STAT 425 Final Project

Ishan Nagpal

16 May 2020

Section 1: Introduction

In this project, I will be analyzing the Hotel Booking Demand dataset, which can be found on Kaggle and ScienceDirect. I will only be focusing on City Hotels for this project.

The aim of this project is to build a prediction model that predicts the Average Daily Rate of a hotel given other independent variables. Average Daily Rate is calculated by dividing the sum of all lodging transactions by the total number of nights stayed.

The data represents booking from July 1st, 2015 until August 31st, 2017. With this information, I should be able build a model that can predict

Let us begin analyzing this data to get actionable insights.

Section 2: Exploratory Data Analysis

Before we start analyzing the data to find interesting trends, it is important to give a brief overview of the variables in the dataset.

The variables are as follows:

- **is_cancelled** (Categorical): Value indicating if the booking was canceled (1) or not (0)
- **lead_time** (Numerical): Number of days that elapsed between the booking and arrival date
- **arrival_date_year** (Categorical): Year of arrival date
- **arrival_date_month** (Categorical): Month of arrival date with 12 categories: “January” to “December”
- **arrival_week_number** (Categorical): Week number of the arrival date
- **arrival_date_day_of_month** (Categorical): Day of the month of the arrival date
- **stay_in_weekend_nights** (Numerical): Number of weekend nights (Saturday or Sunday) booked to stay at the hotel
- **stay_in_week_nights** (Numerical): Number of week nights (Monday to Friday) booked to stay at the hotel
- **adults** (Categorical): Number of adults
- **children** (Categorical): Number of children
- **babies** (Categorical): Number of babies
- **meal** (Categorical): Type of meal booked. BB = Bed & Breakfast, HB - Half Board (Breakfast + 1 meal), FB = Full Board (All 3 meals)
- **market_segment** (Categorical): Market segment designation. “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **reserved_room_type** (Categorical): Code of room type reserved.
- **customer_type** (Categorical): Type of booking
- **adr** (Numerical): Average Daily Rate
- **total_of_special requests** (Categorical): Number of special requests by the guest

Time/Categorical Variables

We have 4 Time variables in this dataset and my initial thoughts on them are that only month and year variables should be useful in making predictions.

We will plot all the time variables and see if there is much variation and then come to a conclusion about what variables we should use.

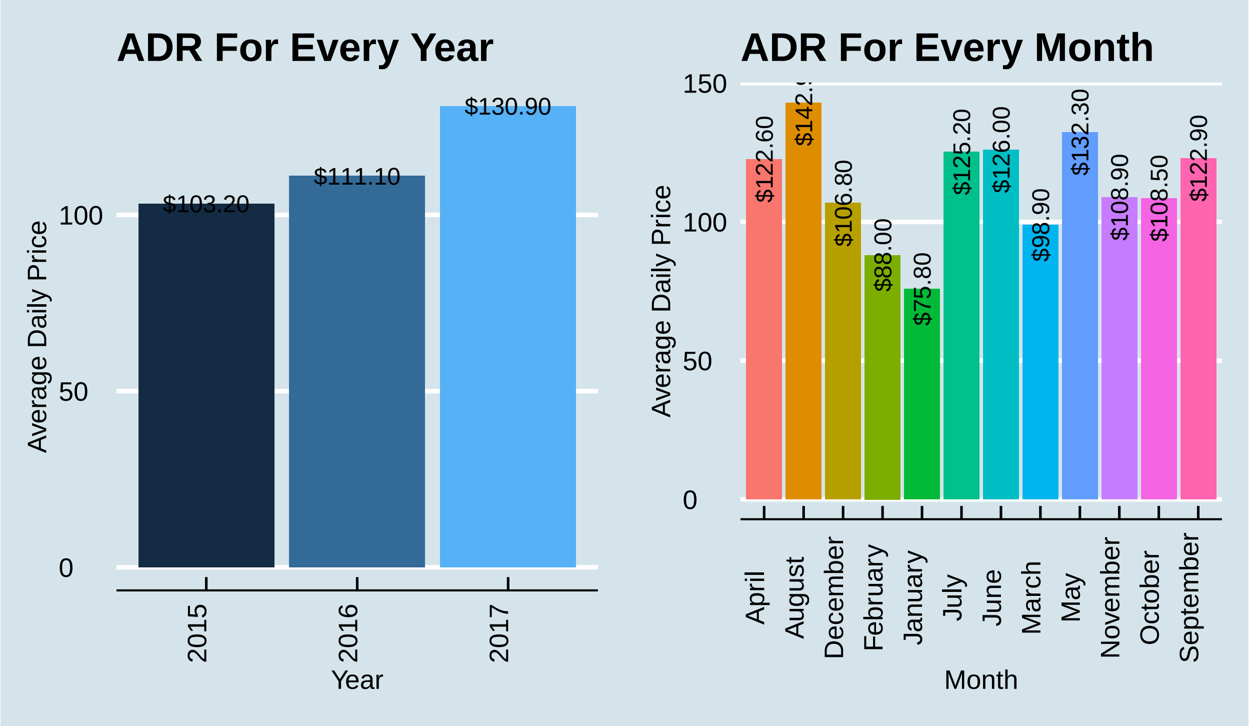

Figure 1: Mean Average Daily Price by Year and Month respectively

Looking at Figure 1, we can tell that our intuition about the **arrival_date_month** and **arrival_date_year** variables being important was true. We can observe that the mean Average Daily Price has increased every year and fluctuates seasonally throughout the year and varies from month to month.

Interestingly, we can observe that the month of December is the month with the highest mean Average Daily Prices. This is probably because of Christmas and New Year's Eve (holiday season) as many people decide to take a trip during these times, leading to high demand for City Hotels.

Now, we will look at the week and day variable.

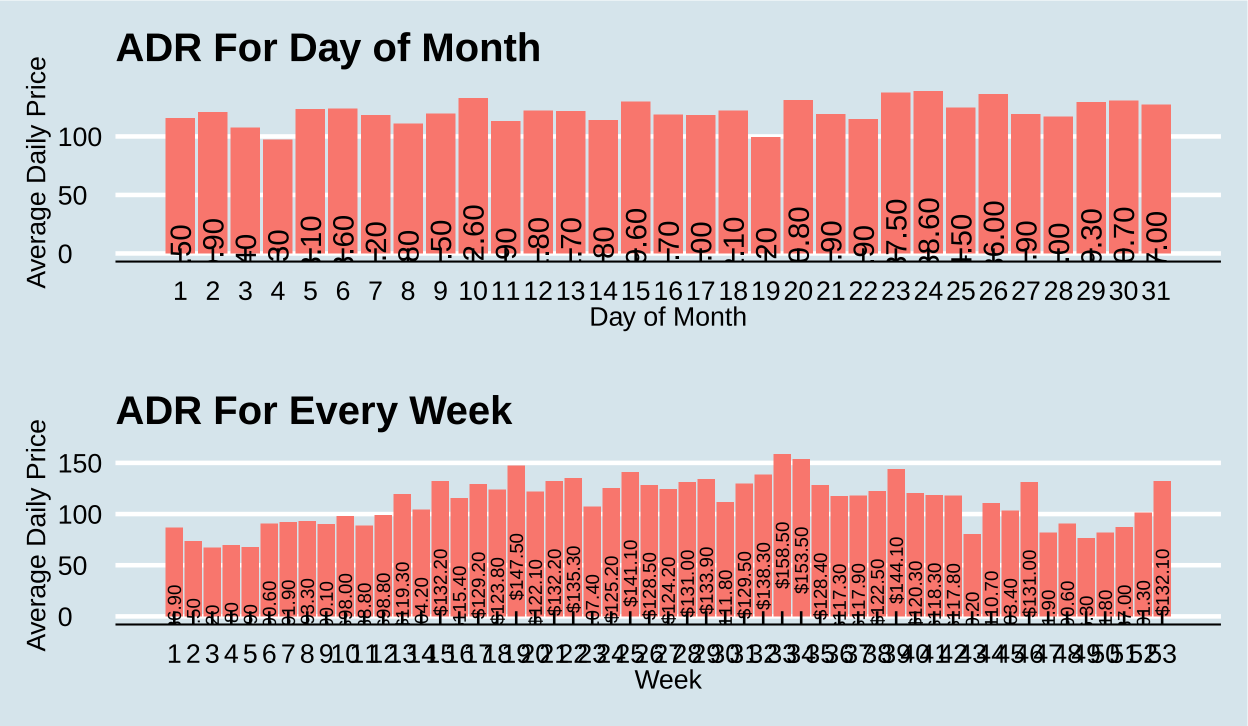

Figure 2: Mean Average Daily Price by Day of Month and Week respectively

Looking at Figure 2, we can tell that there is not much variation between days and it would not make much of a difference on the predictions if we drop the **arrival_date_day_of_month** variable. We can also tell that the **arrival_date_week_number** follows a similar distribution as **arrival_date_month** and that we can keep either of them.

I have decided to drop these two time variables and keep **arrival_date_month** and **arrival_date_year** variables.

For the Time Variables, we can also consider an interaction between the **arrival_date_month** and **arrival_date_year** variables. Let us plot these two variables to compare them.

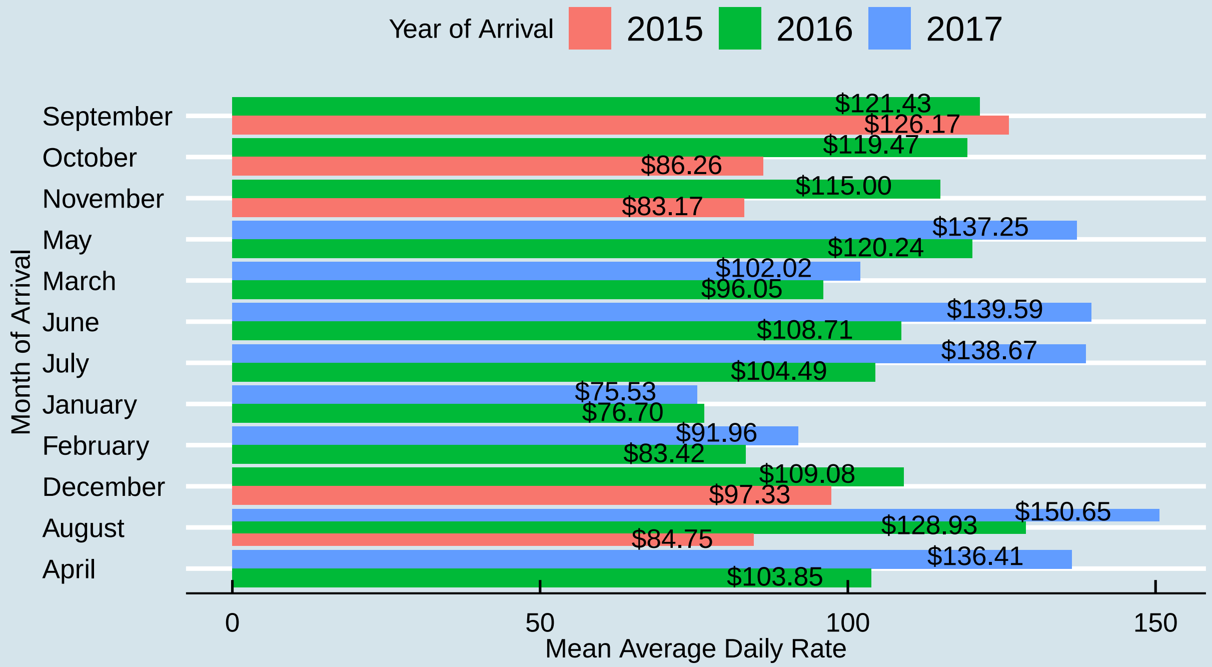

Figure 3: Interaction b/w Year and Month

In Figure 3, we have plotted the interaction between the time variables and there seems to be some sort of an interaction.

We can clearly see that the price increases every year and we can speculate that it is due to inflation or increasing costs.

Now, we will move on to the other variables.

Let us visualize Average Daily Rate by Meal Type, Market Segment, Customer Type, and Room Type.

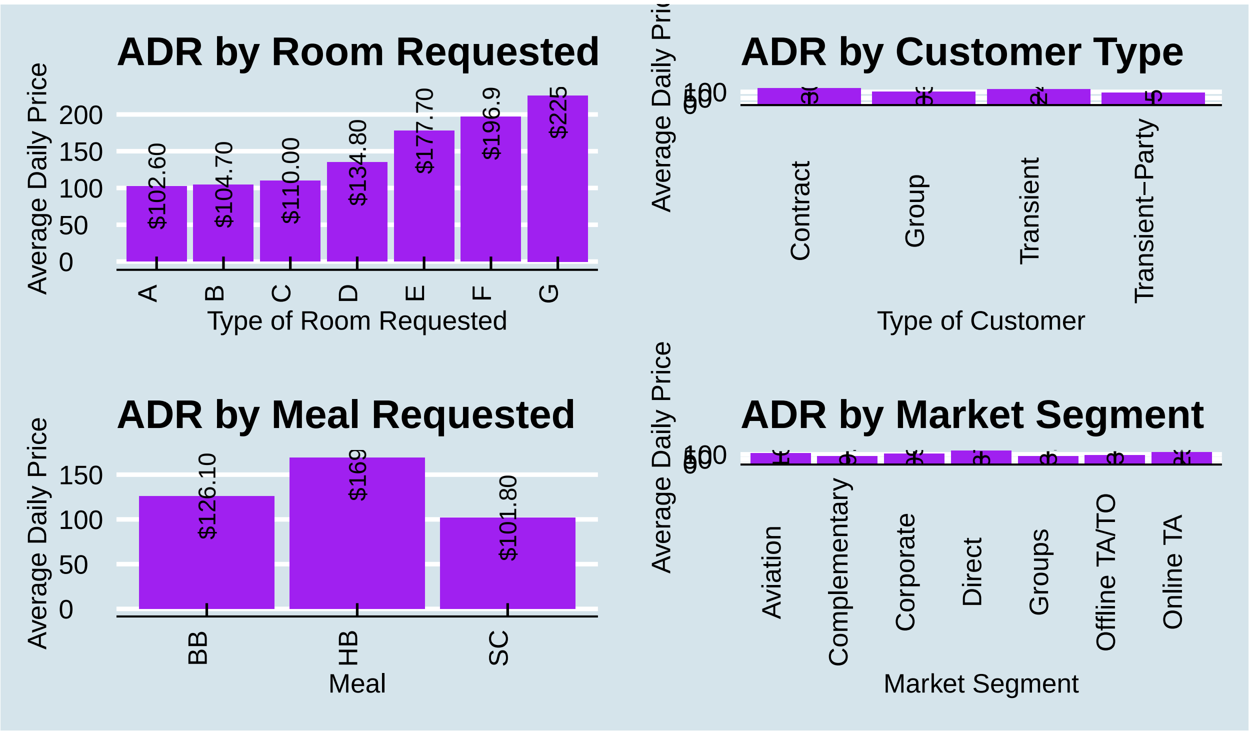

Figure 4: Mean Average Daily Price by Customer Type, Room and Meal Requested

Figure 4 tells us a plethora of things.

Firstly, it is well defined that the Rooms have a hierarchy with the A room being the cheapest and the G room being the most expensive.

Secondly, we can observe that the Contract type of customer pays the most on average compared to other customers.

It is also seen that the HB type of meal is the most expensive, which is expected, as the other two have less meals than this meal plan.

Finally, we also observe that the Direct market segment has the highest ADR while the lowest of the 6 is, obciously, Complimentary. It is also intriguing to see that Online Travel Agencies are the second highest in ADR. It is surprising as I have personally found the best deals through them.

Numerical Variables

Now, we will take a look at numerical variables to find any nonlinear trends. For this, we will use the GGally library and plot a scatter matrix on all the numerical variables of the dataset.

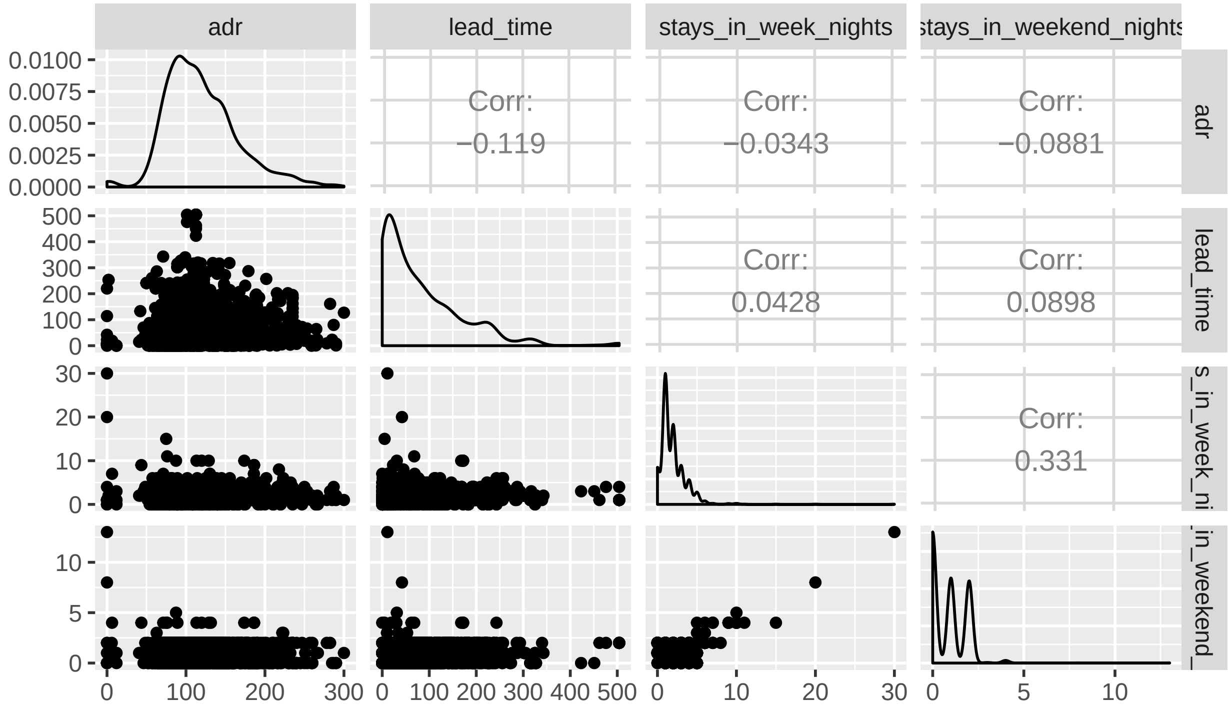

Figure 5: Scatter Matrix

Looking at Figure 10, we can tell it shows us the correlations and scatter plots between all the numerical variables.

We can tell that **lead_time** has a weak negative correlation with **adr**, but **stay_in_week_nights** and **stay_in_weekend_nights** have an even weaker negative correlation with **adr**, if at all.

Therefore, we can conclude that the numerical variables do not have any nonlinear trends and we can drop **stay_in_week_nights** and **stay_in_weekend_nights** variables.

Section 3

Now that we are done with Exploratory Data Analysis, we can now choose the variables we will include in this analysis.

The variables we will include are: - **is_canceled** - **arrival_date_month** - **arrival_date_year** - **lead_time** - **adults** - **children** - **meal** - **market_segment** - **reserved_room_type** - **customer_type** - **total_of_special_requests**

Now, we will build prediction models.

Section 3.1: Linear Regression

Linear Regression is used for finding linear relationship between a target variable and one or more predictors known as independent variables. The population linear regression line is defined as:

$$Y = \beta_0 + ... + \beta_i X_i$$

Linear Regression has a few kew assumptions such as: - Constant error variance - Residuals are normally distributed - Independent errors - There is a linear relationship between Y and the predictors - Predictors are not highly correlated (>0.8)

Table 1: Table 1: Variables Removed

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1573	1073543	10603.00
- customer_type	3	3284.075	1576	1076827	10601.94

Table 2: Table 2: Model Info

BIC	No.of.Parameters
10828.28	42

We will select the model with all the variables including the time interaction term. We will also eliminate any unnecesary variables using the step function.

Let us now take a look at all the variables that were removed from the full model using the step function.

As we can see, the **customer_type** variable was removed from this model.

Let us take a look at the BIC and the Number of Parameters of this model.

Section 3.2: Predict with Linear Regression Model

Now, we will attempt to calculate predictions based on the linear regression model we made in the last part.

The RMSE that we get from this model is **25.79787**, which we can say is quite good for this model.

Now, we will try to predict the Average Daily Rate by setting up an example.

Let's assume I am booking a last minute vacation that will take place **Decemeber 2016** (because it is the most expensive time period for hotels) with a lead time of **15** days. I have **2 adults** and **1 kid** with me, and I want to reserve the most expensive room, that is, **G**. I **do not have any special requests** and opt for **Bed & Breakfast**.

After running the prediction model, it turned out to be pretty expensive, as expected, as we booked it very late during the most expensive time for hotels. The total predicted Average Daily Rate is **$200.8053** in the hypothetical scenario.

Section 3.3: Random Forest

The **Random Forest Algorithm** aggregates predictions made by multiple decision trees of varying depth. Every decision tree in the forest is trained on a subset of the dataset called the bootstrapped dataset. There are no assumptions made in this method.

We will have a huge amount of decorrelated trees which we can use to make predictions to come up with a mean predicted value.

We will now generate a Random Forest Model with 5000 trees using all the variables available to us.

After generating the Random Forest Model, we get an **RMSE of 23.67304** which is slightly better than the RMSE we got from the Linear Regression Model, but not by a lot at all.

Let us try and predict with this model using the same example we used above, with **Transient** as the **customer_type** and look at the results.

The prediction we get from the Random Forest Model is **$128.1448** and we can tell that this is much smaller than the prediction the Linear Regression Model made. As the RMSE is smaller than LR model, we will go with the prediction this model made.

Table 3: Table 3: Final Results

	RMSE
Linear Regression	25.79787
Random Forest	23.68267

Section 4: Results

In conclusion, we find that the Random Forest model was more effective in getting accurate predictions as the RMSE was lower than the one in the Linear Regression model.

It can be seen in the table below:

In my opinion, the results obtained are pretty accurate and are reliable.

References:

- https://www.sciencedirect.com/science/article/pii/S2352340918315191
- https://www.kaggle.com/micahshull/r-hotel-bookings
- https://www.kaggle.com/eduarmma19/hotel-booking-demand-eda-and-logistic-regression