TOURISM IN FUNCTION OF DEVELOPMENT OF THE REPUBLIC OF SERBIA

Tourism in the Era of Digital Transformation

UNIVERSITY OF KRAGUJEVAC
FACULTY OF HOTEL MANAGEMENT AND TOURISM IN VRNJAČKA BANJA
DATA ANALYSIS APPLICATIONS IN TOURISM AND
HOSPITALITY MARKETING

Dragana Ćamilović;

Abstract

The hospitality industry is highly competitive today, and many companies have a customer focus. Successful marketing starts with recognizing and tracking patterns within customer data. For each hospitality company it is essential to determine which customer segment it wants to serve. Marketing segmentation is often performed by using cluster analysis which is presented in this paper. It is a known fact that it costs more to acquire a new customer than to retain an existing one. Thus, a special attention is given to retention models. Models which predict who and why is likely to churn are discussed in the paper. Sentiment analysis becomes especially important with the development of social networks, since it identifies the sentiments and opinions in a text. Data analysis can also be used for predicting next best offers (NBOs). NBOs can boost revenues with cross-selling and up-selling, and improve customer relationships as well.

Key Words: hospitality marketing, data analysis, customer segmentation, retention models, sentiment analysis, next best offers

JEL classification: Z33

Introduction

Today’s businesses generate large amounts of data, such as profile data, purchase history, channel usage, browsing history, Internet clickstreams, social media data, etc. These belong to one of two broad categories:

Structured data: highly organized data that uploads neatly into a relational database (Schaefer, 2016), such as customer buying habits and other transactional data. These are often processed and stored in a data mart or a data warehouse, as described by Ćamilović et al. (2009).

---

1 Dragana Ćamilović, PhD, Professor, The College of Hotel Management, Kneza Višeslava 70, Belgrade, E-mail: dragana.camilovic@vhs.edu.rs
Unstructured data: does not reside in a traditional database and may have its own internal structure (Schaefer, 2016). Unstructured data can be acquired from emails, documents and various social media sites. Analysis of social media data gives us a valuable insight into individual customers’ preferences, their likes and dislikes (Dubey & Nainwani, 2014).

Structured, semi-structured and unstructured data can be stored in a data lake – a storage repository that holds a vast amount of raw data in its native format, without a specific purpose in mind (Harris, 2016). This means that data structures and business requirements do not have to be defined until the data is needed (op. cit.).

Both structured and unstructured data are used extensively in data analysis.

Data Analysis

Data analysis involves employment of different data mining techniques, depending on business objectives. Regardless of the data mining technique used, there are several steps in each data mining project (Čamilović, 2008):

1. Defining the business objectives. Each data mining application has its own business objectives and requirements. Talking to business people is the best way to define them.

2. Assembling and preparing the data. When the data needed resides in multiple data repositories, this step takes a lot of time and effort. Variable selection is a critical step within the data mining process. Usually, the variables have to be transformed in accordance with the requirements of the data mining algorithm chosen.

3. Modelling. Data can be presented to a data mining tool, but modelling phase very often involves coding in some data analysis programming language (Python and R are the most popular).

4. Interpreting the results. The model should be evaluated with respect to problem solving objectives and discussed with business experts.

5. Applying the results. The goal of data mining is to apply what was discovered in order to solve the business problem defined in the first step (e.g. customer churn).

Some data science techniques are used more often than the others. Most popular techniques used in 2017 are presented in Figure 1.
Some of the techniques are very popular in marketing: regression analysis and decision trees are useful for product propensity scoring, clustering is often used for customer segmentation, text mining is valuable to sentiment analysis etc.

As shown in Figure 2, data analysis in marketing can be used for different purposes: customer segmentation, retention models, sentiment analysis, and next best offer. These are described in more detail in the sections that follow.

Figure 2: Data analysis applications in marketing
Customer Segmentation

Most markets are not homogeneous, because different people have different preferences. This is why it cannot be served by one type of product/service offering (Jain, 2000). It is very important for a company to decide which customer segments it wants to connect with and serve.

A market segment is “a portion of a larger market whose needs differ somewhat from the larger market” (Hawkins, 2010, p. 16). Customer segmentation is grouping similar customers together, based on many different criteria and variables (Ćamilović, 2008). Depending on the major segmentation variables, there are four different types of market segmentation: geographic, demographic, psychographic, and behavioural segmentation (Kotler & Keller, 2016).

Geographic segmentation divides a market into different geographical entities. There are several approaches to geographic segmentation. A market can be divided by geographical areas, such as city, state, region, or country. It can also be divided into rural, suburban, and urban market segments. And, finally, a market can be divided by climate or total population in each area (Study.com, 2017). Some methods combine geographic with demographic data in order to get an even better segment description and understanding.

Demographic segmentation divides a market by demographic indicators, such as age, gender, income level, education, family size, life stage, etc. Many purchase behaviours are influenced by how old the customers are, because the purchase of goods and services changes over time. But, most importantly, the year of birth can also predispose the individual’s lifelong preferences (Hsu & Powers, 2002).

When segmenting a business market, restaurants must take age and gender into account. Age is an important factor that determines when and where customers eat in terms of location, time of the day, underlying food principles of health, convenience, variety, etc. (Scott-Thomas, 2012). A marketing research conducted in the USA has shown that marketers should target the over-65s with healthy food and beverages and under-45s with snack and late night meals, and also that Generations X and Y want their meals at their fingertips (op.cit.).
It is also a well-known fact that men and women have different eating preferences. Women are more concerned about nutrition, they are more adventurous in food choices than men, but eat out less often. Men appreciate familiar food, they are meat lovers and do not like fruits as much as women (Hsu & Powers, 2002).

New generations are well educated. There is a tendency for openness to try new foods among customers exposed to education (McCluskey, 2015). Income level impacts how much money customer spends on eating out and influences the choice of a restaurant. Hotels can also identify which customers are most valuable to them through demographic information.

Psychographic segmentation divides a market by personality traits, lifestyle, as well as values and beliefs.

Behavioural segmentation is the process of dividing a market into segments based on customers’ shopping and buying behaviour, the way customers respond to, use or know a product (Bhasin, 2017). There are several behavioural characteristics that can be taken into account (MBASkool):

1. **Occasion segmentation** focuses on dividing the market based on certain events when a customer uses or purchases a product or a service. In restaurants, typical meal occasions are breakfast, lunch and dinner. Snack occasions are midmorning, mid-afternoon, and late evening (Hsu & Powers, 2002).
2. **Usage based segmentation** is grouping based on how much a product is being used/consumed by the customer. For example, pizza delivery operations target the heavy users (Hsu & Powers, 2002).
3. **Loyalty status segmentation** usually divides the market into four segments based on brand loyalty status (Kotler & Keller, 2016): hard-core loyalists, who buy the brand all the time; split loyalists, who are loyal to two or three brands; shifting loyalists, who move from one brand to another; and switchers with no loyalty. Loyal customers are of great importance for chained-brand hotels and this is why they use data analytics to identify this segment and engage it with loyalty programs.
4. **Benefits segmentation** is grouping based on different benefits perceived by different customers.

Depending on the research objective, segmentation can be performed by using different data mining techniques and methods: clustering algorithms, factor analysis or principal component analysis, logistic
Cluster detection is the most important unsupervised learning technique, as there is no output variable. This is unlike the supervised learning algorithms, which prescribe a target variable. In other words, in case of unsupervised learning, an algorithm works only with a set of input variables, trying to find a hidden structure or relationships between different inputs. Cluster analysis divides data into groups – clusters. It finds sets of cases that are more similar to one another than they are to cases in other sets (Čamilović, 2008). This is displayed in Figure 3.

Figure 3: Using clustering for customer segmentation

Source: Author

Clustering algorithms can be directly applied to input data, but it is recommended to employ some data reduction technique first, for example, principal components analysis (PCA). The goal of PCA is to derive the smallest number of components that make up as much information as possible about the original input variables (Tsiptsis & Chorianopoulos, 2009). These components are uncorrelated, and for this reason they are suitable input to many data modelling techniques, including clustering.

There are a number of clustering techniques and the most commonly used is the k-mean algorithm. The “K” in its name refers to the fact that the algorithm looks for a fixed number of clusters (Berry & Linof, 2004). It randomly selects K instances (data points) as initial cluster centres. Instances are assigned to the nearest cluster seed (usually simply by
calculating Euclidean distance), which is then updated to the average position of the data points of each new cluster. This process is carried out iteratively until iteration of the algorithm shows no change in the cluster centres (Roiger & Geatz, 2003).

When using cluster analysis, a valuable issue to consider is the effect of possible outliers. An outlier is something that lies outside the group that it is a part of, a record with extreme value. Outliers can be detected by examining data through simple descriptive statistics or by specialized modelling techniques (Tsiptsis & Chorianopoulos, 2009). In many cases, the differences between the outlier and other records are so immense that they may mask the existing differences among these other “normal” records. As a consequence, algorithm can find the solution that barely separates outliers from the rest of the records. This is why it is a good approach to identify outliers and treat them in a special manner (Tsiptsis & Chorianopoulos, 2009).

After performing clustering, the model should be evaluated. It is a good idea to use supervised learning techniques in this process. Roiger and Geatz (2003) suggest a three step approach:

1. Perform clustering.
2. Choose a random sample of instances for each cluster (initial choice should be two-thirds of all instances in a group).
3. Build a supervised model using randomly sampled instances as a training set and the rest of instances for classification correctness. Decision trees or rule-based structures can be used in the process of supervised evaluation.

Once clusters are identified, their business interpretation comes into play. This process of describing the clusters is known as profiling. The variables used for clustering can be useful in describing each segment, but additional data, such as demographic, can also be beneficial (Venkatesan, 2007).

Usually the profiling process starts with an examination of the cluster centroids. Each cluster should be examined individually and its means should be compared with the input attributes to the overall population means. The analyst should look for large deviations from the “typical” behaviour (Tsiptsis & Chorianopoulos, 2009). After that, additional data (such as customer demographics) can be used in order to fully identify the features of each cluster. This involves examination of the mean of each
continuous attribute for each cluster. For categorical attributes, the procedure includes frequency and percentage comparisons in order to uncover the cluster differentiation (Tsiptsis & Chorianopoulos, 2009).

**Retention Models**

In the highly competitive hospitality industry, customer retention is very important. Losing a customer means more than losing a single sale – it means losing the entire stream of purchases that the customer would make over a lifetime of patronage (Kotler et al., 2014). Customer churn, also known as customer attrition, is the loss of clients or customers. The percentage of customers who are lost within a given time period is known as churn rate (or attrition rate). High churn rate is a sign that customers are not happy with what they are getting.

Churn models, also called retention or attrition models, predict the probability of customer attrition (Parr Rud, 2001). They seek to understand customer behaviours and attributes which indicate that he/she is a potential churner. Churn prediction is one of the most important data mining applications in customer relationship management (CRM). Most advanced casinos and hotels can predict churn (McGuire, 2017).

In order to answer the question who and why is likely to churn, the analyst needs to apply a classification data mining algorithm. Classification algorithms predict the class or category, based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of input attributes and the respective outcome, known as the goal or prediction attribute. The algorithm seeks to discover relationships between the attributes in order to predict the outcome (Voznika & Viana, 2001).

A number of supervised learning techniques can be used for churn prediction: logistic regression, neural networks, decision trees, random forests, support vector machines (SVM), survival analysis.

No matter which algorithm is used, it all begins with assembling and selecting the data. It is very important to take all the important variables into consideration. Business experts’ input and guidance is advised in the input variables selection process. Lazarov and Capota (2007) suggest using four groups of variables (as shown in Figure 4):
- **Customer behaviour**: variables that indicate which products/services a customer is using and how often he/she uses them.
- **Customer perceptions** that show how a customer perceives the product/service. They can be measured by surveys.
- **Customer demographics**, such as age, gender, level of education, geographical data, etc.
- **Macro environment variables** that describe changes in the world and distinct experiences of customers.

Figure 4: *Sets of data variables in churn prediction*

Assembling the data is followed by data pre-processing. Besides data transformations (such as normalization) and sometimes data smoothing, this step includes resolving the problem of missing values. There are several options for dealing with missing data (Roiger & Geatz, 2003): to discard records with missing values, to replace missing values with the class mean, and to replace missing attribute values with the values found within other highly similar instances.

It is important to choose the right set of attributes when building a model. The process of selecting a subset of relevant features (attributes) to be used in model construction is known as features or attributes selection. The optimality of a feature subset is estimated by an evaluation criterion (Tiwari & Singh, 2010). Different algorithms have different evaluation criteria and they fall into one of the three categories (Tiwari & Singh, 2010):

Source: Author
- Filter model: evaluates and selects the right subset of attributes only by their general characteristics. It does not engage any data mining algorithm.
- Wrapper model: requires predetermined mining algorithm and uses its performance as the evaluation criterion.
- Hybrid model: tries to take advantage of the two models by applying their different evaluation criteria in different search stages.

Attributes highly correlated with other input attributes are redundant (Roiger & Geatz, 2003). The correlation between two attributes can be measured by computing the linear correlation coefficient.

Whichever supervised learning algorithm is chosen for churn prediction, the analyst always needs to train and evaluate the algorithm. The data set has to be partitioned appropriately so as to avoid overfitting or underfitting. Overfitting occurs when a model fits the data too well. What happens is that the model captures the noise of the training data in such a manner that it negatively impacts the performance of the model on new data. Underfitting happens when the model does not fit the data well enough, which means that it cannot capture the underlying trend of the data. To prevent this issue, data set should be properly divided into the following (Condamoor, 2015):
- Training set (60% of the original data set), which is used to build up the model.
- Validation set (30% of the original data set), which is used for evaluating effectiveness of the model. The model generating the least amount of error normally gets to the test.
- Test set (10% of the original data set), which is used to test the model that was selected from the validation.

Sannel (2015) emphasizes that churn prediction is an iterative and incremental process. When the model identifies potential churners well and retaining efforts are successful, the customer will not leave the company. This will invalidate the current churn prediction model and may require a new iteration in order to identify the high impact factors that cause customer churn. This process can be repeated numerous times (Sannel, 2015).

Understanding how to reduce churn is critical for the success of a data mining project. The data analyst must be aware of marketing actions that
are going to follow. Cisternas (2010) points out that the best model is not necessarily the one with the best statistical precision, but the one that provides the best insights to prevent churn.

Sentiment Analysis

Sentiment analysis or opinion mining uses different methods to extract and identify the sentiments, attitudes and opinions in a text. It is very useful in analysing opinion-rich resources such as product/service reviews, surveys, customers’ emails, call logs, tweets, Facebook comments or posts, and comments on a blog or a Web site.

Sentiment analysis can be considered as a text classification problem, since an opinion may be positive, negative or neutral. There are three main classification levels (Medhat et al., 2014): document level, sentence level, and entity-aspect level. Overall opinion is examined on the document level, while opinion of a particular sentence is examined on the sentence level. The level which focuses on the opinion itself is known as the entity and aspect level (Vaghela & Jadav, 2016). Most common techniques used in sentiment analysis include Naive Bayes, support vector machines, and entropy classification.

There are many ways in which a business can use sentiment analysis (Sheey et al., 2014): to track what customers say in order to improve marketing efforts, to boost word of mouth marketing, to discover product/service improvement opportunities by analysing customers’ feedback, etc. It is very useful to monitor what customers say, because in this way a company can get real understanding of customers’ attitudes toward its products/services, brand or commercials. These insights are the strengths that can be used for better marketing.

Word of mouth marketing is still a very effective practice for marketers. A company needs to identify its highly influential customers, because it can involve them in advertising at a much lower cost than traditional methods. Analysis of large amounts of data from social networks and across the Web, or from emails and call logs can be very helpful in improving products and services in a timely manner. This is very often an easier and quicker way of identifying customers’ opinions than conducting surveys or focus groups (Sheey et al., 2014).
Next Best Offer

This is a revolutionary data mining application in the travel, tourism and hospitality industry. By analysing the data, from detailed demographics and psychographics to Web clickstream data, marketers create customized next best offers (NBOs) which direct their consumers to the right products/services, at the right moment, at the right price and in the right channel (Davenport et al., 2011).

In the travel industry, NBOs can be used as a secure path to further monetize existing customers’ itineraries, as well as significantly improve customer experience and satisfaction (Kremer, 2014). By predicting which product a particular customer is going to buy next, a company can boost its revenues with personalized cross-selling and upselling (Sheey et al., 2014). But, next best offers should not be confused with in-path upselling implemented by many travel brands. The main difference is that NBO is typically sent not before, but after a purchase and it is designed to be relevant and complementary to that purchase (Kremer, 2014). Figure 5 illustrates this concept.

Figure 5: Next best offers

Source: Author

For the purpose of data mining, a company needs to assemble three types of data (Davenport et al., 2011):

1. Customer data: customer demographic, psychographic, and behaviour data, as well as data about their previous purchases.

270
2. Data about the company’s product offers: information about the company’s products or services. It is advisable to classify products into groups.

3. Data about the purchase context: the channel through which a customer is making contact with a business, reason for purchase, customer emotion, weather, the time of day or the day of the week, etc.

Two most commonly used data mining techniques in predicting what a customer wants to buy next are link analysis and decision trees. Link analysis is used to link the customers’ market basket and historical transactions to products they may want to buy. Then, decision trees can be used for scoring suitable timing of marketing messages (Sheey et al., 2014).

Decision trees can also be used, together with regression analysis, for product propensity scoring (Sheey et al., 2014). Furthermore, link analysis helps in determining what products naturally go well together, which is the key for developing an effective cross- and upselling campaign (Sheey et al., 2014).

Conclusion

Marketing has always been driven by data. In order to segment their customers, predict which customers are at risk of leaving, identify customers’ opinions, and create customized next best offers, companies need to analyse large amounts of data. Besides collecting and integrating the data, creating a model presents an even greater challenge. Choosing the right variables, the right data mining algorithm and the right data analysis tool is not an easy task. This is why data science is the latest in-demand skill within the marketing profession.

Unfortunately, many hotels and restaurants currently do not perform data analysis and thus their data remains an underused asset. This is mainly because they do not understand the potential of data mining applications. The aim of this paper is to emphasize why data analysis is so important and what it can be used for. If hotels, tourist organizations and restaurants start analysing the captured data properly, they will be able to understand their customers’ needs and expectations, create better offers that meet customer needs, keep customers loyal and happy and maybe even manage to turn them into their biggest advocates.
References


