ANALYSIS AND PREDICTIVE MODELING OF HOTEL CANCELLATION AND PRICE TRENDS

Soyoung An

Jun 2021

OVERVIEW

Often, hotel cancellation can lead to the business loss. Having ability to accurately predict future customers' cancellation rates is important. Predicting a hotel price rate is essential to evaluate demand for accommodation and pricing the room rates accordingly.

Predicting the actual demand and price can help to reduce the unexpected profit loss.

Goals

1. Identify the key factors that influence cancellation rate and price
2. Identify whether the average daily rates (ADR) between two groups (cancelled/not cancelled) are same
3. Develop a predictive models to estimate cancellation and price

Data

Hotel booking observations from a resort hotel and city hotel where the resort hotel located in Algarve and the city hotel in Lisbon.

- 40,060 rows 32 features

DATA WRANGLING

1. Processed date types data and stored them into year, month, day separately
2. Dropped any columns with more than 50% missing values
3. Implicated missing values with reasonable data
4. Dropped suspicious data with no additional information available

After Data Wrangling Process,

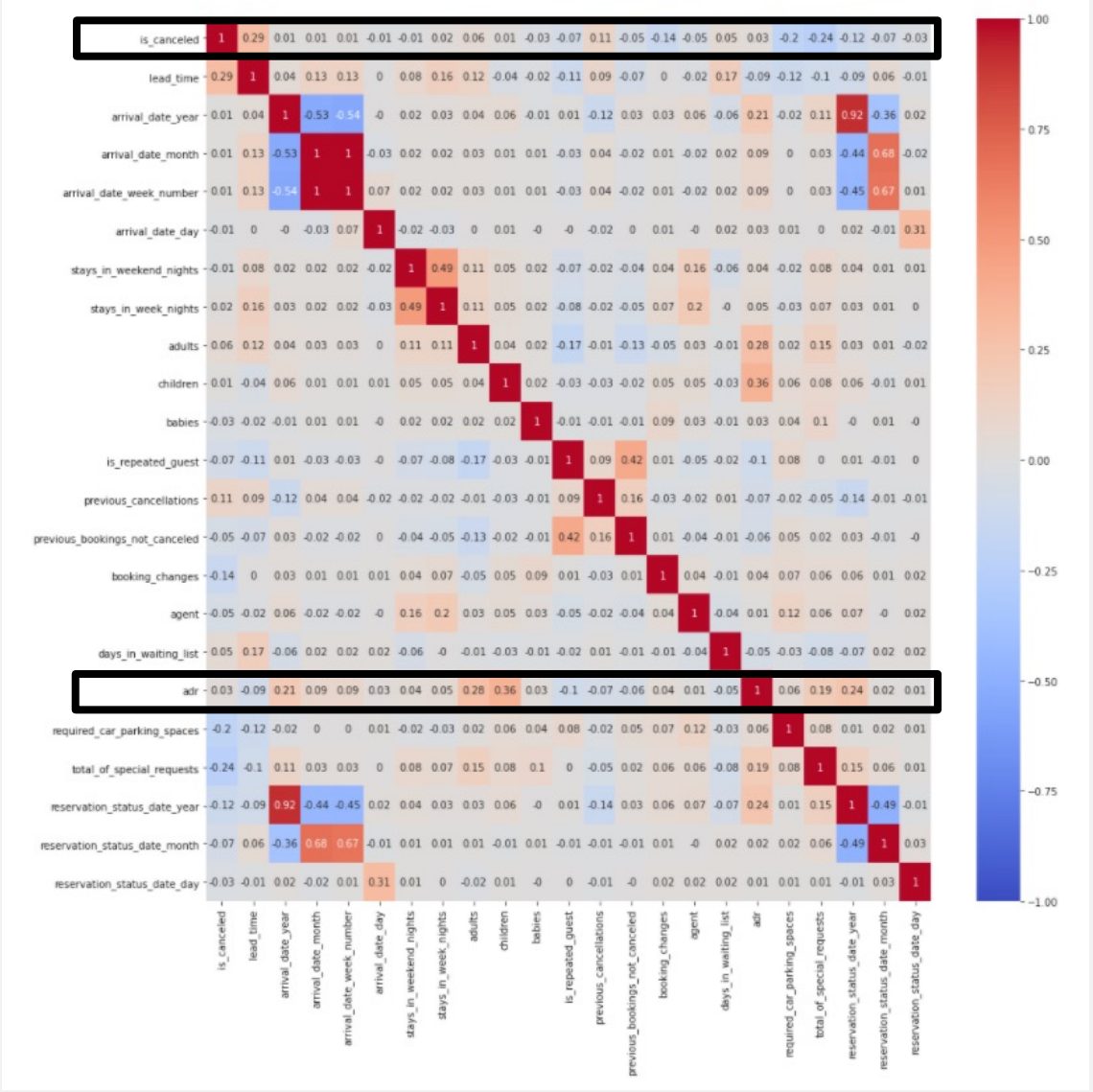
Final dataset left with

116,951 rows and 35 columns

(Original dataset: 119,390 rows and 32 columns)

IMPORTANT FACTORS AFFECTING CANCELLATION AND PRICE

- Cancellation
  - Lead_time
  - Previous_cancellation
  - Days_in_waiting_list
  - adr
- Average Daily Rate (ADR)
  - Children
  - Adults
  - Reservation_status_date_year
  - Arrival_date_year
  - Total_of_special_requests

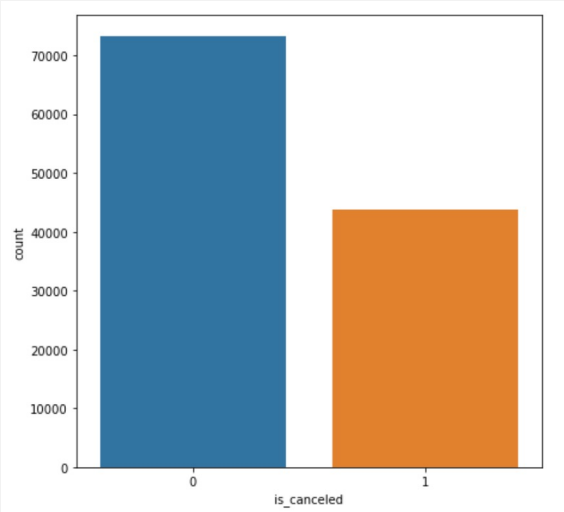


Heatmap of correlation coefficients between variables

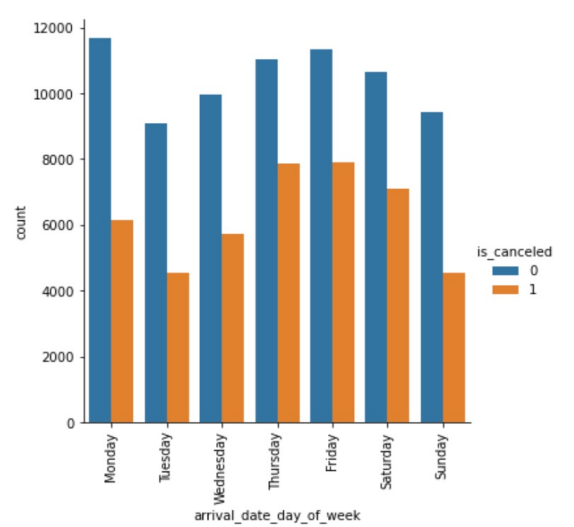
CANCELED VS NON-CANCELED

Cancelled reservations: 43,811

Not Cancelled Reservations: 73,140



Count plot of different cancellation types

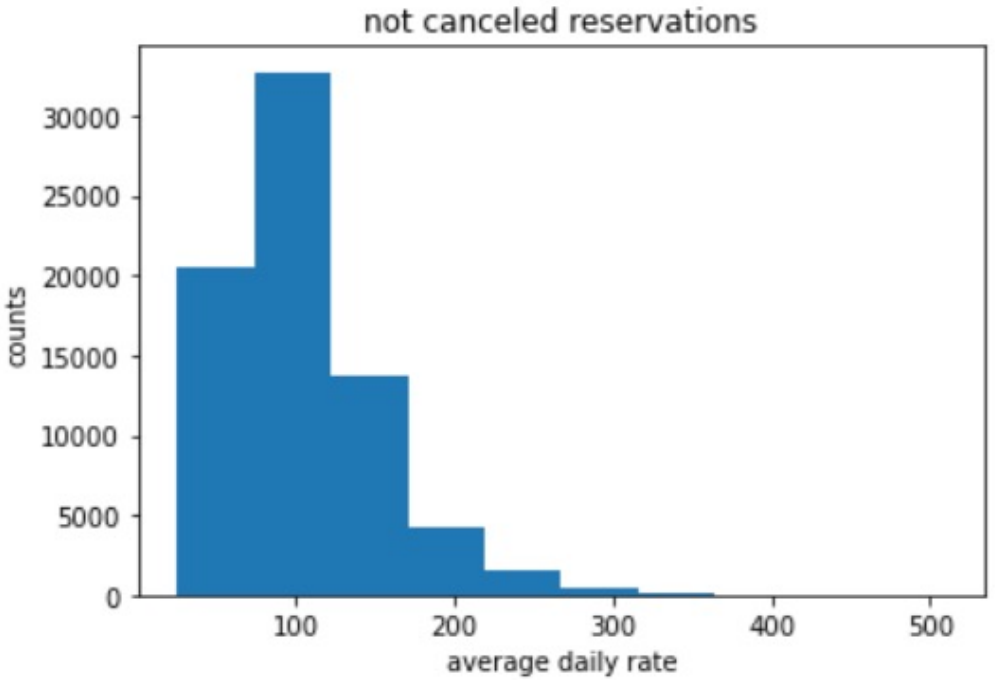


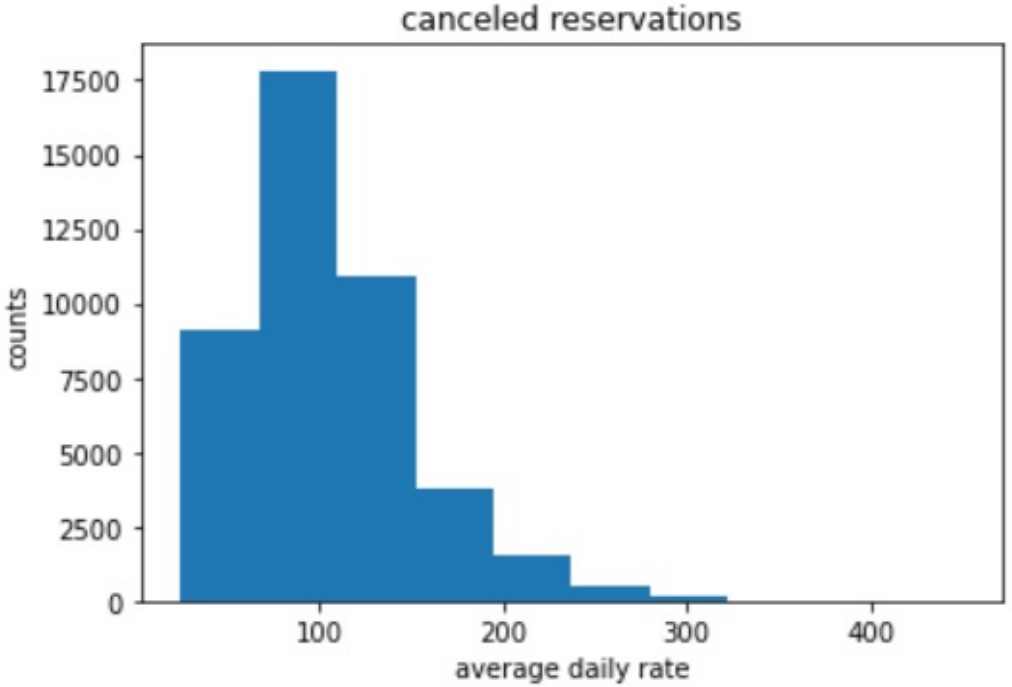
Count plot by day of week and cancellation type

PRICE RATES ARE SAME IN TWO CANCELLATION TYPE GROUPS?

- 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.
- Method
  - Bootstrapping
  - t-test
- Results : p-value < 0.05
  - Reject the null hypothesis and accept alternative hypothesis

Average daily rate (ADR) between two groups are **DIFFERENT!**





MODELING APPROACH

- Logistic Regression (Cancellation) / Linear Regression (Price)
- KNN
- SVM
- Decision Tree
- Gradient Boosting
- Random Forest

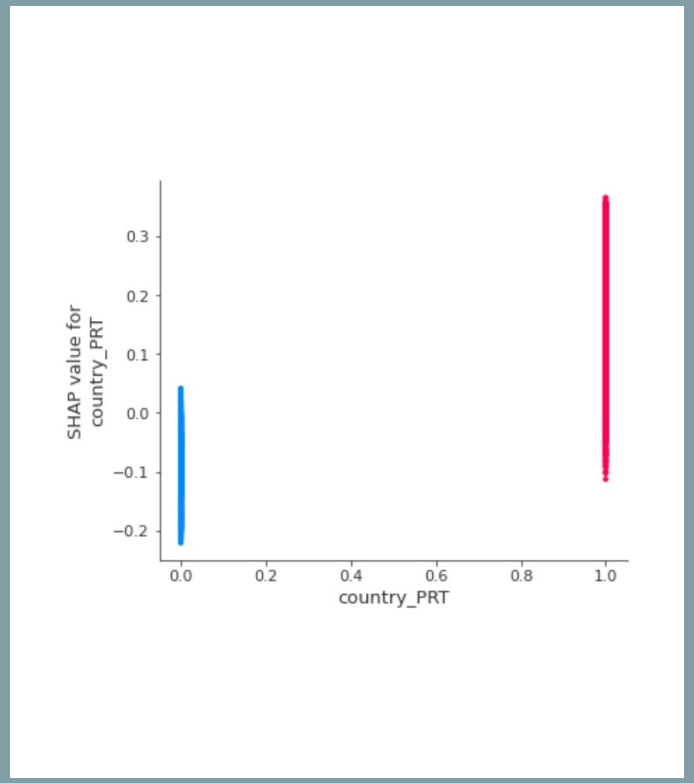
BEST HOTEL CANCELLATION MODEL

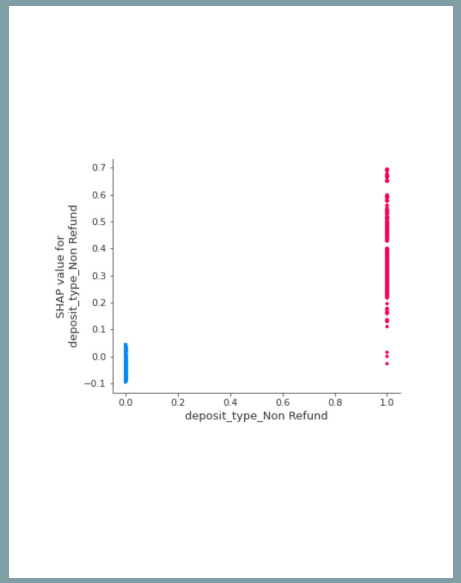
- Decision Tree
  - Criterion = 'entropy'
  - Max_depth = 140
  - Min_samples_leaf = 1
  - Min_samples_split = 3
- Feature Importance
  - Deposit_type_Non Refund
  - Arrival_date_week_number
  - Market_segment_Online TA
- Confusion Matrix

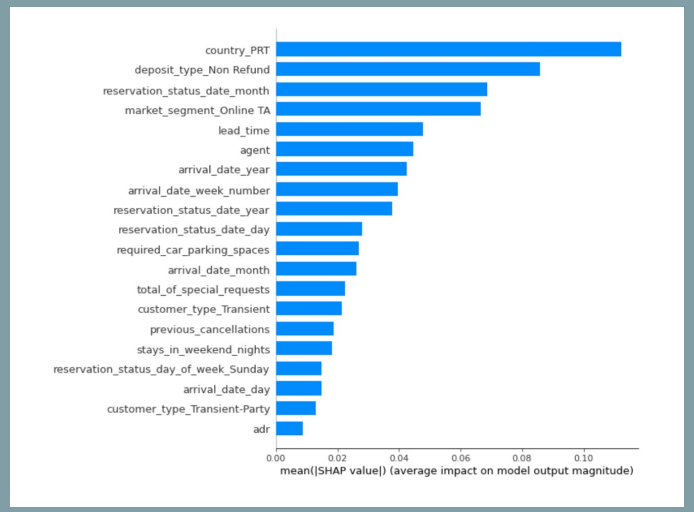
$$\begin{bmatrix} 21166 & 719 \\ 823 & 12378 \end{bmatrix}$$

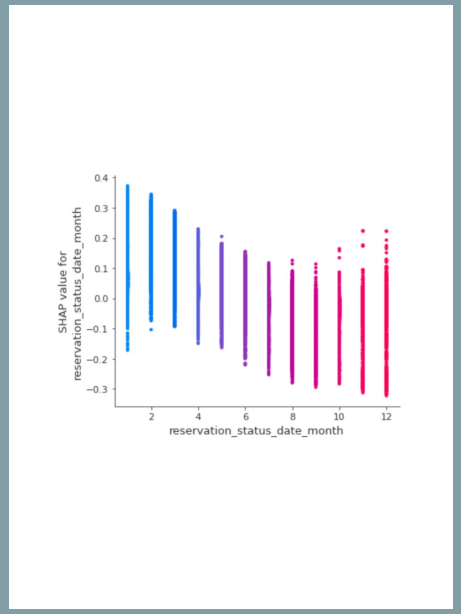
	precision	recall	f1-score	support
0	0.96	0.97	0.96	21885
1	0.95	0.94	0.94	13201
accuracy			0.96	35086
macro avg	0.95	0.95	0.95	35086
weighted avg	0.96	0.96	0.96	35086

Classification Report for Decision Tree with tuned parameters



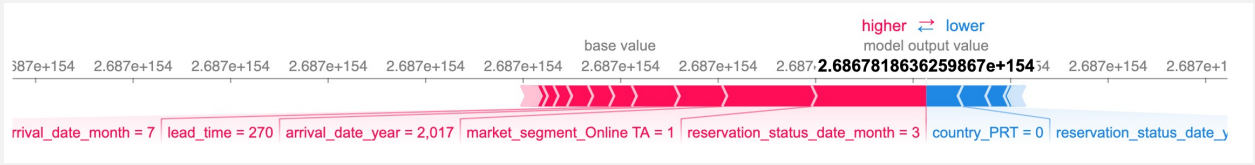




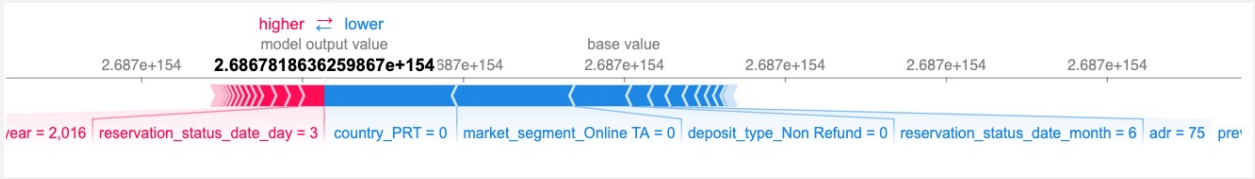


COMPARATIVE STUDIES ON CANCELLATION PREDICTIVE MODELS

- Top 3 Features affecting cancellation (1):
  - Country_PRT (Portugal)
  - Deposit_type_Non Refund
  - Reservation_status_date_month



Cancelled reservation



Not cancelled reservation

BEST HOTEL PRICE MODEL

- Random Forest
  - n_estimator = 1000
  - max_depth = 50
  - max_features = ‘auto’
- Errors
  - R squared: 0.926
  - Mean Squared Error: 12.675
  - Mean Absolute Error: 6.5474
- Top 5 Important features
  - Arrival_date_month
  - Reserved_room_type_A
  - Arrival_date_week_number
  - Children
  - Agent

CONCLUSIONS

- Provided hotels the analysis to reduce cancellation and set reasonable price points which helps to reduce unexpected profit loss
- Developed classification and regression models for hotel using various machine learning algorithms with above 90% average accuracy in model performance
- Performed comparative studies to analyze how feature importance impacts the results of the price and cancellation models for individual transactions for customers