## FORE School (@FORE_Delhi) / X

## 

## **GAI & DLP Project**

**Submitted By :**

Group No.

Aadya Shakti (045001)

Diksha Jain (045018)

Ishaan Aggarwal (045021)

**Submitted To :**

Prof. Ashok Harnal

## **REPORT**

## **Description of Data :**

* **Data Description:** The dataset consists of 83,590 customers' records and includes 31 variables that describe their behavior over a period of three years (2015 to 2018). Along with personal and behavioral data, the dataset also contains demographic and geographical information. All personal related data were transformed or anonymized to guaranty privacy and prevent the hotel or guests’ identification. Time-related variables were accounted for based on the last day of the extraction period. The last day of the extraction period is December 31, 2018.
* **Data source location**: The data came from a four-star hotel located in Lisbon, Portugal, Europe. In Portugal, hotels' star classification scale varies from 1 to 5, with one-star being the low-end quality hotels and five-star being the high-end quality hotels.
* **Descriptive variables :**

ID: Cutomer Id, Numeric

Nationality: country of origin, Categorical

Age: customer's age (in years), Numeric

Name Hash: Name of the customer's SHA2–256 hash string, categorical

DocID Hash: SHA2–256 hash-string of the identification document number the customer provided at check-in (passport number, national ID card number, or other), Categorical

DistributionChannel: Categorical

MarketSegment: Categorical

* **Quantifying Variables :**

Average Lead Time: The average number of days elapsed between the customer's booking date and arrival date. In other words, this variable is calculated by dividing the sum of the number of days elapsed between the moment each booking was made and its arrival date, by the total of bookings made by the customer

* **Numeric :**

**Lodging Revenue**: Total amount spent on lodging expenses by the customer (in Euros). This value includes room, crib, and other related lodging expenses, Numeric

**Other Revenue**: Total amount spent on other expenses by the customer (in Euros). This value includes food, beverage, spa, and other expenses, Numeric

**Bookings Canceled**: Number of bookings the customer made but subsequently canceled (the costumer informed the hotel he/she would not come to stay), Numeric

**Bookings No Showed**: Number of bookings the customer made but subsequently made a “no-show” (did not cancel, but did not check-in to stay at the hotel), Numeric

Bookings CheckedIn: Number of bookings the customer made, and which end up with a staying, Numeric

**Persons Nights:** The total number of persons/nights that the costumer stayed at the hotel. This value is calculated by summing all customers checked-in bookings’ persons/nights. Person/nights of each booking is the result of the multiplication of the number of staying nights by the sum of adults and children, Numeric

**Room Nights:** Total of room/nights the customer stayed at the hotel (checked-in bookings). Room/nights are the multiplication of the number of rooms of each booking by the number of nights of the booking, Numeric

**Days Since Last Stay**: The number of days elapsed between the last day of the extraction and the customer's last arrival date (of a checked-in booking). A value of −1 indicates the customer never stayed at the hotel, Numeric

**Days Since First Stay:** The number of days elapsed between the last day of the extraction and the customer's first arrival date (of a checked-in booking). A value of −1 indicates the customer never stayed at the hotel, Numeric

**SR High Floor**: Indication if the customer usually asks for a room on a higher floor (0: No, 1: Yes), Boolean

**SR Low Floor**: Indication if the customer usually asks for a room on a lower floor (0: No, 1: Yes), Boolean

**SR Accessible Room**: Indication if the customer usually asks for an accessible room (0: No, 1: Yes), Boolean

**SR Medium Floor:** Indication if the customer usually asks for a room on a middle floor (0: No, 1: Yes), Boolean

**SR Bath tub:** Indication if the customer usually asks for a room with a bathtub (0: No, 1: Yes), Boolean

**SR Shower**: Indication if the customer usually asks for a room with a shower (0: No, 1: Yes), Boolean

**SR Crib**: Indication if the customer usually asks for a crib (0: No, 1: Yes), Boolean

**SR KingSizeBed:** Indication if the customer usually asks for a room with a king-size bed (0: No, 1: Yes), Boolean

SR TwinBed Indication if the customer usually asks for a room with a twin bed (0: No, 1: Yes), Boolean

**SR Near Elevator**: Indication if the customer usually asks for a room near the elevator (0: No, 1: Yes), Boolean

**SR AwayFromElevator**: Indication if the customer usually asks for a room away from the elevator (0: No, 1: Yes), Boolean

**SR No Alcohol In MiniBar:** Indication if the customer usually asks for a room with no alcohol in the mini-bar (0: No, 1: Yes),Boolean

**SR Quiet Room:** Indication if the customer usually asks for a room away from the noise (0: No, 1: Yes), Boolean

## **Analysis of Data**

Average age of the respondents is 45 years.Average led time of all the respondents being converted to actual revenue brining customers is 66 days.Average lodging revenue for 3 years is Dolar 299 and average of the revenue coming from other sources is Dolar 67 for 3 years.Booking Cancelled and Booking no show rates are almost negligable, indicating that the conversion rate is high for the hotel.Day since last stay and first stay go hand in hand, there is hardly any difference in their average values.60% of the customers have their nationality as France, Portugal, Germany and Great Britain, with the maximum 17.3% coming from France and closely followed by protugal, 16.2%.Among the four distribution channels through which the customers book their rooms in the hotel, namely, Travel Agent/Operator, Direct, Corporate and Electronic Distribution, Travel Agent/Operator constitute the maxium percentage, 82%.Among seven the Market segments that the customers can belong to, it is observed that the 57% of the customers cannot be classified in any specifoc segment and come under the 'Others' category. Nevertheless, a total of 39% of the customers belong to Travel Agent/Operator, groups and direct customers.

Further, from all the Boolean variables in the dataset only three variables have been analysed, which are

a. requirements of twin size beds

b. requirements for king size beds

c. requiremnt of alcohol minibars in the room

Only these three variables were used because they directly contribute to the revenue for the hotel.

From these three variables the following was observed:

a. Only 35% of the customers asked for king size beds and 14% asked for twin sized beds

b. No customer requested for alcohol minibars in their rooms

**Inferences :**

The hotel can tailor its services and marketing strategies to attract customers in the age group of 45 years. This could include offering amenities and activities that are popular among this age group, such as spa treatments, fitness classes, and cultural experiences.

The reasons for the long led time can be analysed and steps can be taken to reduce it, such as streamlining the booking process, improving communication with customers, and offering incentives for early booking.

The hotel can also analyze the revenue sources and explore ways to increase the revenue from other sources, such as food and beverage sales, event bookings, and merchandise sales. This could include offering unique and high-quality products and experiences that differentiate the hotel from competitors.

The high conversion rate indicates that the hotel is doing a good job of converting bookings into actual revenue. The hotel can continue to focus on providing excellent customer service, personalized experiences, and competitive pricing to maintain and increase its conversion rate.

The similarity in the average values of the day since last stay and first stay suggests that the hotel has a loyal customer base. The hotel can continue to build and strengthen its relationships with existing customers by offering loyalty programs, personalized recommendations, and special promotions. The hotel can also encourage customers to refer their friends and family to increase the customer base.

The hotel can focus its marketing efforts towards the top nationalities, i.e., France, Portugal, Germany, and Great Britain.

Since Travel Agent/Operator is the most preferred booking channel, the hotel can form partnerships with popular travel agencies to attract more customers.

The hotel can analyze the needs and preferences of customers in the 'Others' category and try to create customized packages or offers to attract more customers.

The hotel can consider introducing more room options with king-size beds to cater to the preferences of customers who prefer larger beds.

## **Data Preprocessing**

**Objectives :**

Identifying the null values in both cat and non-cat datasets

1. Treating the null values by imputation

2. Numeric coding of cat variables

3. Identifying outliers using boxplots

4. Treating the outliers

5. Combining the pre-processed cat and non-cat dataframes into one (df\_ppd)

6. Diving the combined pre-processed dataset (df\_ppd) into two, testing and

training datasets (test\_df and train\_df)

**Brief on Analysis:**

The null values in the datasets were identified using the info() functions.These null values were then treated using Simple Imputer with strategy as 'mean'.Numeric encoding of the categorical variables (Distribution Channel and Market Segment) was done using Label Encoder as the data was nominal.

After identifying the outliers in each variable, they were treated using Normalization -> Min-Max Scaling. The cat and non\_cat pre-processed datasets were then merged into one using the 'merge' function.

**Outliers Treatment :**

From the box plots, it can be clearly seen that 'AverageLeadTime', 'Age', 'LodgingRevenue', 'OtherRevenue', 'BookingsCanceled', 'BookingsNoShowed', 'BookingsCheckedIn' and 'PersonsNights' are the variables that contain the outliers. Hence, using normalisation we will be treating these outliers.

## 

## **Formulating The Questions :**

**Customer Demographics:**

What is the age distribution of the customers?

How does customer nationality impact booking patterns?

**Booking Behavior:**

What is the relationship between AverageLeadTime and the likelihood of a booking being canceled or not showing up?

How do the different MarketSegment and DistributionChannel features impact revenue?

**Revenue Insights:**

What is the relationship between RoomNights, PersonsNights, and total LodgingRevenue?

How does OtherRevenue vary with different room preferences (e.g., SRKingSizeBed, SRTwinBed)?

**Customer Loyalty:**

How does DaysSinceFirstStay and DaysSinceLastStay correlate with repeat bookings?

**Feature Interactions:**

How do different service requests (e.g., SRQuietRoom, SRBathtub) interact with the likelihood of customer satisfaction (inferred from BookingsCheckedIn)?

**Target Interaction:**

How do the various features impact the target variables like BookingsCheckedIn and LodgingRevenue?

## **Visualizations :**

### **Distributions**

**Age Distribution :** The age distribution in the dataset reveals a multimodal pattern, with significant peaks around ages 30, 40, and 60, indicating higher representation in these age groups. The distribution is right-skewed, with a concentration of individuals primarily between 20 and 70 years old. The data also suggests a gradual decline in frequency after age 60, with few outliers at extreme ages. These insights highlight the presence of distinct age segments that could be crucial for age-related analysis or targeted interventions.

The normalized age distribution plot illustrates a heavily right-skewed distribution, where the majority of the data is concentrated near the lower end of the normalized age scale (close to 0). This suggests that most individuals in the dataset are relatively young, with fewer older individuals represented as the normalized age increases.

**Revenue Distribution :** The box plot illustrates the distribution of lodging revenue across various market segments. Each segment shows a wide range of revenue, with several outliers present, particularly in the "Other" and "Corporate" segments, where a few extreme values significantly exceed the median. The bulk of the revenue data for all segments is concentrated near the lower end of the scale, indicating that most transactions generate relatively low revenue. Despite this, the presence of outliers suggests that certain segments occasionally generate much higher revenue, contributing to overall variability.

The box plot reveals that normalized lodging revenue varies significantly across market segments. While Market Segment 0 and 1 have lower median revenues and smaller IQRs, Market Segments 2, 3, 4, 5, and 6 exhibit higher median revenues and wider distributions. Additionally, outliers are observed in several segments, indicating extreme revenue values. Overall, the plot suggests that market segmentation is effective in identifying customer groups with distinct spending patterns.

**Heatmap :** The correlation heatmap provides a visual representation of the relationships between different variables in the dataset. The colors indicate the strength and direction of the correlations, with red representing positive correlations and blue representing negative correlations. For example, the strong positive correlation between RoomNights and PersonsNights\_norm suggests that the number of nights stayed is closely related to the number of people staying. On the other hand, the weak negative correlation between SRQuietRoom and BookingsNoShowed\_norm indicates that guests who request quiet rooms are less likely to cancel their bookings. Overall, the heatmap helps identify important relationships and potential dependencies among the variables in the dataset.

**BarPlot :** The bar plot illustrates the average lodging revenue generated by different distribution channels. Corporate, Travel Agent/Operator, Direct, and Electronic Distribution channels have relatively similar average revenues, with minor variations. The error bars indicate the variability within each channel, suggesting that while the average revenues are comparable, there may be some differences in individual revenue values. Overall, the plot suggests that the choice of distribution channel does not have a significant impact on the average lodging revenue.

### 

### **Statistical Tests :**

1. **Hypothesis Testing**

Null Hypothesis (H0):The mean normalized revenue is the same between the two market segments (no significant difference).

Alternative Hypothesis (H1): The mean normalized revenue differs significantly between the two market segments.

T-statistic: The T-statistic of -13.179 indicates that the difference in means between the two market segments is quite large relative to the variability within each segment. A negative value means that the mean normalized revenue of segment 1 is lower than that of segment 2.

P-value:The p-value of 1.7835777357642447e-38 is extremely small, much less than any common significance level (e.g., 0.05, 0.01). This indicates a very strong evidence against the null hypothesis.

Conclusion: Given the extremely low p-value, we can confidently reject the null hypothesis. There is a statistically significant difference in the mean normalized revenue between the two market segments. The negative T-statistic suggests that, on average, segment 1 has a lower normalized revenue compared to segment 2.

1. **Chi Square Test**

Null Hypothesis (H0): There is no association between MarketSegment\_code and BookingsCheckedIn\_norm. The variables are independent.

Alternative Hypothesis (H1): There is an association between MarketSegment\_code and BookingsCheckedIn\_norm. The variables are dependent.

Chi-square Statistic: The Chi-square statistic of 6980.665 is very large, indicating a significant deviation from the expected counts under the null hypothesis of independence.

P-value : The p-value of 0.0 (or essentially zero) is extremely small, much less than any conventional significance level (e.g., 0.05, 0.01). This indicates very strong evidence against the null hypothesis.

Conclusion: Given the very low p-value, you can reject the null hypothesis with high confidence. There is a significant relationship between MarketSegment\_code and BookingsCheckedIn\_norm. This suggests that the distribution of BookingsCheckedIn\_norm is not independent of MarketSegment\_code and that the market segments influence or are associated with different patterns of bookings checked in.

### **Feature Interaction with Target**

1. **Scatter Plot :** The scatter plot illustrates the relationship between the number of room nights and normalized lodging revenue. Overall, there appears to be a weak positive correlation between the two variables, suggesting that as the number of room nights increases, the normalized lodging revenue tends to increase slightly. However, the relationship is not particularly strong, indicating that other factors may also influence lodging revenue. Additionally, there is a cluster of data points at lower room night values, suggesting that a significant portion of the data falls within this range.
2. **Pair Plot :** The pair plot provides a comprehensive visualization of the relationships between the three variables: RoomNights, PersonsNights\_norm, and LodgingRevenue\_norm. The diagonal plots show the distribution of each variable, while the off-diagonal plots illustrate the pairwise relationships. We can observe a positive correlation between RoomNights and PersonsNights\_norm, suggesting that as the number of room nights increases, the number of people staying also tends to increase. Additionally, there appears to be a weak positive correlation between RoomNights and LodgingRevenue\_norm, indicating that longer stays might lead to slightly higher revenue. However, the relationship between PersonsNights\_norm and LodgingRevenue\_norm is less clear, with no strong correlation observed. Overall, the pair plot helps identify potential relationships and dependencies among the variables in the dataset.

### 

### **Regression Model To Predict Revenue**

The residual plot shows the difference between the actual values (y-test) and the predicted values (y-pred) on the y-axis, and the predicted values on the x-axis.

In a perfect model, the residuals would all be zero and the plot would be a horizontal line at zero. In this case, the residuals are scattered around zero with no clear pattern, which suggests that the linear regression model is a reasonable fit for the data. There are a few outliers, but overall the residuals are randomly distributed.

### **Classification Model to Predict Booking Cancellations**

The Logistic Regression model effectively predicts booking cancellations with an accuracy of 99.86%. The confusion matrix reveals a low number of false positives and false negatives, demonstrating the model's ability to accurately identify both canceled and non-canceled bookings. The classification report further supports the model's performance, showing high precision, recall, and F1-score for both classes. These results suggest that the model can be a valuable tool for hotel management in identifying potential cancellations and taking proactive measures to mitigate their impact.

### **Neural Networking Model to Predict Revenue**

The code constructs a Neural Network model to predict booking cancellations in a hotel reservation dataset. The model employs a Sequential architecture with three layers: an input layer, a hidden layer with 64 neurons, and an output layer with one neuron. The model is trained using the Adam optimizer and the binary cross entropy loss function. Evaluation metrics like Mean Squared Error and R-squared are calculated to assess the model's performance. The training and validation loss curves are plotted to visualise the model's learning process, while the training and validation accuracy curves are plotted to assess its classification performance.

The graphs illustrate the performance of the Neural Network model during training. The training and validation loss curves show that the model is learning effectively, with both curves decreasing over epochs. However, a slight gap between the curves suggests potential overfitting. The training and validation accuracy curves also indicate that the model's classification performance is improving. Overall, the model appears to be performing well, but further analysis of evaluation metrics and experimentation with hyperparameters could enhance its performance.

## 

## **6. Results and Observations :**

This analysis explored various aspects of a hotel reservation dataset using data visualization and machine learning techniques. The goal was to gain insights into customer behavior, revenue patterns, and factors influencing booking cancellations.

**Revenue Analysis**

* Market Segment Analysis:
  + Significant variations in revenue across market segments were observed.
  + Aviation and Direct segments demonstrated higher median revenue and wider distributions, suggesting premium pricing or unique offerings.
  + Corporate and Travel Agent/Operator segments had lower median revenues and smaller IQRs, indicating more concentrated revenue patterns.
* Distribution Channel Analysis:
  + No significant differences were found in average lodging revenue across distribution channels.
* Room Nights and Revenue:
  + A weak positive correlation was observed between room nights and lodging revenue, suggesting that longer stays might lead to slightly higher revenue. However, other factors likely influence revenue as well.

**Cancellation Analysis**

* Logistic Regression Model:
  + A Logistic Regression model was built to predict booking cancellations.
  + The model achieved a high accuracy of 99.86%, indicating its effectiveness in identifying potential cancellations.
  + The confusion matrix and classification report confirmed the model's ability to accurately classify both canceled and non-canceled bookings.

**Neural Network Model**

* Predicting Booking Cancellations:
  + A Neural Network model was developed to predict booking cancellations.
  + The model's performance was evaluated using metrics like Mean Squared Error and R-squared.
  + The training and validation loss curves suggested that the model was learning effectively, with some signs of overfitting.
  + The training and validation accuracy curves indicated that the model's classification performance was improving.

**Observations and Recommendations**

* Market Segmentation: The findings highlight the importance of effective market segmentation for understanding customer behavior and optimizing revenue strategies.
* Revenue Drivers: Further analysis is needed to identify specific factors driving higher revenue in certain market segments and distribution channels.
* Cancellation Prediction: The Logistic Regression model provides a valuable tool for predicting booking cancellations and allowing hotels to take proactive measures.
* Neural Network Refinement: The Neural Network model could be further refined by addressing potential overfitting and exploring different model architectures or hyperparameters.
* Additional Factors: Other factors, such as customer demographics, booking lead times, and room amenities, might also influence revenue and cancellations. Future analysis could explore these relationships.

**Conclusion**

This analysis has shed light on key aspects of the hotel reservation data. The findings provide valuable insights for hotel management to optimize revenue, improve customer satisfaction, and reduce the impact of booking cancellations. Further research and analysis could delve deeper into specific areas of interest to gain more comprehensive understanding.