Athens University of Economics and Business (AUEB)

School of Business

DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

MASTER THESIS

ILIAS KATSIS

DATA ANALYSIS

IN A "DARK STORE" ENVIRONMENT

Company : POCKEE

Supervisors :

Angeliki Poulymenakou – Associate Professor

Emmanouil Zachariadis – Assistant Professor

Submitted as part of the acquisition requirements Postgraduate Diploma (MSc) in Management Science and Technology

Athens, January 2022

This page is intentionally left blank.

Certificate of Diploma Thesis

I, hereby, declare that the work presented in this thesis in fulfillment of the requirements for the award of Master of Science (MSc) submitted in the Department of Management Science and Technology at Athens University of Economics and Business has been executed and authored by me and has not been submitted or approved in the framework of someone else's postgraduate or undergraduate degree in Greece or abroad. This thesis, having been executed by me, represents my personal views on the subject. The sources I raised for the elaboration of this thesis are all mentioned and the material used was either given to me by the company or created by me.

Katsis Ilias

MSc student in Management Science and Technology

This page is intentionally left blank.

Abstract

The scope of this thesis is to analyze multidimensional data from a dark store called PockeeMart and derive deep business insights in order to achieve a complete data-driven strategy. The main objective of the dissertation concerns the application of machine learning to the data with purpose to elicit association rules, segment the customers based on their shopping behavior and create a predictive scenario, which classifies the order's placement recency of the customers. PockeeMart is active in the FMCG sector, which is very demanding and competitive. The number of issues, that need to be taken into consideration and the number of tasks that must executed, makes it really challenging. During the pandemic period dark stores became more popular, as they do not require physical presence being part of online retailing and offer various other benefits for the users.

This research methodology employed in this dissertation includes the data collection and the preprocessing for PockeeMart's data and the implementation of the three tasks. For effective and clear results Python is used along with supportive libraries and visualizations. Association rules are extracted with Apriori algorithm, the customer segmentation and the clustering are executed with K-means algorithms, while for the classification process three different models run and are evaluated individually.

The findings of the analysis demonstrate firstly the most frequent itemsets for the dark store and groups of its customers based on the RFM model and their shopping habits. For classification clustering is combined with user's demographic data. The result is a prediction, which refers to whether the customer intends to make purchases within the month or not. From a business view, as the structure of dark store is not the same as a traditional retail store and the results cannot change the layout of the store, all the insights can be proved beneficial for the supply chain, the structure of the application of and the marketing strategy through promotions and discounts. Finally, the delivery management is vital and data-driven decisions can modify it in order to maximize profit.

Keywords:

FMCG, Online Retailing, Dark Store, Association Rules, Customer Segmentation, Classification.

This page is intentionally left blank.

Table of Contents

1. Introduction

1.1 FMCG Sector

Fast Moving Consumer goods sector refers to products, which sell quickly at low cost compared to other types of products. The demand is high and their shelf life is relatively short as they are purchased in daily basis. It is obvious that, the profit margin can be characterized low, but the volume of sales is high

Fast Moving Consuming Good can be divided in categories including processed foods, fresh or frozen products, cleaning products, cosmetics or even medicines(Kenton,2021) . Companies are obligated to focus on the long-term customer relationship, because the consumers have a lot of different options and retention is difficult.

The market size is very large and very competitive and changes are constantly being made in order to satisfy the large number of customers. Online stores are a big part of the FMCG sector and the biggest chains have created a hybrid model of retailing combining traditional with internet shopping. During the pandemic period, a new type of store has evolved, called dark store, changing the balance.

1.2 Thesis Concept

The major goal of this study is to detect and analyze multidimensional and users' characteristics from a dark store. adopt a data-driven policy and orient its decisions on the historical data and characteristics of its customers. Machine learning techniques and data mining support this project and through the implantation of specific algorithms, insightful results are gained. The key difference with the existing

retail channels, is that the business background of a dark store is based on fast delivery, and user-friendly navigation with personalized suggestions for the users, therefore the decisions coming from the data boost these aspects of the store.

1.2.1 Thesis Contribution

1. Observation and data selection from company's database
2. Data transformation techniques
3. Implementation of 2 unsupervised learning algorithms (Apriori, K-Means)
4. Evaluation of 3 classification models (Decision Tree, Random Forest, Naïve Bayes) through accuracy metric.
5. Analysis of the total results from the business view

1.3 Thesis Structure

The first chapter of the thesis is an introduction to the FMCG, as they are main activity of the dark is being analyzed. The second chapter is a complete literature review and refers to machine learning, data mining and all the analytics used in order to explore the data of dark store. Each term is explained from a technical and business point of view as well. Third chapter defines the problem that is discussed in the dissertation, while the fourth one presents the research approach of the thesis. The next part of the thesis is a brief description of the company and the responsibilities of the business intelligence analyst. The three following chapters are the main of the thesis, as they include the analysis, the findings and the discussion about the company's data and users' characteristics. The last part is a conclusion about the efficiency of data-driven policy for dark stores and any business extensions, which can be considered in the future.

2. Literature Review

2.1 Machine Learning

Machine learning is a big part of artificial intelligence, which concerns computer algorithms and how a program can learn and be improved from experience E, keeping up with some tasks T and measured by performance P. These are the three features (class of tasks, source of experience and measure of performance) , which must be identified to have a well-defined learning problem. The best fit to learning problems is determined after a detailed search in a large scale of possible hypotheses, with respect to the observed data and prior knowledge. (Mitchell, 1997). Machine learning is considered as one of the most rapidly evolving sectors of science in general and combines computer science with statistics in order to automate a big variety of research and business procedures . This growth has changed the way of decision making and has reduced uncertainty. (Jordan & Mitchell, 2015) Generally, machine learning has the ability to recognize patterns in data, which can give useful information about a problem. Year on year the rise of machine learning is tremendous and its impact is visible to many areas like health care, finance, manufacturing etc. (Carleo et al., 2019)

Learning problems are categorized in :

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

2.1.1 Supervised Learning

Supervised Learning concerns algorithms, which generate functions to give labels to the data. The program is obligated to learn the labels and the way they are classified from past data. The basic concept of this kind of problem aims to create a fully trained algorithm, which can predict the labels of new data based on the input – output fed data (Oladipupo, 2010). More specifically, each sample of given data Xμ

∈ Rp with μ = 1,…,n has a unique label yμ ∈ Rd, most commonly d = 1. This kind of samples are gathered in a set called training, which has the goal to learn the function f and assign a label to every new sample Xnew. In between there is the evaluation of the classification with the help of test set (Carleo et al., 2019). Supervised learning systems have applications to several areas . The most famous classifiers are those dealing with face recognition, spam classification , medical diagnosis for patients and advertisement popularity. (Jordan & Mitchell, 2015) .

The first step of the supervised machine learning process is about data collection and the sorting of the most important attributes of the data. The second step is related to data preparation and preprocessing. At this stage, data cleaning is vital and missing data should be handled properly, especially in large datasets. Algorithm selection is critical and is after the separation of the set in training and test. Finally, the evaluation of the classifier is based on metrics like accuracy, recall and precision. If the outcome is acceptable, the process is finished. Otherwise, some stages should be reexamined and repeated. (Kotsiantis, 2007).

2.1.2 Unsupervised Learning

In unsupervised learning, data is given to the algorithm without labels and every process is responsible to find the regularities in the input. The program in unsupervised machine learning can learn to cluster, organize or group the data, recognizing patterns that occur more often than others. (Alpaydin, 2014). The construction of an unsupervised model is based on a collection of unlabeled examples {x1, …, xm}, which belong to some classes with unknown characteristics. The goal is to specify the properties of these classes and categorize them. The majority of unsupervised algorithms have been developed to calculate the distance between two examples to group them together. (Dietterich, 1997). Some of the most famous areas that unsupervised learning has applications are recommendations systems and consumer buying habits (Alpaydin, 2014).

2.1.3 Reinforcement Learning

The last basic concept is reinforcement learning, where the aim of an autonomous agent is to choose the right actions and maximize the “reward”, as the goals of the agent are named. (Mitchell, 1997). These are the cases where sequence of

actions is more important than a single one. As a result, the state of the environment should be observed, identify good actions and generate good policies. A classic area of reinforcement learning research is game playing, where a good move is associated with a good policy. Therefore, the algorithms learn how to play well and extend their usage to other business and economic sectors (Alpaydin, 2014). The main type is called Q learning. Q learning creates a matrix, in which qualities are graded under specific conditions – state of the environment. The function is iterative and sometimes it is very difficult to store the whole matrix due to the number of states and actions, which should be examined (Carleo et al., 2019).

2.2 Data Mining

Data Mining is the topic that combines practical techniques with learning theory in order to resolve the situation about data and arrange them in structural patterns, which are easy to be described and interpreted. (Witten & Frank, 2005) [7]. It is true that data mining can be considered as the evolution of information technology and huge collections of row data for example can be transformed to knowledge with the right techniques. (Han et al., 2012) [8]. Data mining consists of complex steps, which are repeated and go through mostly historical data searching for interesting knowledge.

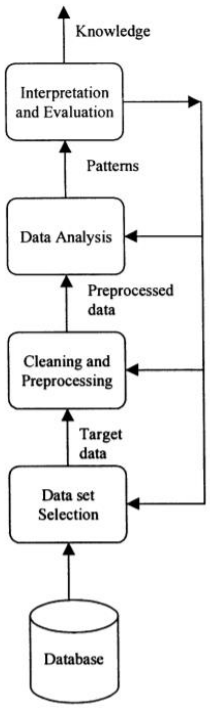


Figure 1 : An overview of the data mining process (Mahapatra, 2001)

Each set of data must be cleaned and underwent preprocessing procedures. Usual abnormalities concern duplicates, missing values or other inconsistency issues. The following is data analysis, where business sets the requests and mining models are obligated to extract useful conclusions. One of the challenges is about the connection between the business sector and analytics one, because the perspective is not the same. Due to this gap, reports and visualizations have been developed in order to make data mining output clearer. One of the most important parts of the process is the evaluation as the results must be precise and as accurate as possible. If evaluation achieves good scores, interpretation will take place leading to knowledge. (Bose & Mahapatra, 2001).

The fact that data predictive modeling, forecasting and descriptive modeling techniques, makes it an ideal tool for businesses. Some great examples, to which there is application are customer segmentation, customer retention and cross-sell and up-sell opportunities. In order to improve customer relationship management, machine learning techniques should be applied and give to the business decisions a data-oriented perspective. (Thanuja, et al., 2011). Machine learning and data mining are inextricably linked, because they are the evolution of traditional statistics, which always are the base for every analysis and model (Han et al., 2012).

The main machine learning techniques used for data mining are :

- Association
- Classification
- Clustering
- Forecasting
- Regression (Thanuja, et al., 2011)

As it was previously mentioned, data-driven management is becoming more and more part of companies' reality. Business analytics can be named the sector, which is being used to make informed business decisions from various data sources. Data science and business analytics work together, utilizing the information in a different way, but having the same target to glean insights from data to inform business decisions (Gavin, 2019). The field being discussed is now a trend and fully embedded in daily routine. Apart from the benefits analytical tools can offer, there are

some serious challenges, which define the gap between analytics processes and real business needs.

Nowadays, there are strict business constraints and timelines are very stressful, therefore cycle time for data collection and analysis should be less. On the other hand, business users should stop having unrealistic expectations. The fact that the majority of decisions are data-oriented does not mean that data mining can always present clear and easily understandable results. One big challenge concerns the way that these results, which are coming from various tools designed for quantitative analysts, are presented to the business end-users. The descriptions of each analysis should be translated into business language and answer to real questions. To make analysis efficient, it is required the data integration from multiple data sources. Extract, transform and load (ETL) processes are often underestimated, but they are the base of the data mining regardless of their difficulty and cost (Sumathi & Sivanandam, 2006).

2.3 Market Basket Analysis

Businesses in the retail sector and especially at Fast Moving Consumer Goods' (FMCG) one, handle large scales of data. Only a proportion can be extracted and give useful information. The step after the extraction and the adjustment of the to the desired format is the analysis (Kaur & Kang, 2016). Market basket analysis is one of the main data mining methods, also known as association rule mining. This method identifies frequently purchased itemset among large databases (Griva et al., 2018). Supermarkets have based their stores' layout on market basket analysis. Before it, they relied on traditional approaches, where products with the same characteristics were placed together. Thus, consumers with limited time were prevented from making their purchases quickly and profits for retail shops were less. Association rules are the outcome of market basket analysis and the guide for every store to be reorganized based on them. The source in order to gather and compute these buying associations can be the receipts, various coupons or surveys (Cil, 2012).

There are four reasons for retailers to include market basket analysis in their daily procedures. Supermarkets are consumer-oriented and the combinations of products on every receipt are very important, more than brands' products list. Therefore, association analysis can identify the most frequent combinations and give to the store helpful information. Moreover, consumers tend to be engaged with

individual shops and less with brands, which means that every basket can be characterized as “product of the store". The main goal of every retailer is to maximize their profits. This can be achieved with a variety of marketing campaigns targeted at combinations of products and not at specific ones. From the perspective of products’ manufacturers market basket analysis is also essential, because they fundamentally have two types of customers. The first one is every retailer and the second one is of course every single consumer. The effect of the analysis shows in what context that every brand is combined. Finally, market basket analysis has the ability to present other important aspects of data about shopping behaviors, filtered by the location or the other descriptive statistics for example and lead to successful customer segmentation as a next step.(Julander, 1992).

2.3.1 Association Rules

Association rules have features and there are metrics, which can be used to evaluate them. The two main metrics are the support and the confidence “Good rules” are called strong and they must satisfy a minimum h a minimum support threshold (min sup) and a minimum confidence threshold (min conf). Support measure (s) represents the percentage of transactions that contain every set (A ∪ B). Each set is a unique rule that consists of predecessors and successors. Confidence is taken to be he conditional probability P(B|A), which describes the percentage of transactions containing A and also contain B. However, not all the strong rules are beneficial for businesses, because “real strength” is not measured by support and confidence. In order to reduce this insufficiency, a correlation measure is included in the association analysis. This measure is called lift and computes the performance of an association. Simply lift represents the importance of each rule and checks if the occurrences of two events A and B are independent or dependent and correlated. Measure of events A and B can be measured from the formula : (A, B) = P(A ∪B) / P(A) * P(B) (Han et al., 2012).

2.3.2 Apriori Algorithm

The algorithm being used to extract association rules is the Apriori algorithm. In the beginning , the algorithm counts large 1-itemsets.Then a following pass k consists of two stages .Fistrly founds the candidate itemsets Ck from the previous

phase (k-1) and count them with the aid of apriori-gen function throughout the whole dataset. Secondly, all the itemsets which have some k-1 subset that are not in the Lk – 1 are deleted, deferring to Ck (Agrawal & Srikant, 1994). Apriori algorithm process is presented below:

```
L1 := {frequent 1-itemsets};
k := 2;  // k represents the pass number
while ( Lk−1 ≠ ∅ ) do
begin
   Ck := New candidates of size k generated from Lk−1.
   forall transactions t ∈ D do
   begin
      Add all ancestors of each item in t to t, removing
         any duplicates.
      Increment the count of all candidates in Ck that
         are contained in t.
   end
   Lk := All candidates in Ck with minimum support.
   k := k + 1;
end
Answer := ⋃k Lk;
```

Figure 2 : Apriori Basics. (Agrawal & Srikant, 1994).

2.3.3 Association Analysis' challenges

As it was previously mentioned, market basket analysis has an effect on the layout of each store. Good and well-organized stores have a positive influence on consumer behaviors and lead to targeted purchases and more profits. Each customer has individual preferences and their daily purchases are differentiated according to their routine. Some of them want to spend less time inside the store, while others prefer to spend more. Products most of the time are placed based on the association rules and depending on the type of the store, the layout is different. The approach of a small local market is completely different from a hypermarket outside of the cities (Al Attal et al., 2018). Subsequently, stores should know their customers in order to make the right decisions about supplies or promotions.

2.4 Customers Segmentation

Customer segmentation is one of the most important parts of customer relationship management (CRM), as it has the goal to divide people – customers into homogeneous groups. The criteria for division vary, but usually relate to purchasing behaviors and patterns. The number of groups depends on the needs and the aim is to

choose the most natural ones. Segmentation is part of descriptive modeling and a powerful tool for business decision-making. (Hand et al., 2001). It is used to apply direct marketing, which is more efficient in contrast to the traditional one, because data is completely integrated to the processes (McCarty & Hastak, 2007).

Concerning retail business analytics and segmentation , stores identify consumers profiles and understand their needs in order to offer the appropriate services. Retailers use many data types to create the groups depending on the occasion. The main are:

- Demographic data
- Geographic data
- Psychographic data
- Attitudinal data
- Sales data
- Behavioral data (Griva et al., 2018)

2.4.1 RFM Model

The most frequently used technique for customer segmentation and direct marketing is the RFM model. Three different measures are combined and compared (recency, frequency and monetary), while recency is considered to be the most important one (Ting We et al., 2010). The main concept of the model is to recognize the good customers and categorize them. Recency shows the interval between the time of the latest purchase of each customer and the present. Frequency represents the number of purchases or generally transactions of the customers in particular periods. Monetary value refers to customers' consumption of money in various periods.

A common way to utilize the analysis is to assign weights to variables and then score each person in the dataset. Weighting is subjective, but generally recency should weigh twice as much as frequency and monetary value. RFM is used for marketing campaigns and helps companies to prioritize their customers based on these three behavioral variables (McCarty & Hastak, 2007). In the retail industry, target customers are identified and the combinations of the metrics can show which ones can visit the store again. Furthermore, stores search for churners, the people who have abandoned the relationship with a service provider, and try to bring them back with

promotions or for people who are not very good consumers and try to improve their scores (Ranjan & Agarwal, 2009) .

Except for the benefits that RFM model offers to businesses, there are some disadvantages. Firstly, the RFM model focuses only on current customers and not the future ones. Predictions are a big part of business analytics, therefore other variables and behaviors should be taken into consideration to make them. On the same wavelength, RFM analysis has a limited number of variables and it is true that there are also numerous characteristics that affect consumer behaviors. Lastly, the discussed model ignores the 1-1-1 type of customers. These customers buy once in a small period, placing small orders. RFM shows the importance only to the best people, who buy often, spending a lot of money, but they do not represent the biggest percentage of the total sales. As a result, low scoring consumers may have the greatest untapped potential (Ting We et al., 2010).

2.4.2 Clustering

Clustering is the assignment of observations or data items into groups based on patterns and similarities .It is a big part of unsupervised machine learning and a key element of business making. In case of unlabeled data, the problem is that they are not grouped in a meaningful way and they do not give the necessary conclusions. The goal is to obtain data points in the same groups as similar as possible and reduce the uncertainty in every data-driven situation.

Pattern representation is the first step of cluster analysis and clarifies the number of patterns, the type and the scale of the available features. Then feature selection follows, which is the process of identifying the most effective subset to use in clustering. Along with selection, is the extraction, where one or more transformations are taking place in order to extract new useful ones. Next part of the process is the pattern proximity. Pattern proximity is measured with aid of distance functions and finds similarities and differences between patterns. The classic and one of the simplest measures is the Euclidean distance. The grouping step can be performed in different ways. Clustering can be hard, where a part of the data is assigned to groups, or fuzzy, where each pattern has a variable degree of membership in each of the output clusters. Final steps of the analysis are data abstraction for

detailed description of the clusters and cluster validity in order to evaluate the output. (A. Jain et al., 1999)

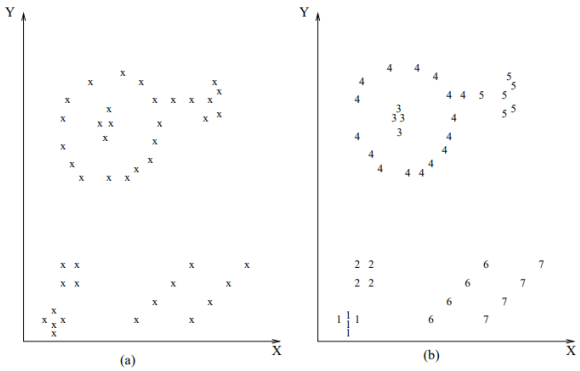


Figure 1. Data clustering.

Figure 3: Data Clustering (A. Jain et al., 1999)

2.4.3 K-means algorithm

Clustering data and segmentation can be achieved using many algorithms and techniques. K- means algorithm is one standard way to group points of data and extract conclusions about them. The algorithm is iterative and considers the parameters and the number of every problem as inputs and separates data into a specific number of clusters in order to achieve high intra-cluster similarity. Before each iteration the value of the centroids is computed and the data points are moved among the clusters depending on these calculations (Christy et al., 2018).

Generally, the first part of K means concerns the partition of the items into K clusters. Elbow curve, a heuristic method, helps the final determination. Secondly all points are checked and assigned to clusters, based on their distance from the nearest centroid. As it was previously mentioned, the centroid of the receiving or losing cluster is calculated again until no more reassigning (Cheng & Chen, 2009).

RFM analysis and clustering can be combined through the K-Means algorithm. The results are very helpful for businesses, especially in the retail industry, and can group customers to interesting groups like loyal, seasonal, churners or even lost ones. The values of the RFM table are the objects of the clusters, therefore the Euclidian distance is calculated among them. The main groups of customers are based on the amount generated with recent transactions or the amount generated in frequent

transactions. This combination of two analyses can achieve effective customer segmentation and organize marketing campaigns to establish a strong relationship between the business and the customers (Anitha & Patil, 2019) .

2.5 Classification

Market basket analysis, RFM table, and clustering are some very important techniques to identify patterns, learn the current customers and assign them to groups. All these can offer the retailers the flexibility and make data-driven management decisions in order to get the competitive advantage. Retailers gather a lot of information, but they do not know which of them are actually important and single out the "really good customers" or the bad ones. Generally good customers are labeled data, who are not in data classes, so classification is the form of data analysis technique to predict them and create the right models to categorize them.

Classification is a supervised learning task, where a model, also called classifier, is used to predict a class. The prediction is categorical, such as "yes" or "no" and "good" or "bad". The basic approach of classification consists of two steps, the learning and classification one. In the learning step, the selected algorithm is learning from a training set and its associated class labels, which are a subset of the initial dataset. The second step concerns the classification, where another independent subset is used, that does not participate in building the classifier. The associated labels of the test set are predicted based on the previous learning process. Test set is used because the classifier tends to overfit the data and is not reliable. Finally, a very important part of the classification is the accuracy of the model, which is the percentage of test set points that are correctly classified by the classifier (Han et al., 2012).

2.5.1 Decision Tree

The decision tree classifier (DTC) is one of the main algorithms and the key idea is to break up one complex situation to simpler ones and finally obtain the final most effective combination of these smaller decisions. DTC objectives are to correctly classify a big percentage of the training set with high accuracy and create an easily updatable model as the training set becomes bigger. Trees' structure has to be simple and follow the appropriate tasks.

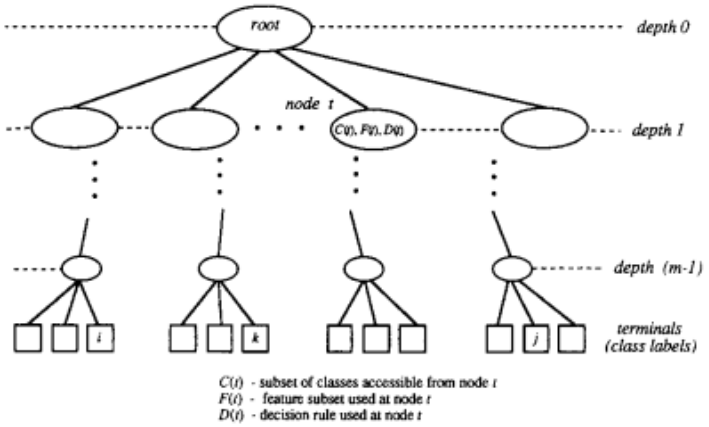


Fig. 1. Example of a general decision tree.

Figure 5: Example of general decision tree (*Safavian & Landgrebe, 1991)*

The common characteristics of the structure are minimum error rate, min-max path length, minimum number of nodes in the tree, minimum expected path length, and maximum average mutual information gain (Safavian & Landgrebe, 1991). Each internal node of the tree represents a test on an attribute, each branch shows the result of this test, while each leaf node, also called terminal node, is the label of a class. The first node called root node is vital and determines the shape of the DTC (Han et al., 2012). Furthermore, information gain is simply how much information can be obtained from a variable, by observing another one and helps to the feature selection.

Typically, decision trees are getting bigger at least to overfit the training set. Then tree pruning follows to reduce overfitting and a test set is used to evaluate the initial decision rules. Every time a rule is removed the connected branch node is replaced with a leaf node. This process is more acceptable than stopping criterion, because more possible combinations are checked. Moreover, pruning can be automated with the aid of various techniques or manually. By implanting it, the size of the tree is obviously reduced, but the quality of the initial decision rules is not affected. The general goal is to create a robust and accurate decision tree (Myles et al., 2004).

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition, *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split-point* or *splitting subset*.

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Figure 5:Basic algorithm for inducing a decision tree from training tuples (Han et al., 2012)

2.5.2 Random Forest

Random forest classifier is a combination of trees, where every classifier is created from a random sample vector. Each individual tree represents a classification vote for the most popular class. The concept is that some features of the dataset are selected randomly or even a combination of them to create trees. From the initial training set with size N, a new training dataset is created by randomly drawing with replacement N examples. This method is called Bagging and is responsible for the feature selection and the combinations. All the examples are classified by taking the most voted class in the forest (Breiman 1999).

Two parameters are mandatory in order to create a random forest. The first one is the number of decision trees (Ntrees). Generally, if it is computationally efficient and does not overfit, there is no limitation to the number of the generated trees. 500 is commonly used as a default value, but actually a lot of variations are encountered like 100, 1000 or even 5000 trees. The first suggestion is the most famous for Ntree, when using RF classifiers. The second parameter is the number of variables to be selected and tested when growing the trees, to achieve the best split

(Mtry). Mtry is mostly defined as the square root of the number of inputs (Belgiu & Drăguţ, 2016).

Individual classifiers of the random forest determine the accuracy of the random forest. For strong RF classifiers, correlation of the individual ones should be at low levels. They are ideal for large databases and since they are not sensitive to the number of selected attributes, they are fast and good internal estimators of variable importance (Han et al., 2012).

2.5.3 Naïve Bayes

Naïve Bayes is a powerful classifier, fits easily to data and its interpretation is clear. For every class variable Y that is to be predicted, there predictors x and the naïve Bayes classifier has the form :

$$P(Y = y \mid x) \propto P(Y = y)P(x \mid Y = y) = P(Y = y) \prod_{j=1}^{d} P(x_j \mid Y = y)$$ (Ye, 2003)

The general concept of the Bayesian classifiers is to assign the most likely class to an example along with the description of its feature vector . In Naïve Bayes one every feature represents an independent class. Even though this assumption is unrealistic, this classifier is effective and the results in practice are successful. This success is due to the fact that the classification error is not necessarily related with the quality of the fit to a probability distribution. Along with optimality of the zero-one loss, Naïve bayes does not require much training data and handles continuous and discrete data as well (Rish, 2001).

2.6 Online Retailing at FMCG sector

Shopping was always one of the most important parts of everyone's routine and online retailing is not something but developed a lot during the pandemic. Prior COVID-19 it was obvious that there is a mass shift to Internet shopping, but it was not it was not a health reason to avoid physical presence in the store. People changed their shopping habits a lot and tried to find alternative and safe ways to procure the necessary things. These changes caused reclassifications to the retailers, who searched for new types of markets and turned to home delivery (Bitterman & Hess, 2021)

In comparison with physical stores, online ones offer various benefits. The main one concerns time saving as there are no waiting lines or traveling. Moreover, clients can access them whenever they want from everywhere with any constraints. The information provided for the products is a lot and there is no time limitation for observing and reading them. In contrast to traditional shopping, customers can compare in real time prices, features or any other field they want with the aid of the numerous online tools that they are given. Finally, another important advantage for the customers is that the transaction costs and generally the prices are most of the time lower at online retailing, which makes it more attractive (Moshrefjavadi et al., 2012).

Apart from the advantages for the clients, online shopping can be also beneficial for the professionals. Recent years, the biggest retailers started to combine the classic store model with the home delivery one to help their customers and be more versatile. This extension gave them the opportunity to expand their network and deliver products to people, who did not have access to their stores. Therefore, they expanded their customer database and they followed the norm of digital transformation. This model is called multichannel and it's the first evolution of physical retail.

2.7 Dark Stores

As already mentioned, the pandemic has changed the situation in many areas. Fast moving consumer goods area was one of the most affected. A new way of store was developed and shopping habits were differentiated, because the demand increased greatly and safety health barriers blocked what is known so far concerning shopping. New stores are called Dark Stores and are similar to typical supermarkets or smaller shops , but they are not open to the public. Online delivery is the only option and delivery "pickers" or proxy shoppers from external companies execute numerous digital orders (Bitterman & Hess, 2021).

This type cannot be compared with traditional shopping, because the format is completely different. For example, large FMCG chains locate their stores outside the center of the towns. As a result, shipping times are bigger. Dark stores are inside the cities and offer swift deliveries with minimal transportation expenses. What is more, the fact that this type of store is not accessible to the public, makes it a perfect place

for product maintenance. Temperatures are at the right level, while the goods remain untouched until the delivery (Shaleva, 2020).

For many retailers, dark stores are considered to be the only way to do successful business and is probably one of the best ways to compete with the big chains. It is an undeniable fact that this type of project requires adjustments. Completely new businesses need the mandatory infrastructure and organize their whole supply chain on this. The changes do not concern only a good online network and a modern delivery strategy. Dark stores strategy is data-driven and includes automated processes as much as possible. The concept of these new stores can improve customer experience and give new opportunities to the retailers (Morgan, 2020).

First important part of dark stores is the CMS Systems, which update in real time the inventories and help pickers to collect the right products in time. Moreover, content management systems are a good tool for the customers, because they are always updated about the preferable products. ERP systems play the same role as in other businesses and plan the deliveries. One of the promises that dark stores give is the quick and instant delivery. Therefore, the routes of the distributors must be studied in detail. Finally, as there is no personal contact between the retailers and the customers and the ways of communication are only digital, CRM-systems are designed differently (Shaleva, 2020).

The majority of dark stores are smaller than the traditional markets and do not have the same equipment. As a result, analytics and generally machine learning are a possible way to help them to organize their daily routine and have the right products at the right time. All the discussed branches of machine learning can be proved extremely helpful for this new retailing style. Association Rules can be extracted from the transactions and use the patterns for promotions and effective supply. Clustering based on RFM analysis can lead to productive customer segmentation. The segmentation can also have an effect on the delivery process and priorities can be depended on the metrics of the table. On the other hand, classification is more demanding, because it mostly answers a question about a target variable. The technical analyses have similarities with the other retail styles, but the differences have to do with business decisions.

Coronavirus has changed consumers' behaviors and generally there is a new era in the shopping sector. Regardless of the power of big chains and the habits that have been established, dark stores are not a trend only for the pandemic period. They offer numerous benefits for both customers and retailers and old school ways of thinking should be reconsidered. The need for physical space will shift and more and more new technologies will enter the dark store concept.

3. *Problem Analysis*

As in regular stores in the FMCG sector, dark stores need the right analytical tools to extract vital business decisions and organize successful campaigns. As already mentioned, strategies of this new retailing style are different and do not follow the existing path. People who prefer them instead of a physical shop, maybe have shopping behaviors based on specific patterns or the fact that they shop online and get the good only home delivered, is an extra reason for changes.

"Dark retailers" want to have a 360-degree view of each customer and create an optimized and automated processes. Relied on the general concepts of dark stores, all the analyses can be executed in real time. The base can be the historical data and then perform the same tasks on current and possible future users. The identified problem is that a dark store should find ways to be differentiated and take the competitive advantage against other similar shops or even shops, which combine both ways of retailing. The issues to be addressed in order to achieve that concern the identification of popular patterns among the transactions of dark store, of different groups of customers and of the important features that create a "good customer". These tasks may have an extension to other important activities of the retailer. A well implemented analytical process can help to optimize supply and demand or adjust their prices instantly and maximize profits.

It is therefore concluded that as there is no physical presence in the dark stores, the main guide to achieve the desired results is the total of the daily online receipts data. The analysis of them can answer a lot of questions about the customers and their preferences. A sample of customers along with their transactions can give the first results and explain the processes of market basket analysis, clustering and classification in a dark store environment

Conclusively the problem identified, concerns the search for useful business information for a dark store, through its receipt data and the definition of the most important. In combination with the above, beneficial for the dark store would be the utilization of general or demographic features of the customers. These two sources make a good base for creating analytical models. The dark store that is going to be

analyzed is PockeeMart, which is based in Thessaloniki and is active in the FMCG sector.

4. Research Review

4.1 Data Collection

For a successful analysis, the most important issue is the sample. Data from the months of October,November,December and January have been extracted from the PockeeMart's database together with the characteristics of the customers. The selection is random and it is completely approved from all management levels, in line with all the existing provisions.

The first sample dataset consists of the PockeeMart's transactions for the aforementioned months and all the details about them. Specifically, it contains the receipt id, product id, Products Cost and Quantity of the products. The second one is about the customers of the dark stores. The available information concerns their age, gender, device of usage. Further features about the users result from the analysis. Moreover, a supportive file is being used, which contains product details. This is very important, because the analysis is being executed at category level and at the product one. As already discussed, the data has been extracted from the database of the dark store in xlsx format and has been stored locally. More details about the data are described with aid of the analytical tool.

	Order ID	Product ID	Product Cost	Quantity
2	2659	7765	€ 3.88	1
3	2659	11103	€ 0.94	1
4	2701	7765	€ 3.88	1
5	2705	11101	€ 0.86	1
6	2705	11103	€ 0.94	1
7	2710	7765	€ 3.88	1
8	2711	18983	€ 1.31	1
9	2751	3556	€ 0.87	1
10	2751	5491	€ 0.74	1
11	2751	9745	€ 1.16	1
12	2752	3457	€ 0.85	1
13	2752	3555	€ 0.96	1
14	2753	7441	€ 1.65	1
15	2754	5491	€ 0.74	1
16	2760	757	€ 1.37	1
17	2760	1923	€ 1.37	1
18	2760	1955	€ 1.36	1

Screenshot 1: Data Snapshot (PockeeMart_receipts.xlsx)

	A	B	C	D	E	F
1	User ID	Gender	Age	Age Groups	Activation Device	
2	126757	Male	62	56 (+)	PC	
3	1550314	Male	67	56 (+)	Android	
4	291	Female	61	56 (+)	Android	
5	50465	Female	73	56 (+)	iOS	
6	218542	Female	60	56 (+)	Android	
7	82044	Male	56	56 (+)	Android	
8	94406	Female	59	56 (+)	Android	
9	125563	Male	60	56 (+)	PC	
10	707140	Male	56	56 (+)	Android	
11	3228	Male	46	46-55	Android	
12	27595	Female	48	46-55	Android	

Screenshot 2: User details Data Snapshot(PockeeMart_user_details.xlsx)

4.2 Design Approach

The sample is used to create a data-driven environment for the dark store and help it manipulate daily situations through business analytics. Every aspect of the data is important and can be proved important and each feature affects the daily transactions, especially for this different type of store.

Firstly, data preprocessing and data cleaning take place, because missing values and other abnormalities distort the results of every analysis. Therefore, handling this type of data is the first move. Then, exploratory data analysis takes place to better understand the data and locate the key points. These two steps are almost the same for every kind of analytical process and they can be characterized as mandatory.

After them, the transaction data is transformed and it is prepared for the association analysis. The appropriate form of the data consists of binary values, which are the right type for the application of the Apriori algorithm and the extraction of the association rules. These rules show which categories are purchased together and the number of famous baskets in a dark store. The three main metrics of the association rules (lift, support, confidence) are applied and evaluate their "power".

It is true that every transaction belongs to a specific customer. As a result, this dataset is also used for clustering. Important features for this are the receipt date, the ids of users and the amount of money they spent on PockeeMart. By combining them, an RFM table is created, which gives to the customers 3 new labels about their shopping behaviors. The outcome of the RFM table, after some obligatory transformations, becomes the input for the cluster analysis. K-means algorithm is used and each customer is assigned into one cluster based on his metrics.

Classification is the last process and needs both sources. The important features are selected from the demographics dataset along with the outcome from previous analyses. Three different algorithms run for this data and they are evaluated in order to distinguish the possible churners of PockeeMart . All the results are visualized in many ways for better understanding

4.3 Technical Overview

The programming language to execute the data analysis for this project is Python, because it is highly productive and completely compatible. The whole process is executed in Jupyter notebooks with Visual Code. For a complete analysis various machine learning and graphics libraries are used. The principal are pandas, NumPy and SciKitLearn, while the supportive are seaborn and matplotlib.

5. *Working Experience*

5.1 Company Overview

Pockee was founded in 2014 as a startup company building technologies around analytics in the FMCG sector. The company is active on three different business levels.

The first level is the initial project of Pockee and consists of an application, which can be characterized as a shopping assistant for every consumer. Pockee app includes all the offers and the promos in the biggest chains of the country with all the necessary information about them. More specifically, consumers are informed about the current retailing promos , as well as the amount of their discount. As a result, they make a good preparation before visiting the respective store. Pockee app has another extension. After visiting the store, consumers have the opportunity to scan their receipt and collect points as a loyalty reward. These points can be transformed to cashback and credited to users' accounts. What is more, Pockee users have the option to redeem coupons through the app, as there are collaborations with many leading industry companies.

The second level refers to business to business (B2B) services. Pockee has built an end-to-end retail platform, which gives to companies and retailers access to data and insights analytics about offers or promos. Business intelligence reports are embedded to the platform and make the competition comparison easier and data oriented. The content of the reports can be filtered in many ways and separate the promos per category, company, brand and many other options.

Pockee’s latest business activity is inextricably linked to the analysis that follows, as it concerns the dark store "PockeeMart". PockeeMart started to operate in 2021 and is based at Thessaloniki. Customers prefer the dark stores as already mentioned because of the situation created by the pandemic.

5.2 Position Description

The title of the position is Business Intelligence Analyst and is a main member of the data team. The analyst is responsible for handling the data and extracting useful information from them. The aim is to increase productivity and improve efficiency through data from many different sources. Moreover, the BI analyst is capable of communicating with customers or other people internally and identifies opportunities for improvements by spotting data trends. In order to perform all these tasks, a combination of business and technical background is required. For this reason, knowledge of many analytical tools, data modeling, critical thinking, decision making and reporting are vital skills. Due to the dominance of technology nowadays more and more companies are hiring BI analysts, thus trying to get rid of subjectivity and turn to data-driven decisions.

5.3 Responsibilities

At “Pockee” the responsibilities of the business intelligence analyst vary and are involved in all three business levels. The analyst reviews and validates the promo data through the application, searching for interesting patterns in order to create basket analytics for future usage. The analyst processes the scanned receipts and tries to extract information from them. A very interesting part of that is the search for the effects of different offers and promos on consumer behaviors. On the same wavelength are the duties of the analyst at the market intelligence platform. Apart from learning and understanding the data landscape, the analyst is responsible for the regular communication with the cooperating companies. Successful communication is achieved by sending reports on a monthly or quarterly basis. These reports are actually a year-to-date overview and give to the receiver an extra tool to organize the marketing strategy. The last responsibilities deal with the dark store of the company. Reporting and creating insightful dashboards, mainly about descriptive analytics, are the main ones. Supportive tasks are about predictive analytics or customers behaviors and popular patterns.

6. *Analysis*

6.1 Sample Description

As already mentioned, PockeeMart is the dark store that is analyzed through its customer data. More specifically, a sample of 990 receipts has been collected from 510 unique customers for the period from October to January. Each receipt has 5 features and a unique key-id. The second dataset contains customers' characteristics and has 3 features along with the ID.

Order ID	Unique receipt number
User ID	Unique customer number
Order Date	Date of the transaction
Product ID	Unique product number
Quantity	Number of products
Total Cost	Cost of purchased product

Matrix 1: Receipt key and features

User ID	Unique customer number
Gender	Gender of the customer
Age	Age of the customer
Activation Device	Customer device for orders

Matrix 2 : Receipt key and features

6.2 Association Rules

The first part of the analysis is the extraction of association rules for the dark store. As there are a lot of similar products, a supportive dataset is imported which assigns a category to every product. This transformation creates larger correlations within the baskets and improves rules' "power". The necessary features for this part of the analysis are the category id and the order id.

The first action is a join, which combines categories with products. The new table has fewer lines as many products are grouped, but it is more maneuverable.

```
receipts_category = pd.merge(products_category,receipts[['Order ID','Product ID','Quantity']],on='Product ID', how='right')
receipts_category
```

	Product ID	Category	Order ID	Quantity
0	77684	285.0	3018	1
1	77526	714.0	3016	1
2	72278	714.0	3016	1
3	72257	714.0	3016	1

Snapshot 1 : Tables Merge (Receipts & Category)

The most important part of the association analysis' preprocessing is the one-hot encoding. One-hot encoding is a transformation, with which raw data receive binary values. As a result, it is easier to understand which categories are in the same basket. If one of them is inside the basket is scored as 1, if not as 0.

```
# Initialize and fit the transaction encoder
online_encoder = mlxtend.preprocessing.TransactionEncoder()
online_encoder_array = online_encoder.fit_transform(receipt_category_list)

# Recast the encoded array as a dataframe
online_encoder_df = pd.DataFrame(online_encoder_array, columns=online_encoder.columns_)
online_encoder_df
```

	34.0	35.0	46.0	48.0	49.0	50.0	54.0	184.0	185.0	186.0	...	665.0	685.0	693.0	694.0	698.0	714.0	720.0	721.0	778.0	782.0
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
1	True	False	False	False	True	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False

Snapshot 2 : One-hot encoding

After the encoding, the next step is the implementation of the Apriori algorithm and the identification of the frequent itemsets, which are the base for the extraction of the association rules.

```
frequent_itemsets = apriori(encoded_transactions_df, min_support=0.07, use_colnames=True)
frequent_itemsets
```

✓ 0.1s

	support	itemsets
0	0.227593	(34.0)
1	0.093656	(35.0)
2	0.257805	(46.0)
3	0.082578	(48.0)
4	0.324270	(49.0)
...	...	...
103	0.088620	(227.0, 229.0)
104	0.088620	(233.0, 227.0)
105	0.099698	(233.0, 229.0)
106	0.077543	(714.0, 252.0)
107	0.070493	(608.0, 49.0, 46.0)

108 rows × 2 columns

Snapshot 3 : Frequent Itemsets

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
rules
```
✓ 0.1s

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(34.0)	(46.0)	0.227593	0.257805	0.092649	0.407080	1.579024	0.033974	1.251762
1	(46.0)	(34.0)	0.257805	0.227593	0.092649	0.359375	1.579024	0.033974	1.205708
2	(49.0)	(34.0)	0.324270	0.227593	0.125881	0.388199	1.705670	0.052080	1.262513
3	(34.0)	(49.0)	0.227593	0.324270	0.125881	0.553097	1.705670	0.052080	1.512030
4	(34.0)	(197.0)	0.227593	0.267875	0.090634	0.398230	1.486626	0.029668	1.216619
...	...	...	...	...	...	...	...	...	...
101	(608.0, 46.0)	(49.0)	0.118832	0.324270	0.070493	0.593220	1.829403	0.031960	1.661170
102	(49.0, 46.0)	(608.0)	0.135952	0.157100	0.070493	0.518519	3.300570	0.049135	1.750639
103	(608.0)	(49.0, 46.0)	0.157100	0.135952	0.070493	0.448718	3.300570	0.049135	1.567343
104	(49.0)	(608.0, 46.0)	0.324270	0.118832	0.070493	0.217391	1.829403	0.031960	1.125937
105	(46.0)	(608.0, 49.0)	0.257805	0.085599	0.070493	0.273438	3.194393	0.048426	1.258530

106 rows × 9 columns

Snapshot 4 : Association Rules

6.3 Clustering

Clustering is the next analytical process for the dark store. For this part it is important to keep the user id instead of the order id, because the goal is effective customer segmentation.

As every data line refers to the individual products' cost and their quantity, so it is mandatory to aggregate them and calculate the total receipt cost.

```
transcations_agg = users_transcations.groupby(['Order ID', 'User ID','Order Date'], as_index=False)['Total Cost'].sum()
transcations_agg.head()
```
✓ 0.1s

	Order ID	User ID	Order Date	Total Cost
0	2990	750620	2021-10-02	57.84
1	2996	1545278	2021-10-01	79.19
2	2998	1556345	2021-10-02	30.42
3	3003	1558130	2021-10-01	30.98
4	3006	175579	2021-10-01	47.73

Snapshot 5 : Receipts Aggregation

Clustering analysis for PockeeMart is based on a RFM table. As already discussed above, recency is considered the most vital metric of the table. Therefore, time constraints must be implemented. The first refers to the earliest transaction date, the second to the latest and the third to the execution date.

```
transcations_agg['Order Date'].min()
✓ 0.8s
Timestamp('2021-10-01 00:00:00')

transcations_agg['Order Date'].max()
✓ 0.6s
Timestamp('2022-01-20 00:00:00')

today = pd.to_datetime("today")
Timestamp('2022-01-20 19:36:14.000782')
```

Snapshot 6 : Time constraints

The next step before the clustering is the creation of the RFM table, which assigns metrics to every customer of the dark store and gives a first impression about their shopping habits.

```
#RECENCY (R): Days since last purchase
#FREQUENCY (F): Total number of purchases
#MONETARY VALUE (M): Total money this customer spent
rfm_table = transcations_agg.groupby('User ID').agg({'Order Date': lambda x: (today - x.max()).days, 'Order ID': lambda x: len(x), 'Total Cost': lambda x: x.sum()})
rfm_table['Order Date'] = rfm_table['Order Date'].astype(int)
rfm_table.rename(columns={'Order Date': 'recency',
                          'Order ID': 'frequency',
                          'Total Cost': 'monetary_value'}, inplace=True)
rfm_table.head()
✓ 0.2s
```

User ID	recency	frequency	monetary_value
141	79	2	129.10
174	12	7	500.31
203	85	1	79.83
206	4	30	1669.40
213	35	4	134.38

Snapshot 7 : RFM Table

The next step before the clustering is the creation of the RFM table, which assigns metrics to every customer of the dark store and gives a first impression about their shopping habits.

```
#The data should meet assumptions where the variables are not skewed and have the same mean and variance.
# Use square root transformation to the data.
customers_fix = pd.DataFrame()
customers_fix["recency"] = pd.Series(np.sqrt(rfm_table['recency'])).values
customers_fix["frequency"] = pd.Series(np.sqrt(rfm_table['frequency'])).values
customers_fix["monetary_value"] = pd.Series(np.sqrt(rfm_table['monetary_value'])).values
customers_fix.head()
[12] ✓ 0.1s
```

	recency	frequency	monetary_value
0	8.888194	1.414214	11.362218
1	3.464102	2.645751	22.367611
2	9.219544	1.000000	8.934764
3	2.000000	5.477226	40.858292
4	5.916080	2.000000	11.592239

Snapshot 8 : Skewness removal

```
#Normalize Data in order to have mean =  and variance = 1,
# Import library
from sklearn.preprocessing import StandardScaler
# Initialize the Object
scaler = StandardScaler()
# Fit and Transform The Data
scaler.fit(customers_fix)
customers_normalized = scaler.transform(customers_fix)
# Assert that it has mean 0 and variance 1
print(customers_normalized.mean(axis = 0).round(2)) # [0. 0. 0.]
print(customers_normalized.std(axis = 0).round(2)) # [1. 1. 1.]
```

✓ 0.2s

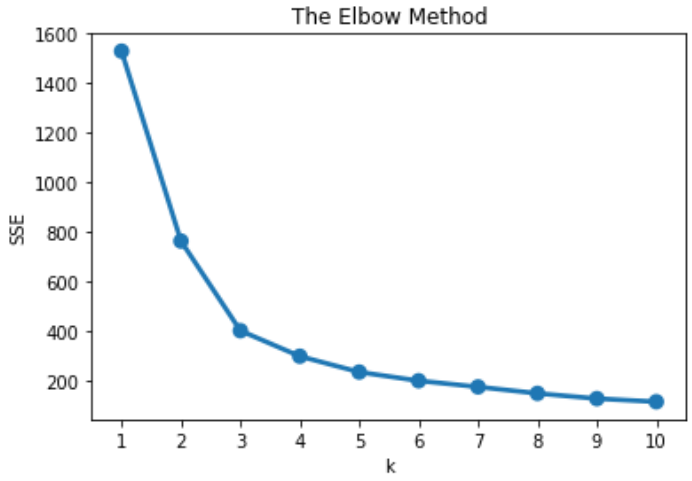
```
[-0. -0.  0.]
[1. 1. 1.]
```

Snapshot 9 : RFM table normalization

After data preparation, K-means algorithm is applied to the data in order to create clusters. The decision on the number of clusters is made by the elbow curve. Specifically, 4 clusters are created and PockeeMart's customers are distributed to them.

```
from sklearn.cluster import KMeans
sse = {}
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(customers_normalized)
    sse[k] = kmeans.inertia_ # SSE to closest cluster centroid
plt.title('The Elbow Method')
plt.xlabel('k')
plt.ylabel('SSE')
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.show()
```

Snapshot 10 : K-Means



Snapshot 11 : Elbow Curve

Each cluster has its own characteristics and the values are a combination of RFM table values according to their shopping behaviors.

```
rfm_table["Cluster"] = model.labels_
rfm_table.groupby('Cluster').agg({
    'recency':'mean',
    'frequency':'mean',
    'monetary_value':['mean']}).round(1)
```

[17] ✓ 0.1s

	recency	frequency	monetary_value
	mean	mean	mean
Cluster			
0	58.6	1.1	27.7
1	16.0	3.7	159.3
2	10.5	1.4	49.8
3	7.1	12.9	666.2

Snapshot 12 : Clusters' Details

```
rfm_table["Cluster"] = model.labels_
rfm_table.head()
```

✓ 0.9s

User ID	recency	frequency	monetary_value	Cluster
141	79	2	129.10	0
174	12	7	500.31	3
203	85	1	79.83	0
206	4	30	1669.40	3
213	35	4	134.38	1

Snapshot 13 : Customers' Assignment to Clusters

6.4 Classification

Classification is the last part of the analytical process for the dark store and tries to give answers to a business problem. It is very important for the dark store to know, which customers will continue to prefer or they will stop doing business with it.

Model comparison is a vital part of the classification, in order to find the most suitable for the PockeeMart's data. First of all, 3 different datasets are combined. Users' transactions, users' details and clustering results. The new DataFrame contains the User ID, the age, the gender, the device, the RFM table metrics from clustering analysis and the total quantity of purchased products. Recency is transformed to a

binary variable and marked as the target one for the classification. Customers who preferred PockeeMart in the last 30 days get the value 0, while customers who visited and bought from the dark store in more than 30 days get the value 1. As a result, customers with recency value over the 30 days are characterized as potential churners.

```
users_rfm = pd.merge(users,rfm_clusters[['User ID','recency','frequency','monetary_value']],on='User ID', how='right')
users_rfm_quantity = pd.merge(users_rfm,users_transcations[['User ID','Quantity']],on='User ID', how='right')
users_final = users_rfm_quantity.groupby(['User ID','Gender','Age','Activation Device','frequency','monetary_value','recency'], as_index=False)['Quantity'].sum()
users_final = users_final.rename(columns={'recency': 'Churn', 'frequency': 'Times', 'monetary_value':'Total Amount'})
users_final.loc[users_final['Churn'] <= 30, 'Churn'] = 0
users_final.loc[users_final['Churn'] > 30, 'Churn'] = 1
users_final.head()
```

✓ 0.3s

	User ID	Gender	Age	Activation Device	Times	Total Amount	Churn	Quantity
0	141	Male	25	Android	2	129.10	1	53
1	174	Female	23	Android	7	500.31	0	239
2	203	Male	41	iOS	1	79.83	1	36
3	206	Male	34	iOS	30	1669.40	0	1022
4	213	Male	43	iOS	4	134.38	1	63

Snapshot 14 : Aggregations for Classification

In the same way as in the clustering analysis, data is normalized. Categorical variables are transformed into binary and the rest are parameterized with a scaler.

```
users_final_copy=pd.get_dummies(users_final,drop_first=True)
users_final_copy.head()
```

✓ 0.1s

	User ID	Age	Times	Total Amount	Churn	Quantity	Gender_Male	Activation Device_PC	Activation Device_iOS
0	141	25	2	129.10	1	53	1	0	0
1	174	23	7	500.31	0	239	0	0	0
2	203	41	1	79.83	1	36	1	0	1
3	206	34	30	1669.40	0	1022	1	0	1
4	213	43	4	134.38	1	63	1	0	1

Snapshot 15 : Categorical Values' transformation

```
from sklearn.preprocessing import MinMaxScaler
features = users_final_copy.columns.values
scaler = MinMaxScaler(feature_range = (0,1))
scaler.fit(users_final_copy)
users_final_copy = pd.DataFrame(scaler.transform(users_final_copy))
users_final_copy.columns = features
users_final_copy.head()
```

✓ 0.1s

	User ID	Age	Times	Total Amount	Churn	Quantity	Gender_Male	Activation Device_PC	Activation Device_iOS
0	0.000000	0.109091	0.028571	0.075589	1.0	0.050930	1.0	0.0	0.0
1	0.000020	0.072727	0.171429	0.298371	0.0	0.233105	0.0	0.0	0.0
2	0.000038	0.400000	0.000000	0.046020	1.0	0.034280	1.0	0.0	1.0
3	0.000040	0.272727	0.828571	1.000000	0.0	1.000000	1.0	0.0	1.0
4	0.000044	0.436364	0.085714	0.078758	1.0	0.060725	1.0	0.0	1.0

Snapshot 16 : Values' normalization

The main concept of the classification requires a target variable and the features. Historical data are split to train and test set in order to build a model to

predict feature records. The separation of the data is the same for all the different models, which are applied.

```
X =  users_final_copy.drop(['Churn'], axis=1)
y = users_final_copy['Churn']


from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=41)
```

Snapshot 17 : Target variable selection & Train and Test set separation

Three different algorithms are applied to the PockeeMarts's data. The first one is a decision tree, which is optimized as much as possible with pruning technique. The second is the random forest and the last one is the Naïve Bayes algorithm. The selected model is that with the highest accuracy metric.

```
#Decision Tree
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
# Create Decision Tree classifer object
decision_tree = DecisionTreeClassifier()

# Train Decision Tree Classifer
decision_tree = decision_tree.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = decision_tree.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

Snapshot 18 : Decision Tree process

```
#Optimize Decision Tree
decision_tree_optimal = DecisionTreeClassifier(criterion="entropy", max_depth=3)

decision_tree_optimal = decision_tree_optimal.fit(X_train,y_train)

y_pred = decision_tree_optimal.predict(X_test)

decision_tree_opt_acc = metrics.accuracy_score(y_test, y_pred)
decision_tree_opt_acc
```

Snapshot 19 : Optimal Decision Tree process

```
#Naive Bayes
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()

#Train the model
gnb.fit(X_train, y_train)

#Predict
y_pred = gnb.predict(X_test)

#Model Accuracy
gnb_acc = metrics.accuracy_score(y_test, y_pred)
gnb_acc
```

Snapshot 20 : Random forest process

```
#random forest
from sklearn.ensemble import RandomForestClassifier
rf_c=RandomForestClassifier()
random_forest=RandomForestClassifier(n_estimators=100)

#Train the model
random_forest.fit(X_train,y_train)

y_pred=random_forest.predict(X_test)

# Model Accuracy,
random_forest_acc = metrics.accuracy_score(y_test, y_pred)
random_forest_acc
```

Snapshot 21: Naïve Bayes process

7. Findings

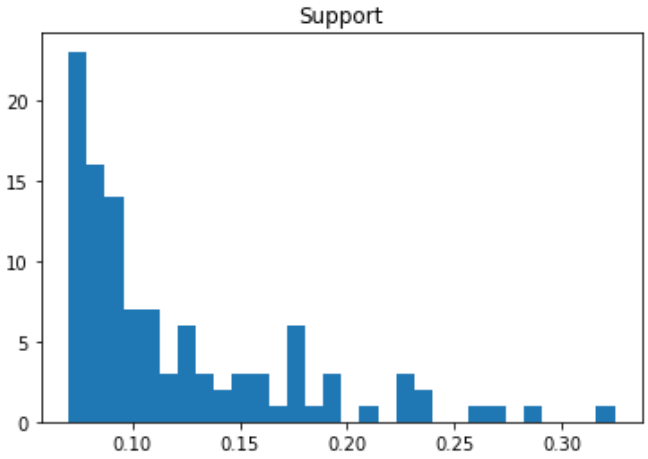
It is an undeniable fact that the general management and the marketing strategy of dark stores is differentiated from a traditional one. Data-oriented decisions are the daily guide and all the techniques explained above are part of the process.

From the association rules analysis, 106 unique rules are extracted with minimum support set at 0.07. This means that the results describe rules that appear at the 7% or more of the total transactions. In retail shops with physical presence association rules are helpful, because they contribute to the stores' design and layout. In the dark stores era, the approach is different and business moves concern mostly the delivery services, the online site browsing and the supply chain management.

Antecedents	Consequents	Support	Confidence	Lift
34.0	**46.0**	0.092649	0.407080	1.579024
608.0	**49.0 - 46.0**	0.070493	0.448718	3.300570
714.0	**252.0**	0.077542	0.418478	2.175648
205.0	**49.0**	0.111782	0.569230	1.755422
222.0	**221.0**	0.082578	0.436170	2.433241
229.0	**233.0**	0.099697	0.426724	1.507961

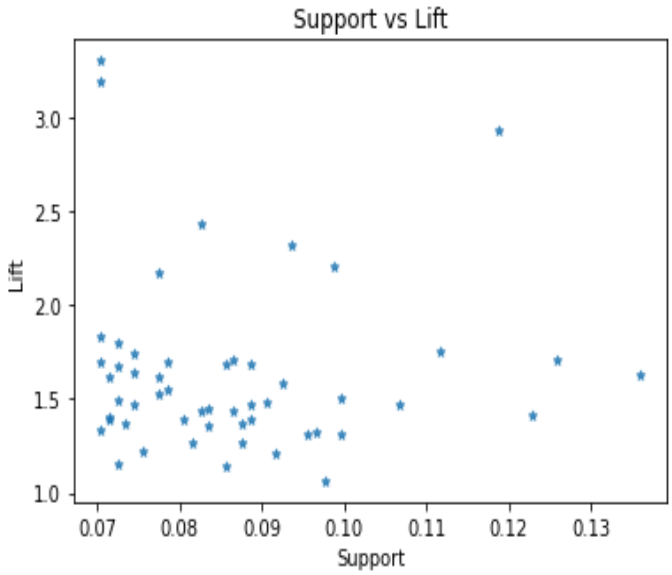
Matrix 3 : Important Association Rules

The rules above are some important results from the market basket analysis. If a customer buys products from category 205.0 for example, then buys the category 49.0. Total transactions contain this specific combination in percentage of 7.7%, with 57.0% confidence. Lift metric is measured at 1.75, therefore there is positive correlation between the two categories and the rule can be characterized as strong.



Visualization 1 : Rules' Support

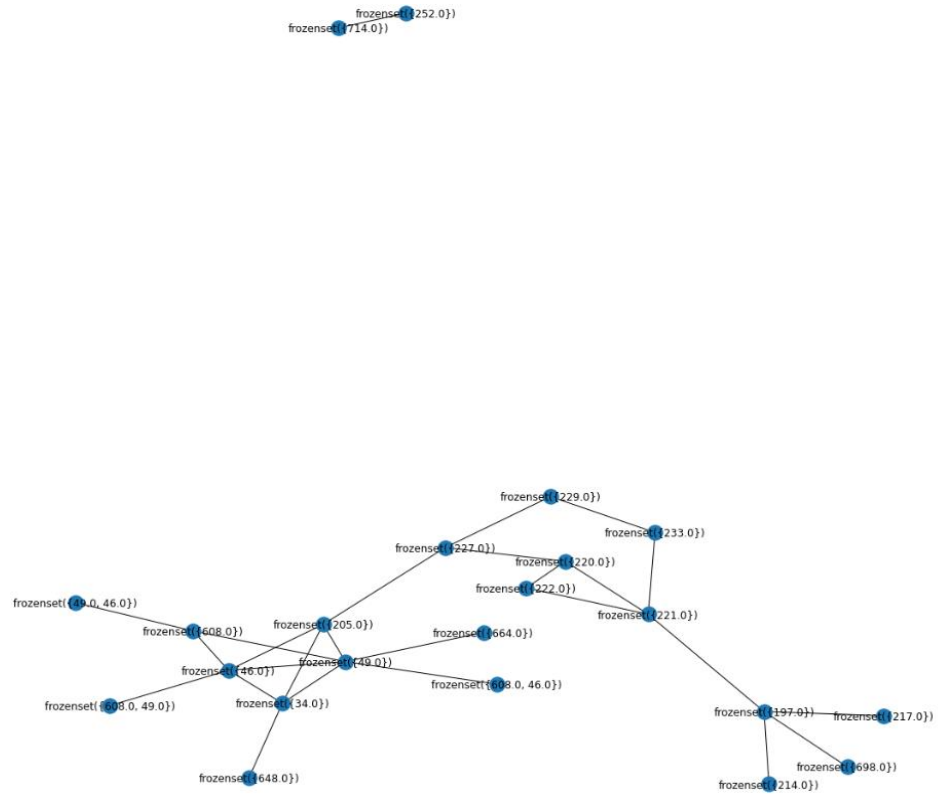
The majority of the rules are found in less than 10%, but there are others that have a high support metric. The fact that there are many different categories, creates variety and gives to the dark stores many options to choose the best combinations to promote. In order to have a more complete view about the results, the support metric is examined in parallel with the lift metric.



Visualization 2: Support vs Lift

The rules discussed before have a lift metric measured between 1.2 and 1.8, which means that the likelihood of buying these categories is 1.2 – 1.8 times more than the likelihood of just buying the antecedent. There are a group of rules which are much "stronger" and are evaluated as well with different criteria. Building a graph is an

efficient way to check and visualize the associations. For clearer results, rules with lift value over 1.5 are presented.



Visualization 3 : Network association graph

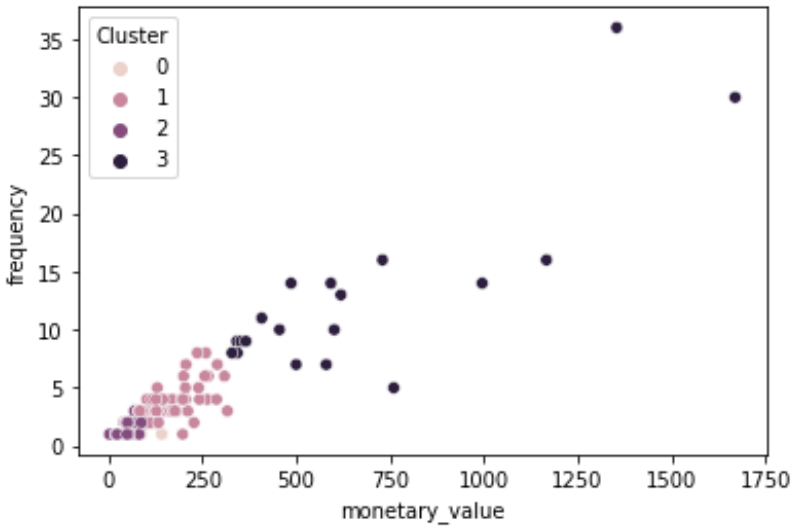
Clustering analysis is also vital, because it helps the dark store to distinguish the customer groups relied on their habits and not on subjectivities. The RFM clustering for PockeeMart has created 4 clusters. The table below presents the means values of the cluster analysis and is the base for the customer segmentation.

	Recency	**Frequency**	**Monetary**
Cluster 0	58.6	1.1	27.7
Cluster 1	16.0	3.7	159.3
Cluster 2	10.5	1.4	49.8
Cluster 3	7.1	12.9	666.2

Matrix 4 : Cluster Analysis results

Each cluster requires different manipulations. The dark store can evaluate customer retention by identifying loyal users and churners or even increase profits, by choosing targeted promotions for right customers without neglecting the

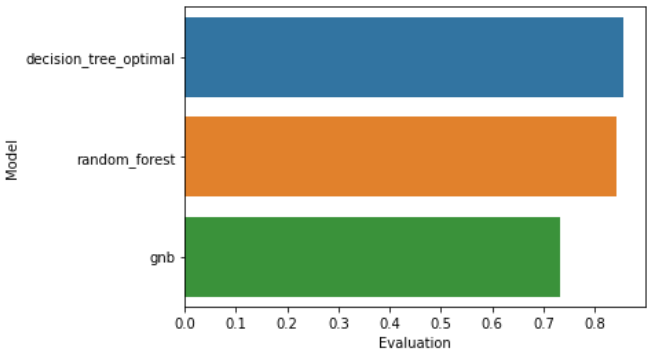
association rules. As a result, the two analyses can work effectively together and offer to the dark store data-oriented options for its marketing strategy.



Visualization 4 : Frequency vs Monetary

The visualization above shows how the cluster labels are distributed according to their frequency and monetary value. Since recency is presented as the most important measure it is examined differently and is trying to be predicted during classification.

Among the three models that have been applied to the combined data from the 3 DataFrames, decision tree optimal is considered as the best. It has the highest accuracy metric, measured at 85.6%, while random forest is really close with accuracy measured at 84.31%. The last algorithm, Naïve Bayes, is the less effective for the PockeeMart's model and the accuracy level is 73.20%



Visualization 5 : Classification Models Comparison

8. Discussion

All the presented analytical processes have a business impact on the dark store. Due to the nature of the store, the results should be utilized in a different way rather than a typical one. The classic concept for market basket analysis refers for example to the layout of the shop and the and the placement of famous itemset nearby. Concerning clustering, retailers use it in order to "learn" their customers, identify their target audience and create in-store, leaflet or internet promotions. Last but not least, retention in traditional stores is estimated and predicted in groups, as most of customer data is not retained.

8.1 Market Basket Analysis – Business Value

The trading and shopping environment of a dark store is a website or an application. Association analysis can be the base for the recommendation system of the dark store. Popular consequents can be recommended to the users and increase the basket size.

Recommendation systems relied on individual historical data, the new or the nearly new users have not placed enough orders. Association rules can be a reliable beginning to inform users for possible combinations. For example, every time a PockeeMart's user adds a product from category "34", a notification can inform him if he wants to add a product from category "46". This recommendation is one of the "strong" rules, which is extracted with the aid of Apriori algorithm. Furthermore, the rules can be a guide for the supply of products. Most of the time, dark stores are small places inside city centers, therefore the supply chain management should be targeted and methodical based on patterns and strict timelines. The study of the rules can show the popularity of certain categories and organize the negotiations with the suppliers. Several itemsets can be included in offers. As a result, unpopular products can be boosted from the combination's power and be part of market baskets.

8.2 Customer Segmentation – Business Value

The competition in the retailing industry is constantly growing and each year and every year more and more options are presented to customers. Dark stores seem to be the latest and require a complete view of their customers. It is known that customers create relationships of trust and loyalty with their retailers and it is mostly difficult to convince them to change their preference.

After the implementation of K-means algorithm, 4 groups have been generated. Each cluster refers to PockeeMart's users with several characteristics in connection with their shopping behavior. Specifically, the type of each cluster is presented below:

Cluster 0 : Refers to customers who preferred the dark store once and then returned to their previous retailer. It is possible that they took advantage of an offer or promotion just to save an amount of money. One potential strategy is to create a new promotional campaign for them and check if they will order one time again. The loyalty should be based on the brand awareness and trust and not on individual situations.

Cluster 1: Refers to "newcomers" with good consuming potential. These customers seem to like the services of the dark and should be handled carefully. They should receive a lot of recommendations based on the association rules and newsletter about everything concerning the dark store. Social media also play vital role and competitions through them are a great way to strengthen the business relationship.

Cluster 2 : Refers to completely new customers, who started to buy products from the dark store. Their habits are not clear and their preferences have not been revealed yet. Promo codes and coupons are a good option in order to encourage them to place more orders, but patience until more data is available may be proved more useful for this cluster.

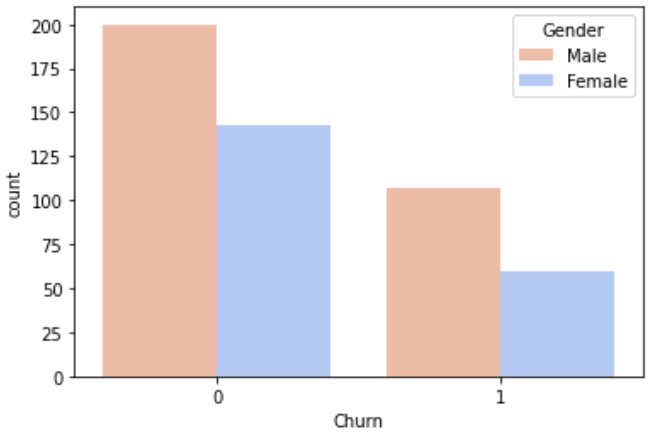
Cluster 3: Refers to the established customers. They prefer the dark store frequently without big time gaps and the amount of money they spend is above the acceptable level. Loyal users should be rewarded and sometimes have special treatment. In this way, they will feel important and the transactions become less impersonal. Delivery priority should be examined for these users and updates on new offers need to be made faster for them.

8.3 Classification – Business Value

Classification process works proactively for the examined dark store. It is an extra tool for the management and marketing team to check the retention and the effect of various campaigns. The time threshold of the one month is representative in order to identify which users can be characterized as churners. The existing historical data has built a base for the future customers. The decision tree classifier will give them a label resulting from their characteristics and their RFM scores.

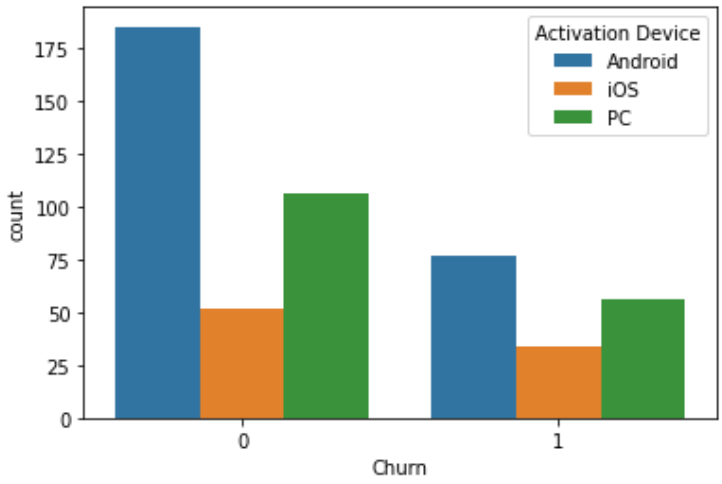
The business objective of the classification is to create an automated tool to categorize new customers. The demographics variables concern the age, the gender and the device of every customer. Customers can be grouped by these characteristics without the clustering results , but this join gives a more complete description about them.

As the company works along with predictive analytics, it has the ability to avoid the churn of a customer. If a new customer after his first few orders is scored with label "1", it is more possible to stop preferring the dark store. As a result, the marketing team has more time to prevent this type of shopping behavior and organize campaigns to increase retention. Potential churners can receive discounts for their favorite products or coupons for their next order. Furthermore, the marketing campaigns who have been affected from classification can become even more personalized.



Visualization 6 : Churn count by Gender

The visualization above presents the churn rate by gender. Specific categories can be promoted for each gender and give them extract incentive to continue to shop from the dark store.



Visualization 7 : Churn count by Device

In the same wavelength, churn rate can be grouped by users' devices. This grouping is very important, because the combination the device's type and the spending score can show really insightful results. More premium devices with high scores may require a different marketing policy than those who have cheaper ones.

The multidimensional data and their flexibility give to the dark store the opportunity to handle numerus situations and offer to its customers quality services. The challenge is to consider which analytical tool and constraints are more suitable for every occasion.

9. Conclusions

According to the analysis above, it is concluded that business analytics have the ability to completely change the strategy of every company. It is true that in the data-oriented era, business decisions are based on real facts and insights. The transformations and the algorithms, which have been implemented to the PockeeMart's multidimensional data , are really important as it is confirmed from the literature review have a huge impact on the business structure.

The pandemic has shaken the retail word and apart from the harmful effect has helped new types of businesses to evolve. During COVID-19, dark stores have been established and have improved customer experience, providing easy access to essential goods. After the normalization of the situation, dark stores will continue to be part of consumers' shopping routine.

9.1 Future Work

Association analysis, customer segmentation and churn prediction with classification techniques will drive the dark store to the initial business-oriented decisions and campaigns. In the future and after the collection of more multidimensional data, a complete recommendation system will be a perfect match for the existing analytical process. A successful engine of this type should be personalized and meet the needs of each customer individually. As already mentioned, the first recommendations are based on association rules. Historical data would help to create more complex variables in order to design the system. One important fact is that the recommendation algorithm should be flexible enough so that it would not be influenced by individual product purchases but by shopping patterns.

10. Bibliography

Al Attal, D., Naser, M., AlBaghli, N., Al Muhaimeed, N., & Al Awadh, S. (2018). *Redesigning a Retail Store Based on Association Rule Mining.* Proceedings Of The International Conference On Industrial Engineering And Operations Management.

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules.* IBM Almaden Research Center.

Anitha, P., & Patil, M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal Of King Saud University - Computer And Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.12.011

Alpaydin, E. (2014). *Introduction to machine learning* (3rd ed.). The MIT Press.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal Of Photogrammetry And Remote Sensing*, *114*, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bitterman, A., & Hess, D. (2021). Going dark: the post-pandemic transformation of the metropolitan retail landscape. *Town Planning Review: Volume 92, Issue 3*, *92*(3), 385-393. https://doi.org/10.3828/tpr.2020.57

Bose, I., & Mahapatra, R. (2001*). Business data mining — a machine learning perspective*. Information & Management, 39(3), 211-225. https://doi.org/10.1016/s0378-7206(01)00091-x

Breiman, L. *Random forests (1999) —Random features.* Technical Report 567, Statistics Department, University of California, Berkeley,

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., & Tishby, N. et al. (2019). *Machine learning and the physical sciences*. DOI:https://doi.org/10.1103/RevModPhys.91.045002

Cheng, C., & Chen, Y. (2009). *Classifying the segmentation of customer value via RFM model and RS theory.* Expert Systems With Applications, 36(3), 4176-4184. https://doi.org/10.1016/j.eswa.2008.04.003

Christy, A., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). *RFM ranking – An effective approach to customer segmentation*. Journal Of King Saud University -

Computer And Information Sciences, 33(10), 1251-1257. https://doi.org/https://doi.org/10.1016/j.jksuci.2018.09.004

Cil, I. (2012*). Consumption universes based supermarket layout through association rule mining and multidimensional scaling*. Expert Systems With Applications, 39(10), 8611-8625. https://doi.org/10.1016/j.eswa.2012.01.192

Dietterich, T. G. (1997). *Machine-Learning Research. AI Magazine, 18*(4), 97. https://doi.org/10.1609/aimag.v18i4.1324

Gavin, M. (2019). *Business Analytics: What It Is & Why It's Important | HBS Online*. Business Insights - Blog. Retrieved from https://online.hbs.edu/blog/post/importance-of-business-analytics.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). *Retail business analytics: Customer visit segmentation using market basket data*. Expert Systems With Applications, *100*, 1-16. https://doi.org/10.1016/j.eswa.2018.01.029

Han, J., Kamber, M., & Pei, J. (2012). *Data mining* (3rd ed.). Elsevier.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. The MIT Press.

JAIN, A., MURTY, M., & FLYNN, P. (1999). *Data Clustering: A Review*. ACM Computing Surveys, 31(3)

Jordan, M., & Mitchell, T. (2015*). Machine learning: Trends, perspectives, and prospect. Science, 349*(6245), 255-260.

Julander, C. (1992). *BASKET ANALYSIS: A NEW WAY OF ANALYSING SCANNER DATA*. International Journal Of Retail & Distribution Management, 20(7). https://doi.org/10.1108/09590559210022362

Kaur, M., & Kang, S. (2016*). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining*. Procedia Computer Science, 85, 78-85. https://doi.org/10.1016/j.procs.2016.05.180

Kotsiantis, S. (2007). *Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249–268*. Citeseerx.ist.psu.edu. Retrieved 6 December 2021, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.9683.

McCarty, J., & Hastak, M. (2007). *Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression*. Journal Of Business Research, 60(6), 656-662. https://doi.org/10.1016/j.jbusres.2006.06.015

Mitchell, T. (1997). *Machine learning*. MacGraw-Hill.

Morgan, B. (2020). *Dark Stores Are The Future Of Post-Pandemic Retail*. Forbes. Retrieved fro,m https://www.forbes.com/sites/blakemorgan/2020/04/25/dark-stores-are-the-future-of-post-pandemic-retail/.

Moshrefjavadi, M., Rezaie Dolatabadi, H., Nourbakhsh, M., Poursaeedi, A., & Asadollahi, A. (2012). *An Analysis of Factors Affecting on Online Shopping Behavior of Consumers.* International Journal Of Marketing Studies, 4(5). https://doi.org/10.5539/ijms.v4n5p81

Myles, A., Feudale, R., Liu, Y., Woody, N., & Brown, S. (2004). *An introduction to decision tree modeling.* Journal Of Chemometrics, *18*(6), 275-285. https://doi.org/10.1002/cem.873

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. T.J. Watson Research Center.

Safavian, S., & Landgrebe, D. (1991). *A survey of decision tree classifier methodology*. IEEE Transactions On Systems, Man, And Cybernetics, 21(3), 660-674. https://doi.org/10.1109/21.97458

Shaleva, O. (2020). *Ensuring socio-economic efficiency of retail in the conditions of crisis on the basis of the dark store format*. Theoretical and empirical scientific research: concept and trends - volume 1. https://doi.org/10.36074/24.07.2020.v1.03

Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its applications ; with 23 tables*. Springer.

Taiwo Oladipupo Ayodele (2010). *Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.)*, ISBN: 978-953-307-034-6, InTech, Available from: http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms

Ting We, J., Yen Lin, S., & Hung Wu, H. (2010). *A review of the application of RFM model*. African Journal Of Business Management, 4, pp. 4199-420.

V. Thanuja, et al.(2011) *Applications of Data Mining in Customer Relationship* Management J. Comp. & Math. Sci. Vol.2 (3), 423-433 (2011) 433

Will Kenton. (2021). *Fast-Moving Consuming Goods (FMCG)* Received from http: https://www.investopedia.com/terms/f/fastmoving-consumer-goods-fmcg.asp

Witten, I., & Frank, E. (2005). *Data mining* (2nd ed., p. 9). Morgan Kaufmann.

Ye, N. (2003). *The Handbook of data mining*. Lawrence Erlbaum Associates.