International Journal of Hospitality Management 89 (2020) 102546

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhm

Using machine learning and big data for efficient forecasting of hotel booking cancellations

Agustín J. Sánchez-Medina*, Eleazar C-Sánchez

Instituto Universitario de Ciencias y Tecnologías Cibernéticas (IUCTC), University of Las Palmas de Gran Canaria, Despacho C-2.21, Ed. de Económicas y Empresariales, Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain

ARTICLE INFO

Keywords:
Cancellation forecasting
Hotel booking
Artificial neural network
Genetic algorithm
Tree decision algorithm

ABSTRACT

Cancellations are a key aspect of hotel revenue management because of their impact on room reservation systems. In fact, very little is known about the reasons that lead customers to cancel, or how it can be avoided. The aim of this paper is to propose a means of enabling the forecasting of hotel booking cancellations using only 13 independent variables, a reduced number in comparison with related research in the area, which in addition coincide with those that are most often requested by customers when they place a reservation. For this matter, machine-learning techniques, among other artificial neural networks optimised with genetic algorithms were applied achieving a cancellation rate of up to 98%. The proposed methodology allows us not only to know about cancellation rates, but also to identify which customer is likely to cancel. This approach would mean organisations could strengthen their action protocols regarding tourist arrivals.

1. Introduction

Tourism is one of the most expanded industries in the world and its importance in the global economy does not raise any doubts. Evidence of its growth can be seen when considering that in 1990 the number of international tourists slightly exceeded the 400 million mark, whereas in 2017 this number increased to 1300 million. In 2017 alone the number of international tourists grew by 7%, which represents the highest increase since the 2009 global economic crisis (World Tourism Organization, 2018). In terms of economic development, the tourism industry generates income to nations through the consumption of goods and services by tourists, taxes, development of enterprises and employment opportunities, among others. However, the environment where the tourism develops its activity forces it to face constant uncertainty. In this sense, the tourism industry is extremely sensitive to multiple external factors that may have a significant impact on income, such as political instability, weather and natural disasters to mention but a few (Chow et al., 1998). Furthermore, the hotel industry offers a product, which does not become stockpiled, thus, at the end of the day, each unoccupied room means a lost revenue opportunity (Chu, 2009; Frechtling, 2001; Witt and Witt, 1995; Heo and Lee, 2009). This situation makes it extremely important to understand short-term demand, as well as plan future demand in advance in order to maximise the occupancy of a fixed inventory (Frechtling, 2001; Rajopadhye et al., 2001). Therefore, organisation and planning are essential and having an accurate forecasting tool is of paramount importance (Chow et al., 1998; Haensel and Koole, 2011; Hassani et al., 2017; Pereira, 2016; Schwartz and Cohen, 2004; Teixeira and Fernandes, 2012; Weatherford and Kimes, 2003; Yüksel, 2007). In fact, the decision-making process is influenced by forecasts, which may be classified according to the time horizon. Along this line, Hassani et al. (2017) classify forecasts within the tourism industry according to three temporal horizons: long-term forecasting, which may help to plan for global infrastructure and services (e.g. airports or highways); medium-term forecasting, which may help with market analysis (e.g. establish sales strategies) and short-term forecasting, which may provide more operational flexibility (Gunter and Önder, 2015; Park et al., 2017; Song and Li, 2008). Along these lines, hotel industry revenue is strongly affected by demand so having an accurate forecasting system becomes critical. As stated by Kourentzes et al. (2017), forecasting demand plays a crucial role in modern organisations because of its impact on a variety of business decisions, from operational, to tactical, to strategic level, such as capacity planning, resource planning, advertising and promotional planning or tactical production planning, among others. However, hotel demand is difficult to forecast (Li et al., 2018; Pan and Yang, 2017), because of its complex and dynamic behaviour (Ostaijen et al., 2017), which makes it a very challenging task (Li et al., 2018). As long as hotel accommodation demand varies and reservations may be cancelled, an

* Corresponding author.
E-mail addresses: agustin.sanchez@ulpgc.es (A.J. Sánchez-Medina), eleazar.caballero101@alu.ulpgc.es (E. C-Sánchez).

https://doi.org/10.1016/j.ijhm.2020.102546
Received 15 September 2019; Received in revised form 26 April 2020; Accepted 27 April 2020
Available online 31 May 2020
0278-4319/

accurate method to determine effective hotel occupancy is crucial for decision making in terms of successful revenue management (Haensel and Koole, 2011). Management decisions are clearly influenced by customer demand, but not just regarding reservations as both reservations and cancellations are the main components of effective demand or net-demand (Rajopadhye et al., 2001; Zakhary et al., 2011). In this regard, it is estimated that about 20% of income is lost because of not considering cancellations as part of room reservation management systems (Sierag et al., 2015). However, literature available on cancellation forecasting within the industry is underdeveloped (Antonio et al., 2017a, b; Zakhary et al., 2011) and little is known about the reasons that lead guests to cancel or how to avoid it (Hajibaba et al., 2016). Moreover, the majority of publications in this area attempt to forecast the cancellation ratio, but very few address the prediction of cancellations by individual customers (e.g Antonio et al., 2017a, a; Falk and Vieru, 2018). According to Antonio et al. (2019b) only two publications focused on the importance of identifying which individual is likely to cancel have been released. Following these two publications, only one more by Antonio et al. (2019a) has been issued.

In recent decades the cancellation problem has become even more dramatic with the increasing use of internet for placing bookings as it is far easier to cancel the reservation (Park et al., 2007). Among the multiple reasons for cancelling a reservation, such as illness, bad weather or natural disasters, the ease of looking for new opportunities and cancelling previous reservations, as well as the possibility of booking several places at the same time to keep options open until making a final decision, have a big impact on cancellations (Antonio et al., 2017a; Chen et al., 2011; Mayr and Zins, 2009). Despite the scarce amount of literature addressing this topic, the limited number of research papers that undertake the prediction of hotel booking cancellations through PNR data disagree about the usefulness of this kind of data for predicting individual cancellations.

This research contributes to the hotel and lodging industry by proposing and testing an empirical model based on artificial intelligence for forecasting hotel cancellations using personal name records (PNR). The use of PNR data with the aim of forecasting within the tourism industry, which supposes a relatively new approach (Gorin et al., 2006), and the use of big data techniques for this matter makes for a novelty approach in contrast to the traditional techniques (Pereira, 2016). However, not all research published in this field reaches the same conclusions with regards the validity of PNR data. In fact, one of the main aims of this research is to shed light on the usefulness of PNR data for forecasting individual hotel cancellations after academic discussion arose on the research carried out by Romero Morales and Wang (2010), who concluded that it was not possible and Antonio et al. (2017a) & Antonio et al. (2019a), who concluded the opposite. In addition, another novelty of this approach is that individual hotel booking cancellations are forecast using a reduced number of variables in comparison with other researches (Antonio et al., 2019a, a), achieving a similar high success rate, thus avoiding the need for large historical booking datasets by either building complementary variables through additional calculation based on database queries or getting access to external databases. In fact, one of the main assets of this approach is the simplification of the procedure for building the dataset, as well as, the dataset itself, which makes for a faster and simpler training phase of the probabilistic models meaning models can be trained more frequently, thus allowing a better following of market trends. For this reason, additional variables that could be extracted from the individuals after historical analysis of the database were not considered, as this would have increased the complexity for the hotels to use this procedure. Furthermore, the variables used for this purpose coincide with those commonly requested from customers by popular web booking platforms such as Booking, Tripadvisor or Trivago, as well as large hotel chains like Radisson Hotels & Resorts, Riu Hotels & Resorts or Meliá Hotel Resorts among others, therefore the proposed methodology works with those variables that hotels may have access to. Finally, another novelty in the proposed model is that genetic algorithms are used for configuring the structural parameters of the artificial neural network.

2. Literature review

2.1. Difficulties and needs for the hotel and lodging industry

There is a strong association between demand forecast and revenue management (Tse and Poon, 2015). As long as hotels are required to manage room occupancy within an uncertain environment, they are exposed to unclear incomes and forced to assume business risks. Moreover, unexpected reduced demand often generates a crisis in the hospitality industry because of the high sensitivity to fluctuations in demand (Yüksel, 2007). In addition, demand uncertainty does not only affect the organisation and scheduling of occupancy, but also internal issues such as budget planning, which is highly dependent on future demand forecasting (Tang et al., 2016). However, demand may not be properly forecasted if the cancellation rate is not taken into account, hence, the variable net-demand should be considered (Rajopadhye et al., 2001), which is defined as “the number of demand requests minus the number of cancellations” (Romero Morales and Wang, 2010, pp.1). Consequently, hotels affront this situation by implementing their own approaches in order to handle the associated risk management, such as overbooking strategies, cancellation policies or pricing strategies. On the one hand, overbooking strategies consist in accepting reservations over and above the capacity of the establishment, assuming that some bookings will fail. Nevertheless, extra costs may be incurred when the actual hotel occupancy exceeds the capacity due to guest compensation or relocation, which may also lead to a negative impact on the reputation. On the other hand, cancellation policies try to mitigate revenue loss because of cancellations, which are specially high when referring to last minute cancellations and no-shows (Chen et al., 2011). According to Zakhary et al. (2011) cancellations are found to decrease dramatically if penalties are imposed for cancelling beyond a certain day. However, imposing rigid cancellation policies can affect not only, corporate social reputation but also income because it has a discouraging effect on clients or, due to the application of significant discounts on price (Antonio et al., 2017b). Other strategies, such as price wars are discouraged due to the fact that they may affect the business strategy in the long-term (Gehrels and Blanar, 2013). In any case, the effectiveness of the different prevention approaches varies across the tourist segments (Hajibaba et al., 2016).

A reliable and accurate cancellation forecast may help in the managerial decision taking process by reducing the risk of cancellation as well as helping establish a proper cancellation policy or pricing strategy. Moreover, Pan and Yang (2017) explain that the tourism sector needs accurate forecast performance in specific destinations to benchmark their properties and optimise operations. Additionally, they state that as competition increases, accurate short-term forecasts become essential for hotel managers. Accordingly, Kouprioushina et al. (2014, pp.2) state than when hotels face elevated levels of risk and distress (e.g. intensified competition or highly volatile markets), “more pressure is placed on the revenue manager to ensure that the forecasts are accurate” and reliable. So, by reducing uncertainty in future net-demand, occupancy rates can be increased while costs depending on idle capacity can be handled more efficiently. Likewise, an accurate cancellation model may prevent hotels from implementing rigid cancellation policies and overbooking strategies which have a negative influence on revenue and reputation (Antonio et al., 2017a).

Finally, it should be highlighted that some issues have arisen in recent decades that have given even more importance to accurate cancellation forecasting. Information technologies have changed customer behaviour and have made it even more difficult to predict future demand and cancellation rates. Now customers have more information about the establishments and the services they offer, for example, they can read previous customer experiences which makes it easier to

compare different offers. Web portals also make it easier to book and cancel hotel services, which has encouraged people to place several bookings on similar dates in different hotels, looking for the most convenient options, only to finally choose one of them and cancel the rest (Chen et al., 2011). Consequently, demand performed by websites seems to increase, but cancellation rates do also. Secondly, another phenomenon that has influenced net-demand forecast is the growth of last-minute bookings whereby customers attempt to take advantage of these kinds of opportunities by postponing their booking. This causes a reduction in the length of the booking window and has an impact on forecasting accuracy (MacCarthy et al., 2016). In the same manner, such policies lead to other reactions; the probability of cancellations by guests who have already booked may also increase as they may be inclined to change to a more economic option, which generates more cancellations. In this way, if last-minute chances are offered by the hotel itself or by competing companies a few days before the time of service, the probability of cancelling increases as the time gets closer.

2.2. Methods and techniques for demand and cancellation forecasting

Looking for patterns and forecasts is always supported by previous experience in the service, the challenge is how to perform them in the best way considering "data availability, time horizons and objectives" (Lee et al., 2008, pp.2). In this section, previous approaches regarding demand and cancellation forecasting methods applied within the hospitality industry are reviewed. These techniques are classified in qualitative and quantitative techniques; the main characteristics of both are presented in the following subsections. For a more extensive review, specific research in this field such as Song and Li (2008) or more recently Song et al. (2019) is highly recommended.

2.2.1. Qualitative techniques

Qualitative techniques employ a team of experts who determine tendencies and probabilities based on available data, own experience and knowledge in the field. These techniques are recommended when unprecedented changes are to come and therefore, historical data does not contain information about future events, are unsuitable or not sufficient enough to perform an appropriate forecast. For example, long-term forecasts with large and/or extraordinary changes (Lee et al., 2008) such as the growing interest in nature-based tourism or short-term forecasts in which unprecedented events are expected to have an impact on the business (Uysal and Crompton, 1985), such as the emergence of new competitors or natural disasters. Among the most relevant techniques, Delphi and scenario writing are the most popular (Lin and Song, 2015). However, qualitative techniques have less presence in literature (Song and Li, 2008) and do not get accurate results if they are not based on quantity (Yüksel, 2005). As an example of qualitative techniques applied within the hotel and lodging industry, Moutinho and Witt (1995) used a consensus approach to forecast long-term tourism environments. More recently, Kaynak and Cavlek (2007) applied the Delphi technique in order to forecast the potential tourism market in Croatia, while Lee et al. (2008) forecasting the demand for the International Expo tourism held in Korea in 2012.

Other authors have proposed mixed approaches, which attempt the combination of quantitative and qualitative techniques in a "quasi-Delphi process" by the integration of statistical methods and judgement of experts with the aim of forecasting tourism arrivals (Tideswell et al., 2001). Other related research papers propose the use of quantitative methods in order to forecast hotel demand and average nights of stay, which is subsequently adjusted by experts periodically (Yüksel, 2005).

2.2.2. Quantitative techniques

Quantitative approaches require the existence of sufficient and appropriate historical data (Uysal and Crompton, 1985). These techniques are an optimal option if past information can be quantified and past patterns can be reasonably extrapolated to the future (Lee et al., 2008). This section reviews the most relevant quantitative forecasting models, which are categorised, according to Peng et al. (2014) and Song and Li (2008), in non-causal time-series models, econometric models and models based on artificial intelligence. It is worth mentioning that most of the cases reviewed are focused on tourism destination forecasting, while research focused on forecasting hotel booking arrivals have less presence in literature (Lee, 2018).

Non-causal time-series models attempt to reveal future patterns based on historical data. The Integrated Autoregressive Moving Average model (ARIMA) has been the most widely time-series model used for demand forecasting in the past decades, although seasonal ARIMA (e.g. SARIMA) models have increased in popularity over the years because of the strong relationship between tourism and seasons (Song and Li, 2008). Chu (2009) used the ARMA-based methods in the context of predicting tourist demand, as represented by the number of world-wide visitors to Hong Kong, Japan, Korea, Taiwan, Singapore, Thailand, the Philippines, Australia and New Zealand. In this research, ARIMA-based models were applied, concluding that all methods provided a good performance using monthly and quarterly data sets. Cho (2003) evaluates the application of three time-series forecasting techniques: exponential smoothing, univariate ARIMA, and Elman's Model of Artificial Neural Networks (recurrent neural networks); to predict travel demand from different countries to Hong Kong. This study concludes that neural networks are the best method for forecasting visitor arrivals, especially those series without obvious patterns. Claveria and Datzira (2010) introduce consumer expectations in time-series models with the aim of analysing their usefulness in the forecast of tourism demand applied to the four main visitor markets (France, the UK, Germany and Italy) in Catalonia. The paper uses combining qualitative information with quantitative models: Auto Regressive (AR), Auto Regressive Integrated Moving Average (ARIMA), Self-Exciting Threshold Auto Regressions (SETAR) and Markov Switching Regime (MKTAR) models. In addition, models are evaluated for different time horizons (one, two, three, six and 12 months). Conclusions support that ARIMA and Markov Switching Regime models outperform the rest of the models and models that consider consumer expectations do not give best results for all time horizons analysed. With regards hotel demand forecasting, Pfeifer and Bodily (1990) applied a space-time ARMA (STARIMA) approach with the aim of forecasting hotel arrivals for 8 different hotels belonging to the same hotel chain in a single metropolitan US city. STARIMA assumes that a special dependency among multiple points exists and attributes more weight to the closer ones. They finally concluded that STARIMA outperformed one single ARIMA time series model.

On the other hand, econometric models can be classified into static models, such as the traditional regression method, gravity models or the static Almost-Ideal Demand System (AIDS); and dynamic, such as Vector Auto Regressive (VAR), Time Varying Parameters (TVP) or the Error Correction Models (ECM) (Peng et al., 2014). Song et al. (2011) examine the factors that influence the demand for hotel rooms in Hong Kong to generate quarterly forecasts of that demand and to assess the impact of the financial/economic crisis. The paper uses econometric approaches to calculate the demand elasticity and its corresponding confidence intervals. Both indicators are then used to generate interval demand predictions.

As examples of cancellation forecasting, Falk and Vieru (2018, pp.2) studied the factors that influence cancellation behaviour with respect to hotel bookings. In this study, variables such as length of stay, hotel, category, booking time or arrival month, among others, were used. The probability of cancellation is estimated by a probit model with cluster adjusted standard errors at the hotel level. Results show that cancellation rates are higher for online bookings than offline bookings and travel agency bookings. Additionally, they found that "booking lead time and country of residence play the largest role, particularly for online bookings".

As noted by Wu et al. (2017), artificial intelligence techniques have

been used in tourism and hotel demand forecasting with satisfactory performance. These authors note that Artificial Neural Networking (ANN) is the most frequently used method, although Vector Support Machines (SVM) or fuzzy methods have also been applied in this field. Along these lines, Claveria et al. (2015) propose three different architectures of artificial neural networks with the aim of forecasting tourist arrivals to Catalonia attending different time horizons (one, three and six months) and main visiting nationalities, concluding that multi-layer perceptron and radial basis function networks have a good performance. Huang (2014) also applied Artificial Neural Networking with a back-propagation architecture for forecasting tourism demand at a resort in Taiwan. This research used several local and international variables for the purpose, such as unemployment rate, international oil prices or the number or foreign visitors to Taiwan among others, concluding that it was an excellent method. Later, Huang and I Hou (2017) applied similar methodology adding genetic algorithms for optimising ANN settings with the aim of forecasting the sales revenue of a travel agency, achieving good results. Other related studies, for instance the research conducted by Moutinho et al. (2008) uses a neural network based fuzzy time series with the aim of forecasting the arrival of Chinese tourists to Taiwan. They conclude that this method outperformed previous research in which only fuzzy time-series were applied (Jeng-Ren Hwang et al., 1998) and others in which the same methodology was applied but only with the maximum degree of memberships (Huarng et al., 2007). Along these lines, Yu and Schwartz (2006) used two artificial intelligence forecast methods – fuzzy time series and grey theory – to predict annual U.S. tourist arrivals. They suggest that given the complexity and cost associated with the application of these two methods, it is imperative to compare their performance with the accuracy of more traditional and easier methods of forecasting. More recently, Hu et al. (2019) modelled tourism demand by incorporating neural networks into Grey–Markov models to forecast the number of foreign tourists using historical annual data from the Taiwan Tourism Bureau and China National Tourism Administration. The paper confirms that the proposed model outperforms other Grey–Markov models.

In terms of hotel cancellation forecasting, there is less amount of literature that addresses cancellation forecasting. Romero Morales and Wang (2010) forecasted cancellation rates for services booking revenue management using data mining. They used 14 variables, such as price of the booking, room type, channel used to make the booking or reservation system used. They stated that tree-based methods and kernel methods are the most popular for hotel cancellation forecasting, specifically, they note that Support Vector Machine (SVM) is the most notable method used for this purpose. Huang et al. (2013) used Back Propagation Neural Networking (BPN) and General Regression Neural Networking (GRNN) for forecasting booking cancellations, concluding that both methods revealed high potential for this purpose. Later Antonio et al. (2017a) applied diverse two-class classification algorithms: Boosted Decision Tree, Decision Forest, Decision Jungle, Locally Deep Support Vector Machine and Neural Network; with the aim of forecasting cancellation rates using data sets from four hotels located in the resort region of the Algarve (Portugal). For this study they use variables such as number of previous bookings not cancelled, previous stays, distribution channel or days of week of booking dates, among others, concluding that machine learning algorithms, specifically decision forest, are good methods for modelling hotel cancellations.

3. Methodology

This research has been developed using real booking records provided by a hotel partner located in Gran Canaria (Spain) with the aim of forecasting future cancellations. According to CRISP-DM process (CRoss Industry Standard Process for Data Mining) (Wirth and Hipp, 2000), before any data preparation takes place, it is necessary to understand the business and data itself, then, models may be built and tested. In this section, the data preparation and modelling process used for this research are detailed.

3.1. Understanding and preparing data

Although seasonal models were commonly used, mainly because it was the only information available, when revenue management systems started to include historical booking records, forecasting methods were developed using such information (Romero Morales and Wang, 2010). These data are known as Personal Name Records (PNR) that are composed of the information provided by guests at the time a reservation is placed, such as the sale channel, additional hotel services, number of customers and others. For this research only two years of booking records with more than 10,000 bookings including all of 2016 up to April 2018 for a four-star hotel partner located in Gran Canaria (Spain) were used with the aim of forecasting cancellations. Some relevant descriptive information of the used dataset is that customers come from more than 30 countries worldwide, of which German and UK nationalities fall slightly short of 50% of all reservations; followed by Spain and Holland with around 8% each and Switzerland and Sweden with around 5.5% each. The length of stay for 97% of clients varies from one single day to 14 days, of which around the 40% stay for 7 days. With regards cancellations, around 30% of reservations are cancelled prior to the consumption of the service. Similarly, it is important to note that, while other locations are more dependent on seasonal demand, the good weather conditions of this location encourage customers to place reservations during the whole year, so that, demand is less sensitive to seasonality.

One of the main objectives of this research is to build a hotel cancellation model using the most common variables requested from customers when they place a booking. Therefore, the most popular web booking portals were reviewed in the search for those variables frequently requested from customers when they place a reservation, such as Booking, Tripadvisor or Trivago, as well as large hotel chains like Radisson Hotels & Resorts, Riu Hotels & Resorts or Meliá Hotel Resorts among others. Among the requested variables, all websites consulted used at least the variables proposed for this research (Table 1). In the same manner, as extensive information about the selected variables, was required, those items with missing values were removed. Furthermore, as a simple and fast procedure to build the dataset was intended, it was consequently not necessary to access individual historical records, which makes for a more efficient procedure in terms of timing and computational resources.

The number and type of available variables are crucial for this kind of research. Accordingly, new variables can be extracted from the

Table 1
Explanatory available variables of the database.

Name	Description	Type
Status	Booking status: on place, cancelled	Categorical
Adults	Number of adults	Numeric
Entity	Enterprises through which customers may book rooms	Categorical
Nationality	Nationality of the guest	Categorical
Advance payment	If require advance payment or not	Categorical
Nights	Number of nights to be spent in the hotel	Numeric
Notice period	Difference between booking date and arrival date	Numeric
Day of creation	Day in which booking was created	Numeric
Month of creation	Month in which booking was created	Numeric
Day of check in	Effective check in day	Numeric
Month of check in	Effective check in month	Numeric
Mean price	Room mean price	Numeric
Sales channel	Sales channels used for the reservation. Entities are organised into 9 channels (e.g. business to business, hotel website)	Categorical
Weekend	Number Saturdays and Sundays during the stay	Numeric

original database, which is the case for the number of weekend days that is calculated considering entry and departure dates. Additionally, rows with omitted values were removed and only closed bookings were considered, as long as there was no evidence of the reservation being consumed for current services or future services. For this research, the booking status is the target variable, and so, the problem was treated as a binary classification in which targets may reach two possible values "cancelled" or "not cancelled". On the other hand, the range of each variable differs from the others, so that predictors were normalised in order to make models less sensitive to scales and maintaining consistency when comparing results across them.

3.2. Models and validation

R statistical software (R Core Team, 2013) was used for applying different supervised learning algorithms. In this regard, several packages were used: two trees decision-based algorithms, C5.0 (Kuhn et al., 2018) and random forest (Liaw and Wiener, 2002), support vector machine (SVM) (Meyer et al., 2019), artificial neural networks (ANN) (Fritsch et al., 2019) and genetic algorithm (GA) (Scrucca, 2013). In the following paragraph these methods are briefly explained.

Decision tree approaches may extract patterns from a given dataset by the formulation of rules (Mingers, 1989a) and their application in several fields have delivered promising results for information extraction, machine learning or pattern recognition (Rokach and Maimon, 2015). These techniques look for the feature which best segregates the classes and divides the data according to the values of this feature (Minz and Jain, 2003). This process is repeated with the new subsets, so that the training set is recursively split into smaller subsets in order to find partitions which contain only one class (Mingers, 1989b). Once the process is completed, the result may be represented as a tree structure in which the features are graphed as nodes joined with branches that represents its possible values. Therefore, decision trees allow translating the knowledge extracted from the training data into a graphical tool that can be easily interpreted by humans, unlike other big data techniques, such as artificial neural networks or SVM, which act as a black box. One of the most popular decision tree algorithms is the *C5.0,* an updated version of the previous *C4.5* algorithm, which runs faster and needs lower memory usage by using less rules (Pandya and Pandya, 2015). *C5.0* model applies the maximum information gained as splitting criteria, which is based on the decrease in entropy after a dataset is split by attribute (Patil et al., 2012). Once built, the decision tree has a post pruning phase, this means that those splits that do not provide relevant information are removed from the tree structure, reducing the complexity of the model as well as overfitting problems. Another popular tree decision based algorithm is the Random Forest, which is an ensemble learning method that generates multiple distinct decision trees (Oshiro et al., 2012). Using a random selection of data subsets for training each individual model the accuracy of the final combination can be improved as the heterogeneity of each one is ensured, and at the same time overfitting problems can be reduced (Pumpuang et al., 2008). On the other hand, the Support Vector Machine technique (SVM) attempts to fix a boundary surface between two classes according to the data's features (Romero Morales and Wang, 2010) using the structure risk minimisation principle. For two linearly separable classes, among the infinite number of linear classifiers, the support vector classifier technique looks for the one that outputs the smallest generalisation error by maximising the margin, which is defined as "the smallest distance between the decision boundary and any of the samples" (Bishop, 2006, pp.326). However, it is not common to find linear separable classes in the nature, so, in order to overcome this issue the data space is transformed into a potential high-dimensional space where they become linearly separable (Amari and Wu, 1999). This can be achieved using a kernel function, which is an inner product in the feature space (Campbell et al., 2006; James et al., 2013). The last probabilistic method used in this study is the Artificial Neural Networks (ANN) which is a biologically-inspired model composed of a specific number of simple computing cells, known as neurons or processing elements (PE), organised in layers and connected among them by a complex communication structure (Hyndman and Athanasopoulos, 2018). Each neuron receives the weighed output of those units connected to it and reverts with one single output after passing through a transfer function (Agatonovic-Kustrin and Beresford, 2000; Shalev-Shwartz and Ben-David, 2014). For this approach, feed-forward neural networks will be used, which are characterised for using a network structure where the connections do not contain cycles. Thus, the architecture of the network is defined by three components: number of layers, connections between nodes and activation function (Shalev-Shwartz and Ben-David, 2014). Finally, it should be noted that these kinds of techniques have been increasingly used in tourism related papers. For instance, tree decision techniques have been used by Brida et al. (2018); Chattopadhyay and Mitra (2019) and Pantano et al. (2017); the SVM technique can be found in works such as Martin-Fuentes et al. (2018); Martinez-Torres and Toral (2019) or (Zhang et al., 2020); while ANN is present in Chatterjee (2019); Shi (2019); Youn and Gu (2010).

All cited methods have been successfully applied in several areas, but there is no systematic procedure for setting architectural parameters, so they are usually adjusted by trial and error (Song and Li, 2008). For this research, tree decision models were set manually, however artificial intelligence models, which are more complex methods were set using tuning tools for SVM and genetic algorithms (GA) for ANN. The GA algorithm is an evolutionary algorithm that seeks an optimal solution for a given function. Initially, a random population is generated and during the process several reproduction, cross over and mutation operations are carried out with the aim of finding the one which best fits the function (Murali et al., 2010). The use of GA algorithm for optimising the ANN model has been used in previous publications, resulting in the combination of both methods outperforming that using conventional ANN (Arifovic and Gencay, 2001; Nasseri et al., 2008; Momeni et al., 2014). In the reviewed literature about hotel booking cancellations no previous uses of GA algorithms for setting the structural parameters have been noted, therefore we can suppose that this is a methodological innovation in this field.

With regards the method's validation, contrary to previous research in the field (Romero Morales and Wang, 2010; Antonio et al., 2017a), who use the K-fold technique, the validation of the model's performance was conducted though repeated random subsampling validation. This validation technique is specially recommended when the dependent variable is dichotomous and raw data are unbalanced (Khakifirooz et al., 2018), which is the case of this study. This technique consists in splitting whole data in a training set for building the model and the testing set for validating the model, in which the process is repeated several times and the mean value is taken as a final evaluation. Applying this method, the unreliability of one single-run training and testing is avoided (holdout method) while the number of runs is not limited, as in the K-fold method. On the other hand, some observations may not be tested, and others may be tested more than once.

The repeated random subsampling validation technique involves randomly splitting the dataset into a training set, in order to fit the model's parameters, and a testing set, to provide an unbiased evaluation. The unbalanced nature of the dataset under study, should be considered as it may produce a poor training set as the classification with a reduced number of examples is likely be ignored by the models during this phase. In order to avoid this problem, although the overall cancellation rate state around 30%, both datasets were balanced, so that each one contained 50% of each class. This procedure is repeated a hundred times using different randomly selected test and train datasets each, so that models are trained and tested once per run. During this process, the results provided by each method are saved, as well as the actual values, which we tried to predict, so that the performance could be calculated by comparing both lists.

4. Cancellation forecasting results

In this section the results of previous techniques are shown and discussed. First, performance measures of several techniques are presented and secondly, the models and results are discussed.

4.1. Results and performance measures

Results extracted from the classification methods are commonly analysed by the use of a confusion matrix, which is a contingency table that shows the difference between the actual class and the predicted one for the test set in a labelled table (Bradley, 1997). The relationship between these classes is expressed through two main performance measures, recall, also known as sensitivity, which is "the proportion of true positives correctly detected by the test", and specificity which is "the proportion of true negatives correctly identified by the test" (Altman and Bland, 1994, pp.1). On the other hand, the accuracy is expected to measure the reliability of the model for both categories, the precision summarise the true positives forecasted over the total positive items predicted by models and f-score "is a fundamental and simple method that measures the distinction between two classes with real values" and combines the precision and recall (Güneş et al., 2010). An easy way to analyse forecasted results is the ROC curve (Fig. 1), which is a graphical support for evaluating the performance classifier giving the relationship between true positives and true negatives predicted by the model (Bradley, 1997). All cited performance metrics have been calculated for each method and can be found in Table 2.

Results show that the SVM technique outputs good results, while tree-based methods improve the accuracy even more. While C5.0 algorithm outputs better results in terms of the area under the ROC curve (AUC), Random Forest shows better accuracy. Nevertheless, the SVM outperforms the Random Forest method in terms of AUC, but shows the lowest accuracy overall. For all three, it can be stated that they deliver good results. When ANN is optimised with GA, this method delivers all metrics with values above 0.95 and has the best performance overall, which can be appreciated in the ROC curve (Fig. 1). This extraordinary performance can be also seen in the ROC curve graph. On the other hand, in all cases specificity and recall are balanced, which means that the applied methods can predict both, negative and positive cases, with high levels of accuracy and therefore, the training phase is considered to be correctly carried out. When models tend to predict with high accuracy only one single class in detriment to the other, specificity or recall reach a much higher value, which is not the case for this study. As regards the precision measure, it behaves according to the rest of parameters, showing a high rate of true positive items forecast by the models over the whole positives predicted. Finally, the f-score, which is the harmonic means of precision and recall, expresses how balanced these measures are. Following the same values achieved by the other performance measures, SVM shows the lowest f-score value, followed by tree decision techniques and ANN with the highest value.

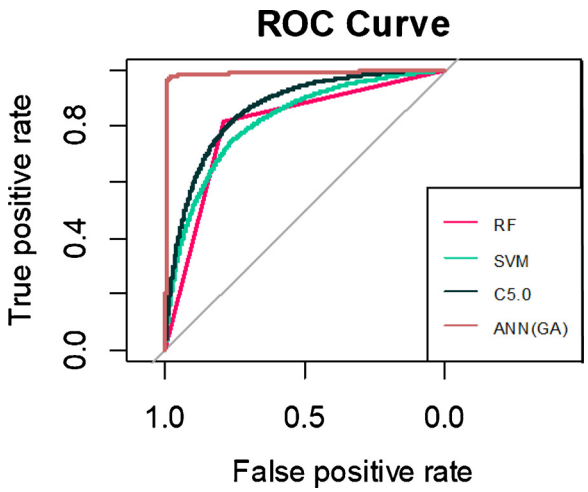

Fig. 1. ROC curves. Comparison of different methods.

Table 2
Performance measures for each method are presented.

	Accuracy	Precision	F1 score	Specificity	Recall	AUC
Random forest	0.804	0.813	0.806	0.809	0.799	0.804
Support vector machine	0.753	0.733	0.748	0.743	0.764	0.820
C 5.0	0.790	0.818	0.796	0.807	0.774	0.864
ANN (GA optimized)	0.980	0.972	0.979	0.972	0.987	0.989

5. Conclusions

This research has contributed to expanding the scarce amount of literature available on the hotel and lodging industry by proposing a methodology for forecasting hotel booking cancellations using artificial intelligence.

The main theoretical contribution of this research was the intention of shedding light on the discussion arising from the research carried out by Romero Morales and Wang (2010) and Antonio et al. (2017a) about the utility of PNR data for forecasting individual hotel cancellations, concluding that results are in line with the suggestion made by Antonio et al. (2017a) and Antonio et al. (2019a). In addition, this research shows that it is possible to forecast cancellations with a high level of accuracy using a reduced number of variables, specifically 13 independent variables, a number significantly lower than seen in previous researches such as Antonio et al. (2017a) and Antonio et al. (2019a) which use at least 37. This fact reinforces the idea that the use of a high number of variables does not necessarily imply better performance, which is aligned with the conclusions of Antonio et al. (2019a) who conclude that no significant improvement is achieved by adding more variables in the model. Therefore, the proposed methodology makes for a simple and efficient model which can forecast cancellations with a lower number of inputs than previous research in this area (Antonio et al., 2017a, a). On the other hand, some techniques proposed by Antonio et al. (2017a) and Antonio et al. (2019a), such as tree-decision based algorithms or SVM, have been tested for available datasets, resulting in a lower performance. This might be down to two possible reasons firstly, the use of a smaller number of variables and secondly databases are different. Nevertheless, with 80% accuracy, the results of this method are good for providing an initial approximation, as long as it requires less timing and computing resources. However, another new slant to this research is the use of genetic algorithms for setting ANN, which allows obtaining 98% accuracy even when using only 13 independent variables. On the other hand, determining the architecture of the artificial neural network is a complex problem and there is no systematic procedure to carry out this task, forcing it to be adjusted by trial and error (Song and Li, 2008). Thus, the use of genetic algorithms for setting the structural parameters of the network technique allows finding an optimal solution, and reduces the risk of finding another local maxima (Jiang et al., 2003). Using genetic algorithms for this purpose can be considered a methodological innovation in this field as no previous uses were found during the literature review regarding cancellation forecasting.

Secondly, the main practical contribution is the proposed methodology that allows forecasting hotel cancellations with a very high level of accuracy using variables that hotels have easy access to. Moreover, one of the key aspects of this procedure is the ease for exporting the methodology to similar businesses, as it has been developed using real booking data, and most importantly, with the most common variables used on multiple booking platforms, used by the most popular online

travel agencies such as Trivago, Tripadvisor or Booking, to large hotel chains like Radisson Hotels & Resorts, Riu Hotels & Resorts or Meliá Hotel Resorts among others. This facilitates the creation of a database, which is aimed at identifying those customers likely to cancel as well as the training sample when it is necessary to update the model. In fact, it was not necessary to extract complementary variables, as previous research does, by querying historical records of the database and nor was it necessary to consider specific information obtained using the guest's identity for that matter (Antonio et al., 2017a). It allows implementing proposed methodology even in those companies with a relative short booking history. In the same way, external data sources were not employed as other authors propose (e.g Antonio et al., 2019a), thereby reducing the complexity of the present procedure. Therefore, this methodology allows to train the models and provide the forecasts using only the information provided during the booking process, which makes it simpler and faster to deliver the forecasts and so, forecasts can be aligned closer to the latest market trends. In fact, one of the main assets of this proposal is that it allows forecasting individual cancellations with high accuracy, which has become more important in recent years because of the increasing use of online bookings. This trend has led customers to placing multiple reservations when planning their holidays for them in the end to choose only one and cancel the rest. This situation becomes even worse when the hotels themselves or competing companies, provide "last minute" offers, because customers have the chance to cancel previous reservations and choose a more economic option.

From a managerial point of view, the results achieved indicate that a customer's historical records are essential for hospitality enterprises and should be treated as a key asset. Along these lines, the treatment of these data through ANN optimised with GA generates considerable value to the organisation, because of the difficulties and revenue loss that cancellations generate. On the one hand, accurate cancellation forecasting leads hoteliers to take proper managerial decisions and provides organisational advantages for the industry. These techniques give management the opportunity to have information in advance, so that they can establish appropriate overbooking policies, cancellation policies and take advantage of proper pricing strategies among others. With regards overbooking policies specifically, if hoteliers have reliable information about the cancellations, they may avoid overbooking, meaning it would not be necessary to relocate guests, which causes revenue loss and has a negative impact on reputation. On the other hand, this methodology represents a considerable competitive advantage because it can forecast the cancellation rate with a level high of accuracy, but it can also determine which customer is likely to cancel. This would allow the hoteliers to take proactive actions in order to encourage clients to maintain their reservation, such as sending reminders or contacting directly with them. For instance, depending on how profitable the guest is, "special gifts" may be offered, such as a free dinner or free access to some additional services. In the same manner, individual cancellation policies could be applied to the customers when they place the reservation, for example, not allowing free cancellations to those clients likely to cancel.

Future research could consider testing the same methodology proposed in this paper on other PNR databases with different characteristics (location, weather conditions, hotel rating, prices, market segment, distribution channels, cancellation policies, etc.). In this way additional variables may be included in the study by easily modifying the reservation form, such as booking purpose (e.g. business or pleasure), as well as others to be added from external sources, such as weather forecasts or economic index of the origin countries (recession index, GDP, etc.) may be interesting to be used as predictor variables in order to improve the results.

With regards the limitations of the proposed methodology, while forecasts are based on past data, abrupt changes in the market could not be initially reflected in the model.

References

Agatonovic-Kustrin, S., Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. 22 (5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1.

Altman, D.G., Bland, J.M., 1994. Diagnostic tests 1: sensitivity and specificity. BMJ 308 (6943), 1.

Amari, S., Wu, S., 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Netw. 12 (6), 783–789. https://doi.org/10.1016/S0893-6080(99)00032-5.

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

Antonio, N., de Almeida, A., Nunes, L., 2017b. Predicting hotel bookings cancellation with a machine learning classification model. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) 1049–1054. https://doi.org/10.1109/ICMLA.2017.00-11.

Antonio, N., de Almeida, A., Nunes, L., 2019a. Big data in hotel revenue management: exploring cancellation drivers to gain insights into booking cancellation behavior. Cornell Hosp. Q. https://doi.org/10.1177/1938965519851466.

Antonio, N., de Almeida, A., Nunes, L., 2019b. Predictive models for hotel booking cancellation: a semi-automated analysis of the literature. Tour. Manag. Stud. 16.

Arifovic, J., Gencay, R., 2001. Using genetic algorithms to select architecture of a feed-forward artiÿcial neural network. Physica A 21.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30 (7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Brida, J.G., Lanzilotta, B., Moreno, L., Santiñaque, F., 2018. A non-linear approximation to the distribution of total expenditure distribution of cruise tourists in Uruguay. Tour. Manag. 69, 62–68. https://doi.org/10.1016/j.tourman.2018.05.006.

Campbell, W.M., Sturim, D.E., Reynolds, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. 13 (5), 308–311. https://doi.org/10.1109/LSP.2006.870086.

Chatterjee, S., 2019. Drivers of helpfulness of online hotel reviews: a sentiment and emotion mining approach. Int. J. Hosp. Manag. 102356. https://doi.org/10.1016/j.ijhm.2019.102356.

Chattopadhyay, M., Mitra, S.K., 2019. Do airbnb host listing attributes influence room pricing homogenously? Int. J. Hosp. Manag. 81, 54–64. https://doi.org/10.1016/j.ijhm.2019.03.008.

Chen, C.C., Schwartz, Z., Vargas, P., 2011. The search for the best deal: how hotel cancellation policies affect the search and booking decisions of deal-seeking customers. Int. J. Hosp. Manag. 30 (1), 129–135. https://doi.org/10.1016/j.ijhm.2010.03.010.

Cho, V., 2003. A comparison of three different approaches to tourist arrival forecasting. Tour. Manag. 24, 323–330. https://doi.org/10.1016/S0261-5177(02)00068-7.

Chow, W.S., Shyu, J.-C., Wang, K.-C., 1998. Developing a forecast system for hotel occupancy rate using integrated ARIMA models. J. Int. Hosp. Leis. Tour. Manag. 1 (3), 55–80. https://doi.org/10.1300/J268v01n03_05.

Chu, F.L., 2009. Forecasting tourism demand with ARMA-based methods. Tour. Manag. 30 (5), 740–751. https://doi.org/10.1016/j.tourman.2008.10.016.

Claveria, O., Datzira, J., 2010. Forecasting tourism demand using consumer expectations. Tour. Rev. 65 (1), 18–36. https://doi.org/10.1108/16605371011040889.

Claveria, O., Monte, E., Torra, S., 2015. Tourism demand forecasting with neural network models: different ways of treating information: tourism demand forecasting with neural network models. Int. J. Tour. Res. 17 (5), 492–500. https://doi.org/10.1002/jtr.2016.

Falk, M., Vieru, M., 2018. Modelling the cancellation behaviour of hotel guests. Int. J. Contemp. Hosp. Manage. 30 (10), 3100–3116. https://doi.org/10.1108/IJCHM-08-2017-0509.

Frechtling, D.C., 2001. Forecasting Tourism Demand: Methods and Strategies. Butterworth-Heinemann, Oxford; Boston.

Fritsch, S., Guenther, F., Wright, M.N., Suling, M., Mueller, S.M., 2019. neuralnet: Training of Neural Networks (Version 1.44.2). Retrieved from https://CRAN.R-project.org/package=neuralnet.

Gehrels, S., Blanar, O., 2013. How economic crisis affects revenue management: the case of the Prague Hilton hotels. Res. Hosp. Manag. 2 (1–2), 9–15. https://doi.org/10.1080/22243534.2013.11828284.

Gorin, T., Brunger, W.G., White, M.M., 2006. No-show forecasting: a blended cost-based, PNR-adjusted approach. J. Revenue Pricing Manag. 5 (3), 188–206. https://doi.org/10.1057/palgrave.rpm.5160039.

Güneş, S., Polat, K., Yosunkaya, Ş., 2010. Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome. Expert Syst. Appl. 37 (2), 998–1004. https://doi.org/10.1016/j.eswa.2009.05.075.

Gunter, U., Önder, I., 2015. Forecasting international city tourism demand for Paris: accuracy of uni- and multivariate models employing monthly data. Tour. Manag. 46, 123–135. https://doi.org/10.1016/j.tourman.2014.06.017.

Haensel, A., Koole, G., 2011. Booking horizon forecasting with dynamic updating: a case study of hotel reservation data. Int. J. Forecast. 27 (3), 942–960. https://doi.org/10.1016/j.ijforecast.2010.10.004.

Hajibaba, H., Boztuğ, Y., Dolnicar, S., 2016. Preventing tourists from canceling in times of crises. Ann. Tour. Res. 60, 48–62. https://doi.org/10.1016/j.annals.2016.06.003.

Hassani, H., Silva, E.S., Antonakakis, N., Filis, G., Gupta, R., 2017. Forecasting accuracy evaluation of tourist arrivals. Ann. Tour. Res. 63, 112–127. https://doi.org/10.1016/j.annals.2017.01.008.

Heo, C.Y., Lee, S., 2009. Application of revenue management practices to the theme park

industry. Int. J. Hosp. Manag. 28 (3), 446–453. https://doi.org/10.1016/j.ijhm.2009.02.001.
Hu, Y.-C., Jiang, P., Lee, P.-C., 2019. Forecasting tourism demand by incorporating neural networks into Grey–Markov models. J. Oper. Res. Soc. 70 (1), 12–20. https://doi.org/10.1080/01605682.2017.1418150.
Huang, H.-C., 2014. A study on artificial intelligence forecasting of resort demand. J. Theor. Appl. Inf. Technol. 70, 265–272.
Huang, H.-C., I Hou, C., 2017. Tourism demand forecasting model using neural network. Int. J. Comput. Sci. Inf. Technol. 9 (2), 19–29. https://doi.org/10.5121/ijcsit.2017.9202.
Huang, H., Chang, A.Y., Ho, C., 2013. Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model (1). pp. 178–180.
Huarng, K.-H., Moutinho, L., Yu, T.H.-K., 2007. An advanced approach to forecasting tourism demand in Taiwan. J. Travel Tour. Mark. 21 (4), 15–24. https://doi.org/10.1300/J073v21n04_03.
Hwang, Jeng-Ren, Chen, Shyi-Ming, Chia-Hoang, Lee, 1998. Handling forecasting problems using fuzzy time series. Fuzzy Sets Syst. 100 (1–3), 217–228. https://doi.org/10.1016/S0165-0114(97)00121-8.
Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: Principles and Practice, 2nd ed. Otexts.
James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), 2013. An Introduction to Statistical Learning: With Applications in R. Springer, New York.
Jiang, N., Zhao, Z., Ren, L., 2003. Design of structural modular neural networks with genetic algorithm. Adv. Eng. Softw. 34 (1), 17–24. https://doi.org/10.1016/S0965-9978(02)00107-2.
Kaynak, E., Cavlek, N., 2007. Measurement of tourism market potential of Croatia by use of delphi qualitative research technique. J. East-West Bus. 12 (4), 105–123. https://doi.org/10.1300/J097v12n04_05.
Khakifirooz, M., Chien, C.F., Chen, Y.-J., 2018. Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0. Appl. Soft Comput. 68, 990–999. https://doi.org/10.1016/j.asoc.2017.11.034.
Kouprioushina, L., van der Rest, J.-P., Schwartz, Z., 2014. On revenue management and the use of occupancy forecasting error measures. Int. J. Hosp. Manag. 41, 104–114. https://doi.org/10.1016/j.ijhm.2014.05.002.
Kourentzes, N., Rostami-Tabar, B., Barrow, D.K., 2017. Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? J. Bus. Res. 78 (October 2016), 1–9. https://doi.org/10.1016/j.jbusres.2017.04.016.
Kuhn, M., Weston, S., Culp, M., Coulter, N., Quinlan, R., 2018. C50: C5.0 Decision Trees and Rule-Based Models (Version 0.1.2). Retrieved from https://CRAN.R-project.org/package=C50.
Lee, M., 2018. Modeling and forecasting hotel room demand based on advance booking information. Tour. Manag. 66, 62–71. https://doi.org/10.1016/j.tourman.2017.11.004.
Lee, C.K., Song, H.J., Mjelde, J.W., 2008. The forecasting of International Expo tourism using quantitative and qualitative techniques. Tour. Manag. 29 (6), 1084–1098. https://doi.org/10.1016/j.tourman.2008.02.007.
Li, S., Chen, T., Wang, L., Ming, C., 2018. Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. Tour. Manag. 68, 116–126. https://doi.org/10.1016/j.tourman.2018.03.006.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22. [R News]. Retrieved from https://CRAN.R-project.org/doc/Rnews/.
Lin, V.S., Song, H., 2015. A review of Delphi forecasting research in tourism. Curr. Issues Tour. 18 (12), 1099–1131. https://doi.org/10.1080/13683500.2014.967187.
MacCarthy, B.L., Blome, C., Olhager, J., Srai, J.S., Zhao, X., 2016. Article information. Int. J. Oper. Prod. Manage. 36 (12), 1696–1718. https://doi.org/10.1108/02656710210415703.
Martinez-Torres, M.R., Toral, S.L., 2019. A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. Tour. Manag. 75, 393–403. https://doi.org/10.1016/j.tourman.2019.06.003.
Martin-Fuentes, E., Fernandez, C., Mateu, C., Marine-Roig, E., 2018. Modelling a grading scheme for peer-to-peer accommodation: stars for Airbnb. Int. J. Hosp. Manag. 69, 75–83. https://doi.org/10.1016/j.ijhm.2017.10.016.
Mayr, T., Zins, A.H., 2009. Acceptance of online vs. traditional travel agencies. Anatolia 20 (1), 164–177. https://doi.org/10.1080/13032917.2009.10518902.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chih-Chung, C., Chih-Chen, L., 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (Version 1.7-2). Retrieved from https://CRAN.R-project.org/package=e1071.
Mingers, J., 1989a. An empirical comparison of pruning methods for decision tree induction. Mach. Learn. 4, 17.
Mingers, J., 1989b. An empirical comparison of selection measures for decision-tree induction. Mach. Learn. 3 (4), 319–342. https://doi.org/10.1007/BF00116837.
Minz, S., Jain, R., 2003. Rough set based decision tree model for classification. Data Warehousing Knowledge Discov. 2737, 172–181. https://doi.org/10.1007/978-3-540-45228-7_18.
Momeni, E., Nazir, R., Jahed Armaghani, D., Maizir, H., 2014. Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN. Measurement 57, 122–131. https://doi.org/10.1016/j.measurement.2014.08.007.
Moutinho, L., Witt, S.F., 1995. Forecasting the tourism environment using a consensus approach. J. Travel. Res. 33 (4), 46–50. https://doi.org/10.1177/004728759503300407.
Moutinho, L., Huarng, K.-H., Yu, T.H.-K., Chen, C.-Y., 2008. Modeling and forecasting tourism demand: the case of flows from Mainland China to Taiwan. Serv. Bus. 2 (3), 219–232. https://doi.org/10.1007/s11628-008-0037-3.
Murali, R.V., Puri, A.B., Prabhakaran, G., 2010. GA-Driven ANN Model for Worker Assignment Into Virtual Manufacturing Cells. pp. 5.
Nasseri, M., Asghari, K., Abedini, M.J., 2008. Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. Expert Syst. Appl. 35 (3), 1415–1421. https://doi.org/10.1016/j.eswa.2007.08.033.
Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest? Machine Learn. Data Min. Pattern Recogn. 7376, 154–168. https://doi.org/10.1007/978-3-642-31537-4_13.
Ostaijen, V., Bruno, F., Van Ostaijen, T., Santos, B.F., 2017. Delft university of technology dynamic airline booking forecasting. Proceedings of the 21st Air Transport Research Society World Conference.
Pan, B., Yang, Y., 2017. Forecasting destination weekly hotel occupancy with big data. J. Travel. Res. 56 (7), 957–970. https://doi.org/10.1177/0047287516669050.
Pandya, R., Pandya, J., 2015. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. Int. J. Comput. Appl. 117 (16), 18–21. https://doi.org/10.5120/20639-3318.
Pantano, E., Priporas, C.-V., Stylos, N., 2017. 'You will like it!' using open data to predict tourists' response to a tourist attraction. Tour. Manag. 60, 430–438. https://doi.org/10.1016/j.tourman.2016.12.020.
Park, Y.A., Gretzel, U., Sirakaya-Turk, E., 2007. Measuring web site quality for online travel agencies. J. Travel Tour. Mark. 23 (1), 15–30. https://doi.org/10.1300/J073v23n01_02.
Park, S., Lee, J., Song, W., 2017. Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. J. Travel Tour. Mark. 34 (3), 357–368. https://doi.org/10.1080/10548408.2016.1170651.
Patil, N., Lathi, R., Chitre, V., 2012. Comparison of C5.0 & CART Classification algorithms using pruning technique. Int. J. Eng. Res. 1 (4), 6.
Peng, B., Song, H., Crouch, G.I., 2014. A meta-analysis of international tourism demand forecasting and implications for practice. Tour. Manag. 45, 181–193. https://doi.org/10.1016/j.tourman.2014.04.005.
Pereira, L.N., 2016. An introduction to helpful forecasting methods for hotel revenue management. Int. J. Hosp. Manag. 58, 13–23. https://doi.org/10.1016/j.ijhm.2016.07.003.
Pfeifer, P.E., Bodily, S.E., 1990. A test of space-time arma modelling and forecasting of hotel data. J. Forecast. 9 (3), 255–272. https://doi.org/10.1002/for.3980090305.
Pumpuang, P., Srivihok, A., Praneetpolgrang, P., 2008. Comparisons of classifier algorithms: Bayesian network, C4.5, decision forest and NBTree for course registration planning model of undergraduate students. 2008 IEEE International Conference on Systems, Man and Cybernetics 3647–3651. https://doi.org/10.1109/ICSMC.2008.4811865.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Retrieved from; Available online: http://www.R-project.org/.
Rajopadhye, M., Ghalia, M.B., Wang, P.P., Baker, T., Eister, C.V., 2001. Forecasting uncertain hotel room demand. Inf. Sci. 11.
Rokach, L., Maimon, O., 2015. Data Mining With Decision Trees: Theory and Applications, second edition. World Scientific, Hackensack, New Jersey.
Romero Morales, D., Wang, J., 2010. Forecasting cancellation rates for services booking revenue management using data mining. Eur. J. Oper. Res. 202 (2), 554–562. https://doi.org/10.1016/j.ejor.2009.06.006.
Schwartz, Z., Cohen, E., 2004. Subjective estimates of occupancy forecast uncertainty by hotel revenue managers. J. Travel Tour. Mark. 16 (4), 59–66. https://doi.org/10.1300/J073v16n04_08.
Scrucca, L., 2013. GA: A Package for Genetic Algorithms in R (Version). J. Stat. Softw. 53 (4), 1–37. Retrieved from http://www.jstatsoft.org/v53/i04/.
Shalev-Shwartz, S., Ben-David, S., 2014. Understanding Machine Learning: From Theory to Algorithms. https://doi.org/10.1017/CBO9781107298019.
Shi, X., 2019. Tourism culture and demand forecasting based on BP neural network mining algorithms. Pers. Ubiquitous Comput. https://doi.org/10.1007/s00779-019-01325-x.
Sierag, D.D., Koole, G.M., van der Mei, R.D., van der Rest, J.I., Zwart, B., 2015. Revenue management under customer choice behaviour with cancellations and overbooking. Eur. J. Oper. Res. 246 (1), 170–185. https://doi.org/10.1016/j.ejor.2015.04.014.
Song, H., Li, G., 2008. Tourism demand modelling and forecasting-A review of recent research. Tour. Manag. 29 (2), 203–220. https://doi.org/10.1016/j.tourman.2007.07.016.
Song, H., Li, G., Witt, S.F., Athanasopoulos, G., 2011. Forecasting tourist arrivals using time-varying parameter structural time series models. Int. J. Forecast. 27 (3), 855–869. https://doi.org/10.1016/j.ijforecast.2010.06.001.
Song, H., Qiu, R.T.R., Park, J., 2019. A review of research on tourism demand forecasting: launching the annals of tourism research curated collection on tourism demand forecasting. Ann. Tour. Res. 75, 338–362. https://doi.org/10.1016/j.annals.2018.12.001.
Tang, C.M.F., King, B., Pratt, S., 2016. Predicting hotel occupancies with public data. Tour. Econ. 23 (5), 135481661666667. https://doi.org/10.1177/1354816616666670.
Teixeira, J.P., Fernandes, P.O., 2012. Tourism time series forecast -different ANN architectures with time index input. Procedia Technol. 5, 445–454. https://doi.org/10.1016/j.protcy.2012.09.049.
Tideswell, C., Mules, T., Faulkner, B., 2001. An integrative approach to tourism forecasting: a glance in the rearview mirror. J. Travel. Res. 40 (2), 162–171. https://doi.org/10.1177/004728750104000207.
Tse, T.S.M., Poon, Y.T., 2015. Analyzing the use of an advance booking curve in forecasting hotel reservations. J. Travel Tour. Mark. 32 (7), 852–869. https://doi.org/10.1080/10548408.2015.1063826.
Uysal, M., Crompton, J.L., 1985. An overview of approaches used to forecast tourism

demand. J. Travel. Res. 23 (4), 7–15. https://doi.org/10.1177/004728758502300402.
Weatherford, L.R., Kimes, S.E., 2003. A comparison of forecasting methods for hotel revenue management. Int. J. Forecast. 19 (3), 401–415. https://doi.org/10.1016/S0169-2070(02)00011-0.
Wirth, R., Hipp, J., 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. pp. 11.
Witt, S.F., Witt, C.A., 1995. Forecasting tourism demand: a review of empirical research. Int. J. Forecast. 11 (3), 447–475. https://doi.org/10.1016/0169-2070(95)00591-7.
World Tourism Organization (Ed.), 2018. World Tourism Organization. https://doi.org/10.18111/9789284419876.
Wu, D.C., Song, H., Shen, S., 2017. New developments in tourism and hotel demand modeling and forecasting. Int. J. Contemp. Hosp. Manage. 29 (1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249.
Youn, H., Gu, Z., 2010. Predicting Korean lodging firm failures: an artificial neural network model along with a logistic regression model. Int. J. Hosp. Manag. 29 (1), 120–127. https://doi.org/10.1016/j.ijhm.2009.06.007.
Yu, G., Schwartz, Z., 2006. Forecasting short time-series tourism demand with artificial intelligence models. J. Travel. Res. 45 (2), 194–203. https://doi.org/10.1177/0047287506291594.
Yüksel, S., 2005. An integrated forecasting approach for hotels. The International Symposium on the Analytic Hierarchy Process (ISAHP) 2005 10.
Yüksel, S., 2007. An integrated forecasting approach to hotel demand. Math. Comput. Model. 46 (7–8), 1063–1070. https://doi.org/10.1016/j.mcm.2007.03.008.
Zakhary, A., Atiya, A.F., El-Shishiny, H., Gayar, N.E., 2011. Forecasting hotel arrivals and occupancy using Monte Carlo simulation. J. Revenue Pricing Manage. 10 (4), 344–366. https://doi.org/10.1057/rpm.2009.42.
Zhang, X., Qiao, S., Yang, Y., Zhang, Z., 2020. Exploring the impact of personalized management responses on tourists' satisfaction: a topic matching perspective. Tour. Manag. 76, 103953. https://doi.org/10.1016/j.tourman.2019.103953.