Instituto Universitário de Lisboa

Department of Information Science and Technology

Study about customer segmentation and application in a real case

Luma Prianca Salman Mendes

Dissertation submitted as partial fulfillment of the requirements
for the degree of Master in Computer Engineering

Supervisor:
Professor Doctor Ana Maria de Almeida, Assistant Professor
ISCTE-IUL

September, 2019

Acknowledgment

Firstly, I want to thank God for always being with me on my walk, protecting me and loving me.

Secondly, I want to thank my supervisor Ana Almeida for all the help, knowledge, patience she had with me, for her perfectionism that motivated me to further improve my thesis and for not letting me give up.

To my mother Samira Mendes, being my biggest motivation, without her I would never get here. Thank you for all the teachings of life, for the education, for the example.

To my father, Henrique Mendes, who always motivated me to take the master's degree and always provided me with all the necessary means for my education and always supported me in the moments I needed most. Thank you for being more than a father, a friend of all time.

To my brother, Henrique Júnior, who is with me always and for who I always want to be the best to always give him the best example.

To my best friend Leovigildo Turé, who I met at this institution and I will take for life, which was one of the people who motivated me to finish the thesis. To my boyfriend Márcio Barros, who in one way or another always encourages me to do my best. My friends, who often took me for a walk so as not to go crazy.

To all who have mentioned my sincere "Thank you".

"Success isn't about the end result, it's about what you learn along the way."

Vera Wang

Resumo

O setor de hospitalidade gera uma enorme variedade de dados que crescem a cada dia, tornando-se fisicamente impossível analisar esses dados manualmente a fim de construir um bom modelo de dados. Um profundo entendimento dos perfis dos atuais clientes permite uma melhor alocação de recursos e leva a uma melhor definição das estratégias de desenvolvimento de produtos e mercados. A divisão dos clientes em grupos semelhantes para ajudar a desenvolver mensagens de marketing mais objetivas e focadas para cada um dos seus segmentos.

Desse modo na presente dissertação são estudados métodos de classificação e segmentação de dados existentes na revisão da literatura. De seguida, procede-se à apresentação de um estudo de um caso real, usando dados pertencentes a Sistemas de Gestão de Propriedade de oito hotéis portugueses, quatro hóteis de cidade e quatro hóteis de resort, este conjunto de dados é composto por quarenta e um atributos, mas, após uma selecção das variáveis com maior poder preditivo, apenas um subconjunto de atributos é utilizado para a modelação dos dados. Em seguida, são aplicados os métodos de classificação e segmentação estudados na revisão de literatura de modo a extrair informação relevante. Os resultados são analisados e discutidos para entender sua adequação ao estudo das características particulares das reservas de hotéis.

Palavras-Chave: Segmentação de Clientes, Mineração de Dados, Gestão da Receita de Hospitalidade, Estudo de Caso.

Abstract

The hospitality industry generates a huge variety of data that grows by the day, becoming incrinsingly difficult to analyse this data manually in order to build a good data model. A thorough understanding of current customer profiles enables better resource allocation and leads to better definition of product and market development strategies. Dividing customers into similar groups to help develop more objective and focused marketing messages for each of the segments. Thus, in the present dissertation methods of classification and segmentation of existing data in the literature review are studied. Then, a real case study is presented, using data from Property Management Systems of eight Portuguese hotels, four city hotels and four resort hotels. This data set consists of forty-one attributes but, after selection of the most predictive variables, only a subset of attributes is used for data modeling. Next, the classification and segmentation methods studied in the literature review are applied for extracting the relevant information. The results are analyzed and discussed to understand their suitability to study the particular characteristics of hotel reservations.

Keywords: Customer Segmentation, Data Mining, Hospitality Revenue Management, Case-Study.

Index

Index of Tables

Index of Figures

List of Abbreviations and Acronyms

ANN – Artificial Neural Network

ARPU - Average Revenue Per User

ATM - Automated Teller Machine

CART – Classification and Regression Trees

CHAID - Chi-squared Automatic Interaction Detection

CT – Classifition Trees

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

DSRM - Design Science Research Methodology

EAP - Exploratory Acquisition of Products

EBBT - Exploratory Buying Behaviours Tendency

EIS - Exploratory Information Seeking

LR – Logistic Regression

LTV - Lifetime Value

NB – Naive Bayesian

PMS – Property Management Systems

RFM - Recency, Frequency and Monetary

SVM – Support Machine Vector

SOM – Self Organizing Maps

1. Chapter 1 – Introduction

1.1.Relevance of theme

It is extremely important for an enterprise to know and study its target customers, at an early stage, understanding what their needs are to match their expectations. A deep understanding of current customers' profiles allows for a better allocation of resources and leads to a better definition of product and market development strategies, avoiding dispersion of the company's objectives, which can then focus efforts on its target segments (Thakur & Mann, 2014).

Customer segmentation consists of dividing existing customers (or potential customers) into differentiated groups. This division is based on a measure of similarity between the customers regarding the company's products or even the information of interests demonstrated by the clients independently of the company. The identification of existence of a product or service that satisfies the needs of the customers so to establish group membership is the major goal of segmentation.

The hospitality industry collects a massive volume of data which makes the process of data analysis a difficult one. Every day, huge amounts of data come into the databases from online platforms, smartphones, etc. A successful mining and knowledge extraction of the data can bring many advantages like product improvement, marketing message focus, seizing of opportunities and increase top-quality revenue. For the treatment of large amounts of data big data clustering algorithms should be used.
In this document, the results from the research made about what methods of data clustering are being used in different areas like banking, health, management, marketing and hospitality as well as the results of a segmentation case study over real data. The main methodologies employed came from the areas of Data Mining.

1.2.Data Mining

There is a lot of knowledge that can be extracted from data that comes from databases and datawarehouses. As an example, analysts use this knowledge to support decisions in their business. The key issue is that of how to extract the most relevant information from data. That's when the Data Mining concept comes along. Data Mining consists of an analytical process designed to exploit large amounts of data (typically related to business, market or scientific research), searching for consistent patterns and/or relationships between variables, and then validating them by applying the detected patterns to new subsets of data. The process consists basically of 3 steps: exploration, model construction and validation (Ziafat & Shakeri, 2014)

Data mining refers to the process that uses a variety of techniques to identify information or decision making knowledge in the database and extracting it in a way that it can be put to use in areas such as decision support, prediction, forecasting and estimation (Kaur & Paul, 2014). Thus, data mining can be considered as the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both (Thakur & Mann, 2014). Discovering

relations that connect variables in a database is crucial for knowledge extraction, that is, the non-trivial extraction of implicit, previously unknown and potentially useful information from data. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data (Kaur & Paul, 2014). In current days, Data Mining is an indispensable tool in business models, making it possible to transform business data into concrete and reliable information to support decision-making processes, translating, in practice, into strategic business advantages.

According to (Thakur & Mann, 2014), Data Mining techniques can be grouped into the following classes: Classification, Estimation, Prediction, Association rules, Clustering and Description. Classification is a process of generalizing the data according to different instances. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. Prediction it is a statement about the way things will happen in the future, often but not always based on experience or knowledge. Association Rules is a rule which implies certain association relationships among a set of objects in a database. Clustering can be considered the most important unsupervised learning problem so, as every other problem of this kind, it deals with finding a pattern in a collection of unlabeled data.

1.3.Customer Segmentation

We could find authors that defend that market segmentation is a set of concepts and models that guides management thinking and leads to new profitable product/service offerings (Smeureanu, Ruxanda, & Badea, 2013). Customer Segmentation is a division of potential customers into different groups based on similarities found among customers.

There are three main approaches to customer segmentation. The first, Apriori segmentation, is the simpler approach, based on publicly available characteristics, such as the size of the company, to classify customers into distinct groups within a market. This approach is not always reliable, since companies of the same size, in the same industry may have different needs. This idea brought forth the Needs-based segmentation, which is based on needs presented by customers regarding a specific product or service offered by the company. These needs are discovered through primary market research. Lastly, Value-based segmentation, is an approach that differentiates customers by their economic value.

Usually, most companies already have a market study and, thus, already understand what the possible segments of clients they have, at least, for the more profitable customers. Even so, it is necessary and indispensable to develop hypotheses and variables of customer segments and validate them through research processes, since new segments might be indentified, whose costumers, when targeted, may bring higher gains. This happens in Needs-based and Value-based segmentation approaches, where it is necessary to establish clear hypotheses to define the research to be done. Assumptions must consider the characteristics of the customer so that customers can be divided into groups according to their needs and according to their value.

Segmentation variables can be defined as factors or characteristics that help differentiate customers. The establishment of assumptions and variables is important primarily because it helps analysts to define a framework for the customer segmentation research process. Once this structure is prepared, it can begin to develop the process to identify the segment of customers.

Benefits of Customer Segmentation

At an early stage, it is extremely important for the company to know and study its target customer, understanding their needs in order to match their expectations. A deep understanding of the segmentation of the best current customers allows for a better allocation and spending of human resources and capital.
Defining the focus of the current customer leads to a better definition of product and market development strategies and there is no diffusion in the company's objectives in relation to its target segments. Otherwise it can prevent the growth of the company.

A Customer Segmentation done correctly will only benefit the company in question, for example: Improving the company's product: Studying the potential customer, their needs, what they want to buy, what they need, helps differentiate the company as the consumer's first choice. Revealing better performance and results compared to competitors. In addition to that since providing information about the customer will ensure the company a better service, services and other types and offers related to the company's product.
Focus on marketing message: A customer targeting project will help to develop more objective and focused marketing messages for each of the segments. This will result in a higher quality of the company's product. Seizing opportunities: By spending less time on less profitable opportunities such as unsuccessful segments, the company will increase its success rate as well as gain more ground and increase revenue. Increasing top-quality revenue: Focusing on the wrong segment can lead to time-consuming and difficult-to-maintain sales, which will result in lower revenues. That way the best way is to focus on the current customer segment and promote the stability of the customer base.

1.4.Goals and Motivation

Nowadays, in the hospitality industry, a great amount of data come from online plataforms, contrasting with what happened a decade or so ago, where operators where the main source of costumer data. Therefore, an effective customer segmentation is needed to better know each group of customers and improve their satisfaction. A few works can be found in hospitality area, so one of the primary goals of this dissertation is that of mining the existing segmentation case applications and the methods employed in the related litterature. Nevertheless, segmentation approaches in other domains will also be described.
In a second phase, a real case study was investigated. Real data pertaining to Property Management Systems from eight Portuguese hotels from a Portuguese hotel chain was analysed mad modelled to understand what costumer s'ments emerged. Futhtermore, and since four of the hotels are situated in Lisbon and the remaining four are resort hotels in the Algarve region, the same data was employed for an exploration towards its predictive power concerning the type of hotel (city or resort) and, since the records identified

cancelations of previous bookings, the likelihood of a booking ‘s cancelation was also investigated.

1.5.Research Questions

In order to guide the research, after a small study and a first analysis of the data, the following Research Questions where established:

1. What methods are used for client segmentation and how they change depending on the area of application (in the data / business domain)?
2. What classification methods exist and if there is a method with better accuracy?
3. Study of a real case in the hotel industry: is it possible to obtain a segmentation based on the available data?
4. Can a multi-method approach improve results?
5. Would it be possible to obtain a general model for application in any hotel or should there be individualized models?

1.6.Structure and organization of the dissertation

The present study is organized in five chapters that intend to reflect the different phases of research progress until its conclusion. The first chapter introduced the research theme and objectives, as well as a brief description of the work structure. The next chapter reflects methodological approach used. The thirth chapter reflects the related literature that was found. Next, the case study is presented, and the data used is described, along with an exploration of this data, the cleaning and processing, and methods of analysis used. The fifth chapter, reflecting the experiences found in tha related literature, presents diverse attempts at modelling this data and the respective results obtained, concluding with a final discussion of the several tests made.
The final chapter presents the conclusions, recommendations, limitations and future work.

2. Chapter 2 – Methodological Approach

2.1. Introduction

In the development of this dissertation, a methodology for the implementation of the Data Mining process was followed: Design Science Research Methodology (DSRM). This is a set of synthetic and analytical techniques for performing research in Information Systems (Kuechler & Petter, 2017). The authors of (Geerts, 2011) define DSRM with three objectives in mind: "(1) provide a nominal process for the conduct of DS research, (2) build upon prior literature about DS in IS and reference disciplines, and (3) provide researchers with a mental model or template for a structure for research outputs.". DSRM in Information Systems (IS) seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts (Geerts, 2011).

Design science research can be divided considering two primary activities to improve and understand the behavior of aspects of Information Systems: "the creation of new knowledge through design of novel or innovative artifacts (things or processes) and (2) the analysis of the artifact's use and/or performance with reflection and abstraction" (Kuechler & Petter, 2017).
The research described in this dissertation considers five phases presented in Figure 1. While the first two phases are described in the previous chapter (Chapter 1), the next ones are presented fully in the following chapters. Notwithstanding, a resume of what can there be found is presented in the next Sub-Sections.

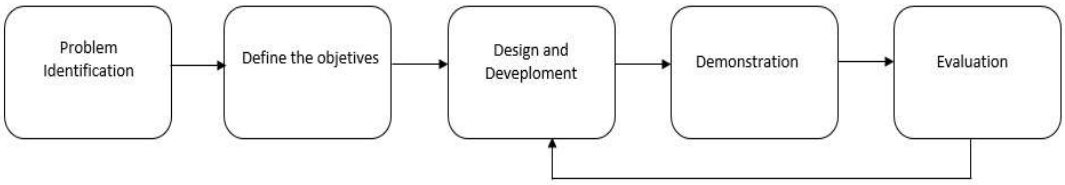

Figure 1 - Design Science Research methodology adapted to the present dissertation

2.2.Data Exploration, Cleaning and Processing

Millions of data come every day from online platforms, quizzes, etc., although these sources usually have redundant data in different representations. In order to improve the accuracy of predictive models and to obtain consistent data, consolidation of different data representations and elimination of duplicate information becomes an important step. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data (Rahm & Do, 2000).
To explore the dataset, it was used histograms and boxplots that are most useful in descriptive statistics. Histograms are commonly used diagrams which show a graphical representation of a data set in which class frequencies are represented by the areas of

rectangles centered on the class interval. Boxplots are diagrams for presenting necessary information to see the center, spread, skew, and length of tails in a data set. This type of graph allows us to compare many distributions in one figure (Ouchi & Takato, 2011).
An important step in processing dataset is the characteristics selection. The dataset usually comes with variables that are statistically uncorrelated with each other. The process of selection of characteristics consists in removal of characteristics in the set of train that are uncorrelated with the category target. This process reduces the set of train and improve the accuracy of the model (J. Han, Kamber, & Pei, 2011).
To make this selection process, it was used methods of correlation. Correlation is a bivariate analysis that measures the relationship between two variables. The value of the correlation coefficient varies between +1 and -1, where values with a positive sign indicates a positive relationship between the variables and values with a negative sign indicates a negative relationship. Usually, in statistics, there are four types of correlations: Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation.
In this thesis, Pearson correlation and Spearman correlation were used. Pearson's correlation measures a linear dependence between two continuous variables (x and y). It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. It can be calculated using the following formula, where m_x and m_y, are the means of x and y variables:

$$p = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}} \quad (1)$$

Spearman's correlation coefficient is a statistical measure of the strength of the relationship between two sets of data and it can be calculated using the following where summation d^2 is the sum of the squared differences between the pairs of ranks, and n is the number of pairs.

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad (2)$$

2.3.Model development, experimentation and testing

After treating the dataset, the next phase is the model development this step consists into the application of data mining methods in order to achieve the goals already defined. In this phase it was used the platform RStudio to develop, test and evaluate the algorithms. This phase can be divided in four steps, the selection of the methods that are going to be applied, the test design, the application of the methods and the evaluation of this methods. The selection of the methods that are going to be used was made with basis in a previous the literature review that it will be present on the next Chapter, between the various methods presented in the literature review it was chosen the most suitable to the case of study and the dataset available. The test design consists of the definition of method to test the algorithm applied, this is important step to understand the capacity of the algorithm must classify a new dataset. In the application of the methods, the dataset is first dividing into a set of train and a set of tests, the train set is the one that will be used to develop the model and the test set will be used to test the model. Finally, the evaluation of the

algorithms used and the critical analysis of the results will be performed using K-Fold Cross.

2.4.Critical analysis of results and validation

The evaluation of the model is an important step of data mining process, being the process of assessing a property or properties of an algorithm. It is often used to understand the relative performance of the algorithm.
As previously refered, to evaluate the methods used in this thesis the K-Fold Cross Validation was used. In K-Fold Cross Validation, the data is randomly divided into k subsets of approximately equal size. At each of k steps, one of the k subsets is used as the set for test, while the other k-1 subsets are put together to form a training set. The cross-validation estimate of accuracy is the overall number of correct classifications, divided by the number of observations in the dataset. The results will be present in this thesis in form of a Confusion Matrix, that is, a table used to describe the performance of the classification model on a set of test data for which the true values are known. Consider the example present in Table 1, having two possible classes YES and NO. An algorithm is classifying 50 observations. Let's say that the algorithm predicted 20 observations as YES and they were a YES and it predicted 5 observations as a YES, but they were not. When the algorithm predicted correctly, we say the rrsponse is a True Positive. When the algorithm predicted that the observation was NO, and it was a NO is termed True Negative. Observations predicted as YES but it that were not are called False Positives and observations predicted as NO but it was a YES is called False Negatives.

Table 1 - Example of Confusion matrix

Nº total: 50	**Predicted: YES**	**Predicted: NO**
Actual: YES	20 (True Positives)	10 (False Negatives)
Actual: NO	5 (False Positives)	15 (True Negatives)

The evalution metrics that are computed using this confusing matrix and that will be presented in this thesis are:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ of\ Observations} \qquad (3)$$

which measures the number of observations that were correctly predicted by the model on the total number of observations.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \qquad (4)$$

That is measuring the number of observations that were correctly predicted by the model on the sum of the observations correctly predicted with the observations that the algorithm predicted as YES, but they are actually NO.

and

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \qquad (5)$$

to measure the number of observations that were correctly predicted by the model on the sum of the observations correctly predicted with the observations that the algorithm predicted as NO, but they are actually YES.

$$F_1 = \ 2\ \frac{precision\ .\ recall}{precision + recall} \qquad (6)$$

is the mean between precision and recall. It tells you how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

3. Chapter 3 – Literature Review

This chapter is divided in several Sub-Sections of which the first presents a brief description about Customer Segmentation and Clustering. Then a review of the related literature divided into the diverse applications domain it's presented.

The methods to be reviewed next can be considered as Data mining techniques. Data mining consists in the process of extracting useful patterns from large data sets, that is, applying data analysis and discovery algorithms to detect patterns over data for prediction and description (Liao & Chueh, 2011). Clustering is a method of data mining to group data according to their similarity. For this case, we have Customer Segmentation that is a particular case of clustering that search for patterns that enables the division of potential or actual customers into differentiating relevant groups. This division is based on similarity among customers, like a need for a product, a service that meets the needs of customers in the same group. In the segmentation process it is difficult to identify the criteria you should use to segment the data. In case of hotel customer segmentation tipically you think of segmenting the customers by characteristics of their reserves. Some of the methods can be classified as soft computing techniques. Soft computing that consistis into tecniques to resolve complex computational problems. Jointly with clustering methods have been used to produce segmented profiles in several areas like health, management, marketing, hotels. But the big question is which method is better (if any) and if just one clustering method it is enough to provide hotels with a meaningful segmentation. In the related literature next presented, we can see that different areas either apply the same or diverse methods.

3.1. Costumer Segmentation in Banking

In the banking business, Smeureanu et al. used both Artificial Neural Networks (ANN) and Support Vector Machine (SVM) to segment customers profiles in a private bank (Smeureanu et al., 2013). The dataset used consisted in 2,783 observations representing active cardholders at an important commercial bank from Romania and reported both methods as showing good performance. The performance of ANN was measured using gradient descent and the logistic function for hidden and output layers while SVM shows good results in terms of misclassifications. In (Davies, Moutinho, & Curry, 1996) the authors also applied Neural Networks to a dataset of ATM users in order to understand how the ATM services can match expectations of their customers and were able to distinguish four distinct user types. For this study, data were gathered by personal interview from a total sample of 380 ATM users. Using network connection weights between five input nodes, four hidden nodes, and four output nodes the authors concluded that clear divisions between four different attitudinal types of ATM consumer were defined. An application of classification and regression trees (CART) and multivariate adaptative regression splines to explore the performance of credit scoring can be found in (Lee, Chiu, Chou, & Lu, 2006). Credit scoring tasks are performed in one bank credit card data set. Analytic results show that the CART and the multivariate adaptative regression splines outperformed traditional discriminant analysis, logistic regression, neural networks, and support vector machine approaches in terms of credit scoring accuracy and misclassification costs thus providing efficient alternatives to conduct credit scoring tasks.

3.2. Costumer Segmentation in Health

In the Health sector, Artificial Neural Networks (ANN) and Support Vectors Machine were used for the recognition of skin diseases (Antkowiak, 2006). For this study the authors developed the Skinchecker-DataSet module with images of all the diseases to be studyed. In this case, the ANN models outperformed the SVM ones in terms of classification. The authors of (Swenson, Bastian, & Nembhard, 2016) applied K-Means methods to a patient data set from electronic medical records to divide the data in six distinct segments. Then used four classification algorithms to classify the patients: Decision Trees, Random Forests, Bootstrap Aggregation, and Non-Linear Support Vector Machines. They used the first set of methods to segment the patients in different clusters and the second set to classify them. The conclusion is that all these classification methods showed good results in terms of accuracy, with the highest being the Random Forest model (92, 8%). In (Wei, Lin, Weng, & Wu, 2012) the authors applied a LRFM (length, recency, frequency and monetary) model extended from marketing RFM (recency, frequency and monetary) model adding to it a length parameter. The authors began by using Self-Organizing Maps (SOM) to segment patients of a dental clinic in Taiwan to which LRFM was applied in each cluster. The results show that three clusters having the above average LRF values (454 patients) can be viewed as core patients.

3.3. Costumer Segmentation in Marketing

In the Marketing sector, we can find works, like the one in (Linder, Geier, & Kölliker, 2004), using Artificial Neural Networks, Classification Trees and Linear Regression in a dataset taken from a random sub-sample of an anonymous data collection of a real Swiss population. The authors state that Classification Trees and Linear Regression show better results for high complexity data. Artifical Neural Netwotrk outperformed both Classification Trees and Linear Regression in predictive accuracy and simultaneously suffered least from overfitting. In (Chiu, Chen, Kuo, & Ku, 2009) Particle Swarm Optimization with K-means was used to decide cluster centres. In addition, a mix technique using Self-Organizing Maps plus K-means was employed to produce prototypes (centres) and then to cluster these prototypes in customers' purchasing behavior data. The authors conclude expressing that Particle Swarm Optimization with K-Means is preferred over K-Means, and Self Organizing Maps with K-Means for their case. Among the three methods, the Mean Square Error value and the intra-cluster distance was the lowest while the inter-cluster distance was the highest for Particle Swarm Optimization with K-Means.

In a more innovative view, a genetic algorithm to select more appropriate customers for each campaign strategy was presented in (Chan, 2008), using a Nissan automobile retailer to segment over 4000 customers. The results demonstrate that the proposed approach can increase potential value, customer loyalty and customer lifetime value. The authors in (Wang, Tu, Guo, Yang, & Huang, 2014) studied the users behaviours in mobile networks by processing big data with clustering methods. The authors thought users with similar ARPU (Average Revenue Per User) may have the same communication behaviour. Using calling records of about one million users in metropolis in China, the users were divided based on their ARPU into tree different groups: high, medium and low ARPU. Then they applied fuzzy c-means in each group to study their behaviours, the conclusion was interesting since they found out that users with same ARPU have different

communication behaviour. For example, one of the clusters generated had peak calling traffic at night and in high ARPU level the proportion of this cluster was small but in medium and low ARPU levels it was large.
According to (Hu & Yeh, 2014) RFM (recency, frequency and monetary) analysis is a powerful tool for assessing customer lifetime value (Hu & Yeh, 2014) . The authors in (Hu & Yeh, 2014) propose a tree structure, called an RFM-pattern-tree, to compress and store an entire transactional database, which receives the fields RScoreDB, FScoreDB, and ttaDB to store the recency score, frequency score, and total transaction amount of certain item. Then a pattern growth-based algorithm, called RFMP-growth, is developed to discover all the RFM-patterns in an RFM-pattern-tree. They used a real dataset called SC-POS-all and apply those methods. Three tests were used to evaluate the proposed method. The first two to evaluate effectiveness, while the third test evaluates efficiency. As a conclusion, they affirmed that RFMP-growth can efficiently and effectively discover more valuable patterns than conventional frequent pattern mining algorithms.
In (Hu, Huang, & Kao, 2013) it is stated that RFM analysis is a well-known and powerful tool in database marketing and is widely used in measuring the values of customers according to their prior purchasing history. They suggest an algorithm to discover sequential patterns with high recency, frequency and monetary scores called RFM-PostfixSpan. This algorithm was developed by modifying the PrefixSpan algorithm, which divides one sequence database into several databases, and returning the RFM-SP (sequential pattern) by exploring only the local pattern in each database. The RFM-PostfixSpan partitions sequence database uses just the postfix to efficiently retrieve the recency score of a pattern. They applied these two algorithms to the sales data of a supermarket chain in Taiwan, called SC-POS. Using the same features, the RFM-PostfixSpan out-performs the PrefixSpan in both runtime and the numbers of generated patterns

3.4. Costumer Segmentation in Management

For the Management area of research, another two-stage method combining Self-organizing maps (SOM) and K-means can be found (Kuo, Ho, & Hu, 2002). The authors proposed the use of SOM to determine the number of clusters and the starting point and then employ the K-means method to find the final solution. Three types of segmentation methods were applied: the conventional two-stage method (self-organizing feature maps and Ward's minimum variance method), self-organizing feature maps and the proposed two-stage method. The authors used a data collection for 3C (computer, communication, and consumer electronics) market. The conclusion was that the proposed two-stage clustering method is slightly better than the conventional two-stage method, except when the number of clusters was two. They also concluded that the use of self-organizing feature maps alone cannot provide a feasible solution.
Customer Relationship Management has become a leading business strategy, with vast benefits like: increased customer retention and loyalty, higher customer profitability, creation value for the customer. In order to this, the authors in (Kim, Jung, Suh, & Hwang, 2006) proposed a new Lifetime Value (LTV) model. LTV was defined as the sum of the revenues gained from company’s customers over the lifetime of transactions, after the deduction of the total cost of attracting, selling, and servicing customers, and taking into account the time value of money. They divided the customers by their Current Value, Potential Value and Customer Loyalty but, for marketing strategies, they need to know

the characteristics of each group. To achieve that, a decision tree was applied to real data of a wireless company.

3.5. Costumer Segmentation in Telecommunications

For the Telecommunication sector, to segment customers from a elecom industry in China a novel customer segmentation method based on a customer lifecycle using a Decision Tree model was proposed (S. H. Han, Lu, & Leung, 2012). The authors in (Liao & Chueh, 2011) used Decision Trees and Neural Networks to segment the costumers from a wireless telecommunication service. Both techniques generate models with a hit rate of 98%.
Hwang et al. suggested a customer Lifetime Value Model considering past profit contribution, potential benefit and defection probability of a customer (Hwang, Jung, & Suh, 2004). They affirmed that the customer value is classified into three categories: current value, potential value and customer loyalty. For testing that model, a six-month service data of one wireless communication company in Korea was used. They considered the current value as the average amount of service charge asked to pay for a customer minus the average charge in arrears for a customer. To calculate the potential value, they first used R2 method to find the variables that affect the fact whether customers use an optional service or not. Then decision trees, neural networks and logistic regression were used for classification. After evaluationg the models by using misclassification rate and lift chart, the most powerful method was selected to calculate the potential value. And finally, customer loyalty was calculated by one minus the churn rate, which was also calculated by using decision trees, neural networks and logistic regression.

3.6. Costumer Segmentation in Hospitality Industry

For the Hospitality industry, there are few studies carried out in this area. However, we can find a work that uses the Chi-square automatic interaction detection (CHAID) as a decision tree technique for the market segmentation (Chung, Oh, Kim, & Han, 2004) The authors collected monthly data from all super deluxe hotels in Seoul for three years. The rationale behind was that the technique used can help researchers trade off variance against segment size to find the most suitable one.
Several density based algorithms - DBSCAN, OPTICS, EnDBSCAN – can be found in (Bose, Munir, & Shabani, 2017) with the purpose of segment differents datasets. The results reveal that EnDBSCAN out-performs DBSCAN and OPTICS in terms of identifying nested and embedded clusters. Similarly, OPTICS out-performs DBSCAN in identifying adjacent nested cluster for different datasets.The authors of (Legohérel, Hsu, & Daucé, 2015) used Variety-seeking to study the behaviour of international travellers. “Variety-seeking refers to the tendency for consumers to alternate between makes of the same product or to the quest for diversity in selecting goods and services”. To evaluate Variety-seeking, the authors used the Exploratory Buying Behaviours Tendency (EBBT) scale, which consists in two variables: Exploratory Acquisition of Products (EAP) which relates to purchase experiences and Exploratory Information Seeking (EIS) which considers the consumer need for cognitive stimulation. To evaluate the significance of variety-seeking they applied the chi-squared automatic interaction detection (CHAID) to 482 interviews collected from travellers in the departure hall of the Hong Kong

International Airport, to segment the travellers based in their criterion to choose a hotel and restaurant. The objective was to identify for each product, hotel and restaurant, the ability of EBBT variables to explain the traveller's decision-making process. In case of the hotel, three segments were formed. The best predictors of hotel preference were brand loyalty (an EAP variable) and window-shopping behaviour (an EIS variable).

3.7.Conclusions drawn from the related literature

According to previous literature review, data mining is used in several areas of application to perform customer segmentation. Methods like Artificial Neural Networks, Classification Trees, Linear Regression, K-Means, Support Vector Machine and others are the most common algorithms used in the surveyed areas: Banking, Health, Telecommunication, Marketing and Management. It should be noted that, although some of the works found were studies for the hospitality area, these were far less when compared with published research for other areas of application.

Table 2 - Summary of the diverse methods found in the literature review, by area

Area	Method	Data Source
Banking	• **Artificial Neural Networks** • **Support Vector Machine**	Active cardholders at an important commercial bank from Romania (Smeureanu et al., 2013)
	• **Classification and Regression Trees**	Bank credit card data set (Lee et al., 2006)
Health	• **Artificial Neural Networks** • **Support Vectors Machine**	Skinchecker dataset (Antkowiak, 2006)
	• **K-Means**	Patient dataset from electronic medical records (Swenson et al., 2016)
	• **Self-Organizing Maps** • **LRFM**	Patients of a dental clinic in Taiwan (Wei et al., 2012)
Marketing	• **Artificial Neural Networks** • **Classification Trees** • **Linear Regression**	Anonymous data collection of a real Swiss population (Linder et al., 2004)
	• **Self-Organizing Maps** • **K-means**	Customers' purchasing behavior data. (Chiu et al., 2009)
	• **Genetic algorithm**	Nissan automobile retailer to segment over 4000 customers (Chan, 2008)
	• **RFM**	Real dataset SC-POS-all (Hu & Yeh, 2014)
Management	• **Self-organizing maps (SOM),** • **K-means** • **Ward's minimum variance method**	Data collection for 3C (computer, communication, and consumer electronics) market (Kuo et al., 2002)
	• **Decision Tree**	Real data of a wireless company(Kim et al., 2006)
Telecommunications	• **Decision Tree**	Customers from a telecom industry in China (S. H. Han et al., 2012)
	• **Decision Trees** • **Neural Networks**	Customers from a wireless telecommunication service (Liao & Chueh, 2011)

	• **Lifetime Value Model**	Six-month service data of one wireless communication company in Korea (Hwang et al., 2004)
Hospitality	• **Chi-square automatic interaction detection**	Data from all super deluxe hotels in Seoul (Chung et al., 2004)
	• **Density based algorithms**	Different datasets (Birant & Kut, 2007)
	• **Variety-seeking**	482 interviews collected from travelers in the departure hall of the Hong Kong International Airport (Legohérel et al., 2015)

Table 2 presents some of these works which are the most similar to the case of study present in this thesis because they are customer-related data. As we saw in the review of the literature, the algorithms differ according to the application area, type and amount of data. In this sense, and with the support of the literature review, a set of customer classification and segmentation algorithms was chosen and the best method for our case was studied.

In the present dissertation, the classification algorithms chosen to be tested are Decision Trees, Naïve Bayesian, Random Forest, Logistic Regression, Support Vector and Artificial Neural Network, which are presented in the following Sub-Sections.

3.8.Classification Algorithms

Classification algorithms are supervised algorithms, algorithms that learn to classify input data and then use this learning to classify a new input set. This classification consists in the assignment of a class according to the data set to be classified, which could be bi-class (it will rain or not) or multi-class (it is red, blue, yellow or green). There are several types of classification algorithms: Logistic Regression, Naive Bayes Classifier, Support Vector Machines, Decision Trees, Boosted Trees, Random Forest, Neural Networks, Nearest Neighbor, among others.

3.8.1. Decision Trees

Decision Trees build classification or regression models in the form of a tree. This algorithm uses a set of rules to divide a dataset into smaller groups which justifies the method's name since tree decisions is developed and can be used to classify new samples. Figure 2 adapted https://www.saedsayad.com/decision_tree.htm presents an example by representing a dataset with the predictor variables and the target at the table at the left. This example shows a binary class (e.g Yes or No) and, at the right of the figure, a decision tree can be seen, with the decision nodes (attribute names) and leaf nodes (class values). The decision node is an attribute with each branch being a possible value of the attribute (e.g Weather, Windy), the leaf node represents the classification or decision (e.g. Rainy, False, Normal). The root node corresponds to the best predictor, in this case it is the class Weather. Decision Trees can fit categorical and numerical type of data.

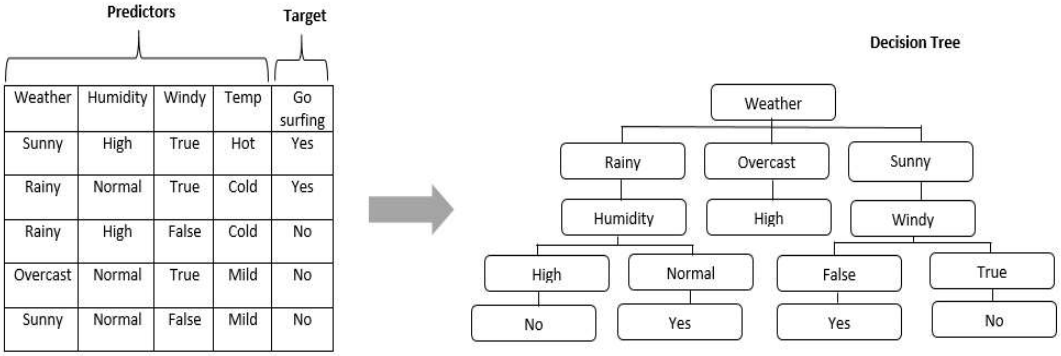

Figure 2 - Example of a Decision Tree model

In this study, the R package rpart was used, which builds decision tree models for classification or regression of a very general structure using a two-stage procedure. The tree is built by the following process: first the single variable is found which best splits the data into two groups. The data is separated, and then this process is applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size or until no improvement can be made. The resulting models can be represented as binary trees.

3.8.2. Naive Bayes

The Naive Bayes classifier is based on Bayes's theorem and assuming independence assumptions between predictors. A Naive Bayes model is easy to build, with no complicated iterative parameter estimation, which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayes classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.
This classifier method is founded on Bayesian probability, which originated from the work of the mathematician (Gaigerov, Elkina, & Pushkin, 1982).

3.8.3. Random Forest

Random Forests are an ensemble of Decision Trees (trees with only a root node) built with bootstrap samples trained using a variant of the random subspace method or feature bagging method. Random forests are a combination of those predictors trees such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The simplest random forest with random features is formed by selecting at random, at each node, a small group of input variables to split on (Breiman, 2001).

3.8.4. Logistic Regression

Logistic Regression is a predictive analysis used to explain the relationship between a binary dependent variable and one or more independent variables. It is considered a statistical method for analyzing a dataset that has one or more dependent variables that determine a binary result. The objective of logistic regression is to find the best fit model to describe the relationship between the independent variables that represent the predictors and the dependent variable that represents the outcome. The logistic function calculates probabilities for each new observation class that, by using a threshold, become binary values in order to make the prediction.
Logistic regression is a widely used statistical modeling technique in which the probability of a dichotomous outcome is related to a set of potential independent variables (Hutcheson, 2011).

3.8.5. Support Vector Machine

Support Machine Vector is a supervised learning model used for classification and regression analysis. The purpose of this model is to find an optimal hyperplane that best divides a dataset into two classes like the example presented in Figure 3 adapted from https://www.saedsayad.com/support_vector_machine.htm in a two dimentional space.

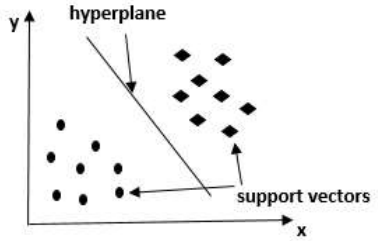

Figure 3 - Example of support sectors for a SVM model

3.8.6. Artificial Neural Network

An Artificial Neural Network is a supervised method for classification that consists of input and output layers, as well as one or more hidden layers, consisting of units that transform the input into something that the output layer can use to classify.
A neural network consists of units (neurons) organized in layers that transform an input vector of attribute values to an output of target values. Each unit receives an input value, applies a an activation function to that input and then passes the result to the next layer. The simplest networks are defined as feed-forward, which means that the values' flow is unidirectional. A unit sends information to other unit from which it does not receive any information. There are no feedback loops. Figure 4 adapted from https://www.saedsayad.com/artificial_neural_network.htm presents an example of a (feed-forward) Artificial Neural Network with one hidden layer.

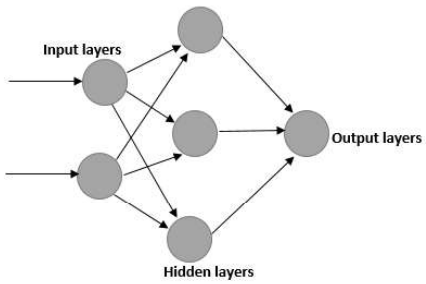

Figure 4 - Example of an Artificial Neural Network

3.9.Segmentation Algorithms

Segmentation or Clustering Algorithms are unsupervised machine learning techniques that consist in grouping data points. For a determined set of data points, this algorithm is used to cluster each data point into a specific group where each point is considered more similar with its mate points and the points between different groups are considered as less similar. The quality of a clustering depends on both the similarity measure used by the method and its implementation. It is measured by the ability of the system to discover some or all of the hidden patterns (Chiu et al., 2009).
There are many clustering tecniques, like K-Means, Fuzzy C-means, Hierarchical clustering, Mixture of Gaussians. In this thesis, only K-Means and Hierarquical Clustering will be used.

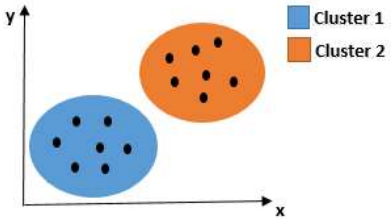

Figure 5 - Example of clustering

3.9.1. K-Means

K-Means is an unsupervised learning model for clustering problems. It assigns a given data point to one of a set of clusters, each set representing a segment. Each of these segments or groups cluster observations with similar caracteritics whitin the same group trying to maximize inner group similarities and outer-group dissimilarities. The algorithm transforms a given dataset into a set of k segments, where k stands for the number of groups pre-specified by the analyst.

3.9.2. Hierarquical Clustering

Hierarchical clustering is also a clustering method for identifying groups in the dataset. While k-Means needs a pre-specification of the number of clusters, this method does not. Hierarchical clustering can be divided into two types: Agglomerative or Divisive Clustering. The Agglomerative clustering, also called AGNES, works in a bottom-up manner. At each step of the algorithm, the two clusters that are most similar are combined into a new bigger cluster. Divisive clustering, also called DIANA, works in a top-down manner: in each step of the algorithm, the most diversified cluster is divided into two. The result for these two types of hierarchical clustering is a tree that can be plotted as a dendrogram.

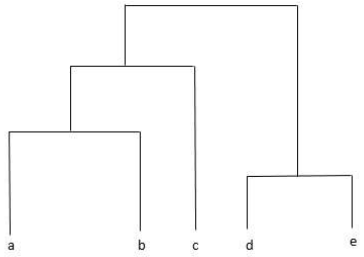

Figure 6 - Example of a hierarchical clustering

In the clustering process, an important step is the definition of the method of similarity or distance matrix that is going to be used. Ssimilarity measures the distance between each pair of observations. Between the many methods to calculate this distance matrix, Ward's minimum distance is the one used in this thesis work, following one of the references previously surveyed (Punj & Stewart, 1983).

3.9.3. Self-Organizing Maps

Self-Organizing Maps (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning. SOM is a clustering concept that group similar data.
In the SOM process, it first generated weight vectors from the observations, then a vector is chosen randomly, and the map of weight vectors is searched to find which weight best represents that vector. The neighborhood is being formed; each neighbor is considered to be similar to the other one. In the train process, the algorithm adjusts the values for the input variables trying to preserve neighbourhood relationships of the input data. The unit weights are adjusted they as well neighbourhood.

3.9.4. Density-Based Spatial Clustering of Applications with Noise

DBSCAN (Density-Based Spatial Clustering and Application with Noise), is a density-based clusering algorithm (Ester et al. 1996), which can be used to identify clusters of any shape in a data set containing noise and outliers.
Clusters are dense regions in the data space, separated by regions of lower density of points. The DBSCAN algorithm is based on this intuitive notion of "clusters" and "noise".

The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.
DBSCAN is designed to discover arbitrary-shaped clusters in any database and at the same time can distinguish noise points. More specifically, DBSCAN accepts a radius value Eps based on a user defined distance measure and a value MinPts for the number of minimal points that should occur within Eps radius (Birant & Kut, 2007).
The goal is to identify dense regions, which can be measured by the number of objects close to a given point. Two important parameters are required for DBSCAN: epsilon ("eps") and minimum points ("MinPts"). The parameter "eps" defines the radius of neighborhood around a point x. It's called called the ϵ-neighborhood of x. The parameter MinPts is the minimum number of neighbors within "eps" radius.

Tha advantage of using DBSCAN rather than K-Means is does not require the user to specify the number of clusters to be generated. This algorithm can find clusters of any shape and can identify outliers from the dataset.

4. Chapter 4 – Study case dataset preparation

The following chapter it is composed in the following sections: Section 4.1 where an analysis and description of the dataset is made, Section 4.2 that presents an exploratory analysis of the dataset and lastly, this chapter conclusions are presented in Section 4.3.

4.1. The Dataset Description

The dataset to be used consists in real hotel booking data obtained from PMS from 8 Portuguese hotels: 4 in a city and 4 in a resort. In this dataset, to preserve anonymity, each hotel is identified by the attribute HotelID, designated here as H1 to H8. The attribute HotelType classifies each hotel as City or Resort.
Data ranges over a period of two and a half years, namely from 01-01-2013 to 30-06-2016. Each example concerns a booking that has existed and data specifying if it was cancelled or not. No values are missing in the data set.
For identifying different types of customers important from the hotel point of view, the categorical attribute CustomerType has the values Trasient, Group, Contract and Trasient-Party. These terms were defined as an initial and usual segmentation that hotels use, where the Trasient type is used for the customers who are predominantly going to stay a short period of time in the hotel. Customers with a last-minute booking, an individual costumer with a short staying can be considered a Trasient customer. Transient Party is almost the same, the unique different being that it is used for small groups, such as families, who require more than one room. The Contract type is usually for enterprises or for customers that are travelling on business. Often a previous contract with the hotels exists. At last, the Group, like the name says, belongs to the customers that travel in group, like families, friends, they usually request more than one room and stay for a long period of time.
The numeric attribute ArrivalDateDayOfMonth corresponds to the day of month the customer arrived or would have arrived at the hotel, thus being a numeric attribute in $\{1, 31\}$. The categorical attribute ArrivalDateDayOfWeek corresponds to the day of week the customer arrived at the hotel. The values for this attribute could be Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. The categorical feature ArrivalDateMonth corresponds to the month the customer arrived at the hotel, with possible values going from January to December, and the numeric attribute ArrivalDateWeekNumber corresponds the week the customer arrived at the hotel, with values ranging from 1 to 52.
The categorical attribute BookingDateDayOfWeek corresponds to the day of week the booking entered the PMS (Monday to Sunday). When the booking was cancelled, the binary attribute IsCanceled has value 1 otherwise it has the value 0. The numeric attribute CanceledTime shows the number of days prior to arrival when the booking was cancelled or -1 if the booking was not cancelled. Based on the guest's history record, the binary attribute IsRepeatedGuest is 0 if the guest has already stayed in that hotel or 1 otherwise. The numeric attributes PreviousBookingsNotCancelled, PreviousCancellations, PreviousStayss are also based on the guest's history record. The variable PreviousBookingsNotCancelled corresponds to the number of bookings the guest has made that weren't cancelled. PreviousCancellations corresponds the number of bookings the guest had made and were cancelled. PreviousStayss corresponds the number of

bookings placed at the hotel prior to the current booking. The number of days the booking was placed at the hotel prior to the current date is shown by the numeric attribute LeadTime. The numeric attribute LenghtOfStay corresponds to the number of days of the guest's stay at the hotel.
Not all customers are the same. Although, of course, all guests must receive the same high standard of service at a hotel or other kind of facility within the Hospitality Sector, there are definitely different types of customers. And offering them a hospitality experience that entirely meets their type and exact needs is the key to a providing an enjoyable hotel stay, where satisfaction is virtually guaranteed. In this sense,
Market Segment are group of customers with similar characteristics, this segmentation can be Geographic, Demographic, Behavioural, Psychographic, Occasional and Cultural second Xhotels enterprise. Each hotel has his way to segment his customers, this can be notice for example in the dataset used in the present thesis, each hotel has a different set of Market Segmentation.
Another variable that is differentiating for hotel is DistributionChannel, this term describes the different platforms in which bookings for a hotel are made. A Channel can be a hotel's booking engine a direct phone reservation or a specific stream of revenue such as a 3rd party website.
Other attributes with more detailed information concerning the booking can be found in Tables 1 and 2, together with the data format (Type) and data's short description (Description).

Name	Type	Description
ADR	Numeric	Average daily rate
Adults	Number	Number of adults
AgeAtBookingDate	Number	Age in years of the booking holder at the time of booking
Agent	Categorical	ID of agent (if booked through an agent)
ArrivalDateDayOfMonth	Numeric	Day of month of arrival date (1 to 31)
ArrivalDateDayOfWeek	Categorical	Day of week of arrival date (Monday to Sunday)
ArrivalDateMonth	Categorical	Month of arrival date
ArrivalDateWeekNumber	Numeric	Number of week in the year (1 to 52)
AssignedRoomType	Categorical	Room type assigned to booking
Babies	Numeric	Number of babies
BookingChanges	Numeric	Heuristic created by summing the number of booking changes (amendments) prior to arrival that could indicate cancellation intentions (arrival or departure dates, number of persons, type of meal, ADR, or reserved room type)
BookingDateDayOfWeek	Categorical	Day of week of booking date (Monday to Sunday)
CanceledTime	Numeric	Number of days prior to arrival that booking was canceled; when booking was not canceled it had the value of –1
Children	Numeric	Number of children
Company	Categorical	ID of company (if an account was associated with it)
Country	Categorical	Country ISO identification of the main booking holder
CustomerType	Categorical	Type of customer (group, contract, transient, or transient-party); this last category is a heuristic built when the booking is transient but is fully or partially paid in conjunction with other bookings (e.g., small groups such as families who require more than one room)
DaysInWaitingList	Numeric	Number of days the booking was in a waiting list prior to confirmed availability and to being confirmed as a booking
DepositType	Categorical	Because no specific field in the database existed with the type of deposit, based on how hotels operate, a heuristic was developed to define deposit type (nonrefundable, refundable, no deposit): payment made in full before the arrival date was considered a "nonrefundable" deposit, partial payment before arrival was considered a "refundable" deposit, otherwise it was considered as "no deposit"

Figure 7 - Descriptive table of the dataset (part I)

Name	Type	Description
DistributionChannel	Categorical	ID of the distribution channel used to make the booking
HotelID	Categorical	ID of hotel
HotelType	Categorical	Hotel type (City or Resort)
IsCanceled	Categorical	Outcome variable; binary value indicating if the booking was canceled (0: no; 1: yes)
IsRepeatedGuest	Categorical	Binary value indicating if the booking holder, at the time of booking, was a repeat guest at the hotel (0: no; 1: yes); created by comparing the time of booking with the guest history creation record
IsVIP	Categorical	Binary value indicating if the guest should be considered a Very Important Person (0: no; 1: yes)
Kardex	Categorical	Kardex ID (history record file ID – if guest has one)
LeadTime	Numeric	Number of days prior to arrival that the booking was placed in the hotel
LenghtOfStay	Numeric	Number of nights the guest stayed at the hotel
MarketSegment	Categorical	ID of the market segment to which the booking was assigned
Meal	Categorical	ID of meal the guest requested
PreviousBookingsNotCanceled	Numeric	Number of previous bookings to this booking the guest had that were not canceled
PreviousCancellations	Numeric	Number of previous bookings to this booking the guest had that were canceled
PreviousStays	Numerical	Number of nights the guest had stayed at the hotel prior to the current booking
RequiredCarParkingSpaces	Numeric	Number of car parking spaces the guest required
ReservedRoomTypes	Categorical	Room type requested by the guest
RoomsQuantity	Numeric	Number of rooms booked
StaysInWeekendNights	Numeric	From the total length of stay, how many nights were in weekends (Saturday and Sunday)
StaysInWeekNights	Numeric	From the total length of stay, how many nights were in weekdays (Monday through Friday)
TotalOfSpecialRequests	Numeric	Number of special requests made (e.g., fruit basket, sea view, etc.)
WasInWaitingList	Categorical	Binary value indicating if the guest was in a waiting list prior to confirmed availability and to being confirmed as an effective booking (0: no; 1: yes)

Figure 8 - Descriptive table of the dataset (part II)

4.2.Data exploratory analysis

An exploratory data analysis was performed with this dataset using the open source software R and RStudio platform.
Using the function *table* provided by the R package stats, it is possible to see that both types of hotels (city and resort) contribute to the dataset with roughly the same amount of observations (bookings): from the 136887 observations in the set, 69103 come from city hotels and 67784 from resort hotels.
Overall, regarding the attribute CustomerType, the Transient type presented the highest number of observations: 92425 observations against 12837 for Contract, 6192 for Group observations and 25433 for Trasient-Party. If we partition for the type of hotel (Fig. 10), we can see that the Trasient type is the most regular for both types of hotel, which is already expected since most of the observations belong to this type. But a curious fact is that the Group type does not appear in City hotels. On the other hand, it presents a significant number of observations for Resort hotels. We can also see that observations with category Contract are more frequent for observations with category Resort than for City hotels.

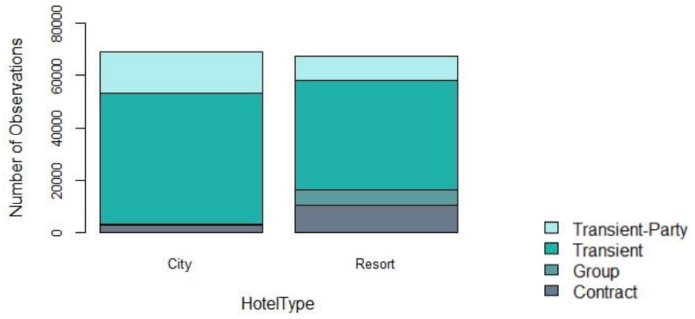

Figure 9 - Costumers classes by hotel type

In City hotels the length of stay is usually short than in Resort hotels which is completely expected as City hotels are in high demand for business trips, or by tourists who stay for a few days and decide to change hotels or locations. In Resort hotels are most in demand for long vacations, often with family, large groups. In Figure 10 we can see that for Resort hotels represented with the blue bar we have almost 15000 bookings with a 7 day stay while for City hotels represented with the pink bar the number of bookings registered with 7 days stay is much smaller.

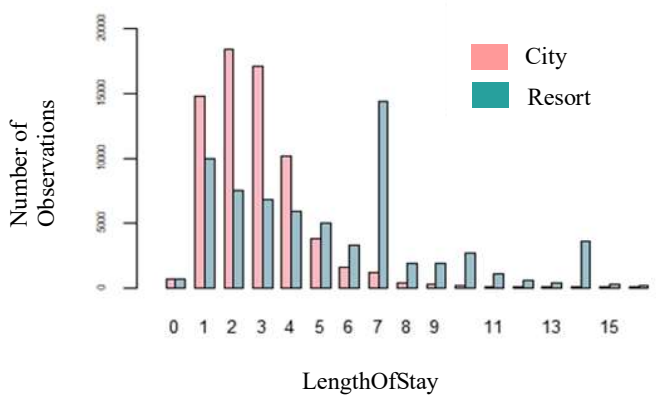

Figure 10 - Lenght of stay by hotel type

Regarding the variable IsCanceled, in the available dataset more than 30000 observations were canceled, near 23% of cancellations. Diving the numbers of cancelations by HotelType, city hotels have a higher rate of cancellations (pink bar in Fig. 11) than resort hotels in a relative difference of 13%. In fact, for resort hotels, the number of cancelled reservations is almost half of the bookings.

Figure 11 - Numbers of cancellations by hotel type

Figure 12 and Table 3 relate CustomerType with the respective number of cancelations. As expected, observations belonging to the Transient present a higher number of cancellations. While the percentage of cancelled bookings in the Transient type is of 71,59% and of 19,17% for the Transient party, for Contract is of 7,29% and for Group is only of 1,94%. Relatively the no cancellations, the Transient type stills have the bigger percentage with 66,07% and Transient-Party have 18,37% of the no cancellation, for

Contract is 10,11% and Group type stills have the lower percentage with only 5,44% of the no cancellations.

Table 3 - Cancellation numbes by costumer type

	Canceled	NotCanceled
Contract	2614	10223
Group	697	5495
Transient	25657	66768
Transient-Party	6872	18561

Figure 12 - Numbers of cancellations by customer type

Analysing the variables CustomerType, HotelType and IsCanceled, Transient customers are the ones with the higher percentage of cancellations while group are the ones who show the smalest percentage of cancellations, regardeless of the hotel's type. We have seen that in Resort hotels there is a higher percentage of group customers and therefore tend to have a longer stay, these customers have a lower percentage of cancellations while in City hotels there will be more cancellations since this type of hotel is often characterized by tourists who stay for a few days and decide to change hotels or locations, or also by business travelers, so they are shorter stays and may suddenly change.

Other variables were also explored. For example, by exploring the attribute BookingDateDayOfWeek, it was found that the days with the highest number of bookings were Friday, Monday and Thursday.
Using the attributes LenghtOfStay and CustomerType, we perceived that Transient type were those who stayed the longest period in any type of hotel and that the Contract type are he ones who stayed the shortest period.

Regarding the attribute RepeatGuest it was found that, for city hotels, costumers tend to choose an hotel that they had already booked before, while for resort hotels, this happens less frequently.

Figure 13 presents a histogram of the variable CanceledTime or the number of days prior to arrival at the date the booking was canceled. The largest bar on the negative side of the x-axis ($x = -1$) refers to bookings that were not canceled, thus it should be disregarded. In fact, most of the bookings were not cancelled but, for those that were cancelled, it is

possible to see that a significant number of bookings were cancelled 1 to 8 days before the day of the arrival in the hotel. The number of observations that were cancelled with the largest distance from the arrival day is very small.

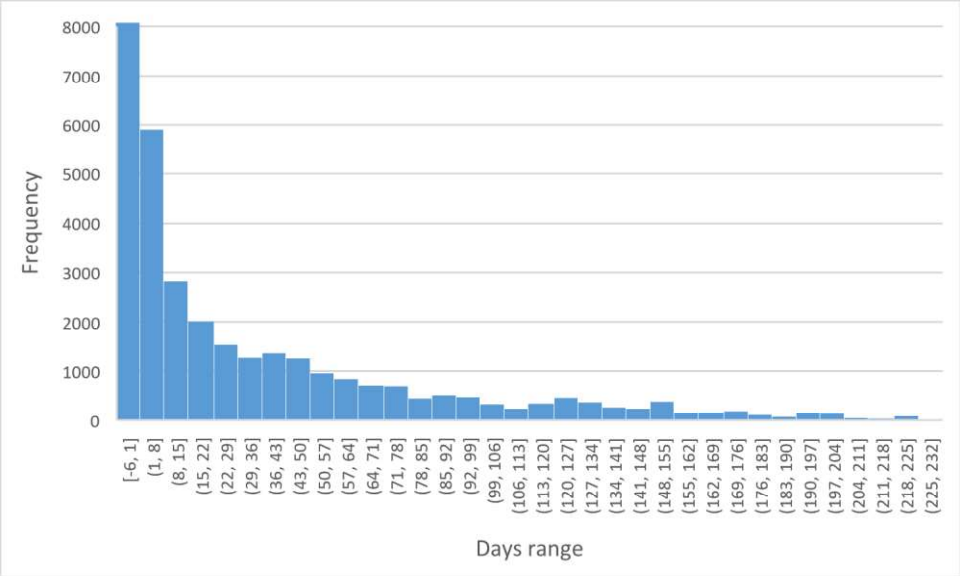

Figure 13 - Histogram of the variable CanceledTime

Figure 14 presents the histogram for the variable LeadTime, i.e., the number of days before the arrival the booking was placed. It is possible to notice that most of the observations were booked close to the day of arrival. But it's also possible to see that many other bookings are made several days in advance, even though they are the ones with the lowest frequency. For resort hotels, LeadTime values are higher than for city hotels. Nevertheless, the reservations with higher LeadTime were also the ones that cancelled more. Curiously, resort hotels presenting an interval between 150 and 350 days for LeadTime were also the ones with higher number of cancelations. It might seem that planning vacations with 6 months ahead might be spurious. However, the cause may be tied to the fact that costumers were waiting for a price downgrade or chose other hotels due to better pricing conditions.
Note that there are LeadTime values above 600 days, which is not normal since this means that the booking was placed almost two years before. Therefore, these observations were considered as outliers.

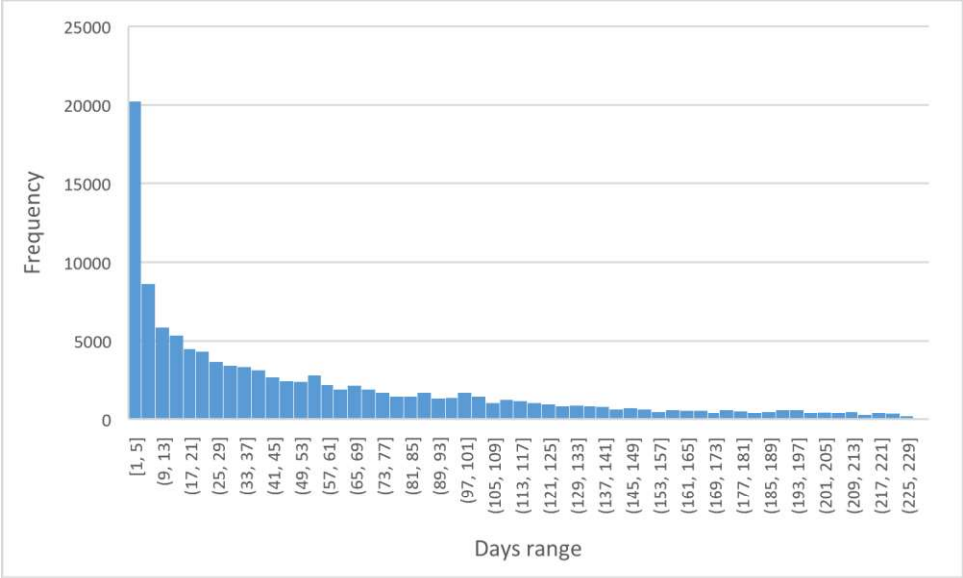

Figure 14 - Histogram of variable LeadTime

To verify this, the boxplot presented in Figure 15 was generated, allowing to account for outliers. As treatment of these same outliers, we decided to restrict the interval of LeadTime from 0 to 600 days and the result is presented in Figure 16 where it is possible to see that no more outliers are present.

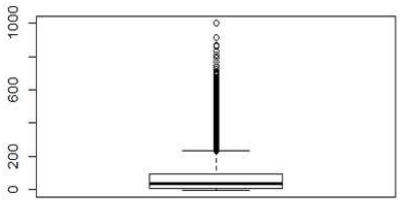

Figure 15 – Boxplot for LeadTime before treating outliers

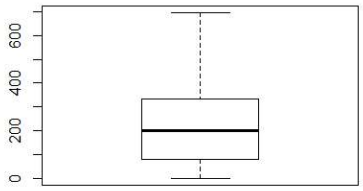

Figure 16 - Boxplot for LeadTime after removing outliers

Feature Selection

In the present case study, we intend to test the datset to answer two different questions: (a) the possibility of predicting a booking to be canceled, which is a classification problem; and (b) is the predefined segmentation of hotel costumer by the attribute CostumerType adequate, i.e., if we explore the data to segment it, will different profiles emerge? In order to proceed with testing the dataset to address question (a), the most predictive or relevant features should be chosen. In this case, the target is the variable isCancelled. Like already mentioned in Chapter 2 (Section 2.3), feature selection is an important step in Data Mining. The process of selection of characteristics consists in the removal of atributes in the training set that appear uncorrelated with the target. This process reduces the training set and improves the accuracy of the model (J. Han et al., 2011).

Categorical variables analysis

To visually study the association between two or more categorical variables, a mosaic plot was used. This plot is a rectangular area, subdivided into rectangular tiles, the area of which represents the conditional relative frequency for an observation in the contingency table.

Figure 17 shows the predictor variables HotelType and IsRepeatedGuest (horizontally) against BookingDateDayOfWeek (vertically) having the variable IsCanceled as a target. If IsRepeatedGuest equals 1, this means that the guest has already stayed at that hotel. This quantity is smaller than the occurrence of new guests (IsRepeatedGuest equals 0). Aditionally, the relationship between cancellations and non-cancellations of repeated guests is similar to the non repeated guests. For an example, in the bottom row we can see that the percentage of cancellations is higher than the percentage of non cancellation in both city and resort hotels and for repeated and non repeated guests. So, the influence of this variable in a prediction of cancellations is practically null. Comparing now the first and the second rows (Friday and Monday) it is possible to see that the percentage of cancellations is bigger on Friday's than on Monday's and the percentage of non-cancellations increase a lot on Monday's. So, in a forecast of cancellations, this variable could be a good predictor. Looking to the variable HotelType, for example: first row, column 0 and 2, it is possible to see that the percentage of non-cancellations is much bigger in Resort hotel than in City hotels, so this variable is also a good predictor in a forecast of cancellations.

Figure 17 - Association analysis for HotelType, IsRepeatedGuest, BookingDateDayOfWeek and IsCanceled.

In Figure 18, the association between the categorical variables HotelID, CustomerType and the target it is being analysed. Regarding to HotelId H1, it possible to see that there are very few observations of Contract type of customers for HotelID H1. The same happens to the CustomerType Group, with Transient and Transient-Party being the main costumers of this hotel. For Trasient type (4th column), which is the most common, the quantity of non-cancellations is higher for most of the hotels. However, this is not the case for hotels H6 and H3. While H3 has only very few observations for this type of

client, H6 shows a higher percentage of cancelations, and for H7, non-canceled booking barely surpasses the canceled ones. Note that HotelID H1-4 are Resort hotels and H5-8 are City hotels. As a whole, Resort hotels show the higest percentages of Contract and Group types, which can be explained by families in vacation booking and either Tourism operators or compamies that use Resorts to join all its collaborators for motivational days.

Figure 18 – Association analysis for CustomerType, HotelID and IsCanceled.

Figure 19 shows a comparison between both the hotel's type in relation ons and weekdays (categorical variables CustomerType, HotelID and Hot king at IsRepeatedGuest = 0 it is possible to see that cancellations for City hotels is higher than for Resort hotels whatever the day of the week. While for IsRepeatedGuest = 1 the percentage of cancellations is higher in City hotels than in Resort hotels every day of the week except Wednesdays. In fact, the percentage of canceled bookings for City hotels ranges from 63% to 81% when the booking is placed by first time customer and from 44% to 84% for repeated guests.

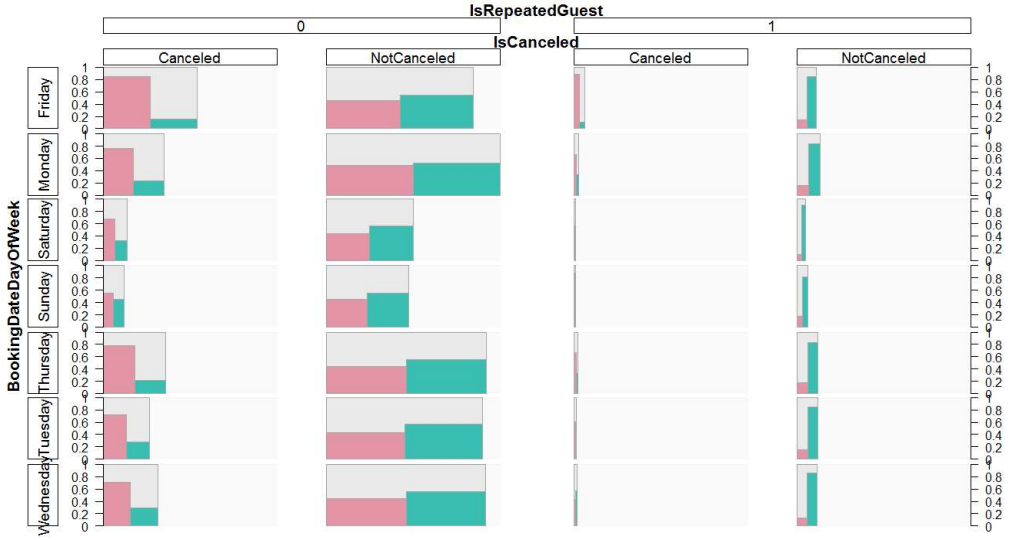

Figure 19 – Association analysis for IsRepeatedGuest, IsCanceled, BookingDateDayOfWeek(and HotelType.

Keeping the analysis in the variable HotelType (Fig.20) and still looking into the canceled bookings and has expected after the previous day-of-the-week analysis, most of the canceled bookings belong to City hotels whenever the month scheduled for arrival date. Notwhitstanding, the difference between, for instance, August and December it is considerable (2nd and 3rd column): in August, the cancellations for City hotels are of 58% of and of 45% in Resort hotels while, in December, the cancellations registered for City hotels were 84% and 16% in Resort hotels. Similarly of what can be seen for the day-of-the-week analysis, most of not canceled bookings belongs to Resort hotels, but in December, February, January and November the quantity of not canceled bookings registered for City hotels were bigger than for Resort hotels. City hotels are highly sought after by people who travel for business, as this happens in months that are not characterized by months of holidays except December, may be the cause of not having so many cancellations at those time.

The remaining categorical variables in the dataset were not considered for this exploratory study since they may not have predictive power or for operational reaons. Varibles IsVip and WasInWaitingList were excluded because only a small number of VIP bookings exists, and the same happens for the variable WasInWaitingList. The variable Country is also excluded because it is only validated upon arrival of the guests, thus, only for bon canceled bookings. This means that most of the canceled bookings show as default value Portugal, which may cause a classifier to become biased upon the Country value. However, the cause for exclusion of variables like Agent, AssignedRoomType, Company, Country, Kardex, MarketSegment, Meal and ReservedRoomTypes, the quantity of possible values for these variables is huge, which leaves the algorithms much slower in case of ANN the algorithm was running for days. So, it was decided excluded them from the predictors group.

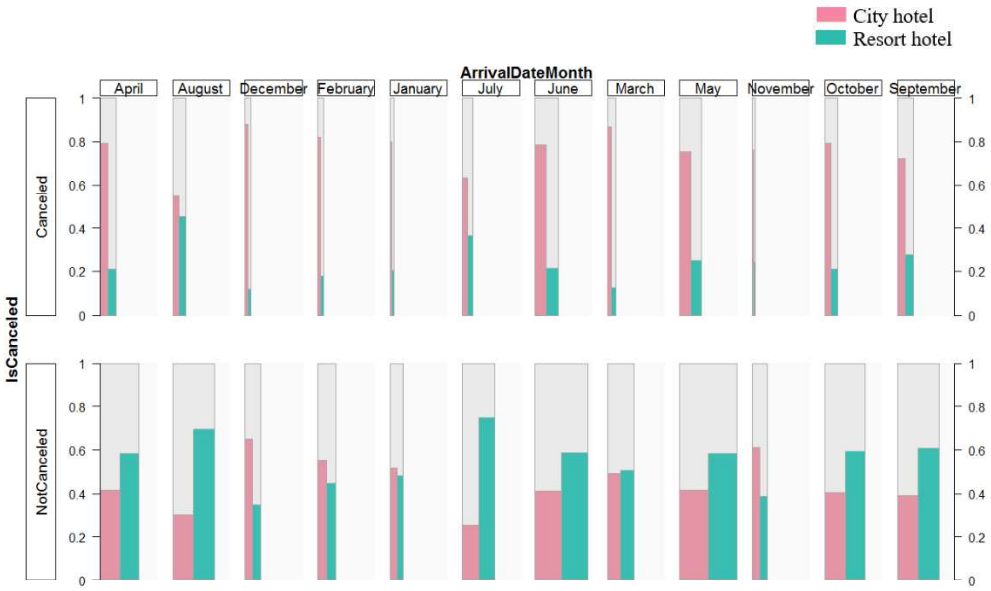

Figure 20 – Association analysis for ArrivalDateMonth, IsCanceled and HotelType.

Numerical variables analysis

For numerical variables the tipical analysis involves correlation plots. Figure 21 presents the correlation matrix (measurement of dependences between two numeric variables) for the numerical variables in the dataset. The value of the correlation coefficient varies between +1 and -1, where values with a positive sign indicate a positive relationship between the variables and values with a negative sign indicate a negative relationship.

It is possible to conclude that the most correlated variables are LenghtOfStay, StaysInWeekNights and StaysInWeekendNights, and PreviousStayss with PreviousBookingsNotCanceled and, in a lesser sense also with PreviousCancelations, PreviousCancelations with PreviousBookingsNotCanceled, and finally CanceledTime and LeadTime and CanceledTime inversely correlated with AgeAtBookingDate.

In fact, LenghtOfStay is the sum of the values for StaysInWeekendNights and StaysInWeekNights for each booking. Thus, only the variable LenghtOfStay (the days the customer was in a determined hotel) was chosen to be used as a predictor variable.

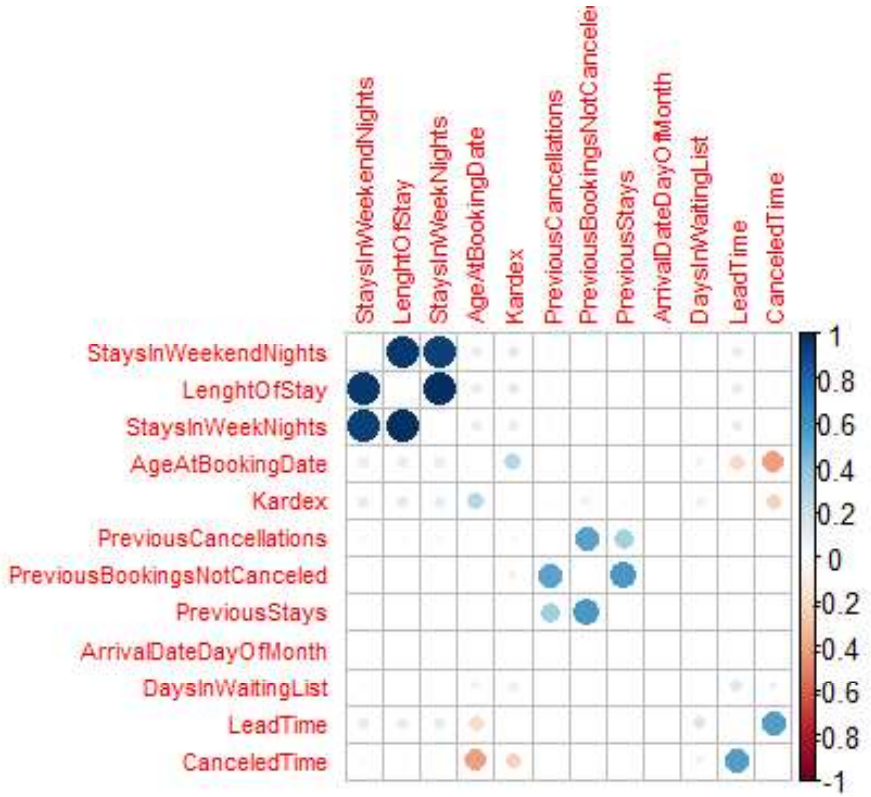

Figure 21 - Correlation matrix for the continuous variables

4.3. General conclusions for this chapter

After the exploratory study presented in the previous Sub-Sections, where visual tools were employed to study the association between the categorical variables IsRepeatedGuest, BookingDateDayOfWeek, WasInWaitingList, CustomerType, ArrivalDateMonth, IsVip, Agent, ArrivalDateDayOfWeek, AssignedRoomType, Company, Country, Kardex, MarketSegment, DistributionChannel, Meal,

ReservedRoomTypes. Variables like IsVIP and WasInWaitingList showed so few positive observations that we considered that they could be removed. In regard with variables Company, Country, Meal and ReservedRoomType show too many variations for the possible values that became meaningless in terms of predictive power. For example, each hotel has his specific values for the variables Meal and ReservedRoomType, so these variables also won't have a predictive power. Relatively to the categoric variables MarketSegment and DistributionChannel, these variables are specific by hotel. Each hotel has its sets of values that characterize these variables, so these variables only was used to study the cancellations for one HotelId specific.

The categorical variables chosen to be used as predictors for the two variables that will be studied as classification targets, (HotelType and IsCanceled) are: IsRepeatedGuest, BookingDateDayOfWeek, CustomerType, ArrivalDateMonth, MarketSegment and DistributionChannel. As for the numeric variables these are: LenghtOfStay, PreviousStayss, PreviousCancellations, PreviousBookingsNotCanceled, CanceledTime and LeadTime.

About segmentation modelling, the variables IsRepeatedGuest, IsCanceled, CustomerType HotelID, HotelType, BookingDateDayOfWeek, LenghtOfStay, CanceledTime, LeadTime, PreviousStayss, PreviousCancellations, PreviousBookingsNotCanceled and ArrivalDateMonth will be used and segmentation models.

5. Chapter 5 – Predictive modelling

One of the research questions of the present thesis is: What are the methods most used for client segmentation and how they change depending on the particular area of application? In Chapter 3, with the related literature review, it was possible to see that the most common methods used for Customer Segmentation (Table 2). Now, for the hotels' dataset described in the previous chapter (Chapter 4), the development of the refered models is presented and the achieved results are discussed individually. The models' results are compared with the ones found in the literature review.

This Chapter is divided into two main sections: Section 5.1 presents Classification modelling and Section 5.2 describes Segmentation models. All the models were developed using R and R packages.

5.1. Classification Models

In the following Section we will try to understand the predictive power of the variables in the dataset for classification tasks using two different perspectives: a) the classification of observations into a Hotel Type - City or Resort - using the predictors previously selected, and b) classification of bookings as canceled or not canceled. Taking in consideration both the previous exploratory analysis and the clustering results, the features used for predictive analysis will vary. In Sub-Section 5.1.1 uses all the dataset as like the predictors selected in Chapter 4 for the both perspectives, HotelType and IsCanceled bookings. Sub-Section 5.1.2 studies the classification of canceled or not canceled bookings using two different scenarios: the first uses observations separating the cancellation prediction by HotelType; the second scenario makes predictions for each hotel (HotelID). In both scenarios, the categoric variables MarketSegment and Distribution Channel were added as predictors since, as already foreseen in Chapter 4, these two variables appear as good predictors for these scenarios.

The classification modelling is explored using several techniques: Decison Trees, namely CART (CT), Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), Support Vectors Machines (SVM) and Artificial Neural Networks (ANN). These algorithms were chosen with basis in the literature review presented in Chapter 3 of the present thesis.

The results presented in the following Sub-Sections and Tables were obtained using R and considering the algorithms already refered and the different targets. Evaluation metrics for comparative performance evaluation are the ones described in Chapter 2: Precision, Recall, F-Measure and Accuracy.

For each training, the dataset used was always randomly divided into a train set (70%) and a test set (30%). The models' performance was evaluated using K-Fold Cross Validation, by employing the R function *train* with k=10. The package used in R to build the models was the *caret* package.

5.1.1. Classification using all observations

This Section explores the previously stated classification tasks: a) the performance of the dataset for the classification of HotelType (binary classification: City or Resort hotel), b) the predictive power for the booking's cancellations (again a binary task: Canceled or NotCanceled).
As alredy refered, the dataset consists in 136887 observations, that were divided in 102665 observations for the train set and 34222 observations for the test set. The predictors used in this context were IsRepeatedGuest, IsCanceled, CustomerType HotelID, HotelType, BookingDateDayOfWeek, LenghtOfStay, CanceledTime, LeadTime, PreviousStayss, PreviousCancellations, PreviousBookingsNotCanceled and ArrivalDateMonth, in accordance with the feature selection made in Chapter 4.
Next, we can see the results of the models obtained from the application of the classification methods earlier mentioned for prediction of the two different targets: prediction targeting the type of the hotel in scenario 1 and cancellation of bookings in scenario 2.

Scenario 1: Classification of the type of the hotel (HotelType)

Table 4 – Results of classification models for HotelType prediction

Algorithms	**Target (HotelType)**	**Evaluation metrics**			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (average) (%)
CT	City	90.14	67.71	77.33	73.36
	Resort	56.3	84.89	67.7	
NB	City	70.12	68.16	69.13	68.42
	Resort	66.7	68.7	67.69	
RF	City	86.85	84.57	85.7	85.38
	Resort	83.88	86.25	85.05	
LR	City	78.83	72.49	75.52	74.24
	Resort	69.57	76.38	72.81	
ANN	City	83.13	77.76	80.35	79.51
	Resort	75.82	81.55	78.58	
SVM	City	84.09	71.45	77.26	74.93
	Resort	65.53	80.07	72.07	

Table 4 presents evaluation results for the different employed techniques for modelling the classification of the target categorical variable, HotelType. The rows stand for the algorithms used for classification, further subdivided as the classes of hotel (Cuty and Resort).

The evaluation metrics of the algorithms are presented as columns (Precision, Recall, F-Measure and Accuracy). As it is possible to see that, not only the Random Forest is the model that shows the best performance regarding accuracy, with 85.38%, But also it is the most balanced one in terms of precision and recall, which reflects in the best AUC value. Naïve Bayes is the one presenting the worst performance, with 68.42% of accuracy. In fact, the top-three results are achieved by the RT, ANN and SVM models (respectively), the latter closely followed by the Logistic Regression model (LR).

Scenario 2: Classification of Cancelations
The dataset employed for studying this scenario is the same as the one used in the previous scenario. The results of the classification models are presented in Table 5.

According to the evaluation metrics used, most of the algorithms presented good performance, highlighting RF with an average accuracy of 99.98% and 99.6% AUC average.
The model built using the Naïve Bayes technique is still the one presenting the worst performance, 74.02% of accuracy, far from the remaining models' values.

Table 5 - Results of classification models in scenario 2 – canceling prediction

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	99.89	100	99.94	99.97
	NotCanceled	100	99.96	99.98	
NB	Canceled	0.146	52	0.29	74.02
	NotCanceled	99.95	74.03	85.07	
RF	Canceled	99.93	99.97	99.95	99.98
	NotCanceled	99.99	99.97	99.98	
LR	Canceled	99.89	100	99.94	99.97
	NotCanceled	100	99.96	99.98	
ANN	Canceled	99.87	100	99.93	99.97
	NotCanceled	100	99.96	99.98	
SVM	Canceled	99.89	100	99.94	99.97
	NotCanceled	100	99.96	99.98	

Comparison of Scenarios 1 and 2
According to Tables 4 and 5, better results were obtained for scenario 2 than for scenario 1, that is, using the same data, classifications for target HotelType are not as good as classifications for canceled bookings (variable IsCanceled). As it has been already seen, for both scenarios, RF is the algorithm presenting the best performance and NB the worst. Tables 6 and 7 show the confusion matrices obtained for RF algorithm and NB algorithm for scenario 1, respectively. It's possible to see that RF misclassified 6,63% of the observations as being Resort when they were City, and misclassified 7,99% of the

observations as City hotels which were Resort hotels, otherwise NB misclassified 15,06% of the observations as Resort hotels which were City hotels and 16,5% of the observations as City hotels which were Resort hotels.

Table 6 – Random Forest confusion matrix in scenario 1

Prediction	Actual	
	City	**Resort**
City	14985	2734
Resort	2268	14235

Table 7 - Naïve Bayes confusion matrix in scenario 1

Prediction	Actual	
	City	**Resort**
City	12098	5651
Resort	5155	11318

Note that, the fact that RF algorithm outperformed the remaining algorithms, for both scenarios, is inline with the findings in (Nguyen, King, & Subramanian, 2016) that compared the performance of RF, CT and NB for a census income prediction.

5.1.2. Study of cancellations

In this sub-section, the cancellations of bookings are studied considering two perspectives. First, in scenario 3, the results of the application of the classification algorithms to a dataset with only one type of hotel are presented. In this case, the dataset consisted in 67784 observations, divided into 50828 for the train set and 16946 for the test set. In scenario 4, other perspective is presented: the performance for the classification of the cancellations considering only the specific hotels is studied. For this case, a dataset with only one HotelID was used. Since the remaining results are very similar, we will only present the performance values for some of the hotels, namely: two for city hotels and two for reort hotels. The variables used in this case will be the same as the ones employed for scenario 3, and the categorical variables MarketSegment and DistributionChannel were also used.
The goal was to study the performance of the models when considering only each hotel type and each hotel id, that is, the difference between using the specific hotel PMS dataset against using the merged datasets by type.

Scenario 3: Observations with one HotelType (Resort hotel)
In this scenario, cancellations were studied considering only one specific HotelType, either City or Resort. The structure of the tables is still the same of the previous scenarios and the target is the categoric variable IsCanceled.
According to the evaluation metrics' results presented in Table 8, most of the models persent accuracies above 99%, with the already usual exception of the Naïve Bayes

model, with 86,54%. These accuracy values are again very high, which can be taken as indicating that the models might be overfitting.

Table 8- Results of classification models in scenario 3

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	99.7	100	99.85	99.96
	NotCanceled	100	99.95	99.97	
NB	Canceled	0	100	0	86.54
	NotCanceled	100	86.54	92.78	
RF	Canceled	99.87	100	99.99	99.98
	NotCanceled	100	99.97	99.99	
LR	Canceled	99.78	100	99.89	99.97
	NotCanceled	100	99.97	99.98	
ANN	Canceled	99.78	100	99.89	99.97
	NotCanceled	100	99.97	99.98	
SVM	Canceled	99.65	100	99.83	99.95
	NotCanceled	100	99.95	99.97	

In this view, it was decided to investigate what might be the cause for this behavior and it was found that the cause may be related with the numeric variable CanceledTime, since it might be influencing the training causing leakage. This new training version will be called scenario 3.2 and the previous scenario will be called scenario 3.1.

Table 9 - Results for classification in scenario 3 after removing CanceledTime

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	18.94	32.06	23.82	81.97
	NotCanceled	92.98	86.78	89.77	
NB	Canceled	0	0	0	85.11
	NotCanceled	99.98	85.12	91.96	
RF	Canceled	5.68	38.46	9.9	89.09
	NotCanceled	98.92	89.89	94.19	
LR	Canceled	0	0	0	85.09
	NotCanceled	99.97	85.12	91.95	
ANN	Canceled	16.59	28.07	20.85	81.27
	NotCanceled	92.57	86.4	89.38	
SVM	Canceled	0	0	0	85.12
	NotCanceled	100	85.12	91.94	

Table 9 shows the results of the application of the methods to the new predictor dataset (that is, without the variable CanceledTime) and it is possible to see that overfitting as been overcome, since the accuracy values that are now around 83%, and this time for all the methods.

The class Canceled presents the worst precision values for all algorithms, with Classification Trees being the model with the best precison for this class. This means that the models classified most Canceled bookings as NotCanceled when they were, in fact, Canceled. In Figure 22 the accuracy of the train set for each fold is presented. Almost all algorithms show an accuracy above 90%, except NB which falls behind by a large margin. Figure 23 presents the accuracy for each fold in scenario 3.2, showing that the leakage has been contained: the accuracy values are always below 90%, which indicates that the model will generalize better.
NB and LR not only present the worst accuracy in each fold, but it is due to the fact that the Recall and Precision are null for the Canceled class, i.e., the models considered that all the canceled bookings would not be not canceled. This may be due to fact that there is an important balancing issue, with the number of canceled reservations being much lower than the number of canceled reservations. So, the probability of a reservation being canceled is very small compared to the probability of the reservation not being canceled. Although, this seems not to have influenced so severly the better performing algorithms, it should be take into account.
Note also that, contrary of what happens in Figure 23, for scenario 3.2, the remaining algorithms present very similar accuracy values.

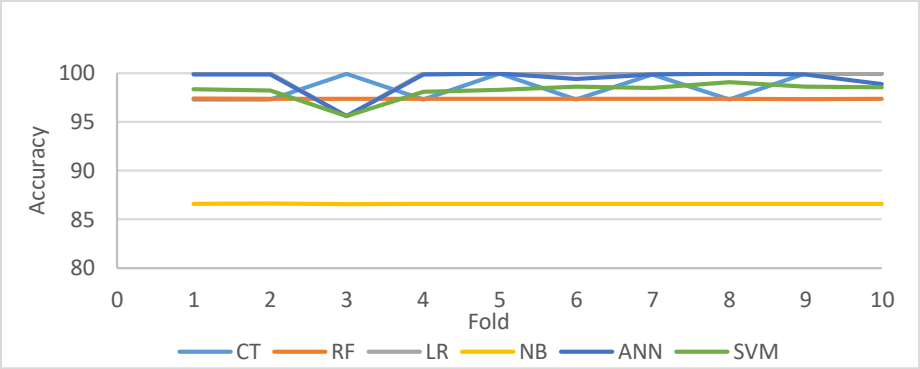

Figure 22 - Accuracy for scenario 3.1 before removing CanceledTime (by fold)

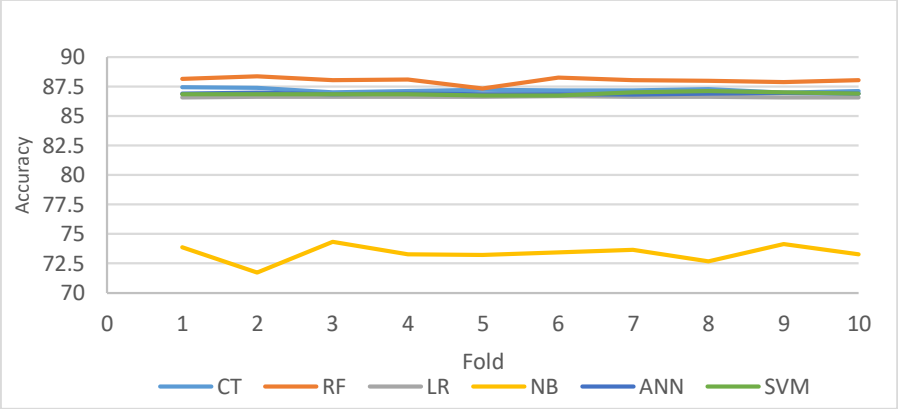

Figure 23 – Accuracy for scenario 3.2 after removing CanceledTime (by fold)

Scenario 4: Observations with one HotelID

In the following scenario, the cancellations were studied considering for each HotelID. Since the remaining results were very similar, we are only presenting the application of the algorithm in two City hotels and two Resort hotels.

In Table 10 the results of the application of the same algorithms shown in the previous Section to the observations of HotelId H4 are presented. This is a Resort hotel, with 8782 observations that were split in 5918 observations for the training phase and 2864 observations for the test. Note that these last observations of the test correspond to bookings dated after the bookings of the training set to prevent leakage, that is, information about the future to be envolved in the training.

Table 10 - Results of classification models in scenario 4 – H4

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	99.13	87.19	92.78	96.93
	NotCanceled	96.38	99.77	98.05	
NB	Canceled	97.37	84.22	90.32	95.84
	NotCanceled	95.46	99.32	97.35	
RF	Canceled	95.09	93.13	94.09	97.63
	NotCanceled	98.26	98.77	98.51	
LR	Canceled	99.29	88.85	93.78	97.38
	NotCanceled	96.9	99.82	98.34	
ANN	Canceled	98.07	90.59	94.18	97.59
	NotCanceled	97.47	99.51	98.48	
SVM	Canceled	99.29	88.85	93.78	97.38
	NotCanceled	96.90	99.82	98.34	

According to the evaluation values presented in Table 10, the performance of the algorithms increased when compared with scenario 3.2. For this specific hotel, the average accuracy was of 97,13%. This means that some variables in the dataset must differ between hotels (or, at least, between hotel's types) and, by biuiding an individual model, that is, a model per hotel, the variables predictive power increases.
We can see in Figure 24, for the training phase, the accuracy in each fold was close to 100%. In this case CT shows the lowest values of accuracy in folds 4, 5, 6 and 7. In fold 9, SVM was the algorithm with the lowest accuracy.

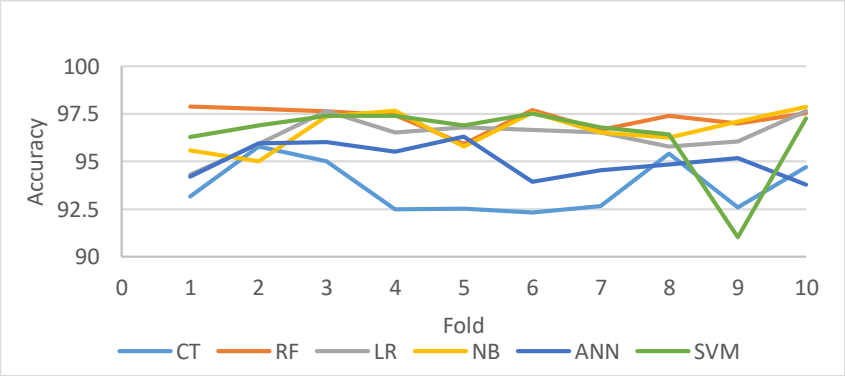

Figure 24 - Accuracy for scenario 4 – H4 (by fold)

This experiment was repeated for HotelId H8, which is still a Resort hotel, with 10758 observations that were splitted in 7033 observations for the training phase and 3725 observations for the test. The results, compared with the previous scenario, were not as good. For example, for the Classification Trees, we had an 88.46% of accuracy, with 21.06% of Precision in Class Canceled what means that 356 observations that were Canceled were classified as NotCanceled otherwise for this same method the Precision for Class NotCanceled was high, only 74 observations were classified as Canceled when

they were actually NotCanceled. In this the algorithm with best results was SVM with an accuracy of 97.34%.

Table 11 - Results of classification models in scenario 4 - H8

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	21.06	56.21	30.65	88.46
	NotCanceled	97.74	89.98	93.70	
NB	Canceled	0.22	11	0.43	87.7
	NotCanceled	99.76	87.89	93.45	
RF	Canceled	15.08	48.92	23.05	87.81
	NotCanceled	97.83	89.32	93.38	
LR	Canceled	13.74	63.27	22.59	88.59
	NotCanceled	98.9	89.27	93.84	
ANN	Canceled	17.74	69.57	28.27	89.1
	NotCanceled	98.93	89.73	94.12	
SVM	Canceled	93.89	100	96.85	97.34
	NotCanceled	100	95.49	97.69	

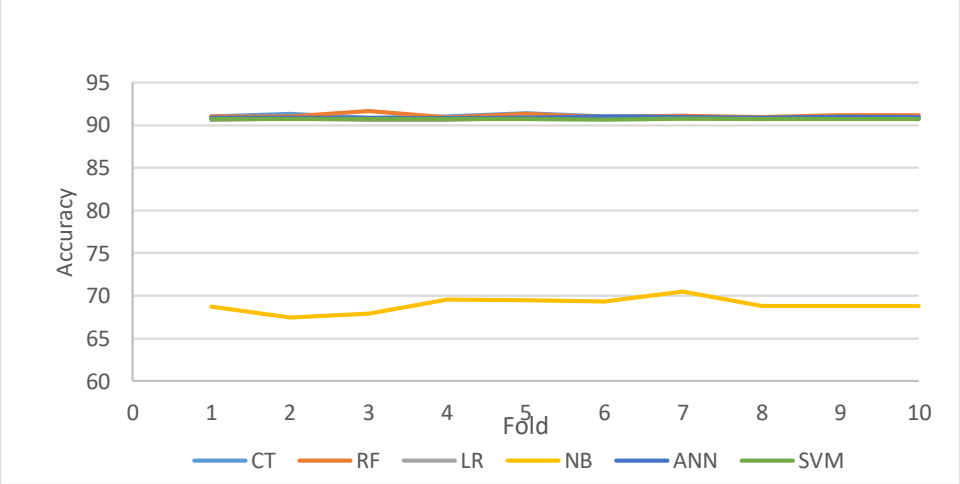

Figure 25 - Accuracy for scenario 4- H8 (by fold)

In the Table 12 it is presented the results of application of those methods in observations of HotelId H2, this hotel is a City hotel with 8782 observations splitted in 1869 observations for the test train and 6913 observations for the train set. In this case the mean of accuracy was 75,24% low compared to the HotelId previously analysed. The method

with the best accuracy was LR with 77.64% of accuracy, for all method the Class Canceled show low values of Precision which result in these low accuracy values.

Table 12 - Results of classification models in scenario 4 - H2

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	10.68	36.72	16.55	74.64
	NotCanceled	94.33	77.43	85.05	
NB	Canceled	0.14	85.71	0.27	76.73
	NotCanceled	99.93	7.7	86.8	
RF	Canceled	29.55	35.91	32.42	71
	NotCanceled	83.76	79.43	81.54	
LR	Canceled	18.86	57.64	28.42	77.64
	NotCanceled	95.73	79.3	86.75	
ANN	Canceled	17.95	42.47	25.24	74.96
	NotCanceled	92.51	78.55	84.96	
SVM	Canceled	0.22	100	0.45	76.51
	NotCanceled	100	76.5	86.68	

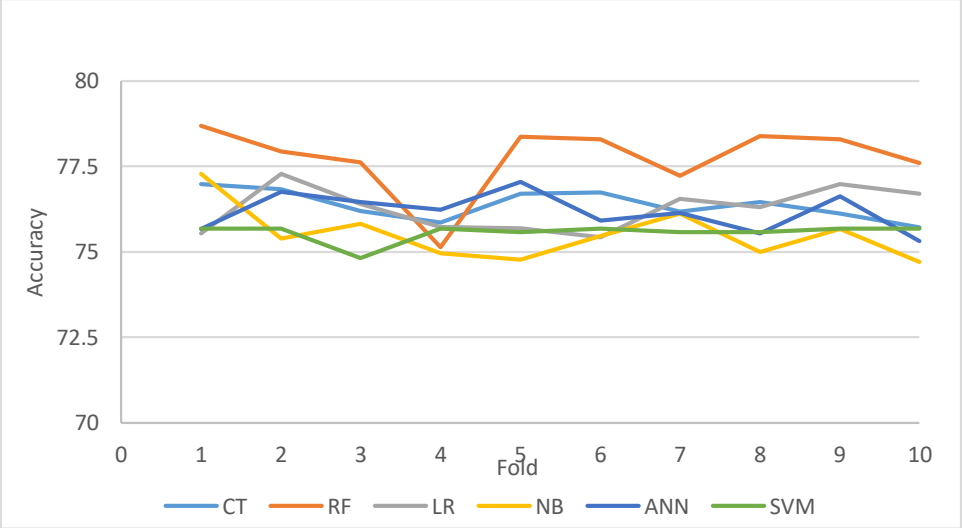

Figure 26 - Accuracy for scenario 4 - H2 (by fold)

Lastly, the same methods were used for observations of HotelId H6, a City hotel with 10758 observations, splitted in 2534 observations into test and 8224 into train. The accuracy average in this case was 73,07%, which is low when compared with the previous case. However, in this case, the precision values were highest than the previous case. For

example, for the LR method, the class Canceled showed a precision of 52,82% while in scenario 4 using HotelId H2, that is a City hotel too, the precision for the Canceled class was of 18,86%.

Table 13 - Results of classification models in scenario 4 - H6

Algorithms	Target (IsCanceled)	Evaluation metrics			
		Precision (%)	Recall (%)	F-measure AUC (%)	Accuracy (%)
CT	Canceled	42.22	86.72	56.79	74.9
	NotCanceled	95.85	72.12	82.31	
NB	Canceled	33.03	96.75	49.25	73.4
	NotCanceled	99.29	69.81	81.98	
RF	Canceled	47.58	66.24	55.38	70.05
	NotCanceled	84.45	71.53	77.46	
LR	Canceled	52.82	68.09	59.49	71.9
	NotCanceled	84.13	73.56	78.49	
ANN	Canceled	46.26	79.65	58.53	74.39
	NotCanceled	92.42	72.84	81.47	
SVM	Canceled	49.49	74.92	56.61	73.79
	NotCanceled	89.37	73.40	80.61	

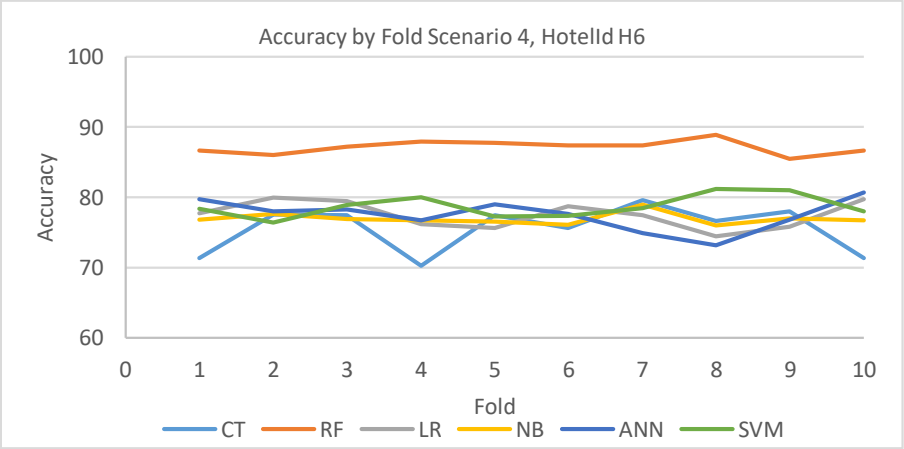

Figure 27- Accuracy of the models for scenario 4 – H6 (by fold)

Comparison of the cancelation classification scenarios

Comparing only the scenarios where cancellations were studied, is scenario 3.1, it was found that, after having obtained very high percenatge values for the performance of the methods, the data would probably be undergoing leakage, and, after some exploration of

where this leakage might be ocorring, it was decided to remove the numeric variable CanceledTime.
A new approach was taken, scenario 3.2, and comparing these two scenarios we can conclude that 3.2 shows accuracy values more realistic than scenario 3.1. RF was the method with the highest accuracy value, 89,09%, and NB was the method with the worst accuracy.
In scenario 4, four different hotels are studied *per se*: two resort hotels and two city hotels. The HotelId with the best average accuracy result was H4 with 97,125% and the method with the best accuracy for this HotelId was SVM, presenting 100% value for precision for the NotCanceled class and 93.89% for the Canceled class. The HotelId with the lowest average accuracy was H6, with 73,07%.
Comparing the performance of SVM that was the method with the best accuracy for HotelId H4. Using HotelId H6, SVM achieved a precision of 49.49% for the Canceled class and 89.37% for the NotCanceled class.
Now, comparing the scenario 3.2 and scenario 4 (using the HotelId H4), it's important to rembember that scenario 3.2 concerns the merged dataset for all the hotels with HotelType Resort while, in scenario 4, only the H4 dataset is considered. For scenario 3.2, the methods presenting the best results were RF and SVM, with 89.09% and 85,12% respectively. For scenario 4, the same methods, RF and SVM, present 97.63% and 97.38%, respectively. Thus, when hotels are studied individually, that is, using only the respective PMS data, better models are built in the sense that more particularities are learned.

For a particular note, according to the authors of (Linder et al., 2004), ANN outperformed CT e LR when applied to a small dataset. They affirmed that CT and LR should be used for larger datasets. Aanalysing the performance of these algorithms for our case we get that, in scenario 1, which is a scenario where a larger dataset was used, ANN outperformed both CT and LR. In scenario 2, which had the same quantity of data as scenario 1, these three algorithms presented the same performance. In scenario 3.2, CT outperformed both LR and ANN. Finally, in scenario 4, with the "smallest" data, ANN outperformed LR and CT. In fact, when a small quantity of data was used, ANN outperformed both CT and LR.

5.2. Segmentation Models

In the following section, segmentation models are studied. A combination of Self-Organizing Maps and K-Means has been proposed [14] under the conjecture that it will prove to be slightly better than a conventional combination of Ward's minimum variance and K-Means and this section will explore this claim.

In Sub-Section 5.2.1, the K-means algorithm is applied. In Sub-Section 5.2.2. the combination of K-Means and SOM is experimented with and, in Sub-Section 5.2.5, the combination of Ward's minimum variance and K-Means is studied.

The features considered in this Section were limited to LeadTime, BookingDateDayOfWeek, ArrivalDateMonth, LengthOfStay and PreviousStays since, during the experiments it was found that the remaining variables that were previously

signaled, in Chapter 4, for segmentation modelling were not relevant because they presented similar averages in the final analysis of the clusters obtained.

5.2.1. K-Means

The first step when using K-means clustering is to indicate the number of clusters, k that will be generated in the final solution. The algorithm starts by randomly selecting k objects from the dataset to serve as the initial centers for the clusters. To determine the optimal number of clusters it should be used in our cases, it will be used the Elbow method and Silhouette method.

Elbow method

Elbow criterion is a way to define the best k in K-means clustering algorithm. Using the method *wss* in RStudio, the graphic presented in Figure 28 was obtained, where the average within Centroids Distance for each k is plotted. Then we picked the k where the graph shows a more notorious angle and after which it is mostly stable. As it possible to see (Figure ~~28~~), this method suggested that 3 is the best value for the number of clusters.

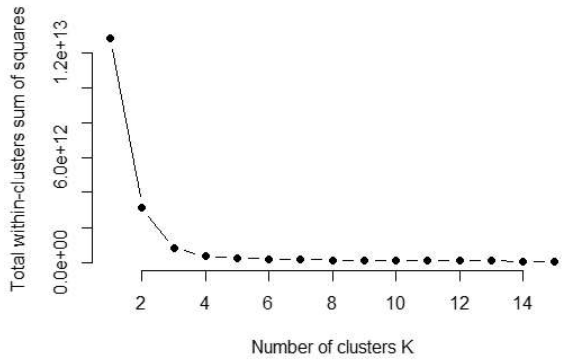

Figura 28 – Elbow method study

Silhouette method

Silhouette is another method that helps to indicate the number of clusters that might be more appropriate for k-Means. The method *silhouette* in the *cluster* package was used to compute the average silhouette width. The results presented in Figure 29 show that using k = 2 maximizes the average silhouette values, with k = 3 coming in as second indication for the number of clusters to be searched for.

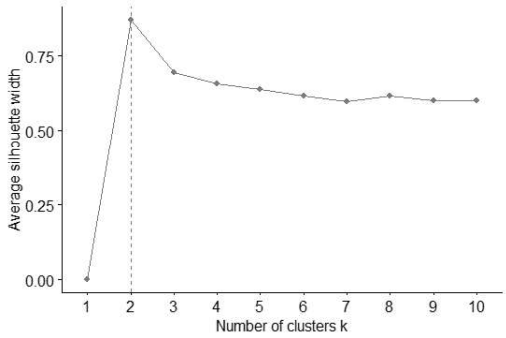

Figura 29 - Silhouette method study

The values of 2, 3 and 4 were chosen as parameter values for k. In order to try and get more informative attributes, k = 5 was also tested. The original dataset was limited so that only City hotels and not canceled bookings were used.

Using k = 2, the higest difference between the two clusters obtained is found in the feature LeadTime. Cluster 1 is characterized by customers who made the booking with a higher number of days prior to the arrival, with an average of 148 days. Otherwise, cluster 2 showed an average of 21 days. Another difference between these two clusters is the month of the arrival. For cluster 1, the customers arrive mostly in July and, for cluster 2, the customers arrive mostly in June.

Table 14 - K-means centers values for the city dataset (k = 2)

Attribute	cluster 1	cluster 2
LeadTime	148	21
BookingDateDayOfWeek	Wednesday	Wednesday
ArrivalDateMonth	July	June
LenghtOfStay	3	3
PreviousStays	4	4

Then, k = 3 was used and the results are shown in Table 15. We can see the average for each attribute by cluster. Analysing each of the clusters, we can higlight the following characteristics:

- *cluster 1*: In this cluster, the number of days prior to the arrival is 221. The day of week when the booking is made is Wednesday and the month of the arrival is August. The number of nights the customer stays at the hotel is three and the number of nights the customer stayed at the hotel prior to the current booking is four.
- *cluster 2*: The LeadTime in this cluster is 88 days. The day of week the booking was made is Wednesday and the arrival month is July. The number of nights the customers stayed at the hotel is three and most of the customers are not repeated

guests, but if he/she is a repeated guest, the number of nights the customer had stayed at the hotel prior to the current booking was three.

- *cluster 3*: The number of days prior to the arrival is 14, the day of week when the booking wis made is Wednesday and the arrival month is July. The number of nights the customer stays at the hotel is three, and most of the customers are not a repeated guest. In case of repeated guests, the number of nights the customer had stayed at the hotel prior to the current booking was four.

Table 15 - K-means centers values for city dataset (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	221	88	15
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	August	July	July
LenghtOfStay	3	3	3
PreviousStays	4	3	4

Table 16 presents the characterization of the clusters obtained using k=4. The feature LeadTime still makes the big difference between clusters: while cluster 1 presents an average of eight days between the day when the booking was made and the arrival day, cluster 2 shows an average of 239 days, cluster 3 an average of 54 days and cluster 4 an average of 118 days. The arrival date month in cluster 2 shows bookings placed in August, while in the rest of clusters show July as the arrival date month. The number of nights the customers stay at the hotels in cluster 2 is two and in the remaining clusters is three. The number of times the customer stayed at the hotel previous the current booking is the same in clusters 1 and 2 and difers from clusters 3 and 4 where the number of stays is three.

Table 16 - K-means centers values for city dataset (k = 4)

Attribute	cluster 1	cluster 2	cluster 3	Cluster4
LeadTime	8	239	54	118
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	July	August	July	July
LenghtOfStay	3	2	3	3
PreviousStays	4	4	3	3

After applying the 3 different k values for the K-Means algorithm, using a limited dataset with only not canceled bookings with the City hotels, it is possible to identify two different segments: customers who made the booking long ahead of time or close to the arrival date. The first ones usually stay at the hotel in July, for three days and had already

stayed at the same hotel three times before in average. The latter are customers who make their bookings closer to the date of the arrived, usually stay in June, with a average of nights of four. The average of previous stays in this segment is four times.

Now, we studied only one of th hotels in this limited data set, HotelId H5, still considering not canceled observations only. K-Means was used with k = 3 because of the previous analysis where we've seen three clusters as very different segments. The results presented in Table 17 seem similar to the previous case, since the three segments appear differentiated mainly by the LeadTime variable: for cluster 1 the average for the number of days booked prior to arrival was 12, for cluster 2 and cluster 3 were 28 and 97, respectively. The LengthOfStay variable also represents a differentiating feature in the segments. The arrival month does not vary much between segments: for cluster 1 and cluster 3 the arrival months are June and for cluster 2 the arrival month is June.

Table 17- K-means centers values - H5 (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	12	28	97
BookingDateDayOfWeek	Wednesday	Tuesday	Wednesday
ArrivalDateMonth	July	June	July
LenghtOfStay	3	2	4
PreviousStays	5	5	4

This analysis is similar when other city hotels are considered, as we can see from Table 18 for a different HotelId.

Table 18 - K-means centers values - H6 (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	22	18	12
BookingDateDayOfWeek	Wednesday	Wednesday	Tueday
ArrivalDateMonth	July	May	June
LenghtOfStay	3	3	3
PreviousStays	0	0	3

Now, the same experience was made but using data of Resort hotels. Table 19 presents the results obtained for k = 2, where the differences between the segments are not significative and LeadTime and ArrivalDateMonth are the variables that present more differences between the clusters.

Table 19 - K-means centers values for resort hotels dataset (k = 2)

Attribute	cluster 1	cluster 2
LeadTime	57	61
BookingDateDayOfWeek	Wednesday	Wednesday
ArrivalDateMonth	July	June
LenghtOfStay	6	6
PreviousStays	3	3

Using k = 3, a more significative difference can be seen in the clustering results. In cluster 1, we have a segment with a longer range of days since the reservation was made until the customer arrives at the hotel, the month of arrival is usually June, customers in this segment stay, on average, 9 nights in the hotel and have never been to that hotel before. In cluster 2 we have customers with a lower LeadTime and a shorter LengthOfStay, but the number of times they have been to the hotel is higher, five. The characterization of cluster 3 is similar to that of cluster 2, also presenting a lower LeadTime when compared with cluster 1.

Table 20 - K-means centers values for resort hotels dataset (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	103	49	57
BookingDateDayOfWeek	Wednesday	Wednesday	Tuesday
ArrivalDateMonth	June	July	July
LenghtOfStay	9	5	6
PreviousStays	0	4	3

Finally, k = 4 was tested and the results are shown in Table 21. Compared to the previous case, it appears that cluster 3 from Table 20 was divided into two (sub)clusters, represented in Table 21 as clusters 1 and 4. Looking at these two clusters in particular, we see that LeadTime on cluster 1 is higher than on cluster 4 and the number of times the customer has been to the hotel before is higher for cluster 4 than for cluster 1.

Table 21 - K-means centers values for resort hotels dataset (k = 4)

Attribute	cluster 1	cluster 2	cluster 3	cluster 4
LeadTime	58	49	103	52
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	July	August	July	July
LenghtOfStay	6	5	9	6
PreviousStays	2	3	0	6

Again, resort hotels were studied individually to understand if the segmentation varies between hotels. Table 22 presents the results of the application of K-Means using k = 3 to HotelId H4. We obtained differents segments, mainly distinguished by the LeadTime variable and the LengthOfStay variable. At cluster 2, we have customers who book with a larger number of days in advance and stay longer when compared to the other clusters. While cluster 1 customers make their reservation closer to their arrival date, they also present a significant length of stay.

Table 22 - K-means centers values for H4 dataset (k = 3)

Atributo	cluster 1	cluster 2	cluster 3
LeadTime	18	240	102
BookingDateDayOfWeek	Wednesday	Wednesday	Tuesday
ArrivalDateMonth	June	June	July
LenghtOfStay	7	8	6
PreviousStays	4	4	4

Table 23 shows the results of applying K-Means for HotelId H8. For this hotel, and unlike the previous hotel HotelId H4, customers with a higher LeadTime present a longer length of stay but fewer previous stays.

Table 23 - K-means centers values for H8 dataset (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	26	17	49
BookingDateDayOfWeek	Wednesday	Wednesday	Tuesday
ArrivalDateMonth	June	June	July
LenghtOfStay	4	3	6
PreviousStays	2	5	1

According to the analysis performed we can identify the following patterns:

- Customers who were previously at this hotel prior to a booking make the reservation with less time in advance and have a shorter stay at the hotel.
- Customers who have never been to this hotel before this booking make the reservation more days in advance and tend to book for a longer stay.

5.2.2. K-Means and SOM

This Sub-Section concerns the combination of K-means with SOM. Initially, SOM technique used, not only for the sake of reducing dimensions, but also for visualizing the clustering in a readable, easy and fast way. Then, we re-clustered the features resulting

from SOM using the K-means algorithm. The data set is the same that has been described in the previous Sub-Section.

Self Organizing Maps

Since present Sub-Section shows the application of SOM, the *kohonen* package from RStudio was used. To apply the *som* algorithm, you need to select the size and type of the map to be used. The shape can be hexagonal or squared, depending on the shape of the nodes you require. Typically, hexagonal grids are prefered since then the interior nodes have six immediate neighbours. Figure 30 presents a neighbour distance graphic. In this graphical representation, white areas are representing dissimilarity (large Euclidean distances between objects), while the red areas are representing similarity (small Euclidean distances between objects). The color degrades between white to red showing the reduction in distance between nodes. Thus, it is obvious that, from the algorithm's point of view, the observations are all very similar.

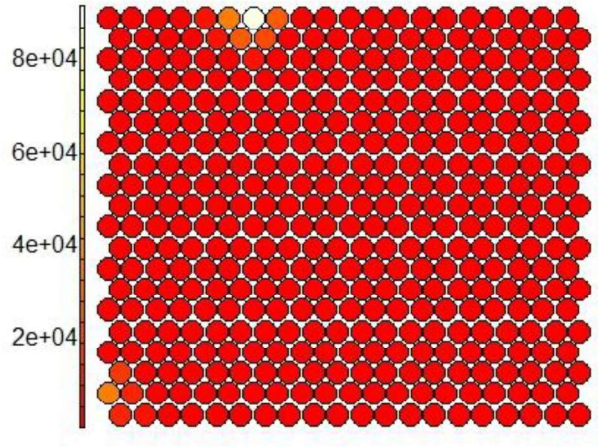

Figura 30 - Neighbour Distance plot for the SOM model

With an analysis of the results obtained from the junction of these two algorithms, we can see the following characteristics in each cluster:

- *cluster 1*: the customers show a longest period of Lead Time, what means they made their booking very early. Tipically, the day that the booking was made is Wednesday, the arrival day of week is also Wednesday and the arrival month is July. These customers stay a short period of time, usually three days, and they have been to the hotel an average of seven times before. The number of days prior to the arrival that the booking was made is high and the month of the arrival is an holidays month. We believe that the customers in this cluster are what the hotel calls transient customer, who could be in passing to other destinations and so they stay few days at the hotel. Another reason may be that, these being City hotels, where usually tourists stay less time before going to either resort hotels or simply move on to other cities.
- *cluster 2*: the number of days prior to the arrival where the booking was made is smaller when compared to cluster 1. The day of the week when the booking was made is still Wednesday. The number of nights the customer stayed at the hotel are five and most customers are not repeated guests. The number of nights the

customers stayed at the hotel prior to the current booking is nine. This cluster presents a higher length of stay when compared with cluster 1, and the same applies to the number of previous stays. We may conclude that this cluster is constituted by customers that want to stay for a longer period in the city or are business travelers.

Compared to Sub-Section 6.2.2, using these two algorithms together was obtained better results. For example, using only city dataset and not canceled bookings using and for K-Means (k = 2) both clusters showed the same LeadTime average, while using K-Means with SOM shows different values for this variable in the cluster centers.

Table 22 - K-means+SOM center values for city dataset (k = 2)

Attribute	cluster 1	cluster 2
LeadTime	151	29
BookingDateDayOfWeek	Wednesday	Wednesday
ArrivalDateMonth	July	June
LenghtOfStays	5	3
PreviousStays	7	9

Using k = 3, once again we have more different results compared to the same experience using just K-Means, in this case we have 3 segments with different average for the LeadTime variable and the PreviousStay variable.

Table 23 - K-mean+SOM center values for city dataset (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	17	87	211
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	July	July	August
LenghtOfStays	5	3	4
PreviousStays	8	10	7

In Table 24 it was presented the results using a k = 4, comparing with the results presented in Table 23 the results seems worst because when it was used k = 3 the clusters showed more relevant differences, for example using k = 4 for the variable PreviousStays all the 4 clusters had an average _ _of 5 days, but using k = 3 the 3 clusters obtained had a different values for the variable PreviousStays. For this reason, it was chosen a k equals 3 for the next experiences.

Table 24 - K-means+SOM center values for city dataset (k = 4)

Attribute	cluster 1	cluster 2	cluster 3	Cluster4
LeadTime	213	86	16	70
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	July	August	July	July
LenghtOfStay	3	4	5	3
PreviousStays	5	5	5	5

In Table 25 it is presented the results of application of K-means and SOM to HotelId H6 using a k = 3, this hotel is a City hotel and the results compared with when it was used just K-means for this hotel is quite better. For example, in the approach using only K-means the three clusters had the same length of stay while using K-Means and SOM cluster 3 had a length of stay smaller than the other clusters.

Table 25 - K-means+SOM center values for H6 (k = 3)

Attribute	cluster 1	cluster 2	cluster 3
LeadTime	13	14	20
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	June	May	July
LenghtOfStays	4	4	3
PreviousStays	3	5	3

In case of Resort, it was used one hotel id used before, H4 and compared with the same HotelId but with only K-means the variable PreviousStays is different in cluster 2 with the biggest average for the variable LeadTime and it was equals in cluster 1 and cluster 3 while when it was used only K-means this variable was equal in the three clusters.

Table 26 - K-means+SOM center values for H4 (k = 3)

Atributo	cluster 1	cluster 2	cluster 3
LeadTime	27	290	115
BookingDateDayOfWeek	Wednesday	Wednesday	Wednesday
ArrivalDateMonth	June	July	July
LenghtOfStays	5	7	7
PreviousStays	5	3	5

Comparison of the application of K-means alone and the join of K-means with SOM

Using only K-means it was obtained good clustering results, the clusters generated presents different patterns for the customers, these results were slightly better when it was added SOM algorithm to K-means. For example, using only K-means, some clusters had the same average for some variables, which does not allow distinguishing these clusters from each other. While using K-means with SOM it was possible to get more diverse clusters.

5.2.3. Agglomerative Hierachical Clustering

To perform Agglomerative Hierachical Clustering, it was first computed the dissimilarituy values with function dist and then the values were feeded into hclust function using the *ward.D2* method. Only 1000 random observations from train dataset were considered because of memory problem. In Figure 31 it is presented the resultant dendrogram, it is possible to see that the algorithm suggested three visible clusters. In our case, since the previous study (Section 5.2.1 from present Chapter) already had pointed 3 segments, this confirms that the most suitable k should be three.

Then, it was applied to the dataset the function kmeans and used the function table to analyzes the clusters. In Table 14, it is presented the table of clusters that were segmented by the variables: HotelType, IsCanceled and CustomerType, it is possible to see that in cluster 1 there is more observations with the category Resort hotel than City hotels, the number of observations not canceled is higher than the number of observations that were canceled. Transient customers are the ones that are most prevalent on this cluster and the Contract customers are the ones that are minus prevalent. It was analysed as well the attribute ArrivalDateMonth and it was possible to see that in this cluster, August, May and June are the months with greater affluence.

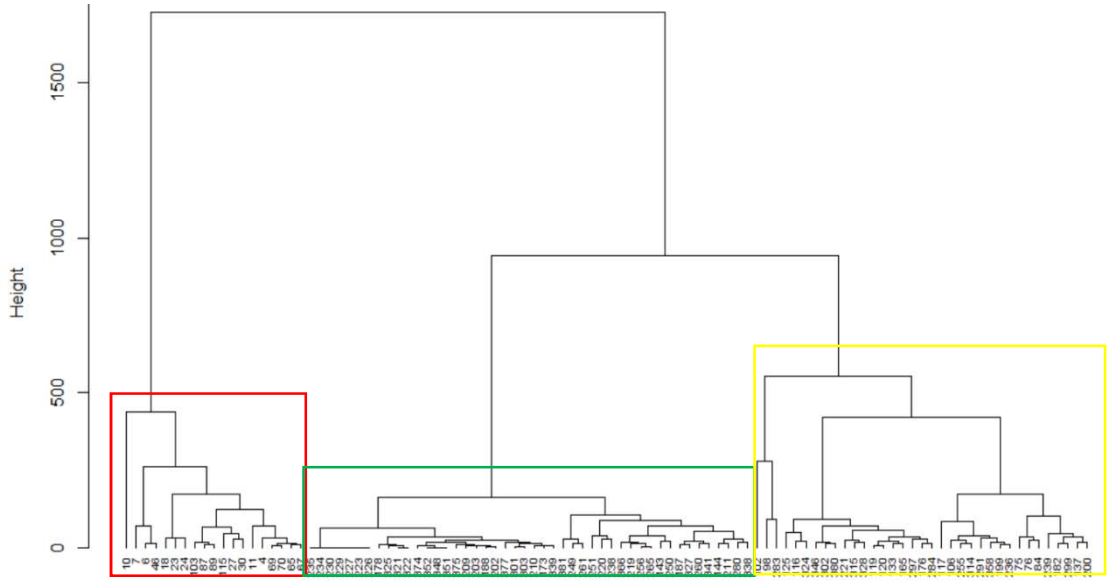

Figure 31 - *Agglomerative Hierachical Clustering considering 3 segments*

Relatively to cluster 3, the number of observations with category City hotels is higher than the Resort hotels, the number of cancelations is higher than the numbers of observations that were not canceled but comparing with the others clusters the number of cancelations is higher. In this cluster, the Transient customer still have the highest number of observations and the Group customers are the least prevalent comparing with other clusters. Relatively to the attribute ArrivalDateMonth, May, June, August and April are the months with the biggest affluence. Considering the attribute IsRepeatGuest, this cluster have the higger number of repeated customers comparing with the other clusters. Finally, the cluster 2, this cluster have less observations than the others, the number of observations with the category Resort hotels is higher than the number of observations with the category City hotels. The number of cancelations is less than the number of observations that were not canceled and the customers with the category Contract are the prevalent one in this cluster.

Table 22 - K-Means Clustering for target HotelType

Segment	Class	
	City	Resort
1	5508	17562
2	1511	4718
3	27522	11622

Table 23 - K-Means Clustering for target IsCanceled

Segment	Class	
	NotCanceled	Canceled
1	22419	651
2	5249	980
3	22899	16245

6. Chapter 6 – Conclusions and Recommendations

In the hospitality industry, a substancial amount of bookings comes from online plataforms and thus an effective customer segmentation is needed to better know each group of customers and improve their satisfaction and thus, their fidelity. In the present dissertation, a real case study was been researched, using a PMS dataset pertaining to eight Portuguese hotels from a Portuguese hotel chain: four hotels situated in Lisbon and four resort hotels in the Algarve region.

According to the previous literature review, data mining is used in several areas of application to perform customer segmentation. For the segmentation modelling of our case study, after an exploratory analisys, it was found that only a subset of the PMS variables showed greater predictive power, namely IsRepeatedGuest, BookingDateDayOfWeek, ArrivalDateMonth. The variable LenghtOfStay was choosed instead of StaysInWeekendNights and StaysInWeekNights, the variables PreviousCancellations and PreviousBookingsNotCanceled were excluded and it was only used the variable PreviousStays which results in the grouping of the other two variables.

The study of customer segmentation began by employing a K-means technique to study the characteristics of the dataset of bookings. In a second experiment, K-means and SOM where both used. Comparing the results obtained using only K-means and then using K-means and SOM, the results were slightly improved by joining the two algorithms.

Two different patterns have been identified: customers who were previously staying at the hotel prior to the present booking, that make the reservation closer to the arrival date and book a shorter stay at the hotel where a type of customers that booked more often in City hotels. They are tourists who stay for a short period of time or business travelers and, in this case, with last minute reservations in most cases. Since these customers tend to stay in the same hotel often, this may be justified by the fact that customers who often go to the same city on business and already stay at the hotel will make reservations at that same hotel. The other striking pattern identified was that of customers who have never been in that hotel before the present booking, make the reservation with more days in advance and tend to book a longer stay. Most of the bookings with these characteristics belong to resort hotels. This fact could be justified for example because that this type of customers is in vacations and usually came from countries other than Portugal and, to secure the booking, they book in great advance. The number of previous stays for this segment is smaller, as customers typically do not choose the same resort twice because they prefer to try different things on such long trips.

The predictive power of PMS system data was also investigated. The categorical variables chosen to be used as predictors for the classification targets – the type of the hotel and if a booking is likely to be canceled – are IsRepeatedGuest, BookingDateDayOfWeek, CustomerType, ArrivalDateMonth, LenghtOfStay, PreviousStays, CanceledTime and LeadTime. The variables MarketSegment, DistributionChannel was used only when it was analysed one specific hotel id.

The experiments allowed to provide answers for this dissertation's research questions, namely:

1. What methods are used for customer segmentation and how do they change depending on the specific application area (in the data / business domain)?

Several segmentation methods exist and with the review of the related literature we found that the most common are K-Means, SOM, Agglomerative Hierachical Clustering, and Density Based methods. No predominant technique per area was found, thus, there is no area-specific method.

2. What classification methods exist and if there is a method with better accuracy?

The classification methods found in the literature review and used were Decision Trees, Naive Bayes, Random Forest, Logistic Regression, Support Vectors Machines and Artificial Neural Networks. There is no method with better accuracy. The accuracy varies according to the scenario being studied and the size of the data being used.

3. Real case study in the hotel industry: is it possible to obtain segmentation based on the available data?

Yes, it was possible to get a segmentation model with the available data. Two major segments were identified. The first segment consists in the type of customer called by the hotels Transient, which are customers who stay for a short time in the hotel. In this segment it was found also the customer type Contract, which are business travelers and therefore already have contracts with the hotels. Other features were found in this segment, bookings last-minute, ie very close to the arrival date. The length of stay in this segment is very short compared to the other identified segment. On the other hand, another segment was identified with reservations made in advance, these reservations are very often in resort hotels and usually the customer's stay is much longer. These features point to customers who travel more with in family, a segment identified by hotels such as Group or Transient-Party.

4. Can a multi-method approach improve results?

We fiound out that, in the case of segmentation, using K-Means plus SOM has improved the findings and characterization of the clusters when compared to using K-Means alone. This indicates that this type of methodoly may, in fact, bring advantages for segmentation studies.

5. Would it be possible to obtain a general model for application in any hotel or should there be individualized models?

Our results confirm the findings in (Antonio, Almeida, & Nunes, 2017) that hotels should be studied separately, that is, each one studied independently of the others. Each hotel has its own particularities that may influence a model's predictive power when studied individually.

For future work, and in view of the these conclusions, the first recommendation is that, when studying hotel's predicitive models, each hotel must be studied separately. On

another hand, future studies must take into account the fact that unbalanced observations for canceled and not canceled bookings' classes may severely affect predictive classification models. Furthermore, in order to obtain a better segmentation, other variables, either internal or external should be considered to provide better insights on the costumer's drivers and habits.

6. References

Antkowiak, M. (2006). Artificial Neural Networks vs . Support Vector Machines for Skin Diseases Recognition. *Neural Networks*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.1206&rep=rep1&type=pdf

Antonio, N., Almeida, A. De, & Nunes, L. (2017). *Predicting hotel booking cancellations to decrease uncertainty and increase revenue*. *13*(2), 25–39. https://doi.org/10.18089/tms.2017.13203

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Bose, A., Munir, A., & Shabani, N. (2017). *A Comparative Quantitative Analysis of Contemporary Big Data Clustering Algorithms for Market Segmentation in Hospitality Industry*. 1–8. Retrieved from http://arxiv.org/abs/1709.06202

Breiman, L. (2001). (impo)Random forests(book). *Machine Learning*, 5–32. https://doi.org/10.1023/A:1010933404324

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications*, *34*(4), 2754–2762. https://doi.org/10.1016/j.eswa.2007.05.043

Chiu, C.-Y., Chen, Y.-F., Kuo, I.-T., & Ku, H. C. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. *Expert Systems with Applications*, *36*(3), 4558–4565. https://doi.org/10.1016/j.eswa.2008.05.029

Chung, K. Y., Oh, S. Y., Kim, S. S., & Han, S. Y. (2004). Three representative market segmentation methodologies for hotel guest room customers. *Tourism Management*, *25*(4), 429–441. https://doi.org/10.1016/S0261-5177(03)00115-8

Davies, F., Moutinho, L., & Curry, B. (1996). ATM user attitudes: a neural network analysis. *Marketing Intelligence & Planning*, *14*(2), 26–32. https://doi.org/10.1108/02634509610110778

Gaigerov, B. A., Elkina, L. P., & Pushkin, S. B. (1982). Metrological characteristics of a group of hydrogen clocks. *Measurement Techniques*, *25*(1), 23–25. https://doi.org/10.1007/BF00824883

Geerts, G. L. (2011). A design science research methodology and its application to accounting information systems research. *International Journal of Accounting Information Systems*, *12*(2), 142–151. https://doi.org/10.1016/j.accinf.2011.02.004

Han, J., Kamber, M., & Pei, J. (2011). Data Transformation by Normalization. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Han, S. H., Lu, S. X., & Leung, S. C. H. (2012). Segmentation of telecom customers based on customer value by decision tree model. *Expert Systems with Applications*, *39*(4), 3964–3973. https://doi.org/10.1016/j.eswa.2011.09.034

Hu, Y. H., Huang, T. C. K., & Kao, Y. H. (2013). Knowledge discovery of weighted RFM sequential patterns from customer sequence databases. *Journal of Systems and Software*, *86*(3), 779–788. https://doi.org/10.1016/j.jss.2012.11.016

Hu, Y. H., & Yeh, T. W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge-Based Systems*, *61*, 76–88. https://doi.org/10.1016/j.knosys.2014.02.009

Hutcheson, G. D. (2011). (2011). *Hutcheson, G. D. (2011). Logistic Regression. In L. Moutinho and G. D. Hutcheson, The SAGE Dictionary of Quantitative Management Research. Pages 173-175*. 3–5.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, *26*(2), 181–188. https://doi.org/10.1016/S0957-4174(03)00133-7

Kaur, D., & Paul, A. (2014). *Performance Analysis of Different Data mining Techniques over Heart Disease dataset*. *4*(1), 220–224. Retrieved from http://inpressco.com/wp-content/uploads/2014/02/Paper40220-224.pdf

Kim, S.-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, *31*(1), 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

Kuechler, B., & Petter, S. (2017). Design Science Research in Information Systems. *Springer, Berlin, Heidelberg*, 1–66. https://doi.org/10.1007/978-3-642-29863-9

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of Self-Organizing Feature Map and K-Means Algorithm for Market Segmentation. *Computers & Operations Research*, *29*(11), 1475–1493. https://doi.org/10.1016/S0305-0548(01)00043-0

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113–1130. https://doi.org/10.1016/j.csda.2004.11.006

Legohérel, P., Hsu, C. H. C., & Daucé, B. (2015). Variety-seeking: Using the CHAID segmentation approach in analyzing the international traveler market. *Tourism Management*, *46*, 359–366. https://doi.org/10.1016/j.tourman.2014.07.011

Liao, K. H., & Chueh, H. E. (2011). Applying Fuzzy Data Mining to Telecom Churn Management. *Communications in Computer and Information Science*, *134*(PART 1), 259–264. https://doi.org/10.1007/978-3-642-18129-0_41

Linder, R., Geier, J., & Kölliker, M. (2004). Artificial neural networks, classification trees and regression: Which method for which customer base? *The Journal of Database Marketing & Customer Strategy Management*, *11*(4), 344–356. https://doi.org/doi:10.1057/palgrave.dbm.3240233

Nguyen, N., King, B., & Subramanian, A. (2016). *Benchmarking Random Forest against Naive Bayes*. 1–12.

Ouchi, S., & Takato, S. (2011). A T X for Mathematics High-Quality Statistical Plots in L Education Using an R-Based KETpic Plug-In. *Yang, WC and Majewski, M and DeAlwis, T and Hew, WP*, 1–10.

Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, *20*(2), 134. https://doi.org/10.2307/3151680

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *Bulletin of the Technical Committee on Data Engineering*, *23*(4), 3–13. https://doi.org/10.1145/1317331.1317341

Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*, *14*(5), 923–939. https://doi.org/10.3846/16111699.2012.749807

Swenson, E. R., Bastian, N. D., & Nembhard, H. B. (2016). Data analytics in health

promotion: Health market segmentation and classification of total joint replacement surgery patients. *Expert Systems with Applications*, *60*, 118–129. https://doi.org/10.1016/j.eswa.2016.05.006

Thakur, B., & Mann, M. (2014). Data Mining for Big Data : A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(5), 469–473.

Wang, Z., Tu, L., Guo, Z., Yang, L. T., & Huang, B. (2014). Analysis of user behaviors by mining large network data sets. *Future Generation Computer Systems*, *37*, 429–437. https://doi.org/10.1016/j.future.2014.02.015

Wei, J.-T., Lin, S.-Y., Weng, C.-C., & Wu, H.-H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, *39*(5), 5529–5533. https://doi.org/10.1016/j.eswa.2011.11.066

Ziafat, H., & Shakeri, M. (2014). *Using Data Mining Techniques in Customer Segmentation*. *4*(9 OP-International Journal of Engineering Research and Applications, Vol 4, Iss 9, Pp 70-79 (2014)), 70. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=edsdoj&AN=edsdoj.f6c32f5214418687339b910d6ed864

Kassambara, 2013