Data Analysis on Hotel Reservations

Jerry Fang

Abstract

In this project, I analyzed a hotel reservation dataset downloaded from Kaggle. I think it is an interesting topic to find out some patterns on hotel reservation and cancellation information, which can help hotels manage their bookings better. The first step of my analysis is to visualize the data – it provides a straightforward way to view the pattern and complex relationship in hotel data. Visualization shows us an easy-to-understand information. After preprocessing the data (such as dropping empty rows, converting categorial information and data slicing), I applied different machine learning algorithms to make predictions on booking cancellations. The machine learning algorithms I tried are Gaussian Naive Bayes model, KNN, and logistic regression classifier. I compared the accuracy, precision, recall, and processing time using these algorithms. After processing the data, I recommend the possible applications of such analysis and prediction. At the end, I proposed better future data collection to help make our predictions better, as well as ways to help hotels make the profit by managing their bookings better.

Introduction

The dataset that I analyzed for this project is the Hotel Reservations Dataset that I found on Kaggle. The data was collected during the years 2017 and 2018, and it contains information about the customers, timing, and the hotel rooms. There are 36275 entries and 19 columns in the dataset. I think it is interesting to find patterns of hotel reservation and cancellation. These problems are important for the hotel booking management. On the one hand, the hotel can maximize its profit if it has as many rooms booked as possible. It is inevitable that some guests need to cancel the bookings due to change in plans, scheduling conflicts, etc. In this case, hotels can slightly overbook their rooms to make sure rooms are utilized. On the other hand, customers will be unhappy if the hotel is overbooked too much and needs to turn away guests. Thus, a good prediction algorithm can help hotels decide how much to overbook and keep a nice balance between profit and customer satisfaction. In this report, several machine learning algorithms are utilized for prediction. In the end, the results from the algorithm are compared.

Data

The dataset that I am working on for this project is the hotel reservations dataset, which can be found on Kaggle (https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset), which contains numerical and categorical attributes. The data was collected during the years 2017 and 2018, and it contains different types of information about the guests, timing, and the hotel rooms. Information of the customer includes number of adults/children, special requests, parking, meal. Timing information includes lead time, number of weekday/weekends, arrival month, etc. The hotel room information includes hotel price, meal plan and room type. There are a total of 36275 entries (data rows) and 19 columns in the dataset. The data size is reasonably good for the machine learning algorithms. I preprocessed the data by dropping the rows where at least one element is missing and some of the rows that

contain empty values and unused columns, and then used the data as is without any further preprocessing. Categorical data are labeled into integers before the data is supplied to machine learning algorithms. The customer id was dropped since it is irrelevant for booking status prediction. The arrival year was also dropped when slicing the data since it will be different for future predictions.

Problems and Methods

The dataset is a big table that contains a lot of columns (features) and entries. To gain some ideas about the data, I used python visualization and used the python libraries Matplotlib and Seaborn. A box plot is a good way to show the count of entries based on certain attributes. Reservation count vs room type gives us an idea of which room type gets the most reservations, while reservation count vs number of adults gives us an overview of the number of adults in the reservations. A box plot shows the distribution of hotel room price. When plotting distribution of price for different room types, we get a nice comparison on different room types. The correlation map can help identify the correlation among features, such as number of children, number of weekend nights, and room type.

The goal of data processing in this project is to train a machine learning model to accurately predict cancellation. I used the following machine learning algorithms: simple but powerful Gaussian Naïve Bayes, KNN, logistic regression classifier, linear regression, support vector machine (SVM), etc.

I used the train_test_split function in Python to split the dataset into training and test sets (70% training and 30% test). I tried different subsets of features for different algorithms. I also built a hotel cancellation prediction model using the logistic regression classifier which predicts if the guest will cancel the booking or not, this will include every column except for booking status.

I also evaluated the performance of each model based on the accuracy, precision, and recall. Accuracy provides us with an insight of how often the classifier is correct. Precision is about being precise, i.e., how accurate your model is. In other words, when a model makes a prediction, how often it is correct. Recall is the ability of a model to find all the relevant cases within a data set.

I used the Receiver Operating Characteristic (ROC) curve, a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity. An AUC score of 1 represents a perfect classifier while 0.5 represents a worthless classifier.

The K-Nearest Neighbor (KNN) classifier, a supervised machine learning model, is used in the project to predict the accuracy of the model. The Logistic Regression classifier was also used to predict the model, where we once again would split the data into training and testing sets and then evaluate the model into the form of a confusion matrix. Later in the project, I will visualize the confusion matrix using the heatmap.

I trained and predicted the model using the SVM model to predict the accuracy, precision, and recall using the scikit-learn metrics module to determine how accurate the model is and ran a linear regression to find the intercept and the coefficients of the SVM model as well as the Errors and the correlation coefficients.

Results and Discussion

Data Visualization Results and Discussion

Visualization of data makes it easier to identify patterns, trends, and outliers in large datasets. I used matplotlib and seaborn to visualize the data. Below are a few examples of the useful data visualization.

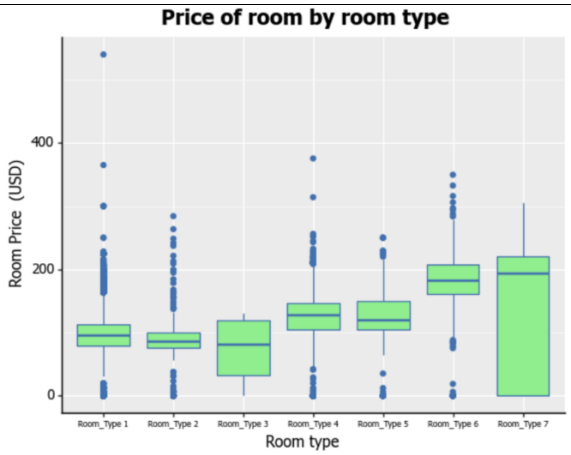

Figure 1. Room price vs. Room type

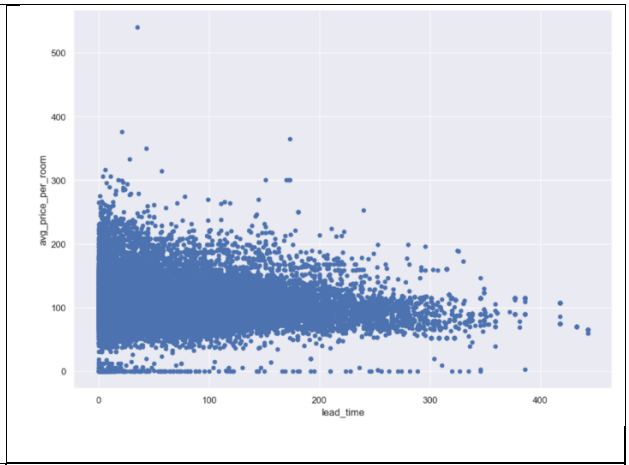

Figure 2. Average price vs. lead time

Figure1 shows the statistics on room price by room type. On average, Room type 7 Is more expensive, and Room Types 1, 2 and 3 are cheaper. There are more outliers for the prices of Room Type 1, where some appear to cost more than 400 (USD) while the average is only about $100. For room type 7, there is a big interquartile range.

The scatterplot above (Figure 2) represents average price per room vs lead time, showing that the average price of the room is lower if the bookings were made early (the booking that was made several days before arrival would have a lower price of the room). Thus, there is a negative correlation between lead time and average price per room.

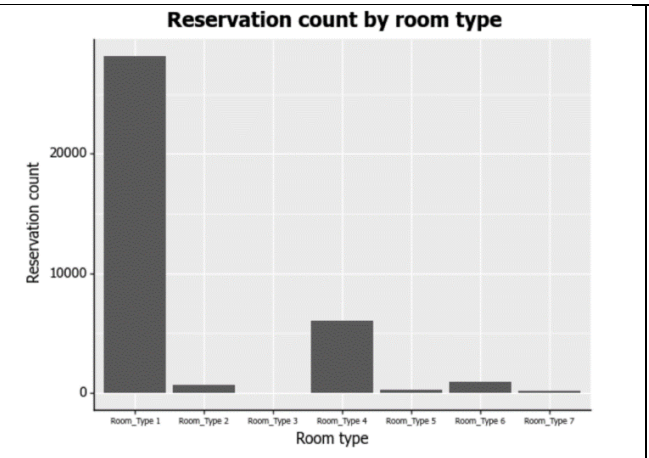

Figure 3. Reservation count by room type

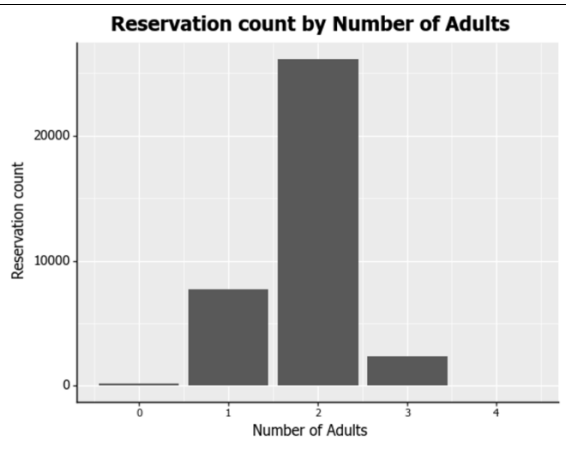

Figure 4. Reservation count by number of adults

Figure 3. shows the bar chart of the number of reservations for room type – more individuals have made reservations for Room Type 1. There are not many reservations on Room Types 3, 5 and 7. From what I observed from Figure 4, it appears that there are more reservations made with 2 adults.

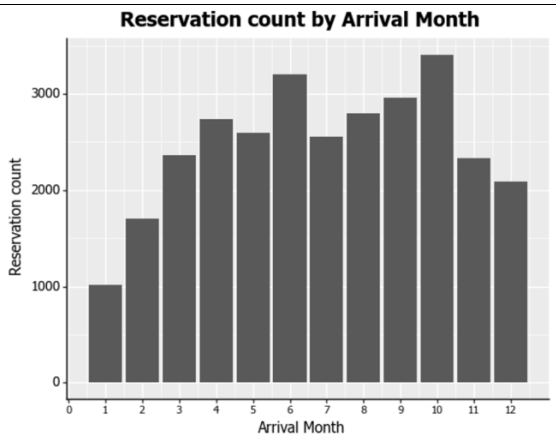

Figure 5. Reservation vs Arrival Month

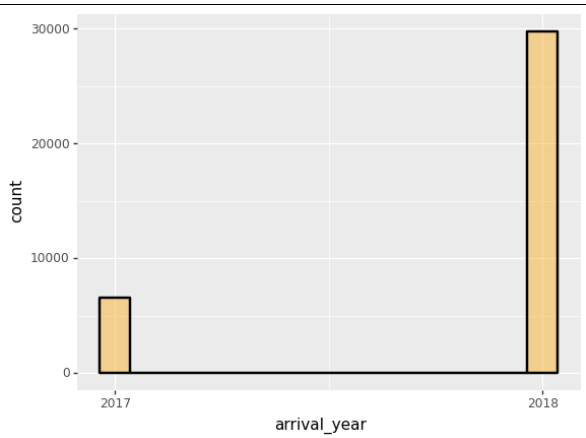

Figure 6. Reservation vs Arrival Year

Figure 5 shows that there are more reservations in June and October than in any month of the year. Figure 6 shows that there are more hotel reservations in the year 2018 than in 2017. However, looking at the data again, 2017 does not have the full year data. For example, it seems that the data from January to July 2017 is missing.

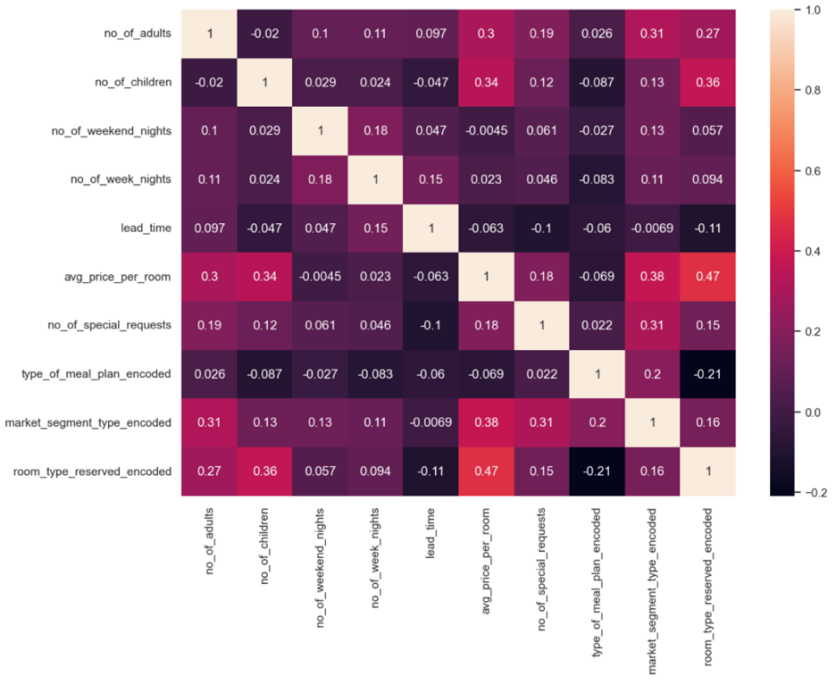

Figure 7. Correlation among features

I have studied the correlation among features by representing a heatmap of the correlation between the variables, where the darkest colors represent a negative correlation while brighter colors represent a positive correlation between the variables. It shows that the room_type_reserved is quite correlated with price of the room and number of adults and number of children.

Data Processing Results and Discussion

Preprocessing

First, I tried to get an idea of the columns and rows of data. There are 19 columns and 36275 rows. There are no rows with missing elements. The categorical data is converted to integer by using the LabelEncoder from sklearn.preprocessing. The following string labels are converted: type_of_meal_plan, market_segment_type, room_type_reserved. The following columns are dropped: booking ID and arrival year (future year is going to be different so that it is irrelevant for prediction)

Machine Learning Algorithms

I split the dataset into training and test sets (70% training and 30% test) and started out by running a Gaussian Naïve Bayes with multiple labels. When evaluating the model, I found the accuracy to be approximately 0.773 with the first subset of data, which included the number of adults/children, number of weekend/weeknights, lead time, average price per room, number of special requests, type of meal plan encoded, market segment type encoded, and room type reserved encoded. I also ran the Gaussian Naïve Bayes using a second subset of data which included the number of adults/children, number of weekend/weeknights, and number of special requests, and I found that the accuracy came out to be approximately 0.676.

The K-Nearest Neighbor (KNN) classifier, the supervised machine learning model also involved splitting the training and test data using the second subset. I built and trained the model by creating the KNN classifier from scipy import stats, and then fit the classifier to the data. Next, I tested the model and showed the first 5 model predictions on the test data and found the accuracy to be approximately 0.815. I then created the knn_cv with n_neighbors=3, and train the model with cv of 5, and printed and averaged each cv score (accuracy) and found the mean to be 0.807. When finding the gridsearch cv, we need to test all values for n_neighbors and then fit the model to the data. I found the best score to be 0.818.

I also built a cancellation prediction model that predicts if the guest will cancel the booking using the logistic regression classifier, and we use the Confusion Matrix to evaluate the model. The dataset features every column except booking status.

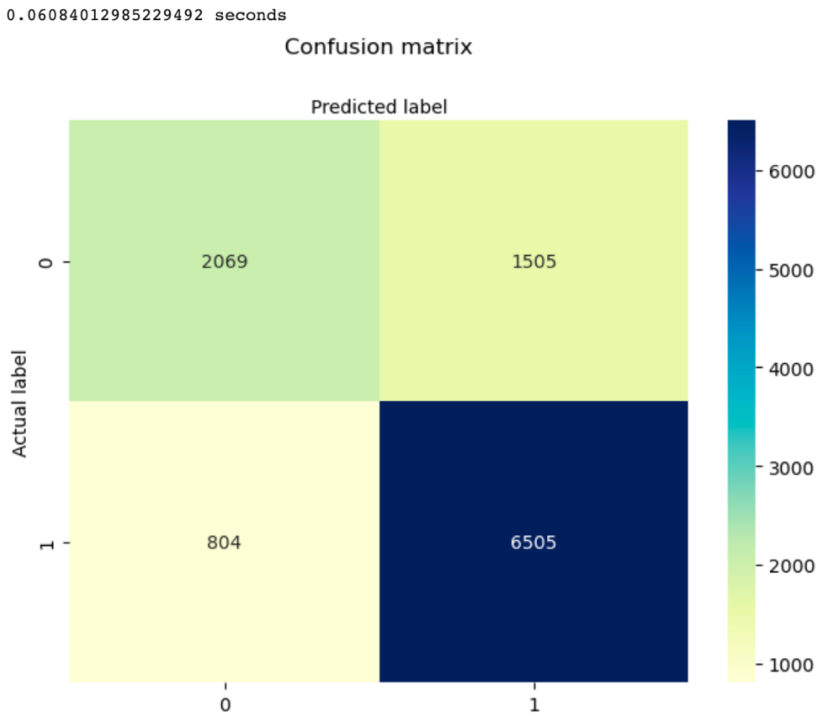

Figure 8. Conclusion Matrix

As represented in the Figure 8 above, 2069 and 6505 are actual predictions while 1505 and 804 are incorrect predictions. 0 represents bookings that are not cancelled while 1 represents bookings that are cancelled.

When evaluating the logistic regression classifier model using model evaluation metrics such as accuracy, precision, and recall, I found the accuracy, precision, and recall of the logistic regression classifier to be 0.788, 0.182, and 0.89 respectively.

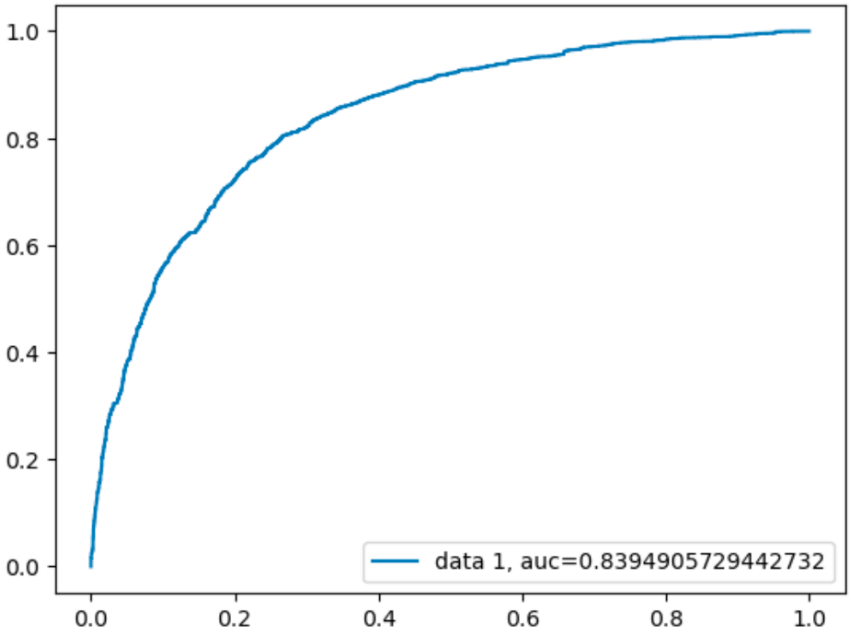

Figure 9. Receiver Operating Characteristic (ROC) curve

The Receiver Operating Curve (ROC) is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity. The AUC score of 1 represents a perfect classifier while an AUC score of 0.5 represents a worthless classifier. As represented above, the AUC score is approximately 0.839, thus making it a reasonably good classifier.

Using the SVM model, we found the accuracy, precision, and recall are respectively 0.799, 0.826, and 0.892. I noticed that the SVM model takes a much longer time to finish. (More than 10 minutes)

Using the linear regression model, I found that the predicted y-values are somewhat close to the actual values (test values). The Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are respectively 0.326, 0.151, and 0.389.

Conclusion

In this study, I have explored the use of many different machine learning algorithms including logistic regression, linear regression, KNN, Gaussian Naïve Bayes, support vector machine (SVM). The results from the machine learning models show that the KNN classifier is the most accurate with an accuracy score of approximately 0.815. With the help of python visualization, I found that most rooms are Room Type 1. The boxplot of room prices vs room type provides nice statistical comparisons of price and room types. I also learned that most of the reservations are made by 2 adults and there are more hotel reservations in the month of June and October. Also, there are more bookings in 2018 than in 2017 (from August only).

All the machine learning methods can potentially help hotels get an idea of possible cancellations. Since there is a possibility that a customer may cancel the booking due to plan changes or other reasons, and it is not always possible to impose a fee on such cancellations. It is a good idea to allow overbooking to see if the model predicts some possible cancellations on a given day. Of course, there is a risk that a reservation may not be honored in this case. Thus, it is a good idea to team up with similar nearby hotels. In case there are more reservations than the room, the hotel can then offer customers alternatives.

It may be tempting to give individualized pricing based on customer information. However, I think it is unethical to do that. A comprehensive model can be derived from data to offer customers with discounts for non-cancellable reservations, but this does not need to be individualized, to avoid discrimination.

It will be helpful to calculate the profit loss due to cancellation, this may be more accurate if some additional data are available. For example, for the rooms that were cancelled, how many days before the arrival date were they cancelled? Are those cancelled rooms rebooked later by a different customer? Also, maybe the location of the hotel (center city, or rural, close to attraction) and big event dates could also be helpful. If there was a big event at certain dates of the year, then those dates may need to be treated differently. (Such big sports events, concerts, trade shows).

From this class and this project, I have learned data visualization in python as well as different machine learning methods. I feel confident that I applied what I learned to extract useful information by analyzing the real-world data.