Hotel Booking

Group Assignment

Course: ALY6015 Intermediate Analytics

Instructor: Fidel Rodriguez

Team - 1

Pragati Koladiya | NUID: 00102944

Tanvi Bhagat | NUID: 001083830

Priyanka Kanukollu | NUID: 001021111

Agenda

- Introduction
- Exploratory Data Analysis (EDA)
- Pre-processing
- Models
- Comparisons of models
- Conclusion

Aim

- Figure out the standard patterns of booking a hotel room.
- Show some booking information details with Exploratory Data Analysis(EDA)
- Find best model to predict hotel booking cancellations.
- The goal of predictive analysis is to avoid overfitting and find the model that gives the highest accuracy.

Dataset Information

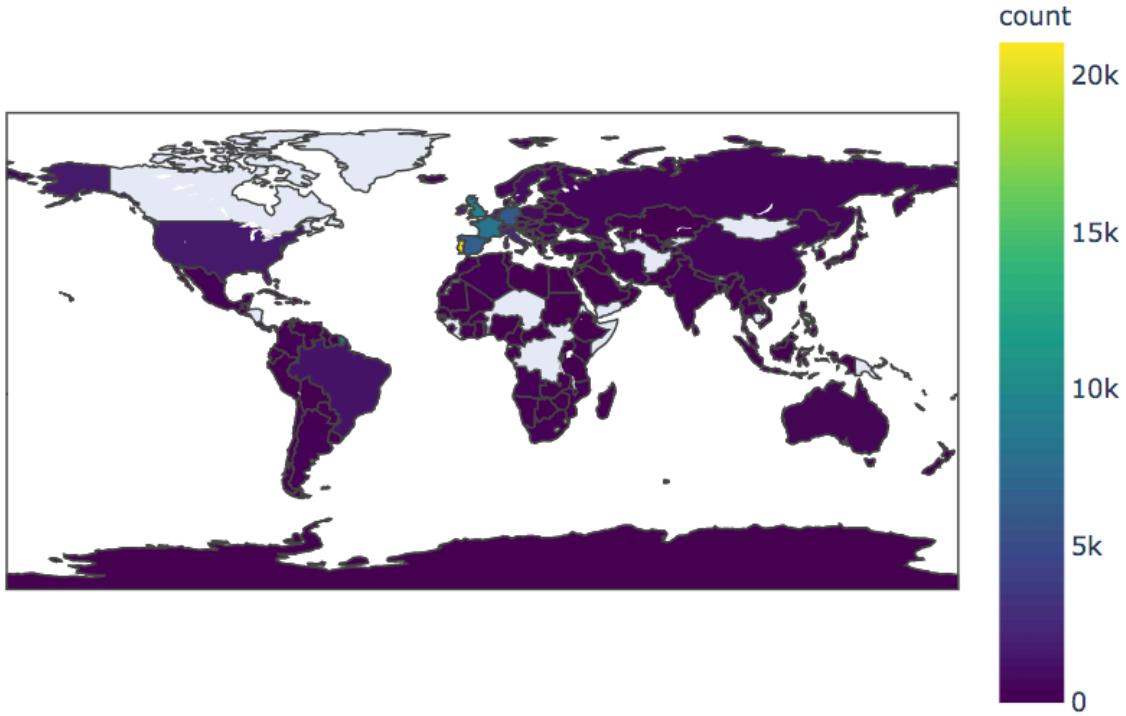
Data which compares various booking information between two hotels: a city hotel and a resort hotel.

Includes 31 features such as when the booking was made, length of stay, average daily rate, room type and the booking cancellations, among other things.

The data contains "bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017".

Exploratory Data Analysis (EDA)

1. Booking data by country of origin
2. Confirmation vs Cancellations?
3. Which year had most canceled bookings?
4. Which month had the most bookings?
5. Per month arrivals?
6. Per day arrivals?
7. How many number of stays on weekend and weeknights?
8. Price: Resort vs City?
9. How much do guests pay for a room per night?
10. How long do people stay at the hotels?

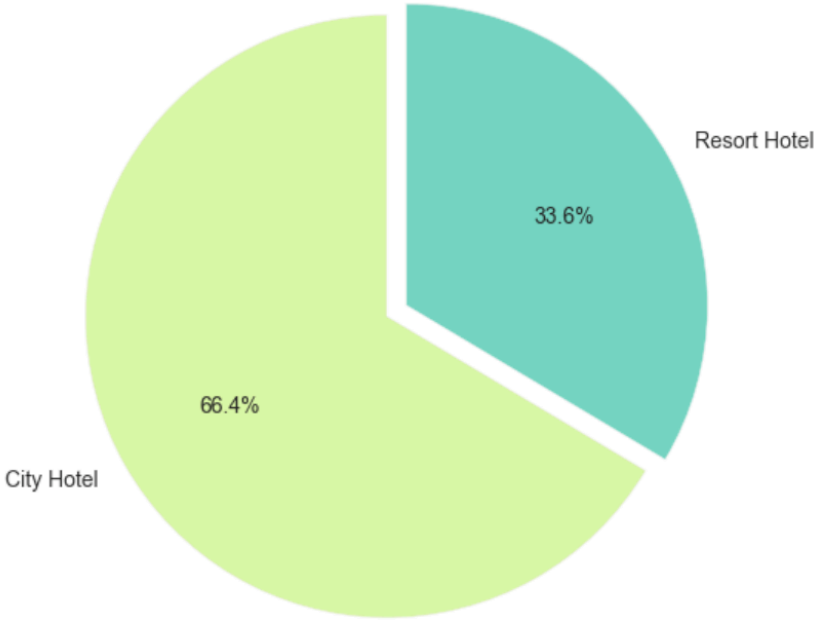


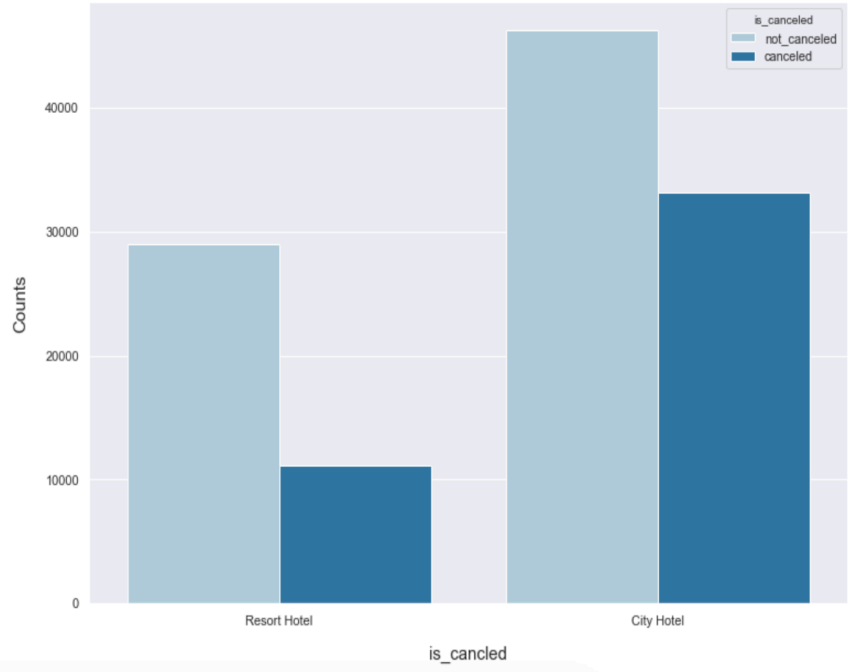
- People from all over the world are staying in these two hotels.
- Most guests are from Portugal and other countries in Europe.

Booking data by country of origin

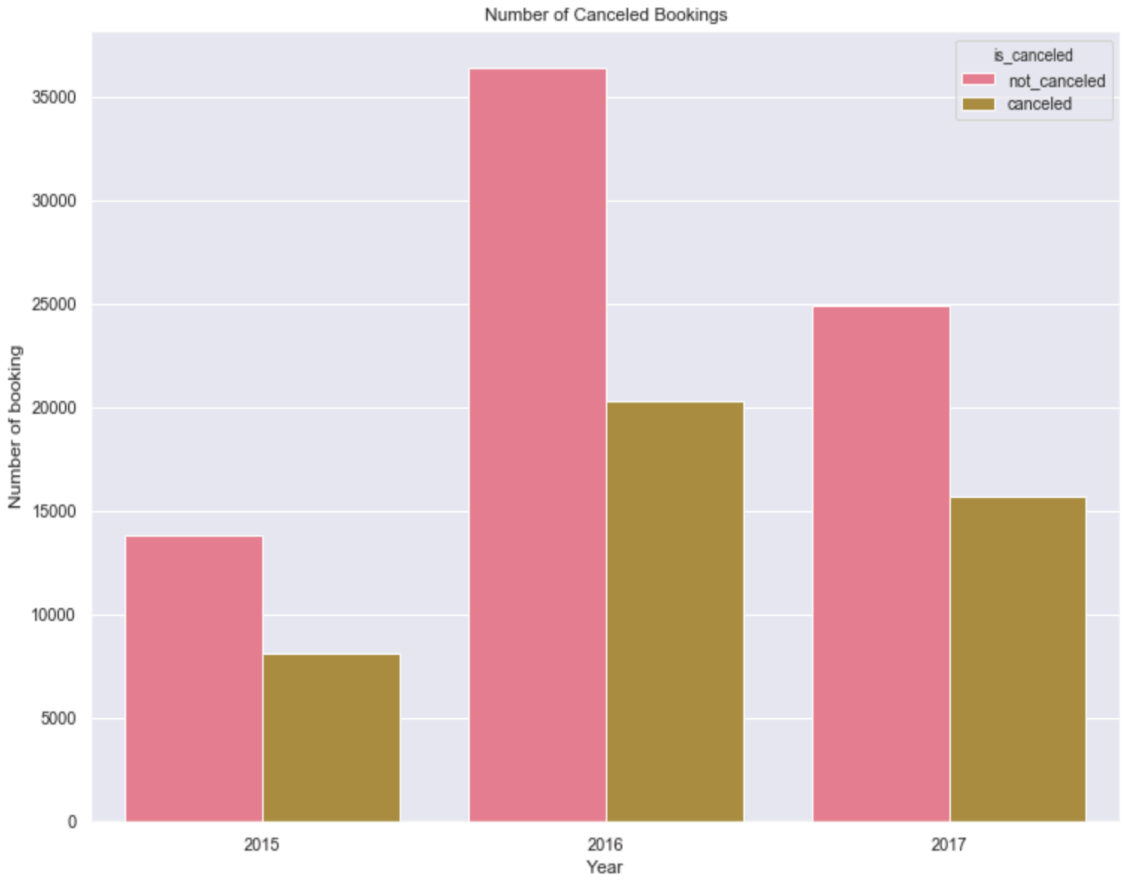
- Count of top 10 countries visitors:
- Portugal(PRT) - 48590
- United Kingdom(GBR) - 12129
- France(FRA) - 10415
- Spain(ESP) - 8568
- Germany(DEU) - 7287
- Italy(ITA) - 3766
- Ireland(IRL) - 3375
- Belgium(BEL) - 2342
- Brazil(BRA) - 2224
- Netherlands(NLD) - 2014

Confirmation vs Cancellations





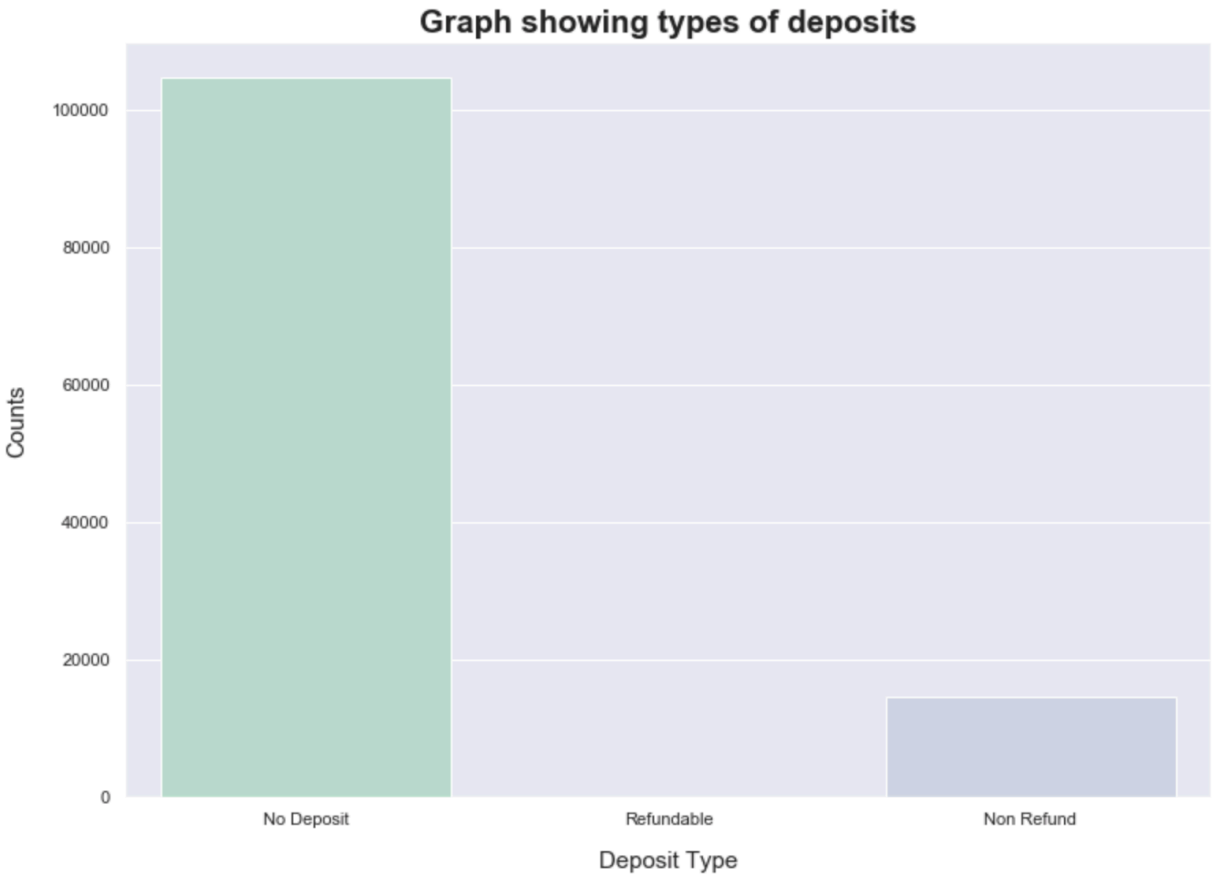
- From the pie chart, we can see that two third of people chose the city hotel option.
- As a greater number of people prefer the City Hotel, it has the largest number of canceled and confirmed bookings.



Total bookings canceled: 44,224 (37 %)
Resort hotel bookings canceled: 11,122 (100 %)
City hotel bookings canceled: 33,102 (100 %)

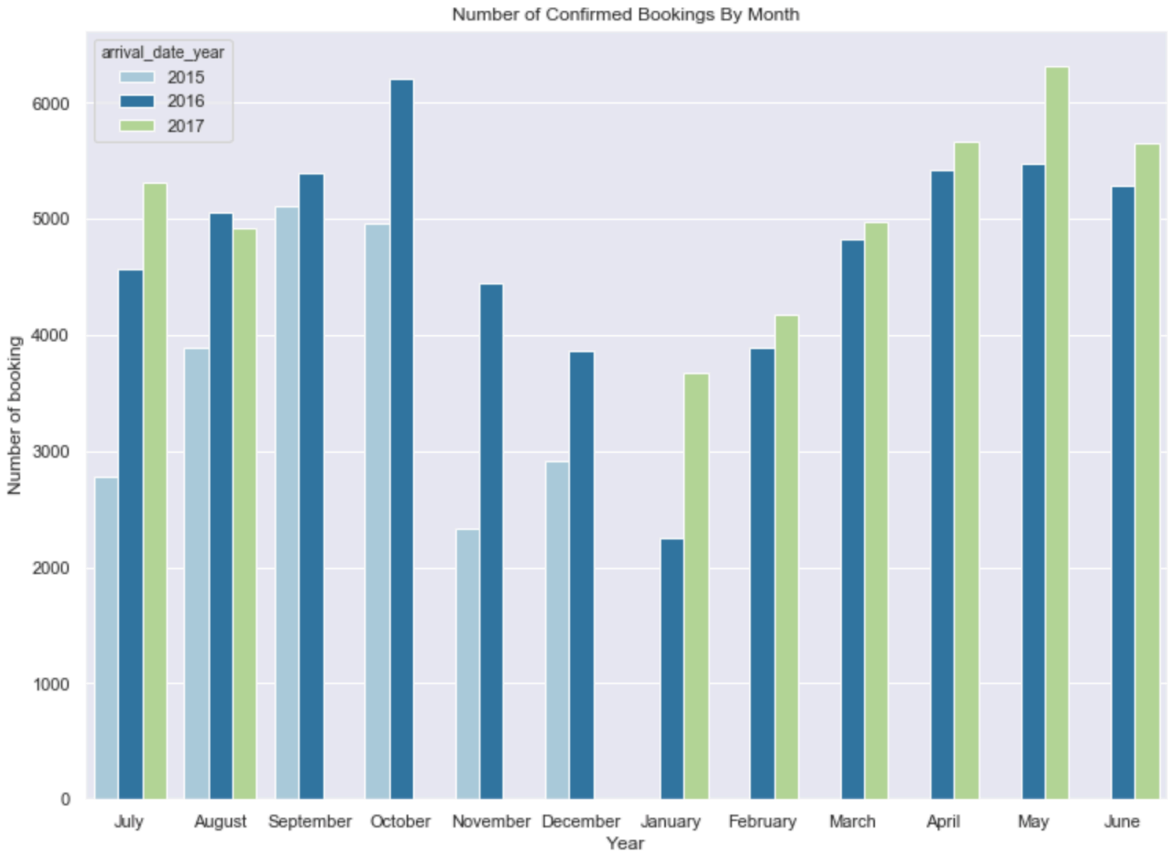
Which year had most canceled bookings?

- 2016 is the year that had the greatest number of canceled bookings as well as confirmed bookings.
- Overall, 37% of the total bookings for both the hotels were canceled.



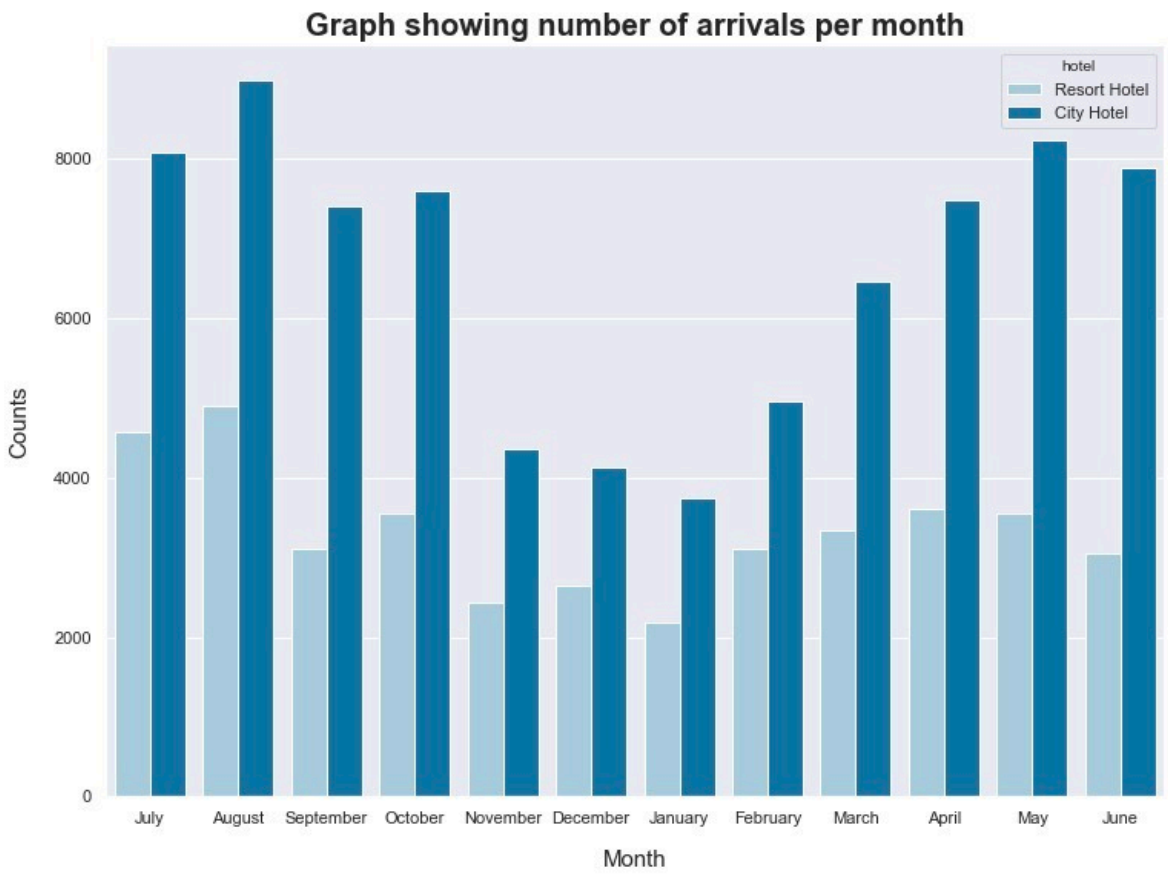
Deposit types

- Few bookings were non-refundable.
- Majority of the bookings did not need any deposits; this explains the high number of cancellations.



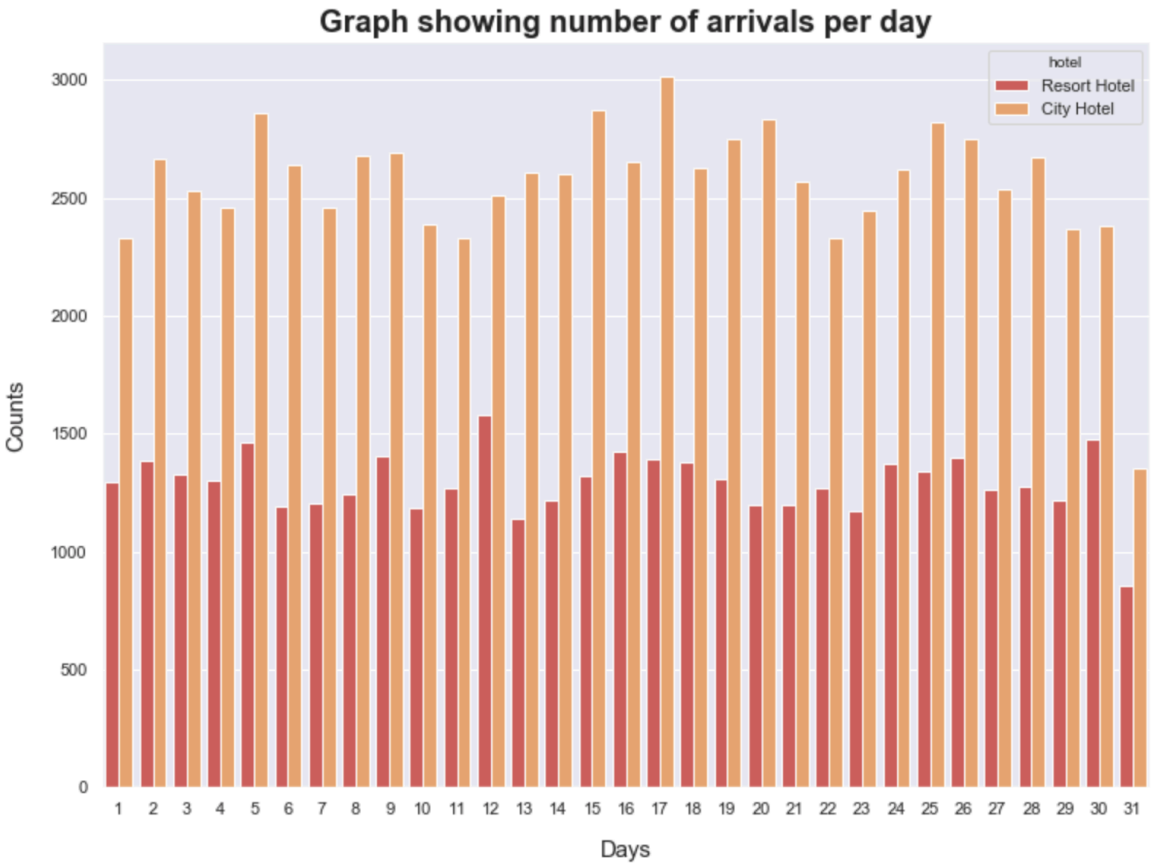
Which month had the most bookings?

- May 2017 had the greatest number of bookings followed by October,2016.
- No Bookings from September to December for the year 2017.



Per month arrivals?

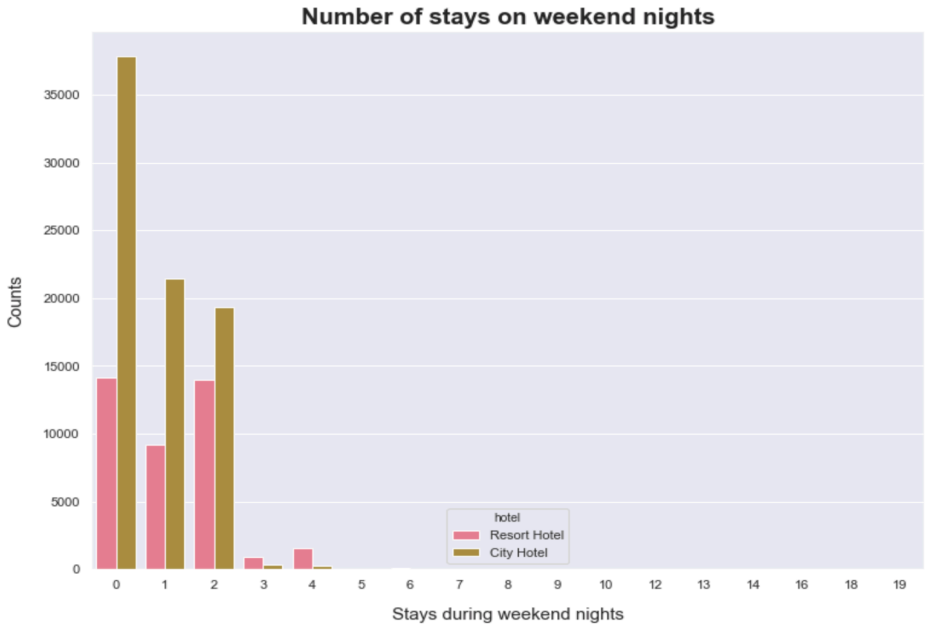
- The trend shows that bookings occur at the highest rate around the middle of year, with August being the highest.
- City hotels have more customers in all months. Considering proportionally, resort hotels seem to be a little closer to city hotels in summer.
- Fewer customers come in the winter months, so when we look at the cancellation rates, it is quite normal that it appears less in the winter months.

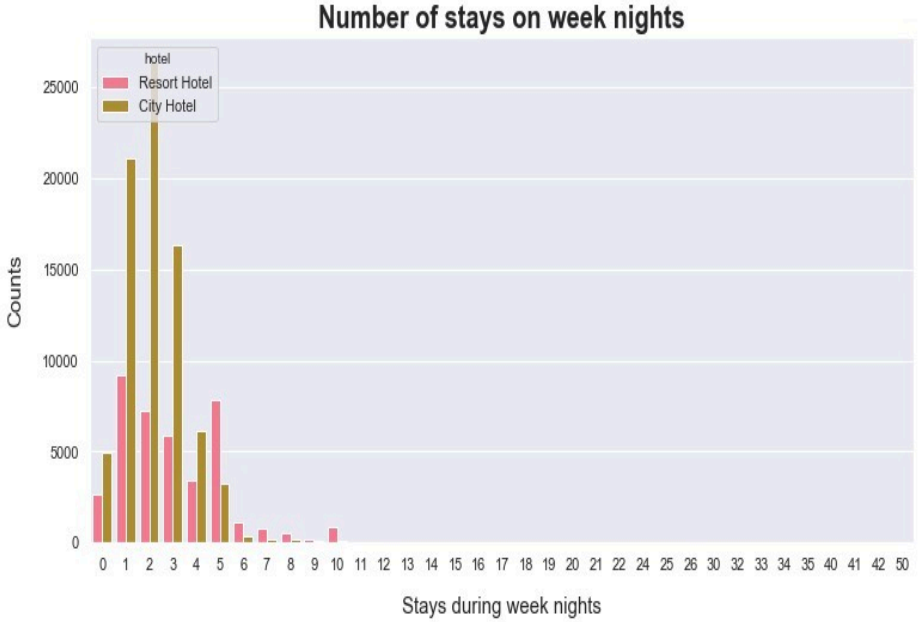


Per day arrivals?

- There is a wave like structure to arrivals by day.
- My speculation is that these peaks depict hotel bookings on the weekends at a higher rate.

How many number of stays on weekend and weeknights?



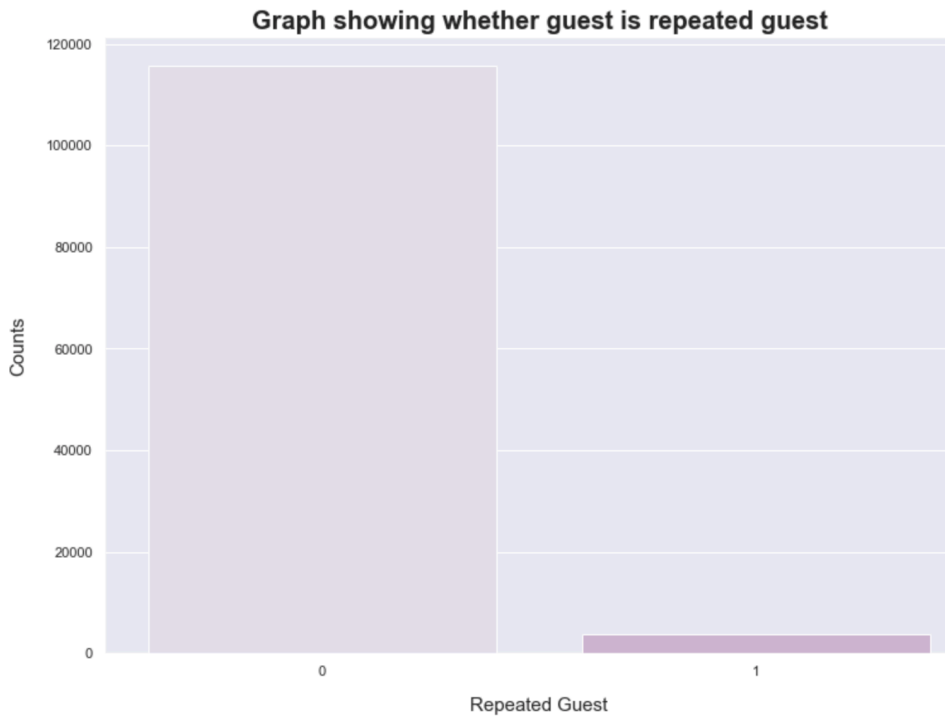


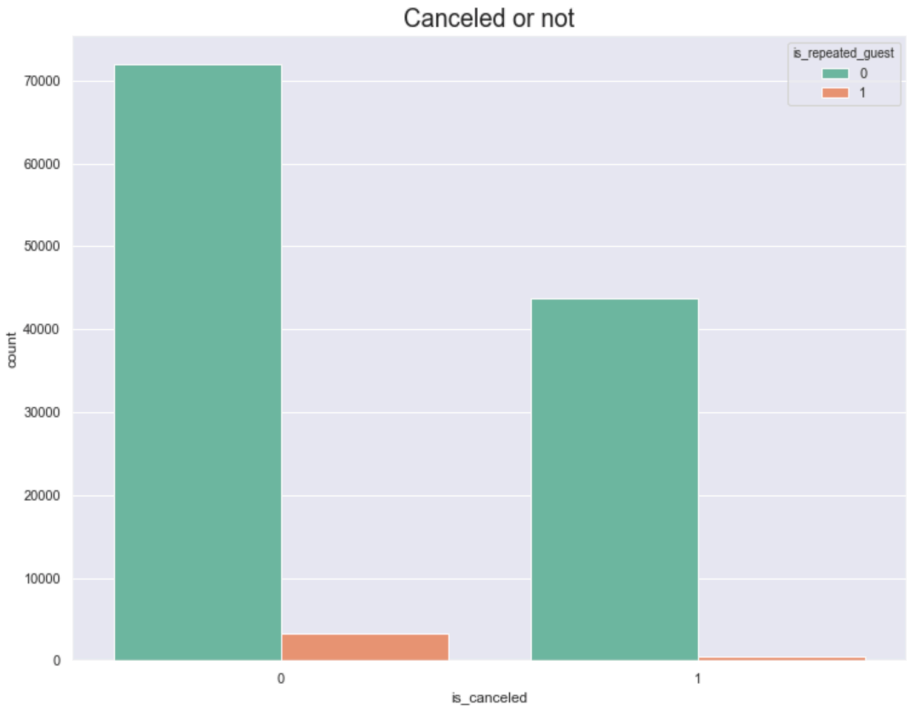
- Our hypothesis was proven false as most stays were on weekday nights.

Price: Resort vs City

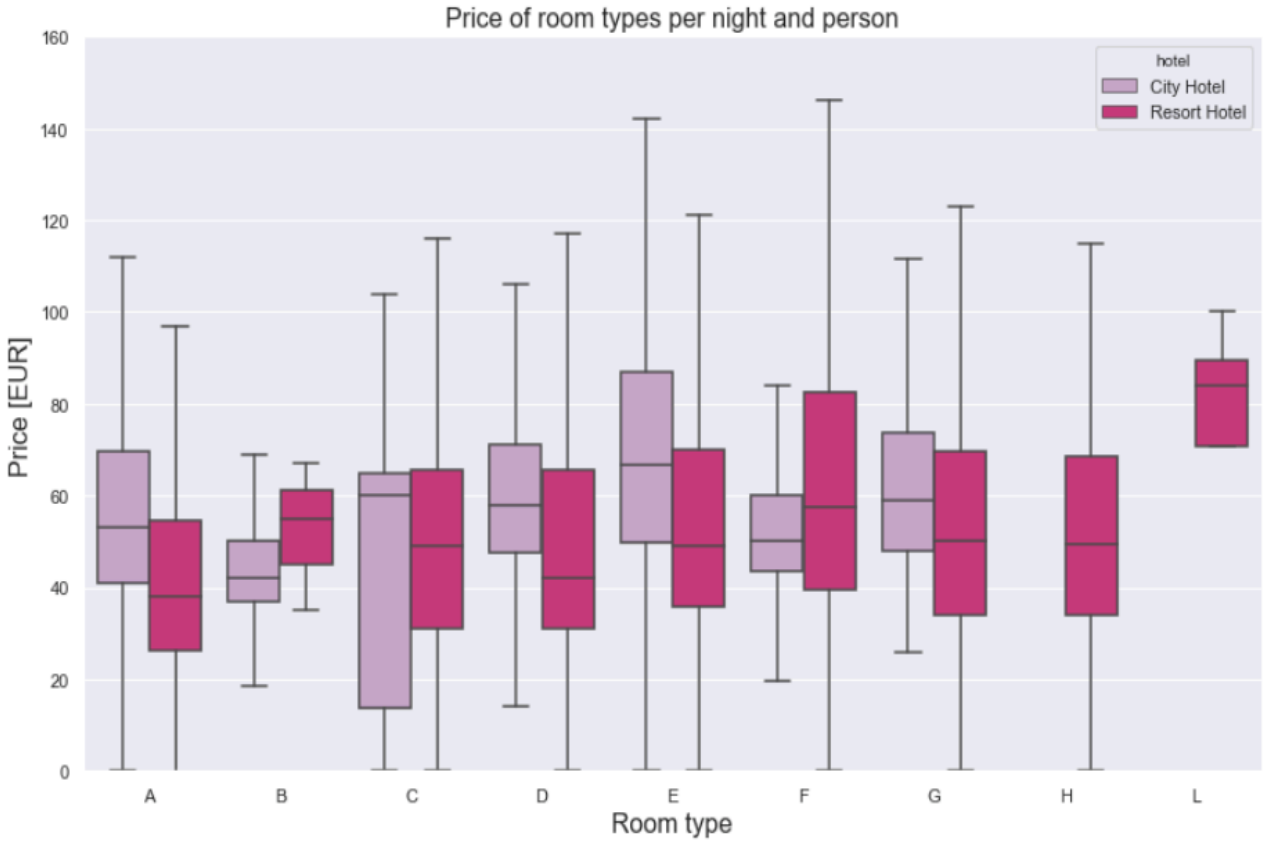
- Prices of the resort hotel are typically higher than the city hotel.
- With highest rates during busy months of travel in the summer, ie. August, June, and July.

Repeated Guests?





- There is no surprise that repeated guests do not cancel their reservations.
- Of course there are some exceptions. Also most of the customers are not repeated guests.

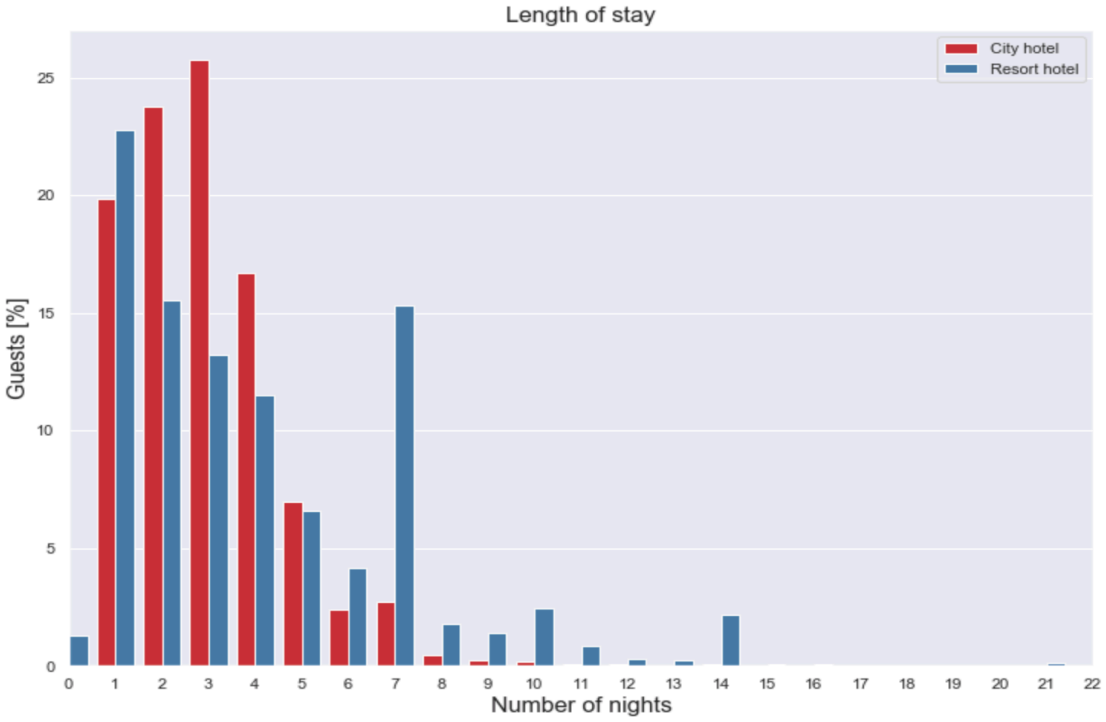


How much do guests pay for a room per night?

- Both hotels have different room types and different meal arrangements.
- Seasonal factors are also important. So the prices vary a lot.
- From all non-cnceled bookings, across all room types and meals, the average prices are:
- Resort hotel: 47 € per night and person.
- City hotel: 59 € per night and person.

How long do people stay at the hotels?

- For the city hotel there is a clear preference for 1-4 nights.
- For the resort hotel, 1-4 nights are also often booked, but 7 nights also stand out as being very popular.



Other Observations

- People usually made booking three months ahead of their arrival time.
- Almost 80% of bookings reserved for Bed&Breakfast.
- 67% of Bed&Breakfast booking made for City Hotel and almost every Full Board bookings made in the Resort Hotel.
- More than 45% of bookings are made via Online Travel Agents and less than 20% of bookings made directly without any agents.
- Half bookings without any special requests have been canceled and another half of them have not been canceled.
- More children are booked into resort hotels.

Preprocessing

- **Checking missing values:**
  94% of the data from the '*company*' column and 13% from *'Agent'* column is missing.

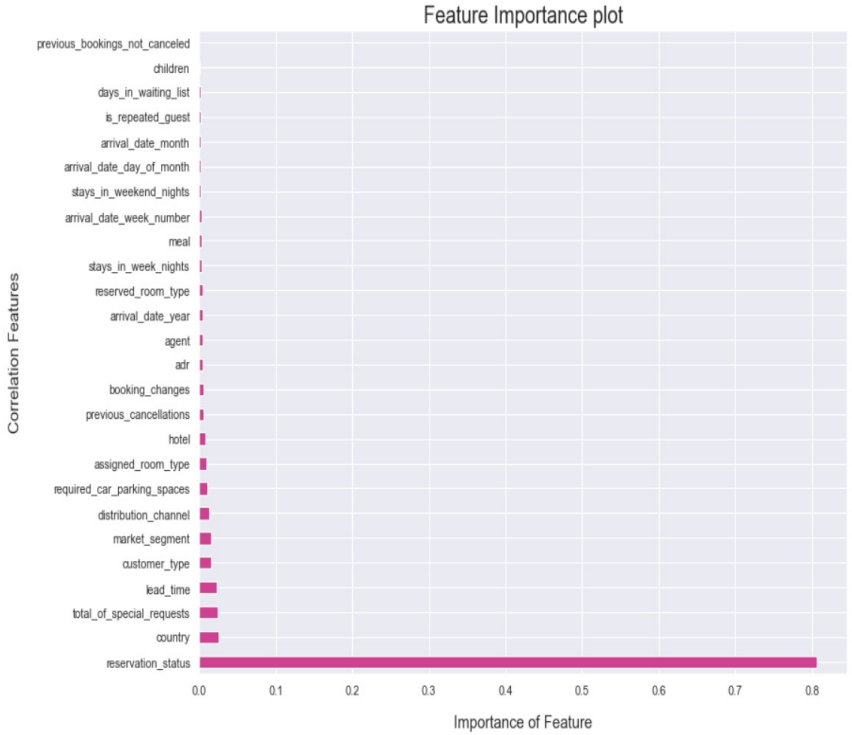
- **Dropping columns:**
  To leverage data for better efficiency and accuracy, we dropped few columns which are high in missing value rate.(We dropped 7 columns from the dataset)

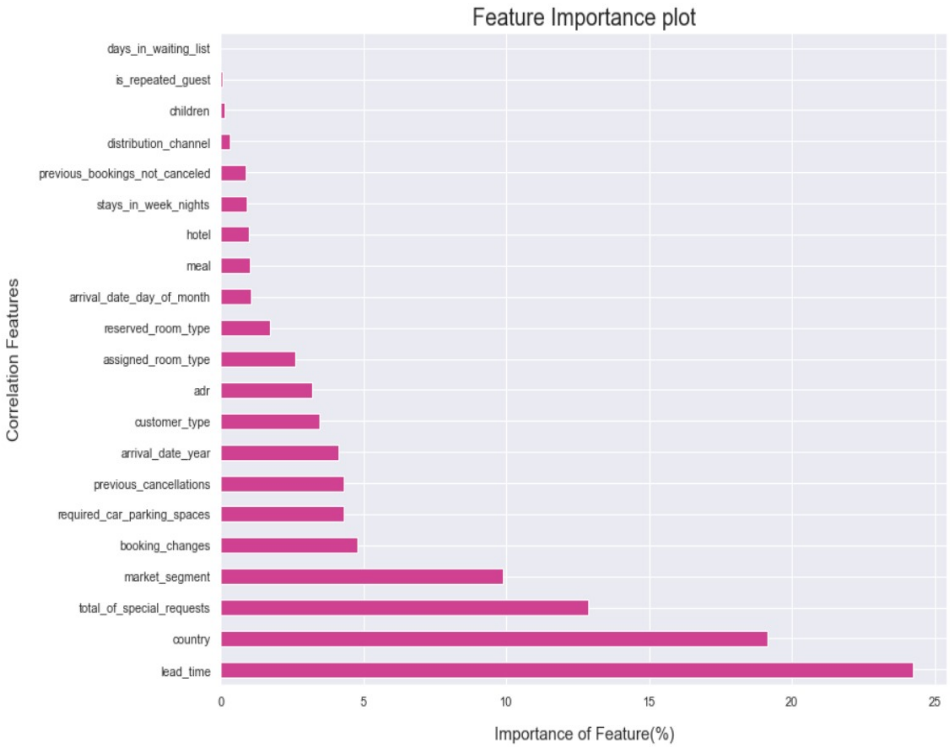
- **Label Encoder:**
  To transform non-numeric values in data such as *hotel*, *reservation_status* etc.

- **Feature Selection:**
  To identify the correlation between the columns.

- **Final arrangements before model comparison:**
  Sorting important features and checking null values.

Feature Importance





- By keeping reservation_status in data, it is possible to achieve 100% accuracy rate because that feature is direct way to predict cancellations.
- For the sake of analysis we droped reservation_status and continue analysis without it.

Models Implemented

- **Decision Tree:**

Used to solve regression and classification problems and can handle colinearity better than logisyic regression

- **Logistic Regression:**

Liblinear: Scikit solver functions for multi class logistic regression with one vs rest.
Lbfgs: Overcome the drawbacks of liblinear and performs fast by saving memory.

- **Knn Classifier:**

Faster compared to other classification models.
Knn optimum - To check how accurately the classifier or model can predict.

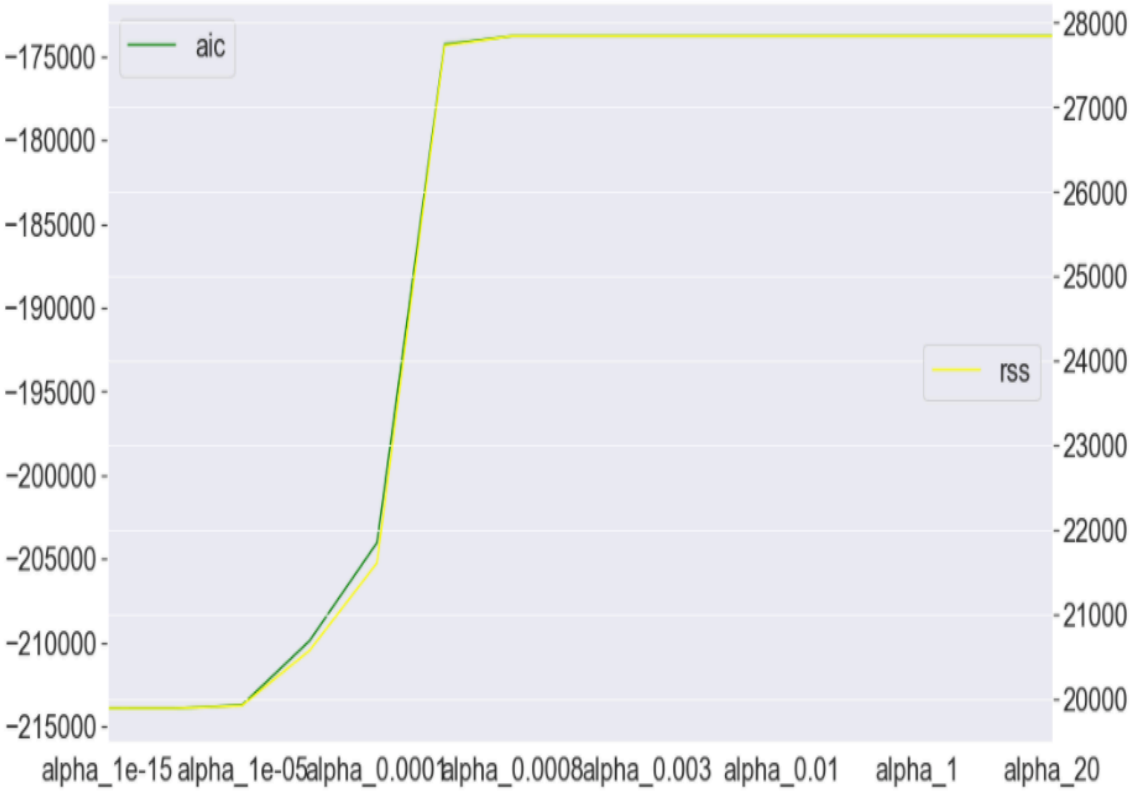
- **Random Forest Classifier:**

To cancel the biases, for robust method and accuracy.
Grid search - Tuning the model by using few parameters.

- **Lasso Regularization model:**

For best feature selection.

Plot for RSS vs AIC with different alpha values

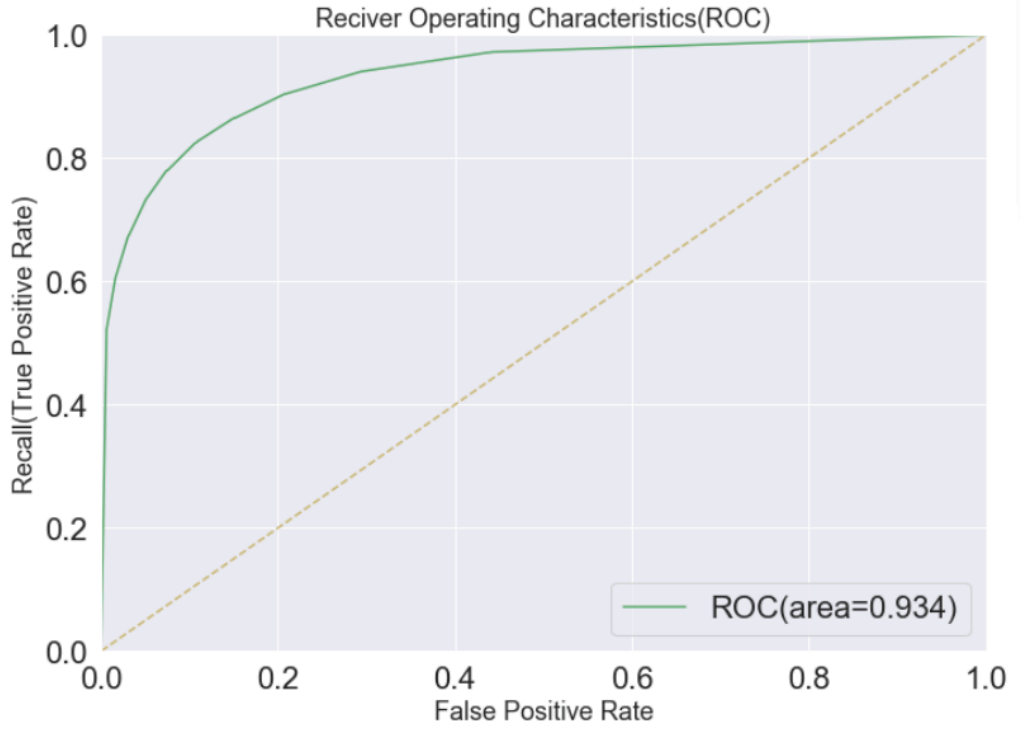


Lasso Regularization

Comparison Of Models

MODEL	ACCURACY SCORES	ROC/AUC
Decision Tree(reservation_status included)	1.00	1.00
Decision Tree(reservation_status excluded)	0.84	0.92
Logistic Regression(Liblinear)	0.78	0.83
Logistic Regression(Lbfgs)	0.73	0.77
KNN Classifier	0.79	0.85
KNN Classifier(with optimum k value)	0.78	0.83
Random Forest	0.87	0.93
Random Forest Tuned	0.84	0.95

We observe that, the best algorithm is random forest for this data set.



Random Forest

ROC/AUC

Conclusion

- All the models with *'reservation_status'* column included gives 100% accuracy score.
- After excluding *'reservation_status'* column we observed a decrease in accuracy score.
- We created arbitrary values for our dataset and predicted *'is_canceled'* by using the highest accuracy score classifier(Random Forest).

Link & References

[1]Machine Learning Basics with the K-Nearest Neighbors Algorithm
https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[2] Logistic Regression — Detailed Overview
https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[3]Understanding Random Forest
https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[4] Ridge and Lasso Regression: L1 and L2 Regularization
https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

[5] **Comparative Study on classification algorithms**
https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

- **Link to Jupiter Notebook**

http://localhost:8888/notebooks/OneDrive%20-%20Northeastern%20University/Intermediate%20Analysis/IntermediateAnalysis/ALY6015_SpringA2020_Group_Assignment_Team_1.ipynb

Thank you

Be Home, Stay Safe.