Prediction of hotel booking cancellations

Suraj Rana Vujjini

GJ58372

Department of Data Science

Data 602 – Intro to Data Analysis and Machine Learning

Christopher McGraw

Introduction

- Hotels thrive on advance reservation systems. Due to the current completive market, hotels need to have an easy cancellation policy for guests.
- The average percentage of canceled reservations is 24%. Cancellations effect the revenue of the hotel will lose potential revenue customers who will not cancel.
- As a part of the project, I will be analyzing bookings data from Microtel BWI and will be creating a model to predict if the reservation will cancel or not. This can help the hotel forecast the future revenue and also help price the rooms accordingly.

Aim of this project

- The main aim of this project is to create a model that finds reservations that have a high chance of getting cancelled. This solves problems like room management and forecasting income for the hotel management.
- The final model can predict the reservations that could be canceled with a good accuracy. These reservations are the ones mostly made by guests who are not sure about their stay at the hotel.

About the dataset

No. of Rows

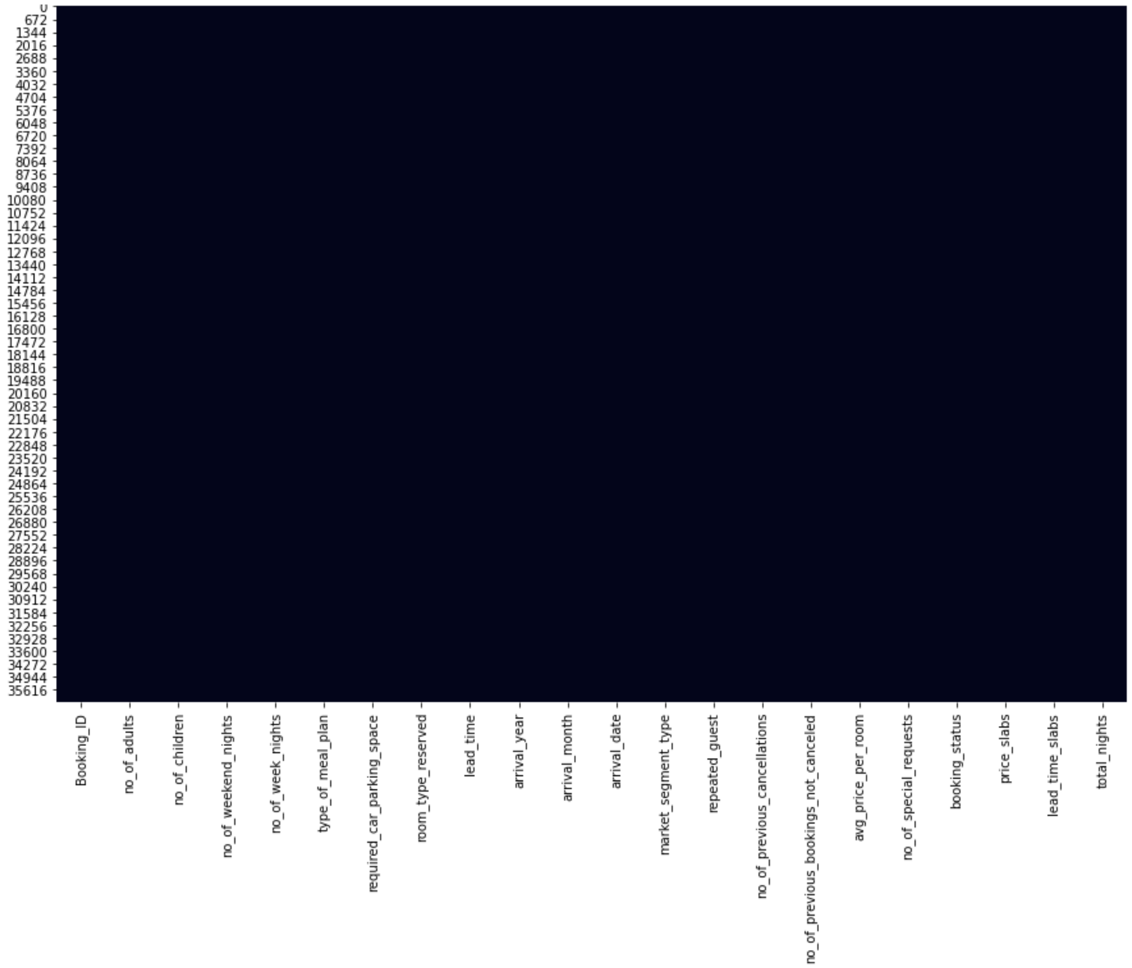
36,275

No. of columns

19

Total number of values

689,225



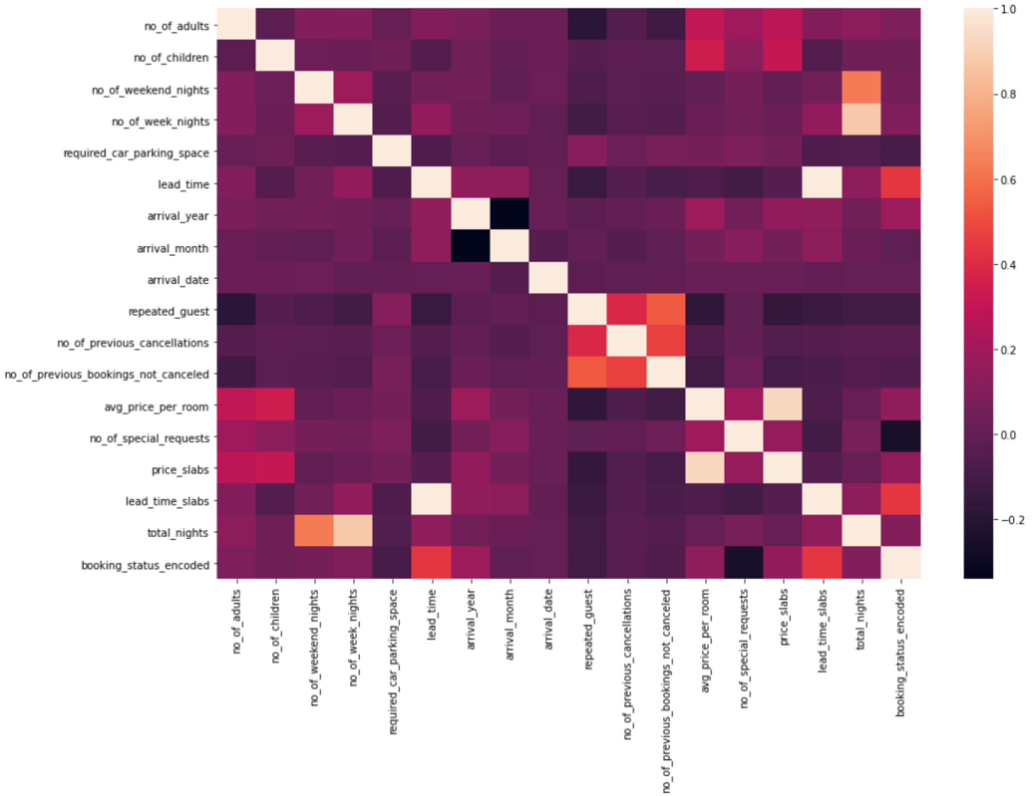
SNS heatmap of null values in the dataset

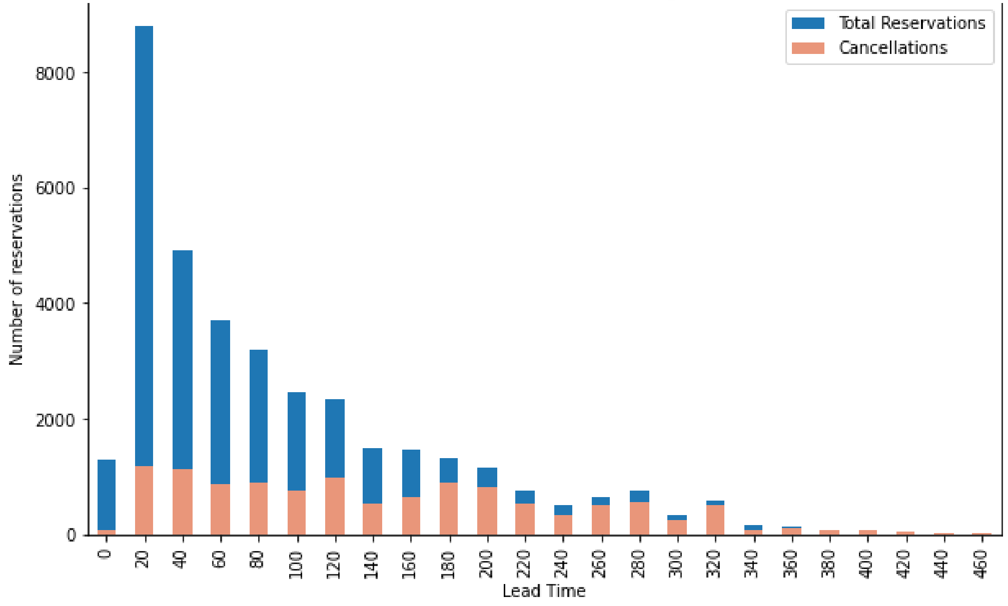
Columns in the dataset

```
#      Column                              Non Null value  Dtype                      Description
# 0   Booking_ID                        36275 non-null  object  - Unique identifier for a booking
# 2   no_of_children                        36275 non-null  int64   - No of children the guest specified in the booking
# 3   no_of_weekend_nights                  36275 non-null  int64   - No of weekend nights the reservation has
# 4   no_of_week_nights                     36275 non-null  int64   - No of weeknights the reservaiton has
# 5   type_of_meal_plan                     36275 non-null  object  - Type of meal plan chosen in the reservation
# 6   required_car_parking_space            36275 non-null  int64   - Number of parking spots chosen by the guest
# 7   room_type_reserved                    36275 non-null  object  - Type of reoom reserved by the guest
# 8   lead_time                             36275 non-null  int64   - The gap between day of booking to day of arrival
# 9   arrival_year                          36275 non-null  int64   - Arrival year of the reservation made
# 10  arrival_month                         36275 non-null  int64   - Arrival month of the reservaton made
# 11  arrival_date                          36275 non-null  int64   - Day of the month of arrival of reservation
# 12  market_segment_type                   36275 non-null  object  - Denotes if the market is online or offline
# 13  repeated_guest                        36275 non-null  int64   - This column shows if the guest is
# 14  no_of_previous_cancellations          36275 non-null  int64   - This column shows the number of times the guest cancelled
# 15  no_of_previous_bookings_not_canceled  36275 non-null  int64   - This column shows the number of times the guest did not
#                                                                     cancel
# 16  avg_price_per_room                    36275 non-null  float64 - Average price for the rooms reserved
# 17  no_of_special_requests                36275 non-null  int64   - Number of special requests made by the guests
# 18  booking_status                        36275 non-null  object  - Booking status canceled or not
# 19  price_slabs                           36275 non-null  int64   - Average price of rooms sorted into respective price slabs
# 20  lead_time_slabs                       36275 non-null  int64   - Lead times of reservations sorted into respective
#                                                                     price slabs
# 21  total_nights                          36275 non-null  int64   - Sum of week nights and weekend nights
```

Correlation among the columns

- I have created a new column “booking_status_ecoded” to use in the correlation heatmap.
- Comparatively there is a good correlation of the booking status column with lead time, as the lead time increases, there is a high chance of a booking getting cancelled.



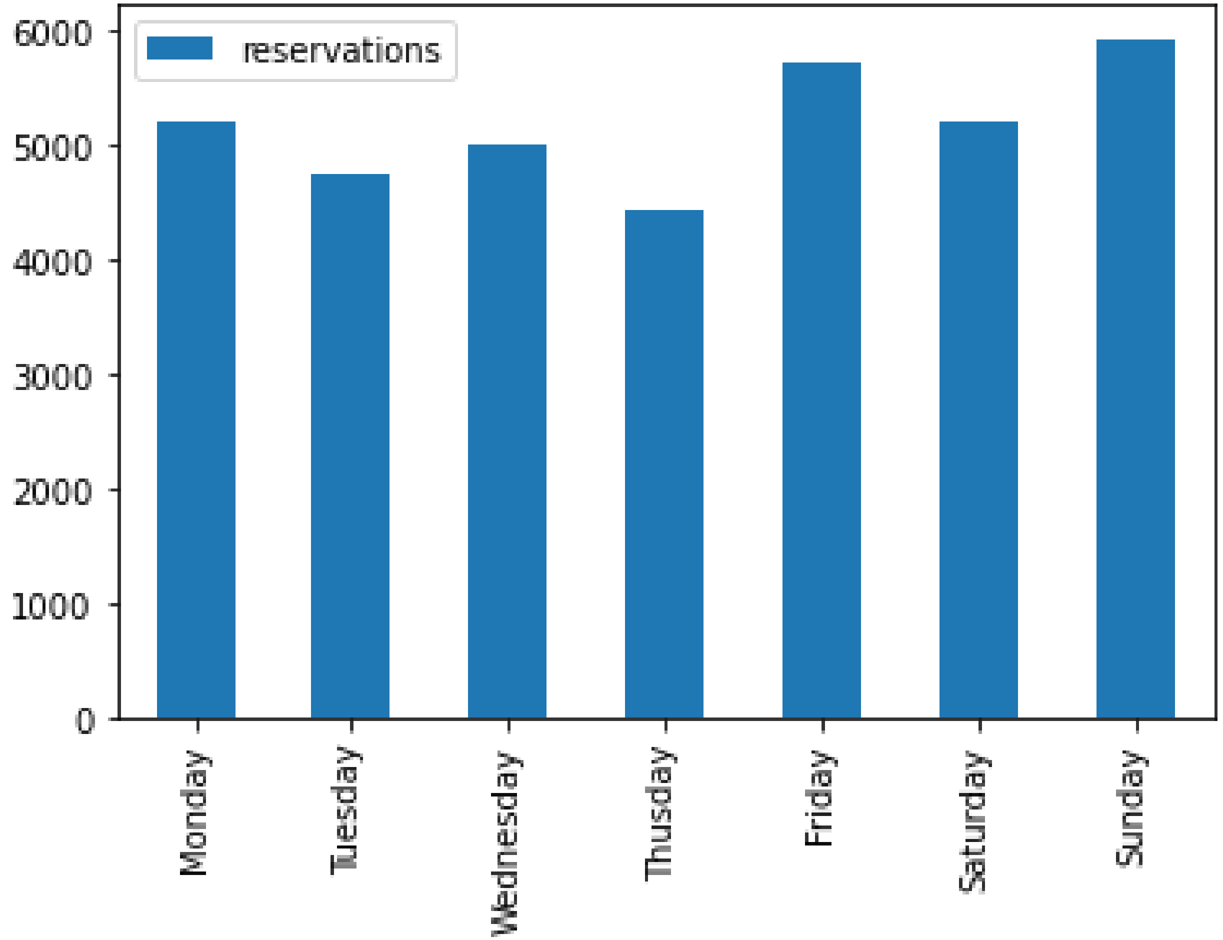


Cancellations WRT months

- As expected, the bookings increase for summer months and fall during wintertime.
- For this hotel, there was an increase in reservations for the month of October as well.
- The cancellations tend to vary with reservations as seen from the graph.

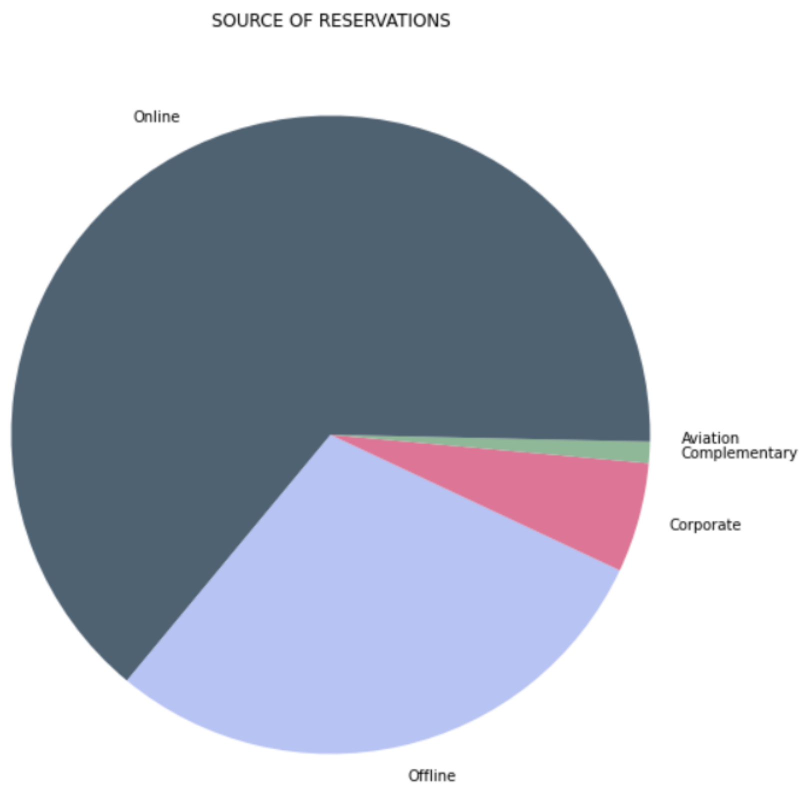
Reservations WRT day of the week

- The booking activity is highest on the weekends and lowest on the weekdays as expected.
- Friday and Sunday are the ones with highest number of reservations



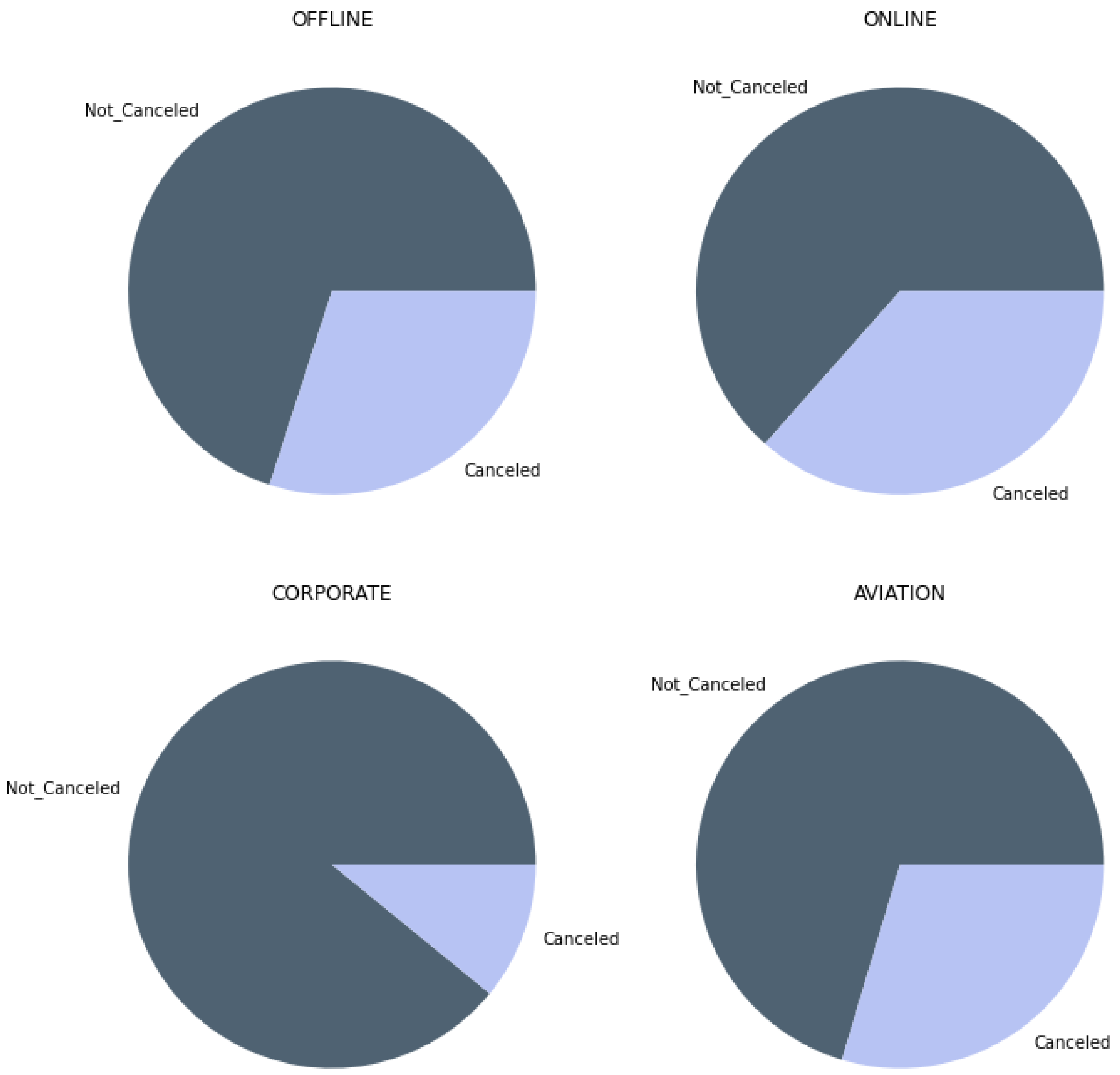
Source of reservations

- Online is the major source for reservations at Microtel followed by reservations made offline.
- This is due to the ease of use and lucrative offers given by the booking companies.



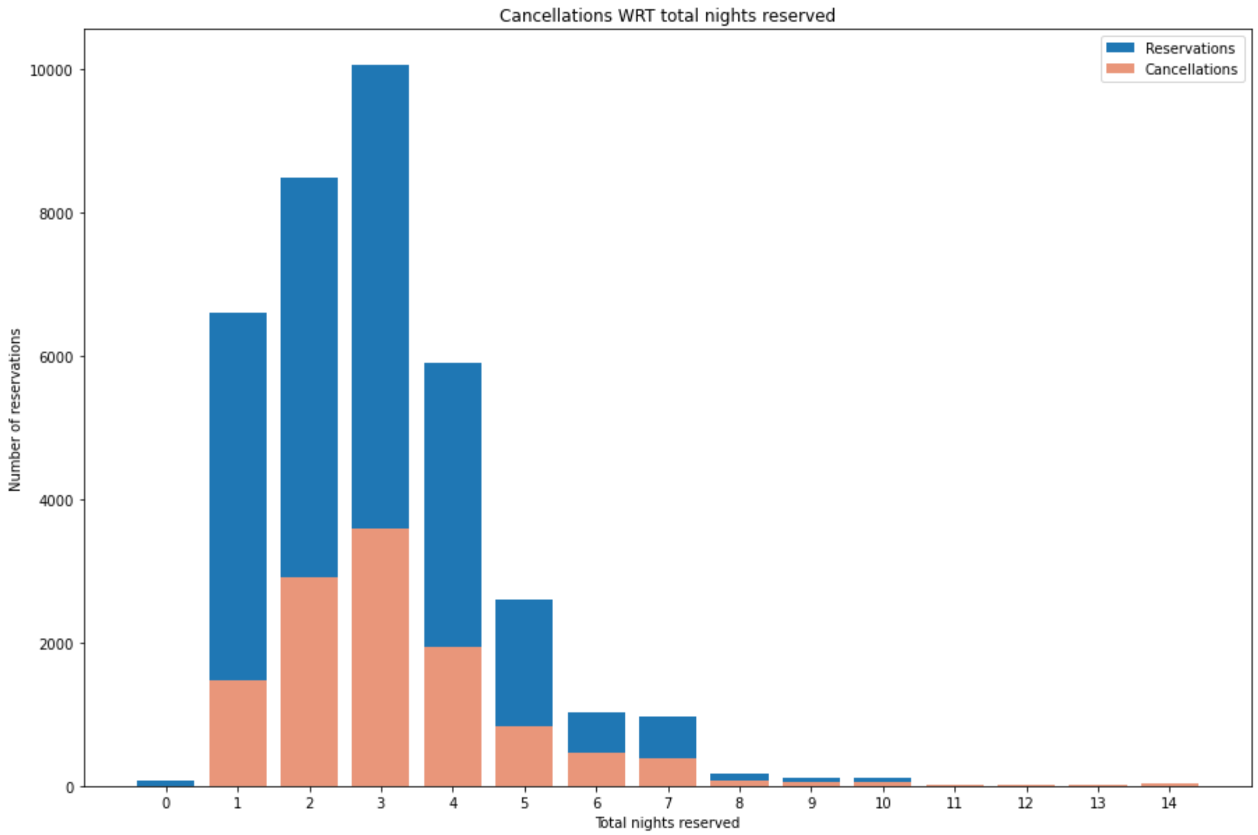
Cancellations WRT Source of reservations

- The easy cancellation policy of online booking sites allow users to cancel at their ease.
- This can be the reason for the high cancellation ratio for online bookings.
- It is followed by offline and aviation.
- Corporate has relatively low cancellations.



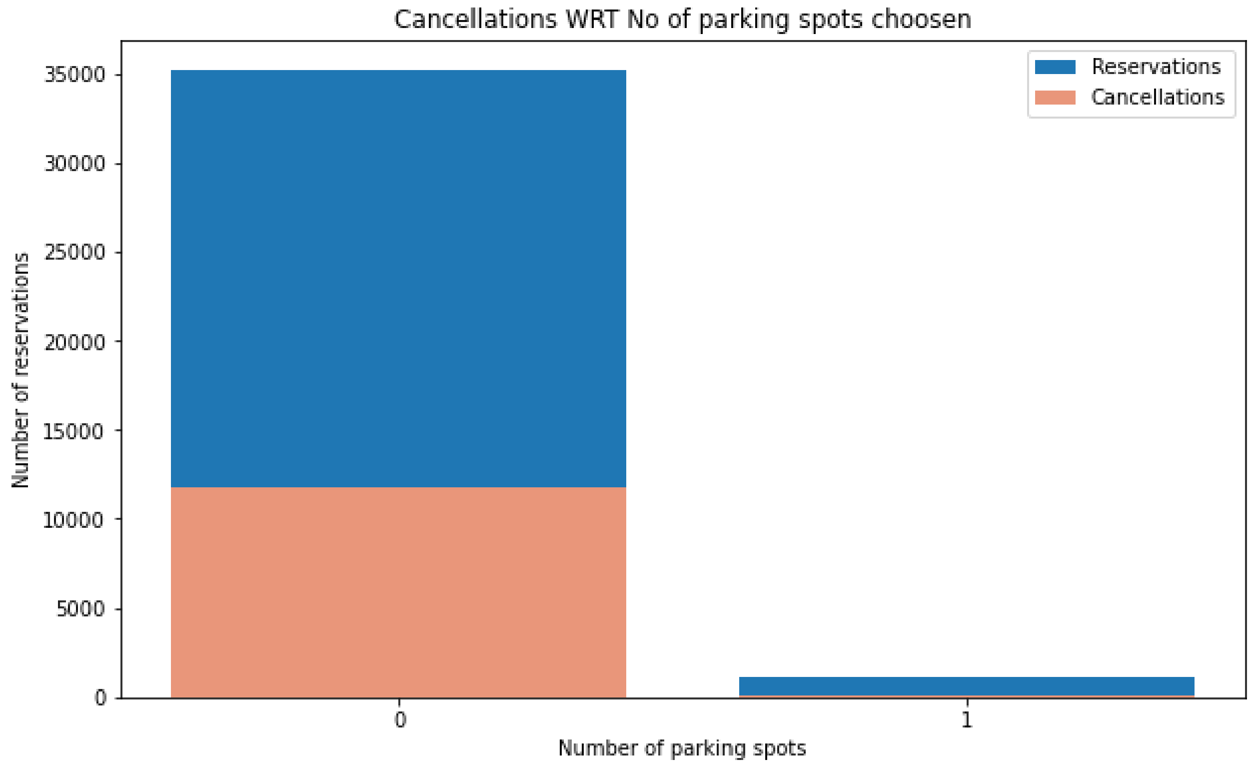
Cancellations WRT Total nights

- Most of the guest chose to stay for 3 nights.
- There is no good information for cancellations as the cancellations are higher when the reservations are high.



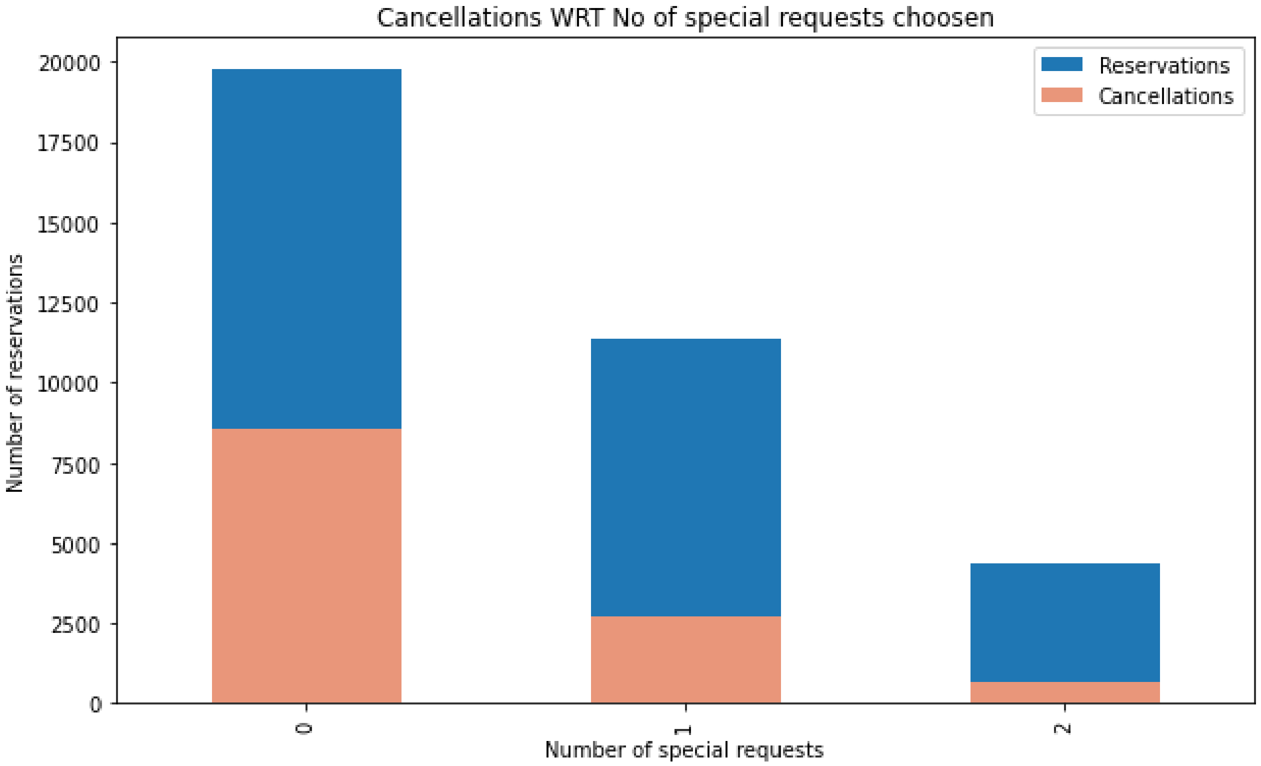
Cancellations WRT no. of parking spots chosen

- We can see that if guests choose to reserve a parking spot, they will most likely not cancel the reservation.
- If they do not reserve a parking spot, there is a higher chance of the reservation getting cancelled.



Cancellations WRT no. of special requests made

- Special requests are made for online order where a guest can request certain things during their stay at the hotel.
- This shows that a guest is serious about his/her stay at the hotel.
- As seen from the graph, as the number of special requests increases, there is less chance of a booking getting cancelled.



Splitting the data & Preprocessing pipeline

```
def generate_splits():
    y = df['booking_status']
    X = df[[x for x in df.columns if x != 'booking_status']]

    return train_test_split(X, y, test_size=0.2, random_state=124)

X_train, X_test, y_train, y_test = generate_splits()
```

Training examples: 29,020

Test examples: 7,255

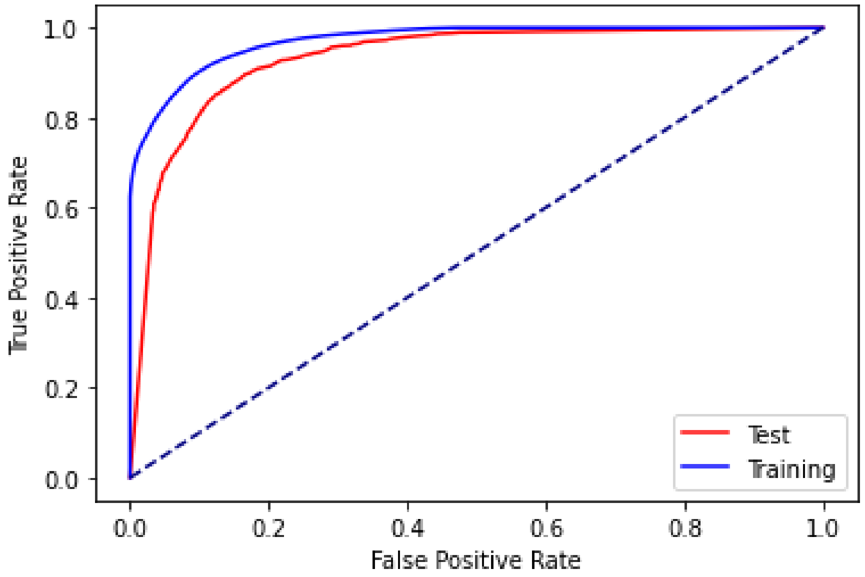
```
num_pipeline = Pipeline([('standardize_num', StandardScaler())])

cat_pipeline = Pipeline([('create_dummies_cats', OneHotEncoder(handle_unknown='error', drop='first'))])

processing_pipeline = ColumnTransformer(transformers=[('proc_numeric', num_pipeline, numerical_list),
                                                      ('create_dummies', cat_pipeline, categorical_list)])
```

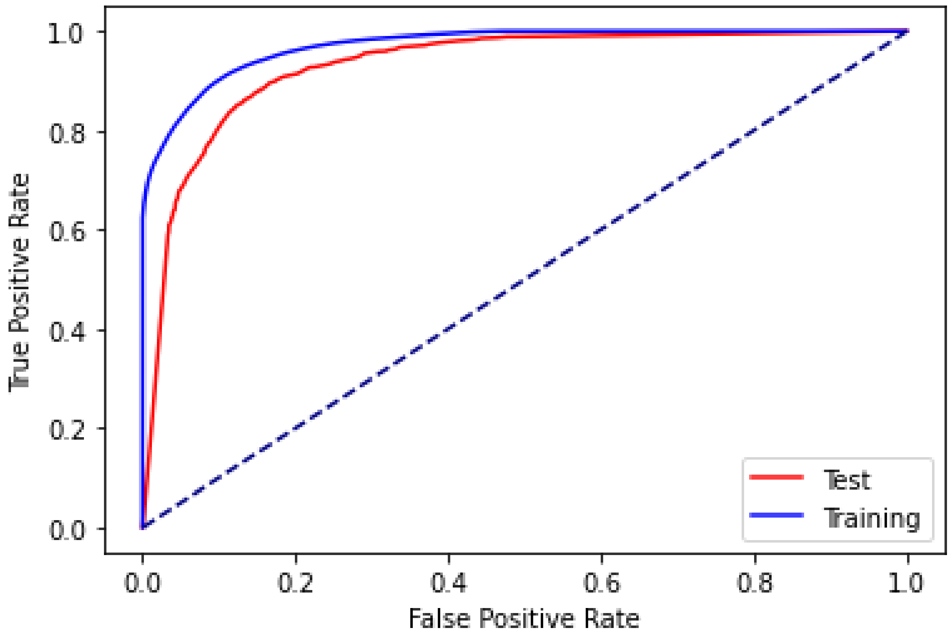
Classification – Logistic regression

	precision	recall	f1-score	support
Canceled	0.72	0.49	0.58	2395
Not_Canceled	0.78	0.90	0.84	4860
accuracy			0.77	7255
macro avg	0.75	0.70	0.71	7255
weighted avg	0.76	0.77	0.75	7255



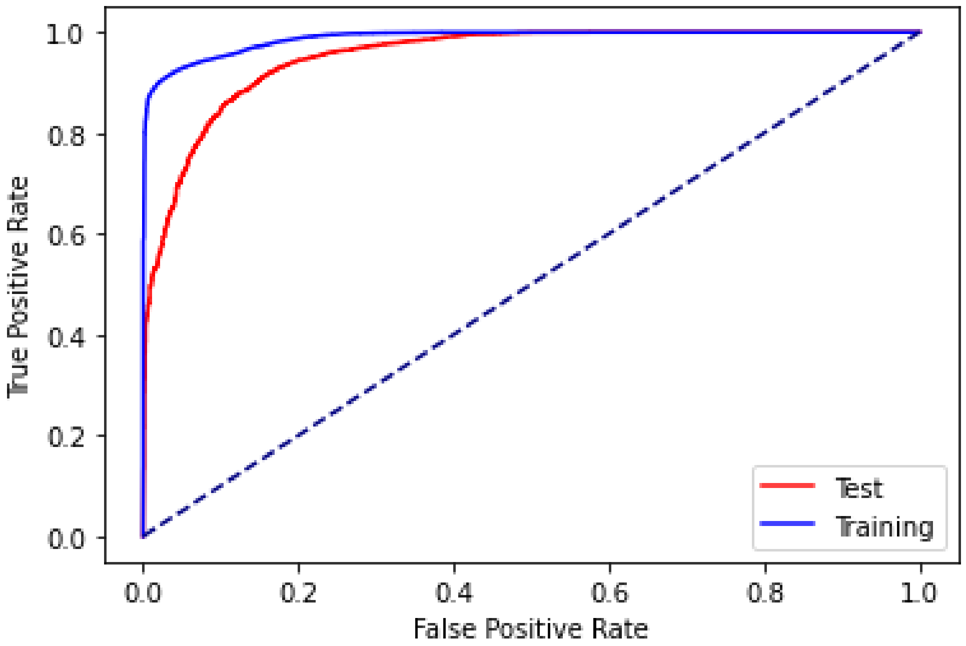
Classification – Decision Tree

	precision	recall	f1-score	support
Canceled	0.83	0.79	0.81	2395
Not_Canceled	0.90	0.92	0.91	4860
accuracy			0.88	7255
macro avg	0.86	0.86	0.86	7255
weighted avg	0.88	0.88	0.88	7255



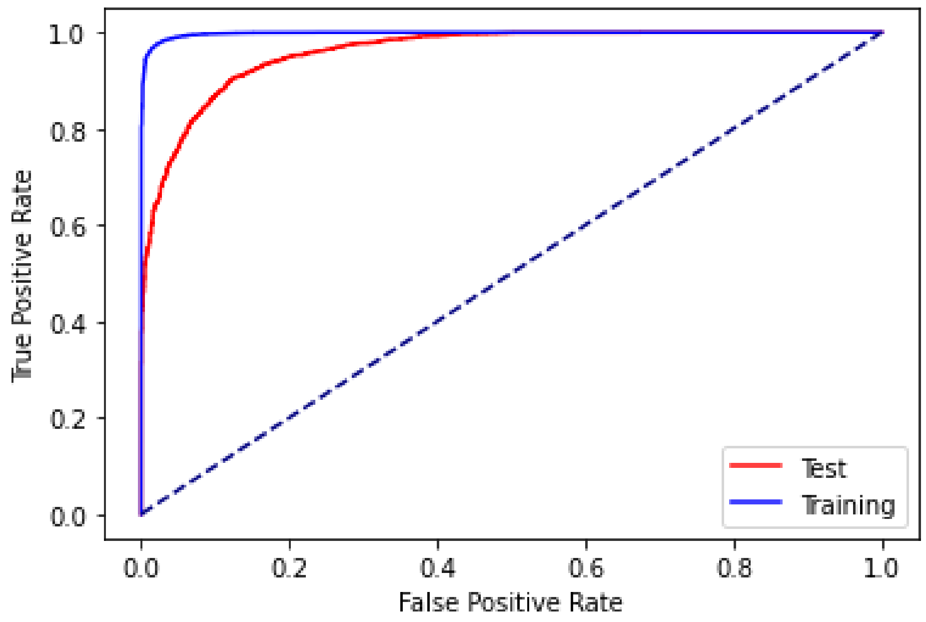
Classification – Random Forest

	precision	recall	f1-score	support
Canceled	0.88	0.79	0.83	2395
Not_Canceled	0.90	0.95	0.92	4860
accuracy			0.90	7255
macro avg	0.89	0.87	0.88	7255
weighted avg	0.89	0.90	0.89	7255



Classification – Gradient Boosting

	precision	recall	f1-score	support
Canceled	0.87	0.82	0.84	2395
Not_Canceled	0.91	0.94	0.92	4860
accuracy			0.90	7255
macro avg	0.89	0.88	0.88	7255
weighted avg	0.90	0.90	0.90	7255



Conclusion

To conclude, irrespective of the price, hotel reservations get cancelled due to several reasons. From these findings, it can be understood that if a guest has more requests or book a parking spot thereby validating their stay and tend to cancel less.
From all the classifiers I have used, I found that Random forest is the best classification algorithm that worked best for this dataset.

Thank you