# Hotel reservation

In this project, we aim to perform EDA visualizations and build a predictive model to determine whether a hotel booking would be canceled, which is crucial for hotels as cancellations affect revenue and operational planning. The dataset contains a high number of features related to booking, such as lead time, deposit type, and special requests, which adds to the complexity of the model. The challenge lies in the data preprocessing steps, which include feature selection and engineering, handling missing values, and noise in the data. Additionally, we are going to train different models, evaluate their performance using the right metrics, and interpret the model by analyzing the most important features in the context of hotel booking cancellations.

Number of Entries: The dataset consists of 119,390 entries.

Columns: The dataset contains 34 columns, which represent various attributes related to hotel bookings.

# Dataset Description:

| **Index** | **Variable** | **Description** |
| --- | --- | --- |
| 1 | **hotel** | Type of hotel (Resort Hotel, City Hotel) |
| 2 | **is\_canceled** | Reservation cancellation status (0 = not canceled, 1 = canceled) |
| 3 | **lead\_time** | Number of days between booking and arrival |
| 4 | **arrival\_date\_year** | Year of arrival |
| 5 | **arrival\_date\_month** | Month of arrival |
| 6 | **arrival\_date\_week\_number** | Week number of the year for arrival |
| 7 | **arrival\_date\_day\_of\_month** | Day of the month of arrival |
| 8 | **stays\_in\_weekend\_nights** | Number of weekend nights (Saturday and Sunday) the guest stayed or booked |
| 9 | **stays\_in\_week\_nights** | Number of week nights the guest stayed or booked |
| 10 | **adults** | Number of adults |
| 11 | **children** | Number of children |
| 12 | **babies** | Number of babies |
| 13 | **meal** | Type of meal booked (BB, FB, HB, SC, Undefined) |
| 14 | **country** | Country of origin of the guest |
| 15 | **market\_segment** | Market segment designation |
| 16 | **distribution\_channel** | Booking distribution channel |
| 17 | **is\_repeated\_guest** | If the guest is a repeat customer (0 = not repeated, 1 = repeated) |
| 18 | **previous\_cancellations** | Number of previous bookings that were canceled by the customer |
| 19 | **previous\_bookings\_not\_canceled** | Number of previous bookings that were not canceled by the customer |
| 20 | **reserved\_room\_type** | Type of reserved room |
| 21 | **assigned\_room\_type** | Type of assigned room |
| 22 | **booking\_changes** | Number of changes made to the booking |
| 23 | **deposit\_type** | Type of deposit made (No Deposit, Refundable, Non Refund) |
| 24 | **agent** | ID of the travel agent responsible for the booking |
| 25 | **company** | ID of the company responsible for the booking |
| 26 | **days\_in\_waiting\_list** | Number of days the booking was in the waiting list |
| 27 | **customer\_type** | Type of customer (Transient, Contract, Transient-Party, Group) |
| 28 | **adr** | Average Daily Rate |
| 29 | **required\_car\_parking\_spaces** | Number of car parking spaces required |
| 30 | **total\_of\_special\_requests** | Number of special requests made |
| 31 | **reservation\_status** | Last reservation status (Check-Out, Canceled, No-Show) |
| 32 | **reservation\_status\_date** | Date of the last reservation status |
| 33 | **name** | Guest's name |
| 34 | **email** | Guest's email address |
| 35 | **phone-number** | Guest's phone number |
| 36 | **credit\_card** | Last four digits of the guest's credit card |

Problem statement:

~ Key features that affect hotel reservation

~How to make hotel cancellations index better

~How will hotels be assisted in making pricing and promotional decisions

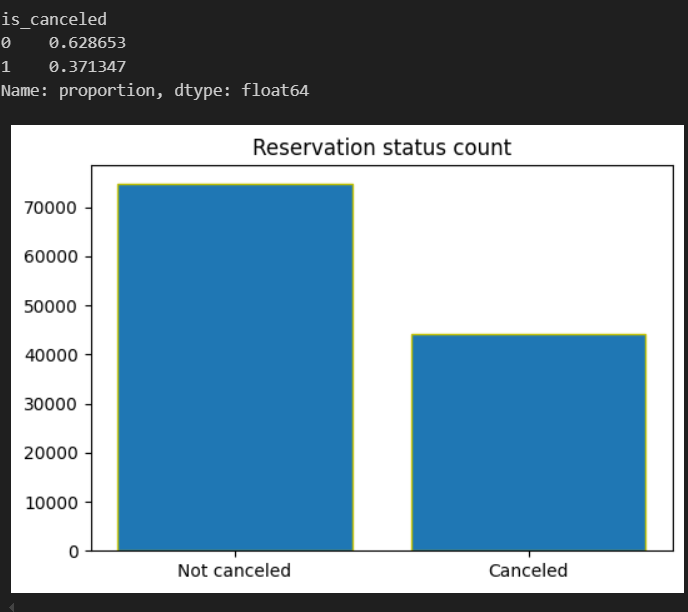


Figure 1: Percentage and count of cancelled and not cancelled

Here, we can observe from bar graph that the count of cancel is lesser but comparatively to the business scenario, this metrics is lot higher. As 37% of bookings are getting canceled by the customers. So, there might be various issues customers are considering to be cancelled, let’s identify the measures moving onwards. Which of the hostels are getting higher cancellation let’s check by visualizing each in bar graphs.



Figure 2: Cancellation count for each hotel

Here by, we can see the cancellation rate of city is a lot bigger than that of the city hotel. Different factors of cost, maintenance, water problem and unhygienic may be the issues. Moving on, we are browsing on the daily rates of hotels, to determine whether cost is one of the factors or not.

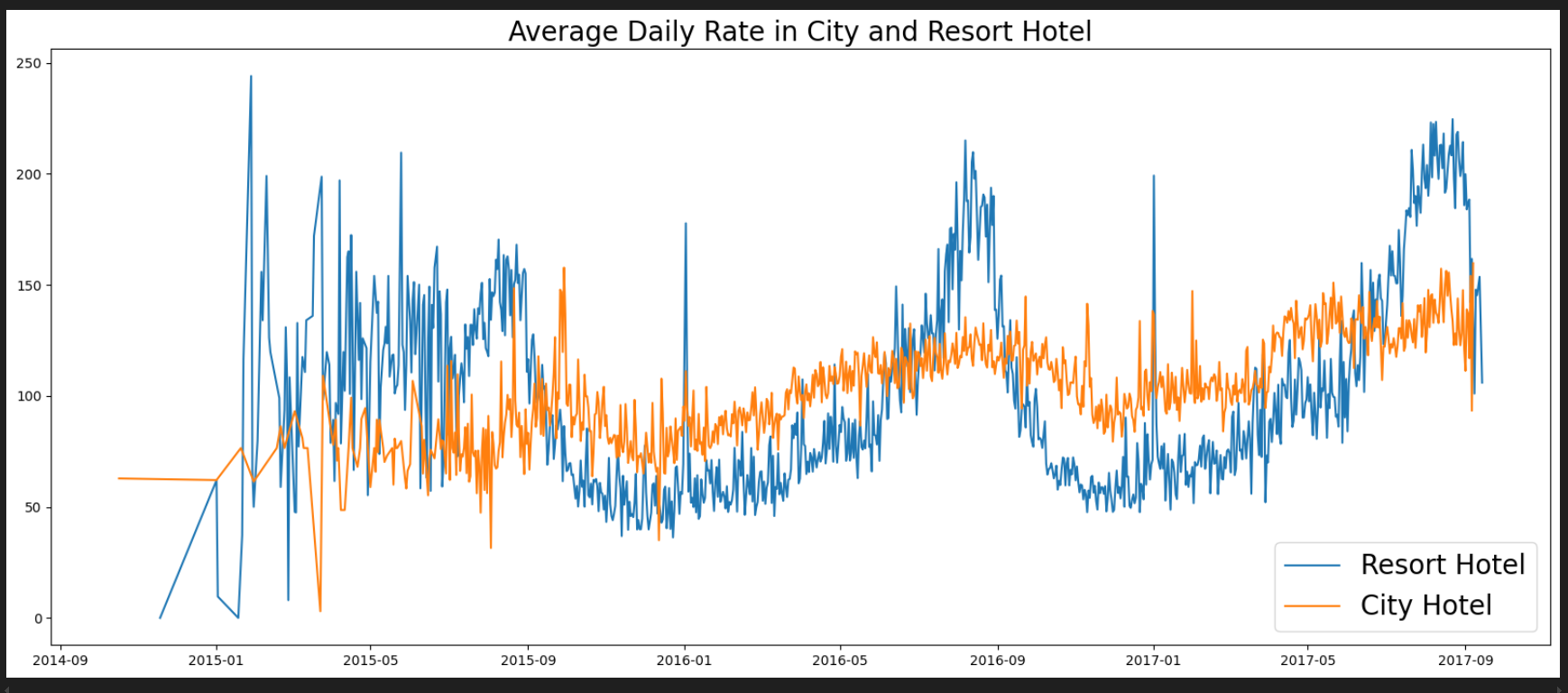


Figure 3: Daily rates of hotels

This figure represents the daily fluctuating rates of hotel rooms. We can observe the different price rate scenarios every time. This might be the result of occasions and weather conditions of the place. From this we can observe whether the price is an influential factor in cancellation or not. So, let’s check the daily rate of rooms with respect to months when the rooms are getting cancelled.

A graph of blue bars

Description automatically generated

Figure 4: Average daily rate of hotel

We can observe clearly that the cancellation rates are higher in those months with higher rates so yes price can be considered as one prime feature that affects the cancellation rate of customers. Hotels must decrease the rates for those seasons where the rooms are getting cancelled which might ultimately result in an increase in customer flow.

Now, check for the country with respect to the cancellation rates. There are lots of countries listed so let’s only visualize in pie chart for top 10 countries.

A screenshot of a graph

Description automatically generated

Figure 5: Cancellation rates from top 10 country

We can observe that 70 % of total cancellation comes from the customers of Portugal followed by Britain, Spain, France and so on, which might be because of the language barrier or the irrelevant environment for them. Hotel are considered to introduce new schemas for more attraction from Portugal and necessary to keep the staff relevant to most of foreign language.

We are going to build two models of decision tree and random forest classification  
Here, I built models on different basis that I am going to show on the tabular form.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Basis** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
|  | Model | 0.9169 | 0.8870 | 0.8899 | 0.8884 |
| Random Forest | Model and hyperparameter | 0.8255 | 0.9147 | 0.5860 | 0.7144 |
|  | Feature selection | 0.7662 | 0.8016 | 0.4934 | 0.6108 |
|  | Feature selection and hyper parameter tuning | 0.8346 | 0.9053 | 0.6200 | 0.7360 |
|  | Model | 0.93464 | 0.9664 | 0.8538 | 0.9066 |
|  | Model and hyperparameter | 0.9021 | 0.9255 | 0.8011 | 0.8588 |
| Decision Tree | Feature selection | 0.7668 | 0.8086 | 0.4882 | 0.6088 |
|  | Feature selection and hyper parameter tuning | 0.82477 | 0.8431 | 0.6495 | 0.7337 |

In this project, we aimed to build predictive models to determine hotel booking cancellations using Random Forest and Decision Tree classifiers. Our evaluations covered various scenarios: raw model performance, hyperparameter tuning, feature selection, and a combination of both feature selection and hyperparameter tuning.

The Random Forest model initially achieved high accuracy (91.69%) and F1 score (88.84%), demonstrating its effectiveness in handling complex data. However, hyperparameter tuning alone led to a notable drop in accuracy (82.55%), indicating potential overfitting or instability. Feature selection alone decreased accuracy (76.62%) but combining it with hyperparameter tuning improved results to an accuracy of 83.46% and an F1 score of 73.60%.

The Decision Tree model exhibited strong initial performance with an accuracy of 93.46% and high precision but lower recall. Hyperparameter tuning slightly reduced accuracy to 90.21%. Feature selection alone resulted in a lower accuracy (76.68%) but combining it with hyperparameter tuning led to an accuracy of 82.48% and an F1 score of 73.37%.

In conclusion, both Random Forest and Decision Tree models show that while initial performances were high, careful tuning and feature selection are crucial for optimizing model performance. The results emphasize the need for a balanced approach in model development, considering both feature selection and hyperparameter tuning to achieve the best predictive accuracy and generalizability.