engineering proceedings

MDPI

Proceeding Paper

Detecting Short-Notice Cancellation in Hotels with Machine Learning [†]

Eleazar C-Sánchez [1,2,*] and Agustín J. Sánchez-Medina [2]

[1] Department of Communications, Faculty of Social and Legal Sciences, Mid-Atlantic University, 35017 Las Palmas de Gran Canaria, Spain
[2] University Institute of Cybernetic, Business and Society, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain; agustin.sanchez@ulpgc.es
[*] Correspondence: eleazar.caballero@pdi.atlanticomedio.es
[†] Presented at the 10th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 15–17 July 2024.

Abstract: Cancellations play a critical role in the lodging industry. Considering the time horizon, cancellations placed close to check-in have a significant impact on hoteliers, who must respond promptly for effective management. In recent years, the introduction of personal name records (PNR) has brought innovative approaches to this domain, but short-notice cancellation prediction is still underdeveloped. Using real PNR data with more than 10k reservations provided by a four-star hotel, this research aims to combine fuzzy clustering with tree decision techniques and random forest under R software version 4.3.3 to forecast cancellations placed close to the entry day, slightly improving the performance of individual techniques.

Keywords: hotel cancellation forecast; decision tree algorithm; fuzzy C-means clustering; machine learning; random forest

check for updates

Citation: C-Sánchez, E.; Sánchez-Medina, A.J. Detecting Short-Notice Cancellation in Hotels with Machine Learning. *Eng. Proc.* **2024**, *68*, 43. https://doi.org/10.3390/engproc2024068043

Academic Editors: Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Hector Pomares and Ignacio Rojas

Published: 15 July 2024

1. Introduction

Tourism stands out as a globally expanding industry with huge relevance to the economies of multiple countries worldwide. Its economic significance is undeniable, contributing to national economies through expenditures, taxes, job creation, or business development. However, the dynamic nature of the tourism environment forces it to deal with constant uncertainties, including political instability, weather events, or natural disasters [1]. All these factors have an impact on the hospitality sector, especially within the lodging industry, where effective room occupancy management becomes crucial for revenue optimization. Unoccupied rooms translate to missed revenue opportunities. In fact, some studies affirm that cancellations can contribute to a substantial income loss, potentially reaching up to 20% [2]. Hence, having an accurate forecasting system helps to enhance hotels' operational efficiency. However, forecasting hotel demand presents challenges due to the changing nature of this industry. With the aim of addressing this issue more effectively, some authors point out the importance of "net-demand" [3], which consider not only the raw demand, but also the cancellation of reservation previously placed.

Despite its relevance, the literature on cancellation forecasting remains underdeveloped, particularly on individual cancellations [4]. In addition, the factors that traditionally motivated hotels' cancellations, such as guest illness or adverse weather conditions, now become more complex with the increasing prevalence of internet bookings and the ease of canceling reservations [5]. In fact, during the last decades, the increasing use of online travel agencies has further intensified this issue. As these platforms allow for easy cancellation, cancellation rates have increased [6,7]. This customer behavior forces hoteliers to act rapidly to avoid unsold inventory, normally requiring a reduced price [6]. Hence, the notice period when the cancellation is placed becomes a critical parameter to reduce the risk of revenue loss.

In recent years, the use of artificial intelligence (AI) for forecasting within the tourism industry has experienced exponential growth [8]. Regarding cancellation forecast, the use of the information given by the clients during their reservation, known as personal name records (PNR), has brought excellent results in detecting customers likely to cancel, reaching up to 90% accuracy [4,6,9,10]. However, these works do not focus on cancellations placed a few days before the check-in date but cancellations in general. On the other hand, researches addressing short-notice cancellations require assembling multiple supervised models [11]. This study aims to contribute to the literature on short-notice cancellations in hotels using a forecasting model that combines unsupervised and supervised techniques. The goal of this approach is to detect which clients are likely to cancel close to the check-in day by exploring if the outcome of fuzzy clustering techniques can be used as an input in later models for improving their results. To address this problem, the literature suggests the use of binary classification techniques, such as decision tree algorithm and random forest. For this purpose, more than 10k real reservations recorded from 2016 to 2018 were used. These data were provided by a four-star hotel in Gran Canaria, one of the most important touristic locations in Spain.

2. Literature Review

2.1. The Cancellation Issue in Hotels

In the hospitality industry, where hotels grapple with managing room occupancy in an uncertain environment, unclear incomes and inherent business risks become unavoidable. The impact of demand uncertainty extends beyond occupancy organization and scheduling, affecting internal matters like budget planning, which is heavily reliant on accurate future demand forecasting [12]. With the aim of reducing this uncertainty, the lodging industry has implemented overbooking strategies, cancellation policies, and pricing strategies. Overbooking strategies involve accepting reservations exceeding the establishment's capacity, assuming that some bookings will fail. However, this may lead to extra costs if actual hotel occupancy surpasses its capacity, resulting in guest compensation or relocation and potential damage to the hotel's reputation. On the other hand, cancellation policies aim to mitigate revenue loss from cancellations, particularly last-minute cancellations. Imposing penalties for cancellations beyond a certain day has been shown to significantly decrease cancellations [13]. Nevertheless, strict cancellation policies may adversely affect corporate social reputation and income due to a discouraging effect on clients or significant price discounts [6]. Strategies like price wars are discouraged due to potential long-term impacts on business strategy [14], with varying effectiveness across tourist segments [15]. Attending to this problem, a reliable cancellation forecast is crucial for managerial decision-making, reducing cancellation risks, and facilitating the establishment of appropriate cancellation policies or pricing strategies.

Recent decades have witnessed changes in customer behavior due to information technologies, making it more challenging to predict future demand and cancellation rates. Customers now have access to more information about establishments and services, enabling them to compare different offers based on previous customer experiences. Web portals facilitate easy booking and cancellation, encouraging multiple bookings on similar dates across different hotels, with customers ultimately choosing one option and canceling the rest [7]. This has led to an apparent increase in demand on websites, coupled with higher cancellation rates. This circumstance has occurred simultaneously with a significant increase in last-minute offers, leading to a rise in cancellations as customers attempt to take advantage of a more economical option [16]. These trends have added complexity to net demand forecasts, requiring adaptation to evolving customer behaviors and preferences.

2.2. Demand and Cancellation Forecasting with AI in the Lodging Industry

While non-causal time series models and econometric models have been widely used within the loading industry for forecasting purposes, the latest studies using AI methods have achieved excellent results [17]. According to the objective of this research, this literature review is focused only on AI developments in the hotel industry.

One of the first machine learning approaches within the cancelation forecast proposed several techniques in order to compare their performance across different time frames [18]. More specifically, tree decision-based methods, naive Bayes, and support vector machine (SVM) were employed in their research, concluding that SVM showed promising results in the matter. Subsequent research in this domain has extended the application of similar AI techniques to forecast individual cancellations, achieving a high level of accuracy [4,6,9,10,16]. While some of them advocate for the use of tree-based models, others conclude that artificial neural networks (ANN) are the best candidate to address this problem. These approaches show excellent results on general cancellation forecast; however, short-notice cancellation studies within the lodging industry are still underdeveloped. The articles that address last-minute cancellations within the hotel industry resort to assembling models of different nature [11]. This analysis becomes more challenging due to the time window constraints, which require reducing the number of positive cases, thus affecting the training processes of the techniques. Recent studies on hotel cancellation forecast propose the use of cluster techniques to improve their predictions, improving the accuracy for all time horizons tested [19].

3. Model Development

In order to ensure a systematic approach, this research has been developed following the standard Cross Industry Standard Process for Data Mining (CRISP-DM), which delineates the life cycle of a data-mining project across six phases [20]. The objective was to construct an AI model for predicting short-notice cancellations. This section specifies the preprocessing steps undertaken, followed by a detailed account of the techniques employed in the research.

3.1. Data Source and Pre-Treatment

This study applied diverse machine learning techniques to real data provided by a hotel chain, encompassing over 10k booking records collected from 2016 to 2018. These data contain personal information given by the client during the reservation process (e.g., number of guests, personalized requests, or their nationality). This information is saved for each reservation, forming a set of feature vectors called personal name records (PNR) file. For this research, the PNR data were provided by a four-star hotel located in Gran Canaria (Spain), a prominent European sun and beach tourism destination [21]. Typically, these kinds of hotels offer swimming pools, fitness areas, outdoor sports tracks, food services, saunas, spas, and additional services such as bike rentals or massages.

Following previous studies [9,11], this research intends to use only the most commonly requested variables by customers when booking through online travel agencies, the hotel, or external platforms like Booking or Trivago. Even guest identities were not used during the dataset construction, avoiding database queries, and thereby significantly reducing computational time and resource requirements. Specifically, 13 variables were used (Table 1), including nationality, number of nights, and channel, with the “weekend” variable, representing the number of weekend days during the stay period, derived from the original dataset.

Table 1. Description of the variables used in this study, followed by the type of data (C: categorical; N: numeric).

Name	Description	Type
Status	Canceled within the specific range (0 to 7 days)	C
Adults	Adult headcount	N
Entity	Entities offering room reservation services for customers.	C
Nationality	Guest's nationality	C
Advance payment	Whether advance payment is required	C
Nights	Duration of stay in the hotel in nights	N
Notice period	Days between the booking date and the arrival date.	N
Day of creation	When reservation was placed (day)	N
Month of creation	When reservation was placed (month)	N
Day of check in	Entry day on the reservation	N
Month of check in	Entry month on the reservation	N
Mean price	Average price	N
Sales channel	Channels utilized for reservations, organized into nine entities (e.g., business to business, hotel website, call center or tour operator)	C
Weekend	Count of Saturdays and Sundays during the stay	N

The reservation status was designated as a dependent variable, indicating whether the booking was canceled "with sufficient time" or "close to the entry day" among the existing cancellations, exploring different horizons ranging from 0 to 7 days before the service date. In this manner, our dataset is transformed into a labeled dataset where the dependent variable is binary, with "one" indicating cancellations made within the specified time horizon, and "zero" representing cancellations occurring further in advance. It is worth mentioning that while prior studies have achieved excellent results on general cancellations [4,6,9,10,16], with some of them exceeding 90% of accuracy, the aim of this research is to concentrate on short-notice cancellations. Consequently, this study focuses on cancellations made close to the check-in day among general cancellations, excluding other categories. According to the existing literature in the field, this issue can be addressed using binary classification techniques, as proposed in the next section. Lastly, all variables were numerically coded and normalized so that each variable is standardized with mean zero and a standard deviation of one. This procedure is useful for diminishing model sensitivity to diverse scales, ensuring consistency when comparing outputs across different models.

3.2. Methods and Validation

The problem of short-notice cancellations implies a reduced number of positive cases, typically leading researchers to deal with a highly imbalanced dataset. This research aims to contribute to forecast this type of cancellation by using clustering techniques to complement later methods. Clustering fits into the unsupervised machine learning category, which uses all given data as input with the aim to discover patterns, relationships, or groups [22]. Differently to other works in which raw data are divided into multiple groups for individualized treatment [19,23], this research applies clustering techniques as the input for later supervised models. This approach may contribute to improving the forecasting capabilities of the models to be trained with scarce positive cases by providing a more complete dataset. Unlike traditional "hard" clustering methods (such as kmeans), where each data point is assigned exclusively to a single cluster, this research uses fuzzy C-means clustering (FCM), which provides membership degrees for each group. As a result, we obtain the percentage of membership for a given point, allowing a single item to belong to different clusters simultaneously [24]. This approach allows for a more flexible representation of complex relationships within the data, making it well suited for scenarios where data points may exhibit ambiguity in their cluster assignments [25].

Once fuzzy clusters are calculated and integrated with the original data, the subsequent step involves the deployment of supervised models. According to the characteristics of this problem several binary classification techniques can be used. In this vein, artificial neural networks have achieved excellent results in general cancelation forecast [6,9], but they have proven to be not effective for small training sets [11]. Therefore, this research proposes to use decision tree algorithm and random forest (RF) as they have output good results in similar problems [9,16,18]. The C5.0 algorithm falls into the decision tree models, where the goal is to identify the feature that effectively separates classes and organizes the data based on the values of this feature [26]. This process iteratively divides the training set into smaller subsets, seeking partitions containing only one class [27]. On the other hand, RF is also a tree-based method, which ensemble several trees using the bagging method [28].

The implementation of the models proposed in this work were coded using the following packages: C5.0 [29], E1071 [30] and randomForest [31]; all of them are available for R statistical software in version 4.3.3 [32]. The model was validated using the repeated random subsampling technique, recommended especially in cases involving a dichotomous dependent variable and an imbalanced distribution of raw data, as evident in the present study [33]. This method advocates dividing the entire dataset into a training set for model construction and a testing set for model validation. The process is iteratively repeated 100 times, and the final evaluation is based on the mean value obtained from these repetitions. Considering the imbalanced nature of the dataset, the subdivision process involved balancing both sets to guarantee an equitable representation of each class. Consequently, each dataset comprises 50% of instances from each class for every single run. This balanced subdivision process ensures that each class is equally represented in both the training and testing datasets, thereby preventing the model from being biased towards the majority class and allowing a model generalization. By maintaining this equilibrium, the model can learn from a more diverse range of examples, leading to better performance and reliability in real-world scenarios.

4. Results

This section aims to show the measures used to evaluate the performance of the proposed models.

As commonly used in two-class problems, a confusion matrix was used to evaluate the performance of the models. This matrix serves as a contingency table, illustrating the disparities between the actual and predicted classes calculated during the test phase [34]. Several performance measures were then applied to assess the models, including sensitivity, which expresses the proportion of true positives correctly detected by the test; specificity, intended to measure the proportion of true negatives correctly identified by the test [35]; precision, which calculates the ratio of true positives forecasted to the total positive items predicted by models; and accuracy, which expresses the proportion between the true positives forecasted and the total positive items [36]. Additionally, a widely used method to evaluate model performance involves the ROC curve, illustrating the relationship between true positives and true negatives forecasted across different cutoff points in a graph [34]. The Area Under this Curve (AUC) is a key metric in two-class forecasting problems, and it was also assessed in this research. This parameter ranges from 0 to 1, with 1 representing a perfect forecast.

Various time horizons were explored, ranging from cancellations generated on the same day as the check-in to those occurring a week in advance. For each scenario, up to 10 fuzzy clusters were tested, which allows to explore the behavior of the proposed methodology without incurring an excessive number of groups. This choice is made with consideration to previous findings, which suggest that adding a large number of clusters may diminish the discriminative power of the tool [37] and potentially introduce noise for subsequent techniques. Therefore, the selection of 10 clusters aims to achieve a balance between a comprehensive exploration and maintaining the effectiveness of the analysis. For clarity, all measures are presented for each time horizon without using any cluster (Table 2), while the effectiveness of fuzzy clusters are presented by the variation of AUC, in percentage, compared to the model trained without clusters (Table 3).

Table 2. Performance measures for cancellations placed multiple days in advance. Each row corresponds to a different time horizon specified in the first column.

Days before Cancellation	Accuracy		Precision		Specificity		Sensitivity		AUC	
	C5.0	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**
0	0.658	0.691	0.622	0.650	0.647	0.676	0.670	0.708	0.709	0.755
1	0.693	0.717	0.651	0.695	0.678	0.708	0.710	0.727	0.749	0.783
2	0.694	0.717	0.646	0.674	0.677	0.700	0.715	0.737	0.758	0.777
3	0.705	0.723	0.673	0.685	0.693	0.707	0.719	0.742	0.771	0.795
4	0.701	0.722	0.644	0.671	0.680	0.701	0.727	0.746	0.766	0.794
5	0.729	0.740	0.686	0.687	0.711	0.717	0.751	0.769	0.803	0.818
6	0.732	0.741	0.691	0.696	0.714	0.721	0.753	0.765	0.813	0.828
7	0.729	0.746	0.686	0.710	0.711	0.730	0.751	0.765	0.809	0.830

Table 3. Variation of AUC in percentage, respective to the case without cluster calculated independently for each time horizon. The highest value in each row is represented with bold text.

No. of FCM	Number of Days before the Check-In															
	0		**1**		**2**		**3**		**4**		**5**		**6**		**7**	
	C5.0	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**	**C5.0**	**RF**
2	2.59	1.26	1.70	0.17	−0.42	−0.22	0.00	0.94	0.38	0.41	−0.56	0.55	−0.30	0.16	0.12	−0.35
3	2.79	2.82	1.53	1.63	1.40	2.40	1.46	**1.86**	1.27	0.78	0.41	0.52	0.34	0.04	0.85	−0.28
4	0.61	−2.15	0.94	1.41	1.16	1.41	0.78	0.45	1.91	0.81	0.95	0.53	0.50	0.43	1.66	−0.09
5	1.54	−3.40	0.87	0.04	1.63	1.47	1.40	0.31	2.27	1.40	1.23	0.88	0.79	0.72	1.54	−0.09
6	−0.91	1.56	1.28	0.16	1.75	2.46	1.32	0.66	1.89	1.67	2.07	0.77	0.17	1.17	1.11	0.33
7	1.38	**4.14**	1.68	2.02	**2.54**	1.18	1.07	1.35	2.77	**1.68**	1.46	1.50	0.88	0.71	2.15	1.27
8	2.08	0.18	1.35	1.88	2.20	**3.17**	**2.37**	1.38	**3.77**	1.52	1.69	1.85	**1.95**	**1.84**	1.82	0.99
9	1.67	1.46	2.47	**2.17**	1.47	3.03	1.76	0.99	2.81	1.03	1.85	**2.35**	1.92	1.21	**2.82**	1.37
10	**2.79**	0.32	**2.47**	1.34	1.81	1.96	2.05	1.21	3.03	1.12	**2.20**	1.64	1.76	1.20	1.34	**1.43**

The accuracy values range from 65.8% to 73.2% for C5.0 algorithm and from 69.1% to 74.6% for RF, tending to increase with higher time windows for both cases. The sensitivity and specificity values are balanced across the various time frames, meaning that the proposed methodology does not predict one category over the other consistently. The same can be observed for AUC values, ranging from 70.9% to 81.3% in the first case and 75.5% to 83.0% in the second case. Generally, as the number of days before cancellation increases, there is an improvement in the performance metrics. This suggests that the model tends to perform better for larger time horizons. This phenomenon can be explained considering the low number of positive cases for the shortest time frame (t = 0 days), which increases as the time window expands.

Regarding the utilization of fuzzy clustering as an input for the supervised model, it can be observed that the results can be improved by up to 2.6% for C5.0 and 2.3% for random forest on average. The findings indicate that better AUC values can be achieved for all cases by employing several clusters for both supervised techniques. Specifically, the optimal AUC values are obtained when more than seven clusters are utilized. These results suggest that the information extracted during the clustering process is beneficial for feeding both tree-based models.

5. Conclusions and Future Works

This work aims to contribute to the short-notice cancellations in hotels combining fuzzy C-means clustering with two different tree-based algorithms for classification. Given the limited existing literature on the subject [4,38], this work intends to expand the knowledge on the field. Moreover, it focuses on individual short-notice cancelations, which has been occasionally addressed [11].

From a theoretical perspective, the methodology proposed allows to detect individual guests likely to cancel within short period of notice before the check-in day without having to resort to complex methodologies such as ensembles of multiple techniques [11]. In addition, the utilization of variables commonly required in the reservation process and the lack of necessity for customer history led to an efficient methodology.

Several practical perspectives can be drawn. First, the presented methodology enables an easy implementation in hotel chains, allowing hoteliers to retrieve information from the data they store with low investment. Second, this approach focuses on short-notice cancellations, which are particularly problematic for hotel managers [7,39] as they leave them with little margin, often compelling them to resort to price reductions to avoid idle capacities. Considering the increase in last-minute cancellations during the last years [16] due to the convenience provided by online platforms to book and cancel easily, this approach comes to address the challenge. Third, reliable cancellation forecasts may help managers to handle the risk of taking overbooking strategies. This way, they can maximize the utilization of their capacities, mitigating the probability of reputation deterioration or incurring additional costs for guest relocation. It can be extended to pricing strategies, in which cancellations' predictions can be used to improve revenue management. Some authors recommend offering only top-class rooms instead of standard accommodation in certain periods, assuming that more economic options are almost guaranteed [40]. While such actions can enhance profitability, they come with the risk of losing already-placed standard room reservations without access to a reliable short-term cancellation forecasting tool. Fourth, having insights into guests likely to cancel close to the service time enables hoteliers to take proactive actions to retain the client. Reminders or special offers of additional services provided by the hotel (e.g., meals or sauna) are just examples of measures that can be addressed. Finally, the promising results obtained in this study underline the importance of maintaining a reliable historical database in the lodging industry and the additional value it provides to organizations.

Future research could involve applying the presented methodology to datasets from other hotels with different characteristics, including location, customer target, market niche, or hotel policies (e.g., pricing or overbooking strategies). On the other hand, incorporating new variables may improve the results. New data related to the number and kind of special requests or booking purpose may be beneficial for a better model fit.

Regarding limitations, forecasting models relies on historical records, thus, the models must be trained frequently, so they can capture the latest market trends. However unprecedented events (e.g., pandemics) may affect the effectiveness of the predictions as the new situation may change the customer's pattern dramatically.

Author Contributions: Conceptualization, E.C.-S. and A.J.S.-M.; methodology, E.C.-S.; validation E.C.-S.; investigation, E.C.-S. and A.J.S.-M.; data curation, E.C.-S.; writing—original draft preparation, E.C.-S.; writing—review and editing, E.C.-S. and A.J.S.-M.; supervision, E.C.-S. and A.J.S.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used in the study are restricted by a formal agreement with the provider.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chow, W.S.; Shyu, J.-C.; Wang, K.-C. Developing a Forecast System for Hotel Occupancy Rate Using Integrated ARIMA Models. *J. Int. Hosp. Leis. Tour. Manag.* **1998**, *1*, 55–80. [CrossRef]
2. Sierag, D.D.; Koole, G.M.; Van Der Mei, R.D.; Van Der Rest, J.I.; Zwart, B. Revenue Management under Customer Choice Behaviour with Cancellations and Overbooking. *Eur. J. Oper. Res.* **2015**, *246*, 170–185. [CrossRef]
3. Rajopadhye, M.; Ghalia, M.B.; Wang, P.P.; Baker, T.; Eister, C.V. Forecasting Uncertain Hotel Room Demand. *Inf. Sci.* **2001**, *132*, 1–11. [CrossRef]

4. Antonio, N.; De Almeida, A.; Nunes, L. Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hosp. Q.* **2019**, *60*, 298–319. [CrossRef]
5. Park, Y.A.; Gretzel, U.; Sirakaya-Turk, E. Measuring Web Site Quality for Online Travel Agencies. *J. Travel Tour. Mark.* **2007**, *23*, 15–30. [CrossRef]
6. Antonio, N.; De Almeida, A.; Nunes, L. Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 1049–1054.
7. Chen, C.-C.; Schwartz, Z.; Vargas, P. The Search for the Best Deal: How Hotel Cancellation Policies Affect the Search and Booking Decisions of Deal-Seeking Customers. *Int. J. Hosp. Manag.* **2011**, *30*, 129–135. [CrossRef]
8. Song, H.; Qiu, R.T.R.; Park, J. A Review of Research on Tourism Demand Forecasting: Launching the Annals of Tourism Research Curated Collection on Tourism Demand Forecasting. *Ann. Tour. Res.* **2019**, *75*, 338–362. [CrossRef]
9. Sánchez-Medina, A.J.; C-Sánchez, E. Using Machine Learning and Big Data for Efficient Forecasting of Hotel Booking Cancellations. *Int. J. Hosp. Manag.* **2020**, *89*, 102546. [CrossRef]
10. Chen, S.; Ngai, E.W.T.; Ku, Y.; Xu, Z.; Gou, X.; Zhang, C. Prediction of Hotel Booking Cancellations: Integration of Machine Learning and Probability Model Based on Interpretable Feature Interaction. *Decis. Support Syst.* **2023**, *170*, 113959. [CrossRef]
11. Sánchez, E.C.; Sánchez-Medina, A.J.; Pellejero, M. Identifying Critical Hotel Cancellations Using Artificial Intelligence. *Tour. Manag. Perspect.* **2020**, *35*, 100718. [CrossRef]
12. Tang, C.M.F.; King, B.; Pratt, S. Predicting Hotel Occupancies with Public Data: An Application of OECD Indices as Leading Indicators. *Tour. Econ.* **2017**, *23*, 1096–1113. [CrossRef]
13. Zakhary, A.; Atiya, A.F.; El-Shishiny, H.; Gayar, N.E. Forecasting Hotel Arrivals and Occupancy Using Monte Carlo Simulation. *J. Revenue Pricing Manag.* **2011**, *10*, 344–366. [CrossRef]
14. Gehrels, S.; Blanar, O. How Economic Crisis Affects Revenue Management: The Case of the Prague Hilton Hotels. *Res. Hosp. Manag.* **2013**, *2*, 9–15. [CrossRef]
15. Hajibaba, H.; Boztuğ, Y.; Dolnicar, S. Preventing Tourists from Canceling in Times of Crises. *Ann. Tour. Res.* **2016**, *60*, 48–62. [CrossRef]
16. Antonio, N.; De Almeida, A.; Nunes, L. Instituto Universitário de Lisboa Predicting Hotel Booking Cancellations to Decrease Uncertainty and Increase Revenue. *Tour. Manag. Stud.* **2017**, *13*, 25–39. [CrossRef]
17. Peng, B.; Song, H.; Crouch, G.I. A Meta-Analysis of International Tourism Demand Forecasting and Implications for Practice. *Tour. Manag.* **2014**, *45*, 181–193. [CrossRef]
18. Romero Morales, D.; Wang, J. Forecasting Cancellation Rates for Services Booking Revenue Management Using Data Mining. *Eur. J. Oper. Res.* **2010**, *202*, 554–562. [CrossRef]
19. Viverit, L.; Heo, C.Y.; Pereira, L.N.; Tiana, G. Application of Machine Learning to Cluster Hotel Booking Curves for Hotel Demand Forecasting. *Int. J. Hosp. Manag.* **2023**, *111*, 103455. [CrossRef]
20. Wirth, R.; Hipp, J. CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000.
21. Pérez-Rodríguez, J.V.; Acosta-González, E. Cost Efficiency of the Lodging Industry in the Tourist Destination of Gran Canaria (Spain). *Tour. Manag.* **2007**, *28*, 993–1005. [CrossRef]
22. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*; Berry, M.W., Mohamed, A., Yap, B.W., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 3–21. ISBN 978-3-030-22475-2.
23. Tianyang, W. A K-Means Group Division and LSTM Based Method for Hotel Demand Forecasting. *Teh. Vjesn.* **2021**, *28*, 1345–1352. [CrossRef]
24. Cebeci, Z.; Yildiz, F. Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures. *J. Agric. Inform.* **2015**, *6*, 13–23. [CrossRef]
25. Yu, B.; Zheng, Z.; Cai, M.; Pedrycz, W.; Ding, W. FRCM: A Fuzzy Rough c-Means Clustering Method. *Fuzzy Sets Syst.* **2024**, *480*, 108860. [CrossRef]
26. Minz, S.; Jain, R. Rough Set Based Decision Tree Model for Classification. In *Data Warehousing and Knowledge Discovery*; Kambayashi, Y., Mohania, M., Wöß, W., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2737, pp. 172–181, ISBN 978-3-540-40807-9.
27. Mingers, J. An Empirical Comparison of Selection Measures for Decision-Tree Induction. *Mach. Learn.* **1989**, *3*, 319–342. [CrossRef]
28. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How Many Trees in a Random Forest? In *Machine Learning and Data Mining in Pattern Recognition*; Perner, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7376, pp. 154–168. ISBN 978-3-642-31536-7.
29. Kuhn, M.; Weston, S.; Culp, M.; Coulter, N.; Quinlan, R. C50: C5.0 Decision Trees and Rule-Based Models. 2023. Available online: https://CRAN.R-project.org/package=C50 (accessed on 7 July 2024).
30. Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. Misc functions of the department of statistics (e1071), tu wien. *R Package* **2008**, *1*, 5–24.
31. Liaw, A.; Wiener, M. Classification and Regression by random Forest. *R News* **2002**, *2*, 18–22.
32. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
33. Khakifirooz, M.; Chien, C.F.; Chen, Y.-J. Bayesian Inference for Mining Semiconductor Manufacturing Big Data for Yield Enhancement and Smart Production to Empower Industry 4.0. *Appl. Soft Comput.* **2018**, *68*, 990–999. [CrossRef]

34. Landgrebe, T.C.W.; Duin, R.P.W. Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 810–822. [CrossRef]
35. Altman, D.G.; Bland, J.M. Diagnostic Tests 1: Sensitivity and Specificity. *BMJ Br. Med. J.* **1994**, *308*, 1552. [CrossRef] [PubMed]
36. Güneş, S.; Polat, K.; Yosunkaya, Ş. Multi-Class f-Score Feature Selection Approach to Classification of Obstructive Sleep Apnea Syndrome. *Expert Syst. Appl.* **2010**, *37*, 998–1004. [CrossRef]
37. Chen, J.; Pi, D. A Cluster Validity Index for Fuzzy Clustering Based on Non-Distance. In Proceedings of the 2013 International Conference on Computational and Information Sciences, Shiyang, China, 21–23 June 2013; pp. 880–883.
38. Hassani, H.; Silva, E.S.; Antonakakis, N.; Filis, G.; Gupta, R. Forecasting Accuracy Evaluation of Tourist Arrivals. *Ann. Tour. Res.* **2017**, *63*, 112–127. [CrossRef]
39. Koide, T.; Ishii, H. The Hotel Yield Management with Two Types of Room Prices, Overbooking and Cancellations. *Int. J. Prod. Econ.* **2005**, *93–94*, 417–428. [CrossRef]
40. Abrate, G.; Viglia, G. Strategic and Tactical Price Decisions in Hotel Revenue Management. *Tour. Manag.* **2016**, *55*, 123–132. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.