitive landscape placement.

In the above example we are looking at the analgesics industry layout in US. This industry is broadly classified into private label, internal and external analgesics. When we did a deep dive into internal analgesics we see that it is further broken down as prescriptive and non-prescriptive. Then within prescriptive if we continue we will see the first level is different brands followed by usage pattern, form and finally the lowest level is pack size.

This gives the manufacturer the idea about how the customers are behaving when they have to buy analgesics from any retail outlet. So the first priority in this case is brand then the customer looks for the usage form and pack size respectively before making the final purchase. The manufacturer will get the idea where they are facing tough competition and also provide them knowledge of areas they have no presence. So our analysis will help the manufacturer with the insight on future product development.

10. Conclusion

The final outcome will give for a category/market the high level breaks and as we move down the hierarchy we will have finer details of the products segments. To arrive at the structure we will use unsupervised Clustering technique along with other clustering techniques. The solution will capture the major segments in the data and will also provide actionable insights to both retailer/manufacturer.

Decision tree models will be helpful to develop a deeper understanding of consumers' hierarchical purchase pattern. Decision trees will further reflect on which of the product attributes trump one another and how, for instance, these dynamics relate to the shelf organization in store environment, puts a fine point on consumer insight. Decision tree models can be manipulated to focus on either brand or product perspectives. Decision tree models can often capitalize on visual representation of the products considered in order to facilitate decision making.

Manufacturer armed with such detailed information about their product and attributes helps them to plan for future products. It provides them information about the competitive landscape and the scope or opportunity given the current market structure.

The future of decision tree is immense only if we can mine the data correctly and at the right time [10]. The next step in this direction would to use artificial intelligence to train our models based on the historic data. Manufactures getting real time feedback to evaluate each customer even before they visit any outlet. The trees will provide us best possible future state and build action plan to drive real change[11].

11. References

1. Wansink B. New techniques to generate key marketing insights. Marketing Research (Summer 2000), 2000; 28–36.
2. Lim JM. How Does International and Format Diversification Affect the Financial Performance of Retailers? PhD diss. University of Florida, 2011.
3. Thompson S, O'Callaghan C. Decision making in music therapy: The use of a decision tree. Australian Journal of Music Therapy. 2013; 24:48. Available from: http://www.brighthubpm.com/project-planning/106005-disadvantages-to-using-decision-trees/
4. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of Behavioral Medicine. 2003; 26(3):172–81. https://doi.org/10.1207/S15324796ABM2603_02
5. Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research. 2003; 4(Jun):211–55.
6. Dawes J. The structure of switching: an examination of market structure across brands and brand variants. PhD diss. University of Otago. 2007.
7. Kalwani MU, Morrison DG. A parsimonious description of the Hendry system. Management Science. 1977; 23(5):467–77. https://doi.org/10.1287/mnsc.23.5.467
8. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1979; 28(1):100–8. https://doi.org/10.2307/2346830
9. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
10. Chen Y, Tu L. Density-based clustering for real-time stream data. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2007. p. 133–42. https://doi.org/10.1145/1281192.1281210
11. Boutilier C, Dean T, Hanks S. Decision-theoretic planning: Structural assumptions and computational leverage. Journal of Artificial Intelligence Research. 1999; 11(1):94.

Citation:

Abhisek Mukherjee
"Knowledge and Choice based Decision Trees for Manufactures",
Global Journal of Enterprise Information System. Volume-8, Issue-4, October-December, 2016. (http://informaticsjournals.com/index.php/gjeis)

Conflict of Interest:

Author of a Paper had no conflict neither financially nor academically.