Journal of
Information Systems Engineering
and Business Intelligence

Vol.6, No.2, October 2020
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Booking Prediction Models for Peer-to-peer Accommodation Listings using Logistics Regression, Decision Tree, K-Nearest Neighbor, and Random Forest Classifiers

Mochammad Agus Afrianto[1)]*, Meditya Wasesa[2)]

[1)2)] School of Business and Management, Institut Teknologi Bandung, Indonesia
Jl. Ganesa 10, Bandung
[1)] mochammad-agus@sbm-itb.ac.id, [2)] meditya.wasesa@sbm-itb.ac.id

Article history:

Received 23 June 2020
Revised 28 August 2020
Accepted 10 September 2020
Available online 28 October 2020

Keywords:

Airbnb
Classification
Machine Learning
Peer-to-peer Accommodation
Prediction
Revenue Management

Abstract

Background: Literature in the peer-to-peer accommodation has put a substantial focus on accommodation listings' price determinants. Developing prediction models related to the demand for accommodation listings is vital in revenue management because accurate price and demand forecasts will help determine the best revenue management responses.
Objective: This study aims to develop prediction models to determine the booking likelihood of accommodation listings.
Methods: Using an Airbnb dataset, we developed four machine learning models, namely Logistics Regression, Decision Tree, K-Nearest Neighbor (KNN), and Random Forest Classifiers. We assessed the models using the AUC-ROC score and the model development time by using the ten-fold three-way split and the ten-fold cross-validation procedures.
Results: In terms of average AUC-ROC score, the Random Forest Classifiers outperformed other evaluated models. In three-ways split procedure, it had a 15.03% higher AUC-ROC score than Decision Tree, 2.93 % higher than KNN, and 2.38% higher than Logistics Regression. In the cross-validation procedure, it has a 26,99% higher AUC-ROC score than Decision Tree, 4.41 % higher than KNN, and 3.31% higher than Logistics Regression. It should be noted that the Decision Tree model has the lowest AUC-ROC score, but it has the smallest model development time.
Conclusion: The performance of random forest models in predicting booking likelihood of accommodation listings is the most superior. The model can be used by peer-to-peer accommodation owners to improve their revenue management responses.

I. INTRODUCTION

Revenue management refers to the process of organizing or controlling prices and supplies to maximize revenue [1] by matching the right product to the right customer at the right time [2]. Demand level prediction can be set so that prices will be accepted by customers who are sensitive or insensitive to prices at a particular time [3]. Effectiveness can be achieved when the operation considers aspects like relatively fixed capacity; variable and uncertain demand; perishable inventory; a high fixed cost structure, and varying customer price sensitivity [4]. Revenue management has been applied in various industries like the airline, automobile rental, broadcasting, cruise lines, the Internet service provision, lodging and hospitality, and passenger railways; and even in the non-profit sector [5].

Over sixty percent of research on revenue management focus on the hotel business contexts [6]. In the new peer-to-peer business models that use electronic platforms to connect landlords and guests such as Airbnb, the use of revenue management would help landlords to increase their revenue [7].

Two common strategies to increase revenue are the pricing and the non-pricing strategy [8]. The former includes demand-based pricing, which could be efficient in providing competitive advantage in the market. However, it relies on the accuracy of a demand prediction [9]. A non-pricing strategy includes capacity management. Market demand forecast would help the decision-makers determine the allocation of capacity, whether to sell now or later, depending on their decision-making rules and the estimation of costumers' willingness-to-pay [10].

[*] Corresponding author

doi: http://dx.doi.org/10.20473/jisebi.6.2.123-132

In popular peer-to-peer accommodation locations, the prices of accommodation during the holiday season, such as during the new year's holiday, are always higher than in other periods. This is common because capacity remains the same while demand increases, so property managers search strategy to maximize the revenue by increasing the prices [11]. Many famous cities, such as London, implement a policy on short term rental to protect the availability of housing for long-term residents. The policy stated that the host listings could only rent their property to guests for no more than 90 nights in a year [12].

Previous studies in peer-to-peer accommodation business context have explored various dimensions of the pricing issues [13]–[16]. To the best of our knowledge, articles exploring the demand for Airbnb listings are still lacking. Therefore, this study aims to develop prediction models for the booking likelihood. The findings can help accommodation hosts determine profitable pricing and the capacity strategies. To develop the prediction models, machine learning techniques were used, namely Logistics Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest.

II. Literature Review

Previous studies in peer-to-peer accommodation business have explored various dimensions of the pricing issues [13]–[16]. Many of them use the Airbnb business context as a case study. Exploring price determinants using the hedonic pricing model is popular in the property market [17]. Later, this method is also applied in Airbnb listings business context to find the price determinants [18]–[20]. Variables related to the price include the environment, the social aspect, the accessibility, and the spillover impacts [21]. A study from [22] involved twelve countries in the Caribbean and macro-financial data. A study from [23] involved eleven cities in the US and focused on how 137 amenities factors influence pricing. The characteristic of the city may also be used as price determinants [24]. Another study comparing Airbnb listings characteristics explain how pricing was different between urban city and sun-beach holiday destinations [25]. The study about price determinants of Airbnb listings also uses the market demand to explain the price [26]. Another study uses text data from the guest's reviews to know the guests' sentiments [27].

A study using sequential Bayesian [28] aims to understand the booking probability of listings and to know the posterior distribution. Demand forecast is an essential part of revenue management [29] because it maximizes revenue gain [30]. In a restaurant business, demand forecast figures can minimize operating costs [31]. Revenue optimization measures should be implemented after establishing an accurate demand forecasting system [32].

The type of prediction model influences the forecast accuracy. In the revenue management context, there are three different models for forecasting the booking process [33], i.e., (1) historical booking models that focus on the total booking figures, (2) advanced booking models that focus on elapsed reservations aspect, and (3) the combination between historical booking model and advanced booking model. The historical booking models employ same-day, last year, moving average, exponential smoothing, and other time-series forecasting methods. The advanced booking models use a classical pickup, advanced pickup, synthetic booking curve, and other time-series approaches. The combined model uses regression and weighted average of historical and advanced booking forecasts.

Table 1 gives an overview of three studies that analyze forecasting topics using time series data in the hotel industry. The first study compared different forecasting methods to predict the booking reservation and room occupancy accurately [34]. The second study used various forecasting methods and concluded that the pickup, moving average, and exponential smoothing models was the best. The third study compared different forecasting methods using hotel occupancy data (three different room types) [35]. Another analyzed the time series method [36] and used monthly observations of hotel and motel guest' nights in New Zealand. The results show that Holt-Winters method and ARMA model were better than the Box-Jenkins seasonal-autoregressive–moving-average (SARMA) model.

The target variables in the previous study were guest arrival, hotel occupancy, and duration of stay. The current study aims to predict whether an accommodation listing will be booked or not. In terms of method, Logistics Regression is relatively easy to use and does not need any hyper-parameter optimization setup. The model can also compete with more sophisticated machine-learning models [37]. A Decision Tree model is a non-parametric approach that can adapt to any kind of dataset and can deal with nonlinear relationships well [38]. KNN is a popular algorithm among the top 10 algorithms in data mining [39] due to its simplicity and significant performance [40]. Lastly, the Random Forest is an improvement of the Decision Tree by combining several Decision Trees, which then provides good predictions, and tends not to overfit because it is compatible with large numbers [41].

TABLE 1
REVIEWED STUDIES ON FORECASTING IN THE HOTEL INDUSTRY USING TIME SERIES DATA

Article	Model	Forecasted Value	Methods	Business Context
[33]	Time series	Guest's arrival	Simple exponential smoothing; moving average; linear regression; logarithmic; linear regression; additive or pickup; multiplicative; holt's double exponential smoothing	Hotel
[34]	Time series	Hotel occupancy	TBATS, DSHW, BATS, Standard HW model, Same day of the last year (SdLy), average of same-day of last three years (ASdL3y)	Hotel
[35]	Time series	Guest nights	Holt winters and Box-Jenkins ARMA SARMA	Hotel

III. METHOD

Fig. 1 shows the working framework in this study. It is adapted from the standard method to build a predictive analytics model [42]. There are five stages: collecting data; selecting relevant predictor variables; determining the potential prediction method; evaluating, validating, and selecting the best prediction model; and finally reporting the research result.

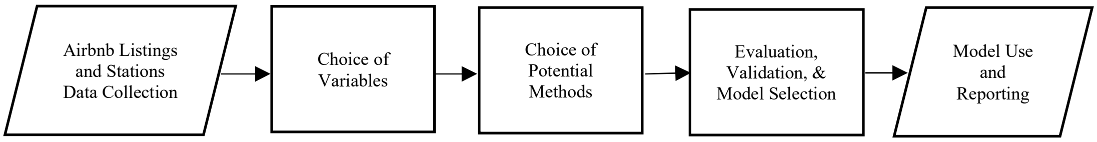
Airbnb Listings and Stations Data Collection
Choice of Variables
Choice of Potential Methods
Evaluation, Validation, & Model Selection
Model Use and Reporting

Fig. 1 The Research Framework

A. Airbnb listings data collection

In this study, we utilize available listing data from the InsideAirbnb.com platform and the doogal.co.uk [43]. InsideAirbnb.com is a website that uses a web scraping technique to gather data from the Airbnb website and it provides open data to the public. In this research, we used the Airbnb listings data from December 2018 consisting of 77,096 Airbnb listings and 96 data variables. Several listings were removed from the dataset because they indicated illogical inferences, i.e., the listings were booked for an entire year, duplicate records or missing values. The filtered dataset consists of 53,514 listings. Another dataset used was doogal.co.uk platform, which provides information about the London stations. Table 2 shows the descriptive statistics of the datasets.

TABLE 2
DESCRIPTIVE STATISTICS OF THE DATASETS

Data	Attributes	Statistics	Raw Data	Filtered Data
Airbnb listing data	Number of Records	Count	77.096	53.514
	Number of Property listing	Count	77.096	53.514
	Number of Hosts	Count	50.098	31.925
Airbnb calendar data	Number of Records	Count	28.139.675	53.514
	Period	Min	12/7/2018	12/31/2018
		Max	12/31/2018	12/31/2018
Station data	Number of stations	Count	652	610

B. Choice of variables

The predictor variables were based on the findings of previous studies [28] [7] but five predictor variables, namely the number of neighboring listings, available neighboring listings, house rules, property description, and the number of listing pictures were excluded from this study because of data and computing limitations. Instead, we added other variables such as total host listings, host verifications, accommodates, the guests included, minimum nights, and maximum nights. The host total listing variable indicates if the host is professional or not. If a host has more than one listing, we categorized the host as professional [26]. The more professional the host is, the better the services. The host verification variables indicate the reliability. Tables 3 shows the detailed information on the predictor variables.

TABLE 3
CHOICE OF PREDICTOR VARIABLE

Category	Variables/ Data Types	Variable Description
Property Function	Bedrooms/Integer	The quantity of bedrooms for listing
	Bed/Integer	The quantity of bed per bedroom for listing
	Bathroom/Double	The quantity of bathroom for listing
	Amenities/Integer	The quantity of amenities for listing
	Internet/Integer	1 means with internet and 0 without internet
	Kitchen/Integer	1 means with kitchen and 0 without a kitchen
	Experiences offered /Integer	Family, business, romantic, social, or standard
	Property type/Integer	Apartment, host and townhouse, B&B-guesthouse-hostel, or others
	Room type/Integer	Shared room, private room, or entire home/apartment
	Bed type/Integer	Real bed, couch/futon, or others
Location	Tube station (km)/Double	Listing to the nearest tube station (distance in km)
	The city center (km)/Double	Listing to city center (distance in km)
Ease of booking	Instant reservation/Integer	1 means with instant reservation and 0 without instant reservation
	Refund policy/Integer	1 means with refund policy and 0 without a refund policy
	Guest verification required/Integer	1 means need guest verification and 0 no need guest verification
Price	Price/Double	Price for rent a listing per night
	Security deposit/Double	Security deposit to book the listing
	Cleaning fee/Double	Cleaning fee after booking the listing
	Fee for extra person (£)/Double	Fee for extra person (£)
	Weekly discount/Integer	1 means a weekly discount and 0 no weekly discount
	Monthly discount/Integer	1 means a monthly discount and 0 no monthly discount
Platform signal	Superhost/ Integer	1 means host is superhost and 0 not superhost
	Host with verified ID/Integer	1 means with host verified ID and 0 hosts not verified
Host Signal	Description/Integer	1 means the description of space and 0 no description
	Host profile/Integer	1 means with host profile and 0 no-host profile
	Host total listings/Integer	The total listings per host.
	Host verification/Integer	Number of sources that is verified by Airbnb, such as email, identity card, and so on.
Peer-guest signal	Reviews/Integer	Number of reviews for listing
	Scores rating/Double	Total score from guests for listing
	Review per month/Double	Review per month for a specific listing
	Months since last review/Double	Months since the last review for listing
Host's rules	Accommodates/Integer	The number of guests that are allowed to stay
	Guest Included/Integer	The number of guests from guests that are allowed to stay
	Minimum nights/Integer	The minimum number of stays for guest
	Maximum nights/Integer	The maximum number of stays for guest
Availability	Booking status/Integer	1 means listing is booked, and 0 means listing available

C. Choice of potential methods

The focus of this research is to develop prediction models with binary classification that can give accurate predictions on whether an Airbnb listing will be booked or not. Table 4 shows the prediction models employed in this study. Looking at the number of subordinate models in a single machine learning model, we investigate both ensemble models and singular models. In general, ensemble models predict more accurately than singular models [44]. However, this research still investigates the application of singular models due to their simplicity and implementation easiness. Singular models can still outperform ensemble models [37]. We used Logistic Regression, K-Nearest Neighbors, and Decision Trees/Classification and Regression Tree (CART). In the ensemble group, we used Random Forest approach.

TABLE 4
PREDICTION MODELS CHOICE

Classifiers Category	Classifiers by Groups	Model
Ensemble	Parallel/ Bagging	Random Forest
Singular	Regression	Logistic regression
	Distance	K-Nearest Neighbors (KNN)
	Trees	Decision Trees and Regression Trees

C.1. Logistic Regression

Commonly, Logistic Regression is used to describe and test the hypotheses [45]. Choosing the right variables and avoiding the highly correlated variables must be observed when using Logistic Regression [46]. The predictor variables in Logistic Regression can be categorical or numerical; and the target variable of Logistic Regression is binary or dichotomous. Therefore, Logistic Regression cannot predict target variables of more than two classes. Although Logistic Regression may have several weaknesses, it can often compete with other machine learning techniques, such as neural networks, support vector machine, random forest, and gradient boosting [37]. The formalization of logistic regression is stated as follows [45]:

$$\pi = Probability(Y = outcome\ of\ interest|X = x, a\ specific\ value\ of\ X) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}} \quad (1)$$

where:

- π is the probability of the outcome of interest
- e is 2.71828 (the base of the system of natural logarithms)
- α is Y intercept
- β_n is the regression coefficients
- x_n is set of predictor variables

C.2. Decision Trees/Classification and Regression tree (CART)

CART can solve a classification problem. Like its name, CART algorithm looks like a tree structure. It has a root node, leaf nodes, and branches; and several advantages, such as nonparametric, adaptive with any dataset, and can deal with non-linear relationship [38]. CART is an algorithm used in a decision tree [47] and it uses the Gini index to evaluate the split. The best score is 0, and the worst score is an equal value for each class. The formalization of the Gini index is stated as follows [48]:

$$i(t) = \sum_{i=1}^{k} \sum_{j=1,j\neq i}^{k} p(c_i|t)p(c_j|t) = 1 - \sum_{j=i}^{k} (p(c_j|t))^2 \quad (2)$$

where:

- $i(t)$ is the estimated probability of misclassification under the Gini Index
- c_j is the classes j
- c_i is the classes i
- P is probability
- t is node
- k is classes

C.3. K-Nearest Neighbors (KNN)

KNN calculates the distance between samples and determines the class for each value. It has three essential parts; first, a collection of labelled objects; second, a distance between objects; third, the number of nearest neighbors. The formalization of the KNN classification (Euclidean distance) is stated as follows [49]:

$$\mathrm{Dist(X,Y)} = \sqrt{\sum_{i=1}^{D} (Xi - Yi)^2} \quad (3)$$

where:

- X is class for not booked
- Y is class for booked
- Xi (i=1….N) is an attribute of sample instance X
- Yi (i=1….N) is an attribute of sample instance Y
- D is the distance for the nearest neighbors

C.4. Random Forest

The ensemble method uses a Random Forest for classifiers, which consists of Decision Trees that are formed randomly and independently from the sampled dataset. It uses the law of large numbers, so it does not overfit and can be good for prediction [41]. Furthermore, it can be used for any dataset because it does not need a distribution assumption [38] but the weaknesses is that can be biased because the samples consist a different composition of the label prediction [50]. The formalization of the random forest classifier is stated as follows [51]:

$$\hat{Y}_i = mode_{n=1....N_{trees}}\hat{Y}_n \tag{4}$$

where:

- $\hat{Y}_i$ is the score of Random Forest
- N_{trees} is the total number of trees used in the Random Forest
- $\hat{Y}_n$ is the score of a single tree
- $mode$ is the score that most often occur

D. Evaluation, Validation and Model Selection

To assess the prediction performance of the models, we used two different evaluation methods, namely the ten-folds three-way split and ten-fold cross-validation procedures [52]. In the ten-fold three-way data split procedure, we did two data groupings. For the first grouping, we divided the dataset into ten equal sections/folds. The dataset was split into ten folds, and were not equally divided. From a total of 53,514 records, we grouped the dataset for fold number one until fold number nine consisting of 5,352 records. Fold number ten consists of 5,346 records. The second grouping was more functional.

First, the training set was used to fit the data points with the proposed model. Second, the validation set was used to evaluate the most accurate model trained in the training set. Third, the testing set was used to generate the final prediction score for each generated model. The number of data records utilized in the training, validation, and testing sets was adjusted based on the fold number category. If the testing was set to fold number ten (5,345 records), the training set consisted of 42,816 records (5,352 x 8 folds) and the validation set consisted of 5,352 records. If the testing was not set to fold number ten (5,352 records), the training set consists of 42,810 records (5,352 x 7 folds + 5,346 records from previously fold number ten) and the validation set consisted of 5,352 records. In total, there were 90 testing combinations.

In the second procedure, the ten-fold cross-validation, we divided the data to be nine-folds for training and one-fold for testing. In total, there were ten testing combinations. The prediction score of the evaluated models using the ten-fold three-way split and ten-fold cross-validation procedures were compared. The model with the highest prediction score was selected. In this study, the receiver operating characteristics (ROC) or simply AUC value was used to determine the prediction score because it was better than accuracy [53]. Mathematically, we formalize the AUC score as follows:

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \tag{5}$$

where:

- $\hat{A}$ is the AUC score
- n_0 is the number of negative class
- n_1 is the number of positive class
- S_0 is $\sum r_i$ and r_i is the rank of the i-th positive example in the ranked list

E. Model Use and Reporting

The performance of each model in terms of model development time and prediction score were compared. The best prediction model with the best prediction AUC-ROC score was selected and used to help decision-makers to formulate their corresponding revenue management response in a better way.

IV. Results

Table 5 and Table 6 show the evaluation results of the constructed machine learning classification models. Supervised machine learning models constructed the model automatically from the training dataset. Through its learning algorithm, it tried to identify and construct a generalizable pattern that reflected the relationship between the dependent (target) and independent variables. Based on the constructed pattern, the model then could build predictions on the target variable based on the observed independent variables.

To test the accuracy of the model, the prediction results of the constructed model was then compared with the actual value of the target variable. If the target variable was categorical, the AUC-ROC score was commonly used to evaluate how good the model could differentiate among different categorical variables (classes). The AUC-ROC score is a simple evaluation method [54] deemed better than the prediction accuracy score as an evaluation

TABLE 5
EVALUATION RESULT OF TEN-FOLD THREE-WAY SPLIT PROCEDURE

	Logistic Regression						Decision trees						K Nearest Neighbors						Random Forest					
	Not-Standardized Data			Standardized Data			Not-Standardized Data			Standardized Data			Not-Standardized Data			Standardized Data			Not-Standardized Data			Standardized Data		
Fold	VF	MVS	TS	VF	MVS	TS	VF	MVS	TS	VF	MVS	TS	VF	MVS	TS	VF	MVS	TS	VF	MVS	TS	VF	MVS	TS
1	2	0.790	0.760	2	0.799	0.770	1	0.628	0.604	1	0.627	0.604	2	0.739	0.707	2	0.793	0.759	2	0.824	0.785	2	0.824	0.785
2	2	0.778	0.761	2	0.799	0.778	2	0.636	0.639	2	0.636	0.638	2	0.736	0.722	2	0.791	0.760	2	0.822	0.794	2	0.822	0.794
3	2	0.794	0.778	2	0.799	0.783	1	0.629	0.626	2	0.628	0.630	2	0.740	0.723	2	0.790	0.770	2	0.817	0.804	2	0.817	0.804
4	2	0.778	0.794	2	0.783	0.799	1	0.639	0.636	1	0.638	0.636	2	0.723	0.740	2	0.770	0.790	2	0.804	0.817	2	0.804	0.817
5	3	0.794	0.752	3	0.799	0.766	2	0.629	0.621	3	0.630	0.610	3	0.734	0.710	3	0.790	0.754	3	0.821	0.788	3	0.821	0.788
6	3	0.793	0.741	3	0.800	0.748	3	0.631	0.607	3	0.632	0.608	3	0.741	0.715	3	0.788	0.749	3	0.823	0.774	3	0.823	0.774
7	3	0.786	0.708	3	0.799	0.741	3	0.650	0.603	3	0.649	0.604	3	0.742	0.682	3	0.793	0.738	3	0.826	0.765	3	0.826	0.765
8	3	0.787	0.717	3	0.800	0.752	3	0.643	0.602	3	0.643	0.601	3	0.743	0.685	3	0.791	0.749	3	0.823	0.794	3	0.824	0.794
9	3	0.786	0.704	3	0.800	0.741	3	0.643	0.610	3	0.643	0.610	3	0.742	0.669	3	0.794	0.741	3	0.828	0.769	3	0.828	0.769
10	3	0.785	0.640	3	0.802	0.676	3	0.651	0.585	3	0.651	0.584	3	0.747	0.609	3	0.796	0.659	3	0.823	0.706	3	0.823	0.705
AVG		0.775	0.739		0.781	0.755		0.699	0.672		0.699	0.672		0.758	0.725		0.779	0.751		0.803	0.773		0.803	0.773
STD		0.006	0.040		0.004	0.027		0.008	0.021		0.008	0.021		0.003	0.045		0.003	0.032		0.003	0.028		0.003	0.028
SPT		2.8			2.63			1.35			1.33			13.08			101.3			13.817			17.53	

VF = Validation Fold, MVS = Maximum Validation AUC-ROC Score, TS = Test AUC-ROC Score, AVG = Average AUC-ROC Score, STD = Standard Deviation, SPT = Sum of Processing Time in Minutes

TABLE 6
EVALUATION RESULT OF 10-FOLD CROSS VALIDATION PROCEDURE

	Logistic Regression		Decision tree		K Nearest Neighbors		Random Forest	
Fold	Not-Standardized Data	Standardized Data	Not-Standardized Data	Standardized Data	Not-Standardized Data	Standardized Data	Not-Standardized Data	Standardized Data
1	0.748	0.771	0.609	0.609	0.714	0.756	0.780	0.780
2	0.750	0.779	0.620	0.619	0.727	0.764	0.800	0.799
3	0.759	0.784	0.630	0.630	0.729	0.772	0.810	0.810
4	0.782	0.800	0.652	0.652	0.740	0.793	0.826	0.826
5	0.732	0.767	0.621	0.620	0.718	0.759	0.791	0.791
6	0.714	0.748	0.611	0.611	0.713	0.751	0.777	0.777
7	0.648	0.741	0.613	0.612	0.684	0.738	0.764	0.765
8	0.681	0.752	0.617	0.617	0.687	0.753	0.795	0.795
9	0.669	0.741	0.600	0.600	0.675	0.739	0.767	0.767
10	0.626	0.675	0.580	0.580	0.606	0.658	0.703	0.703
Average AUC-ROC	0.711	0.756	0.615	0.615	0.699	0.748	0.781	0.781
Standard Deviation	0.052	0.034	0.019	0.019	0.039	0.036	0.034	0.034
Model Development Time (in seconds)	19.00	16.00	10.00	10.00	53.00	376.00	112.00	121.00

method [53]. Logistic Regression, Decision Trees, KNN, and Random Forest methods are evaluated with ten-fold three-way split and ten-fold cross-validation procedures.

Tables 5 shows the results for the three-way split and the comparison between data with and without the data standardization process. Data standardization increases the average and decreases the standard deviation prediction of the AUC-ROC scores of the logistic regression. KNN modelled both in the validation and testing conditions. However, the data standardization process did not increase the prediction AUC-ROC scores of the decision tree and random forest models. The decision tree had the lowest AUC-ROC score, but it had the fastest model development time. Furthermore, Random Forest classifier had a 15.03% higher AUC-ROC score than decision tree, 2.93 % higher than KNN, and 2.38% higher than Logistics Regression. Therefore, using ten-fold three-way split procedure, we concluded that Random Forest performed best.

The results for the ten-fold cross-validation procedure are shown in Table 6. Fold column shows the sequencing fold, and the rest of the columns show the score for each technique. Lastly, the average score and standard deviation are at the bottom of the table. The Logistics Regression and KNN yielded a better AUC-ROC score after undergoing a data standardization process. The Decision Tree yielded the fastest processing time, but it yielded the lowest score. Random Forest classifier yielded a 26,99% higher AUC-ROC score than Decision Tree, 4.41% higher than KNN, and 3.31% higher than the Logistics Regression models. Therefore, using a ten-fold cross-validation procedure, we also concluded that the Random Forest performs best.

V. Discussion

From the average AUC-ROC score, Random Forest models performed superior in both evaluation procedures. The random forest models reach 0.773 average AUC-ROC scores in a ten-fold three-way split condition and reach 0.781 in the ten-fold cross-validation condition. From the category of the classifiers, the ensemble methods outperformed the singular methods, which means the ensemble methods was better than singular methods in dealing with bias, noise, and variance. The Decision Tree or single tree model yielded the lowest score because of the inaccuracy [55], but it had the fastest processing time because of their simplicity.

The data standardization process increased the AUC-ROC score of Logistics Regression and KNN. Interestingly, the processing time after data standardization was reduced in Logistic Regression models. It was affected by outliers, and data standardization could handle the negativity of outlier cases. That was why the score increased and the model development time decreased in Logistics Regression models. Furthermore, the highest AUC-ROC score of the random forest model was in line with the findings of the earlier study [41]. There were many advantages of the Random Forest model, such as being able to handle outliers and noise.

The evaluation performance methods showed different average AUC-ROC scores in the same model. The Random Forest and Logistics Regression yielded a higher average score in the ten-fold cross-validation method rather than in a ten-fold three-way split procedure. However, the other models produced a higher AUC-ROC score in the three-ways split procedure. Besides, the difference in the average AUC-ROC score among the models was higher in cross-validation methods rather than in three-ways split procedure. It means that the number of training set affects the testing score of each model.

VI. Conclusion

Considering the importance of demand forecasts in the revenue management context, this study analyses four machine learning techniques to predict the booking likelihood of accommodation listings. We evaluated the AUC-ROC score of each model using two different evaluation methods, i.e., the ten-fold three-way split and the ten-fold cross-validation procedures.

In terms of the AUC-ROC score, the Random Forest classifiers outperformed the other models, i.e., Logistics Regression, Decision Tree, and K-Nearest Neighbor. The Decision Tree model had the lowest AUC-ROC score, but it had the lowest processing time. The performance of random forest models in predicting the booking likelihood of accommodation listings is the most superior. The findings can inform peer-to-peer accommodation owners to improve their predictions and the revenue management responses. In terms of contribution to literature, this study informs the prediction method of the booking likelihood.

References

[1] Q. Meng, H. Zhao, and Y. Wang, “Revenue Management for Container Liner Shipping Services: Critical Review and Future Research Directions,” *Transp. Res. Part E Logist. Transp. Rev.*, vol. 128, no. June, pp. 280–292, 2019, doi: 10.1016/j.tre.2019.06.010.

[2] S. E. Kimes, “The Basics of Yield Management,” *Cornell Hotel Restaur. Adm. Q.*, vol. 30, no. 3, pp. 14–19, 1989, doi: 10.1177/001088048903000309.

[3] S. E. Kimes and K. A. McGuire, "Function-Space Revenue Management: A Case Study from Singapore," *Cornell Hotel Restaur. Adm. Q.*, no. December, pp. 33–46, 2001.

[4] S. E. Kimes and J. Wirtz, "Has Revenue Management Become Acceptable?: Findings From an International Study on the Perceived Fairness of Rate Fences," *J. Serv. Res.*, vol. 6, no. 2, pp. 125–135, 2003, doi: 10.1177/1094670503257038.

[5] J. I. McGill and G. J. Van Ryzin, "Revenue Management: Research Overview and Prospects," *Transp. Sci.*, vol. 33, no. 2, pp. 233–256, 1999, doi: 10.1287/trsc.33.2.233.

[6] B. Denizci Guillet and I. Mohammed, "Revenue Management Research in Hospitality and Tourism," *Int. J. Contemp. Hosp. Manag.*, vol. 27, no. 4, pp. 526–560, May 2015, doi: 10.1108/IJCHM-06-2014-0295.

[7] Y. Koh, A. Belarmino, and M. G. Kim, "Good Fences Make Good Revenue: An Examination of Revenue Management Practices at Peer to-Peer Accommodations," Tour. Econ., 2019, doi: 10.1177/1354816619867579.

[8] S. H. Ivanov, *Hotel Revenue Management: From Theory to Practice*, vol. 00, no. October. 2014.

[9] G. K. Nair, "Dynamics of Pricing and Non-Pricing Strategies, Revenue Management Performance and Competitive Advantage in Hotel Industry," *Int. J. Hosp. Manag.*, vol. 82, no. June, pp. 287–297, 2019, doi: 10.1016/j.ijhm.2018.10.007.

[10] C. C. Queenan, M. E. Ferguson, and J. K. Stratman, "Revenue Management Performance Drivers: An Exploratory Analysis Within the Hotel Industry," *J. Revenue Pricing Manag.*, vol. 10, no. 2, pp. 172–188, 2011, doi: 10.1057/rpm.2009.31.

[11] G. Bitran and R. Caldentey, "Commissioned Paper An Overview of Pricing Models for Revenue Management," vol. 5, no. 3, pp. 203–229, 2003.

[12] "Short Term and Holiday Lets in London - London.gov.uk." [Online]. Available: https://www.london.gov.uk/what-we-do/housing-and-land/improving-private-rented-sector/short-term-and-holiday-lets-london.

[13] L. (Rebecca) Tang, J. Kim, and X. Wang, "Estimating Spatial Effects on Peer-to-Peer Accommodation Prices: Towards an Innovative Hedonic Model Approach," *Int. J. Hosp. Manag.*, vol. 81, no. August 2018, pp. 43–53, 2019, doi: 10.1016/j.ijhm.2019.03.012.

[14] R. Deboosere *et al.*, "Regional Studies , Regional Science Location , Location and Professionalization : a Multilevel Hedonic Analysis of Airbnb Listing Prices and Revenue Location, Location and Professionalization : Prices and Revenue b," vol. 1376, 2019, doi: 10.1080/21681376.2019.1592699.

[15] B. Tong and U. Gunter, "Current Issues in Tourism Hedonic Pricing and the Sharing Economy : How Profile Characteristics Affect Airbnb Accommodation Prices in Barcelona , Madrid , and Seville," *Curr. Issues Tour.*, vol. 0, no. 0, pp. 1–20, 2020, doi: 10.1080/13683500.2020.1718619.

[16] P. Arvanitidis, A. Economou, and G. Grigoriou, "Current Issues in Tourism Trust in Peers or in the Institution ? A Decomposition Analysis of Airbnb Listings ' Pricing," *Curr. Issues Tour.*, vol. 0, no. 0, pp. 1–18, 2020, doi: 10.1080/13683500.2020.1806794.

[17] H. Selim, "Determinants of House Prices in Turkey : Hedonic Regression Versus Artificial Neural Network," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2843–2852, 2009, doi: 10.1016/j.eswa.2008.01.044.

[18] Y. Chen and K. Xie, "Consumer Valuation of Airbnb Listings: A Hedonic Pricing Approach," *Int. J. Contemp. Hosp. Manag.*, vol. 29, no. 9, pp. 2405–2424, Sep. 2017, doi: 10.1108/IJCHM-10-2016-0606.

[19] Z. Zhang, R. J. C. Chen, L. D. Han, and L. Yang, "Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach," *Sustain.*, vol. 9, no. 9, pp. 1–13, 2017, doi: 10.3390/su9091635.

[20] D. Wang and J. L. Nicolau, "Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com," *Int. J. Hosp. Manag.*, vol. 62, pp. 120–131, 2017, doi: 10.1016/j.ijhm.2016.12.007.

[21] J. Chica-Olmo, J. G. González-Morales, and J. L. Zafra-Gómez, "Effects of Location on Airbnb Apartment Pricing in Málaga," *Tour. Manag.*, vol. 77, no. March 2019, p. 103981, 2020, doi: 10.1016/j.tourman.2019.103981.

[22] T. Lorde, J. Jacob, and Q. Weekes, "Price-Setting Behavior in a Tourism Sharing Economy Accommodation Market : A Hedonic Price Analysis of AirBnB Hosts in the Caribbean ☆," *Tour. Manag. Perspect.*, vol. 30, no. February, pp. 251–261, 2019, doi: 10.1016/j.tmp.2019.03.006.

[23] M. Chattopadhyay and S. K. Mitra, "Do Airbnb Host Listing Attributes Influence Room Pricing Homogenously?," *Int. J. Hosp. Manag.*, vol. 81, no. September 2018, pp. 54–64, 2019, doi: 10.1016/j.ijhm.2019.03.008.

[24] C. Gibbs, D. Guttentag, U. Gretzel, and J. Morton, "Pricing in the Sharing Economy : A Hedonic Pricing Model Applied to Airbnb Listings," *J. Travel Tour. Mark.*, vol. 35, no. 1, pp. 46–56, 2018, doi: 10.1080/10548408.2017.1308292.

[25] L. Moreno-Izquierdo, A. B. Ramón-Rodríguez, M. J. Such-Devesa, and J. F. Perles-Ribes, "Tourist Environment and Online Reputation as A Generator of Added Value in the Sharing Economy: The case of Airbnb in urban and sun- and-beach holiday destinations," *J. Destin. Mark. Manag.*, vol. 11, no. January 2018, pp. 53–66, 2019, doi: 10.1016/j.jdmm.2018.11.004.

[26] F. Magno, F. Cassia, and M. M. Ugolini, "Accommodation Prices on Airbnb: Effects of Host Experience and Market Demand," *TQM J.*, vol. 30, no. 5, pp. 608–620, 2018, doi: 10.1108/TQM-12-2017-0164.

[27] A. Lawani, M. R. Reed, T. Mark, and Y. Zheng, "Reviews and Price on Online Platforms: Evidence from Sentiment Analysis of Airbnb Reviews in Boston," *Reg. Sci. Urban Econ.*, vol. 75, pp. 22–34, 2019, doi: 10.1016/j.regsciurbeco.2018.11.003.

[28] B. Yao, R. T. R. Qiu, D. X. F. Fan, A. Liu, and D. Buhalis, "Standing Out from the Crowd – An Exploration of Signal Attributes of Airbnb Listings," *Int. J. Contemp. Hosp. Manag.*, vol. 31, no. 12, pp. 4520–4542, 2019, doi: 10.1108/IJCHM-02-2019-0106.

[29] C. Cleophas, M. Frank, and N. Kliewer, "Recent Developments in Demand Forecasting for Airline Revenue Management," *Int. J. Revenue Manag.*, vol. 3, no. 3, pp. 252–269, 2009, doi: 10.1504/IJRM.2009.027386.

[30] R. CROSS, "Launching the Revenue Rocket How Revenue Management Can Work for Your Business," *Cornell Hotel Restaur. Adm. Q.*, vol. 38, no. 2, pp. 32–43, Apr. 1997, doi: 10.1016/S0010-8804(97)81474-7.

[31] J. Shields, "Restaurant Revenue Management: An Investigation Into Changing Standard Operating Procedurs to Maximize Revenue," *J. Small Bus. Strateg.*, p. 77, 2006.

[32] A. Lahoti, "Why CEOs should Care About Revenue Management.," 2002. [Online]. Available: https://www.informs.org/ORMS-Today/Archived-Issues/2002/orms-2-02/Why-CEOs-Should-Care-About-Revenue-Management.

[33] A. O. Lee, "Airline Reservations Forecasting: Probabilistic and Statistical Models of the Booking Process," no. September 1990, p. 266, 1990.

[34] L. R. Weatherford and S. E. Kimes, "A Comparison of Forecasting Methods for Hotel Revenue Management," *Int. J. Forecast.*, vol. 19, no. 3, pp. 401–415, 2003, doi: 10.1016/S0169-2070(02)00011-0.

[35] L. N. Pereira, "An Introduction to Helpful Forecasting Methods For Hotel Revenue Management," *Int. J. Hosp. Manag.*, vol. 58, pp. 13–23, 2016, doi: 10.1016/j.ijhm.2016.07.003.

[36] C. Lim, C. Chang, and M. McAleer, “Forecasting h(m)otel Guest Nights in New Zealand,” *Int. J. Hosp. Manag.*, vol. 28, no. 2, pp. 228–235, 2009, doi: 10.1016/j.ijhm.2008.08.001.

[37] S. Nusinovici *et al.*, “Logistic Regression Was as Good as Machine Learning for Predicting Major Chronic Diseases,” *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[38] X. E. Pantazi, D. Moshou, and D. Bochtis, *Artificial intelligence in agriculture*. 2020.

[39] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[40] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN Classification with Different Numbers of Nearest Neighbors,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.

[41] L. Breiman, “Random Forest,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[42] G. Shmueli and O. R. Koppius, “Predictive Analytics in Information Systems Research,” *MIS Q. Manag. Inf. Syst.*, vol. 35, no. 3, pp. 553–572, 2011, doi: 10.2307/23042796.

[43] C. Bell, “London stations.” [Online]. Available: https://www.doogal.co.uk/london_stations.php. [Accessed: 30-Apr-2020].

[44] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Emperical Study,” J. Artif. Intell. Res., vol. 1, 1999, doi: https://doi.org/10.1613/jair.614.

[45] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, “An Introduction to Logistic Regression Analysis and Reporting,” J. Educ. Res., vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[46] C. S. P. Priya Ranganathan and R. Aggarwal, “Common Pitfalls in Statistical Analysis: Logistic Regression,” Perspect. Clin. Res., vol. 10, no. 2, pp. 51–56, 2017, doi: 10.4103/picr.PICR.

[47] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and regression trees. 1999.

[48] L. E. Raileanu and K. Stoffel, “Theoretical Comparison Between the Gini Index and Information Gain Criteria ∗,” no. 2100, pp. 77–93, 2004.

[49] S. Zhang, “Cost-Sensitive KNN Classification,” Neurocomputing, vol. 391, no. xxxx, pp. 234–242, 2020, doi: 10.1016/j.neucom.2018.11.101.

[50] S. J. Winham, R. R. Freimuth, and J. M. Biernacka, “A Weighted Random Forests Approach to Improve Predictive Performance,” Stat. Anal. Data Min., 2013, doi: https://doi.org/10.1002/sam.11196.

[51] E. Izquierdo-Verdiguier and R. Zurita-Milla, “An Evaluation of Guided Regularized Random Forest for Classification and Regression Tasks in Remote Sensing,” Int. J. Appl. Earth Obs. Geoinf., vol. 88, no. February, p. 102051, 2020, doi: 10.1016/j.jag.2020.102051.

[52] B. Park and J. Kwon Bae, “Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data,” Expert Syst. Appl., vol. 42, no. 6, pp. 2928–2934, 2015, doi: 10.1016/j.eswa.2014.11.040.

[53] J. Huang and C. X. Ling, “Using AUC and Accuracy in Evaluating Learning Algorithms,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 3, pp. 299–310, 2005, doi: 10.1109/TKDE.2005.50.

[54] A. Geršl and M. Jašová, “Credit-Based Early Warning Indicators of Banking Crises in Emerging Markets,” Econ. Syst., vol. 42, no. 1, pp. 18–31, 2018, doi: 10.1016/j.ecosys.2017.05.004.

[55] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.2307/2699986.