HOTEL BOOKING DEMAND DATASET ANALYSIS

➔ INTRODUCTION: ABOUT THIS DATASET:

Have you ever wondered when the best time of the year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?This hotel booking dataset can help you explore those questions**!**

➔ HOW DATA WAS ACQUIRED:

Data was extracted from hotels' Property Management System(PMS) database, and downloaded from the Kaggle website.

➔ DATA SOURCE LOCATION:

Both hotels are located in Portugal: H1 at the resort region of Algarve and H2 at the city of Lisbon.

➔ CONTENT:

The dataset for this project contains booking information for a City hotel and a Resort hotel, and includes information as when the booking was made, number of adults,children and/or babies, length of stay, and the number of available parking spaces, among other things.

➔DETAIL DESCRIPTION:

We have two criteria hotel demand data. One of the hotels is a Resort hotel (H1), while the other is a City hotel (H2). Both datasets share the same structure, with 31 variables describing the 40060 observations of H1 and 79330 observations of H2. Each observation represents a hotel booking.

Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July,2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled.Since this is a hotel real data, all data elements pertaining to hotel or customer identification were deleted.

However, the remaining tourism and travel industries like hospitality, cruising, theme parks, etc., have different requirements and peculiarities that cannot be fully explored without industry's specific data. Hence, two hotel datasets with demand data are shared to help in overcoming this limitation.

The datasets now made available were collected aiming at the development of prediction models to classify a hotel booking's likelihood to be canceled. Nevertheless, due to the characteristics of the variables included in these datasets, their use goes beyond this cancellation prediction problem.

A word of caution is due for those not so familiar with hotel operations. In the hotel industry, it is quite common for customers to change their booking's attributes, like the number of persons, staying duration, or room type preferences, either at the time of their check-in or during their stay. It is also common for hotels not to know the correct nationality of the customer until the moment of check-in. Therefore, even though the capture of data took into consideration a timespan prior to the arrival date, it is understandable that the distribution of some variables differ between non cancelled and cancelled bookings.

➔ END-USER OBJECTIVE FOR HOTEL BOOKING DEMAND IS:

By utilizing the predictive model, hotels will be able to identify the reason why customers cancel their bookings and their reasons for doing so, also, what time of the year these cancelations are experienced. It would be grateful to identify the root cause and better strategy for hotel management teams.

We are going to extract insight about most of the customers' preferred choice between Resort or City hotel.

By using Exploratory Data Analysis(EDA), we shall be analysing and visualising the different prospective, which will give us different insights into

1. What is the percentage of bookings for each year?,
2. Which are the busiest months?,
3. How many bookings were cancelled?, etc).

Visualizing more insights by performing Exploratory Data Analysis, we shall be helping hotels to get prepared with adequate and timely arrangements on heavy and low tourist visit periods.

➔ IN SUMMARY, WE WILL TRY TO PROVIDE ANSWERS TO THE FOLLOWING QUESTIONS:

1. How many bookings were cancelled?
2. What is the percentage of booking for each year?
3. Which are the busiest months for hotels?
4. From which country does most guests come from?
5. How long does a guest stay in the hotel?
6. What is the booking ratio between Resort hotel and City hotel?
7. Which was the most booked accommodation type(single, couple, family)?

When we are done with providing answers to these questions, we will build a predictive model to make predictions in the future whether a booking will be cancelled or not.

➔ WE WILL:

1. Perform Data Exploration
2. Data Cleaning

3. Feature Engineering / Feature Selection to Make New Features and also select only relevant ones.
4. Transform the Data(Categorical to Numerical)
5. Split the Data(Train Test Split)
6. Model the Data(Fit the Data)
7. Evaluate the Model

➔ TOOLS AND LIBRARIES USED:

We will deploy Python 3 and some of its packages:

1. Pandas
2. Matplotlib
3. Seaborn
4. Sklearn

➔ Dataset contains the following features:

1. Hotel
2. Is_canceled
3. Lead_time
4. Arrival_date_year
5. Arrival_date_month
6. Arrival_date_week_number
7. Arrival_date_day_of_month
8. stays_in_weekend_nights
9. Stays_in_week_nights
10. Adults
11. Children
12. Babies
13. Meal
14. Country
15. Market_segment
16. Distribution_channel

17. Is_repeated_guest
18. previous_cancellations
19. previous_bookings_not_canceled
20. Reserved_room_type
21. Assigned_room_type
22. Booking_changes
23. Deposit_type
24. Agent
25. Company
26. Days_in_waiting_list
27. Customer_type
28. Adr
29. Required_car_parking_spaces
30. Total_of_special_requests
31. Reservation_status
32. Reservation_status_date

MODEL DEFINITION

➔ Decision Tree Algorithm:

Decision trees come under the supervised learning algorithms category. It is primarily used for regression and classification in machine learning models. It provides transparency by offering a single view of all traces and alternatives. Decision trees also assign specific values to problems and decisions, enabling better decision-making.Decision tree is very easy to understand and communicate

➔ Advantages of Using Decision Tree Algorithm:

- Interpretation and visualization is made easy when Decision trees are used.

- Capturing Nonlinear patterns is easier.
- Normalization of columns is not needed as negligible data preprocessing is required from the user.
- Variable selection can be more efficiently done.
- Feature engineering such as predicting missing values can be done very efficiently using this algorithm.
- There are no assumptions about distribution because the decision tree has a non-parametric nature.

➔ **Disadvantages of Using Decision Tree Algorithm:**

- Overfitting noisy data and sensitivity to noisy data.

➔ **LOGISTIC REGRESSION:**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. Because of its efficient and straightforward nature, it doesn't require high computation power, easy to implement, easily interpretable, used widely by data analysts and scientists. Also, it doesn't require scaling of features. Logistic regression provides a probability score for observations

➔ **Advantages of Logistic Regression**:

- It is a widely used technique because it is very efficient, does not require too many computational resources, it’s highly interpretable, it doesn’t require input features to be scaled, it doesn’t require any tuning, it’s easy to regularize, and it outputs well-calibrated predicted probabilities.
- Another advantage of Logistic Regression is that it is incredibly easy to implement and very efficient to train.
- Because of its simplicity and the fact that it can be implemented relatively easily and quickly, Logistic Regression is also a good baseline that you can use to measure the performance of other more complex Algorithms.

➔ **Disadvantage of Logistic Regression:**

- A disadvantage of it is that we can't solve non-linear problems with logistic regression since it's decision surface is linear.

➔ **Gaussian Naive Bayes Model:**

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

➔ **Random Forest Classifier:**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

➔ **Advantages of Random Forest :**

1. It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
2. It runs efficiently on large databases.
3. It can handle thousands of input variables without variable deletion.
4. It gives estimates of what variables that are important in the classification.
5. It generates an internal unbiased estimate of the generalization error as the forest building progresses.
6. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

➔

➔ Disadvantages of Random Forest :

1. Random forests have been observed to overfit for some datasets with noisy classification/regression tasks.
2. For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

➔ Conclusion:

- We can observe that the best algorithm is Random Forest for this dataset analysis.
- 0 values are uncalculated ones.
- We did not count the decision tree with Reservatiton_Status because all the algorithms would give 100% accuracy scores when Reservation_Status is included.

- Furthermore, we learned that:
    A. Bookings got canceled 37% of the time. While booking guest checked-in (did not cancel the booking) almost 63% of the time.
    B. More than double bookings were made in 2016, compared to the previous year of 2015. But the bookings decreased by almost 15% the next year 2017..
    C. Most bookings were made from July to August. While the least bookings were made at the start and end of the year.

D. Portugal,UK,France,Spain and Germany are the top countries from which most guests come from. This means that more than 80% of guests come from these 5 countries..

E. Most people stay Two,Three,One and four. Therefore,more than 60% of guests come under these options.

   a. For Resort hotels, the most popular stay duration is one, two, three, and four days respectively.

   b. For City hotels, the most popular stay duration is one, two, three, and seven(weeks) respectively.

F. More than 60% of the population booked a City hotel.

G. We observe that couples (or 2 adults) is the most popular accommodation type. So hotels can make plans accordingly.

➔ **References:**

- https://www.kaggle.com/vssseel/eda-various-ml-models-and-nn-with-roc-curves#Preprocessing
- https://medium.com/swlh/random-forest-classification-and-its-implementation
- https://www.datacamp.com
- https://www.machinelearning-blog.com
- https://www.researchgate.net/publication/329286343_Hotel_booking_demand_datasets