Tourism Management Perspectives 35 (2020) 100718

Contents lists available at ScienceDirect

Tourism Management Perspectives

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

Identifying critical hotel cancellations using artificial intelligence

Eleazar C. Sánchez[a], Agustín J. Sánchez-Medina[a,*], Mónica Pellejero[b]

[a] 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
[b] Mid Atlantic University, Carretera de Quilmes, 37, 35017 Tafira Baja, Las Palmas de Gran Canaria, Spain

ARTICLE INFO

Keywords:
Forecasting models
Hotel
Artificial intelligence
Cancellations
Revenue management

ABSTRACT

Cancellations have a significant impact on the hotel and lodging industry because they directly affect income and are thus considered critical in revenue management. Specifically, cancellations made close to the time of service are the most damaging for hotels because they leave management with no time to react. The use of Personal Name Records (PNR) has led to new approaches in this field, however despite this novel research area there are no investigations focusing on forecasting for individual hotel cancellations made close to the time of service. With the aim of filling this gap, this research is intended to identify those individuals likely to make cancellations in a short-horizon of time using Artificial Intelligence (AI) techniques through PNR data. Promising results have been achieved with 80% accuracy for cancellations made 7 days in advance. By taking this approach, booking management systems, as well as cancellation policies may be optimised.

1. Introduction

One of the most significant issues that hotel managers face is trying to match the hotel's capacity with demand. In the hospitality industry the product cannot be stockpiled, which forces hoteliers to deal with demand and the limited number of rooms in a specific time frame in such a way that each unoccupied room does not result in a loss of revenue {Citation}. This leads to hoteliers attempting to increase their revenue by maximising occupancy, which involves dealing with future demand. However, demand is subject to several external factors, such as weather, political stability, high competitiveness and others, which make it difficult to forecast. Reservations already placed may be cancelled, adding even more complexity to the planning and organising process of the hotel's capacity. Indeed, the importance that cancellations have within the hotel and lodging industry is not just reflected in terms of inventory management, but also in pricing strategies (Chen & Xie, 2013). Cancellations may, at times, suppose a revenue loss accounting for about 20% of income (Sierag, Koole, van der Mei, van der Rest, & Zwart, 2015). The huge importance that cancellations have for the lodging industry has led some authors to talk about the analysis of "net-demand", instead of just demand (Rajopadhye, Ghalia, Wang, Baker, & Eister, 2001). Bearing in mind that hotels work with reservations for future services, which may be cancelled, requested bookings do not reflect the real number of services to be provided by the hotel, therefore "net-demand" represents the number of reservations requested minus the number of cancellations (Antonio, de Almeida, & Nunes, 2017a, 2019b; Romero Morales & Wang, 2010). This approach allows segregating the problem of demand, so that instead of treating it as a whole, apparent demand and cancellations can be studied separately. Therefore, research can focus on one part of the problem, in this case, the analysis of cancellations, which is a critical aspect for the hotel and lodging industry.

The impact that cancellations have on hotel chains has resulted in the development of strategies designed specifically with the goal of reducing cancellations. One such strategy is overbooking, which consists in accepting more bookings than the hotel has capacity for, relying on the fact that some cancellations will occur. The main aim of this strategy is to match the number or clients who do not appear at the time of service (no-show), last minute cancellations, as well as other cancellations notified in advance (Ivanov & Zhechev, 2011), so that hotels can avoid having idle capacity as far as possible. If the number of arrivals is below the forecast, hotels lose income because of unsold rooms. Similarly when the capacity of a hotel is not large enough to assume the demand, it incurs a loss of revenue because of having to relocate guests, which can also affect their reputation and corporate image. Other strategies, such as cancellation policies attempt to encourage customers to cancel in advance (Chen & Xie, 2013) through the imposition of restrictions within a specific period before the time of service (Law & Wong, 2010). This is common practice in the hospitality industry (Chen et al., 2011) and helps to reduce the number of no-shows and last

* Corresponding author.
E-mail address: agustin.sanchez@ulpgc.es (A.J. Sánchez-Medina).

https://doi.org/10.1016/j.tmp.2020.100718
Received 15 November 2019; Received in revised form 14 June 2020; Accepted 24 June 2020

minute cancellations (Zakhary, Atiya, El-Shishiny, & Gayar, 2011). On the other hand, cancellation policies can have a negative impact on the hotel's corporate social reputation, and at the same time may create a discouraging effect on clients (Antonio, de Almeida, & Nunes, 2017b). For these reasons, having information in advance about cancellations is crucial for hotel management and different approaches have been developed with this in mind (Antonio et al., 2017b; Antonio, de Almeida, & Nunes, 2019a; Falk & Vieru, 2018; Romero Morales & Wang, 2010). In fact, forecasting plays a crucial role in the operations of modern organisations because of supporting "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" (Kourentzes, Rostami-Tabar, & Barrow, 2017, pp1). Accurate cancellation forecasts may support management in the decision-making process, as well as in the design of optimal strategies to reduce the impact that cancellations have on income. Nevertheless, the literature which addresses hotel cancellations is underdeveloped (Antonio et al., 2019b; Zakhary et al., 2011) and little is known about how to prevent them (Hajibaba, Boztuğ, & Dolnicar, 2016). Moreover, literature about short-term hotel demand forecasting has less presence than other revenue management related research (Pereira, 2016).

The period of notice for a cancellation must be considered a critical aspect, as cancellations placed close to the time of service may produce a particularly high loss in revenue (Chen et al., 2011; Koide & Ishii, 2005). These cases leave hotel management with little margin to react and may result in unsold inventory or substantial discounts having to be made on price (Antonio et al., 2017a). This has become even more critical in recent years because of the increasing use of online travel agencies for hotel room bookings, which has led to customers making several reservations before finally choosing one and cancelling the rest (Antonio et al., 2017a; Chen et al., 2011). Moreover, e-commerce allows customers to easily compare different offers and even read about the experience of previous clients, thus adding to the risk of cancellation (Koide & Ishii, 2005). Likewise, the growth of last-minute bookings encourages customers to take advantage of more economical offers to the detriment of previous reservations that are cancelled, with this having a direct impact on cancellations.

Previous research has managed to forecast which individuals are likely to cancel with a very high level of accuracy, above 90% (Antonio et al., 2017a; C-Sánchez, 2019), however there is no research addressing the forecasting of individual hotel cancellations made close to the time of service. In order to fill this gap and, considering that this kind of cancellation generates high revenue loss (Chen et al., 2011; Koide & Ishii, 2005), the aim of this research is to forecast individual hotel cancellations made close to the time of service through Artificial Intelligence (AI) techniques based on real Personal Name Records (PNR) provided from a four-star hotel chain located in one of the most touristic places in Spain, the island of Gran Canaria. For this purpose, the most commonly requested data by online booking agencies when customers place a reservation were used. Accordingly, several existing online booking agencies were consulted, such as Booking, Tripadvisor and Trivago, together with large hotel chains like Radisson Hotels & Resorts, Riu Hotels & Resorts or Meliá Hotel Resorts, among others. The findings confirm that these companies use, at least, the variables selected for this research (Table 1). Thus, by using the most common variables we avoid using others that hotels might not have access to or might be too complex to use. Client identification was not available in the provided database and therefore, it cannot be used for the model development. However, it can be seen as an advantage of the present procedure as it makes for a more efficient data treatment process. The use of this variable makes the performance of a specific search into the whole database per customer obligatory. This, of course, would require more time and computational resources. For the same reasons, the use of external data sources, such as weather data sources or economical indexes outside the organisation have not been used either. Along these lines, it is worth noting that choosing a database is not an easy task and looking for a convenient method to integrate different heterogeneous data structures may complicate this stage (Antonio et al., 2019b; Haller, Pröll, Retschitzegger, Tjoa, & Wagner, 2000). Therefore, the methodology presented in this paper allows us to simplify the forecasting process in terms of building the dataset and the data itself, which represents one of the main advantages of this approach. It also leads to more frequent training so that hoteliers are better able to follow the market trend, which of course is especially important for forecasting cancellations near to the time of service.

Table 1
Explanatory variables used for developing forecasting model.

Name	Description	Type
Status	Booking status: active, cancelled	Categorial
Adults	Number of adults	Numeric
Entity	Entity through which booking was made	Categorical
Nationality	Nationality of the guest	Categorical
Advance payment	If require advance payment (1) or not (0)	Categorial
Nights	Number of nights to be spent at 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
Average price	Average room price	Numeric
Channel	Channel used for booking classified	Categorial
Weekend	Number of Saturdays and Sundays during the stay	Numeric

The main purpose of this research is to develop a model for addressing cancellations made in a short window of time using AI techniques. In that sense, the proposed model attempts to detect individuals likely to cancel close to the time of service, according to their characteristics and the historical reasons they have for changing their minds. The validation of this model has been carried out using real PNR data provided by a hotel located in Gran Canaria (Spain), thus, the configuration of the model has been carried out addressing the particularities of this hotel. In this regard, this approach attempts to forecast individual cancellations likely to be made very close to the entry day, from 4 day to 7 days in advance, and which can be considered "critical cancellations" in that they leave management with no time to react.

2. Literature review

This section is composed of two subsections. In the first, the relationship between hotel revenue management and forecasting is explained, while in the second previous forecasting approaches are presented for both hotels and airlines in which PNR data are used.

2.1. Revenue management and forecasting

One of the main aspects for maximising hotel revenue management lies in the efficiency of the organisation and planning procedures for the available rooms, not an easy task because of the uncertain environment that the sector is exposed to. Economic crisis, inflation, environmental changes, wars, regulatory changes, new client demand or technological changes are factors that management must take into consideration in this industry, among others (Yüksel, 2005).

Moreover, an unexpected reduction in demand often generates a crisis in this sector because of the high sensitivity to fluctuations in demand (Yüksel, 2007). For this reason, it is essential to understand the environment in which hotels operate, as well as develop a strategy for future room allocations (Mubiru, 2014). In this context, performing accurate forecasts is necessary for optimising operations, as well as supporting the decision-making process. As a general overview,

accurate forecasts help managers with medium- and long-term decisions not only for determining hotel polices, human resources required according to workloads or budget planning, but also for assisting in the development of short-term occupancy schedules (Gunter & Önder, 2015; Hassani, Silva, Antonakakis, Filis, & Gupta, 2017). Bearing in mind that forecasting models are based on historical data (Uysal & Crompton, 1985), a number of investigations have encouraged management to consider the importance of having a reliable revenue management system through which past data can provide value to the organisation (Zhang et al., 2017). However, as mentioned earlier, cancellation models in the hospitality industry are a huge challenge because of the highly volatile and uncertain environment of this sector (Yüksel, 2007). Naturally, unexpected changes (e.g. natural disasters, pandemics or coups among others) can be easily identified as the main reason for an increase in a hotel's number of cancellations. However, trying to identify the cancellation drivers under normal circumstances is more difficult, as they can have very widespread reasons that are unknown to hoteliers and thus, little is known about how to prevent them (Hajibaba et al., 2016). This situation becomes even more dramatic for short-horizon cancellations, because the guest's motivations are constantly changing as the service time approaches (Romero Morales & Wang, 2010). With this in mind, the present model has been built with the intention of identifying short-time horizon cancellations under normal demand conditions, so if customer behaviour is affected by exceptional situations models must be retrained.

Originally for revenue management, forecasts were based on seasonal data (e.g. the month, week or weather), mostly because it was the only information available (Romero Morales & Wang, 2010). However, more recently, Personal Name Records (PNR) have been used for this (Tang, King, & Pratt, 2017). They contain a wide range of information about the customers, which is collated at the time a reservation is made, such as preferences, number of customers, nationality and other personal details. In the context of hotel cancellations, this approach allows to know more about each customer and forecast not just anonymous cancellation rates, but determine which individuals are likely to cancel (Antonio et al., 2019a, 2019b). This kind of information is very valuable for hotel chains that are developing and investing in intelligent systems to provide accurate forecasts (Zhang et al., 2017) and this has prompted the use of “more mathematically sophisticated optimisation engines” (Weatherford, 2016, pp1). Indeed, while traditional methods such as explorative methods (e.g. time series analysis) have been widely used to forecast within the industry, in recent years the amount of research using AI-based models has increased because of the excellent forecasting capacity they present (Song, Qiu, & Park, 2019), often achieving better results than traditional models (Wu, Song, & Shen, 2017). In this regard, several investigations carried out within the industry have concluded that AI-based methods outperform traditional ones (Burger, Dohnal, Kathrada, & Law, 2001; Chen & Wang, 2007; Cho, 2003; Law, 2000; Li, Chen, Wang, & Ming, 2018).

2.2. PRN-based forecasts in the tourist industry and ensemble methods

The use of PNR data for forecasting within the tourism industry is relatively new (Gorin, Brunger, & White, 2006) and research in this area concludes that when using this information a more accurate model can be built. In this section, PRN-based forecasting approaches addressing cancellation and no-shows within the airline and lodging industry are reviewed.

For the airline industry, most published research papers address the no-show problem (Antonio et al., 2019b). In this regard, Garrow and Koppelman (2004) applied a multinomial logit model using disaggregate PNR data in order to forecast no-shows, early standbys and large standbys, concluding that using PNR data and itinerary information is possible to build more accurate models. Later, Gorin et al. (2006) propose a cost-based model to forecast no-show rates in the airline industry in which no-show rates were estimated assuming they followed a normal distribution and, in a second stage, PNR data are used to adjust these forecasts. In this research they conclude that by using this model, income could increase between 15 and 18% for the applied time window. Other strategies propose the mix of a traditional model, based on the analysis of non-causal time series, with PNR-based models. This is the case of the research published by Neuling, Riedel, and Kalka (2004) who compared exponential smoothing results with forecasts performed with tree-based models and a blended solution. It was found that exponential smoothing outperformed tree-based methods in the first stages, but it changed when more data were used for that matter. However, the proposed blended model improved overall. Tree-based models have been also used by Lawrence and Cherrier (2003) who applied C4.5 decision-tree, a segmented Naive Bayes algorithm and an ensemble aggregation method with the aim of forecasting no-show rates in the airline industry. The results showed the method outperformed conventional methods and improved income rates between 0.4% and 3.2%. More recently, Cao, Ding, He, and Zhang (2010) compared three different machine learning techniques, using real data from a Chinese airline to forecast no-show rates. They concluded that the most accurate model was the logistic regression model, followed by artificial neural network and the decision tree model, respectively.

In the hospitality industry recent research papers address the problem of hotel cancellation forecasts by employing PNR data. This is the case of Romero Morales and Wang (2010) who compared several data mining techniques with the aim of forecasting hotel cancellations by applying different time frames. In their research, they used decision tree-based methods, Naive Bayes based methods and support vector machine (SVM), concluding that the latter is a promising method for this task. Another important observation they made about their study of different time frames is that the closer the reservation is to the time of service, the less errors were found in each technique tested. Later research in the field has forecasted individual cancellations using similar AI techniques (Antonio et al., 2019a; 2019a) with a high rate of accuracy. In fact, Wu et al., (2017, pp517) note that techniques based on Artificial Intelligence (AI) have been applied in the hotel and lodging industry with a satisfactory performance. As stated by these authors, Artificial Neural Network (ANN) models appear the most in literature, followed by “Support Vector Regression (SVR), rough set model, fuzzy system methods, genetic algorithms and Gaussian Process Regression (GPR)”.

As it can be appreciated, advanced ensemble methods have been successfully applied in the airline industry with the aim of forecasting cancellations, however, no evidence has been found on the use of these kinds of technique in the hospitality industry, thus providing a novel factor to the present research. The goal of the ensemble learning methods is to collect the output of several individual classifiers for obtaining a more accurate result. The most popular techniques for ensemble are the bootstrap aggregation (bagging) and boosting. While the first one creates multiple versions of a predictor by sampling the training set with a replacement so that the final result is obtained by combining them; the boosting method uses the whole training set in each iteration, assigning a weight for each training instance which is adjusted during the process. As an example, Dietterich (2000) compared bagging, boosting and randomisation as ensemble methods for decision trees. He concluded that boosting obtined the best results in most cases while the others achieved similar results. Opitz and Maclin (1999) studied the effects of applying bagging and boosting methods for ensemble decision trees and neural networks. They noted that while the bagging method normally delivered better results than single classifiers, in some cases boosting got much better results. However, boosting proved to obtain worse results than the single classifier in some tests. Despite the fact that there are multiple ensemble methods (Zhou, 2012), new ensemble models have been published. Most recently, Yu, Lai, and Wang (2008) proposed a multistage nonlinear radial basis function neural network ensemble for forecasting exchange rates,

which uses the output of several neural networks in order to create an ensemble applying another neural network structure. They note promising results considering this proposal outperformed the rest of the methods used.

3. Model development

Data mining is a creative process in which data are treated to find patterns, trends or rules that explain the underlying behaviour behind them (Wirth & Hipp, 2000). In order to follow an orderly process, Cross Industry Standard Process for Data Mining (CRISP-DM) was applied in this study. This is a standard created for industry data mining, which describes the life cycle of a data-mining project broken down in six phases. The intention was to build an AI model with the aim of forecasting individual hotel critical cancellations, which could identify specific customers likely to cancel close to the time of service.

In this section, firstly, the characteristics of the data used in this research are described, as well as how the data was treated before applying any of the methods. Secondly, the techniques applied in the research are detailed.

3.1. Understanding the data set

This research was carried out applying different machine learning techniques to real data provided by a hotel chain, which contain more than 10,000 booking records between 2016 and 2018. In order to better understand the particularities of the hotel collaborator, the main descriptive characteristics related to the facilities, locations and key points of the provided dataset are detailed later. This hotel is located in the centre of Gran Canaria (Spain), one of the most relevant European destinations in terms of the so-called, sun and beach tourism (Pérez-Rodríguez & Acosta-González, 2007) and is ranked with a four-star classification. The excellent weather conditions of Gran Canaria throughout the year attract tourists from around the world but in particular from Europe. Almost 50% of the visitors are German and British, followed by Norwegian, Swedish and Dutch holidaymakers (Medina-Muñoz & Medina-Muñoz, 2012). This data coincides with the reservations at the hotel and therefore, it can be considered a representative local hotel. Commonly, the facilities of this kind of hotel usually consist of at least one swimming pool, a fitness area, outdoor sports tracks, food service, sauna, spa, and business centres, and they usually offer complementary services, such as bicycle renting or massages.

One of the most significant advantages of this proposal is that it has been developed using the most commonly requested variables asked of customers when they place a reservation by online travel agencies, the hotel itself or through external platforms such as Booking or Trivago, among others. Specifically, a total of 13 variables was used independently (Table 1), such as nationality, number of nights or channel and only the variable "*weekend*", which represents the number of weekend days within the period of stay, was calculated from the original dataset. Likewise, the state of the reservation was assigned as a dependent variable. This means if the booking was cancelled "close to the entry day" or "with sufficient time", therefore, this approach allows using two-class probability estimation methods. In addition, different time-horizons were considered, ranging from 4 days to 7 days prior to the time of service. Takinginto consideration the excellent results achieved by Antonio et al. (2017a) with an accuracy rate of above 90%, as well as the work conducted by C-Sánchez (2019) with an accuracy of 98% when forecasting general hotel cancellations, the present research is intended to go one step further, as it attempts to identify cancellations that could be made close to the time of service. For this reason, this approach focuses only on cancelled reservations, removing any other type. It should be noted that during the construction phase of the dataset the identity of the guests was not used, therefore, avoiding the need to query the database, which significantly reduces computational timing and resources. Finally, all variables were coded into numerical numbers and normalised in order to reduce the sensitivity of the model to the different scales and maintain the consistency when comparing the outputs of the different models.

3.2. Methods and model

R statistical software (R Core Team., 2013) was used for the model development in which several AI techniques within the supervised area were applied. Accordingly, the following packages were used: C5.0 (Kuhn et al., 2018), Support Vector Machine (SVM) (Meyer et al., 2019), Artificial Neural Networks (ANN) (Fritsch et al., 2019) and GBM for tree boosting ensemble (Greenwell, Boehmke, & Cunningham, 2019).

All the proposed methods have been used for binary classification in multiple tourism related researches, such as Maeda, Yoshida, Toriumi, and Ohashi (2016), Sung, Chiu, Hsieh, and Chou (2011) for C5.0, Chen and Wang (2007) or Romero Morales and Wang (2010) for SVM, Claveria, Monte, and Torra (2015) or Jun, Yuyan, Lingyu, and Peng (2018) for ANN or even all three within the same research (e.g. Akın, 2015). However, each method works in a different manner. The C5.0 algorithm is a tree decision technique, which attempts to split the training set into smaller groups according to the data features so that each part contains one single type by running an iterative process (Mingers, 1989; Minz & Jain, 2003). One of the major advantages of this technique relies on the possibility it brings to represent the final tree structure, composed by nodes joined with branches, which allows a better understanding of the classification process. This characteristic is not applicable to SVM or ANN as they act as a black box. Specifically, for binary classification, the support vector machine technique uses the structure risk minimisation principle in order to find the hyperplanes, which can better segregate both types according to the data features (Bishop, 2006; Romero Morales & Wang, 2010). As it is not usual to find linear classifications, this technique proposes a data transformation by using a kernel function, which allows finding a more appropriate linear segregation in a higher dimensional space. On the other hand, ANN makes use of a structure composed by a certain number of neurons that connect among themselves and are organised in one or more layers (Hyndman & Athanasopoulos, 2018). The information is transmitted through the network, in which each neuron weighs the sum of the outputs in the previous layer and produces one single output after passing through a transfer function (Agatonovic-Kustrin & Beresford, 2000; Livingstone Livingstone, 2008; Shalev-Shwartz & Ben-David, 2014). In addition, this research makes use of ensemble methods, which lead a combination of multiple single classifiers to increase the forecasting capabilities (Opitz & Maclin, 1999). Specifically, the boosting ensemble, one of the most popular techniques used, proposes to weigh each single classifier in a way that a higher weight can be applied to those misclassified items and therefore, weak learners can be reinforced (Wang & Zhang, 2005).

The present methodology proposes to split the whole dataset in two sets, the first one is used to set the boosting algorithm and the second one to test the ensemble (Fig. 1).

In a first stage, a balanced training and testing set is created from *dataset A*, which are used for training and forecasting critical cancellations with C5.0, R-part, SVM and ANN techniques, as well as performing an initial evaluation of each method. This procedure is repeated a hundred times in order to avoid the lack of reliability of single training and testing, and in each case the training and testing datasets are randomly selected. During this process, the result of each algorithm and the actual values are saved to be used later for setting the ensemble method. After that, *dataset B*, which contains 50% of the original dataset, is used for repeating the same procedure, but in this case, outputs of each individual technique are used for evaluating the ensemble method. Likewise, each individual method is evaluated in this second run in order to properly compare the outputs across the different techniques.

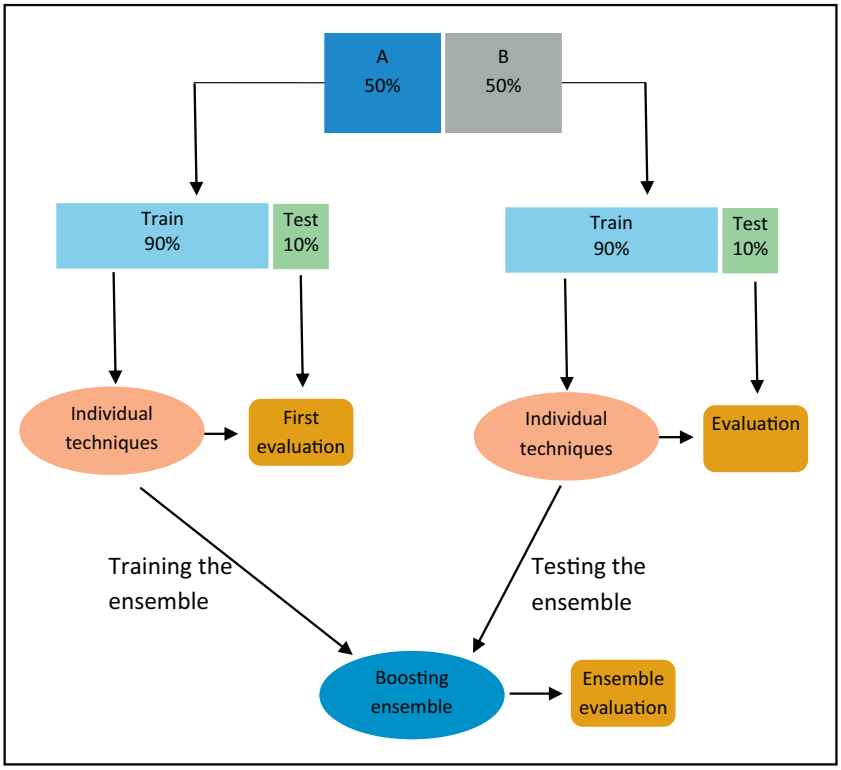

Fig. 1. Flowchart of the forecast model.

4. Results

In order to analyse the results, a confusion matrix was used. This is quite common in two-class problems and consists of a contingency table that shows the difference between the actual and the predicted class calculated in the test phase in a labelled table (Bradley, 1997). In order to evaluate the models, several performance measures were applied, such as the sensitivity "the proportion of true positives correctly detected by the test" expresses, the specificity that is expected to measure "the proportion of true negatives correctly identified by the test" (Altman & Bland, 1994), the reliability of the model and the accuracy of the proportion between the true positives forecasted and the total positive items (Table 1). In addition, a common way to evaluate a model's performance is by using the ROC curve through the relationship between the true positives and true negatives forecasted along different cut off points in a graph (Bradley, 1997). Indeed, the area under this curve is another parameter analysed in this research (Fig. 1). It should be noted that this approach attempts to forecast cancellations in a short time horizon, which means that the data set is highly unbalanced, reaching rates of 5%. This configuration makes it more complicated to achieve accurate models because of the lack of data for training the models.

According to the results (Table 2), the tree-based algorithm C5.0 is the individual technique that shows the best output, followed by SVM and ANN respectively. As can be appreciated, in the most unfavourable case, which is the 4-day window, the least accurate rate of 60% corresponds to the ANN technique while the ensemble method achieves 73%, which is a good level of accuracy. A similar scenario is found for the case of the Area Under Curve (AUC) (Fig. 2) where the ANN technique delivers the lowest value (0.60) and the ensemble method succeeded in improving the AUC at 0.73. This can be explained because of the low number of positive cases that have a negative effect on the training phase, especially for the ANN, which is data hungry. Regards specificity and sensitivity, both are balanced in all cases, except for the case of ANN, in which the specificity is slightly lower than the rest however, sensitivity improves significantly.

Table 2
Summary of results: performance measures for each method.

	KPIs	C5.0	SVM	ANN	Boosting ensemble
4 days	Accuracy	0.685	0.673	0.601	0.730
	Precision	0.568	0.551	0.242	0.658
	Specificity	0.650	0.639	0.559	0.701
	Sensitivity	0.741	0.728	0.859	0.769
	AUC	0.685	0.736	0.683	0.802
5 days	Accuracy	0.738	0.698	0.674	0.793
	Precision	0.635	0.694	0.443	0.778
	Specificity	0.697	0.696	0.619	0.785
	Sensitivity	0.799	0.699	0.825	0.803
	AUC	0.807	0.767	0.729	0.873
6 days	Accuracy	0.739	0.716	0.689	0.805
	Precision	0.634	0.718	0.488	0.797
	Specificity	0.698	0.717	0.635	0.800
	Sensitivity	0.803	0.715	0.817	0.809
	AUC	0.813	0.786	0.745	0.876
7 days	Accuracy	0.736	0.716	0.692	0.805
	Precision	0.700	0.705	0.538	0.808
	Specificity	0.720	0.711	0.647	0.807
	Sensitivity	0.755	0.720	0.776	0.804
	AUC	0.808	0.778	0.740	0.885

In this research, four timeframes were considered, ranging from those cancellations placed 4 to 7 days prior to the entry day, so that when a greater period of time is considered, the number of actual cancellations increases and thus, more positive cases are available for training the models. This is reflected in the results, which improve as more time prior to the entry day is considered, as well as, the specificity and sensitivity, whose trend tends to be more balanced.

On the other hand, the results confirm that the ensemble technique successfully improves the individual techniques in all cases, achieving up to 14% of AUC above the lowest value and also reaching balanced specificity and sensitivity. As an example of the significant improvements achieved with the ensemble technique in Fig. 3 the performance

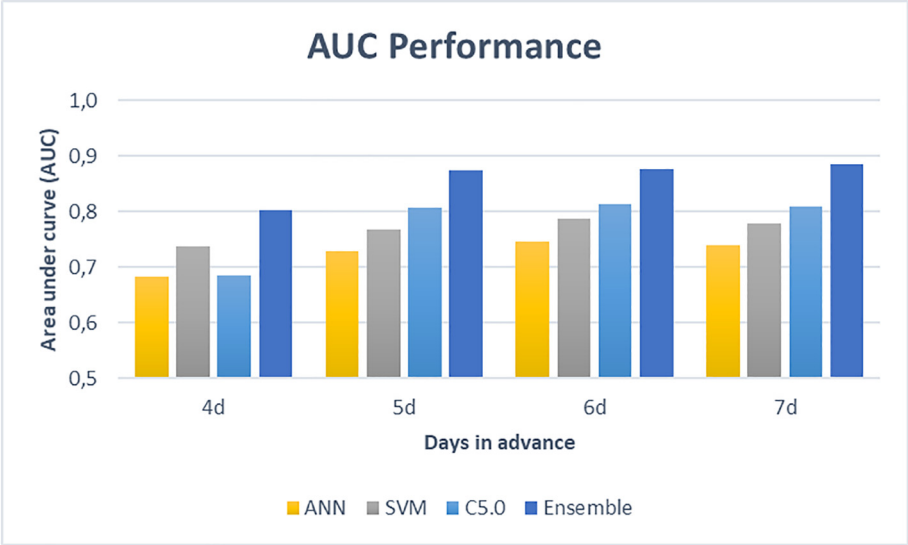

Fig. 2. AUC performance of each method considering number of days in advance.

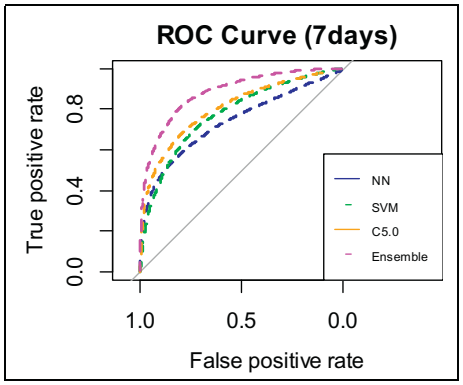

Fig. 3. ROC curve for each method for the case of 7 days prior to the entry day

of the individual and ensemble methods is plotted for the forecasts of 7 days in advance.

5. Conclusion and future work

Having knowledge on potential cancellations is very valuable for the hotel and lodging industry, as it is essential for those that belong to this sector to minimise their idle capacity. In this sense, the data collected during the booking process plays a vital role in helping hoteliers to forecast cancellations and therefore, improve their booking schedules.

This research has focused on forecasting hotel cancellations that are made close to the time of service, using Artificial Intelligence techniques applied to real PNR data, achieving good results. In this regard, PNR data were provided by a four-star hotel located in Gran Canaria (Spain) corresponding to reservations between 2016 and 2018.

Considering the scarce amount of literature in the field (Antonio et al., 2019a; Hassani et al., 2017; Zakhary et al., 2011), this study contributes to expanding knowledge on hotel cancellations, but it also focuses on those cancellations that are more likely to generate a revenue loss for the hotel. This is especially relevant as we are not aware of any previous research that has specifically explored this particular issue within the hotel and lodging industry. Furthermore, this study has focused on analysing individual customer features, thus allowing for a more in-depth study of the customers themselves instead of a mass of clients as has been done traditionally (Antonio et al., 2017a; Antonio et al., 2019a; Falk & Vieru, 2018).

The main theoretical contribution arises in relation to the proposed methodology, which allows forecasting hotel cancellations made with short-time horizons with a good level of accuracy. Moreover, this methodology makes for a more efficient procedure in terms of using a simple data preparation procedure and a reduced number of variables. It is not necessary to integrate external data sources, which could seriously complicate this step, nor is it necessary to look at the individual client history. In addition, it should be noted that this study has been developed using only 13 variables, in comparison with other related research papers, which use 37 variables (Antonio et al., 2017a).

This study has also allowed us to prove the effectiveness of using ensemble algorithms for forecasting individual cancellations close to the time of service, which improve the AUC up to 14% over the individual classifiers. This method can therefore, be considered a novel approach, as we are not aware of it being used previously for forecasting hotel cancellations.

As far as the practical conclusions are concerned, the most noteworthy is the use of AI techniques to PNR databases and the possibility this brings of identifying individuals who are likely to cancel a few days prior to the booking with a good level of accuracy and not just a ratio or percentage of probable cancellations over the total number of bookings.

In terms of managerial implications, this research provides several interesting findings. We have seen that short-time horizon cancellations have the most negative impact for hotels (Chen et al., 2011; Koide & Ishii, 2005), as they leave hoteliers with very little time to react, forcing them in many cases to lose the sale or resell the room at a significant reduction in price. This situation has become even worse in recent years because of new customer behaviour, whereby consumers make several bookings in multiple establishments in order to keep their options open, until finally choosing one at the last minute and cancelling the rest (Antonio et al., 2017a). In the same manner, the increase of last-minute offers means clients take advantage of this situation by booking a cheaper room and cancelling any previous reservations made. To reduce this risk, many hotels impose cancellation policies that usually imply clients paying a specific amount if the cancellation is made after a certain date. Although, this kind of policy has proven to be effective (Zakhary et al., 2011), it can also have a negative impact on the hotel's reputation. Indeed, booking managers try to avoid any arbitration or litigation, especially in the case of groups (DeKay et al., 2004).

To overcome this situation, hoteliers may use specific forecasting tools for short-horizon cancellations. The approach presented in this paper provides information about future guests likely to cancel close to

the time of service and this can help management in improving their strategies in order to maximise their capacity while assuming a lower risk. One of the strategies currently used is overbooking, which aims to sell services over and above the hotel's capacity, with the expectation that some bookings will fall through. If hotel managers make decisions based on reliable forecasts, overbooking risks may be reduced, so that full capacity can be exploited and the relocating of guests can be avoided, which not only impacts negatively on revenue, but also on the hotel's reputation (Dong & Ling, 2015). Pricing strategies can also be improved; for example, Abrate & Viglia (2016) encourage hotels to offer only premium rooms during a certain period of time in which reservations for standard rooms are almost guaranteed for the following days. By taking these actions, hoteliers may increase profit, but they also run the risk of losing some standard room reservations already placed if they do not have access to a reliable short-time horizon cancellation-forecasting tool. In the same manner, having information on the guests likely to cancel close to the time of service allows taking proactive actions, such as contacting the clients directly by phone, sending a reminder or trying to retain the reservation by offering special discounts on additional services (e.g. spa, mini-golf or dinner, among others).

With respect to the implementation of the proposed methodology in a real environment, this procedure is very practical for several reasons. One of the main advantages is the simplicity of the data process, which allows for configuring models faster that, in turn, leads to more frequent training, so that models can better reflect market trends. This also supposes an advantage for hotels because it does not force them to invest in expensive computational resources. At the same time, this procedure works with the most-commonly requested data used in online forms, which in some cases provides the only information available. These variables are presented in all online booking portals reviewed, from online travel agencies (e.g. Booking, Tripadvisor or Trivago) to the hotels themselves (e.g. Radisson Hotels & Resorts, Riu Hotels & Resorts or Meliá Hotel Resorts).

All the improvements that this methodology offers, together with the promising results achieved using data analysis techniques highlight the importance of keeping a reliable historical database in the hotel and lodging industry and the additional value that it provides to organisations.

Finally, future research and the limitations of the present study are presented. Future prospects should include the application of this presented methodology to other data sets provided by other hotels with different characteristics, such as hotel location, customer target (luxury, standard or economic), market niche or hotel policies (e.g. cancellation policies) among others. It would be interesting to add new variables related to customer preferences, such as special requests, types of additional services required or booking purpose (e.g. business or pleasure). In addition, despite the fact that good results have been achieved, it would be interesting to use larger datasets in order for the models to be better trained.

With regards the limitations, as forecasting relies on historical records, if sudden changes are to occur (e.g. natural disasters, pandemics or coups among others), the accuracy of the model could be initially compromised until it is re-trained with the latest records.

Credit author statement

All authors have contributed to all phases of the research.

References

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

Akın, M. (2015). A novel approach to model selection in tourism demand modeling. *Tourism Management, 48*, 64–72. https://doi.org/10.1016/j.tourman.2014.11.004.

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

Antonio, N., de Almeida, A., & Nunes, L. (2017a). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies, 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 Hospitality Quarterly.*. 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. *Tourism & Management Studies,* 16.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer (ISBN 978-0-387-31073-2).

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

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting—A case study of Durban, South Africa. *Tourism Management, 22*(4), 403–409. https://doi.org/10.1016/S0261-5177(00)00068-6.

Cao, R. Z., Ding, W., He, X. Y., & Zhang, H. (2010). Data mining techniques to improve no-show forecasting. *Proceedings of 2010 IEEE international conference on service operations and logistics, and informatics* (pp. 40–45). . https://doi.org/10.1109/SOLI.2010.5551620.

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. *International Journal of Hospitality Management, 30*(1), 129–135. https://doi.org/10.1016/j.ijhm.2010.03.010.

Chen, C.-C., & Xie, K. (2013). Differentiation of cancellation policies in the U.S. hotel industry. *International Journal of Hospitality Management, 34*, 66–72. https://doi.org/10.1016/j.ijhm.2013.02.007 (Lijia).

Chen, K.-Y., & Wang, C.-H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management, 28*(1), 215–226. https://doi.org/10.1016/j.tourman.2005.12.018.

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management, 24*(3), 323–330. https://doi.org/10.1016/S0261-5177(02)00068-7.

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. *International Journal of Tourism Research, 17*(5), 492–500. https://doi.org/10.1002/jtr.2016.

C-Sánchez, E. (2019). Adding value to tourism management through artificial intelligence: A case-based study. Doctoral research. Retrieved from https://accedacris.ulpgc.es/handle/10553/30017.

Dietterich, T. G. (2000). *An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Vol. 19*.

Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management, 30*(10), 3100–3116. https://doi.org/10.1108/IJCHM-08-2017-0509.

Garrow, L. A., & Koppelman, F. S. (2004). Predicting air travelers' no-show and standby behavior using passenger and directional itinerary information. *Journal of Air Transport Management, 10*(6), 401–411. https://doi.org/10.1016/j.jairtraman.2004.06.007.

Gorin, T., Brunger, W. G., & White, M. M. (2006). No-show forecasting: A blended cost-based, PNR-adjusted approach. *Journal of Revenue and Pricing Management, 5*(3), 188–206. https://doi.org/10.1057/palgrave.rpm.5160039.

Greenwell, B., Boehmke, B., & Cunningham, J. (2019). Gbm: Generalized boosted regression models (version 2.1.5). Retrieved from https://CRAN.R-project.org/package=gbm.

Gunter, U., & Önder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management, 46*, 123–135. https://doi.org/10.1016/j.tourman.2014.06.017.

Hajibaba, H., Boztuğ, Y., & Dolnicar, S. (2016). Preventing tourists from canceling in times of crises. *Annals of Tourism Research, 60*, 48–62. https://doi.org/10.1016/j.annals.2016.06.003.

Haller, M., Pröll, B., Retschitzegger, W., Tjoa, A. M., & Wagner, R. R. (2000). Integrating heterogeneous tourism information in TIScover—The MIRO-Web Approach. In D. R. Fesenmaier, S. Klein, & D. Buhalis (Eds.). *Information and Communication Technologies in Tourism 2000* (pp. 71–80). . https://doi.org/10.1007/978-3-7091-6291-0_7.

Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., & Gupta, R. (2017). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research, 63*, 112–127. https://doi.org/10.1016/j.annals.2017.01.008.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice.* (2.ª Ed.). Otexts.

Ivanov, S., & Zhechev, V. S. (2011). Hotel revenue management – A critical literature review. *SSRN Electronic Journal.*. https://doi.org/10.2139/ssrn.1977467.

Jun, W., Yuyan, L., Lingyu, T., & Peng, G. (2018). Modeling a combined forecast algorithm based on sequence patterns and near characteristics: An application for tourism demand forecasting. *Chaos, Solitons & Fractals, 108*, 136–147. https://doi.org/10.1016/j.chaos.2018.01.028.

Koide, T., & Ishii, H. (2005). The hotel yield management with two types of room prices, overbooking and cancellations. *International Journal of Production Economics, 93-94*,

417–428. https://doi.org/10.1016/j.ijpe.2004.06.038.
Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research, 78*, 1–9. https://doi.org/10.1016/j.jbusres.2017.04.016.
Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management, 21*(4), 331–340. https://doi.org/10.1016/S0261-5177(99)00067-9.
Law, R., & Wong, R. (2010). Analysing room rates and terms and conditions for the online booking of hotel rooms. *Asia Pacific Journal of Tourism Research, 15*(1), 43–56. https://doi.org/10.1080/10941660903310102.
Lawrence, R. D., & Cherrier, J. (2003). *Passenger-based predictive Modeling of airline no-show rates. Vol. 10*https://doi.org/10.1145/956750.956796.
Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management, 68*, 116–126. https://doi.org/10.1016/j.tourman.2018.03.006.
Livingstone, D. (Ed.). (2008). *Artificial Neural Networks: Methods and Applications*Totowa, NJ: Humana Press. https://doi.org/10.1007/978-160327-101-1.
Maeda, T. N., Yoshida, M., Toriumi, F., & Ohashi, H. (2016). Decision tree analysis of Tourists' preferences regarding tourist attractions using Geotag data from social media. *Proceedings of the Second International Conference on IoT in Urban Space - Urb-IoT, 16*, 61–64. https://doi.org/10.1145/2962735.2962745.
Medina-Muñoz, D. R., & Medina-Muñoz, R. D. (2012). Determinants of expenditures on wellness services: The case of gran Canaria. *Regional Studies, 46*(3), 309–319. https://doi.org/10.1080/00343404.2010.510501.
Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning, 4*, 17.
Minz, S., & Jain, R. (2003). Rough set based decision tree model for classification. *Data Warehousing and Knowledge Discovery, 2737*, 172–181. https://doi.org/10.1007/978-3-540-45228-7_18.
Mubiru, K. P. (2014). A Markov decision model to optimize hotel room occupancy under stochastic demand. *International Journal of Scientific Research Engineering & Technology, 3*(4), 6.
Neuling, R., Riedel, S., & Kalka, K.-U. (2004). New approaches to origin and destination and no-show forecasting: Excavating the passenger name records treasure. *Journal of Revenue and Pricing Management, 3*(1), 62–72. https://doi.org/10.1057/palgrave.rpm.5170094.
Opitz, & Maclin (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*.
Pereira, L. N. (2016). An introduction to helpful forecasting methods for hotel revenue management. *International Journal of Hospitality Management, 58*, 13–23. https://doi.org/10.1016/j.ijhm.2016.07.003.
Pérez-Rodríguez, J. V., & Acosta-González, E. (2007). Cost efficiency of the lodging industry in the tourist destination of gran Canaria (Spain). *Tourism Management, 28*(4), 993–1005. https://doi.org/10.1016/j.tourman.2006.08.007.
Rajopadhye, M., Ghalia, M. B., Wang, P. P., Baker, T., & Eister, C. V. (2001). Forecasting uncertain hotel room demand. *Information Sciences, 11*.
Romero Morales, D., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research, 202*(2), 554–562. https://doi.org/10.1016/j.ejor.2009.06.006.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* https://doi.org/10.1017/CBO9781107298019.
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. *European Journal of Operational Research, 246*(1), 170–185. https://doi.org/10.1016/j.ejor.2015.04.014.
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. *Annals of Tourism Research, 75*, 338–362. https://doi.org/10.1016/j.annals.2018.12.001.
Sung, R., Chiu, C., Hsieh, P., & Chou, H. (2011). The analysis of hotel customer generated contents in weblogs. *International Conference on Financial Management and Economics, 11*, 5.
Tang, C. M. F., King, B., & Pratt, S. (2017). Predicting hotel occupancies with public data: An application of OECD indices as leading indicators. *Tourism Economics, 23*(5), 1096–1113. https://doi.org/10.1177/1354816616666670.
Uysal, M., & Crompton, J. L. (1985). An overview of approaches used to forecast tourism demand. *Journal of Travel Research, 23*(4), 7–15. https://doi.org/10.1177/004728758502300402.
Wang, S., & Zhang, C. (2005). Network game and boosting J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, & L. Torgo (Eds.), Machine Learning: ECML 2005 (Vol. 3720, pp. 461–472). doi:https://doi.org/10.1007/11564096_44.
Weatherford, L. (2016). The history of forecasting models in revenue management. *Journal of Revenue and Pricing Management, 15*(3–4), 212–221. https://doi.org/10.1057/rpm.2016.18.
Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining. Vol. 11*.
Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management, 29*(1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249.
Yu, L., Lai, K. K., & Wang, S. (2008). Multistage RBF neural network ensemble learning for exchange rates forecasting. *Neurocomputing, 71*(16–18), 3295–3302. https://doi.org/10.1016/j.neucom.2008.04.029.
Yüksel, S. (2005). An integrated forecasting approach for hotels. *The international symposium on the analytic hierarchy process (ISAHP) 2005. Vol. 10*.
Yüksel, S. (2007). An integrated forecasting approach to hotel demand. *Mathematical and Computer Modelling, 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. *Journal of Revenue and Pricing Management, 10*(4), 344–366. https://doi.org/10.1057/rpm.2009.42.
Zhang, G., Wu, J., Pan, B., Li, J., Ma, M., Zhang, M., & Wang, J. (2017). Improving daily occupancy forecasting accuracy for hotels based on EEMD-ARIMA model. *Tourism Economics, 23*(7), 1496–1514. https://doi.org/10.1177/1354816617706852.
Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms.* CRC Presshttps://doi.org/10.1201/b12207.

Eleazar C-Sánchez received his Ph. D. in Enterprise & TICs from the University of Las Palmas de Gran Canaria Spain, in 2019. He combines his professional activity as operations analyst with an active research role in the fields of artificial intelligence & data analysis within business environments.

Gustín J. Sánchez-Medina received the Computer Engineer degree in 1994, the Business Administration and Management degree in 1998 and the Ph.D. degree in 2003 from University of Las Palmas de Gran Canaria (ULPGC-Spain) where he is an Associate Professor in the Department of Economics and Management from 1998. He has published in journals such as Journal of Business Ethics, Neurocumputing, International Journal of Hospitality Management, Regional Studies, PLOS One, International Review of Administrative Sciences, Journal of Small Business Management, Complexity, etc. Since 2012 he is Secretary at University Institute of Cybernetic Science and Technology in ULPGC.

Monica Pellejero-Silva is Ph.D. in Tourism from the University of Las Palmas de Gran Canaria (ULPGC - Spain) since 2013. She is currently Dean of the School of Communication and Director of the Department of Communication at the University of the Middle Atlantic, and professor at the National University of Education Distance (UNED). Her research interests are related to business management in the tourism industry, entrepreneurship and education.