Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/153/2024.19518

Research for Hotel Reservation Cancellation Based on Prediction Model

Yi Liu[1,a,*]

[1]*Zhejiang University of Finance & Economics, Zhejiang, China*
a. nishiyao@ldy.edu.rs
**corresponding author*

Abstract: The rapid changes in the global economic situation have brought many challenges to the hotel industry. Many hotels face operating difficulties due to the lack of a stable source of income. Customer cancellation is one of the main factors leading to revenue instability, however, a large proportion of hotels are still not effectively taking measures to deal with it. The paper uses machine learning methods to fix profitability issues in the hotel industry. First, the paper processes the raw data from the reservation conditions of a hotel. Then, some key variables are selected through a correlation matrix. The paper introduces and explains some machine learning models to show the advantages and disadvantages of each model. After programming, compare the accuracy scores of models. In the last, select the best model. In the paper, logistic regression, single decision tree, random forest, neural net, and boosted tree are alternative options. The random forest model has the highest accuracy score. The reasons for the conclusion and the suggestions for hotels are listed in the section of the discussion.

Keywords: Hotel Reservation Cancellation, decision tree, random forest.

1. Introduction

U.S. hotels have lost more than $15 billion in room revenue during previous years' epidemic, and with occupancy rates projected to be 20 percent or lower in the coming months, further deterioration could be expected in the future [1]. Nowadays, as the world economy rises, coupled with a gradual improvement in the epidemic situation, people's consumption levels have increased. People's willingness to travel increases. Hence, the hotel industry is a popular sector. The demand has reached 3.45 billion room times [2]. The competitive environment for hotels is fierce. Many customers will change their minds in the meantime. It will lead to hotel reservation cancellation. Hotel reservation cancellation is a quite simple phenomenon in the hotel industry, especially when the traveling craze comes.

There is much difficulty in the prediction of hotel reservation cancellation. This is a decision-making issue, so it's easy to mistake the side of deciding between two outcomes. The information of customers cannot be fully collected. Different hotels collect different information about their customers. These differences between customers and hotels lead to the incoherence of prediction.

Hotel reservation cancellation is crucial for revenue management (RM) in the hospitality industry. It is a management study designed to achieve the goal of maximizing revenues by predicting the customers' demand through different factors. In Stanislav Ivanov's meta-analysis, the Hotel revenue management system and RM tools both contribute significantly to hotel revenues [3]. In detail, hotels

DOI: 10.54254/2754-1169/153/2024.19518

can bring the right product to customers at the right time at the right price. Cancellations in the hotel industry can be affected by other factors such as the reputation of the hotel, the price of the booking, and the date of arrival [4]. Research on this issue is not currently extensive. Most academic papers consider random forest and logistic regression as more appropriate methods to solve this type of problem [5, 6].

The main purpose of the paper is to predict hotel reservation cancellations through the methods of machine learning. These methods mainly focus on the binary problems. The paper uses logistic regression, a single decision tree, a random forest, neural net, boosted tree to build the model. Compare the models and find out the best model to fix the problem.

2. Data General Description

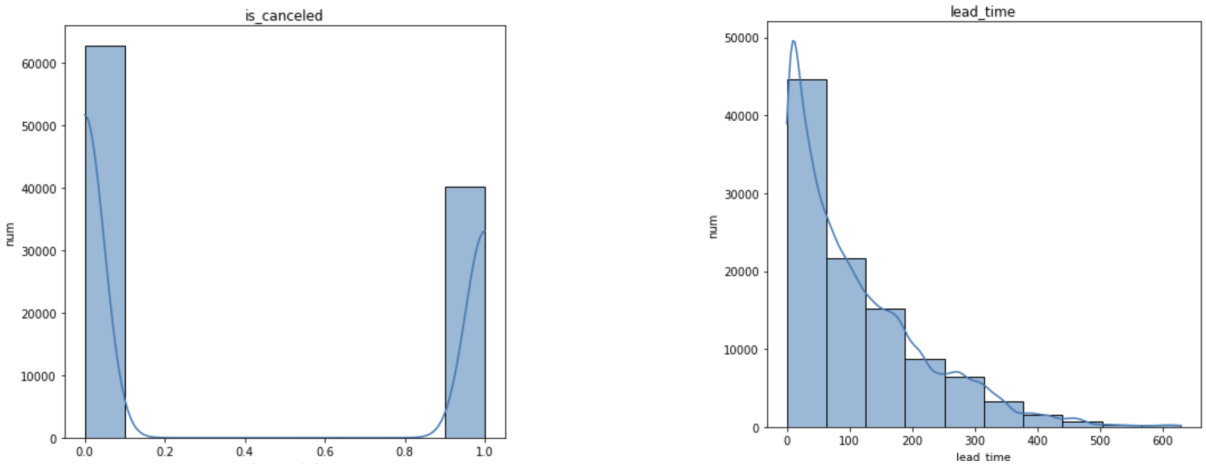
This dataset is about all the customers of a particular hotel who booked that hotel from 2015 to 2017 [7]. Its direct source is from Kaggle, a dataset platform. For the dataset, there are 36 columns and 119390 examples. They are enough to build machine learning models. There are 9 key variables in the dataset. The explanations of variables are listed as follows.

'is_canceled' means whether the reservation is canceled or not. 'lead_time' stands for the number of days from booking date to arrival. 'previous_cancellation' means the number of cancellations before this booking. 'adr' is the short for average daily rate. 'stays_in_week_nights' means the number of guests staying or booked to stay at the hotel on weekday nights. 'adults' and 'children' mean the numbers of adults and children. 'arrival_date_year', 'arrival_data_week_number' and 'arrival_date_day_of_month' show the exact time of the reservation. These 9 variables are important to build the models. The reason why these 9 variables are chosen is listed in the part of Feature Selection.

First, the raw dataset needs to be cleaned up. The variables with too many missing values should be removed. That is because if a variable with a lot of missing values is selected in the model, it will lead to an underfitting problem. The accuracy of the model decreases. Figure 1 shows the number of missing values of each variable in the raw dataset. In the figure, the left column represents the sequence of samples. The column with a lighter color represents there are many samples for which this column is a missing value. The column with a darker color represents there are few samples for which this column is a missing value.

The summary of the dependent variable, 'is_canceled', is listed in Figure 2. Knowing all the information of each variable helps build the model to avoid a lot of problems. Some typical variables' information is as follows. With the information about the data, it is possible to decide whether this set of data is suitable for modeling and what improvements need to be made. Take this set of data as an example. The dependent variable, 'is_canceled', is binary data. It is also the final result of the model. There are many limits to it. First, the number of 'canceled' samples and the number of 'not canceled' samples should be balanced. If the distribution of the dependent variable is unbalanced, it is hard to test the accuracy of the final model. The further step should be applied. SMOTE helps with this problem. SMOTE can be explained briefly as a method to multiply the records with average value of samples. It works significantly with numeric variables. As for the hotel reservation cancellation prediction, the most helpful way can be duplicating the records. The dependent variable distribution in this paper is eligible. Second, 'arrival_date_year', 'Arrival_data_week_number', and 'Arrival_date_day_of_month', these three features are similar. PCA can help with this problem [8]. PCA is a method to combine correlative features into a new feature to avoid overfitting. In the research, only ten key variables are chosen in the model, so PCA is not that essential.

DOI: 10.54254/2754-1169/153/2024.19518



(a) (b)

Figure 1: Illustrates the distribution of the variable, (a) 'is_canceled'; (b)Summary of 'lead_time' (Photo/Picture credit: Original).

In Figure 1, the days from booking date to arrival are strongly related to the number of bookings [9]. The shorter the days are, the more people book the hotel. It is said that people prefer to book hotels when they're sure about it to avoid the uncertainties like bad weather. There is another possible reason. People who make temporary hotel reservations usually use them for business trips, so they don't book particularly luxurious hotels. People who choose to book a hotel a few weeks in advance are people who generally go for traveling and vacation. Their reservations are generally more luxurious and fewer people booked. This reason can be proved by the column of 'stays_in_week_nights.

3. Methodology

3.1. Machine Learning Methods

To predict the cancellation of the hotel reservation, the direct way is to create a machine learning model. Before testing every model, it is essential to learn about the theory of each model.

3.2. Feature Selection

Feature selection is the first step in building a model. The purpose of feature selection is to choose the independent variables for the model. There are two main methods in feature selection. The first is filter, the other is the wrapper. In this paper, the main method is a filter. Put every variable in the correlation matrix. The strength of the association between each variable and the dependent variable can be visualized in Figure 2.

DOI: 10.54254/2754-1169/153/2024.19518

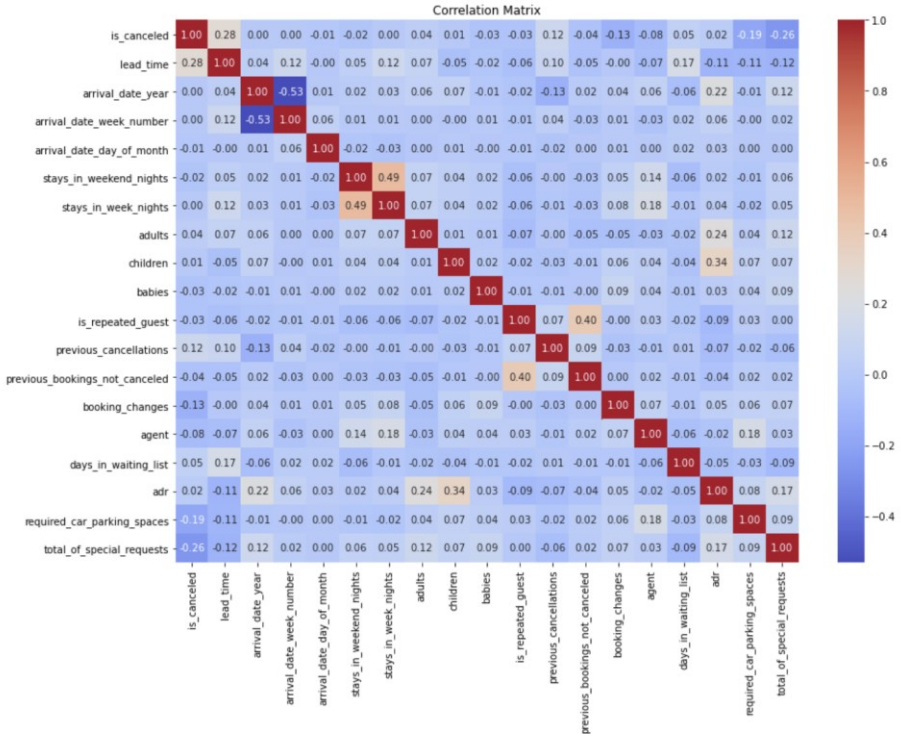


Figure 2: Correlation matrix (Photo/Picture credit: Original).

To avoid problems of overfitting and underfitting, ten variables are chosen as independent variables. It tells the reason why ten key variables are chosen in the section of data general description. The closer the coefficient of the relationship with the dependent variable is to 1, the more the independent variable can be the independent variable of the model.

The result of the correlation matrix also tells other information. The variable with the largest correlation coefficient in the matrix is 'lead_time'. Its correlation coefficient is about 0.28. It means the dependent variable is not directly and necessarily related to the individual variables in the data set. It can be proved by the distribution of lead time with respect to cancellation in Figure 3. The canceled and uncanceled samples are not clearly distributed on both ends of the image, which indicates that 'lead_time' cannot be directly linked to the dependent variable.

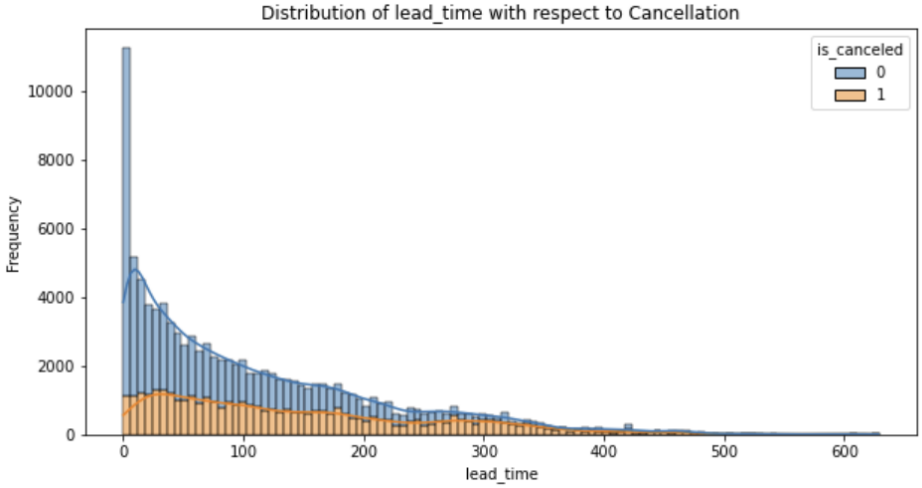


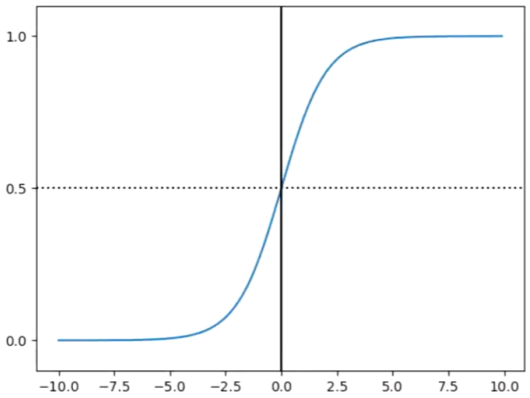
Figure 3: Distribution of 'lead_time' with respect to 'is_canceled' (Photo/Picture credit: Original).

DOI: 10.54254/2754-1169/153/2024.19518

3.3. Logistic Regression

Normally, there are two kinds of regression in machine learning models. One is the linear or non-linear regression, the other is the logistic regression. The purpose of a linear or non-linear regression is to predict the value of the dependent variable. The logistic regression is to solve yes-or-no questions.

The theory of logistic regression is based on the Sigmoid function, which can make the data polarized. It also limits the value of the dependent variable between 1 and -1. Each side of the data represents the result of yes and no. Figure 4 is the Sigmoid function in the coordinate system.



$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

Figure 4: Sigmoid function (Photo/Picture Credit: Original).

3.4. Single Decision Tree

The decision tree is a method to put the examples from the dataset into different columns depending on dividing values. There are three main parts in a decision tree. A root node is the original node in the whole process. Normally, the root node in the decision tree distinguishes whether samples can enter the model. Samples after the root node will result in the subdivision of all records into two or more mutually exclusive subsets. Internal nodes, also called chance nodes, represent one of the possible choices available at that point in the tree structure. Leaf nodes, also called end nodes, normally, represent the final result of the sample in the decision tree.

Decision trees are simple to understand and visualize. Only one hard question in the process of model building. How to choose internal nodes or how do decision trees decide where to cut? The goodness of cuts in a decision is measured by impurity. We use entropy to calculate the impurity of a decision tree. In the formula of entropy, pi means probability density. ΔHe indicates the amount of change in data entropy before and after splitting. The larger the change is, the more likely it is to classify the samples.

3.5. Random Forest

Like decision trees, random forest is a method of bagging decision trees. Before discussing details about the random forest, bagging is an essential concept. For all results of models, there are two main problems. They are bias and variance. Find the balance between them is the key of tree models. High bias causes the predictions of samples to be far from the target. High variance causes the predictions of samples to be dispersedly distributed.

So, bagging is an effective method to reduce variance. The random forest combines the advantages of bagging and decision trees. Figure 5 is an example of a random forest. It builds many independent trees. The dataset in the decision trees turns out to be different results. In the final, combine all the results by averaging the variables across all the trees.

DOI: 10.54254/2754-1169/153/2024.19518

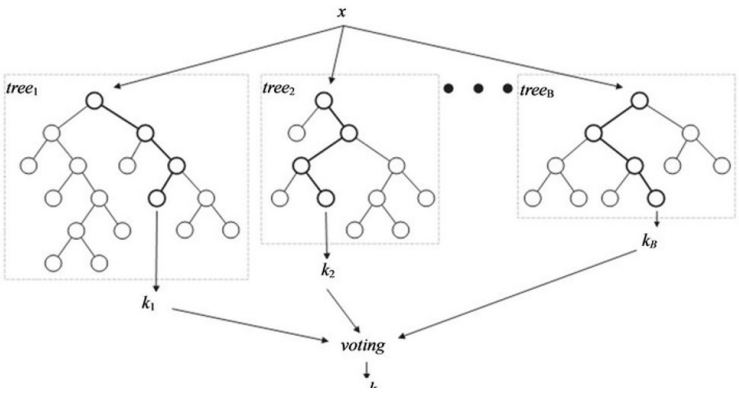


Figure 5: Visualization of random forest (Photo/Picture credit: Original).

3.6. Neural Net

The working principle of the neural net is similar to that of human brain neurons. When an external message stimulates nerve endings, then a message is sent to our brain. So does the neural net. The whole process of a neural net comprises three parts, the input layer, the hidden layer, and the output layer. The input layer and the output layer are understandable, which means the independent variables and the dependent variables. The hidden layers are quite complicated. Each layer contains a set of nodes. The deeper layer receives the previous layer's information. Use the weighting method to form own new data. The output of a neural net can be anything, normally a piece of logistic information. In real life, neural nets can be applied in many fields such as Automatic Speech Recognition (ASR) and picture recognition.

3.7. Boosted Tree

There are many similarities between random trees and boosted trees. They both contain many simple decision trees and result in the combination of decision trees' variables. However, the most notable difference between boosted trees and random forests is that the boosted trees are the sum of weak models typically a few hundred shallow trees. Because of this, the boosted tree can reduce the bias, instead of variance (Figure 6).

Figure 6: Boosted tree (Photo/Picture credit: Original).

4. Model Comparison

Through programming, put processed data in each model. Visualize the results of each model in the confusion matrix. During the model testing process, the data is divided into two groups. One is for training, the other is for testing. To test the applicability of the model, the results of the test were determined using two criteria: check accuracy ('precision') and check completeness ('recall').

In Tables 1-5, the whole table is divided into two Tables. The top Table refers to the accuracy and completeness of the model predictions for real outcomes of 0 and 1. The first row of the Table below shows the predicted results, the first column shows the true results. It shows the numbers of four situations of each sample.

DOI: 10.54254/2754-1169/153/2024.19518

Table 1: Result of logistic regression

percentage	precision	recall
0	67%	91%
1	70%	31%
Confusion matrix		
number	0	1
0	11397	1113
1	5531	2538

Table 2: Result of decision tree

percentage	precision	recall
0	81%	81%
1	70%	71%
Confusion matrix		
number	0	1
0	10080	2430
1	2326	5743

Table 3: Result of random forest

percentage	precision	recall
0	81%	90%
1	82%	66%
Confusion matrix		
number	0	1
0	11295	1215
1	2705	5364

Table 4: Result of neural net

percentage	precision	recall
0	66%	87%
1	61%	32%
Confusion matrix		
number	0	1
0	10887	1623
1	5499	2570

Table 5: Result of boosted tree

percentage	precision	recall
0	69%	95%
1	80%	35%
Confusion matrix		
number	0	1
0	11823	687
1	5237	2832

DOI: 10.54254/2754-1169/153/2024.19518

Based on each model's confusion matrix, the logistic regression model predicts 13935 (11397+2538) samples correctly, the figures of decision tree, random forest, neural net, and boosted tree are 15823 (10080+5743), 16659 (11295+5364), 13457(10887+2570) and 14655(11823+2832) respectively. Through the scores and accuracy of models, random forest is the best solution for the hotel reservation cancellation prediction problem (Table 6).

Table 6: Rank of models

Rank	Model	Accuracy score
1	random forest	0.81
2	decision tree	0.77
3	boosted tree	0.71
4	logistic regression	0.68
5	neural net	0.65

5. Conclusion

Through the whole process of building models for the problem of predicting the cancellation of hotel reservations, it turns out to be two main conclusions. First, there are ten key variables, which can be the factors of the problem. Second, in the last section, comparing the scores of each model, random forest is the best solution for the hotel reservation cancellation prediction problem.

Hotels can predict the possibility of the cancellation of each customer to gain better revenue management. For example, for customers who book the hotel temporarily, the hotel can appropriately increase the price of low-end room types; for customers who book the hotel in advance for a long time, the hotel can appropriately reduce the price of high-end room types. Just as the timing of a customer's hotel reservation is also an important factor, hotels can do a good job of warning customers of cancellations during some of these times. Knowing the ten key variables, hotels can reduce the collection of information on every aspect of customers to prevent leakage of customer information. The hotel can also use these variables about future customers to train the model to make the model more sophisticated and accurate.

Random forest is made up of independent trees. Each tree is a complete, strong model by itself. Integrating predictions can improve model accuracy. Random forests are capable of handling datasets with high-dimensional features and do not require feature dimensionality reduction like PCA.

Even after finding a suiTable model for solving the hotel reservation cancellation problem, this study still left some problems. There can be some methods to upgrade the model. Sometimes, stacking the models can improve the accuracy of the prediction. Some variables may be more relevant to the question to include in the modeling.

References

[1] *YYu, J., Seo, J., & Hyun, S. S. (2021). Perceived hygiene attributes in the hotel industry: Customer retention amid the COVID-19 crisis. International Journal of Hospitality Management, 93, 102768.*

[2] *Zhang, J., Li, D., Lan, H., Tang, L., & Guo, M. (2024). Research on hotel reservation scheme based on random forest model prediction. Advances in Computer and Communication, 6, 1–10.*

[3] *Dawood, M. (2024). Hotel booking cancellations. Kaggle. Retrieved October 8, 2024, from https://www.kaggle.com/datasets/muhammaddawood42/hotel-booking-cancelations.*

[4] *Polemis, M. L., Tzeremes, P., & Tzeremes, N. G. (2023). Hotels' occupancy rates and convergence: Empirical evidence from the first pandemic wave. Tourism Economics, 29(2), 533–542*

[5] *Almotiri, S., Alosaimi, N., & Abdullah, B. (2021). Using API with logistic regression model to predict hotel reservation cancellation by detecting the cancellation factors. International Journal of Advanced Computer Science and Applications, 12(6), 217–222.*

DOI: 10.54254/2754-1169/153/2024.19518

[6] *Antonio, N., de Almeida, A., & Nunes, L. (2019). An automated machine learning-based decision support system to predict hotel booking cancellations. Data Science Journal, 18, 32.*

[7] *Chen, Y., Ding, C., Ye, H., & Zhou, Y. (2022, March). Comparison and analysis of machine learning models to predict hotel booking cancellation. In 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022) (pp. 1363–1370). Atlantis Press.*

[8] *Almotiri, S., Alosaimi, N., & Abdullah, B. (2021). Using API with logistic regression model to predict hotel reservation cancellation by detecting the cancellation factors. International Journal of Advanced Computer Science and Applications, 12(6), 102–109.*

[9] *Nuno, A., Almeida, A. de, & Nunes, L. (2019). Predictive models for hotel booking cancellation: A semi-automated analysis of the literature. Tourism & Management Studies, 15(1), 7–21.*