HOTEL DEMAND

CASE STUDY

Business Problem

The hotel demand stems from business travellers and leisure tourists who travel frequently. With information on the hotel bookings, our goal is to predict booking cancellation. With the increase in flexible cancellation policy through online services, estimating cancellation rates becomes important so that the hotel is not left with unrented rooms for multiple days at a time. Predicting the likelihood that a customer will cancel their reservation could be used to successfully optimize booking service and anticipate when cancellations will occur.

The business questions we are trying to answer are:

- How does cancellation vary across different booking sources?
- Is cancellation trend seasonal? If so, when does it occur?
- Where do most cancellations come from?
- Are bookings with longer lead times more likely to result in a cancellation?
- What other factors influence high cancellation rates?

Data Overview

The data is the hotel demand dataset on Kaggle that contains information on two types of hotels, a resort and city hotel. The dataset has 1,19,390 rows and 32 features. Each observation is a hotel booking done by a customer. The dataset includes bookings by customers who would arrive between 1st July 2015 and 31st August 2017, including information on cancellation status. It also has features indicating the stay information like length of stay, number of adults, the facilities provided by the hotel like parking and details on the country, market segment and arrival date of the booking.

Tools

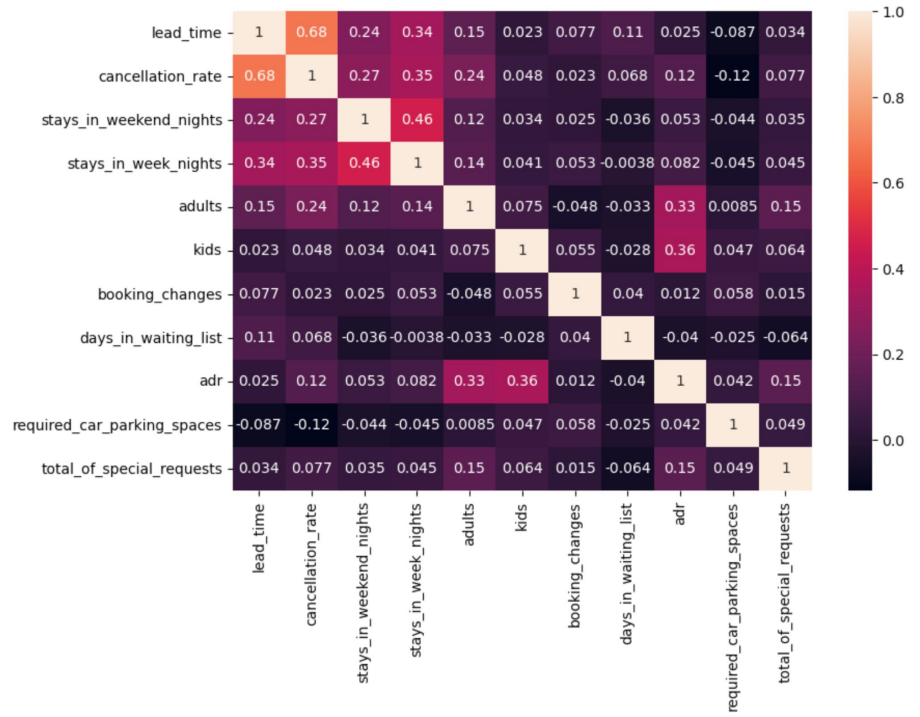
- Python, Jupyter Notebook, Tableau
- Python libraries: Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn, SciPy

Analytical Methods

Data wrangling and cleaning
Deriving variables
Linear regression
Time series analysis
Data consistency checks
Exploratory analysis
Cluster analysis
Visualization

1) Exploratory Data Analysis

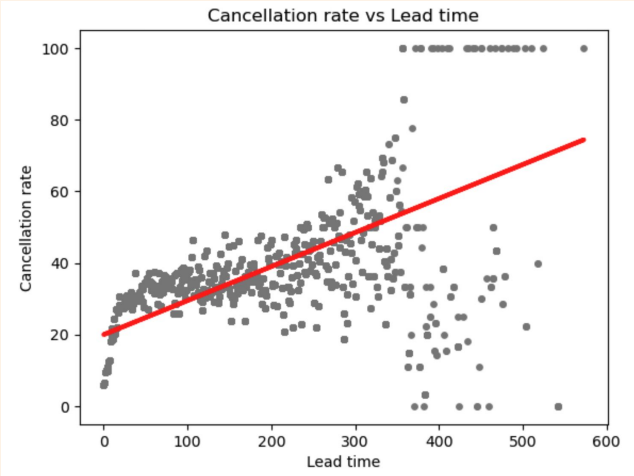
- Explored the data in Python to identify relationship among variables.
- Explored how each variable contributed to cancellation and investigate their relationships.
- This helped me to develop hypotheses on variables that could be an influencing factor for cancellation.



- Created a correlation heatmap with seaborn to determine the critical numeric features influencing cancellations.
- The strongest correlation is between lead time (in days) and cancellation rate. Meaning the higher the lead time, the higher the chance of its booking being cancelled.
- I therefore decided to explore this relationship further by employing regression and K-means clustering.

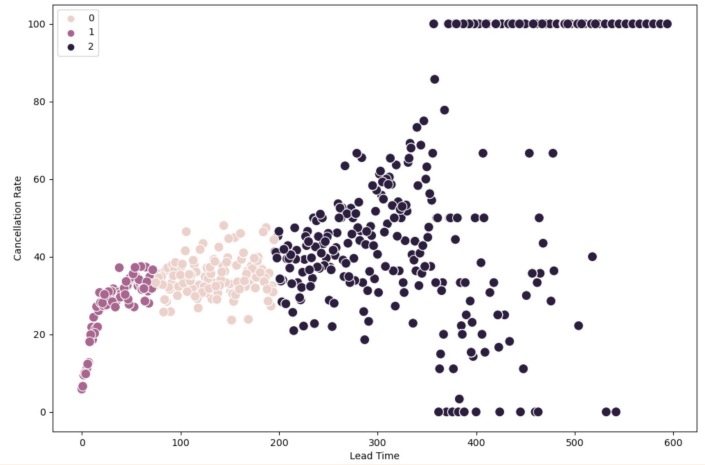
2) Machine Learning

Supervised: Linear Regression



- A moderately strong positive correlation was found between lead time and cancellation rate with an R2 score of 0.48

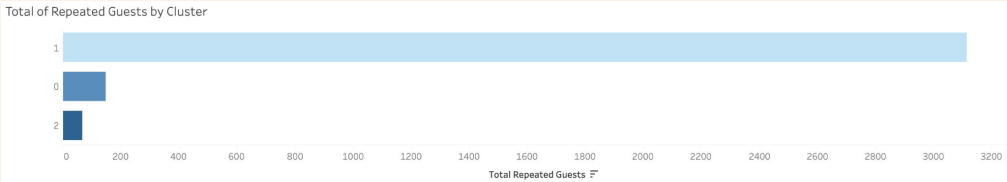
Unsupervised: K-Means Clustering

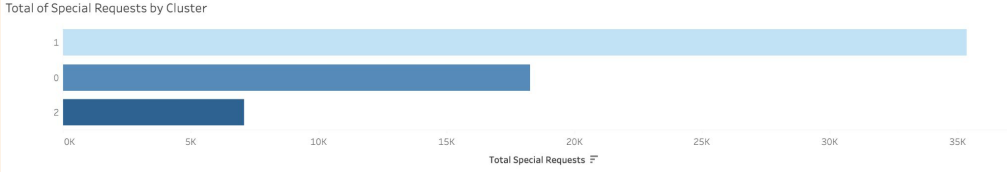


- Elbow technique indicated 3 clusters
- Conducted K-Means Clustering Algorithm using Scikit-learn library.
- Clustering analysis yielded 3 distinct groups of data points, which can be seen represented in different colors on this scatterplot.
- The clusters clearly show the relationship between both variables, supporting the hypothesis.

3) Tableau Visualizations

Conducted further exploratory analysis and created visualisations in Tableau in order to answer the business questions and make recommendations.





- Cluster 2, having the highest cancellation rate, has the lowest number of repeated guests, booking changes and special requests.
  Average cancellation rate: 40%

- Conversely, cluster 1 with the lowest cancellation rate, has the highest number of repeated guests, booking changes and special requests.
  Average cancellation rate: 22%

- Cluster 0, with mid-level cancellation rate, has values between clusters 2 and 1.
  Average cancellation rate: 35%

4) Insights & Recommendations

- OTAs have higher cancellation rates than offline and direct bookings. Focus on driving direct or offline bookings rather than heavily relying on OTAs. The personalized connection felt directly with a hotel versus an impersonal OTA could explain why direct or offline bookings have a lower rate of cancellation. With direct bookings, the hotel is also better able to push rebooking or vouchers as an alternative to cancelling a booking.

- Bookings with no deposit are extremely high in number and bookings with non-refundable deposit type are high in proportion in reference to cancellation. Implement a strict cancellation policy with conditions such as deposits. This way the hotel can ensure guaranteed reservations or atleast be assured they're covered financially even if a guest does cancel the booking.

- Cancellation trend is highly seasonal, peaking during the summer months of July and August.

- Most number of cancellations come from Portugal, whereas China leads the way when it comes to the rate of cancellation.

- Bookings with more lead days have a higher likelihood of being cancelled. This may be tackled if the hotel customer service keeps in touch with the customer in advance of the scheduled stay, especially in summer months, when cancellation peaks. Use pre-arrival emails and follow up with phone calls. It’s also a good opportunity to upsell extras as well as foster guest loyalty and connection, especially by personalizing communications.

- Higher number of repeated guests, booking changes or special requests reduces the likelihood of cancellation.
    - A loyalty strategy could be created since both hotels have a very low number of guests with repeated bookings; especially for domestic guests from Portugal, as they have the highest number of bookings. In addition, their proximity to both hotels would increase the likelihood of repeat stays.

    - Offer upgrades to customers if the hotel is not fully occupied. This reduces the likelihood of cancellation and ensures a part of the revenue is realized.

    - Be flexible on special offerings to keep customers from cancelling. Providing complimentary services will also maximize the guest experience.

5) Limitations

- With the date range from Jul 2015 to Aug 2017, the dataset lacks complete year-on-year data from Jan-Dec time periods, and as such comparisons for trends and seasonality are incomplete.
- The dataset contains booking records for two hotels based in Portugal. It could be that cancellation trends for Portugal are different from that of hotels in other countries. Therefore, the results cannot be generalised to any other group of hotels.

6) Further Research and Analysis

- Investigate why most of the non-refundable bookings end up being cancelled.
- Conduct analysis on hotels based in other countries.
- Explore other machine learning algorithms for predictive analytics.

Reflections

I thoroughly enjoyed working on this project as this was my most challenging one yet. I familiarized myself with the core Python visualization libraries, and created engaging, accessible visuals to form an interactive data dashboard. My most favorite and exciting part of this project - building on my advanced analytics skills by taking a dive into the basics of machine learning and modelling in the form of regression and cluster analysis. I was fascinated by how the algorithms could identify patterns for me, saving me a lot of effort and resulting in more accurate suggestions. It gave me a good basis for generating new hypotheses and finding new directions to further explore the data.