ARTIFICIAL INTELLIGENCE THEORY and APPLICATIONS

More information available at aita.bakircay.edu.tr

Performance Comparison of Classification Algorithms in Hotel Booking Cancellation Prediction

Muhammed TEKİN [1, *], Murat GÖK [1]

[1] *Yalova University, Faculty of Engineering, Dept. of Computer Engineering, Turkey*

* Corresponding Author: Yalova University, Faculty of Engineering, Dept. of Computer Engineering, Turkey
Tel.: +90 226 8155344. E-Mail: muhammed.tekin@yalova.edu.tr

Publication Information

Keywords :
- Hotel Booking,
- Classification algorithms,
- Cancellation prediction.

Category : Full Research Article

Received : 14.03.2021
Accepted : 15.04.2021

A B S T R A C T

Having a high occupancy rate is one of the most important goals of hotel management. However, booking cancellations have a negative effect on the profit rates of the hotels. Although hotel businesses try to develop various solutions to overcome this problem, they cannot achieve the desired result. In this context, it is of great importance for hotels to be able to predict booking cancellations that may occur.

In order to solve this problem, in this study, k-Nearest Neighbors algorithm, Logistic Regression, Artificial Neural Networks, Decision Tree algorithm, Random Forest algorithm and Gradient Boosting algorithm are run on an open shared dataset that includes the reservation information of various hotels between 2015 and 2017. When the results are compared, it has been shown that K-Nearest Neighbors and Random Forest algorithms are the best solutions to the problem with both have 85% accuracy.

1. Introduction

One of the most important needs of humanity since its existence is the need for shelter. Today, with the development of societies, the need for accommodation has arisen not only as a private living space but also for many different needs. The hotel industry, which emerged to meet these needs, ranges from small businesses to international hotel chains. This increase in supply brought competition with it. Hotels with a limited number of capacities are looking for ways to maximize their revenues while using their existing capacities in the most efficient way. Maintaining this competition will only be possible by keeping profits at the highest level [1].

The revenue management system is defined as making use of information systems and pricing strategies in order to provide the right capacity to the right customer at the right price at the right time. This system was first developed for the aviation sector [2] and after its success was seen in this field, it was widely used by the service sector, especially by the hotels. In the hospitality industry, revenue management means making the right room for the right guest, at the right price, at the right time and with the right channel [4]. Thus, profitability can be kept at the highest level [5].

For this purpose, hotels use reservation systems. A reservation symbolizes an agreement between the customer and the hotel. With this agreement, the customer guarantees to receive the service determined at the specified date and price. This situation is beneficial in the first place in terms of revenue management for the hotel. However, most hotels also allow the reservation to be canceled on the customer's side. Although it is beneficial to have such a right by the customer, this creates risks in terms of revenue management for the hotels [3]. The reason for this is that making accurate and realistic predictions is an important factor for revenue management, and canceled reservations have a negative impact on that [6]. This means serious revenue losses for the hotels. Studies show that the cancellation rate in bookings is 20% [7]. This value increases up to 60% in hotels around the airport or on the road [8].

In order to minimize the damage caused by the risks and losses in question, most of the hotels apply the overbooking method [9]. In this method, a room is reserved for more than one customer, as it is predicted that some reservations will be canceled. But sometimes the predictions do not match the real situations and that leads to a loss of customers for the hotel. Because, the customers who cannot get the service in their booking will turn to other hotels. This means a loss of customers for the hotel in question [10, 11]. Strict cancellation policies can be applied to prevent this, but it causes a decrease in the number of bookings [12]. As a result, it causes loss of revenue and reputation for the business in both cases.

From a business point of view, the ability to predict booking cancellations is important both to minimize these negative effects and to have a successful revenue management system and to apply it correctly when a method such as overbooking is desired

There are some studies in the literature about the loss of income caused by booking cancellations and the inefficient use of institutional resources [14-17]. These studies were mostly carried out using statistical methods. For example, Pölt argued that a 20% reduction in estimation errors would provide a 1% increase in operating income [18]. Huang et al., one of the studies for cancellation of booking, were able to reach a specificity value of around 87% in their prediction models that uses artificial neural network [19].

Another study is Morales et al. They tried to examine the cancellation of the bookings over the passenger name record data used by the airline companies [7]. However, in most of the studies, the problem was seen as a regression problem, not a classification problem.

Some of the studies that see it as a classification are as follows:

Antonio et al. In their study using four different hotel data, they developed separate models for each hotel and achieved an average accuracy of more than 90% [20]. One of the more recent studies, Boz et al. They were able to achieve an accuracy rate of 73% by applying various classification methods on a dataset belonging to five different hotels [9]. Antonio et al. build a system that uses the gradient tree boosting method and have achieved an accuracy rate of 85% [3].

In this study, an open shared dataset [3] containing reservation information for resort and city type hotels between 2015-2017 is used to predict booking cancellations by means of various classification methods.

What is expected from the developed system is that hotel businesses can have a pre-warning system to minimize the losses mentioned above. In this way, the managers of the business can take the necessary corrective actions in advance. These actions may include customer-specific cancellation policies, promotional offers, etc.

In order to create the system, k-Nearest Neighbor Algorithm, Logistic Regression, Artificial Neural Networks, Decision Tree Algorithm, Random Forest Algorithm and Gradient Boosting methods were preferred. These methods are also known as supervised learning methods. In these methods, training is

performed first with known data, and then when new unknown data comes, the system is expected to correctly classify it. In addition to these methods, there are also unsupervised methods which are expected to classify a given data correctly without any previous training step. Clustering applications can be given as examples to these [13].

Information on the methods used are discussed in the third part of this document. Performance values and evaluations obtained from the system are presented in the fourth section.

2. Materials and Methods

2.1 Data

The dataset utilized in this study was prepared by Antonio et al. It is the hotel reservation dataset published by [3] in open sharing [21]. In their study, two separate datasets are used for resort and city hotels, respectively, H1 and H2. In this study, both datasets were combined and used in order to avoid the complexity created by the separate datasets in the preprocessing steps. In addition, it was observed that the datasets being separate or unique in performance measurements did not make a significant difference.

There are a total of 119.390 reservation records and 30 attributes with one target field indicating whether the reservation has been canceled or not in the dataset. The historical distribution of the data is between 1 July 2015 and 31 August 2017. Data were collected from hotels located in Portugal. Resort-type hotels are records of some hotels in the Algarve region, while city hotels include some of the hotels in the city of Lisbon. All data are anonymous. In other words, the private information of the hotel or customers is not available in the dataset. Name, data type and explanation information about the fields in the dataset are presented in Table 1.

Table 1. Features, their data types and meanings in the dataset

Feature Name	Type	Info
ADR	N	Average daily room return in monetary terms
Adults	I	Number of adult individuals
Agent	C	ID information of the travel agency. This information is not available in 16,384 records.
ArrivalDateDayOfMonth	I	Data available for hotel arrivals only
ArrivalDateMonth	C	Data available for hotel arrivals only
ArrivalDateWeekNumber	I	Data available for hotel arrivals only
ArrivalDateYear	I	Data available for hotel arrivals only
AssignedRoomType	C	Room type assigned to the reservation
Babies	I	Number of babies
BookingChanges	I	The number of changes from the time the record was created until the entry or cancellation
Children	I	Number of children. There are 4 Null values.
Company	C	Tourism company information. 94% of the data (112,593) contain Null values.
Country	C	There are 488 invalid data. By default, all reservations are saved in Portugal. Country information has been entered after the customer has checked into the hotel.
CustomerType	C	Group, negotiated, temporary, temporary group
DaysInWaitingList	I	The day on the waiting list before the reservation is received
DepositType	C	Refundable, non-refundable, no deposit
DistributionChannel	C	Reservation distribution channel. 'TA': Travel Agency, 'TO': Tour Operator
IsCanceled	C	Target area to be classified
IsRepeatedGuest	C	Is it a permanent customer?
LeadTime	I	The number of days from the time the record was created until the entry or cancellation

Feature Name	Type	Info
MarketSegment	C	The market where the reservation belongs. 'TA': Travel Agency, 'TO': Tour Operator
Meal	C	Undefined / SC: No meal, BB: Breakfast, HB: Half Board, FB: Full Board
PreviousBookingsNotCanceled	I	Number of pre-canceled reservations by the customer
PreviousCancellations	I	Number of previously canceled reservations by the customer
RequiredCardParkingSpaces	I	The amount of vehicle parking space requested by the customer
ReservationStatus	C	Removed because the target field (is_canceled) has a description
ReservationStatusDate	D	The date on which the final status of the reservation is determined. Removed because it depends on the above.
ReservedRoomType	C	Reservation room type
StaysInWeekendNights	I	Number of days spent on weekend nights
StaysInWeekNights	I	Number of days of stay on weekdays
TotalOfSpecialRequests	I	Total number of special requests from customers

Attributes written in bold in Table 1 are attributes not used in the developed system, in other words excluded. The reasons for exclusion are also stated in the explanations of such areas. A similar process is carried out by Antonio et al. [3]. However, in this study, the fields named "required_car_parking_spaces", "assigned_room_type", and "reserved_room_type" that were excluded in that study were not excluded. The reason for this is that these fields are seen to be effective in the correlation and p-value analysis, which will also be mentioned in the data preprocessing section below. In addition, it has been observed that the success in classification training without these areas has decreased, albeit a little. The meanings of data type abbreviations are as follows:

- I : Integer
- D : Date
- C : Categorical

In Table 2, there are some statistical information about the fields that contain numerical values. The first column of Table 2 contains the names of the attributes. The second column contains the correlation coefficient of the related feature. This value expresses the correlation between the target area, "is_canceled", and the feature. Therefore, the higher this value is, the more useful that attribute will be in the training. In this respect, the "deposit type" field, which shows the deposit status of customers in the first row of the table, has the most impact on reservation cancellation. The second most effective area, "lead_time", is the day difference between the reservation date and the reservation date. This seems to be closely related to the cancellation of the reservation.

Along with the correlation values, there is another value that is considered to be statistically significant, p-value, in Table 2. The closer this value is to 0 (zero), the more dominant our hypothesis, that is, the factor that causes reservation cancellation. High p value indicates that the Null hypothesis is correct, that is, it has no effect on reservation cancellation.

Table 2. Some statistical information of the features

Feature Name	Cor.Coef.	P-Value	Count	Mean	Std.	Min	25%	50%	75%	Max
deposit_type	0.47	0.00								
lead_time	0.29	0.00	119,390	104.01	106.86	0	18	69	160	737
total_of_special_requests	0.23	0.00	119,390	0.57	0.79	0	0	0	1	5
required_car_parking_spaces	0.20	0.00	119,390	0.06	0.25	0	0	0	0	8
assigned_room_type	0.18	0.00								
distribution_channel	0.17	0.00								
booking_changes	0.14	0.00	119,390	0.22	0.65	0	0	0	0	21
hotel	0.14	0.00								
previous_cancellations	0.11	8.9327E-319	119,390	0.09	0.84	0	0	0	0	26
is_repeated_guest	0.08	2.31E-189	119,390	0.03	0.18	0	0	0	0	1
customer_type	0.07	7.56E-123								
reserved_room_type	0.06	1.08E-99								
previous_bookings_not_canceled	0.06	1.49E-87	119,390	0.14	1.50	0	0	0	0	72
market_segment	0.06	1.40E-93								
adults	0.06	1.08E-95	119,390	1.86	0.58	0	2	2	2	55
agent	0.05	1.39E-66	103,050	86.69	110.77	1	9	14	229	535
days_in_waiting_list	0.05	2.50E-78	119,390	2.32	17.59	0	0	0	0	391
adr	0.05	9.68E-61	119,390	101.83	50.54	-6	69	95	126	5,400
babies	0.03	2.92E-29	119,390	0.01	0.10	0	0	0	0	10
meal	0.02	1.01E-09								
stays_in_week_nights	0.02	1.15E-17	119,390	2.50	1.91	0	1	2	3	50
children	0.01	0.08	119,386	0.10	0.40	0	0	0	0	10
stays_in_weekend_nights	0.00	0.54	119,390	0.93	1.00	0	0	1	2	19
is_canceled			119,390	0.37	0.48	0	0	0	1	1

Other statistical information is the number of non-Null values in the attribute, arithmetic average, standard deviation, the smallest value, the value in the 25% -50% -75% of the data and the largest value, respectively. 63% of the data (75,166) are reservations that have not been canceled. In this respect, the target area "is_canceled" contains the value 0 (zero) for these lines. The remaining 37% (44,224 units) are canceled reservations and the “is_canceled” fields contain the value 1. These values are also shown graphically in Figure 1.

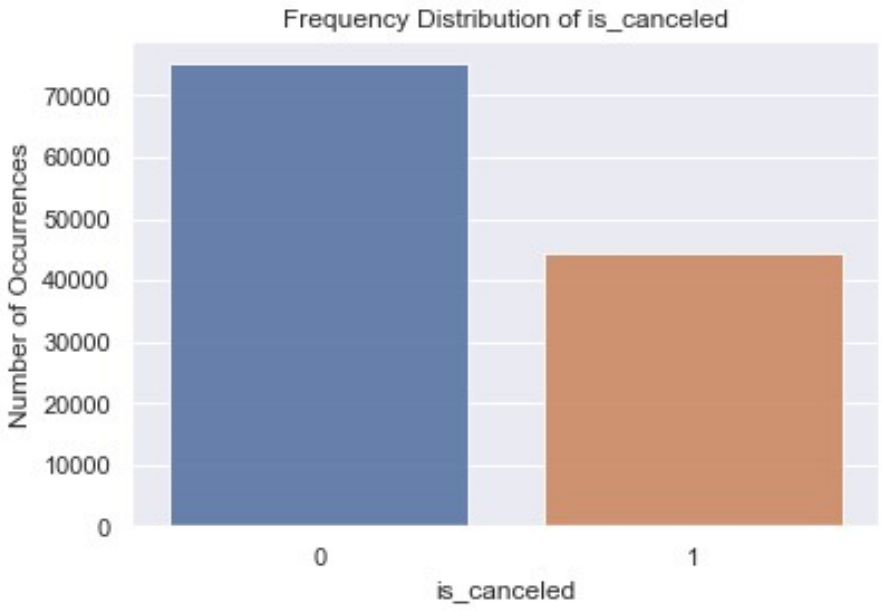

Figure 1. Bar chart for target field "is_canceled"

2.2 Preprocessing

Before building a classification model, the data should be prepared for the methods to be used. Here the steps meant by preparation can be summarized as follows:

i. Attributes that will not benefit or even be potentially harmful should be removed from the data. Measurements such as the correlation coefficient or p-value can be used in this step.

ii. Since the Null value fields among the attributes to be used will negatively affect the results, an action must be taken regarding them. This process can either be the exclusion of the field completely from the training, or it can be like adding another value instead of the Null value, for example, the most repeated value in that field.

iii. Fields containing categorical values must be converted to numeric. Because classification methods only work with numerical type attributes.

iv. Since outlier values in the attribute areas in the training process negatively affect the performance, such values should be corrected with an appropriate method.

v. Subjecting the numerical data within the attribute areas to a certain standardization often benefits educational performance. This process should be done where deemed necessary in this respect.

vi. The fact that the attribute areas are more than necessary can prolong the training process and reduce the performance values. In this respect, the dataset can be made smaller by using feature reduction methods.

In this context, the operations performed on the dataset used in the study are listed below:

- "ReservationStatus" and "ReservationStatusData" fields are in a one-to-one relationship with the "is_canceled" field, which is the target area of the classification operation. Not removing this field may cause the system to deceptively give high accuracy values. Also, this information will not be available as soon as an actual reservation record is added to the hotel database. This information is only entered when the customer cancels the booking. Therefore, it would be extremely wrong to include this field in the training process.

- Since the fields starting with the name of "Arrival" are valid for customers entering the hotel, these fields have been removed from the dataset.

- Since 94% of the data in the "company" field contains Null values, this attribute has been removed from the dataset.

- The data in the "country" field is registered as Portugal by default. However, the actual country information is updated as soon as the customer enters the hotel. In this respect, this area has been removed from the dataset.

- Attributes containing null values are assigned to the most repeated value of the field they belong to.

- For the fields "hotel", "meal", "distribution_channel", "market_segment", "deposit_type", "customer_type", "reserved_room_type", and "assigned_room_type" that contain categorical values, conversion to numeric data type has been made.

- Outlier analysis was conducted, but no significant performance increase was observed at the end of the training, so no corrective action was taken for this step.

- Since all the remaining features at this point are numerical or digitized, standardization and dimension reduction processes have been applied to the entire dataset. While there were 23 features in total before this step, the number of features decreased to 8 with dimension reduction. A variance value of 92% was obtained for these 8 features, which shows that the new features can express 92% of the previous ones.

2.3 Classification Methods

In machine learning applications performance measurements are made by creating more than one model or by changing the parameters of a given model [9]. In this study, six different classification analyses were performed by executing k-Nearest Neighbor (kNN) Algorithm, Logistic Regression (LR), Artificial Neural Networks (ANN), Decision Tree Algorithm (DT), Random Forest Algorithm (RF) and Gradient Boosting on the dataset to estimate booking cancellations. For the training of the models, the dataset was divided into two parts as training and testing with the cross validation method. Here, the k-value of 10 is preferred for cross-validation. Then, the performances of the created models were compared in terms of the determined performance evaluation criteria and the best model was shown. In addition, the Grid-Search method was applied to make the parameters of the models more efficient.

All of the analyzes were carried out using the Python language. The Python language is an interpreted, high-level and open source language [22]. There are libraries in this language in many different fields. Scikit-learn library is one of the most widely used libraries for machine learning applications [23] and in this study it is used.

The classification algorithms executed on the dataset are described below:

i. kNN: In k-Nearest Neighbor (kNN) classification method, in order to determine the class to which a sample in the dataset belongs, the distance to the center points of the nearest k neighbors of the point itself is checked. Whichever class's distance is closer to the point, the point in question is included in that class. [24]. In the event of a tie, a random decision is made. In order to avoid this situation, the value of k is usually chosen as an odd number. There are also studies in the literature about what the k-value should be [25] [26]. In this study, 20 was used as the k value. In Figure 2, the graph of RocAUC performance score for different neighbors, i.e. k values, is presented. It is seen in the graphic that giving more than 20 neighbors does not contribute much to success.

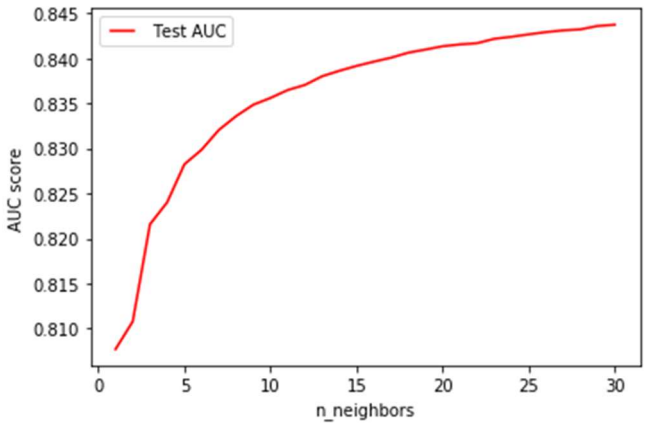

Figure 2. The effect of different neighbor (k) values on success

ii. LR: Logistic Regression is a method frequently used in the field of statistics. While the dependent variable is continuous in linear regression analysis, it is discrete in logistic regression. In this respect, while obtaining the value of the dependent variable in linear regression, the probability of being in one of the classes to which the dependent variable can belong is obtained in logistic regression [27, 28].

iii. ANN: Artificial Neural Networks is a machine learning method that takes the working principle of biology nerve cells as an example. It consists of an input, an output, and a hidden layer. This hidden layer can contain one or more layers in itself. The ANN method is frequently used in classification and regression analysis [29]. In this study, a two-layer structure consisting of 20 neurons each is used as the hidden layer.

iv. DT: Decision Tree algorithm is a method that is frequently used in classification problems due to its simple structure, easy visualization and not needing preliminary data processing like other methods [30-32]. A decision tree consists of root (top), node (middle), branch, and leaf (bottom) components. Nodes represent the attributes in the data [31]. While creating the nodes, the items with the lowest entropy value are selected. There are different algorithms used in creating the tree structure. ID3 (Iterative Dichotomiser) algorithm provides the branching of the tree, taking into account the information gain. The C4.5 algorithm is a more developed version of the ID3 algorithm and takes into account the gain rate. The highest gain rate determines the node where the branch will take place. The CART (Classification and Regression Tree) algorithm is a continuation of Morgan and Sonquist's decision tree algorithm called AID (Automatic Interaction Detection). It was suggested by Breiman etc. in 1984 [33, 34]. This algorithm creates nodes according to the value of the gini coefficient. This coefficient value indicates how often a randomly selected item will be incorrectly identified. The CART method can be used to create both a regression tree and a classification tree. In this study, classification was made using the CART algorithm.

v. RF: The Random Forest (RF) algorithm was first proposed by Tin Kam Ho in 1995 [35]. The definition of the RF algorithm as it is used today is made as it is a classifier consisting of a structured decision tree collection [36]. The algorithm uses multiple decision trees instead of a single decision tree to arrive at the result. Thus, it can obtain more successful results than the decision tree method. The algorithm can be applied to both classification and regression problems [36]. As in the Decision Tree algorithm, the CART method is used here as well.

vi. GB: Gradient Boosting is a model developed to improve the prediction of decision trees [37]. To achieve this improvement, tree branching occurs while trying to minimize the error rate that occurs.

2.4 Cross-Validation

When applying machine learning methods, it is a frequently used method to reserve some of the data (mostly 80%) for training and some for testing. This method is called the holdout method [38]. The aim here is to measure the performance of the created model more accurately. Because when both training and testing are done with the same dataset, overfitting problems may occur [39]. This means that the model has memorized the data. In such a case, when different data come to the model, the performance is lower than the previously predicted values.

Although the holdout method partially solved this problem, randomization of the distribution constitutes the weakness of the method. Because, some data are included in the test set many times, while some data may not be included in the test set at all. For this reason, using k-fold cross validation will give more accurate performance values [40]. In this method, the data are divided into k different groups (usually 10 is chosen for the value of k). One of the groups is used for testing and the other for training. After the model is trained, the performance results are recorded. Then, one of the groups that have never been included in

the test group before is determined as the test group, and the other groups form the training group. The model is retrained and the performance results are also recorded again. When this process is repeated k times, it is ensured that all data groups are included in the test set once. In other words, k different test groups are created and k performance criteria are obtained. When the average of these recorded successes is taken, the performance criterion of the developed model will be obtained. This value represents actual performance more accurately than the exclusion method.

In this study, 10 was preferred as the k value.

2.5 Grid-Search

There are many tuning parameters for machine learning methods. Trying all these manually to find the best performance is a time consuming task. In literature this process is called hyper-tuning the parameters [42] [43]. One of the methods used for this purpose is Grid-Search method. The Grid-Search method is used to obtain the most suitable parameter values for the model by applying the given parameter options on the model.

While listing the machine learning methods in the previous section, some parameters used in the developed system are also specified. These values were obtained by using the aforementioned Grid-Search method. Other parameters not specified are default values of the libraries used.

2.6 Performance Criteria

Accuracy, F1 Score, Cross-Validation accuracy mean and area under the ROC curve (ROCAUC) score are used to evaluate the performance of machine learning methods run on the dataset. Accuracy and F1 Score values are calculated as follows:

$$\text{Accuracy} = \frac{(\text{True\_Positive} + \text{True\_Negative})}{(\text{Total\_Positive} + \text{Total\_Negative})} \quad (1)$$

$$\text{F1} = \frac{2 * \text{True\_Positive}}{(2 * \text{True\_Positive} + \text{False\_Positive} + \text{False\_Negative})} \quad (2)$$

The area under the ROC (Receiver Operating Characteristics) curve (Area Under Curve) shows the performance of a classifier. In short, it is the fraction of true positives to false positives. In other words, it is the ratio of Sensitivity value to False_Positive_Rate (1 - Specificity) values. The closer this value is to 1, the more successful the developed model. In other words, the model can distinguish between classes so well.

The Cross-Validation accuracy mean value is the average of the performance results of 10 different divisions applied in the Cross-Validation method. The 95% confidence interval is also presented here with this information. 95% confidence interval corresponds to two times the standard deviation.

3. Results and Evaluation

In this study, the Scikit-learn library of the Python program language is used. This library has been developed for developing machine learning applications using Python. Table 3 shows the performance values of the classification algorithms mentioned in the previous section.

Table 3. Performance values of the algorithms

	kNN	LR	ANN	DT	RF	GB
Accuracy	0.85	0.75	0.80	0.82	0.85	0.80
F1	0.83	0.67	0.77	0.80	0.83	0.77
Cross-Validation Mean	0.85	0.75	0.80	0.82	0.85	0.80
Confidence Interval	0.01	0.01	0.01	0.01	0.00	0.01
RocAUC	0.83	0.67	0.76	0.80	0.83	0.75

When these values are examined, it is seen that reservation cancellations can be determined between 75% and 85%. The most successful classification algorithms have been RF and kNN. It is seen that both algorithms can achieve 85% accuracy value.

In Table 4, the effect of application of PCA method on data before classification is presented. In order to this, kNN algorithm was run two times separately, without PCA and after PCA was performed on the data. The performance values show that the application of PCA before model training will have a positive effect on performance.

Table 4. The effect of PCA application on success

	kNN	
	WITHOUT PCA (feature = 23)	**WITH PCA (feature = 8)**
Accuracy	0.81	0.85
F1	0.79	0.83
Cross-Validation Mean	0.81	0.85
Confidence Interval	-	0.01
RocAUC	0.79	0.83

In Table 5, the performance values of the systems developed by Antonio et al. [3], where we use the same dataset, and the performance values of the kNN algorithm in this study are presented comparatively. Here, it is seen that the results obtained for Resort (H1) and City (H2) hotel types in the mentioned article and the results of the kNN algorithm in this study are very close to each other.

Table 5. Literature performance comparison

	kNN	**H1 (**Antonio vd.**)**	**H2 (**Antonio vd.**)**
Accuracy	0.85	0.85	0.86
F1	0.83	0.70	0.83
RocAUC	0.83	0.89	0.93

It is very important for the hotel industry to maximize profit rates and have a successful revenue management by making the best of the limited resources available. However, canceled bookings are always an important risk factor in this regard. Since not allowing the cancellation of the bookings is not considered appropriate for customer satisfaction, the only solution for hotels is to predict the cancellations in advance.

In this study, various classification methods are used on the data in hotel information systems to determine the cancellations before they occur. However, the system has aspects that can be improved.

By doing more detailed work on the features, the number of features can be reduced or some detailed features can be collected and a single summary feature can be created.

The number of canceled and uncanceled bookings in the dataset is not balanced. This imbalance can have a negative effect on classification algorithms. It can be overcome by applying various methods.

Better performance can be obtained by testing other classification methods on the dataset.

With the Grid-Search method, fine adjustments can be made for all parameters of the classification models used.

In addition, outside of the engineering field, the effects of pre-booking cancellation estimation on marketing and customer relations may be subject to separate investigation.

References

1. H. Akmeşe and S. Aras, "Otel İşletmelerinde Gelir Yönetimi Uygulamaları: İzmir'de Faaliyet Gösteren 4 ve 5 Yıldızlı Otel İşletmelerinde Bir Uygulama," Int. J. Acad. Value Stud., vol. 3, no. 16, pp. 344–358, 2017, [Online]. Available: https://www.researchgate.net/profile/Halil_Akmese/publication/321113419_Otel_Isletmelerinde_Gelir_Yonetimi_Uygulamalari_Izmir'de_Faaliyet_Gosteren_4_Ve/links/5a2322760f7e9b71dd053d40/Otel-Isletmelerinde-Gelir-Yoenetimi-Uygulamalari-Izmirde-Faaliyet-Goeste.
2. J. C. H. Chen, "An overview of research on revenue management : current issues and future research Wen-Chyuan Chiang * Xiaojing Xu," Oper. Manag., 2007.
3. N. Antonio, A. De Almeida, and L. Nunes, "An automated machine learning based decision support system to predict hotel booking cancellations," Data Sci. J., 2019, doi: 10.5334/dsj-2019-032.
4. R. Mehrotra and J. Ruttley, Revenue management. Washington, DC: American Hotel and Lodging Association, 2006.
5. G. Zöngür, K. G. Yilmaz, and A. Güngördü, "Konaklama İşletmelerinde Dış Kaynak Kullanımı: Ankara İlindeki Dört ve Beş Yıldızlı Otel İşletmelerinde Bir Uygulama," Gazi Üniversitesi İktisadi ve İdari Bilim. Fakültesi Derg., 2016.
6. G. J. van Ryzin and K. T. Talluri, "An introduction to revenue management. In Emerging Theory, Methods, and Applications," INFORMS., 2005, doi: 10.1287/educ.1053.0019.
7. D. Romero Morales and J. Wang, "Forecasting cancellation rates for services booking revenue management using data mining," Eur. J. Oper. Res., 2010, doi: 10.1016/j.ejor.2009.06.006.
8. P. Liu, Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In: Yeoman. 2004, pp. 91–108.
9. M. Boz, E. Canbazoğlu, Z. Özen, and S. Gülseçen, "Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi," Veri Bilim., vol. 1, no. 1, pp. 7–14, Dec. 2018, Accessed: Apr. 18, 2020. [Online]. Available: http://dergipark.org.tr/en/pub/veri/issue/41532/490816#author994713.
10. B. M. Noone and C. H. Lee, "Hotel overbooking: The effect of overcompensation on customers' reactions to denied service," J. Hosp. Tour. Res., 2011, doi: 10.1177/1096348010382238.
11. V. İyitoğlu and N. Tetik, "Fazla Oda Satan Otellerin Kullandığı Yaygın İyileştirme Faaliyetinin Yerli Turistlerin Memnuniyet ve Tekrar Gelme Niyetlerine Etkisinin Bazı Değişkenler Açısından Değerlendirilmesi," Tur. Akad. Derg., vol. 3, no. 1, pp. 57–68, May 2016, Accessed: Apr. 18, 2020. [Online]. Available: http://dergipark.org.tr/tr/pub/touraj/issue/24968/263496.
12. S. J. Smith, H. G. Parsa, M. Bujisic, and J. P. van der Rest, "Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry," J. Travel Tour. Mark., 2015, doi: 10.1080/10548408.2015.1063864.
13. W. Sullivan, Machine learning Beginners Guide Algorithms: Supervised & Unsupervised learning, Decision Tree & Random Forest Introduction. USA,: CreateSpace Independent Publishing Platform, 2017.
14. S. Ivanov, Hotel Revenue Management: From Theory to Practice. 2014.
15. D. K. Hayes and A. Miller, Revenue Management for the Hospitality Industry. Wiley, 2010.

16. W. Caicedo-Torres and F. Payares, "A machine learning model for occupancy rates and demand forecasting in the hospitality industry," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, doi: 10.1007/978-3-319-47955-2_17.

17. L. R. Weatherford and S. E. Kimes, "A comparison of forecasting methods for hotel revenue management," Int. J. Forecast., 2003, doi: 10.1016/S0169-2070(02)00011-0.

18. S. Pölt, "Forecasting is difficult – especially if it refers to the future," in Reservations and Yield Management Study Group Annual Meeting Proceedings, 1998, doi: 10.1002/for.1094.

19. H.-C. Huang, A. Chang, and C.-C. Ho, "Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model," Prz. Elektrotechniczny, vol. 89, pp. 178–180, 2013.

20. N. Antonio, A. de Almeida, and L. Nunes, "Predicting hotel booking cancellations to decrease uncertainty and increase revenue," Tour. Manag. Stud., 2017, doi: 10.18089/tms.2017.13203.

21. N. Antonio, A. de Almeida, and L. Nunes, "Hotel booking demand datasets," Data Br., 2019, doi: 10.1016/j.dib.2018.11.126.

22. "Python." https://www.python.org/ (accessed Apr. 01, 2020).

23. T. Elliott, "The State of the Octoverse: machine learning," 2019. https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/ (accessed Apr. 19, 2020).

24. P. Cunningham and S. J. Delany, "K -Nearest Neighbour Classifiers," Mult. Classif. Syst., 2007, doi: 10.1016/S0031-3203(00)00099-6.

25. P. Hall, B. U. Park, and R. J. Samworth, "Choice of neighbor order in nearest-neighbor classification," Ann. Stat., 2008, doi: 10.1214/07-AOS537.

26. K. A. Heller, "Efficient Bayesian Methods for Clustering," New York, 2007.

27. S. Kilic, "Binary logistic regression analysis," J. Mood Disord., 2015, doi: 10.5455/jmood.20151202122141.

28. "Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama," Kocaeli Üniversitesi Sos. Bilim. Derg., 2004.

29. S. Lek and Y. S. Park, "Artificial Neural Networks," Encycl. Ecol., pp. 237–245, Jan. 2008, doi: 10.1016/B978-008045405-4.00173-7.

30. K. A. Grajski, L. Breiman, G. V. Di Prisco, and W. J. Freeman, "Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART," IEEE Trans. Biomed. Eng., 1986, doi: 10.1109/TBME.1986.325684.

31. J. R. Quinlan, "Induction of Decision Trees," Mach. Learn., 1986, doi: 10.1023/A:1022643204877.

32. Y. Yang, S. S. Farid, and N. F. Thornhill, "Prediction of biopharmaceutical facility fit issues using decision tree analysis," Comput. Aided Chem. Eng., vol. 32, pp. 61–66, Jan. 2013, doi: 10.1016/B978-0-444-63234-0.50011-7.

33. E. Sezer, A. Bozkır, A. S. Yağız, and S. Gökçeoğlu, "Karar Ağacı Derinliğinin CART Algoritmasında Kestirim Kapasitesine Etkisi: Bir Tünel Açma Makinesinin İlerleme Hızı Üzerinde Uygulama," in Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, 2010.

34. Aslı ÇALIŞ, Sema KAYAPINAR, and Tahsin ÇETİNYOKUŞ, "Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama," Endüstri Mühendisliği Derg., 2014.

35. T. K. Ho, "Random decision forests," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1995, doi: 10.1109/ICDAR.1995.598994.

36. L. Breiman, "Random forests," Mach. Learn., 2001, doi: 10.1023/A:1010933404324.

37. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., 2001, doi: 10.2307/2699986.

38. K. A. Ross et al., "Cross-Validation," in Encyclopedia of Database Systems, Springer US, 2009, pp. 532–538.

39. D. M. Hawkins, "The Problem of Overfitting," Journal of Chemical Information and Computer Sciences, vol. 44, no. 1. pp. 1–12, Jan. 2004, doi: 10.1021/ci0342472.

40. T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," Pattern Recognit., vol. 48, no. 9, pp. 2839–2846, Sep. 2015, doi: 10.1016/J.PATCOG.2015.03.009.

41. "Çapraz doğrulama diyagramı" (Accessed Apr. 19, 2020). https://upload.wikimedia.org/wikipedia/commons/a/a6/Çapraz_doğrulama_diyagramı.svg

42. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," J. Mach. Learn. Res., 2012.

43. G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," Netw. Model. Anal. Heal. Informatics Bioinforma., 2016, doi: 10.1007/s13721-016-0125-6.