ISL 439E

Introduction to Machine Learning with Business Applications

2018-2019 Fall Semester

Term Project Final Report

A New Approach to Consumer Segmentation for Turkish Fast Moving Consumer Goods Market

070130262 – Ahmet Talha YİĞİT

070140217 – Barış SAMAK

1. INTRODUCTION

In today's world, the amount of stored data and the computational power to analyse it has become significantly large and beneficial for many sectors. FMCG is one of the sectors that can use data for many purposes such as customer segmentation. The aim of this project is to suggest a new consumer segmentation structure for Turkish Fast Moving Consumer Goods (FMCG) sector using household consumer panel metrics. The main purpose of this project is to find out distinct clusters of consumer panel households based on not only their household specific characteristics like socio-economic status, household size, average education status, etc. but also their consumption patterns. In the first step, unsupervised learning techniques such as K-means clustering will be used to obtain the customer clusters by using the attributes such as price, category, and retailer of the products in the baskets , in the second part of the study, supervised classification methods will be used to allocate new households into the already defined consumer segments. Traditionally, there are consumer segments that are being used in the industry using the characteristics of individuals. However, in this project; using advanced statistical learning techniques, we will be looking to come up with a better segmentation of consumers for FMCG companies to come up with better sales and marketing strategies and provide a more accurate segmentation model than the traditional one.

2. LITERATURE REVIEW

2.1. Market Segmentation

Market or customer segmentation is the activity to divide customers into groups by using some of their characteristics. With this activity, companies aim to gain a deeper understanding of their customers so that they can come up with better marketing and sales strategies. This activity relies on having the necessary data related to the customers. Because this data will be used to divide customer into different segments or groups. Since companies are allowed to target different customer groups, an effective segmentation has the potential to enable the company to use their marketing resources in the most efficient way. Knowing their customers better, they can conduct marketing mixes for the targeted group with a bigger accuracy. This way, customers feel more entitled to the rand and the company. They are more willing to feel valued and respected since companies have the chance to send more personalized messages through their marketing communications. Segmentation also enables companies to stay one step ahead of their competitors. Since they know their customers better, they can seize future trends and opportunities better and calibrate their strategies. Companies also gets the chance to analyze these segments so that they can decide ehich segments would produce a bigger, better, more sustainable revenue streams for the company.

2.2. Types of Market Segmentation

In the world of marketing, there mainly four different ways for segmentation. These are demographic segmentation, geographical segmentation, behavioral segmentation and psychographic segmentation. Demographic segmentation has been the most popular one both globally and in Turkey until now. This type of segmentation uses demographic variables to differentiate different customer groups. These variables are generally age, gender, family size, education status, employment status, religion, race etc. Whereas geographical segmentation differentiates customer groups by using where they live and tries to understand different customer behaviors for different regions. The third type of customer segmentation is the behavioral one. In this type of segmentation, the behavioral patterns of the customers are being used to differentiate like what they do, what they think etc. The last type of segmentation to mention is psychographic one. This type of segmentation uses the science of

psychology to divide customers in to groups and enabling companies to come up with strategies rearding psychological factors.

2.3. FMCG Segmentation

In short FMCG sector which means Fast Moving Consumer Goods is a huge sector both worldwide and in Turkey that aims to make the best use of segmentation. Various FMCG companies use different types of segmentation and often more than one at the same time. Some companies use different types of geographical segmentation which includes location, density, distribution channels and even climate. In the demographical segmentation the most common used variables are age, gender, income, education, family life cycle. It should be noted that geographical and demographical segmentation are the most popular ones in the Turkish FMCG sector still. On the psychological side; interest, ideas, opinions, activities, benefits sought are the variables that are being used to segment the customers in to bases. The last one is behavioral segmentation and it is perhaps the most interesting one for this sector. Having the panel data of consumers allows for agencies to see the behavioral patterns of their customers. Especially for FMCG companies this can be very useful if put in motion.

2.4. Machine Learning Applications on Customer Relationship Management

Machine learning techniques has been used in many sectors by various companies to have a better understanding and control over all customer relationship management areas. These areas are customer attraction, customer retention, customer development, and customer identification and theirs subtopics (Ngai,Xiu, & Chau, 2009). The machine learning techniques which are used in this topic are in a very wide spectrum and almost every technique can be applied to enhace customer relationship management. But the most commonly used techniques are clustering methods because of the topics unsupervised nature. With the applications of machine learning techniques, companies and researchers are aiming to look at this topic from a comtemporary angle instead of the conventional ones. This new angle has created more profitable customer relationship management applications and so a competitive advantage for the companies which integrate it to themselves.

2.5. Machine Learning Applications on Customer Segmentation

Customer segmentation is one of the most important topics of customer relationship management about which there are many researchers has studied. Understanding the underlying characteristics of customers is the main aim of the companies while assigning customers into different segments, because the characteristics of customers in the same segment should be similar, and the characteristics of different segments should be relatively different and distinguishable (Lee, & Park, 2005). By the usage of clustering methods such as K-means clustering, pattern based clustering, and self organising maps in the customer segmentation applications, it has been a subject to unsupervised machine learning (Ngai,Xiu, & Chau, 2009). The application of these methods has created an easy, profitable and reliable alternative for companies which are aiming to understand and recognise their customers in a better, more profitable, and holistic way; moreover, it has created a new area for different and creative approaches and solutions for this problem. In todays world, these applicaitons can be seen in many sectors.

2.6. Machine Learning Applications on Customer Segmentation in FMCG Sector

Fast moving consumer goods (FMCG) sector is a very large sector with massive number of customers which are expected to be loyal, and profitable for the companies. To create this loyalty, and profitability of the customers, FMCG companies should understand and know their customers in various ways; consequently, customer segmentation is an essential and can be a competency for them if it is done in the correct way for the correct goal. Machine learning usage in customer segmentation can exploit some expected or unexpected shopping patterns of different segments of customers. For example, earlier studies of Wal-Mart showed that there is a linkage between nappies and beers on Friday evenings in the USA (Denis, Marsland, & Cocket, 2001). By finding this kind of insights of shopping patterns, retailers can find more profitable ways to design their facilities etc..

As it is seen above review, with the contemporary techniques of using, mining, and understanding the data such as machine learning, companies have a chance to create more effective solutions for the conventional problems. Moreover, these applications can provide

cost minimization and profit maximization for the companies which are using them. Consequently, the literature about this topic has improved and is improving with the developments of data storage and computational technologies in the last decades.

3. CUSTOMER SEGMENTATION FOR FMCG SECTOR

3.1. DATASET DESCRIPTION

The dataset that will be used in this project will be the panel data of IPSOS Turkey. IPSOS Turkey is a leader company in the field of marketing research. IPSOS has a widely distributed household network in their panel services. They collect the consumption data of households periodically. Further in the project, this consumption data of Turkish households will be used to discover distinct clusters for FMCG sector.

In our data, every line consists of a purchasing of an item by a household which belongs to one of the main categories of FMCG products. While column “BASKET” refers to the basket that the item belongs to, “MAIN_CAT” refers to the category of the item. “VOLUP” refers to the volume of the item that has been purchased. On the other hand, “TOTVAL” refers to the total value of the basket which is the total value of every line that belongs to that basket. So every line refers to an item and every basket may consist of multiple lines. While the basket variable shows which basket the line of the item belongs to.

BASKET	MAIN_CAT	VOLUP	TOTVAL
01010341102-JAN-18	MEAT-RED-RAW	1000	52.9555
01010341102-JAN-18	MEAT-RED-RAW	1150	52.9555
01010341106-JAN-18	CONFECTIONERY	238	4.4982
01010341110-JAN-18	CONFECTIONERY	338	16.7622
01010341110-JAN-18	CONFECTIONERY	200	16.7622
01010341110-JAN-18	CONFECTIONERY	116	16.7622
01010341110-JAN-18	CONFECTIONERY	298	16.7622
01010341118-JAN-18	CONFECTIONERY	400	11
01010341118-JAN-18	CONFECTIONERY	292	11
01010341126-DEC-17	CONFECTIONERY	312	3.7128
01010341129-DEC-17	MEAT-RED-RAW	1500	48
010103418722-JAN-18	MILKS	2000	8.1
010103418722-JAN-18	CONFECTIONERY	354	8.1
010103418722-JAN-18	PASTAS	500	8.1
010103418723-JAN-18	HAIRCARE	50	7.9
010103418724-JAN-18	RICE	1000	2.13
010103418728-DEC-17	CONFECTIONERY	55	0.4
010103418730-DEC-17	SOFTDRINKS	1200	3.25
01010342018-JAN-18	MEAT-RED-RAW	1764	67.59648
01010342022-JAN-18	MEAT-WHITE-RAW	708	6.903
01010342607-JAN-18	MEAL RELATED	340	27.75

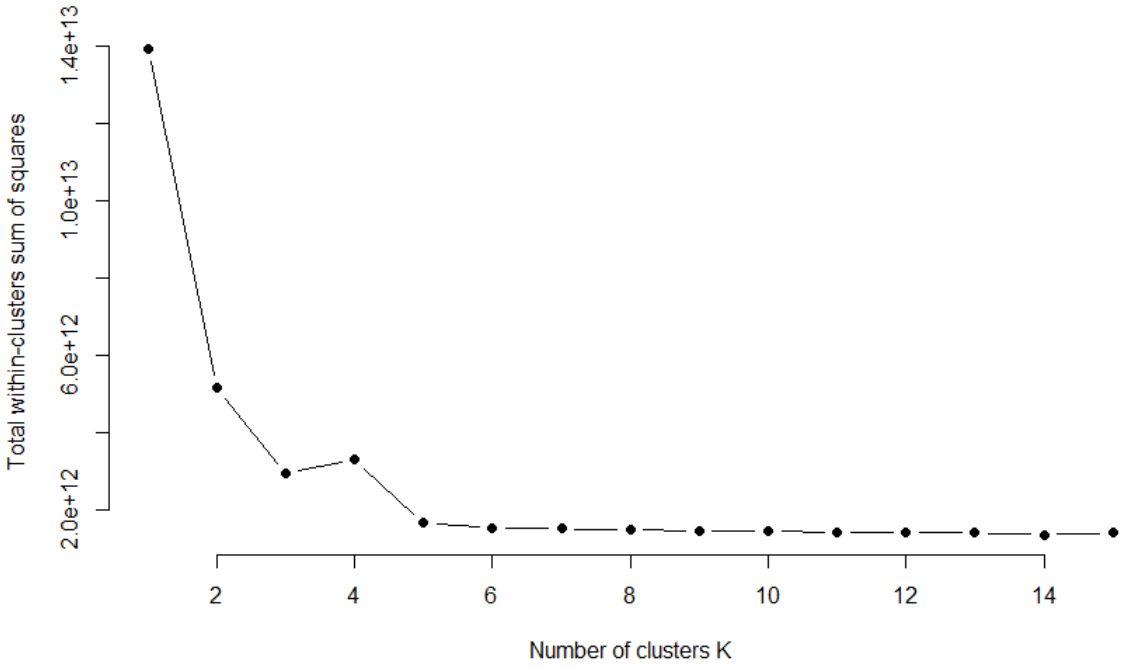
3.2. DATA MANIPULATION

Since it is our aim to define the different segments of consumers. Our aim to manipulate our data into a format that every basket is being described in one line with different variables. To do that a transformation has been made on R programming language. Every category that has been defined under the variable "MAIN_CAT" has been transformed into a variable. So that summing the lines that belongs to the same basket can be possible. After summing the lines that belongs to the basket, every line consists of all items that belogs to the same basket where every line refers to a basket. After that the value of every category have been divided with the total value of that basket. So that under every category it can be seen how many percents of that basket's value belongs to every category. Also another variable has been created which named "VALDIVVOL" which equals value divided with volume so that baskets with more valuable items can be seen.

BASKET	VOLUP	TOTVAL	TOTITEM	VALDIVVOL	MEAT_RED_RAW	CONFECTIONERY	MILKS	PASTAS	HAIRCARE	RICE	SOFTDRINKS	MEAT_WHITE_RAW	MEAL_RELATED	PAPER
01010341102-JAN-18	2150	52,9555	2	0,024630465	1	0	0	0	0	0	0	0	0	0
01010341106-JAN-18	238	4,4982	1	0,0189	0	1	0	0	0	0	0	0	0	0
01010341110-JAN-18	952	16,7622	4	0,017607353	0	1	0	0	0	0	0	0	0	0
01010341118-JAN-18	692	11	2	0,015895954	0	1	0	0	0	0	0	0	0	0
01010341126-DEC-17	312	3,7128	1	0,0119	0	1	0	0	0	0	0	0	0	0
01010341129-DEC-17	1500	48	1	0,032	1	0	0	0	0	0	0	0	0	0
01010418722-JAN-18	2854	8,1	3	0,002838122	0	0,333333333	0,555556	0,111111	0	0	0	0	0	0
01010418723-JAN-18	50	7,9	1	0,158	0	0	0	0	1	0	0	0	0	0
01010418724-JAN-18	1000	2,13	1	0,00213	0	0	0	0	0	1	0	0	0	0
01010418728-DEC-17	55	0,4	1	0,007272727	0	1	0	0	0	0	0	0	0	0
01010418730-DEC-17	1200	3,25	1	0,002708333	0	0	0	0	0	0	1	0	0	0
01010342018-JAN-18	1764	67,59648	1	0,03832	1	0	0	0	0	0	0	0	0	0
01010342022-JAN-18	708	6,903	1	0,00975	0	0	0	0	0	0	0	1	0	0
01010342607-JAN-18	6870	27,75	7	0,004039301	0	0	0,165766	0	0	0	0	0	0,463063063	0,07027
01010342613-JAN-18	4600	18,6	3	0,004043478	0	0	0,247312	0	0	0	0	0	0,389784946	0
01010342626-DEC-17	2000	4,6	1	0,0023	0	0	1	0	0	0	0	0	0	0
01010426707-JAN-18	1762	7,91138	1	0,00449	0	0	0	0	0	1	0	0	0	0
01010426710-JAN-18	1852	10,16748	1	0,00549	0	0	0	0	0	0	0	1	0	0
01010426728-DEC-17	1922	11,4359	1	0,00595	0	0	0	0	0	1	0	0	0	0
0101034804-JAN-18	5000	15	1	0,003	0	0	1	0	0	0	0	0	0	0
0101034806-JAN-18	2345	61,98	5	0,026430704	0	0	0	0	0	0	0	0	0	0
0101034812-JAN-18	5000	17,5	1	0,0035	0	0	1	0	0	0	0	0	0	0

3.3. K MEANS CLUSTERING

To discover different FMCG customer segments using our dataset, k-means clustering method will be used. To determine which k to use would be most appopiate for this problem, the elbow method has been used. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified.



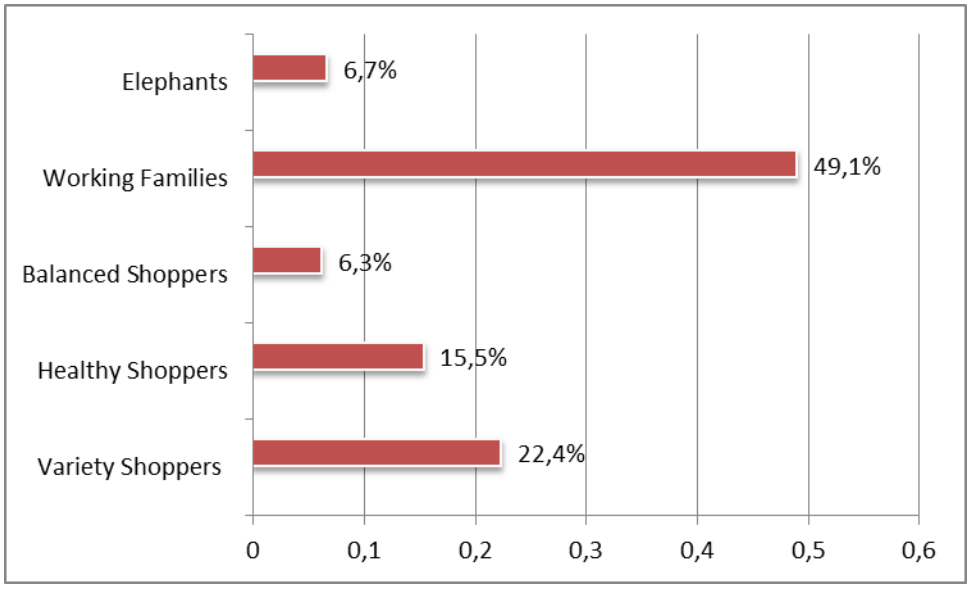
Either k=3 or k=5 has been decided as approipiate for this clustering problem. Since the explained variance is not decerasing after 5, k=5 has been selected as the appropiate k for this problem.

After this the clustering algorithms in R programming language has been used where k=5 and the five distinct clusters of FMCG customers have been found. And these clusters have been named according to their distinctive qualities.

Clustername	Cluster Sizes		VOLUP	TOTVAL	TOTITEM	VALDIVVOL	WATER	CONFECTIONERY	MILKS	SOFTDRIN	YOGHURT	CHEESE	EGGS	MEAL_REL	OILS
Variety Shoppers	59846	22,4%	2669,467	22,406	3,237	0,009	0,56%	12,77%	12,07%	13,73%	4,49%	6,52%	2,33%	5,98%	1,30%
Healthy Shoppers	41325	15,5%	5745,157	31,684	4,048	0,005	5,34%	7,89%	23,89%	6,57%	6,99%	4,71%	2,33%	5,67%	5,31%
Balanced Shoppers	16748	6,3%	11837,438	54,518	6,453	0,005	19,86%	7,42%	8,98%	4,47%	4,38%	4,99%	2,05%	5,64%	5,94%
Working Families	130924	49,1%	618,359	10,223	1,764	0,126	0,81%	29,93%	4,44%	5,55%	1,38%	5,68%	6,26%	4,77%	0,25%
Elephants	17895	6,7%	25191,183	34,295	3,605	0,001	79,54%	2,28%	0,99%	0,93%	0,44%	1,45%	0,59%	1,60%	1,67%

MEAT_WH	LAUNDRY(	PERSONALCARE	CLEANING	MEAT_REI	COFFEE	BREAKFAS	BABY_REL	RICE	LEGUMES	TEA	PAPER	FATS	HOUSECAI	FLOUR	PASTAS	HAIRCARE	MEAT_REI	FOOD_DRI
5,54%	1,37%	1,76%	3,57%	1,01%	3,09%	1,09%	2,45%	2,17%	2,03%	1,16%	1,89%	1,92%	1,17%	1,47%	0,92%	1,67%	1,38%	
2,26%	3,22%	1,43%	1,74%	0,75%	2,04%	0,85%	2,08%	1,56%	1,83%	1,07%	1,78%	1,47%	1,82%	0,95%	0,70%	1,16%	1,49%	
2,00%	5,15%	1,73%	1,71%	0,76%	2,27%	0,94%	2,35%	1,85%	2,16%	1,50%	1,92%	1,63%	1,86%	0,97%	0,98%	1,30%	1,62%	
2,37%	0,36%	4,99%	3,02%	3,28%	2,31%	3,08%	0,64%	1,19%	1,99%	1,94%	1,94%	1,18%	0,45%	1,85%	1,85%	1,94%	1,65%	
0,50%	1,22%	0,54%	0,60%	0,25%	0,66%	0,33%	0,61%	0,59%	0,74%	0,46%	0,57%	0,42%	0,39%	0,28%	0,31%	0,41%	0,46%	

These clusters have been named as Variety Shoppers, Healthy Shoppers, Balanced Shoppers, Working Families and Elephants.



The first cluster which is “Variety Shoppers” consist of 22,4% of the FMCG customers and their shopping behaviours can be described as buying from many categories even the ones that are that popular in other clusters. At this point they distinguish from the second cluster which is “Balanced Shoppers”. They have a similar behaviour on buying from different categories but they do buy from these distinct categories compatible with the overall needs. The third cluster is “Healthy Shoppers” and consumers who belong to this cluster lives a more modest and undemanding lifestyle. They buy the vital categories for a healthy life more and others less. The fourth cluster that has been defined is the one with the most coverage of population with 49,1%: “Working Families”. And the fifth and final cluster has been named as “Elephants”. This may be due to a lack of coverage in the data since this data consists of one month of purchases. “Elephants” spend a lot and they spend mostly on water which is about 80% of their spendings

3.4. ASSIGNMENT OF CUSTOMERS TO DEFINED CLUSTERS

After finding the clusters for all of the data, the allocation will be done by a supervised classification method. Decision tree extreme gradient boosting method with "multi:sofmax" objective will be used as the classification method. While data is prepared, it is seperated into an 80% train set and 20% test set. To find the needed parameters such as "eta", "gamma", "max_depth", and "nrounds", a caret library bas grid search has been used. After finding the optimum parameters, the extreme graident boosting method has been applied with them to the train set. And then made predictions on the test set. Extreme gradient boosting has achieved perfect classification in the test set.

		Predicted Classes				
		Balanced Shoppers	Elephants	Healthy Shoppers	Variety Shoppers	Working Families
True Classes	Balanced Shoppers	3310	0	0	0	0
	Elephants	0	3570	0	0	0
	Healthy Shoppers	0	0	8311	0	0
	Variety Shoppers	0	0	0	12058	0
	Working Families	0	0	0	0	26099

As can be seen on the confusion matrix, all samples have been assigned to the right cluster.

4. CONCLUSION&DISCUSSION

Clustering algorithms have been ran on the dataset and 5 distinct FMCG consumer segments have been found. On the second part of the project decision trees with xG boost have been able to assign the customers into right clusters perfectly. Even though this means that this five clusters can be perfectly discriminated, this does not mean that this segmentation is the most proper one for the FMCG sector. For example the “Elephants” which consist of 6,3% of population who consumes a lot and mostly only water is not realistic to find in the real world. This goes on to show that a wider dataset that consist of a bigger time period may result in better clusters for future researches.

5. References

https://www.marketing91.com/behavioral-segmentation/

http://shodhganga.inflibnet.ac.in/bitstream/10603/16744/12/12_chapter5.pdf

https://www.marketing91.com/4-types-market-segmentation-segment/

https://searchsalesforce.techtarget.com/definition/customer-segmentation

http://cmapspublic3.ihmc.us/rid=1MSYC3Z3W-1B2W04K-15MY/DM-usage.pdf

Lee, J. H., & Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. Expert Systems with Applications, 29, 145–152

Dennis, C., Marsland, D., & Cockett, T. (2001). Data mining for shopping centrescustomer knowledge management framework. Journal of Knowledge Management, 5, 368–374.

https://datascienceplus.com/finding-optimal-number-of-clusters/