Springboard – DSC

Capstone Project II

Analysis and Predictive Modeling of

Hotel Cancellation and Price Trends

Final Report

Soyoung An

Data Science Career Track

TABLE OF CONTENTS

*1 Introduction 3*

1.1 objectives 3

1.2 significance 3

*2 Dataset 3*

2.1 data description 3

*3 package introduction 5*

*4 data wrangling 5*

4.1 data processing 5

*5 exploratory data analysis 8*

5.1 summary statistics 8

5.2 heat map and correlations 9

5.3 dataset explorations 9

5.3 hypothesis test and bootstrapping 10

*6 predictive modeLing #*

6.1 preprocessing and training dataset #

6.2 hotel cancellation predictive model #

6.3 hotel price predictive model #

*7 Results #*

7.1 best hotel cancellation predictive model #

7.2 best hotel price predictive model #

7.3 tuning paramers #

7.4 #

*8 Conclusion #*

8.1 summary #

8.2 future scope #

1 Introduction

Hotels provides services for people on holiday that deals with guest accommodation or lodgings. However, hotel cancellation is one of the biggest challenges to control in hospitality industry. Hotel cancellation can lead to the business loss. Having ability to accurately predict future customers’ cancellation rates can help the business gain and future expected revenue. In addition, when we can use cancellation prediction to forecast the potential cancellation rate of a particular customer, it allows to target the individuals to prevent them from cancellation.

Predicting a hotel price rate is essential to evaluate demand for accommodation and pricing the room rates accordingly. Hotel industry is an area where there is a lot of competitions. Having lower rate with same or advance accommodation will help marketing and acquiring customers. Therefore, predicting the actual demand and price can help to reduce the unexpected profit loss. This can help the industry by reducing the unexpected risk, as well as ready with enough facilities.

1.1 Objectives

The objectives of this project are to:

* Explore and analyze hotel booking data
  + Identify the key factors that influence cancellation rate and price
  + Identify whether the average daily rates (ADR) are same between two groups (cancelled and not cancelled)
* Develop a predictive model to estimate the cancellation
* Develop a model to predict hotel price considering multiple factors

1.2 Significance

By analyzing the historical hotel related data, we will identify the main factors that influence hotel churns and price. Any hospitality industry business such as hotels.com, bookings.com, Airbnb, or hotel booking agencies can use suggested predictive models to check their price and cancellation rates.

2 Dataset

2.1 Data Description

The [dataset](https://www.kaggle.com/jessemostipak/hotel-booking-demand) used in this study are sourced from the website Kaggle, a subsidiary of Google LLC, is an online community of data scientist and machine learning practitioners that allows users to find and publish datasets, explore and build models in a web-based data-science environment. This consist of a dataset: hotel\_bookings.csv file:

* Hotel\_booking.csv file is composed of total of 32 columns of features and 40,060 observations of H1 (resort hotel) and 79,330 observations of H2 (city hotel) collected from the article Hotel Booking Demand Dataset written by Nuno Antonio.
* H1 is a resort hotel located in Algarve, Portugal and H2 is a city hotel located in Lisbon, Portugal. Both datasets comprehend bookings due to arrive between the July 1, 2015 and August 31, 2017, including bookings that effectively arrived and bookings that were cancelled.

A description of each of the features of the dataset are provided in Table 2.1.

Table 2.1 Description of hotel\_bookings.csv

|  |  |  |  |
| --- | --- | --- | --- |
| No | Feature Name | Feature Description | Data Type |
| 1 | hotel | Resort Hotel (H1) or City Hotel (H2) | Object |
| 2 | is\_canceled | Value indicating if the booking was cancelled (1) or not (0) | Integer |
| 3 | lead\_time | Number of days between the booking date to the arrival date | Integer |
| 4 | arrival\_date\_year | Year of arrival | Integer |
| 5 | arrival\_date\_month | Month of arrival | Object |
| 6 | arrival\_date\_week\_number | Week number according to year of arrival | Integer |
| 7 | arrival\_date\_day\_of\_month | Day of arrival | Integer |
| 8 | stays\_in\_weekend\_nights | Number of weekend nights booked (Saturday and Sunday) | Integer |
| 9 | stays\_in\_week\_nights | Number of weeknights booked (Monday to Friday) | Integer |
| 10 | adults | Number of adults | Integer |
| 11 | children | Number of children | Integer |
| 12 | babies | Number of babies | Integer |
| 13 | meal | Type of meal booked | Object |
| 14 | country | Country of customer’s origin | Object |
| 15 | market\_segment | Market segment designation | Object |
| 16 | distribution\_channel | Booking distribution channel (how the booking was made) | Object |
| 17 | is\_repeated\_guest | Value indication if the booking was from a repeated guest (1) or not (0) | Integer |
| 18 | previous\_cancellations | Number of previous cancellations prior to current booking | Integer |
| 19 | previous\_bookings\_not\_canceled | Number of previous booking not cancelled prior to current booking | Integer |
| 20 | reserved\_room\_type | Reserved room type code | Object |
| 21 | assigned\_room\_type | Code for the type of room assigned to the booking | Object |
| 22 | booking\_changes | Number of changes made to the booking since entering the hotel management system | Integer |
| 23 | deposit\_type | Type of deposit made for the reservation (No Deposit, Refundable, Non refund) | Object |
| 24 | agent | ID of the travel agency that made the booking | Float |
| 25 | company | ID of the company/ organization that made the booking or is responsible for payment | Float |
| 26 | days\_in\_waiting\_list | Number of days booking was in the waiting list until it was confirmed | Integer |
| 27 | customer\_type | Type of booking | Object |
| 28 | adr | Average Daily Rate (the sum of transactions divided by the number of nights stayed) | Float |
| 29 | required\_car\_parking\_spaces | Number of car parking spaces requested | Integer |
| 30 | total\_of\_special\_requests | Number of special requests made by the customer | Integer |
| 31 | reservation\_status | Last reservation status (Cancelled, Check-Out, No-Show) | Object |
| 32 | reservation\_status\_date | Date at which the last status was set | Object |

3 Package Introduction

In this study we used Jupyter Notebook to run all the codes for data analyzing and modeling. Numpy, Pandas, Matplotlib, Seaborn were installed as basic package. Datetime was installed to manipulate dates and times. Statistical functions (scipy.stats) were imported for inference statistical test. Scikit-learn was installed as the machine learning library. Tabulate is installed to print tabular data as formatted tables.

4 Data Wrangling

4.1 Dataset Processing

We processed the dates type data by following 5 main steps. More details of each step are included in Table 4.1.

Table 4.1 Data Processing Steps for Dates

|  |  |  |  |
| --- | --- | --- | --- |
| Steps | Action | Variable Names | Detail Explanation |
| Step 1 | Data type Correcting | reservation\_status\_date | Convert object to datetime |
| Step 2 | New Variable Creation | arrival\_date | Add new variable ‘arrival\_date’ in datetime by combining ‘arrival\_date\_year’, ‘arrival\_date\_month’, and ‘arrival\_date\_day of month’ |
| Step 3 | Data  Transformation | arrival\_date\_month | Convert month names to numeric value |
| Step 4 | New Variable Creation | reservation\_status\_date\_year | Add new variable ‘reservation\_status\_date\_year’ by extracting year value from variable ‘reservation\_status\_date’ |
| reservation\_status\_date\_month | Add new variable ‘reservation\_status\_date\_month’ by extracting month value from variable ‘reservation\_status\_date’ |
| reservation\_status\_date\_day | Add new variable ‘reservation\_status\_date\_day’ by extracting day value from variable ‘reservation\_status\_date’ |
| reservation\_status\_day\_of\_week | Add new variable ‘reservation\_status\_day\_of\_week’ by extracting day name from variable ‘reservation\_status\_date’ |
| arrival\_date\_day\_of\_week | Add new variable ‘arrival\_date\_date\_day\_of\_week’ by extracting day name from variable ‘arrival\_date’ |
| Step 5 | Rename | arrival\_date\_day | Rename ‘arrival\_date\_day\_of\_month’ to ‘arrival\_date\_day’ |
| Step 6 | Removal | reservation\_status\_date | At the end of data wrangling, reservation\_status\_date was removed. All information (year, month, day) is stored separately. |
| arrival\_date | At the end of data wrangling, removed arrival\_date. All information (year, month, day) is stored in separately. |

Missing Values

We found some missing values in children, country, agent, and company columns in the dataset.

* ‘company’ column was dropped since it has 94% of columns with missing values.
* Missing values in ‘agent’ column were replaced with a value zero, representing ‘not a third party’.
* Missing values in ‘country’ column was replaced with a string ‘not available’.
* There are 4 missing values in children and filled with 0 since the majority children values are zero.

Suspicious Data

* Average Daily Rate (adr)

We investigate the average daily rate with zero, negative, and very small values. We were not able to find any additional information such as rewards program (free-night) or policy. Histogram below is shown in Figure 4.1 that represent the count numbers for average daily rate under 30 euro. We dropped the 0 adr for the histogram for the scaling reason. We were able to see a big drop at 25 euro adr and decided to investigate further more.

Chart, histogram

Description automatically generated

Figure 4.1 Histogram of average daily rate below 30 euro (dropped 0 adr for scaling)

There are 2437 rows containing adr values less than 25 euro. The size of data volume (adr less than 25 euro) is minor compare to the total size of the dataset. We decided to drop these rows. More detailed information is shown in Table 4.2.

Table 4.2 Number of rows in different conditions

|  |  |  |
| --- | --- | --- |
| Condition | Number of Rows | adr values |
| adr < 0 | 1 | -6.38 |
| adr = 0 | 1959 | 0 |
| 0 < adr < 25 | 477 | 0.26 to 24.95 |
| Total | 2437 | -6.38 to 24.95 |

Also, there was one outlier row with adr value of 5400 euro. We dropped this row for data analysis and modeling.

* No-Show

There was one row with no-show in reservation with different reservation\_status\_date and arrival\_date. No-Show should be considered as not showing on the day of arrival date and the data was ambiguous. We decided to drop this one row.

* Final Shape of Dataset

After all the dataset processing, there are 116,951 rows and 35 columns left from 119,390 rows and 32 columns initially.

5. Exploratory Data Analysis

5.1 Summary Statistics

After cleaning the data, statistics including mean, standard deviation, minimum, maximum, and percentiles for each variable in the hotel\_booking.csv were summarized in Table 5.1. Excluded classification data types.

Table 5.1 Summary Statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Vaiable Name | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| lead\_time | 116951 | 105.188130 | 106.924572 | 0 | 19 | 71 | 162 | 709 |
| stays\_in\_weekend\_nights | 116951 | 0.935358 | 0.993030 | 0 | 0 | 1 | 2 | 19 |
| stays\_in\_week\_nights | 116951 | 2.519517 | 1.885362 | 0 | 1 | 2 | 3 | 50 |
| adults | 116951 | 1.861831 | 0.480757 | 0 | 2 | 2 | 2 | 4 |
| children | 116951 | 0.104223 | 0.398829 | 0 | 0 | 0 | 0 | 10 |
| babies | 116951 | 0.007867 | 0.097193 | 0 | 0 | 0 | 0 | 10 |
| is\_repeated\_guest | 116951 | 0.027730 | 0.164198 | 0 | 0 | 0 | 0 | 1 |
| previous\_cancellations | 116951 | 0.081906 | 0.777167 | 0 | 0 | 0 | 0 | 26 |
| previous\_bookings\_not\_canceled | 116951 | 0.125258 | 1.448456 | 0 | 0 | 0 | 0 | 72 |
| booking\_changes | 116951 | 0.215193 | 0.630094 | 0 | 0 | 0 | 0 | 18 |
| days\_in\_waiting\_list | 116951 | 2.345367 | 17.710354 | 0 | 0 | 0 | 0 | 391 |
| adr | 116951 | 103.863971 | 46.422812 | 25 | 71.1 | 95 | 126 | 510 |
| required\_car\_parking\_spaces | 116951 | 0.062607 | 0.245517 | 0 | 0 | 0 | 0 | 8 |
| total\_of\_special\_requests | 116951 | 0.571718 | 0.791880 | 0 | 0 | 0 | 1 | 5 |

The average leading time between arrival date and booking date is 105 days which is about 3 months and more. The average number of adults is 2 people. The mean of previous cancellations variable is 0.08 however, there are some extreme cases where customers made up to 26 times previous cancellations. The average time waited in the waiting list until it is confirmed is a little bit over 2 days. The maximum waited days in waiting list is 391 days. The mean of average daily rate is 103.86 euro with 46.42 standard deviation. The lowest price is 25 euro and highest is at 510 euro.

5.2 Heatmap and Correlations

Heatmap is shown in Figure 5.1 to see the relationship and correlation amongst the features.

Chart

Description automatically generated

Figure 5.1 Heatmap

Average daily rate (adr) has high correlation with children, adults, reservation\_status\_date\_year, arrival\_date\_year, and total\_of\_special\_requests. Having more people (adults and children) and special orders increases daily rate.

Cancellation (is\_canceled) has high association with lead\_time, previous\_cancellations, days\_in\_waiting\_list, and adr. Having longer waiting time until the booking date and long waiting time until the booking confirmation often cause cancellations. Also, customer’s history and price are important features that affects cancellation.

5.3 Dataset Explorations

There are 73,140 reservations that were checked-in (not cancelled), and 43,811 reservations that were cancelled. From Figure 5.2, you can find the count plot for cancellation.

Also, there were similar numbers of reservation bookings throughout the week for both for cancelled and non-cancelled groups. In Figure 5.2, we visualized the number of reservations by different day of week and cancellation.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 5.2 Figure 5.3 Reservation booking by

Counts of different cancellation types day of week and cancellation

5.4 Hypothesis Test and Bootstrapping

We performed hypothesis test to see whether if the average daily rate price booked by who has cancelled has the same or different rates as people who is not cancelled.

* Null hypothesis: The average daily rate price (adr) booked by who has cancelled the reservation has the same rate as people who is not cancelled.
* Alternative hypothesis: The average daily rate price (adr) booked by who has cancelled the reservation does NOT have the same rate as people who is not cancelled.

We performed the hypothesis test using t-test. The shape of histogram follows normal distribution trends however, it is slightly skewed to the right (Figure 5.4). Therefore, we used bootstrapping method to check the t-test results.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Figure 5.4 Histogram of average daily rate by different cancellation types

Both t-test using original dataset and bootstrapped sample, we obtained p-value lower than 0.05, standard alpha value which also known as significance level. Therefore, we can safely reject the null hypothesis and accept alternative hypothesis which the average daily rate (adr) between two groups (who has cancelled and not cancelled the reservations) are different.

6 Predictive Modeling

6.1 Preprocessing and Training Dataset

6.2 Hotel Cancellation Predictive Modeling

6.3 Hotel Price Predictive Modeling

7