Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241296

Machine learning-based hotel occupancy prediction

Jiana Peng

Faculty of Economics, Wuhan University of Technology, Wuhan, China

Pengjiana0422@whut.edu.cn

Abstract. The online booking system has tremendously facilitated people's ability to travel and make reservations because of the growth of the Internet. However, the unpredictability of travel led to frequent changes in reservations. Frequent order changes can lead to many problems, such as hotels not being able to get order changes in a timely manner and more in-demand customers not being able to stay, which can lead to lower profits and occupancy rates. In addition to this, there are a number of subjective, such as changes in the trip, reasons for work, and reasons for family, and objective, such as weather changes, natural disasters, and Transportation issues, factors that make it more difficult to predict the occupancy rate. In recent years, machine learning has become an increasingly valuable tool for researchers to analyze data. Based on these, this paper summarizes the machine learning algorithms related to hotel occupancy probability prediction and analyzes and compares them. Finally, it gives an outlook on the research of the hotel occupancy rate prediction.

Keywords: Un-subscription, Machine learning, Hotel occupancy prediction.

1. Introduction

Recent technological development enables us to reserve hotels conveniently via E-commerce websites and applications. Besides, with the development of the tourism industry, more and more people book hotels online [1]. However, it can be difficult to forecast occupancy accurately because the demand curve changes over time due to a variety of hotel attributes and outside factors like seasonality and events [2]. For merchants, not having an accurate occupancy rate will be a big risk. With access to accurate occupancy rates, merchants can better maximize their profits. After a customer has made a reservation, knowing exactly whether the customer will check in on time and whether there is a risk of backing out allows the merchant to better deal with the situation and re-sell the room in a timely manner. Estimating hotel occupancy also allows merchants to understand that the number of empty rooms sold online should be a higher percentage than the actual number of empty rooms during that time period, as a way to better hedge against the risk of default while earning higher profits. In anticipation of a high rate of hotel cancellations in the next quarter, merchants can also take measures to prevent excessive loss of profits, such as charging a deposit or increasing the percentage of cancellation fees and other methods. In addition, it also enables merchants to find the main source of customers for better maintenance and new target customer groups to better increase the new source of customers to find, and if you understand the main factors of the customer unsubscribe, it can help merchants better improve the marketing strategy to reduce the unsubscribe rate. A low un-subscription can attract more customers

DOI: 10.54254/2755-2721/46/20241296

because a survey shows t that online reviews and peer recommendations are being relied upon more and more when travelers book hotels [3].

Previously, researchers typically predicted checkout rates by observing hotel sales data over the same period or relied on previous years' hotel sales data to build predictive models to anticipate the probability of consumers checking out. Nowadays, researchers use machine learning methods to build reliable models and predict unsubscribe rates and individual user unsubscribe probabilities based on a large amount of data from previous years [4].

Before building a mathematical model, it is important to understand how to convert the real-world problem of predicting the probability of a single user checking out of a hotel into a mathematical problem so that a mathematical model can be built. Firstly, we need to know what factors (subjective and objective) will affect the customer's unsubscription, secondly, the weight of these factors in the customer's decision to unsubscribe, and finally, we need to select an appropriate and reliable model for prediction. This process involves analyzing historical data, conducting surveys or interviews with customers, and utilizing statistical techniques to establish correlations and relationships between the factors and withdrawal probability. Additionally, it is crucial to regularly update and refine the model based on new data and changing customer preferences to ensure its accuracy and effectiveness in predicting single-user withdrawal from a hotel.

For binary classification modeling, machine learning has shown quite good results in recent years [5]. In order to achieve the goal of predicting whether a single subscriber will unsubscribe, researchers usually use a large amount of previous years' subscriber data to construct a reliable regression model to predict the probability of a single subscriber's unsubscription through regression, the closer the regression result is to 1, the higher the probability of the subscriber's unsubscription, and vice versa, the lower the probability. The model generally consists of the type of room booked by the user in previous years (x1), the number of people traveling with the user (x2), the time of the trip (x3), the booking platform (x4), whether or not the hotel charges a deposit (x5), whether or not the deposit is returned (x6), whether or not the hotel provides meals (x7), and the cost of meals (x8) as the explanatory variables. All the binary variables in the model are True=1 and False=0. The model is constructed as follows:

$$P = \sum_{i=1}^{8} \alpha_i x_i \tag{1}$$

Additionally, a large number of researchers have built decision tree models using machine learning to forecast single-user unsubscribe rates [6]. Data type variables are categorized based on the data range, and categorical type variables are divided based on their own category. Through the haphazard combination of various categories of variables that can be divided into multiple decision trees, which is equivalent to the various needs of various customers, and likes and dislikes are divided into several categories, the user's likelihood of unsubscribing is then located in each category. This algorithm considers both numerical variables, such as age or the number of rooms, and categorical variables, such as customer preferences or product categories. By utilizing decision trees that consider various combinations of these variables, it becomes possible to accurately classify users into different categories based on their likelihood of unsubscribing.

However, there are still a few issues with this problem's solution that need to be fixed. Insufficient samples may skew the model construction, and difficulties with personal privacy may arise during data collection. If the grounds for unsubscribing are filled out incorrectly, the results may not be accurately assessed. In addition to this, a tour group will travel, but the room will enter the basic data of various clients so that there will be a variety of samples for analysis, which could skew the factor weights. Furthermore, the lack of transparency in the data collection process can also raise concerns about the accuracy and reliability of the collected data. These problems have not yet been resolved. Additionally, addressing these issues requires careful consideration of ethical guidelines to ensure that individuals' privacy rights are respected throughout the data collection and analysis process.

This paper starts off with an introduction to the pertinent dataset and some treatments of the pertinent variables. The introduction to the relevant dataset provides crucial background information for understanding the subsequent analysis. Then moves on to an introduction and analysis of classic hotel

DOI: 10.54254/2755-2721/46/20241296

unsubscription rate prediction models and an evaluation of the accuracy of the results predicted by these models, before concluding with a summary of the entire paper and a glance at the view.

2. Dataset

Nuno Antonio, Ana Almeida, and Luis Nunes created this hotel booking dataset for Data in Brief, Volume 22, February 2019. For "TidyTuesday during the week of February 11th, 2020," Thomas Mock and Antoine Bichat downloaded and organized the data.

2.1. Variables

One file in this dataset compares different booking details between two hotels: a city hotel and a resort hotel. This dataset has 119390 samples and 32 columns, including 'Hotel' (H1 = Resort Hotel or H2 = City Hotel), is_cancled(Value indicating whether the booking was canceled (1) or not (0)), lead_time, arrival_date_year, arrival_date_month, arrival_date_week_of_number, arrival_date_of_month, stays_in_weekend_nights(Number of weekend nights the guest stayed), stays_in_week_nights, adults(Number of adults), children, babies, meal(Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)), country, market_segment("TA" means "Travel Agents" and "TO" means "Tour Operators"), distribution_channel, is_repeated_guest, previous_cancellations(Number of previous bookings that were cancelled by the customer prior to the current booking), previous_bookings_not_canceled, reserved_room_type(Code of room type reserved), assigned_room_type, booking_changes, deposit_type(No Deposit: No deposit was made; Non Refund: A deposit equal to the complete cost of the stay was made; Refundable: A deposit equal to less than the total cost of the stay was made.), company, days_in_waiting_list(Number of days between booking day and occupancy day), customer_type, adr(To determine the average daily rate, divide the total number of accommodation transactions by the total number of nights stayed there), required_car_parking_spaces, total_of_special_requests, reservation_status(Cancelled: The consumer canceled the reservation; Check-out signifies that a customer has already checked out. No-Show: The customer failed to check in but did let the hotel know why.) and reservation_status_date(where the previous status was set for dinner).

2.2. Model Evaluation

For a binary classification model, there are many norms to evaluate the predicting ability of the model, such as accuracy rate, precision rate, recall rate, and area under the ROC curve (AUC) [7]. All these norms are such that the higher they are, the better the model. These are the results of a binary classification model and what these norms represent: as shown Table1.

Table 1. The result of Binary Classification

classification	Real value: 1	Real value: 0
Predict value: 1	TP	FP
Predict value: 0	FN	TN

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Racall = \frac{TP}{TP+FN} \tag{4}$$

2.3. Distribution of Data

There is the distribution of the explained variable 'is_cancled'. As show Fig1,

DOI: 10.54254/2755-2721/46/20241296

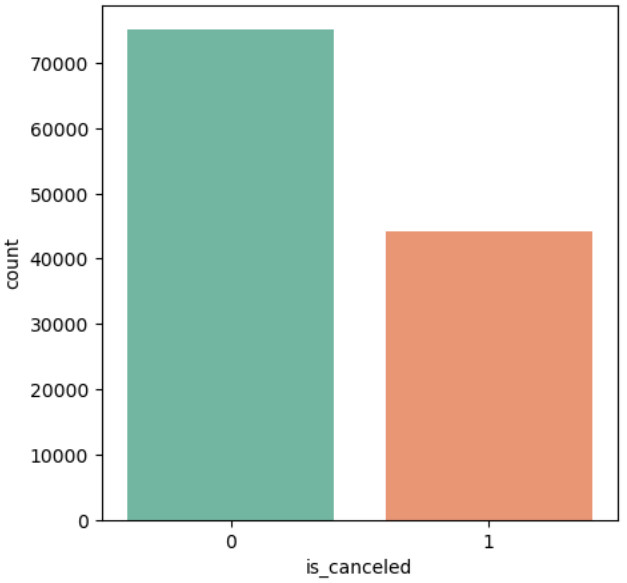

Figure 1. The Distribution of Explained Variable

This distribution demonstrates that even though there are more users who check in than those who unsubscribe, the proportion of users who unsubscribe is still significant, demonstrating the importance of accurately predicting occupancy in order to maximize resource utilization and profitability for the hotel. However, the uneven distribution of data also adds some difficulties to the subsequent model design. The uneven distribution of data poses challenges in developing a reliable predictive model as it may lead to imbalanced training and testing datasets. This necessitates the use of techniques such as oversampling or undersampling to address the issue and ensure accurate predictions. Additionally, incorporating other relevant factors like seasonality, events, and market trends can further enhance the accuracy of occupancy predictions despite the uneven data distribution.

3. Methods

3.1. Traditional Ways

It's well-known that merchants will predict the occupancy rate through priori knowledge or human empirical judgment in the past time. Through the changes in weather, the occurrence of natural disasters, different seasons, the average price per day, and the average daily price of competitor hotels to roughly estimate hotel occupancy for the coming period. While traditional methods of predicting occupancy and machine learning algorithms are very different, both take into account multiple factors to achieve the best possible prediction [8]. Internal and external factors can be categorized into two categories: economic changes, events (sports competitions, concerts, conferences), exchange rates, political situations, competitors' prices, and weather changes. Internal factors include the hotel's location and the size and number of its rooms. External factors include the price of competing hotels and weather changes. Internal factors play a crucial role in determining the prediction accuracy as they directly relate to the hotel's operations and offerings. On the other hand, external factors such as economic changes and weather conditions can greatly impact the demand and pricing of hotels, making them equally important to consider for accurate predictions. The traditional prediction methods are not rigorous, subjective and lack comparisons between the same period. However, the traditional forecasting method still breaks the limitation of people's perceptions allowing the hotel's profits to be improved as a result and the society's resources to be better utilized, which was not the case in the previous period.

In recent years, advancements in technology have paved the way for more sophisticated forecasting models that reconsider various factors such as economic indicators, social trends, and even weather

DOI: 10.54254/2755-2721/46/20241296

patterns. These new methods offer a more comprehensive and data-driven approach to predicting hotel demand and pricing, leading to more accurate forecasts. Additionally, by incorporating machine learning algorithms, these models can continuously learn and adapt to changing market conditions, further enhancing their predictive capabilities.

3.2. Machine Learning based methods

With the development of technology, using machine learning to predict the occupancy rate of hotels has become more and more mature and efficient. A survey used the XGBOOST algorithm to predict the rate, which removed the subjective elements of traditional methods and added more data to perfect the predicting model [9][10]. The XGBOOST method is a popular choice for predicting hotel occupancy rates because of its propensity to handle huge datasets and intricate relationships. By incorporating various factors such as historical booking data, weather conditions, and local events, the model can provide more accurate and reliable predictions. This advancement in technology has not only improved the efficiency of hotel management but also allowed businesses to make informed decisions regarding pricing, staffing, and marketing strategies. Using the XGBOOST algorithm for predicting the occupancy rate of this dataset showed that: A AUC is greater than 0.94, Accuracy Rate is greater than 0.87, and Precision is greater than 0.87. However, because the XGBOOST model consists of many decision trees, it is running slowly.

Logistic regression can also be used to forecast the occupancy rate. Logistic Regression is a basic, useful, and efficient model for binary classification. In contrast, using Logistic Regression to forecast occupancy rate is more efficient and allows us to clearly understand the relative importance of each element, allowing us to identify the primary cause of un-subscription. Using the Logistic Regression algorithm for predicting the occupancy rate of this dataset showed that: AUC is greater than 0.86, Accuracy Rate is greater than 0.80, and Precision is greater than 0.80. This information can help businesses make informed decisions and take necessary actions to improve occupancy rates. Additionally, logistic regression can provide insights into the factors that contribute to changes in occupancy rates over time, allowing businesses to anticipate and address potential issues before they occur. However, compared to XGBOOST model, the accuracy and other norms of the Logistic Regression are lower because the form is very simple. Additionally, the Logistic Regression model is more sensitive to multicollinear data and difficult to deal with imbalance.

Random Forest is the most recent effective algorithm used to forecast hotel occupancy rates. The Random Forest algorithm combines several sample characteristics to produce many decision trees, and it then uses the average of each decision tree value as the sample's forecast outcome. Additionally, the Random Forest algorithm's ability to handle high-dimensional data and missing values makes it a suitable choice for estimating hotel occupancy rates. Its ensemble of decision trees also helps to reduce overfitting and improve generalization performance. Therefore, considering its strong performance metrics and versatility, the Random Forest algorithm is a highly recommended approach for estimating hotel occupancy rates. This dataset's occupancy rate was predicted using the Logistic Regression technique, and the results revealed that: Precision is higher than 0.88, Accuracy Rate is higher than 0.89, and AUC is better than 0.95. Because the Random Forest algorithm's norms are so good, it can be extremely nice used to estimate the hotel occupancy rate. Although Random Forest is a robust and analytically thorough technique for binary classification, its poor performance (lower speed) is caused by the requirement to build numerous decision trees. If this problem can be improved, then the random forest algorithm will be used more widely.

Here are the predicted results of XGBOOST, Logistic Regression, and Random Forest algorithms, as Fig. 2, Table 2, Table 3, Table 4:

DOI: 10.54254/2755-2721/46/20241296

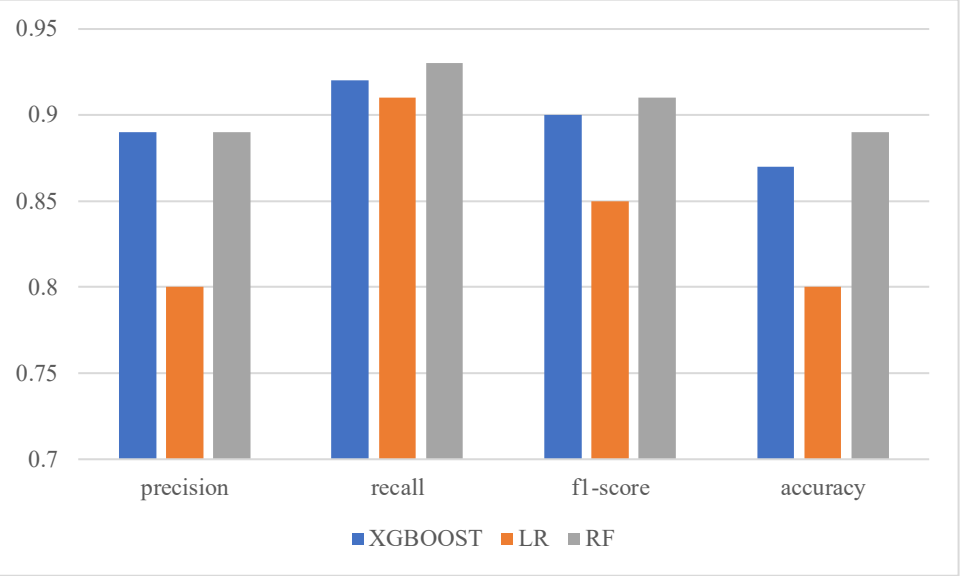

Figure 2. The results of three algorithmns predicting

Table 2. The result of XGBOOST Model

XGBOOST	precision	recall	f1-score	support
0	0.89	0.92	0.90	22684
1	0.85	0.80	0.82	13133
accuracy			0.87	35817

Table 3. The result of Logistic Regression Model

LR	precision	recall	f1-score	support
0	0.80	0.91	0.85	22684
1	0.80	0.60	0.68	13133
accuracy			0.80	35817

Table 4. The result of Random Forest Model

RF	precision	recall	f1-score	support
0	0.89	0.93	0.91	22684
1	0.87	0.81	0.84	13133
accuracy			0.89	35817

3.3. Deep Learning

Nowadays, deep learning algorithms, a new field of machine learning, can be used for binary classification modeling. Typical deep learning algorithm models are Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Stacked Auto-encoder Network [11]. Deep learning algorithms, as a human neural network-based pattern analysis method, have deeper model structures and

DOI: 10.54254/2755-2721/46/20241296

more model parameters compared to traditional methods. The use of deep learning algorithms to predict hotel occupancy not only provides new ideas for making that prediction but also has the potential to make that prediction more accurate, further improving hotel profitability and resource utilization. Furthermore, deep learning algorithms can automatically extract high-level features from unprocessed data, doing away with the requirement for manual feature engineering. In addition to saving time and effort, doing so enables a more thorough and sophisticated investigation of hotel occupancy patterns. Furthermore, the scalability of deep learning models makes them suitable for handling large datasets, enabling more accurate predictions based on a wide range of factors such as seasonality, pricing, and customer preferences.

4. Conclusion

This study examines machine learning-based hotel occupancy prediction, discusses the importance of hotel occupancy prediction, the relevant variables that can affect the magnitude of the occupancy probability, how to transform this practical issue into a mathematical issue, how to make a prediction, and discusses the pertinent datasets and machine learning algorithms. Then, by contrasting the conventional techniques with the machine learning algorithms and the three machine learning algorithms with one another, a thorough comprehension of the development of this predictive model is offered. Additionally, it was claimed that deeper predictions for better outcomes can be made using deep learning algorithms.

In introducing the three algorithms of machine learning (XGBOOST, Logistic Regression, Random Forest), this paper demonstrates and assesses the quality of t the models by comparing the metrics of accuracy, precision, recall, and AUC. Among them, Random Forest and XGBOOST performed well on this dataset, with 89% accuracy and 95% AUC for Random Forest, the model can then be enhanced by modifying the parameters and using additional techniques. However, because the random forest model needs to create a lot of decision trees, there is a subsequent need to reduce the algorithm's running time. Besides, to further analyze hotel occupancy rates, the deep learning algorithm can be used to predict.

It is feasible to see that there are several issues with the present approaches being more accurate after comprehending the significance of projecting hotel occupancy and discussing some of the methods utilized for this purpose. A few issues that need to be highlighted and attempted to be resolved in future research include the dearth of available data, the data's imperfect quality, such as relevant factors' data are missing, critical data are untrue, and the data are too old to be informative, and the privacy of the customers. If these issues can be resolved, predictions of hotel occupancy rates will be more accurate, and hotels will have better access to timely and accurate information about actual user occupancy rates. If these issues can be resolved, the prediction of hotel occupancy rates will be more accurate, and the hotel will be able to make more informed decisions regarding the number of pre-sale rooms, pre-sale room rates, and pre-sale start times, ultimately increasing the hotel's profitability and resource efficiency.

To summarize, this paper focuses on the dataset of hotel booking conditions and the machine learning algorithms related to predicting hotel occupancy and highlights the significance of predicting hotel occupancy and the problems that still exist in this prediction. The subsequent research needs to solve the problem of predicting hotel occupancy under a small amount of data or how to obtain a large amount of valid data for prediction, as well as focusing on personal privacy.

References

[1] Gustavo, Nuno. "Marketing management trends in tourism and hospitality industry: facing the 21st century environment." International Journal of Marketing Studies 5.3 (2013): 13.

[2] Zhu, Fanwei, et al. "Modeling Price Elasticity for Occupancy Prediction in Hotel Dynamic Pricing." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. (2022): 2

[3] Casalo, Luis V., et al. "Do online hotel rating schemes influence booking behaviors?." International Journal of Hospitality Management 49 (2015): 28-36.

DOI: 10.54254/2755-2721/46/20241296

[4] Albanese, Davide, et al. "mlpy: Machine learning python." arXiv preprint arXiv:1202.6548 (2012): 1-3

[5] Kumari, Roshan, and Saurabh Kr Srivastava. "Machine learning: A review on binary classification." International Journal of Computer Applications 160.7 (2017): 2-4.

[6] Zhang, Jialu, et al. "Succinct Explanations With Cascading Decision Trees." arXiv preprint arXiv:2010.06631 (2020): 4-8.

[7] Churcher, Andrew, et al. "An experimental analysis of attack classification using machine learning in IoT networks." Sensors 21.2 (2021): 446.

[8] Zhu, Fanwei, et al. "Modeling Price Elasticity for Occupancy Prediction in Hotel Dynamic Pricing." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. (2022): 3-4.

[9] Antonio, Nuno, Ana de Almeida, and Luis Nunes. "An automated machine learning based decision support system to predict hotel booking cancellations." Data Science Journal 18 (2019): 32-32.

[10] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. (2016): 785-794.

[11] W. J. Zhang, G. Yang, Y. Lin, C. Ji and M. M. Gupta, "On Definition of Deep Learning," 2018 World Automation Congress (WAC), Stevenson, WA, USA, 2018, pp. 1-5.