Star Hotel project

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- Data Processing – Initial Steps
- EDA
    - Univariate Analysis
    - Bivariate Analysis
- Data Processing – Other Steps
    - Outlier Detection and Treatment
- Model Performance Summary
- Model Performance Evaluation
- Decision Tree – Pre_Pruning and Post_Pruning
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

Context

- A significant number of hotel bookings are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.
- The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.
- The cancellation of bookings impact a hotel on various fronts:
- Loss of resources (revenue) when the hotel cannot resell the room.
- Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
- Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
- Human resources to make arrangements for the guests.

Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

Data Overview

Data Dictionary

- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
  - Not Selected – No meal plan selected
  - Meal Plan 1 – Breakfast
  - Meal Plan 2 – Half board (breakfast and one other meal)
  - Meal Plan 3 – Full board (breakfast, lunch, and dinner)

- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by Star Hotels.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

Data Processing – Initial Steps.

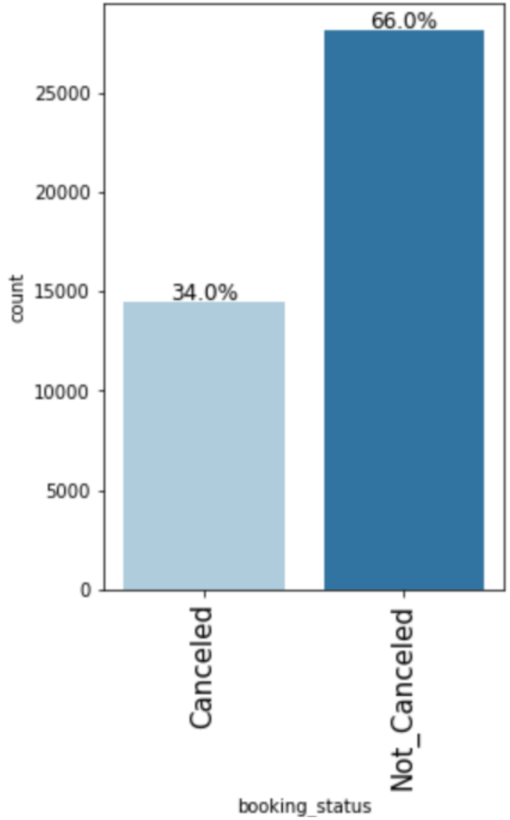
- Dataset has 56926 rows and 18 columns
- Most of the data-types are **int64** and one of the columns has data-type **float64**.
- 4 columns - type_of_meal_plan, room_type_reserved, market_segment_type and booking_status are having data-types as an **object**, this means we need to convert these into suitable data-type before we feed our data into the model.
- There were 14350 duplicate values in the dataset and they are removed
- There are no missing values in the data.

EDA

- **Univariate Analysis of booking_status**

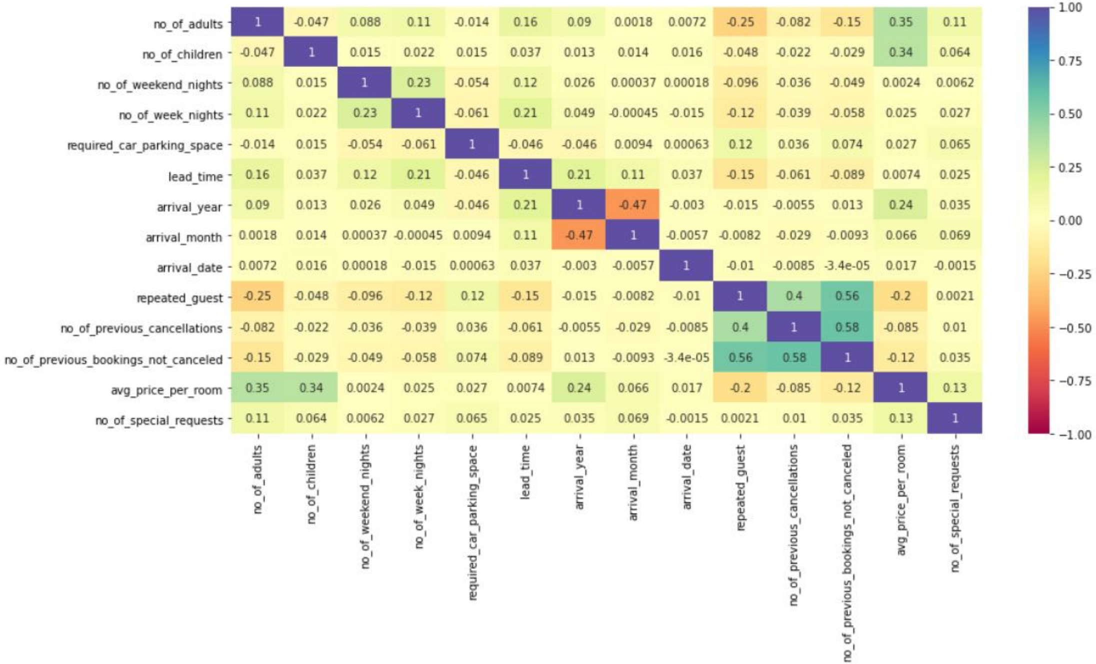
Observations

- 66% of the bookings are not_cancelled.
- 34% of the bookings are cancelled.



Bivariate Analysis

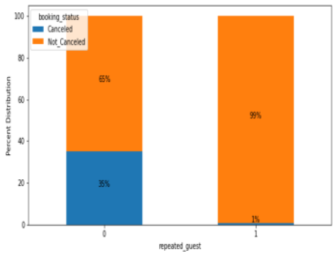
- **Correlation**

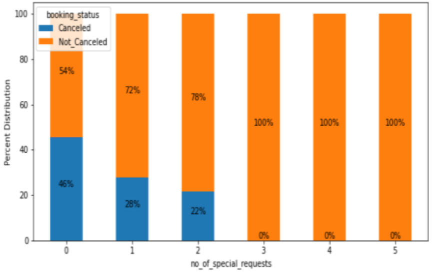


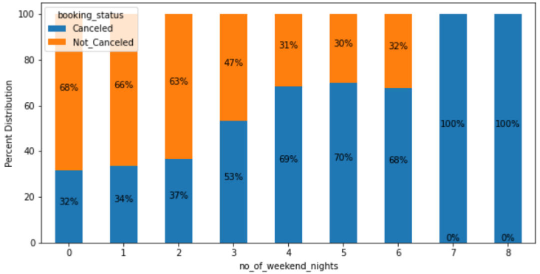
Observations

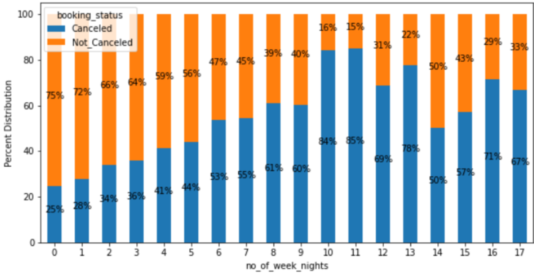
- There are no significant positive or negative correlation between any variables.
- no_of_previous_bookings_not_canceled has a strong positive correlation with repeated_guest and no_of_previous_cancellations.

Relationship of booking_status with repeated_guest, no_of_special_requests,no_of_weekend_nights,no_of_week_nights





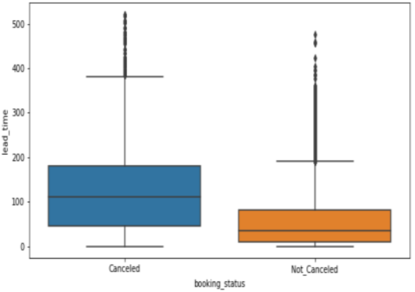


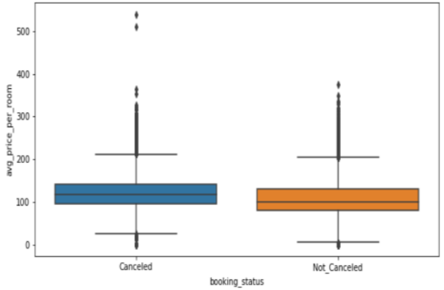


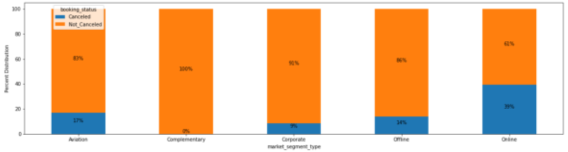
Observations

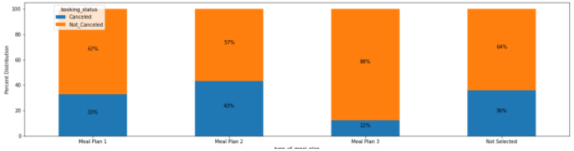
- Only 0.76% of repeated guests have cancelled their bookings.
- If a customer makes special requests, less chances of cancelling it.
- As the no_of_weekend_nights increases, the "cancelled" booking_status also have increased.
- As the no_of_week_nights increases, the "cancelled" booking_status also have increased steadily till 11 nights to 85% cancellations.

Relationship of booking_status with lead_time,avg_price_per_room,type_of_meal_plan,market_segment_type









Observations

- Cancelled bookings have 25% to 75% lead time between 50 and 180 days. There are many outliers after 400.
- 25% to 75% of the avg_price_per_room for cancelled bookings are in the range of 100 to 150 euros.
- Most customers whcih is 43% who have taken Meal Plan 2 have cancelled their bookings.
- Customers who belong to market segment online have the most high percentage to cancel which is 39%.

Data Processing – Other Steps

- **Outlier Detection**

We have only two numerical variable that are continous in nature : lead_time and avg_price_per_room

We Detect their Outliers and Remove them

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a customer will not cancel the booking but in reality the customer cancels the booking. - Loss of revenue
- Predicting a customer will cancel the booking but in reality the customer did not cancel the booking. - spend more money for marketing and reduce prices.

Which case is more important?

- Predicting a customer will not cancel the booking but in reality the customer cancels the booking. - Loss of revenue

How to reduce this loss i.e need to reduce False Negatives?

- recall should be maximized, the greater the recall higher the chances of minimizing the false negatives.

Model Performance Summary

Logistic Regression Model

- All the models are giving a generalized performance on training and test set.
- The highest recall is ~78% on the training set.
- Using the model with default threshold the model will give a low recall but good precision scores - This model will help the Hotel save resources on marketing.
- Using the model with 0.33 threshold the model will give a high recall but low precision scores - This model will help the Hotel identify customers will cancel the bookings effectively but the cost of resources will be high.
- Using the model with 0.42 threshold the model will give a balance recall and precision score - This model will help the Hotel to maintain a balance in identifying potential cancellations and the cost of resources.

Decision Tree Model

* GridSearch was used for Pre-Pruning

* Cost Complexity Pruningwas used for Post_Pruning

- Decision tree model with pre-pruning has given the best recall score 98.87% on training data.
- The pre-pruned and the post-pruned models have reduced overfitting of the model.
- The Pre-Puned model gives a better performance on training and test sets . Hence we can use this model for decion making.

Actionable Insights and Recommendations

- Since repeated guests have very less chance of cancelling the booking, send Special Offers to the email of Special Guests so that they book the same Star hotel again.
- As the guests that make Special Requests show less chances of cancellation, provide the options of Special Requests for all the guests at the time of booking. Provide some complementary offers alongside that they will prefer.
- Provide Special Offers and Rates to Customers who book for long duration that involves more than 2 weekend nights or more than 5 weekdays. Offers such as one weekday/weekend night free with 2 weekend nights/5 weekday nights.
- Provide free Car parking space as Special Offers for few guests and see if that improves the Non Cancellation.
- Regularly send remainder mails to customers with long lead time/ arrival year. Ask if they would like to keep or cancel their booking. So that Hotel can get a status sooner.
- Provide Special offers for Summer months.
- Make some complementary offers for rooms with High Average Price.
- Improve the dishes provided in Meal Plan 2 which has higher cancellations.
- Redesign Room Type 6 which has higher cancellations.

greatlearning
Power Ahead

Happy Learning !