ISSN (PRINT): 1994 - 7658 ISSN (ONLINE): 1314 - 0817

VARNA UNIVERSITY OF MANAGEMENT

European Journal of Tourism Research

http://ejtr.vumk.eu

Airbnb listings' performance: determinants and predictive models

Efstathios Kirkos [1]*

[1] Department of Accounting and Information Systems, International Hellenic University, Thessaloniki, Greece. E-mail: stkirk@ihu.gr

*Corresponding author

Abstract

The present study analyzes Airbnb listings' performance in terms of occupancy rate, number of bookings and revenue, by employing data mining methodologies. The research objective is twofold, to highlight the strongest determinants that influence customers' purchase intentions and to propose reliable models capable of predicting listings' performance. The data set refers to the Airbnb market of Thessaloniki, Greece and contains explanatory variables related to the hosts, lodgings, rules, and guests' ratings. Elaborated classification methods derived from Machine Learning are used as analytical tools. The findings highlight the central role of the host. Superhost badge, rich host presentation and quick response to customers' requests are factors that boost performance. Other determinants are the provision of amenities and high overall rating. In terms of predictive models, Random Forest outperforms its competitors and is proposed as the most suitable classifier for the specific domain. The paper contributes to the existing literature in several ways. It adopts a data-driven research approach, employs machine learning techniques, proposes reliable models capable of predicting listings' performance and highlights the most influential determinants. The results and conclusions can be useful to individual hosts, professional listings' managers, as well as legislative and taxation authorities.

Key words: sharing economy, Airbnb, short-term rentals, peer-to-peer accommodation, purchase intention

Citation: Kirkos, E. (2022). Airbnb listings' performance: determinants and predictive models. *European Journal of Tourism Research* 30, 3012.

Introduction

Sharing economy is a modern term which describes the collaborative usage of under-utilized inventory through fee-based sharing (Zervas *et al.*, 2016). This new business model allows individuals to become microentrepreneurs and earn money from their idle property and spare time (Lutz & Newlands, 2018). Sharing economy is based on digital peer-to-peer marketplaces, where individuals directly sell and buy goods and services, while the marketplace platform is provided and maintained by a third party (Botsman & Rogers, 2011).

A prominent example of peer-to-peer marketplace in the tourism sector is the Airbnb platform. Initially launched in 2008 as a hotel alternative, it matches lodgings' owners with short-term renters. Due to its special characteristics, such as an internet-based business model, massive numbers of hosts and guests, low cost, provision of household amenities and the potential for a more authentic local experience (Guttentag, 2015), Airbnb enjoys an exponential expansion. Other characteristics that have contributed to its rapid increase in popularity are a sense of community, human contact, and shared meanings (Yannopoulou *et al*. 2013).

The rapid rise of Airbnb has triggered a stream of manifold research addressing several issues like pricing, and user demographics, as well as consequences on the performance of the hotel industry, on unemployment, on properties' value, and even on urban planning. As expected, a main branch of Airbnb-related research deals with customers' behavior and examines the determinants that affect accommodation purchase intentions. The topic is of major importance for Airbnb lodgings' providers. Gibbs *et al*. (2018) point out that due to the exponential growth of the platform, the hosts are facing greater competition and thus their marketing practices require more informed approaches.

A critical appraisal of the existing research literature which deals with determinants affecting purchase intentions, and consequently listings' performance, reveals that the typical approach is to concentrate on specific factors including online reviews, user ratings, information quality, perceived value and price, along with host profile and quality. The researchers formulate a priori hypotheses about the significance of specific determinants and then they validate them by using questionnaires or by performing statistical analysis. Zhao and Peng, (2019) propose eight hypotheses to evaluate the impact of on line reviews on users' purchase decisions and perform a questionnaire survey. Chen and Chang (2018) examine the impact of rating, rating volume, review, information quality and media richness on guests' purchase intention and use a questionnaire, ANOVA and path analysis. Nisar *et al*. (2019) analyze perceived value, perceived price, lodging information, lodging reviews, trust towards the host, website usability, and perceived privacy and security in terms of purchase intentions. Their study is based on survey data. Han *et al*. (2019) hypothesize that a guest's purchase would be affected by the host-created information's ethos, pathos and logos and employ Tobit Regression to analyze the data.

Targeted, hypotheses-driven approaches which concentrate on specific issues can highlight influential factors but avoid analyzing the total of the available information in a systematic and coherent manner, a possibility provided by Data Mining (DM) methodologies. The recent data-centric view revises the role of the hypothesis in science. Research is driven not by theoretical expectations. On the contrary, empirical input determines the direction of research. This novel research approach has triggered an interesting disputation among scientists and philosophers of science. Supporters of the theory-centric view of science argue in favor of the central role of theory and hypothesis-driven research. Mazanec (2020) examines big data analytics within tourism design research, and simultaneously addresses several methodological and epistemological issues. He claims that there are elements of theory hidden in data-driven analytical approaches and stresses that 'humans cannot *not* hypothesize'.

Moreover, Kozinets and Gretzel (2021) concentrate on more practical matters and criticize the incomprehensibility of the patterns produced, the disconnection between marketers and customers and the vulnerability of the marketers to changes in algorithms. Paraphrasing the evangelist Matthew, they state that 'He who lives by the algorithm, dies by the algorithm'. However, they also conclude that marketers can diversify their perspectives on understanding, their links to customers and their use of platforms and enjoy the gifts of Artificial Intelligence (AI) empowerment. An interesting epistemological article about big data analytics and related methodological issues in research can be found in the Stanford Encyclopedia of Philosophy (Leonelli, 2020). The author deals with the use of big data within scientific research and discuses a wide range of themes, which include the extrapolation of patterns from data, the role of human intelligence in machine learning, the nature of data as research components, the relation between data and evidence, the theory centric view of knowledge, the relation between prediction and casualty, and finally ethical issues related to data science.

Despite the criticism of AI and big data, it is an undeniable fact that the interest in these topics is steadily growing, as demonstrated by market trends ("Big Data Technology Market Size, Share, Demand & Growth [2027]", 2021). AI methods have advanced capabilities to identify patterns and trends hidden in volumes of data. AI classification methods make no assumption about the independence of the explanatory variables or the distribution of the data and their predictive power is validated against previously unseen observations. Moreover, data mining provides a wide range of methodologies which cover the whole spectrum of the knowledge discovery process. Such methods deal with data preprocessing problems (like feature selection and class balancing), the development and validation of reliable predictive models and the evaluation of the predictors' significance. Remarkably, no AI – data mining methods have yet been used to analyze listings' performance.

The present study contributes to the existing literature in several ways. It analyses the topic of Airbnb listings' performance within a data-centric, data mining framework. The purpose of the research is twofold. First, to reveal determinants that have a strong impact on customers' purchase decisions and second, to propose reliable models capable of predicting listings' performance. To address these objectives, several DM methods are employed. These include data preprocessing (feature selection and class distribution balancing), model development with Machine Learning (ML) algorithms, models' validation against previously unseen data and finally, predictors' significance evaluation through the use of interpretable models, sensitivity analysis and a wrapper evaluator.

Previously conducted research has examined several factors that highly influence customers' decisions. Such factors include perceived value, price, trust, online reviews, lodging information, security, geographical location, amenities, host profile, narratives containing social words, and place pictures (Han *et al.*, 2019; Nisar *et al.*, 2019; Xie & Mao, 2017). However, none of these studies have used the DM methods applied in the present study or claim that they propose reliable models for the prediction of listings' performance.

Another contribution of the present paper is the proposal of approximative measures for lodgings' performance. Unfortunately, it is an established policy of Airbnb not to provide exact information about the lodgings' occupancy rate. We overcome this lack of information by using the number of reviews as a proxy for room reservations. This criterion has previously been used by researchers (Han *et al.*, 2019; Lee *et al.*, 2015; Liang *et al.*, 2020) and legislative authorities (InsideAirbnb.com, n.d.). Based on the number of reviews we introduce performance measures, which refer to lodging occupancy, bookings, and revenue. Mean values are used as a threshold to differentiate high from low performing listings. The classifiers used result in successful models and provide evidence on the predictors' significance.

The data used come from InsideAirbnb, a third-party site which provides data sets about Airbnb. The data set analyzed refers to lodgings located in the city of Thessaloniki, Greece. Thessaloniki is a city with strong touristic appeal. Being the second largest city of Greece, it is an urban, commercial and economic center. Established in 316 BC and named after the sister of Alexander the Great, Thessaloniki has a long ancient, roman, and byzantine history, as the numerous monuments reveal. The city is located close to the northern borders of the country and allows easy access to visitors from neighboring countries. Thessaloniki is also in close proximity to the famous summer resort of Chalkidiki. For the aforementioned reasons, Thessaloniki has become an important tourist destination. The increased tourist traffic has triggered an exponential growth of the local Airbnb market, from 10 listings in 2010 to 2283 listings at present. The Airbnb market in Thessaloniki is an example of a rapidly growing accommodation sharing peer-to-peer network, located in a highly touristic urban area.

The present study highlights determinants and proposes reliable models for estimating Airbnb listings' performance, employing data mining methodologies to analyze publicly available data. Individual hosts, professional listings' managers, legislative and taxation authorities may benefit from the outcomes of this research.

Literature Review

Airbnb in the research literature

The present subsection presents several Airbnb-related research topics. The Airbnb platform is a prominent example of an informal peer-to-peer accommodation sector. Its rapid rise in popularity attracts the interest of many researchers. The existing research literature addresses a wide variety of issues. Guttentag (2015) examines the Airbnb phenomenon through the lens of disruptive innovation theory. Davis (2016) examines the growth of Airbnb and Uber in terms of domestic and international market expansion and of internal platform functions development. Other studies related to the Airbnb platform focus on taxation and the regulatory framework (Allen & Berg, 2014; Koopman *et al.*, 2015; Palombo 2015).

Numerous studies examine the impact of the Airbnb phenomenon on the properties' value and availability (Lee, 2016; Oskam & Boswijk, 2016), on urban planning (Gurran & Phibbs, 2017), on tourism industry employment (Fang *et al.*, 2016), and on the performance of the hotel industry (Dogru *et al.*, 2017; Zervas *et al.*, 2017). Several studies focus on user demographics (Adbar & Yen, 2017; Smith, 2016). Pricing in the Airbnb market is yet another well examined field. Researchers relate prices with lodging and host characteristics (Chen & Xie, 2017; Gibbs *et al.*, 2018), with the presence of hosts' photos (Ert *et al.*, 2016), and with the superhost badge (Liang *et al.*, 2017). Wang and Nicolau (2017) conclude that host attributes, location, accommodation capacity, amenities and services are strong drivers for higher prices. Dudás *et al.* (2020) apply hedonic price modeling and reveal that property-related attributes like the provision of air-condition, free internet and free parking significantly influence Airbnb prices.

Trust is a major issue in all peer-to-peer economic transactions. Yang *et al.* (2018) examine trust factors and conclude that security, privacy and Airbnb traits contribute to users' trust towards Airbnb. Other studies relate trust to information quality and perceived social capital (Chen *et al.*, 2015) as well as to host attributes such as reservation confirmation speed, acceptance rate, and existence of profile page (Wu *et al.*, 2017). Other researchers associate trust with the presence of hosts' photos, their facial expressions (Ert *et al.* 2016; Fagerstrøm *et al.*, 2017) and with users' reviews (Sparks & Browning, 2011).

Airbnb customers' characteristics and satisfaction criteria are another fruitful research area. Guttentag (2016) recognizes five evenly sized customer segments in the Airbnb market, but demonstrates that,

with the exception of home-seekers, low cost is the primary motivation of Airbnb customers. Del Chiappa *et al.* (2020) analyze and segment Italian Airbnb customers according to their motivations. The authors identify three main clusters, i.e. enthusiastic Airbnb lovers, pragmatic Airbnb users and pragmatic authenticity seekers. The discrimination of the clusters is based on marital status, level of education and employment. Lutz and Newlands (2018) compare the characteristics of customers who prefer shared rooms to those who choose entire homes, and argue that there are differences related to gender and socio-economic status. Festila and Müller (2017) differentiate Airbnb customers as 'extrovert' and 'introvert', and as individuals who 'go to see' as opposed to those who 'go to feel'. Customers of each of these categories have different satisfaction criteria. Tussyadiah (2016) examines the factors that influence guests' satisfaction and concludes that highly influential factors are enjoyment, monetary benefits (value) and accommodation amenities. Moreover, Yannopoulou *et al.* (2013) emphasize the social dimension of the Airbnb market and highlight the access to the private sphere, the human dimension, and the meaningful inter-personal discourses as the main emerged themes.

Airbnb purchase intension and listings' performance literature

Recent research effort has been directed towards the detection of factors that determine Airbnb listings' performance. Zhao and Peng (2019) investigate the impact of online reviews on Airbnb users' decisions. The authors use the Stimuli-Organism-Response (SOR) model in a questionnaires-based study. According to the results, the quality of the reviews is a significant factor that positively affects the perceived value, and negatively affects the perceived risk of a listing. Chen and Chang (2018) also employ questionnaires to investigate factors affecting the purchase intentions of Airbnb users and how they are influenced by rating, rating volume, review, information quality and media richness. The authors use ANOVA and conclude that perceived value and satisfaction are determinants of their intention to buy. Rating volume, review, information quality and media richness are important precursors.

In another recent study, Nisar *et al.* (2019) attempt to understand the main determinants that affect accommodation purchase intentions. The factors tested are the perceived lodging value, perceived lodging price, lodging information, online reviews, trust towards the host, and perceived price/security. Data were gathered with online questionnaires. The test outcomes reveal that these factors positively influence purchase intention. The greatest driver of purchase intention is perceived lodging value, followed by perceived lodging price. The factor with the smallest effect is online reviews. Peng *et al.* (2019) draw online reviews of Chinese tenants for Airbnb lodgings located in New York, Paris, Beijing, Shanghai, and Guangzhou and conduct content analysis to understand customer needs. The factors influencing users' choices are mainly measured from three dimensions: internal environment, external environment and cultural needs. Concerning cultural needs, the host is the most important party, since he/she represents the local culture.

Lee *et al.* (2015) analyze a total of 4,178 Airbnb room data to detect features that are strongly associated with room sales. Multiple Linear Regression is used as the analytical tool. Social features and room features are both significant determinants of room sales. Xie *et al.* (2019) examine customers' loyalty in combination with host attributes and travelers' frequency of past stays. The researchers estimate the baseline models by using OLS regressions and they implement probability regressions with random effect estimations. The results suggest that customers exhibit loyalty, and that host relevant attributes like acceptance rate and listing capacity, positively affect the likelihood of repeating a purchase.

A striking study is that of Han *et al.* (2019). By making use of Aristotle's assertions, the authors attempt to detect factors influencing guests' purchase decision. According to Aristotle's appeals, interpersonal

messages can be persuasive through three components: ethos, pathos, and logos. The authors express these appeals by using several independent variables. The data are drawn from Airbnb and Tobit Regression models are developed. The number of reviews is used as proxy for actual purchases. The results suggest that the superhost badge, host review, the use of social words, price, place picture and star-rating positively impact purchases.

Xie and Mao (2017) also search for factors that influence listings' performance. The data refer to Airbnb lodgings located in Austin, Texas and the analytical method used is Linear Regression. The authors conclude that host quality attributes, such as the superhost badge, long operating experience, and high response rate, significantly affect future reservations. On the contrary, identity verification does not have a positive effect. Information quality, expressed by host quality attributes, allows hosts to build trust among Airbnb users.

Liang *et al*. (2020) examine the importance of Marketer-Generated content on Airbnb listings performance, expressed as the number of online comments. The authors hypothesize that descriptions covering more topics or aspects (width) as well as more detailed descriptions (depth) increase the likelihood of bookings. The data is from lodgings located in Hong Kong and the employed analytical method is Multilevel Linear Regression. The results show that both the width and depth of host-provided descriptions for properties are positively associated with their review volume. Textual self-descriptions of the host also increase the likelihood of obtaining a greater number of reviews.

Critical appraisal of the listings' performance literature

A critical appraisal of the presented literature reveals that considerable research effort has been directed towards the examination of factors affecting Airbnb listings' performance. Many researchers use questionnaires, while others use listings' data to examine these factors. The employed analytical tools are solely derived from the field of Statistics and the researchers adopt a standard statistical approach. Typically, researchers formulate prior hypotheses on specific issues and then they validate them against data sets. Remarkably, researchers miss to validate the derived models by testing their performance against unknown (out of the training data set) observations. Such an approach allows researchers to explain the distribution of the training data, but provides no evidence that the models are able to predict the class value of new cases. Learning algorithms tend to overfit, and in this case their predictive power against unknown observations decreases substantially. Thus, no research study thus far claims to propose a reliable model, able to predict listings' performance for new, unknown cases.

So far, no attempt has been made to examine listings' performance within the data mining framework. Data mining allows the construction of data-driven predictive models, without the formulation of prior hypotheses. Already in 2008, the chief editor of 'Wired' Chris Anderson, in his article carrying the subversive title 'The end of theory', discussed the advent of the 'Petabyte Age', where the use of advanced software allows the automatic extraction of patterns without any a priori assumption. As an example he cites Google and mentions the following: 'Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day. And Google was right' (Anderson, 2008).

Based on the above, the aim of the present research study is to examine Airbnb listings' performance within the data mining framework, with the aim of highlighting significant determinants and of proposing reliable, data-driven models, able to predict lodgings' popularity and monetary success. The customers' final decision is the outcome of a combinatorial assessment of several issues. Accordingly,

the simultaneous, fair, and coherent analysis of these issues can reveal significant factors which strongly influence bookings. Moreover, elaborated classification methods, derived from Machine Learning, can offer reliable models which outperform the widely used Logistic Regression.

Research Methodology

Data

The data used in the present study come from InsideAirbnb.com (n.d.). According to the site operators, 'Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world'. The data sets are actually a time snapshot and contain publicly available information about the lodgings, the host, customer reviews etc. We chose the city of Thessaloniki and downloaded the file listings.csv.gz, as it contains the most detailed information. The file was compiled on March 11th 2019 and the data refer to that time period. The selected data set contains 2282 lines, corresponding to the total listings in Thessaloniki, and 106 columns pertaining to the lodging, the neighborhood, the host, availability, customer reviews etc.

A very interesting column in the data set is the column 'availability_365'. A major issue concerning Airbnb hosts is whether they rent a property on a permanent basis, like hotels, or whether they occasionally rent their own homes. As a result, the time period when the lodging is available may range from a few days to the whole year. The column 'availability_365' depicts the number of days a listing is available, without declaring if the rest of the days the lodging is occupied or not made available by the host. It is an established policy of Airbnb not to provide information about the number of days a lodging is actually occupied. This adds complexity to the analysis since an absolute measure for performance is not available. According to the entries in the 'availablility_365' column, half of the properties are available for a time period of up to six months and the other half are available for six to twelve months. Many columns of the data set contain information about the host. One of the most significant host details is whether the host has been granted the 'superhost' badge. The status of superhost is granted to hosts who satisfy certain requirements including high occupancy, high response rate, low cancelation proportion, and high customer evaluation. In terms of customer evaluations and reviews, we observe a very high proportion of listings (80%) with top scores of four stars or more. This is not surprising, since several studies stress the fact that many Airbnb listings achieve considerably high scores. However, the mass assignment of extremely high scores makes this criterion ineffective for lodging selection. The number of reviews is another, more influential measure, as it reflects the number of customers who have previously selected the lodging and may affect the intentions of potential future customers. In the case of Thessaloniki we observe that 45% of the listings receive one or fewer reviews per month.

Columns with a high rate of missing values were removed. For the remaining columns we overcome the problem of lost values by retrieving the correct values from the original entries. Fifteen lines, for which original information could not be recovered, were removed from the data set. Moreover, columns containing no task-relevant information or containing information already included in other columns were also removed. Such columns were 'listing id', 'listing_url', space and description, 'host_id' etc. As a data preprocessing step, we transformed Boolean (true/false) values to binary (1/0) and ordinal values to numeric.

The column 'host_about' contains a textual self-presentation of the host. This column was substituted by a column containing the number of characters in each description, as a measure of the quantity of information a host provides about him/herself. For the attributes 'property_type', 'room_type' and 'amenities' we created dummy variables, one for each value found in these columns. Additional calculated columns accumulate values of other columns. The calculated attribute 'host_total' sums up

the values of the host-related attributes and the attribute 'total_amenities' contains the total number of the provided amenities. The column 'total_price' contains the sum of the price per night and the cleaning fee.

Finally, an important characteristic of a lodging is its location. Research findings suggest that location may influence customers' behavior (Lee *et al.*, 2015). In Thessaloniki, most points of interest are located in or around the city center. Aristotelous square is the most central point and is considered the heart of the city. In an attempt to introduce a location-relevant measure, we calculated the distance between the lodgings and Aristotelous square based on longitude and latitude and using the haversine formula. Since it is meaningless to include non-informative columns such as irrelevant attributes or attributes with high rates of missing values in the analysis, we eliminated such attributes from the data set. Moreover, although some methods, like Decision Trees, appear to be immune to the existence of irrelevant attributes, other methods, like Logistic Regression, assume that only significant and independent (i.e., not correlated) variables are included, and thus the latter methods may be derailed by irrelevant or overlapping attributes. In total 34 independent variables were included in the data set. There are host-relevant variables, lodging-relevant variables, variables related to terms and conditions, and to customers' scores. The variables 'host_total' and 'total_amenities' aggregate the values of the host-relevant variables and the amenities-relevant variables, respectively.

As above stated, the aim of the present study is the construction of models capable of predicting the popularity and the financial returns of Airbnb listings, and at the same time the identification of strong influential factors, related to their success. Unfortunately, Airbnb does not provide information about the occupancy, or even estimated bookings of the lodgings. As a result, there are no absolute measures of popularity and return available. Additionally, the availability calendar, operated and updated by the host, is also not useful, since it does not differentiate the days the lodging is booked by guests from the days the lodging is not made available by the host. Moreover, some hosts may not update their calendar properly. In order to deal with this problem, researchers and authorities use approximations. On the Airbnb platform, only actual customers can publish reviews. Thus, the number of reviews may be used for the approximate estimation of the number of customers (Chen & Xie, 2017).

A model used to estimate the number of bookings based on the reviews is the 'San Francisco model'. This model has been proposed and is used by the Inside Airbnb website. The model assumes that half of the real customers publish a review. Other similar approaches assume different percentages. Alex Marqusee, a legislative and policy director in Oakland, uses a review rate of 72%. The Budget and Legislative Analyst's Office also uses a review rate of 72%, but additionally introduces a higher impact model, using a review rate of 30.5%. Inside Airbnb estimates that a review rate of 30.5% is not conservative enough and proposes a percentage of 50% as it sits almost exactly between 72% and 30.5% (Inside Airbnb, n.d.). In the present study we adopt this suggestion and use a review rate of 50%. However, in our view, the total number of reviews is not the most appropriate measure. A lodging that is available for a long time period will normally receive more reviews (Zhang *et al.*, 2019). The measure 'reviews_per_month' is more suitable, since it normalizes the number of reviews to the number of months the lodging is made available by the host.

Based on the above assumptions, we introduce three dependent variables that reflect the performance of the lodgings. The variable 'occupancy' expresses the number of days the lodging is occupied in terms of its availability. The variable 'bookings_per_year' contains the number of bookings the lodging achieves within a one-year time period. The variable 'estimated_monthly_revenue' stands for the achieved monthly income. The exact formulas for these measures are presented in equations 1, 2, and 3.

$$Occupancy = (reviews\_per\_month * minimum\_nights * 12)/\ availability\_365 \quad (1)$$

$$Bookings\_per\_year = reviews\_per\_month\ * 2 * 12 \quad (2)$$

$$Estimated\_monthly\_revenue = reviews\_per\_month * 2 * total\_price * minimum\_nights \quad (3)$$

Mean values are used to differentiate high from low performing listings. Thus, three dichotomous nominal variables are formulated for occupancy (high occupancy/low occupancy), bookings (high bookings / low bookings) and revenue (high revenue / low revenue). This comparative assessment of performance (i.e., high performing vs. low performing listings) neutralizes deviations which possibly derive from the assumption that half of the guests publish a review. We also have to underline that the variable 'reviews_per_month', although logically significant, has not been included in the final input vector since it was used for the calculation of the dependent variables, and would thus, if used as an independent variable, unfairly boost the models' performance and bias the results of the analysis.

Classifiers and other algorithms

We employ five well known classifiers for the development of the models. These are C4.5 Decision Tree (DT), Logistic Regression (LR), Multilayer Perceptron Neural Network (MLP), Support Vector Machines (SVM) and Random Forest (RF).

C4.5 is a classification method that belongs to the Decision Trees family. A Decision Tree (DT) is a tree structure, where each node represents a test on an attribute and each branch represents an outcome of the test. The goodness of a split is based on the selection of the attribute that best separates the sample. For this selection, C4.5 uses an entropy-based measure, called Gain Ratio. DTs offer considerable advantages. They make no assumptions about the independence of the input variables or the distribution of the data. They produce comprehensible models which can be easily converted to a set of If-Then rules, and they have a fast learning algorithm. A major disadvantage of decision trees is that they are sensitive to changes of the sample.

Logistic Regression (LR) is probably the most widespread classification algorithm. LR has initially been proposed during the 19th century, and since then has been used in a vast amount of studies. Normally suitable for dichotomous problems (two class values), LR models the logarithm of the odds for a class value labeled '1' as a linear combination of the independent variables. LR is considered a classification benchmark, however numerous recent studies claim that new methods, mostly originated from machine learning, usually outperform LR.

A Multilayer Perceptron (MLP) is a feedforward neural network, suitable for modeling relationships between a set of predictors and a response variable. An MLP consists of a number of processing units called neurons. Each neuron is connected to other neurons and each link carries a numerical value called a 'weight'. When a signal is sent from one neuron to another it is multiplied by the weight of the connection. The total input signal for each neuron is calculated and, if it exceeds a threshold value, it is transformed by the transformation function and passed to other neurons. The training of the network is in fact the tuning of the weights, performed using the Backpropagation algorithm. MLPs do not assume a linear or any other relation between the independent and the dependent variables, are capable of handling noisy or inconsistent data and are a suitable alternative for problems where an algorithmic solution is not applicable. They also usually generalize well and achieve high accuracy rates against unknown observations.

Support Vector Machines (SVMs) is a classification method developed by Vapnik (1995) and is mainly suitable for dichotomous problems. The key idea is to transform the input space to a higher dimensionality feature space and create a maximum-margin hyperplane that splits the example classes. In order to find the optimum hyperplane, the notion of margin is introduced. The task of the algorithm is to find the hyperplane with the maximum margin. In real world problems a separating hyperplane may not exist and some examples may be misclassified. To deal with this problem, Vapnik introduces slack variables. The trained classifier should maximize the margin and simultaneously minimize the sum of the slacks. SVM enjoys excellent reputation due to its solid mathematical foundation, its ability to deal with data sets with many columns, and its classification performance.

Random Forest (RF) is a recent Decision Tree based classification algorithm and has been proposed as a response to the tendency of Decision Trees to overfit the training data sets. RF belongs to the family of ensemble classifiers. This means that a committee of individual classifiers is created, and the final classification decision is obtained by aggregating the individual decisions. To build the multiple trees, multiple training data sets are created. These data sets are built by selecting examples through sampling with replacement and simultaneously randomly selecting a subset of the initial attributes. The simultaneous selection of lines and columns ensures the substantial differentiation of the data sets, and thus the differentiation of the produced models, a prerequisite for ensemble classifiers. RF achieves high accuracy rates, does not overfit, and runs efficiently on large data sets.

Feature selection is a standard data preprocessing step in classification problems. The initial columns of the data set (also features or attributes) can be irrelevant, redundant, or useful. The aim of feature selection is to reduce the input space and to select an optimum subset of relevant features, suitable for the classification task. In the present study we use the Correlation Based Feature Selection (CFS) method. CFS is a filter method and uses a correlation based heuristic to evaluate the worth of a subset of features. The heuristic takes into account the usefulness of individual features in predicting the class label along with the level of intercorrelation among them. A feature is selected if it predicts classes in the instance space that are not already predicted by other features. Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other. A detailed description of CFS can be found in Hall (1999).

A common challenge in classification data sets is the class imbalance problem. Class imbalance means that there are many observations belonging to one class (majority class) and significantly fewer observations of the other class (minority class). Class imbalance is a serious problem, since learning algorithms tend to overfit to the majority class. As a result, the produced models achieve high accuracy rates against majority examples but fail to predict the class label of the minority observations. Several methods have been proposed to deal with the class imbalance problem. Some of them adopt a majority class undersampling approach (at the cost of information loss), while others employ minority class oversampling techniques. There is significant class imbalance in the data set employed in the present study. More specifically, the ratio between high occupancy and low occupancy lodgings is 698/1569, the ratio between lodgings with high bookings and low bookings is 856/1411 and the ratio between high revenue and low revenue lodgings is 840/1427. In an attempt to overcome this problem, we employ the Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al.*, 2002). SMOTE is a widely recognized technique and balances the data set by producing synthetic examples derived from existing minority observations.

10-Fold Cross Validation (10FCV) is a well-recognized method for estimating a model's ability to predict the class label for out_of_training_set observations. According to 10-Fold Cross Validation the data set

is divided into ten subsets. In an iterative procedure, nine of the folds are used for training and the remaining fold is used for testing. The process iterates ten times, each time using a different fold for testing. 10FCV provides reliable evidence about a classifier's generalization ability.

Results and discussion

An initial analysis of the data reveals interesting aspects of the Airbnb market in Thessaloniki. The vast majority (about 92%) of available lodgings are entire homes - apartments, as opposed to private or shared rooms. This proportion is extremely high compared to the findings of other studies. For example, Wang and Nicolau (2017) analyze the Airbnb data of 33 different cities and find that the percentage of entire homes - apartments is 65%. Gibbs *et al*. (2018) report a proportion of 58%, in a study which covers five Canadian cities. Therefore, it seems that Greek owners offer their property, but are rather unwilling to share their own accommodation. Remarkably, 39% of the hosts in Thessaloniki have been characterized as superhosts. The host-relevant data also reveal that 70% of the hosts respond to customer requests within one hour. Another interesting issue is that a rather low proportion (19%) of hosts have verified their identity. Research findings indicate that identity verification upgrades the quality perceived by potential customers.

Lodging prices range from 10€ to 700€, with a mean value of 60€ (73 USD). This value is significantly lower in comparison to mean values of other cities, as can be seen in Table 1. Most of the hosts (47%) list only one lodging on the platform, 18% list two lodgings and 35% list three or more lodgings. Remarkably, the percentage of hosts with two or more listings (which in the literature are considered professionals) is high. Comparatively, the percentage of professional hosts in Texas is 12% (Xie & Mao, 2017) and in five Canadian cities it is 37% (Gibbs *et al*., 2018). We thus observe that the majority of Airbnb listings in Thessaloniki is managed by professional hosts.

Table 1. *Average prices of Airbnb lodgings*

City	Price (USD)	Reference
33 cities	117	Wang & Nicolau (2017)
5 Canadian Cities	103	Gibbs *et al*. (2018)
Hong Kong	101	Liang *et al*. (2020)
New York	210	Fagerstrøm *et al*., (2017)

One of the primary goals of the present study is to develop reliable models, capable of predicting listings' performance in terms of occupancy, bookings, and revenue. All the reported results are the outcome of the 10-Fold Cross Validation test, which provides a strong indication on how the model will perform, when used in the 'real world' against unknown observations. The software used is WEKA, a well-known data mining package, developed at the University of Waikato.

Occupancy

Regarding occupancy, the variable 'availablity_365' is not used as a predictor, since it is used for the calculation of the dependent variable. After performing feature selection, seventeen variables were selected to participate in the final input vector. The results of the models are summarized in Table 2. All the predictive models are successful. The achieved accuracy rates are high and range between 73% and 82%. Random Forest outperforms the other classifiers in all used metrics.

Table 2. *Models' performance for occupancy*

Classifier	Accuracy	TP Rate High	TP Rate Low	F-measure High	F -measure Low	Roc Area
C4.5	75.85	0.710	0.802	0.735	0.778	0.791
LR	68.53	0.670	0.699	0.667	0.702	0.755
MLP	72.78	0.765	0.695	0.726	0.730	0,796
SMO	65.19	0.601	0.697	0.619	0.680	0.649
RF	81.59	0.744	0.880	0.792	0.835	0.890

Two of the employed classification methods offer interpretable models and provide evidence regarding the significance of the independent variables. As first level splitter, C4.5 selects the variable 'host_is_superhost'. This means that, according to the Gain Ratio measure, this variable best separates high from low occupancy listings. LR offers additional insights into the predictors' significance. In Table 3 the coefficients of the LR model for high occupancy are reported. As can be seen, the variables 'host_is_superhost', 'security_deposit', and 'instant_bookable' increase the possibility of a lodging being classified as 'high occupancy', while the variables 'host_total_listings_count' and 'host_response_time' decrease this possibility. We conclude that a host can increase the occupancy of his/her lodging by obtaining the superhost badge, allowing instant booking and by responding quickly to the customers' requests. Moreover, the management of many listings appears to be a disadvantage for occupancy.

Table 3. *LR coefficients for high occupancy*

Variable	Coefficient
host_is_superhost	1.0028
security_deposit	0.5360
instant_bookable	0.5317
review_scores_value	0.2874
is_location_exact	0.2255
bedrooms	0.2296
total_amenities	0.1443
calendar_updated	0.1128
minimum_nights	0.0003
total_price	-0.0177
host_identity_verified	-0.0289
extra_people	-0.0353
cancellation_policy	-0.1047
accommodates	-0.1124
distance_calculation	-0.1849
host_response_time	-0.7994
host_total_listings_count	-7.2739
Intercept	3.2962

Bookings

In the second experiment models able to predict the level of bookings are developed. CFS selected eighteen variables to participate in the final input vector. The results of the five classification methods are depicted in Table 4. All models predict how successful lodgings are in securing bookings, at accuracy rates ranging between 74.39% and 82.62%. The tree based methods have more balanced performance in predicting the two class values, while LR, MLP and SMO tend to favor the low booking cases. Again, the RF classifier achieves the best scores.

Table 4. *Models' performance for bookings*

Classifier	Accuracy	TP Rate High	TP Rate Low	F-measure High	F-measure Low	Roc Area
C4.5	76.80	0.773	0.763	0.772	0.764	0.798
LR	75.68	0.816	0.696	0.773	0.738	0.842
MLP	75.79	0.800	0.714	0.770	0.744	0.836
SMO	74.39	0.800	0.686	0.760	0.725	0.743
RF	82.62	0.811	0.842	0.826	0.827	0.915

The DT uses the variable 'host_response_time' as the first level splitter and highlights this variable as being the most influential. Hosts with response time <= 0.32 (within a few hours) obtain more bookings. Probably, potential customers perceive a fast response as an indication of reliability, a factor that increases trust. The variable 'host_is_superhost' is used as the second level splitter. Interestingly, two host related variables are used as high level splitters. LR also offers an interpretable model. The coefficients of the predictors, listed in Table 5, suggest that the variables 'host_is_superhost', 'host_identity_verfied' and 'instant_bookable' have a strong effect leading to more bookings. The variables which decrease the possibility of a lodging being classified as 'high_bookings' are 'host_responce_time' and 'minimum_nights'. Overall, superhosts who respond quickly, allow instant booking, and verify their identity can increase the bookings of their listings. We highlight that these results are in accordance with the findings concerning occupancy.

Table 5. *LR coefficients for high bookings*

Variable	Coefficient
host_is_superhost	1.2204
host_identity_verified	0.5797
instant_bookable	0.5447
security_deposit	0.2449
host_about_character_count	0.2043
is_location_exact	0.1722
beds	0.1380
total_amenities	0.1241
cancellation_policy	0.0789
bedrooms	0.0556
review_scores_rating	0.0320
extra_people	0.0316
calendar_updated	0.0293
availability_365	-0.0012
total_price	-0.0286
distance_calculation	-0.3114
minimum_nights	-0.5481
host_response_time	-2.7987
Intercept	-2.4244

Revenue

The third class attribute used in the present study is revenue. According to the results of CFS, eighteen variables form the final input vector. Table 6 summarizes the performances of the classifiers. All the models successfully predict the two class values. Top scores are achieved by the RF classifier. In all conducted experiments SMO has the lowest scores among the tested classification methods.

Table 6. *Models' performance for revenue*

Classifier	Accuracy	TP Rate High	TP Rate Low	F High	F Low	Roc Area
C4.5	77.44	0.763	0.786	0.772	0.777	0.808
LR	73.84	0.756	0.721	0.743	0.734	0.817
MLP	75.51	0.771	0.739	0.759	0.751	0.831
SMO	72.36	0.661	0.786	0.705	0.740	0.724
RF	82.63	0.793	0.860	0.820	0,832	0.904

In terms of predictors' significance, the Decision Tree uses the variable 'host_is_superhost' as the first level splitter. The Gain Ratio criterion highlights the superhost badge as the best predictor separating high from low revenue listings in coherent subsets. The variables 'host_response_time' and 'entire_home/apt' are used as second level splitters. Additional evidence regarding the predictors' significance is provided by the LR model. The coefficients indicate that the input variables 'entire_home/apt', 'host_is_superhost', 'instant_bookable' and 'host_identity_verified' are strong factors that increase the financial performance of a listing. The variable 'entire_home/apt' is high, thus signaling that privacy is a consideration for customers. Regarding listings with low profitability, the most significant variables are 'host_total_listings_count', 'host_response_time', 'distance_calculation' and 'host_about_ character_count'. Overall, superhosts who respond quickly, allow instant booking, verify their identity and offer privacy can increase the revenue of their lodging.

Table 7. *LR coefficients for high revenue*

Variable	Coefficient
entire_home/apt	1.6149
host_is_superhost	1.3839
instant_bookable	0.7411
host_identity_verified	0.7163
is_location_exact	0.1717
bedrooms	0.1689
total_amenities	0.1521
cancellation_policy	0.1004
security_deposit	0.0895
extra_people	0.0720
review_scores_rating	0.0287
minimum_nights	0.0094
availability_365	-0.0012
calendar_updated	-0.0107
host_about_character_count	-0.1852
distance_calculation	-0.2504
host_response_time	-2.0049
host_total_listings_count	-2.7098
Intercept	-3.8181

Sensitivity analysis

The targeted examination of isolated aspects related to listings' performance can offer successful models. However, the combined analysis of several issues can offer improved models which incorporate more information and model complex relationships. We perform sensitivity analysis to test the superiority of such a holistic approach. Three new data sets were created and are tested regarding their

predictive power. The first data set contains the host-related attributes, the second contains the lodging-related attributes and the third contains the attributes related to customers' ratings. The RF classifier is used to develop models that predict occupancy, bookings and revenue based on these data sets. The results are listed in Table 8

Table 8. *RF accuracies for different data sets*

Data Set	occupancy	bookings	revenue
full data set	81.59	82.62	82.63
host data	71.68	72.52	72.65
lodging data	67.27	62.37	65.64
Customer ratings	67.23	69.43	69.83

The results provide evidence that the coherent analysis of several diversified issues offers models with higher predictive power. Moreover, we observe that the host-relevant data have a stronger effect than the other data sets on occupancy, bookings and revenue, which may be interpreted as an indication that the host plays an enhanced role in a listing's success.

Significant Variables

The detection of significant factors that affect the performance of Airbnb listings is another main goal of the present study. Two classification methods used, i.e. DT and LR, produce interpretable models and provide insights regarding the predictors' significance. However, there are restrictions. The results obtained refer to reduced data sets, which are the outcome of feature selection. Feature selection optimizes the input vector to maximize the performance of the models, not their interpretability. In order to reduce collinearity, attributes that are related to the class attribute, but are also related to other input variables are eliminated. In the case of CFS, attributes that predict classes in the instance space that are already predicted by other attributes are ignored. Thus, variables strongly related to the class attribute may not be revealed following this approach.

With the purpose of assessing the individual predictive power of each input variable we perform univariate analysis on the initial data set, which contains a total of thirty-four variables. For this purpose, we adopt a complex wrapper approach. The predictive power of each independent variable is estimated by using the classifier itself. By using accuracy as a criterion, the method calculates a score, which indicates the degree of relation between an input variable and the class attribute. Each classification method employed in the present study is used to estimate each variable's significance and their individual scores are integrated in a mean value. This ensemble approach integrates several criteria in a common voting scheme and can produce more reliable results. Table 9 depicts the mean score values for the three dependent variables. Bold characters indicate the most significant variables.

A close examination of the significance scores reveals interesting relations. The superhost badge is by far the most influential factor. The variable 'host_is_superhost' achieves the highest score for all three dependent variables. This finding is in agreement with the results of Han *et al.* (2019). Of particular importance is the speed with which the host responds to customers' requests. This is verified by the complex wrapper evaluator, the LR coefficients and the Information Gain criterion. The aggregator 'host_total', which sums up the values of the host related attributes, is found to strongly influence bookings and profitability. From the lodging-related attributes, only the variable 'total_amenities' has a strong positive impact, suggesting that customers do not concentrate on particular amenities but they appreciate the overall provisions. Interestingly, the variable 'distance_calculation' does not strongly

affect the dependent variables. This may be attributed to specific characteristics of Thessaloniki, where the vast majority of Airbnb lodgings are located around the city center.

Table 9. *Variables' significance*

Variable	Occupancy	Bookings	Revenue
host_about character count	0.03788	0.085593	0.0993
host_response_time	0.041597	**0.161623**	**0.137128**
host_is_superhost	**0.137261**	**0.206277**	**0.228861**
host_total_listings_count	0.042146	0.037012	0.064484
host_has_profile_pic	-0.00171	-0.00133	0.002732
host_identity_verified	0.012972	0.057334	0.079195
host_total	0.051534	**0.098893**	**0.127881**
distance calculation	0.003709	0.085752	0.084309
is_location_exact	0.022124	0.031795	0.034788
entire_home/apt	0.00054	0.016435	0.050088
private_room	0.000472	0.015052	0.04788
shared_room	-0.00225	-0.00028	0.001892
accommodates	0.053816	0.03659	0.064414
bathrooms	0.005742	0.021379	0.00063
bedrooms	0.043258	0.059035	0.043908
beds	0.036498	0.048048	0.051664
total_amenities	**0.07051**	**0.112695**	**0.141254**
security_deposit	0.027603	0.027263	0.031769
total_price	**0.073526**	0.091434	0.041604
extra_people	0.02995	0.07211	0.072259
minimum_nights	0.040364	0.073944	0.065254
calendar_updated	0.046442	0.094968	0.098648
availbility_365		0.036073	0.047742
review_scores_rating	**0.076006**	0.083258	**0.126311**
review_scores_accuracy	0.027084	0.054884	0.112365
review_scores_cleanliness	0.03356	0.040968	0.105317
review_scores_checkin	0.018294	0.040361	0.09937
review_scores_communication	0.016398	0.038068	0.093057
review_scores_location	0.048668	-0.00366	0.067972
review_scores_value	**0.061558**	0.043259	0.099419
instant_bookable	0.027664	0.084577	0.106655
cancellation_policy	0.049254	0.091942	0.114268
require_guest_profile_picture	-0.00236	-0.00042	-0.00042
require_guest_phone_verification	-0.00192	0.00035	0.001401

Privacy, as expressed by the variables 'entire_home/apt', 'private_room' and 'shared_room', is not significant, however this could be another peculiarity of the Airbnb market in Thessaloniki, where the vast majority of the listings are entire homes or apartments. In terms of reviews, guests appear to be influenced by the total review score, while they also pay specific attention to particular issues like check-in, communication, accuracy etc. A general conclusion is that the host plays a central role in listings' performance since three host-related attributes have been highlighted as strong influencers. A superhost who responds quickly to customers' requests and communicates a proper image can significantly improve the performance of his/her lodging.

Conclusions

The exponential growth of Airbnb attracts the interest of many researchers and a topic of major importance is the examination of factors that influence listings' performance. In the present study we employ data mining to analyze the success of Airbnb listings, as reflected by occupancy rate, number of bookings, and revenue. The data set used is drawn from Airbnb accommodation offers in Thessaloniki, Greece and contains a wide range of information pertaining to hosts' characteristics, lodgings' characteristics, rules, and customers' ratings. Such enriched data sets can offer improved models which incorporate more information, thus allowing to model complex relationships. By using the number of customer reviews as a proxy for room reservations, we introduce three dependent variables reflecting occupancy, bookings and revenue. The purpose of the study is twofold, to develop models capable of predicting listings' performance and to reveal factors that strongly influence customers' purchase decisions. The predictors' significance is assessed by using interpretable models, sensitivity analysis, and a complex wrapper which integrates individual classifiers in a common voting scheme.

Several conclusions can be drawn from the results of the conducted research. Publicly available data can be used to estimate Airbnb listings' performance. All the employed classifiers offered successful models, however the Random Forest method managed to outperform all the competitors and is therefore suggested as the most suitable classifier for the specific domain.

The interpretation of the comprehensible models, sensitivity analysis and the wrapper evaluator provide evidence on the significance of the predictors. According to the results, the superhost badge emerged as the most influential factor. Airbnb grants the superhost badge to hosts who satisfy certain requirements. These are a) at least 10 bookings per year, b) at least a 90% response rate, c) no cancelations (with the exception of specific cases), d) at least 80% 5-star ratings. Thus, the superhost badge integrates sufficient occupancy, a host's commitment and responsiveness, and guests' high ratings. Furthermore, the speed with which the host responds to customers' requests is of particular importance. The total of host-related information also strongly influences bookings and profitability. We highlight that three alternative analytical methods and experiments converge to the conclusion that the host plays a central role in Airbnb listings' performance.

This outcome deserves further exanimation. Del Chiappa *et al.* (2021) perform a survey among *non* Airbnb users to identify constraints that prevent them from using Airbnb. Their results suggest that distrust is a major concern. The authors claim that 'Distrust was felt by the majority of participants, especially in the hosts, who were perceived to be misleading when advertising their listings'. They also point out that travelers associate Airbnb with declining professionalism and quality of service. On the other hand, Xie and Mao (2017) mention that the superhost program was developed by Airbnb to identify the most *trusted* hosts. The authors also conclude that host-related quality information cues have a significant effect on trust. Moreover, Liang *et al.* (2017) stress that the superhost badge acts as a motivation for hosts to improve or professionalize their services. The theoretical implication that

derives from our results, in combination with the existing literature, is that host quality attributes (i.e. the superhost badge, quick response, rich host presentation) mitigate distrust and signal a more professional attitude, thus leading to improved listings' performance.

In terms of lodging-related attributes, total amenities have a strong positive impact, even though particular amenities were not highlighted as significant. Another determinant of listings' performance is the overall guests' ratings. Remarkably, both provision of amenities and high ratings increase the value perceived by the potential customers. Perceived value emerges as another prominent factor. Interestingly, price marginally affects performance and is not a major customer concern. In short, the superhost badge, quick response, a rich presentation of the host, the provision of amenities and high overall guest ratings seem to constitute the recipe for success.

The present study offers important implications for practitioners. First, hosts are provided with insights into the best tactics to increase their revenue. More specifically, these results suggest that hosts should boost their personal exposure. Moreover, they should improve their response rate, responding to guests' requests as soon as possible. It is also clear that investing in amenities is a tactic that pays off. Above all, according to our analyses, hosts should strive to obtain and retain the superhost badge, a signal of trustworthiness and quality. Second, owners of low performing lodgings can take advantage of the predictive model and perform what-if analysis to evaluate alternative actions allowing them to define an optimal strategy to increase occupancy, bookings and revenue. Markedly, the model can easily be tuned to differentiate the top 30% or 20% of listings. Third, taxation authorities can use the model to tackle tax evasion by locating hosts whose declared profits are low and in disagreement with the models' predictions.

We acknowledge some limitations of the present study. First, the policy of Airbnb not to provide exact information about lodgings' occupancy rate does not allow the definition of an absolute measure of performance, and thus approximations must be used. Second, the data set used refers to the Airbnb market of Thessaloniki and may incorporate some local peculiarities. The examination of Airbnb data drawn from lodgings located in other cities could cross-validate our results and conclusions. As a topic for further research, we propose the enrichment of the input vector with more information, such as the sentiment of the guests' reviews, or the characteristics of lodgings' or hosts' photos. The superhost badge should also be further investigated to clarify the extent to which it boosts lodgings' performance in its own right and independently of the features it summarizes.

References

Adbar, M., & Yen, N. (2017). Sharing economy and its effect on human behaviour changes in accommodation: a survey on Airbnb. *International Journal of Social and Humanistic Computing,* 2(3/4), 203-218. doi: 10.1504/ijshc.2017.084747

Allen, D., & Berg, C. (2014). The sharing economy. How over-regulation could destroy an economic revolution. Retrieved 3 August 2019, from https://collaborativeeconomy.com/wp/wp-content/uploads/2015/04/Allen-D.-and-Berg-C.2014.The-Sharing-Economy.-Institute-of-Public-Affairs.-.pdf

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. Retrieved 15 March 2020, from https://www.wired.com/2008/06/pb-theory/

Botsman, R., & Rogers, R. (2011). *What's mine is yours: how collaborative consumption is changing the way we live*. London: Collins.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. doi:10.1613/jair.953

Chen, C., & Chang, Y. (2018). What drives purchase intention on Airbnb? Perspectives of consumer reviews, information quality, and media richness. *Telematics and Informatics, 35*(5), 1512-1523. doi: 10.1016/j.tele.2018.03.019

Chen, D., Lou, H., & Van Slyke, C. (2015). Toward an Understanding of Online Lending Intentions: Evidence from a Survey in China. *Communications of The Association For Information Systems, 36*, 317-336. doi: 10.17705/1cais.03617

Chen, Y., & Xie, K. (2017). Consumer valuation of Airbnb listings: a hedonic pricing approach. *International Journal of Contemporary Hospitality Management, 29*(9), 2405-2424. doi: 10.1108/ijchm-10-2016-0606

Davis, P. (2016). How Do Sharing Economy Companies Grow? A Comparison of Internal and External Growth Patterns of Airbnb and Uber. Retrieved 10 August 2019, from https://pdfs.semanticscholar.org/09ed/4c7f1123c1cdf9fecb8858ff83b8d1f93bc5.pdf?_ga=2.157513336.371552843.1569162250-2081967136.1568455273

Del Chiappa, G., Pung, J.M., Atzeni, M., & Sini, L. (2021). What prevents customers that are aware of Airbnb from using the platform? A mixed methods approach. *International Journal of Hospitality Management, 93*, 102775. doi: 10.1016/j.ijhm.2020.102775

Del Chiappa, G., Sini, L., & Atzeni, M. (2020). A motivation-based segmentation of Italian Airbnb users: An exploratory mixed method approach. *European Journal of Tourism Research, 25*, 2505.

Dogru, T., Mody, M., & Suess, C. (2017). Comparing apples and oranges? Examining the impacts of Airbnb on hotel performance in Boston. Boston Hospitality Review, Blog Archive, Boston University. Retrieved 4 August 2019, from http://www.bu.edu/bhr/2017/06/07/airbnb-in-boston/

Dudás, G., Kovalcsik, T., Vida, G., Boros, L., & Nagy, G. (2020). Price determinants of Airbnb listing prices in Lake Balaton Touristic Region, Hungary. *European Journal of Tourism Research, 24*, 2410.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and Reputation in the Sharing Economy: The Role of Personal Photos on Airbnb. *Tourism Management, 55*(Supplement C), 62–73. doi: 10.2139/ssrn.2624181

Fagerstrøm, A., Pawar, S., Sigurdsson, V., Foxall, G., & Yani-de-Soriano, M. (2017). That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on Airbnb™. *Computers In Human Behavior, 72*, 123-131. doi: 10.1016/j.chb.2017.02.029

Fang, B., Ye, Q., & Law, R. (2016). Effect of sharing economy on tourism industry employment. *Annals of Tourism Research, 57*, 264-267. doi: 10.1016/j.annals.2015.11.018

Festila, M., & Müller, S. (2017). The Impact of Technology-Mediated Consumption on Identity: the Case of Airbnb. In *the 50th Hawaii International Conference on System Sciences* (pp. 54-63). Hilton Waikoloa Village. Retrieved from https://dblp.org/db/conf/hicss/hicss2017.html

Fortunebusinessinsights.com (2021). Big Data Technology Market Size, Share, Demand & Growth Retrieved from https://www.fortunebusinessinsights.com/industry-reports/big-data-technology-market-100144.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing, 35*(1), 46-56. doi: 10.1080/10548408.2017.1308292

Gurran, N., & Phibbs, P. (2017). When Tourists Move In: How Should Urban Planners Respond to Airbnb? *Journal of The American Planning Association, 83*(1), 80-92. doi: 10.1080/01944363.2016.1249011

Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Current *Issues in Tourism, 18*(12), 1192-1217. doi: 10.1080/13683500.2013.827159

Guttentag, D. (2016). *Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts* (Ph.D.). University of Waterloo.

Hall, M.A. (1999). *Correlation-Based Feature Selection for Machine Learning* (Ph.D.). University of Waikato.

Han, H., Shin, S., Chung, N., & Koo, C. (2019). Which appeals (ethos, pathos, logos) are the most important for Airbnb users to booking? *International Journal of Contemporary Hospitality Management, 31*(3), 1205-1223. doi: 10.1108/ijchm-12-2017-0784

Inside Airbnb (n.d.). Get the data. Retrieved from http://insideairbnb.com/get-the-data.html

Koopman, C., Mitchell, M., & Thierer, A. (2015). The Sharing Economy and Consumer Protection Regulation: The Case for Policy Change. *The Journal of Business, Entrepreneurship & The Law, 8*(2), 529-545. doi: 10.2139/ssrn.2535345

Kozinets, R.V., & Gretzel, U. (2021). Commentary: Artificial Intelligence: The Marketer's Dilemma. *Journal of Marketing, 85*(1), 156-159. doi: 0.1177/0022242920972933

Lee, D. (2016). How Airbnb Short-Term Rentals Exacerbate Los Angeles's Affordable Housing Crisis: Analysis and Policy Recommendations. Retrieved 9 July 2019, from http://blogs.ubc.ca/canadianliteratureparkinson/files/2016/06/How-Airbnb-Short-term-rentals-disrupted.pdf

Lee, D., Hyun, W., Ryu, J., Lee, W., Rhee, W., & Suh, B. (2015). An Analysis of Social Features Associated with Room Sales of Airbnb. In *the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 219-222). Vancouver, BC, Canada.

Leonelli, S. (2020). *Scientific Research and Big Data (Stanford Encyclopedia of Philosophy/Summer 2020 Edition)*. Plato.stanford.edu. Retrieved 23 January 2021, from https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.

Liang, S., Schuckert, M., Law, R., & Chen, C. (2017). Be a "Superhost": The importance of badge systems for peer-to-peer rental accommodations. *Tourism Management, 60*, 454-465. doi: 10.1016/j.tourman.2017.01.007

Liang, S., Schuckert, M., Law, R., & Chen, C. (2020). The importance of marketer-generated content to peer-to-peer property rental platforms: Evidence from Airbnb. *International Journal of Hospitality Management, 84*, 102329. doi: 10.1016/j.ijhm.2019.102329

Lutz, C., & Newlands, G. (2018). Consumer segmentation within the sharing economy: The case of Airbnb. *Journal of Business Research, 88*, 187-196. doi: 10.1016/j.jbusres.2018.03.019

Mazanec, J. (2020). Hidden theorizing in big data analytics: With a reference to tourism design research. *Annals of Tourism Research, 83*, 102931. doi: 10.1016/j.annals.2020.102931

Nisar, T., Hajli, N., Prabhakar, G., & Dwivedi, Y. (2019). Sharing economy and the lodging websites. *Information Technology & People 33*(3), 873-896. doi: 10.1108/itp-06-2018-0297

Oskam, J., & Boswijk, A. (2016). Airbnb: the future of networked hospitality businesses. *Journal of Tourism Futures, 2*(1), 22-42. doi: 10.1108/jtf-11-2015-0048

Palombo, D. (2015). A Tale of Two Cities: The Regulatory Battle to Incorporate Short-Term Residential Rentals into Modern Law. *American University Business Law Review, 4*(2), 287-320.

Peng, N., Xie, Y., Li, Y., Wen, H., Dai, D., & Subinur. (2019). What's Your Ideal Online Short-Term Accommodation? Demand Mining for Chinese Tourists. In *the 8th International Conference on Industrial Technology and Management (ICITM)* (pp. 369-374). Cambridge, United Kingdom: IEEE.

Smith, A. (2016). The New Digital Economy: Shared, Collaborative and On Demand. Retrieved 6 July 2019, from https://www.pewinternet.org/2016/05/19/the-new-digital-economy/

Sparks, B., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management, 32*(6), 1310-1323. doi: 10.1016/j.tourman.2010.12.011

Tussyadiah, I. (2016). Factors of satisfaction and intention to use peer-to-peer accommodation. International *Journal of Hospitality Management,* 55, 70-80. doi: 10.1016/j.ijhm.2016.03.005

Vapnik, V.N. 1995. *The Nature of Statistical Learning Theory*, Springer Verlag, New York , USA.

Wang, D., & Nicolau, J. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management, 62*, 120-131. doi: 10.1016/j.ijhm.2016.12.007

Wu, J., Ma, P., & Xie, K. (2017). In sharing economy we trust: the effects of host attributes on short-term rental purchases. *International Journal of Contemporary Hospitality Management, 29*(11), 2962-2976. doi: 10.1108/ijchm-08-2016-0480

Xie, K., Kwok, L., & Wu, J. (2019). Are consumers loyal to home-sharing services? *International Journal of Contemporary Hospitality Management, 31*(3), 1066-1085. doi: 10.1108/ijchm-09-2017-0552

Xie, K., & Mao, Z. (2017). The impacts of quality and quantity attributes of Airbnb hosts on listing performance. *International Journal of Contemporary Hospitality Management, 29*(9), 2240-2260. doi: 10.1108/ijchm-07-2016-0345

Yang, S., Lee, K., Lee, H., & Koo, C. (2018). In Airbnb we trust: Understanding consumers' trust-attachment building mechanisms in the sharing economy. *International Journal of Hospitality Management, 83*, 198-209. doi: 10.1016/j.ijhm.2018.10.016

Yannopoulou, N., Moufahim, M., & Bian, X. (2013). User-Generated Brands and Social Media: Couchsurfing and Airbnb. *Contemporary Management Research, 9*(1), 85-90. doi: 10.7903/cmr.11116

Zervas, G., Proserpio, D., & Byers, J. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of Marketing Research, 54*(5), 687-705. doi: 10.1509/jmr.15.0204

Zhao, J., & Peng, Z. (2019). Shared Short-Term Rentals for Sustainable Tourism in the Social-Network Age: The Impact of Online Reviews on Users' Purchase Decisions. *Sustainability, 11*(15), 4064. doi: 10.3390/su11154064

Received: 10/12/2020
Accepted: 16/03/2021
Coordinating editor: Giacomo Del Chiappa