Hotel Booking Analysis & Prediction

Vikas Ranjan

DSC680, Summer 2021

Bellevue University, NE

Introduction

The hotel industry is subdivision of the hospitality industry that specializes in providing customers with lodging services. There are a variety of hotel types that typically can be categorized by size, function, service, and cost. Levels of service can usually be split into three options namely limited-service, mid-range service, and full-service. Booking of hotel room are done through various methods such as travel booking websites (such as Priceline, Kayak, Expedia, Booking.com), directly through hotel's website, or various travel agents. Being an avid traveler, I choose to pick up a dataset in hotel bookings. Booking cancellations have significant impact on demand-management decisions in the hospitality industry.

Hotel industry faces a very high cancellation rate. With a global average of almost 40% cancellation rate, this trend produces a very negative impact on hotel revenue and distribution management strategies. To mitigate the effect of cancellations, hotels implement rigid cancellation policies and overbooking tactics, which in turn can have a negative impact on revenue and on the hotel reputation. To make things worse, during this covid crisis, hotel and hospitality industry has taken a massive hit in terms of less bookings and even more cancellations. The impact of COVID-19 on the travel industry so far has been multiple times worse than 9/11. Hotels were one of the first industries affected by the pandemic and it will be one of the last to recover. Therefore, it is more than ever necessary that using data science and ML, potential booking cancellations are identified in advance to allow hotel management allocate resources and plan accordingly. The aim of this project to not only find patterns and key insights related to hotel booking and cancellations but also to develop a few models for predicting a cancellation on a hotel room booking.

Methods

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains following steps which will be followed throughout the project.

- **Business Understanding** – Hotel industry plays a very important role in our economy and it supported one in 25 American jobs before COVID. A very high percentage of booking cancellations was already a major headache for hotel industry, below is the snapshot of cancellation rate (pre Covid 19).

CANCELLATION RATE BY RESERVATION VALUE

Percentage of on-the-books revenue cancelled before arrival in Europe

	2014	2015	2016	2017	2018	Change
Booking Group	43.4%	43.8%	48.2%	50.9%	49.8%	6.4
Expedia Group	20.0%	25.0%	25.8%	24.7%	26.1%	6.1
Hotelbeds Group	33.2%	37.8%	40.3%	38.3%	37.6%	4.4
HRS Group	58.5%	51.7%	55.2%	59.4%	66.0%	7.5
Other OTAs	13.7%	15.2%	27.0%	24.4%	24.3%	10.6
Other Wholesalers	31.2%	30.3%	34.6%	33.8%	32.8%	1.6
Website Direct	15.4%	17.7%	18.0%	18.4%	18.2%	2.8
AVERAGE	**32.5%**	**34.8%**	**39.6%**	**41.3%**	**39.6%**	**7.1**

Covid has not only severely impacted the demand but also spiked the already high cancellations. As part of this project, I'll be looking at a few patterns and insights in the hotel booking dataset. I'll also be developing a few models for predicting a cancellation on a hotel room booking. The dataset I've found and used is one of the clean & reliable dataset. The dataset is moderately imbalanced, and I'll be using methods to overcome that challenge.

- **Data Understanding** - I extracted the dataset from Kaggle from the following url: https://www.kaggle.com/jessemostipak/hotel-booking-demand.

  This dataset includes 119391 rows and 32 feature variables. This data set contains booking information for a city hotel and a resort hotel, and also includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

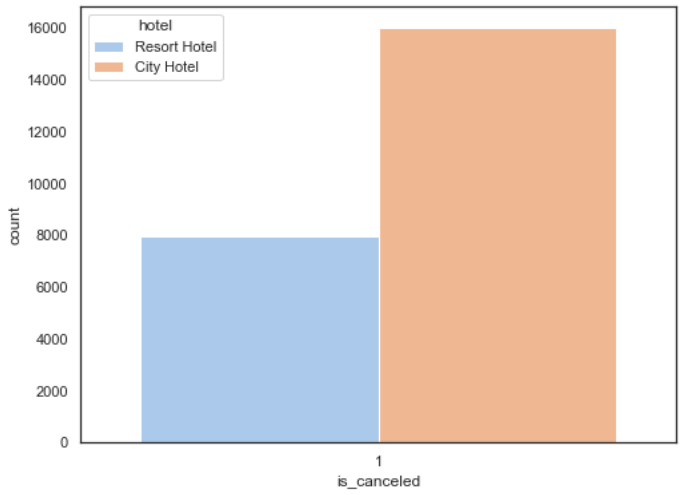
  All personally identifying information has been removed from the data.

  Below are the booking attributes available in the dataset:

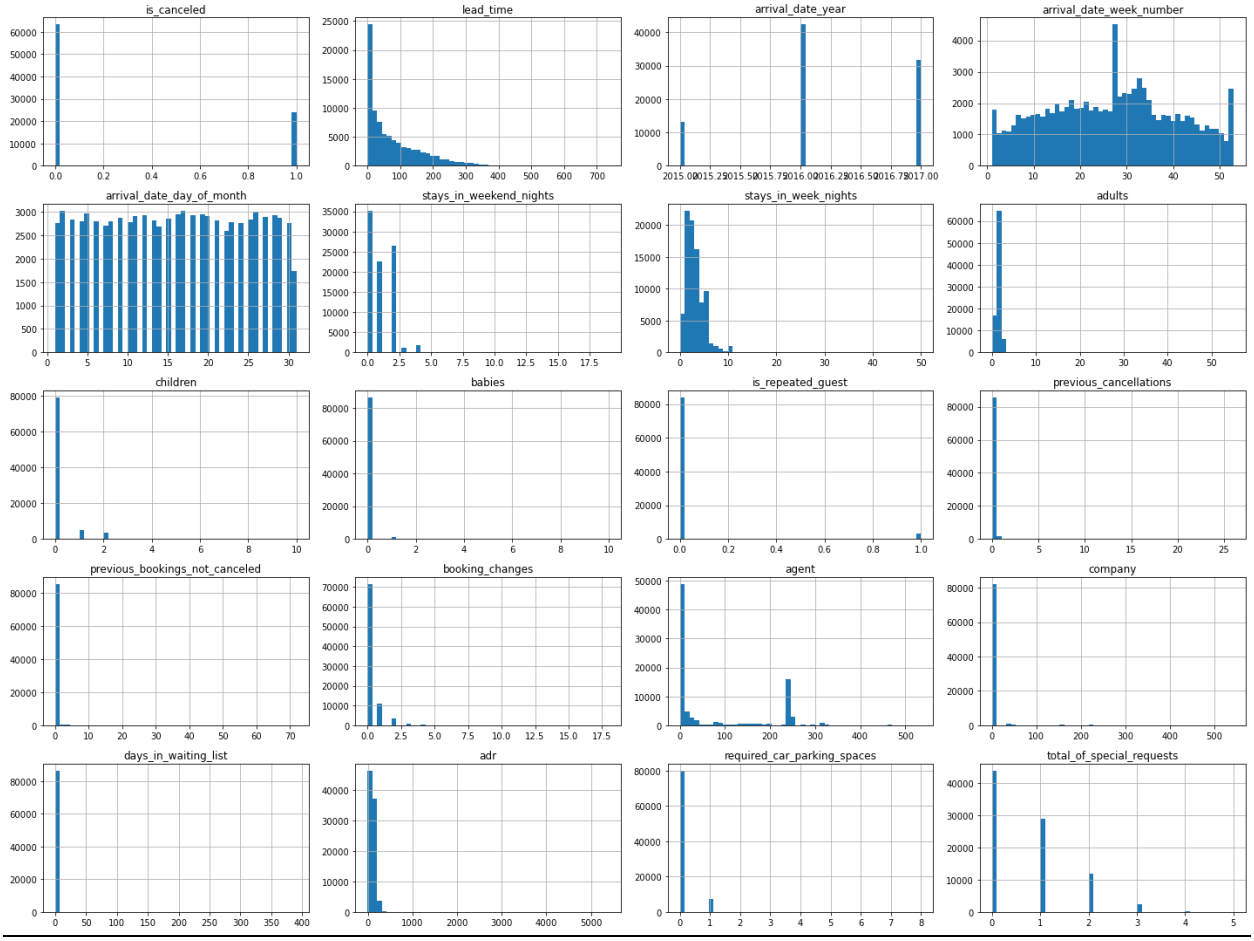
  - hotel
  - is_canceled
  - lead_time
  - arrival_date_year
  - arrival_date_month
  - arrival_date_week_number
  - arrival_date_day_of_month
  - stays_in_weekend_nights
  - stays_in_week_nights
  - adults
  - children
  - babies
  - meal
  - country
  - market_segment
  - distribution_channel
  - is_repeated_guest
  - previous_cancellations
  - previous_bookings_not_canceled
  - reserved_room_type
  - assigned_room_type
  - booking_changes
  - deposit_type
  - agent
  - company
  - days_in_waiting_list
  - customer_type
  - adr
  - required_car_parking_spaces
  - total_of_special_requests
  - reservation_status
  - reservation_status_date

- **Data Preparation** – The source dataset was relatively clean. First step was to perform data cleanup.
    - Validated all the fields for null values and replaced then with either 0 or mean value (for booking rates).
    - There were a few duplicates which were removed as well.
    - Removed the records which had 0s for number of adults, children and babies.
    - Formatted the date fields and split year, month, & day into different fields.
    - Dropped the fields (**'days_in_waiting_list', 'arrival_date_year', 'arrival_date_year', 'assigned_room_type', 'booking_changes', 'reservation_status', 'country', 'days_in_waiting_list')** which didn't seem useful for the model.
    - Performed encoding on categorical variables based.
    - Performed normalization of numeric variables.

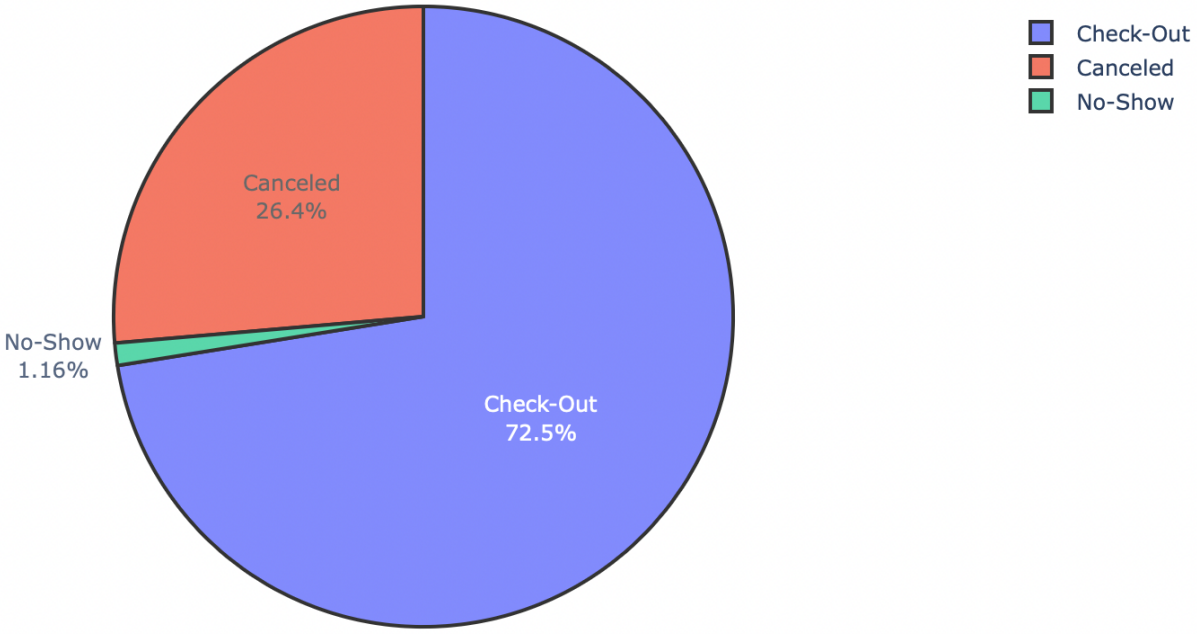
- **Exploratory Data Analysis** - Next step of understanding data is to perform graph analysis. Graph analysis helps significantly with that as it not only shows the patterns and distributions but also would also tell which attributes are corelated. They would also help to guide future business decisions.
    - Below plot shows the comparison of cancellation of bookings for Resort Hotel and City Hotel.



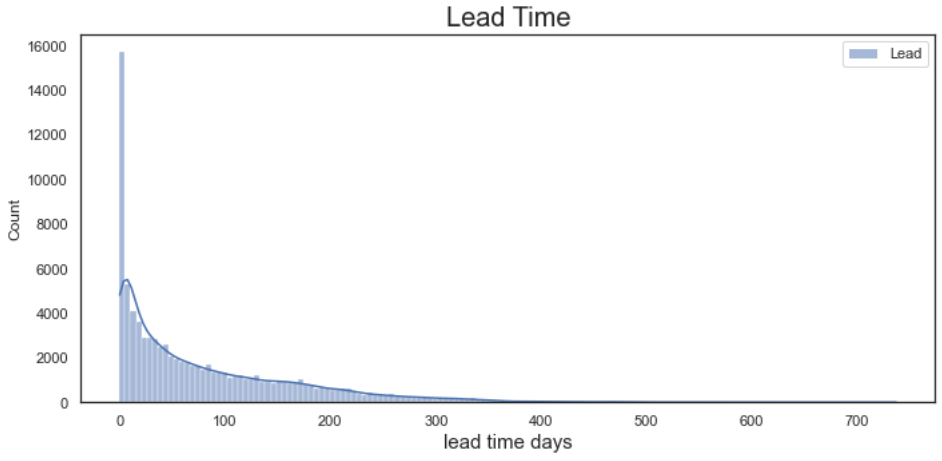
- Below set if plots shows the distribution of 20 different numeric attributes of the dataset. Below are some of the key observations:
  - Week 32, 33 and 34 seems to have highest number of bookings. Seems like summer weeks have higher rates of bookings.
  - Looking at the lead_time plot, most of the bookings were made shortly before arrival.
  - Majority of bookings tend to be without children.
  - It seems that the most accommodations are two weeks long or shorter.
  - While most bookings were not canceled, there are thousands of instances that were.
  - Most of the bookings prices were in a range of 100-200.
  - Most of the bookings were for a period of stay of 2 to 3 week days.
  - Majority of the bookings are for 2 adults.
  - Most of the bookings came with 1 or 2 special requests.
  - Most of the booking required car parking.



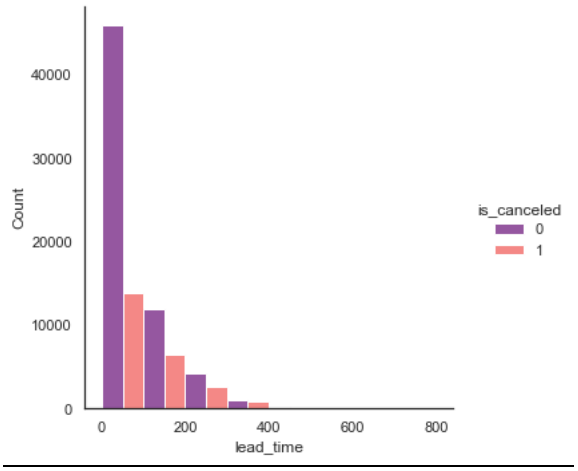
- Below pie chart shows the distribution of the Reservation Status.



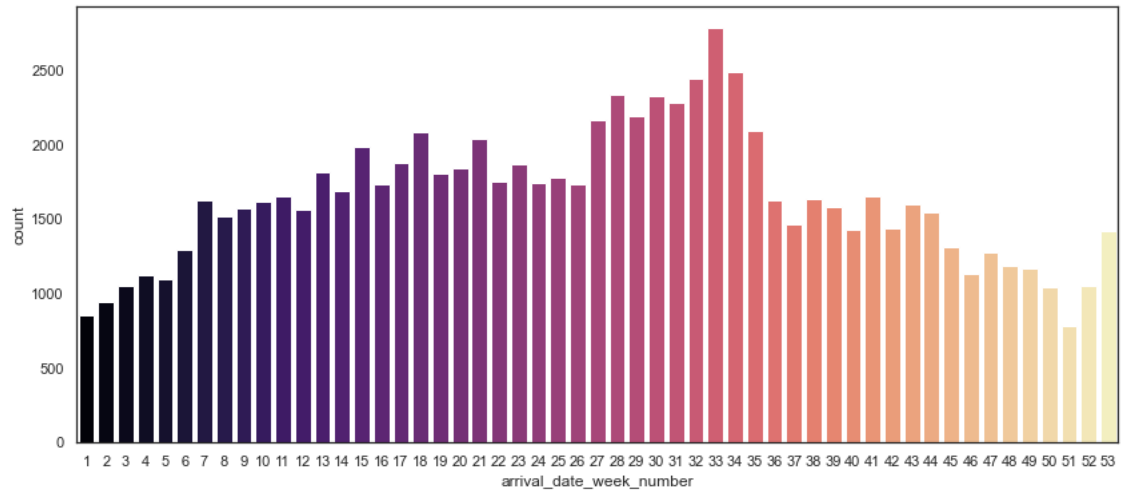
- Below plot shows the distribution and highest concentration points.



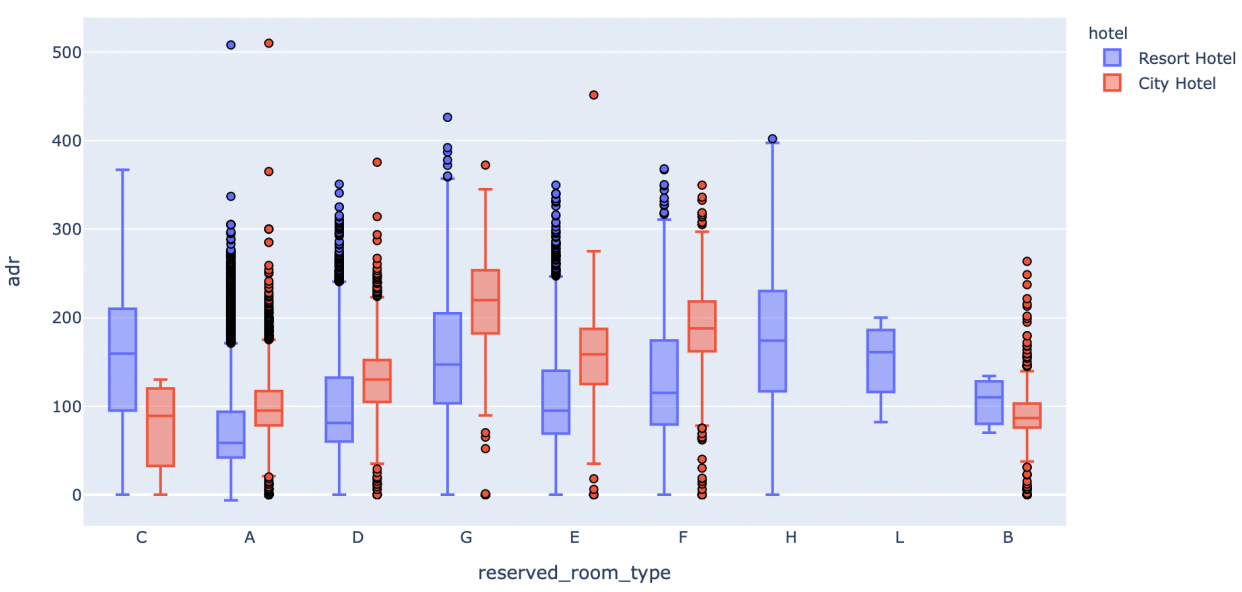
- Below plot shows the distribution of Lead time when compared to cancelled vs non-cancelled bookings.



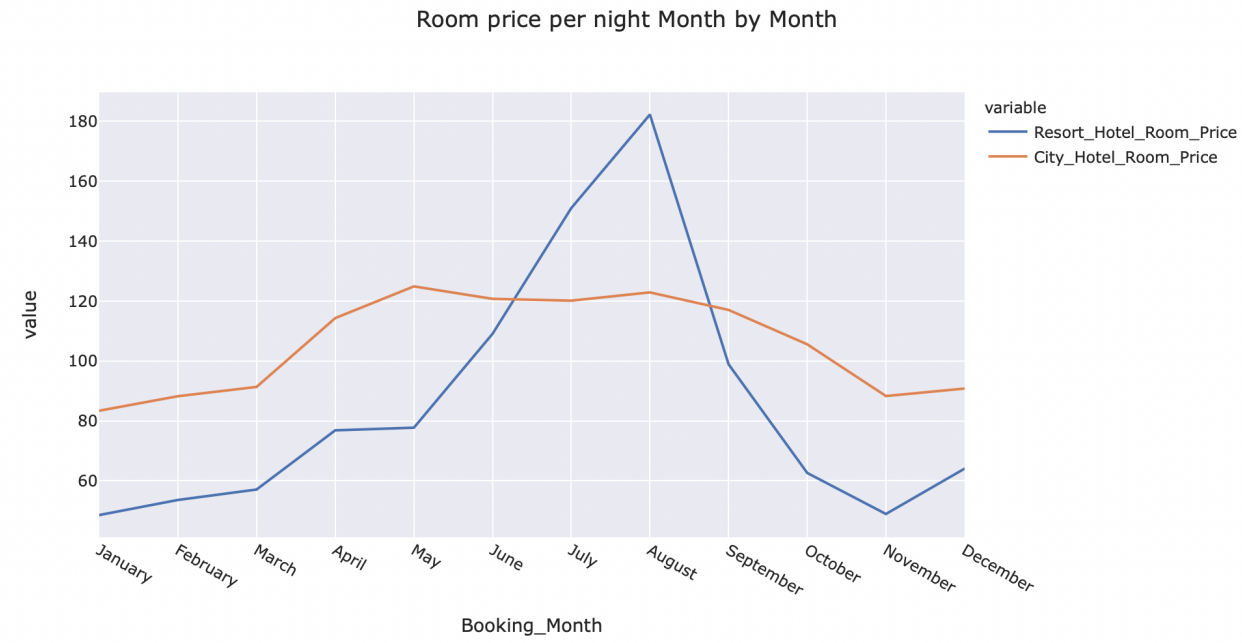
- Below plot shows the distribution of bookings over the weeks of the year.



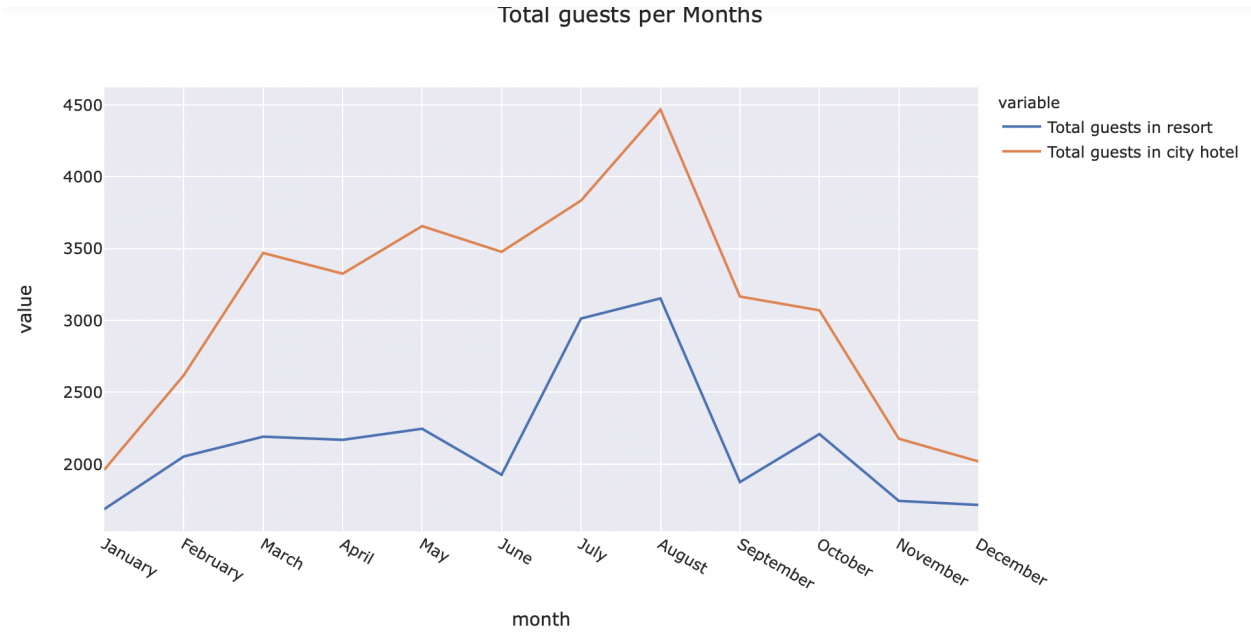
- Below plot the demonstrates the distribution of average daily rates for different room types



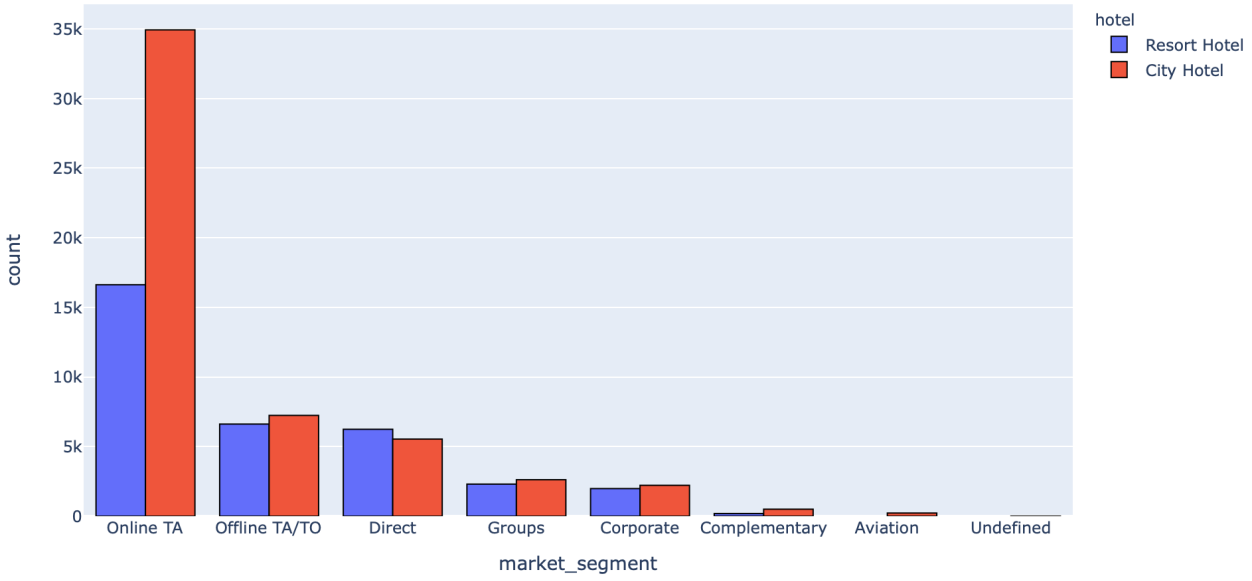
- Below plot shows the comparison of booking price of room over a period of time.



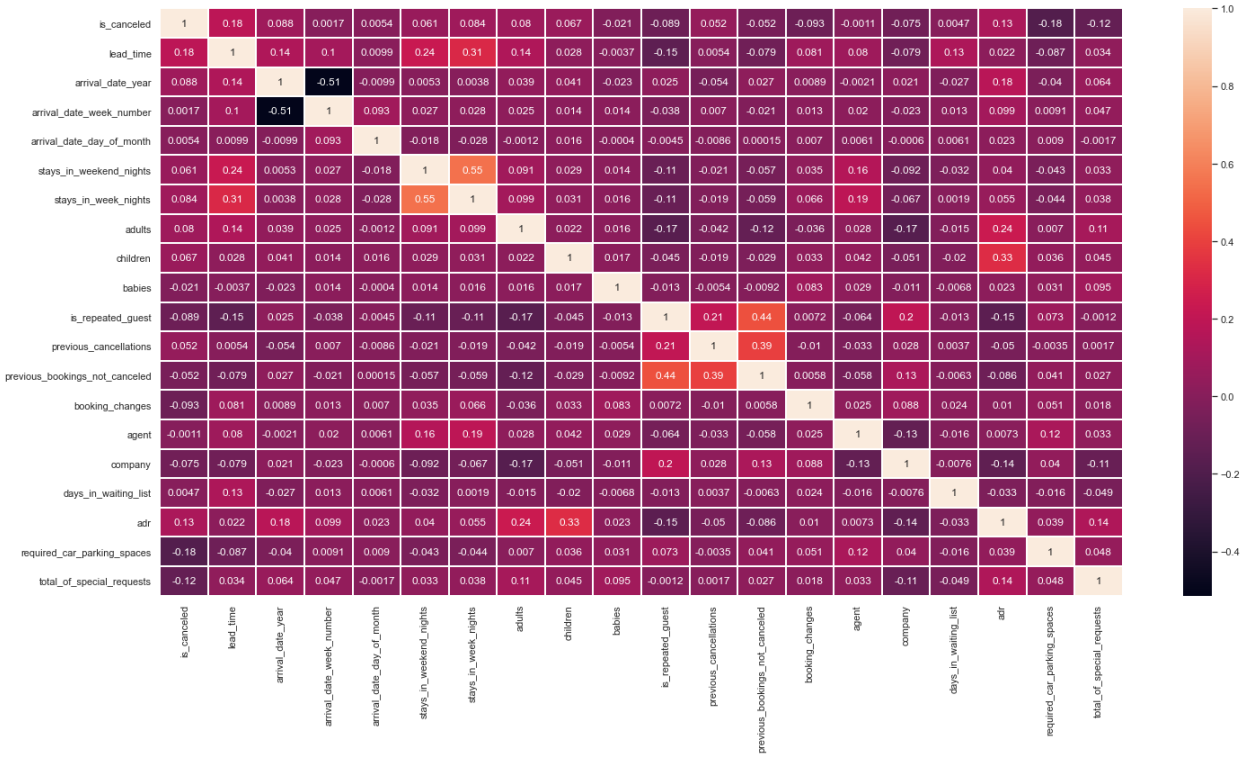
- Below plot shows the comparison of total guests per month.



- Plot showing distribution of Market Segment by different Hotel Types



- Heatmap of the correlation




```
is_canceled                      1.000000
lead_time                        0.184515
adr                              0.127233
arrival_date_year                0.088020
stays_in_week_nights             0.084159
adults                           0.080271
children                         0.067182
stays_in_weekend_nights          0.060992
previous_cancellations           0.051501
arrival_date_day_of_month        0.005449
days_in_waiting_list             0.004710
arrival_date_week_number         0.001691
agent                           -0.001145
babies                          -0.020627
previous_bookings_not_canceled  -0.052178
company                         -0.075314
is_repeated_guest               -0.088764
booking_changes                 -0.093236
total_of_special_requests       -0.120794
required_car_parking_spaces     -0.184456
Name: is_canceled, dtype: float64
```

Correlations Analyis - is_canceled:

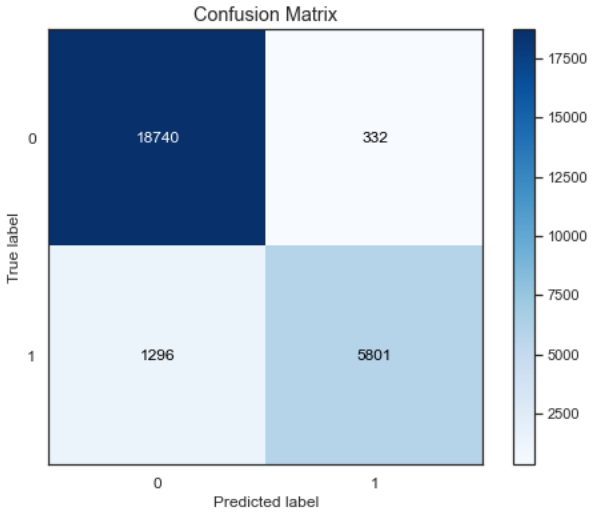
- The strongest positive correlations are with lead_time and adr.
- The strongest negative correlations are with total_of_special_requests, required_car_parking_spaces and booking_changes.

- **Modeling** – As part of modeling, I did notice that this dataset is moderately imbalanced. If not taken care of, results may be biased as the algorithms are much likely to classify new observations to the majority class and high accuracy won't tell us anything. To address the problem of imbalanced dataset, we choose to use oversampling data approach technique. Oversampling increases the number of minority class members in the training set. In order to make our data set balanced, I'm using a type of oversampling called Random Oversampling to overcome using resample. RandomOverSample class is part of imblearn which allows us to over-sample the minority class by picking samples at random with replacement. One of the other popular class would be SMOTE.

1. **Random forest model** creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It provides us a high true prediction values for our dataset. This model when applied on the dataset has an AUROC of 0.93779 which means that the model has very good discriminatory ability.

```
Accuracy Score of Random Forest is : 0.9377889869693148

Confusion Matrix :
[[18740   332]
 [ 1296  5801]]

Classification Report :
              precision    recall  f1-score   support

           0       0.94      0.98      0.96     19072
           1       0.95      0.82      0.88      7097

    accuracy                           0.94     26169
   macro avg       0.94      0.90      0.92     26169
weighted avg       0.94      0.94      0.94     26169
```

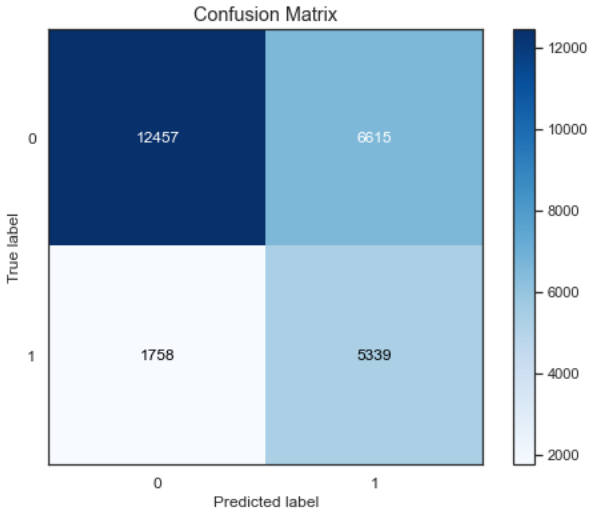
Confusion Matrix:



2. **Logistic regression model** takes a linear equation as input and use logistic function and log odds to perform a binary classification task. I tested this model with the test dataset which resulted in an AUROC of 0.68. Below is the snapshot.

```
Accuracy Score of Logistic Regression is : 0.6800412702052047

Confusion Matrix :
[[12457  6615]
 [ 1758  5339]]

Classification Report :
              precision    recall  f1-score   support

           0       0.88      0.65      0.75     19072
           1       0.45      0.75      0.56      7097

    accuracy                           0.68     26169
   macro avg       0.66      0.70      0.65     26169
weighted avg       0.76      0.68      0.70     26169
```

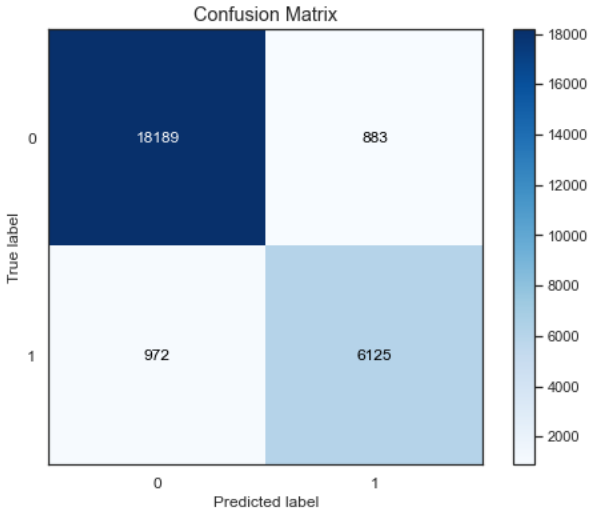
Confusion Matrix:



3. **Decision Tree model** - I tested this model with the test dataset which resulted in an AUROC of 0.92911. Below is the snapshot of results.

```
Accuracy Score of Decision Tree is : 0.9291146012457487

Confusion Matrix :
[[18189   883]
 [  972  6125]]

Classification Report :
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     19072
           1       0.87      0.86      0.87      7097

    accuracy                           0.93     26169
   macro avg       0.91      0.91      0.91     26169
weighted avg       0.93      0.93      0.93     26169
```

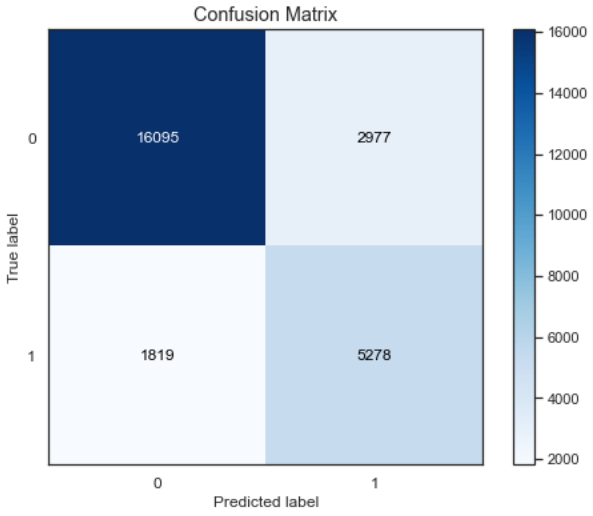
Confusion Matrix:



4. **KNN classification** - Applied KNN model on the test dataset which resulted in an AUROC of 0.81673. Below is the snapshot of results.

```
Accuracy Score of KNN is : 0.8167297183690626

Confusion Matrix :
[[16095  2977]
 [ 1819  5278]]

 Classification Report :
              precision    recall  f1-score   support

           0       0.90      0.84      0.87     19072
           1       0.64      0.74      0.69      7097

    accuracy                           0.82     26169
   macro avg       0.77      0.79      0.78     26169
weighted avg       0.83      0.82      0.82     26169
```

Confusion Matrix:



- **<u>Models Evaluation</u> –** Comparison of the implemented models in terms of accuracy. Here, we can see that Random forest and Decision Tree classification models had best accuracy scores when applied on this dataset.

	Model	Score
1	Random Forest Classifier	0.937789
3	Decision Tree Classifier	0.929115
2	KNN	0.816730
0	Logistic Regression	0.680041

Conclusion

- Booking cancellations has strongest positive correlation with lead time and average daily rates. Therefore, hotels should look at this and try to adjust their pricing and cancellation strategies accordingly.

- Random oversampling technique helped overcome the Imbalanced datasets challenge.
- Using Random Forest Model our model will correctly predict if a hotel room booking will be cancelled or not 93.78% of the time.
- Decision Tree Model our model will correctly predict if a hotel room booking will be cancelled or not 92.91% of the time.
- Random forest model has less false positives than Decision Tree making it a better model.

References

Mostipak, J. (2020, February 13). *Hotel booking demand*. Kaggle. https://www.kaggle.com/jessemostipak/hotel-booking-demand.

Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. IEEE Xplore. (n.d.). https://ieeexplore.ieee.org/abstract/document/8260781.

Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

Machine Learning with Python Cookbook, Chris Albon

Antonio, N., Almeida, A. de, & Nunes, L. (2019, July 8). Data Science Journal. https://datascience.codata.org/articles/10.5334/dsj-2019-032/.

(PDF) Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue. ResearchGate. (n.d.). https://www.researchgate.net/publication/310504011_Predicting_Hotel_Booking_Cancellation_to_Decrease_Uncertainty_and_Increase_Revenue.

Adnan, D. (2020, August 26). *Predicting a Hotel Booking Demand*. Medium. https://towardsdatascience.com/predicting-a-hotel-booking-demand-7608a7dbf5a4.

EmelLike79Comment12ShareLinkedInFacebookTwitter0, M. B. F. D. S. |, & Follow. (n.d.). *Predicting Hotel Booking Cancellations Using Machine Learning - Step by Step Guide with Real Data and Python*. LinkedIn. https://www.linkedin.com/pulse/u-hotel-booking-cancellations-using-machine-learning-manuel-banza.

Vamshi, C. (2020, July 19). *Exploratory Data Analysis(EDA) For Predicting Hotel Booking Cancellations Using Machine Learning*. Medium. https://medium.com/analytics-vidhya/exploratory-data-analysis-eda-for-predicting-hotel-booking-cancellations-using-machine-learning-3990be4af2ff.

Wikimedia Foundation. (2020, December 25). *Online hotel reservations*. Wikipedia. https://en.wikipedia.org/wiki/Online_hotel_reservations.