Enhancing Hospitality Predictions: Analyzing Model Performance Selection for Hotel Booking Data

Group Members:

Tharun Cota

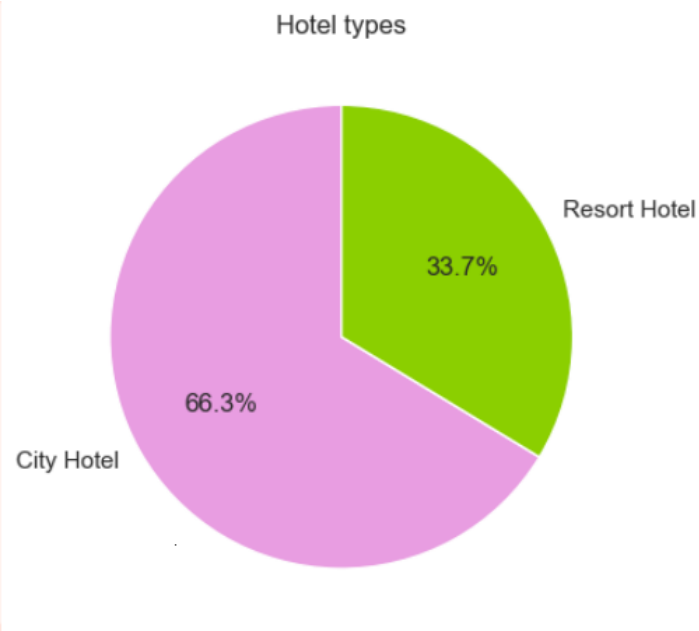
Suhas K M

Introduction:

This data set contains booking information for city and resort hotels, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All the original personal identifying information has been removed from the data and has been fabricated with duplicate values.

Source of the Data and More:

- ➢ Each row is a booking so questions will revolve around answering or predicting what are the possible behavior and features of the booking given a set of input features.
- ➢ Each row is a booking so questions will revolve around answering or predicting what are the possible behaviour and features of the booking given a set of input features.
- ➢ TSQL queries were executed directly on the hotels' PMS databases on SQL Server Studio Manager
- ➢ PMS - A property management system (PMS) is a software application for the operations of hospitality accommodations and commercial residential rental properties. PMS is also used in manufacturing industries, local government and manufacturing.



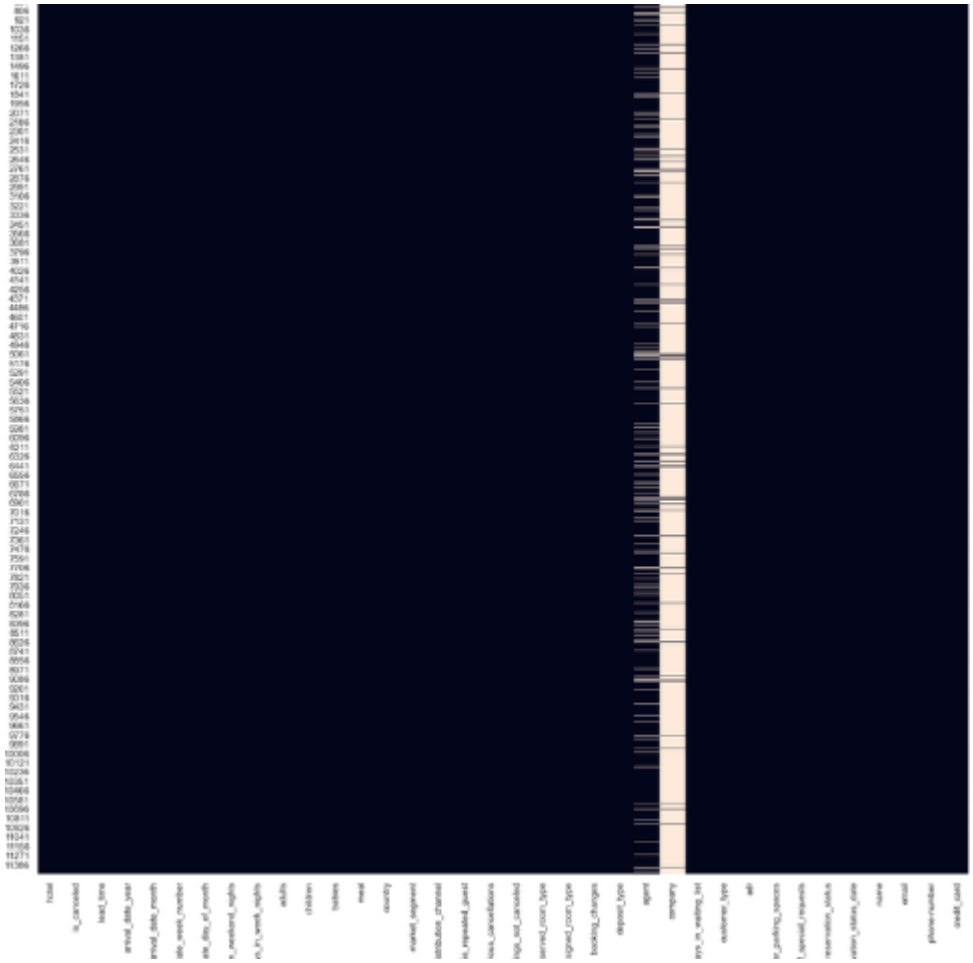
Our Data and More:

➢ Categorical:

hotel, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, agent, country, customer_type, reservation_status_date, name, email, phone_number, credit_card

➢ Ordinal / Numerical:

is_cancelled, lead_time, arrival_date_year, arrival_date_week_number, arrival_date_month, stays_in_week, weekend_nights, adults, children, babies, is_repeated_guests, previou_cancellations, previous_bookings_not_cancelled, booking_changes, company, agent, days_in_waiting_list, adr, required_car_parking_spaces, total_of_requests



EDA:

The Exploratory Data Analysis is the crucial step to gain a comprehensive understanding of the dataset's characteristics, structure, and content. The primary objective is to gain a comprehensive understanding of the hotel booking dataset by delving into its characteristics, structure, and content, where we found there is a need to understand and build analysis and models for two types of Hotels present in the Dataframe, i.e, City Hotel and Resort Hotel. This involves an initial exploration whereby the dataset's dimensions, number of rows and columns, and a cursory review of a few rows to obtain a preliminary sense of the data relied more on the analysis focusing on identifying the data types in each column (numerical, categorical, etc.) and checking for any missing or null values. Descriptive statistics, including mean, median, min, max, etc., are computed for numerical features, while frequency counts are obtained for categorical features. Visualization techniques such as histograms, kernel density plots, and bar charts

provide insights into the distribution of numerical and categorical features, respectively. Correlation analysis, using correlation matrices and heatmaps, helps uncover potential relationships between numerical features. Outlier detection is performed through visualizations like box plots and scatter plots. Feature engineering considerations involve exploring opportunities for creating new features and encoding categorical variables for machine learning models. Domain-specific knowledge is leveraged to understand the significance of specific features and relate insights to the broader context of the business problem. Findings, anomalies, and patterns discovered during exploration are meticulously documented, and a data dictionary is created to define the meaning of each variable. In conclusion, this systematic exploration forms the foundation for subsequent data preprocessing, modeling, and decision-making processes.

The analysis of the hotel booking dataset encompassed a meticulous exploration of both numerical and categorical features. For numerical features, a comprehensive examination of data types, ranges, and distribution characteristics was conducted. This involved assessing the types of numerical features, investigating their ranges to identify potential outliers, utilizing descriptive statistics for central tendency and spread, and employing visualizations such as histograms and box plots for a visual exploration of the data distribution. In parallel, the analysis of categorical features involved understanding their data types, determining unique values, calculating frequency distributions, and visualizing the categorical variable distribution through bar charts or count plots. The inspection of data types, ranges, and the overall structure of the dataset ensured alignment with expectations and modeling requirements. Anomalies and outliers were detected using both statistical methods and visualizations, prompting considerations for appropriate handling strategies based on their impact on analysis and modeling. This data exploration phase provided a holistic understanding of the dataset, laying a robust foundation for subsequent preprocessing and modeling endeavors.

We opted to sample 10% of the original dataset to streamline the modeling process while ensuring a representative subset for our investigation. However, this subsampling revealed instances where certain features contained missing values. The presence of missing data is a critical consideration in EDA, as it has implications for the reliability and completeness of subsequent analyses. Majority of the missing data were in agent and company features.

To address this, our analysis involved a meticulous examination of the sampled data, focusing on identifying the specific features and instances affected by missing values. This scrutiny aimed to understand the nature and patterns of missing data, providing a foundation for making informed decisions on how to handle these gaps in later stages of the analysis.

The commencement of our analytical scrutiny was centered on the examination of the Distribution Channel and Market Segments within the dataset. This deliberate focus aimed to discern underlying correlations between the missing values across these specific columns. An intrinsic observation surfaced, indicating that instances pertaining to certain features were exclusively filled when a customer opted for booking through a Corporate entity or an agent. Conversely, when reservations were made through other channels, the corresponding features remained unpopulated.

In a specific instance from the dataset, James McCann undertook a reservation for a room at the Resort Hotel for the date of 10th March 2016, a Thursday. The reservation was made precisely 219 days in advance, indicating the booking date as 4th August 2015. James planned to stay until Wednesday, departing on the 17th of March 2016. Accompanied by an additional adult, James, a resident of Portugal, falls under the Transient customer type, booking through the market segment Offline TA/TO and

utilizing the corresponding distribution channel.

James's Average Daily Rate (ADR) was recorded at 52.0, leading to an estimated expenditure of 364.0 considering a seven-day stay. As a first-time customer, both previous bookings and cancellations for James were documented as zero. Regrettably, James canceled his reservation. The reserved room type was denoted as "A," coinciding with the room type assigned by the hotel. The transaction involved a non-refundable policy, facilitated through Agent 310.0. Additionally, James indicated zero special requests and opted for a Half-Board (HB) meal plan. This detailed instance offers a contextualized understanding of the dataset, highlighting the diverse parameters associated with a single reservation entry.

This identified correlation became instrumental in our subsequent model-building endeavors. During the data cleaning process, we selectively addressed instances where correlation was absent or where no informative value was derived. This discerning approach allowed us to retain and leverage the data points that contribute meaningfully to the analytical objectives, while omitting those that lacked substantive correlation.Furthermore, a nuanced investigation uncovered a minor fraction of missing values within the country columns. Rather than outright removal of these instances, a strategic decision was made to impute these gaps by assigning the placeholder value 'others.' This nuanced handling of missing values aligns with the broader analytical strategy, preserving the integrity of the dataset and facilitating a more nuanced and comprehensive approach to subsequent modeling and analysis.

Diving Into Visualizations:

In the realm of exploratory data analysis (EDA), visualization emerges as a pivotal tool, serving as a lens through which we gain a nuanced understanding of both numerical and categorical features within the dataset. Through meticulous employment of visualizations, our objective was not only to unravel the inherent distributional patterns but also to establish correlations, enabling a profound comprehension of the underlying problem statement.

The process of visualization played a crucial role in steering our analytical journey toward the formulation of both classification and regression models. The classification model was conceived with the primary aim of predicting whether a customer is likely to cancel or uphold their booking. This predictive insight holds significant strategic importance for hotel management in optimizing resource allocation and enhancing customer service. On the other front, the regression model was tailored to prognosticate the Average Daily Rate (ADR), a pivotal metric influencing sales strategies. By understanding and forecasting the ADR, our aim was to empower the business with actionable intelligence for refining pricing structures and maximizing revenue generation.

The visualizations acted as a compass, guiding us through the intricate landscape of data intricacies and paving the way for informed decision-making. As we delved into the multifaceted facets of the dataset, each chart and graph became a narrative, unraveling the story embedded in the data. This robust analytical foundation, fostered by comprehensive visual exploration, positioned us strategically for the subsequent phases of model development and optimization.

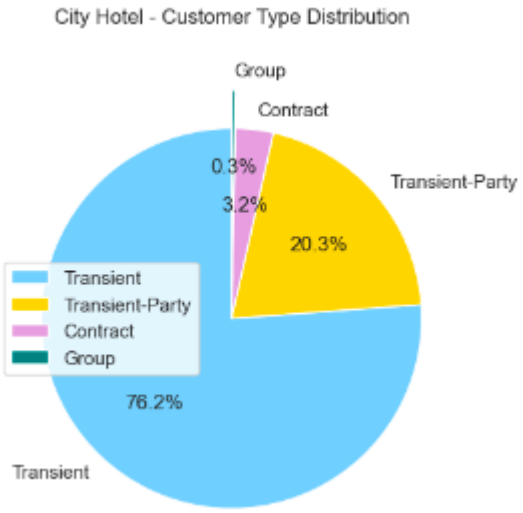


Fig - Customer Type Distribution for City Hotel

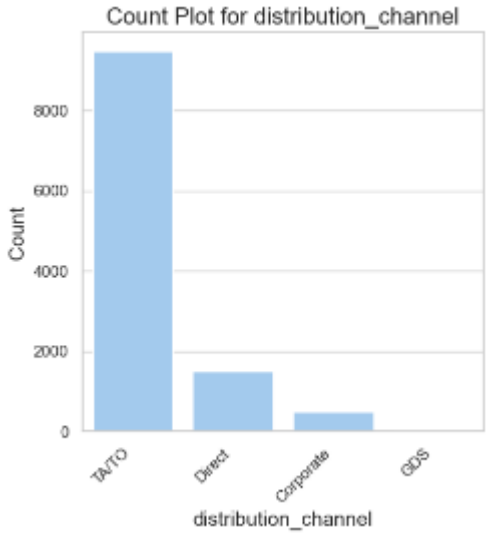


Fig - Visualising Distribution Channel V/S Count

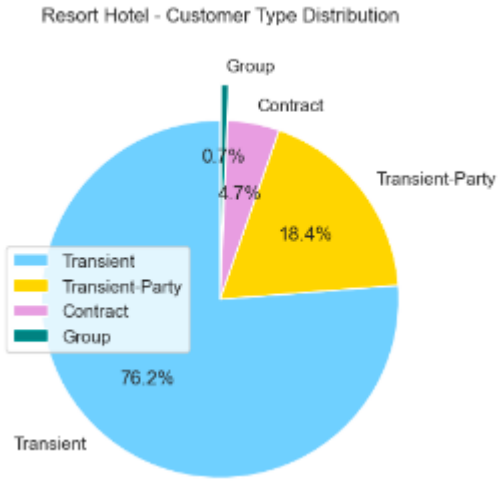


Fig - Customer Type Distribution for Resort Hotel

hotel	City Hotel
is_canceled	1
lead_time	25
arrival_date_year	2017
arrival_date_month	June
arrival_date_week_number	26
arrival_date_day_of_month	27
stays_in_weekend_nights	0
stays_in_week_nights	1
adults	2
children	0.0
babies	0
meal	BB
country	PRT
market_segment	Corporate
distribution_channel	Corporate
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	A
assigned_room_type	A
booking_changes	0
deposit_type	No Deposit
agent	NaN
company	280.0
days_in_waiting_list	0
customer_type	Transient-Party
adr	100.0
required_car_parking_spaces	0
total_of_special_requests	0
reservation_status	Canceled
reservation_status_date	2017-06-02
name	Andrea Parker
email	AParker@att.com
phone-number	404-320-0760
credit_card	************9809

Name: 4130, dtype: object

Fig - The above image is an instance of a City Hotel.

hotel	Resort Hotel
is_canceled	0
lead_time	53
arrival_date_year	2016
arrival_date_month	June
arrival_date_week_number	24
arrival_date_day_of_month	7
stays_in_weekend_nights	0
stays_in_week_nights	2
adults	2
children	0.0
babies	0
meal	BB
country	CHN
market_segment	Online TA
distribution_channel	TA/TO
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	A
assigned_room_type	A
booking_changes	0
deposit_type	No Deposit
agent	240.0
company	NaN
days_in_waiting_list	0
customer_type	Transient
adr	129.0
required_car_parking_spaces	0
total_of_special_requests	1
reservation_status	Check-Out
reservation_status_date	2016-06-09
name	Bryan Brown
email	Bryan_Brown@protonmail.com
phone-number	417-637-3312
credit_card	**********3396

Name: 71, dtype: object

Fig - The above image is a single instance of a Resort Hotel.

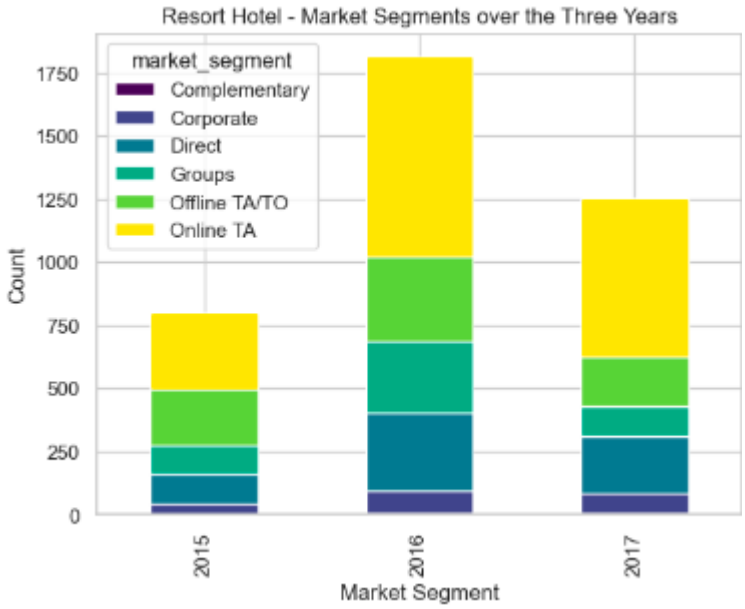


Fig - Margaret Segment for a Resort Hotel

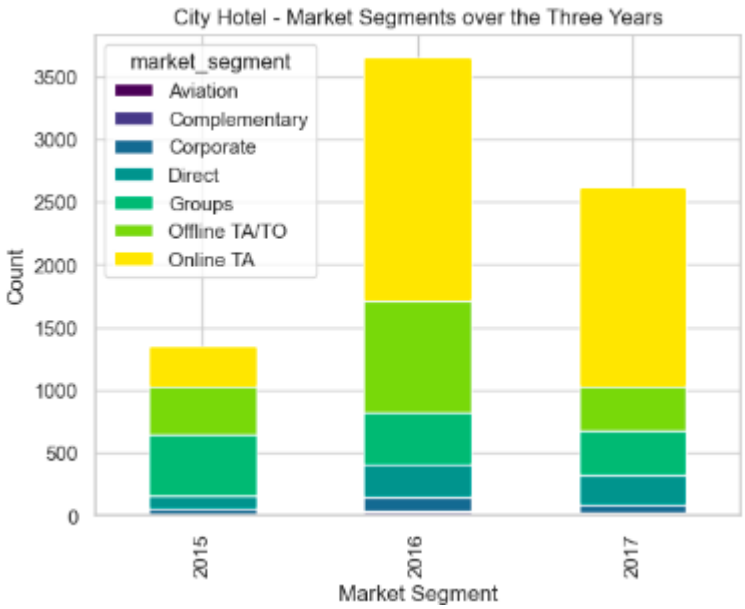


Fig - Market Segment for a City Hotel

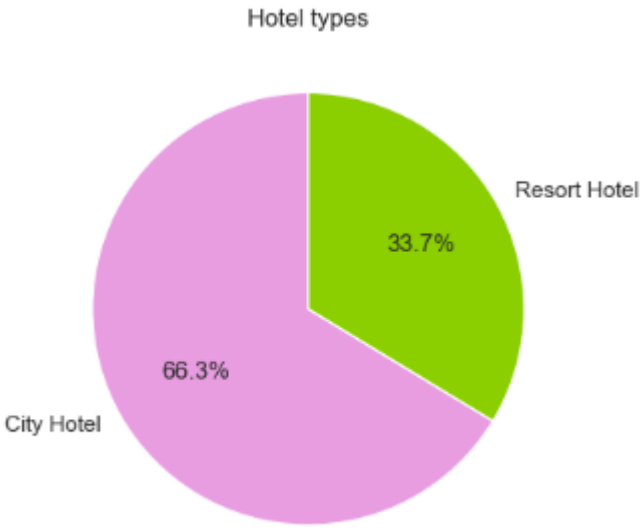


Fig - Split of the Data into City and Resort

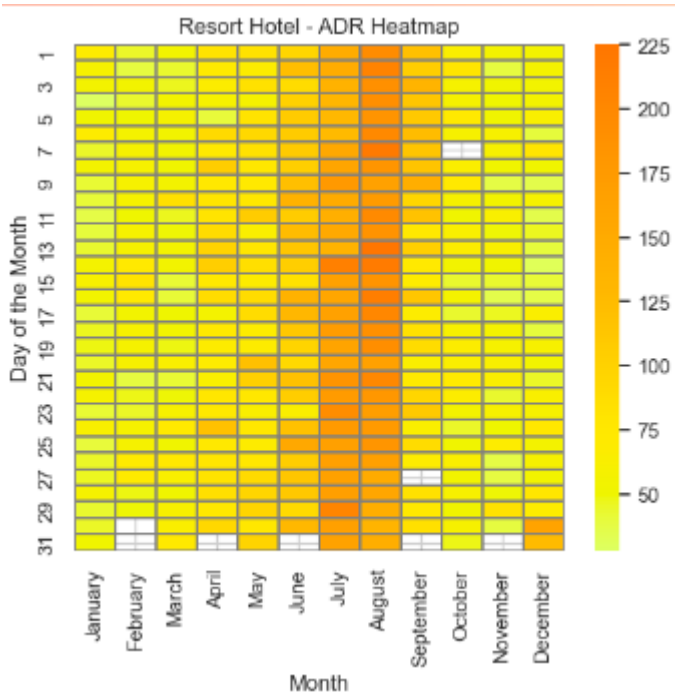


Fig - Seasonal heatmap visualization of Day of Month V/S Month w.r.t to days of the month (Resort Hotel)

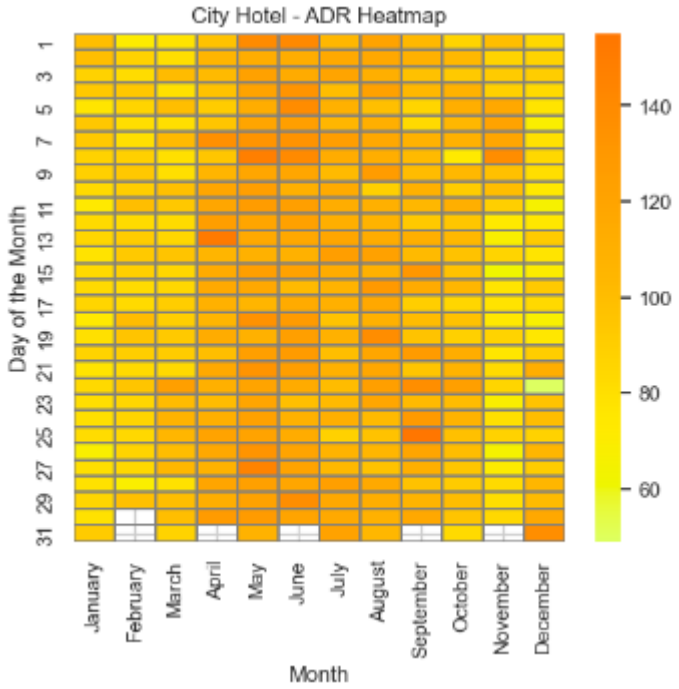


Fig - Seasonal heatmap visualization of Day of Month V/S Month w.r.t to days of the month (City Hotel)

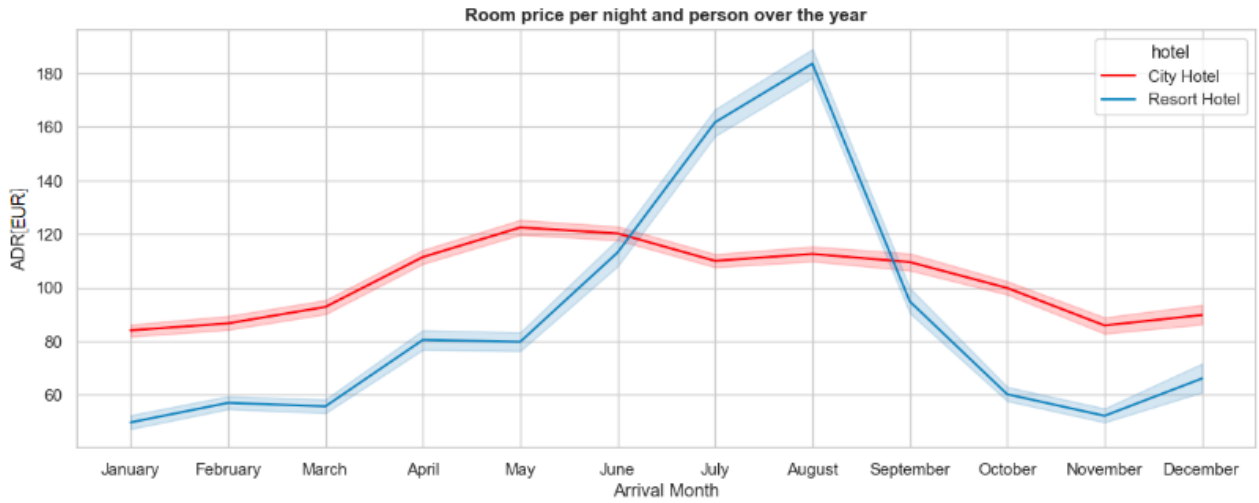


Fig - Seasonal Price Variation - Arrival Month V/S EUR

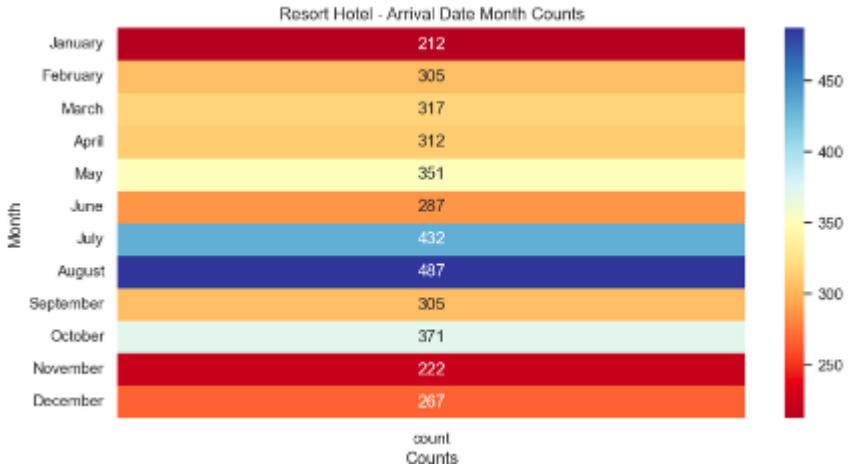


Fig - Count of Arrivals over the months throughout the years (Resort Hotel)

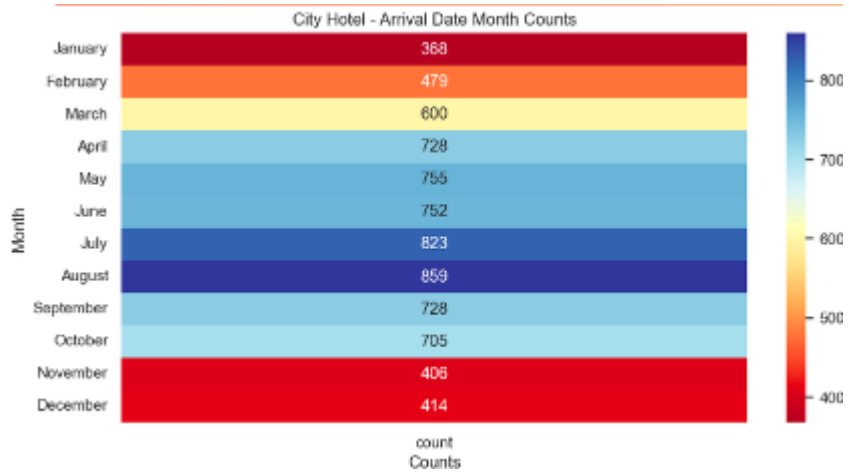


Fig - Count of Arrivals over the months throughout the years (City Hotel)

➢ **Classification and Regression Modelling:**

Classification:

- ➢ FEATURE SELECTION
- ➢ BUILDING THE MODEL
- ➢ EVALUATING THE MODEL & OPTIMIZING MODEL PERFORMANCE

Feature Selection:

Fig - Correlation Table (City Hotel)

In this correlation analysis, several noteworthy patterns emerge regarding the likelihood of booking cancellations. Strong positive correlations indicate that longer lead times, a higher number of special requests, and the need for parking spaces are associated with an increased probability of cancellations. Moderately positive correlations suggest that booking changes, a history of cancellations by guests, and specific company or travel agency associations contribute to a heightened cancellation risk. Additionally, weak positive correlations with factors like being a repeated guest, longer waiting times, and a history of previous bookings not being canceled provide subtle indicators of increased cancellation likelihood. On the contrary, weak negative correlations hint at a slight decrease in cancellation probability concerning the number of weeknights stayed and the month of arrival. Notably, some features, such as arrival date week number, year, children, and arrival date day of the month, exhibit negligible correlations with cancellations. While these correlations offer valuable insights, it's essential to approach the findings cautiously, recognizing that correlation does not imply causation. Further analyses, such as regression modeling or assessing feature importance, are recommended for a more comprehensive understanding of the factors influencing booking cancellations.

Fig - Correlation Table (Resort Hotel)

In this comprehensive correlation analysis, key insights emerge regarding the factors influencing the likelihood of booking cancellations. Strong positive correlations underscore the significance of certain features, such as a higher number of special requests, longer lead times, and a history of previous cancellations, all of which exhibit robust associations with increased cancellation probabilities. Moderately positive correlations highlight the influence of factors like extended waiting times, the status of being a repeated guest, and the number of previous bookings not canceled. Meanwhile, weak positive correlations suggest that subtle effects are present, including the number of adults, the average daily rate, and the presence of children or babies. On the contrary, weak negative correlations indicate minor mitigating influences, such as the year of arrival. Additionally, negligible correlations are observed for features like the day of the month of arrival. It is crucial to approach these correlations with caution, recognizing that correlation does not imply causation. Therefore, further analyses, such as regression modeling or feature importance assessments, are recommended to deepen our understanding of the intricate dynamics contributing to booking cancellations.

IMPORTANCE OF ENCODING:

In our pursuit of refining predictive accuracy, we implemented a strategic feature engineering approach, with a particular focus on encoding the 'arrival_date_month' column. This transformation involved converting textual representations of months into a numerical format, offering our machine learning model the ability to discern temporal patterns in booking cancellations. By incorporating this feature, we aimed to capture seasonal variations, account for special events influencing booking behavior, and understand the temporal dynamics of cancellations. We explored both one-hot encoding and label encoding techniques and seamlessly integrated the encoded feature into our model training process. Through rigorous evaluation and validation, we assessed the impact on accuracy, considering metrics such as precision, recall, and F1 score. This feature engineering initiative represents a deliberate step toward improving our model's predictive capabilities, aligning with our commitment to staying attuned

to the nuanced temporal aspects inherent in booking data.

BUILDING THE MODEL:

Problem Statement Overview:

In the dynamic landscape of the hotel industry, where optimizing decision-making processes is paramount, our analytical journey is centered around developing predictive models. This endeavor is particularly focused on addressing two pivotal facets crucial for the industry's sustainable growth:

Cancellation Prediction: Our primary goal is to construct a predictive model that discerns whether a customer is likely to cancel their booking. This predictive capability is instrumental in empowering hotels to fine-tune resource management, staffing, and overall operational efficiency.

Model Selection Rationale:

Logistic Regression:

- Applicability to Binary Classification: Logistic Regression emerges as a cornerstone, custom-tailored for the binary classification challenge inherent in predicting booking cancellations (1 for cancellation, 0 for no cancellation).
- Interpretability: Beyond its predictive prowess, the transparent nature of Logistic Regression yields easily interpretable results. This interpretability is paramount, offering a profound understanding of each feature's impact on the likelihood of cancellation—a vital consideration in the nuanced realm of the hotel industry.
- Efficiency: Logistic Regression stands as a computational workhorse, showcasing efficiency and robust performance even when grappling with the substantial scale of hotel booking datasets.

Decision Tree:

- Non-linearity and Feature Importance: Decision Trees, with their intrinsic ability to capture non-linear relationships, prove invaluable in unraveling the intricate patterns within customer behavior that may influence cancellations.
- Feature Importance: Similar to a detective unveiling clues, Decision Trees provide insights into the relative importance of features. This feature-centric knowledge aids in identifying the critical factors steering booking outcomes.
- Intuitive Decision-Making: The transparent and intuitive decision-making process facilitated by Decision Trees empowers hotel management to decipher the decision rules and comprehend the factors steering predictive outcomes.

Random Forest Classifier:

- Ensemble Learning: As a symphony of Decision Trees, the Random Forest Classifier harnesses the collective strength of multiple models, elevating predictive accuracy and robustness to new heights.
- Handling Overfitting: Mitigating the overfitting conundrum, a common affliction of individual Decision Trees, the Random Forest approach judiciously aggregates predictions from diverse

trees.

- Improved Generalization: Through the amalgamation of multiple weak learners (trees), the Random Forest Classifier not only mitigates overfitting but also bestows superior generalization prowess. This translates into enhanced predictive capabilities, particularly when confronted with unseen data.

The strategic selection of Logistic Regression, Decision Tree, and Random Forest Classifier is a meticulous alignment with the intricate nuances of the hotel booking problem statement. Each model contributes a unique set of strengths, encompassing interpretability, non-linear pattern recognition, and the ensemble advantage of Random Forest. This holistic modeling strategy positions itself as a beacon of actionable insights for hotel management, poised to significantly augment the sophistication of decision-making processes within the industry.

EVALUATING THE MODEL & OPTIMIZING MODEL PERFORMANCE:

CITY HOTEL:

Model Performance Evaluation: Logistic Regression:

Accuracy and Confusion Matrix Analysis:

The Logistic Regression model demonstrated an accuracy score of 70.76%, indicating its overall effectiveness. A detailed examination of the confusion matrix revealed that 507 instances were accurately predicted as class 0 (no cancellation), while 248 instances were correctly classified as class 1 (cancellation). However, the model misclassified 178 instances as class 0 and 134 instances as class 1, emphasizing the need for a closer scrutiny of false positives and false negatives.

Precision, Recall, and F1-Score Metrics:

Precision, a measure of the accuracy of positive predictions, yielded values of 0.74 for class 0 and 0.65 for class 1. The recall metric, indicating the model's ability to capture true positives, stood at 0.79 for class 0 and 0.58 for class 1. The f1-score, a harmonic mean of precision and recall, provided insights into the balanced performance of the model, with values of 0.76 for class 0 and 0.61 for class 1. These metrics collectively offered a nuanced understanding of the model's discriminatory capabilities.

Macro and Weighted Average Metrics:

The macro and weighted average metrics, both at 0.69, provided a holistic assessment across classes, underscoring the balanced evaluation of the model. This comprehensive approach ensures that the model's performance is scrutinized without bias towards any particular class, offering valuable insights into its overall robustness.

Model Performance Evaluation: Decision Tree Classifier:

The detailed examination of classification reports for selected and non-selected features sheds light on the discriminative power and robustness of the model across diverse feature subsets.

For Selected Features:

The model exhibits commendable proficiency when trained exclusively on the selected features. Achieving an overall accuracy of 71%, the model excels in differentiating between class 0 (no cancellation) and class 1 (cancellation) instances. Delving deeper into class-specific metrics, precision for class 0 stands at 0.70, reflecting the model's accuracy in predicting non-cancellations. The recall for class 0 impressively reaches 0.83, indicating the model's ability to capture a substantial proportion of actual non-cancellation instances. The harmonized f1-score for class 0, registering at 0.76, underscores the balance between precision and recall. Moving to class 1, the model maintains a commendable precision of 0.73, highlighting its accuracy in predicting cancellations. The recall for class 1, at 0.56, signifies the model's capacity to capture a significant portion of actual cancellations, while the f1-score of 0.63 harmonizes these aspects. The macro and weighted average metrics, both resting at 0.71, emphasize the model's consistent and well-rounded performance across the diverse classes.

For Not Selected Features:

The classification report provides a comprehensive evaluation of the model's performance in predicting booking cancellations without selecting specific features. The precision values for both classes (0.70 for non-cancellations and 0.73 for cancellations) indicate a moderate level of accuracy in correctly identifying instances within each category. However, the recall values reveal imbalances, with a notably higher recall for non-cancellations (0.83) compared to cancellations (0.56). This suggests that while the model effectively minimizes false positives for non-cancellations, there is room for improvement in capturing all instances of cancellations, as indicated by a higher number of false negatives. The F1-scores, harmonizing precision and recall, further emphasize this trade-off. The overall accuracy of 71% indicates the proportion of correctly classified instances, providing a global assessment of model performance. Macro and weighted averages provide additional insights, with the macro-average F1-score at 0.69 and the weighted-average F1-score at 0.70. Understanding these metrics and their implications is crucial for refining the model, potentially through feature engineering or hyperparameter tuning, to achieve a more balanced and effective predictive capability for booking

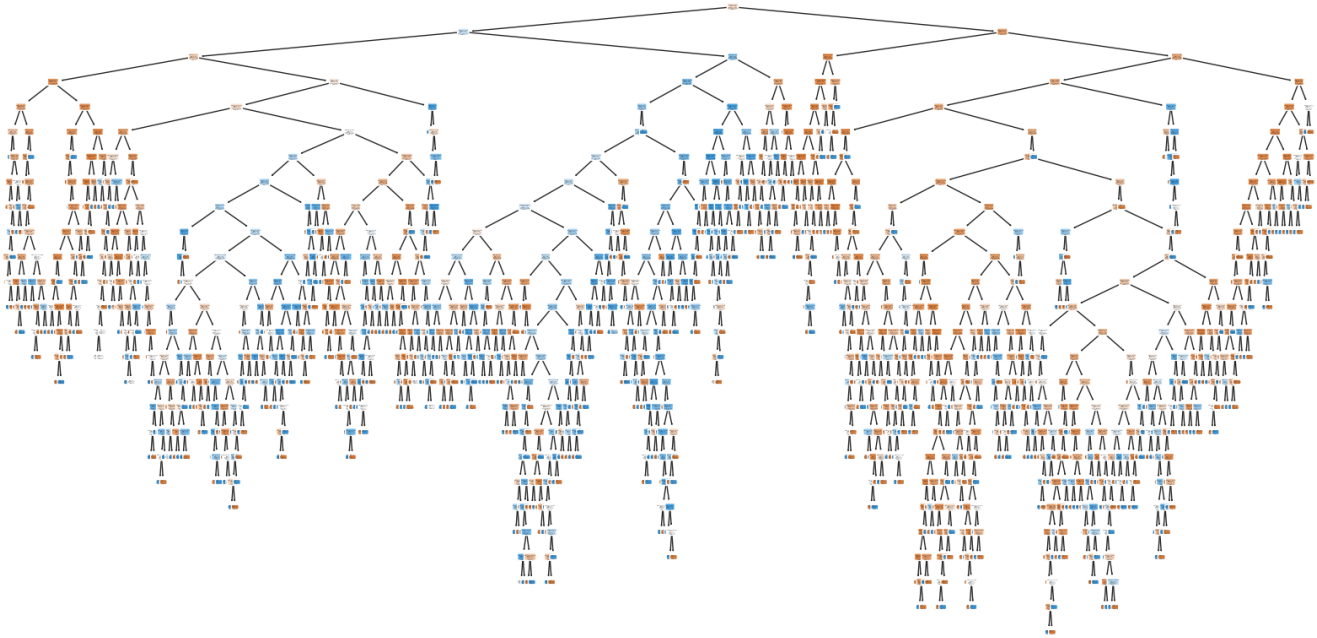


Fig - Decision Tree

Model Performance Evaluation: Random Forest Classifier:

The Random Forest Classifier (RFC) model demonstrates a robust and nuanced performance in predicting booking cancellations, showcasing its ability to effectively balance precision and recall. Precision is a crucial metric representing the accuracy of positive predictions, and the RFC model excels in this aspect, achieving high precision rates of 84% for cancellations (class 1) and 78% for non-cancellations (class 0). This implies that when the model anticipates a cancellation, it is correct 84% of the time, and likewise, when predicting a non-cancellation, it maintains a high accuracy of 78%. The slightly lower recall for cancellations (0.68) compared to non-cancellations (0.89) indicates that while the model adeptly identifies instances of non-cancellations, there is room for improvement in capturing a higher proportion of actual cancellations.

The F1-score, a metric that harmonizes precision and recall, reinforces the model's balanced performance. The scores of 0.75 for cancellations and 0.83 for non-cancellations suggest a harmonious compromise, affirming the model's effectiveness in managing the trade-off between avoiding false positives and capturing true positives. The overall accuracy of 80% reflects the model's success in making correct predictions across both classes.

Delving into the confusion matrix provides a more granular view of the model's performance. True positives (instances correctly predicted as cancellations) and true negatives (instances correctly predicted as non-cancellations) showcase the model's strengths, while false positives (instances incorrectly predicted as cancellations) and false negatives (instances incorrectly predicted as non-cancellations) highlight areas for improvement. This nuanced understanding of the model's strengths and areas for enhancement lays a solid foundation for iterative refinement. Future optimization efforts should focus on elevating the RFC model's predictive capacity, emphasizing both precision and recall to enhance its

ability to accurately identify instances of booking cancellations.

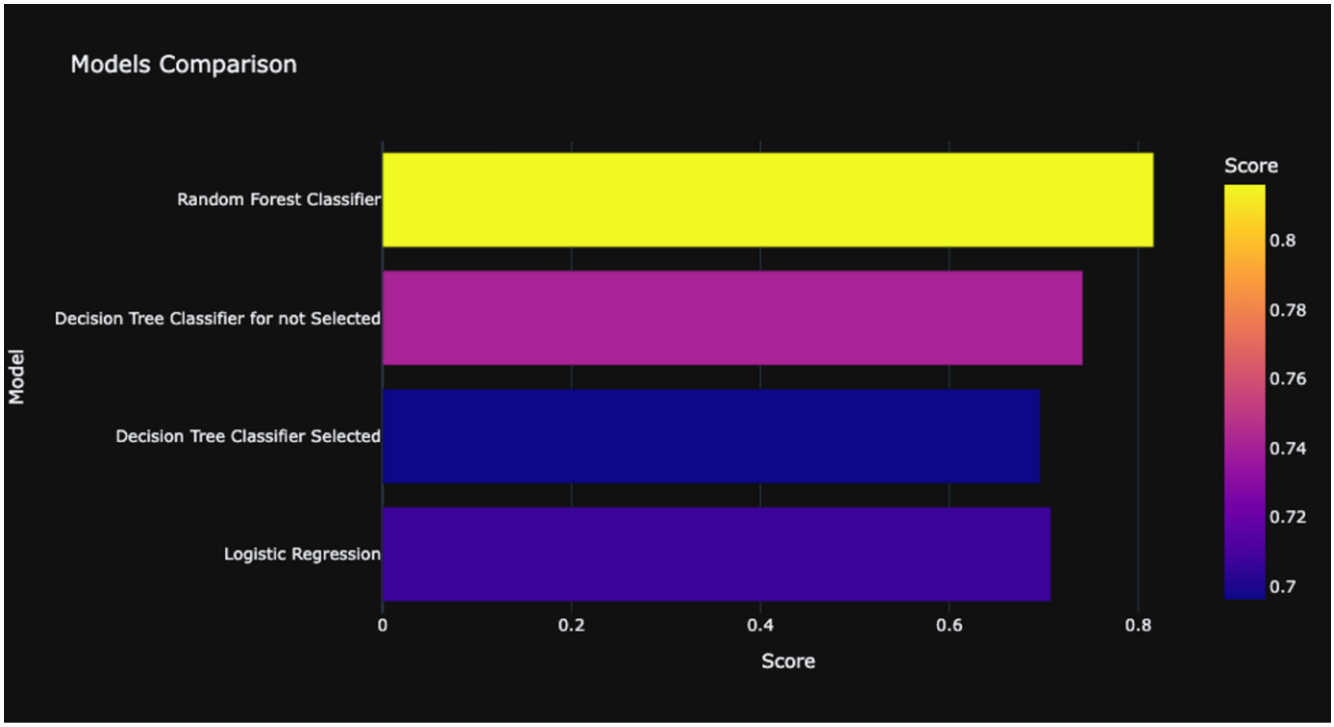


Fig - Model Comparison for City

Cross-validation for Logistic Regression and RandomForestClassifier:

The cross-validation scores provide valuable insights into the performance of two different machine learning models: Logistic Regression and RandomForestClassifier. Cross-validation is a crucial technique for assessing a model's generalizability by training and evaluating it on different subsets of the dataset.

Logistic Regression:

The cross-validation scores for Logistic Regression across five folds are [0.74794842, 0.73153576, 0.75029308, 0.74912075, 0.7370892]. These scores represent the accuracy achieved by the model in each fold. The mean cross-validation accuracy for Logistic Regression is computed as 74.32%. This indicates that, on average, the Logistic Regression model correctly predicted the target variable for approximately 74.32% of the instances during the cross-validation process.

RandomForestClassifier:

For the RandomForestClassifier, the cross-validation scores across the same five folds are [0.82415006, 0.82532239, 0.8042204, 0.8042204, 0.81807512]. The mean cross-validation accuracy for RandomForestClassifier is notably higher at 81.52%. This suggests that the RandomForestClassifier, on average, achieved an accuracy of 81.52% across the different folds, demonstrating a more robust

performance compared to Logistic Regression.

The higher mean cross-validation accuracy for the RandomForestClassifier implies that, in this specific context, the RandomForestClassifier model exhibits better predictive performance compared to Logistic Regression. It's essential to consider the balance between bias and variance when interpreting these results. A higher accuracy suggests that the RandomForestClassifier may capture more complex relationships in the data, making it a potentially more suitable choice for this particular prediction task.

These cross-validation results serve as a guide for model selection and provide a basis for understanding the expected performance of each model on unseen data. However, it's crucial to consider other factors such as interpretability, computational efficiency, and the specific requirements of the problem when deciding on the most suitable model for deployment. Further model tuning and hyperparameter optimization could potentially enhance the performance of both models.

Receiver Operating Characteristic (ROC) Curve for Logistic Regression:

The Receiver Operating Characteristic (ROC) curve and the associated Area Under the Curve (AUC) are crucial metrics for evaluating the performance of binary classification models, such as Logistic Regression. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at various classification thresholds.

In the case of Logistic Regression, the AUC is reported as 0.78. The AUC is a scalar value that quantifies the overall discriminatory power of the model across different threshold settings. A higher AUC indicates better discrimination, with a value of 1.0 representing a perfect classifier and 0.5 indicating a model that performs no better than random chance.

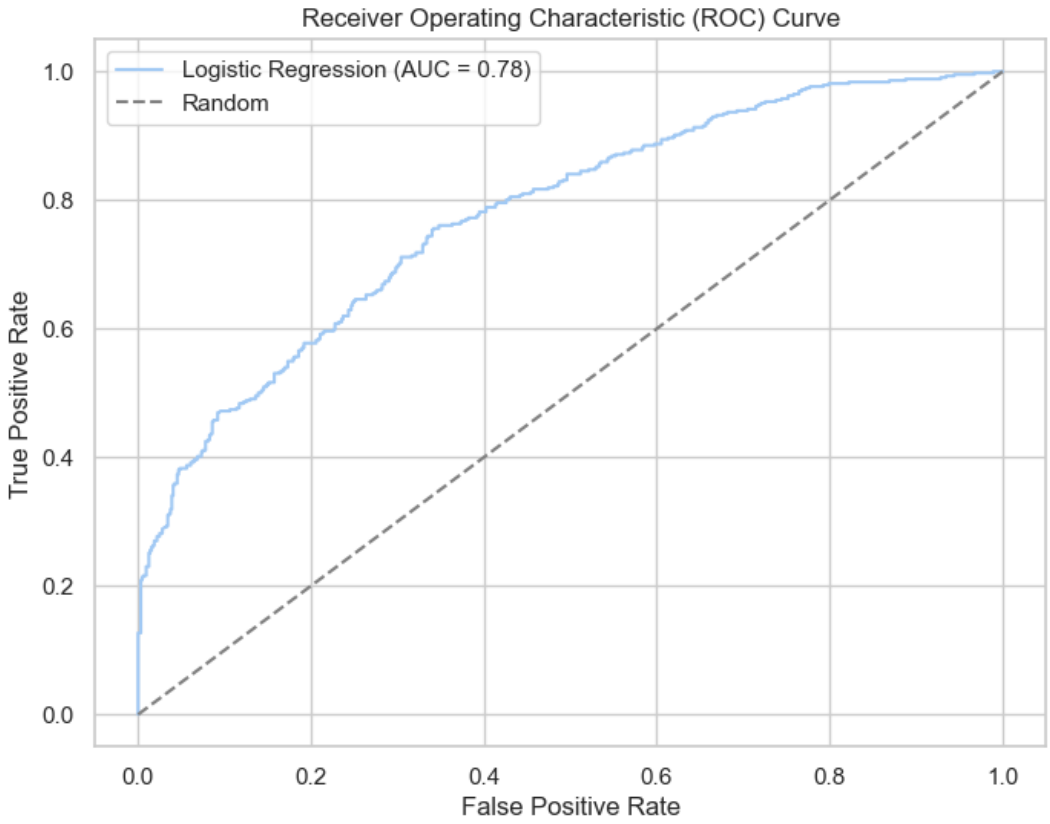


Fig - ROC curve

Hyperparameter Tuning using GridSearchCV for DecisionTreeClassifier:

Decision Tree Classifier - Before Hyperparameter Tuning (Without Feature Selection):

Before hyperparameter tuning and feature selection, the Decision Tree Classifier achieved an accuracy score of 71%. This baseline accuracy represents the model's performance on the dataset without any specialized adjustments. The initial model might exhibit some degree of overfitting or suboptimal parameter settings, leading to moderate accuracy. The goal of hyperparameter tuning is to systematically search through different parameter combinations to identify the configuration that maximizes the model's predictive performance. The starting point of 71% accuracy provides a benchmark against which improvements from hyperparameter tuning can be measured.

Decision Tree Classifier - After Hyperparameter Tuning (Without Feature Selection):

Following the hyperparameter tuning process for the Decision Tree Classifier without feature selection, the accuracy significantly increased to 79.15%. This improvement suggests that the fine-tuning of hyperparameters has led to a more optimized model configuration, resulting in better predictive accuracy. The tuning process might have addressed overfitting issues, enhanced the model's ability to generalize to unseen data, and improved its overall performance on the given dataset.

Decision Tree Classifier - Before Hyperparameter Tuning (With Feature Selection):

Before hyperparameter tuning but with the additional step of feature selection, the Decision Tree Classifier achieved an accuracy score of 70.76%. Feature selection involves identifying and using only the most relevant features for model training, potentially reducing overfitting and enhancing interpretability. The initial accuracy of 70.76% serves as a reference point for evaluating the impact of both feature selection and hyperparameter tuning.

Decision Tree Classifier - After Hyperparameter Tuning (With Feature Selection):

Upon completion of hyperparameter tuning with feature selection, the accuracy slightly decreased to 73.25%. While the accuracy is lower compared to the model without feature selection after tuning, it is essential to consider the trade-off between accuracy and the interpretability of the model. Feature selection might have resulted in a simplified model, potentially sacrificing a small amount of accuracy for a more concise and interpretable set of features. The accuracy of 73.25% reflects the model's performance after considering both hyperparameter tuning and feature selection.

RESORT HOTEL:

Model Performance Evaluation: Logistic Regression:

Accuracy and Confusion Matrix Analysis:

The Logistic Regression model achieved an accuracy score of 72.51%, indicating its overall success in correctly classifying instances into non-cancellation (class 0) and cancellation (class 1) categories. The confusion matrix provides a detailed breakdown: 352 True Negatives (accurate non-cancellations), 31 False Positives (instances incorrectly predicted as cancellations), 118 False Negatives (instances incorrectly predicted as non-cancellations), and 41 True Positives (accurate cancellations). This matrix offers insights into the model's strengths and areas for improvement, highlighting the trade-off between false positives and false negatives.

Precision, Recall, and F1-Score Metrics:

- Non-cancellations (Class 0):
  - Precision: 0.75 (75% of predicted non-cancellations are correct)
  - Recall: 0.92 (92% of actual non-cancellations are correctly identified)
  - F1-Score: 0.83 (harmonizing precision and recall)
- Cancellations (Class 1):
  - Precision: 0.57 (57% of predicted cancellations are correct)
  - Recall: 0.26 (26% of actual cancellations are correctly identified)
  - F1-Score: 0.35 (balancing precision and recall)

These metrics provide a detailed understanding of the model's performance for each class, emphasizing the challenges in correctly identifying cancellations.

Macro and Weighted Average Metrics:

- Macro-Average:
  - Macro-Average Precision: 0.66
  - Macro-Average Recall: 0.59
  - Macro-Average F1-Score: 0.59
- Weighted Average:
  - Weighted Average Precision: 0.70
  - Weighted Average Recall: 0.73
  - Weighted Average F1-Score: 0.69

These global metrics offer aggregated assessments, considering the imbalances in class support. The macro-average F1-score indicates a balanced performance, while the weighted average provides an overall measure accounting for class distribution.

Model Performance Evaluation: Decision Tree Classifier:

For Selected Features:

The model with selected features reflects a robust performance, particularly in predicting non-cancellations (Class 0) where precision, recall, and F1-score are consistently high at 0.75, 0.79, and 0.77, respectively. The model also demonstrates commendable accuracy in predicting cancellations (Class 1), with a precision of 0.72 and a balanced F1-score of 0.70, indicating an effective trade-off between precision and recall. The global metrics further reinforce the model's overall success, achieving an accuracy of 74%, a macro-average F1-score of 0.73, and a weighted-average F1-score of 0.74. These results collectively underscore the efficacy of feature selection in enhancing the model's ability to discern between cancellations and non-cancellations.

For Not Selected Features:

The model without selected features reveals a different scenario. While the model excels in predicting non-cancellations, as evidenced by high precision, recall, and F1-score for Class 0, it faces challenges in accurately identifying cancellations (Class 1). The precision and recall for cancellations are notably lower at 0.51 and 0.49, respectively, resulting in a modest F1-score of 0.50. The global metrics echo these findings, with an accuracy of 71%, a macro-average F1-score of 0.65, and a weighted-average F1-score of 0.71. This indicates that without the guidance of selected features, the model's performance diminishes, particularly in capturing the nuances associated with cancellations. The comparison highlights the critical role of feature selection in optimizing the Decision Tree Classifier's predictive capabilities for the given dataset.

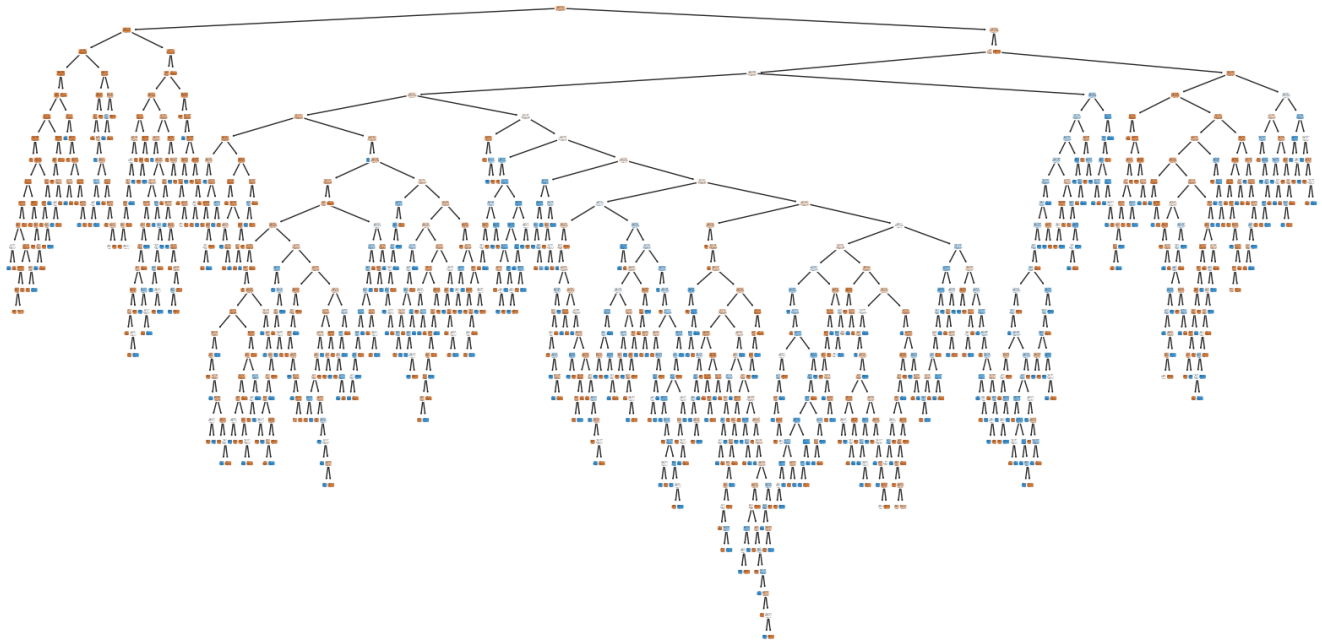


Fig - Decision Tree

Model Performance Evaluation: Random Forest Classifier:

The detailed classification report unveils a thorough examination of the Random Forest_Classifier's effectiveness in a binary classification scenario. Precision, denoting the accuracy of positive predictions, attains notable levels with 0.86 for non-cancellations (Class 0) and 0.59 for cancellations (Class 1). This implies that when the model predicts a non-cancellation, it is correct approximately 86% of the time, while predictions for cancellations are correct 59% of the time.

The model showcases a commendable ability to correctly identify instances of non-cancellations, as evidenced by a recall of 0.90. However, the recall for cancellations is comparatively lower at 0.50, indicating a challenge in capturing a substantial portion of actual cancellations within the predictions. The F1-Score, balancing precision and recall, reflects this trade-off, with a high value of 0.88 for non-cancellations and a lower value of 0.54 for cancellations.

Examining the support metrics, the dataset comprises 421 instances of non-cancellations and 121 instances of cancellations, providing context for the model's predictions. The overall accuracy of 81% signifies the proportion of correctly predicted instances across both classes, demonstrating the model's general effectiveness.

The macro and weighted averages offer a holistic evaluation, accounting for potential imbalances in class distribution. The macro-average F1-Score of 0.71 and the weighted-average F1-Score of 0.81 provide nuanced insights into the model's performance, considering both its ability to capture minority classes and maintain a balanced trade-off between precision and recall.

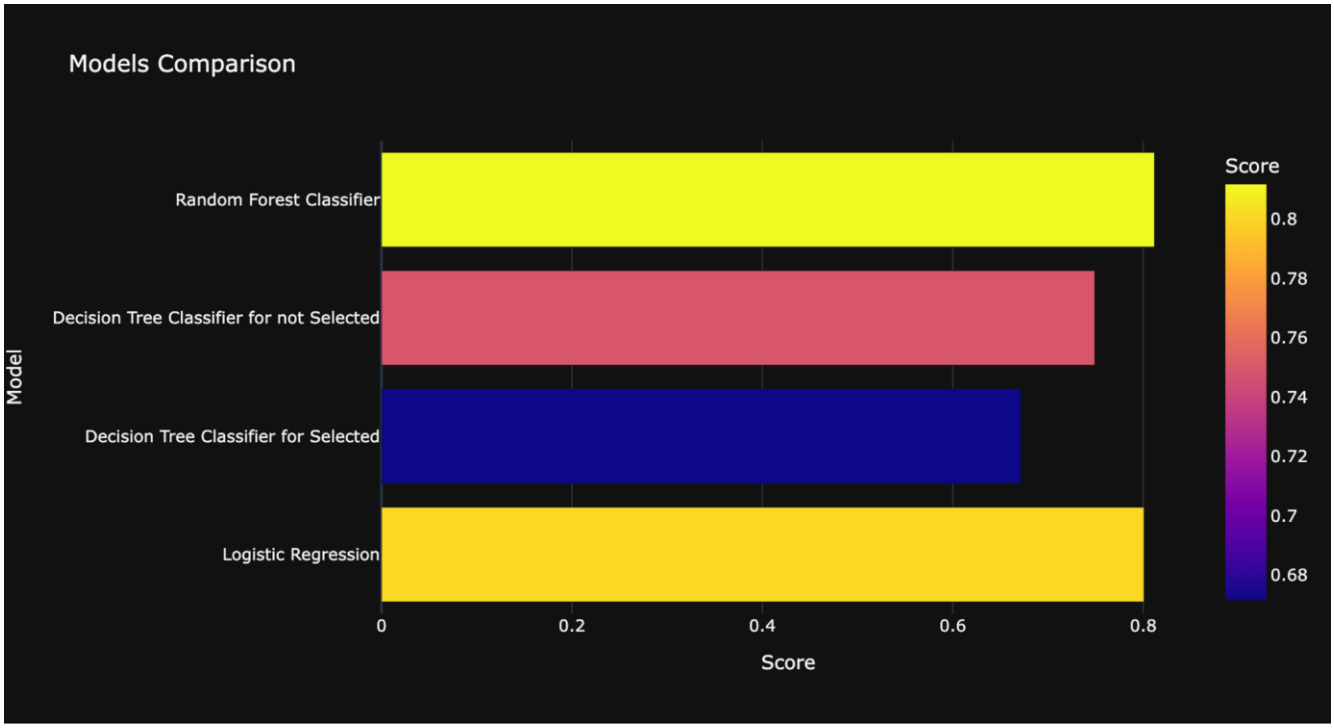


Fig - Model Comparison for Resort

Cross-validation for Logistic Regression and RandomForestClassifier:

The cross-validation scores provide insights into the performance of two different classification models—Logistic Regression and RandomForestClassifier—across multiple folds of the dataset. Cross-validation is a robust technique used to assess a model's generalizability by training and evaluating it on different subsets of the dataset.

Logistic Regression:

The cross-validation scores across five folds range from 0.727 to 0.767. The mean cross-validation accuracy is calculated to be 74.23%. This indicates that the Logistic Regression model consistently achieves accuracies around the 74% mark across different subsets of the data. The relatively narrow range of scores suggests stable performance, and the mean accuracy serves as a representative estimate of the model's overall performance.

RandomForest Classifier:

The RandomForest Classifier exhibits higher cross-validation scores, ranging from 0.792 to 0.843 across the five folds. The mean cross-validation accuracy for the RandomForestClassifier is notably higher at 81.62%. This suggests that the RandomForestClassifier consistently performs well across different subsets of the data, achieving accuracies around the 82% mark. The wider range of scores compared to Logistic Regression may indicate that RandomForestClassifier is more sensitive to variations in the training subsets.

Receiver Operating Characteristic (ROC) Curve for Logistic Regression:

The Receiver Operating Characteristic (ROC) curve with an area under the curve (AUC-ROC) score of 0.82 for the Logistic Regression model is indicative of its strong discriminatory ability in distinguishing between booking cancellations and non-cancellations. The ROC curve visually illustrates the trade-off between true positive rate and false positive rate at different probability thresholds. With an AUC-ROC score of 0.82, the model exhibits a high true positive rate while maintaining a low false positive rate, suggesting that it effectively ranks positive instances (cancellations) higher than negative instances (non-cancellations). This performance metric of 0.82 signifies a robust and reliable model, showcasing its proficiency in capturing the nuances of the underlying dataset and making well-informed predictions.

The AUC-ROC score of 0.82 is particularly noteworthy as it surpasses the random chance of 0.5, indicating that the Logistic Regression model significantly outperforms a random classifier. This level of discrimination is valuable in scenarios where accurately identifying cancellations is crucial. While the AUC-ROC score provides a comprehensive summary of the model's discriminatory power, further examination of precision, recall, and other metrics can offer a more nuanced understanding of its overall performance and any potential areas for refinement.

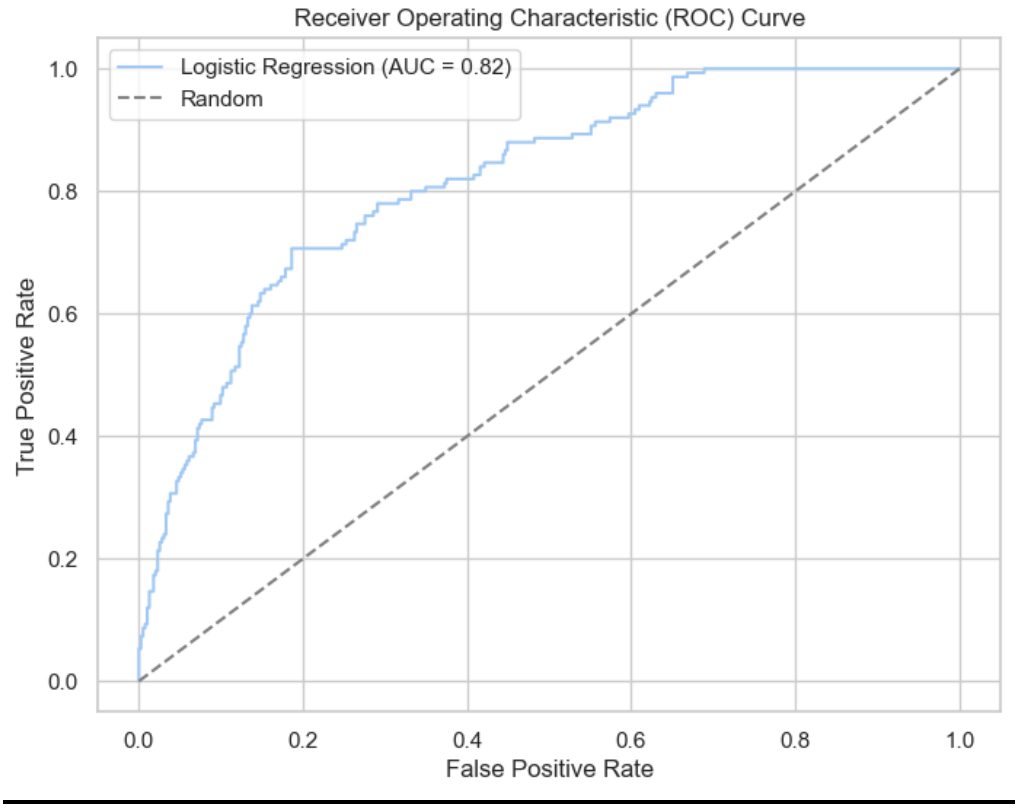


Fig - ROC curve

Hyperparameter Tuning using GridSearchCV for DecisionTreeClassifier:

Before Hyperparameter Tuning - Without Feature Selection:

Before hyperparameter tuning, the Decision Tree Classifier without feature selection achieved an accuracy score of 73.85%. At this stage, the model exhibited a reasonably accurate performance in predicting booking cancellations, but there was room for enhancement. The initial configuration of hyperparameters and the inclusion of all features in the model provided a solid foundation, but opportunities for refinement existed to further optimize its predictive capabilities.

After Hyperparameter Tuning - Without Feature Selection:

Following hyperparameter tuning using GridSearchCV, the Decision Tree Classifier experienced a noticeable improvement in accuracy, reaching 74.72%. This enhancement signifies that the fine-tuning of hyperparameters resulted in a more effective configuration, enabling the model to make better-informed decisions. The optimized hyperparameters likely contributed to a more nuanced decision-making process within the Decision Tree, ultimately leading to increased accuracy in predicting booking cancellations.

Before Hyperparameter Tuning - With Feature Selection:

Initially, before hyperparameter tuning, the Decision Tree Classifier with feature selection achieved an accuracy score of 70.76%. Feature selection involves choosing a subset of the most relevant features, aiming to simplify the model while preserving its predictive power. In this state, the model showed decent accuracy, but the potential impact of hyperparameter tuning had yet to be realized.

After Hyperparameter Tuning - With Feature Selection:

After hyperparameter tuning, the Decision Tree Classifier with feature selection maintained a solid accuracy score of 72.69%. Despite a slight decrease compared to the model without feature selection, this result underscores the robustness of the model even after reducing the number of features. The optimized hyperparameters likely facilitated a more efficient decision-making process, contributing to the model's ability to maintain a high level of accuracy even with a more streamlined set of features.

Regression:

#4 Predicting Average Daily Rate to improve sales:

- FEATURE ENGINEERING
- BUILDING THE MODEL
- EVALUATING THE MODEL
- OPTIMIZING MODEL PERFORMANCE

Feature Engineering:

We selected a set of 13 features from which were correlating linearly and were inferred to effect the Average Daily Rate. Our journey into feature engineering was guided by meticulous research and analysis, with a clear goal in mind: predicting the Average Daily Rate (ADR) of our product. Let me elaborate on the key aspects of our feature selection:

- Thorough Research: We initiated our feature engineering process with an extensive information search. This groundwork led us to identify 13 input variables that we believed would have a significant impact on ADR.
- Correlation and Impact Analysis: We didn't stop at just selecting these variables; we delved deeper into understanding their relationship with ADR through correlation analysis. This step was pivotal, as it ensured that each feature had a statistically significant connection with room rates. This, in turn, established their direct influence on revenue optimization strategies. In essence, our feature selection was data-driven, guaranteeing that we were focusing on the most pertinent aspects of our dataset.
- Our feature selection was a thoughtful process, offering a comprehensive view of booking dynamics. We considered various aspects, such as room types, guest composition, booking times, and special requests, ensuring that we captured both customer preferences and operational factors affecting pricing. This approach made our ADR prediction model robust. We also accounted for temporal factors by including features like arrival_date_year and arrival_date_month. These insights helped us adapt pricing strategies to seasonal trends and market changes over the years, crucial for staying responsive to market dynamics and customer behavior in the hospitality industry.

In summary, our feature engineering process was meticulous, data-driven, and comprehensive. It ensured that our ADR prediction model not only considered a wide range of variables but also adapted to temporal changes, making it a powerful tool for revenue optimization in the ever-evolving hospitality sector.

Visualizing Seasonal Trends for Average Daily Rate (ADR):

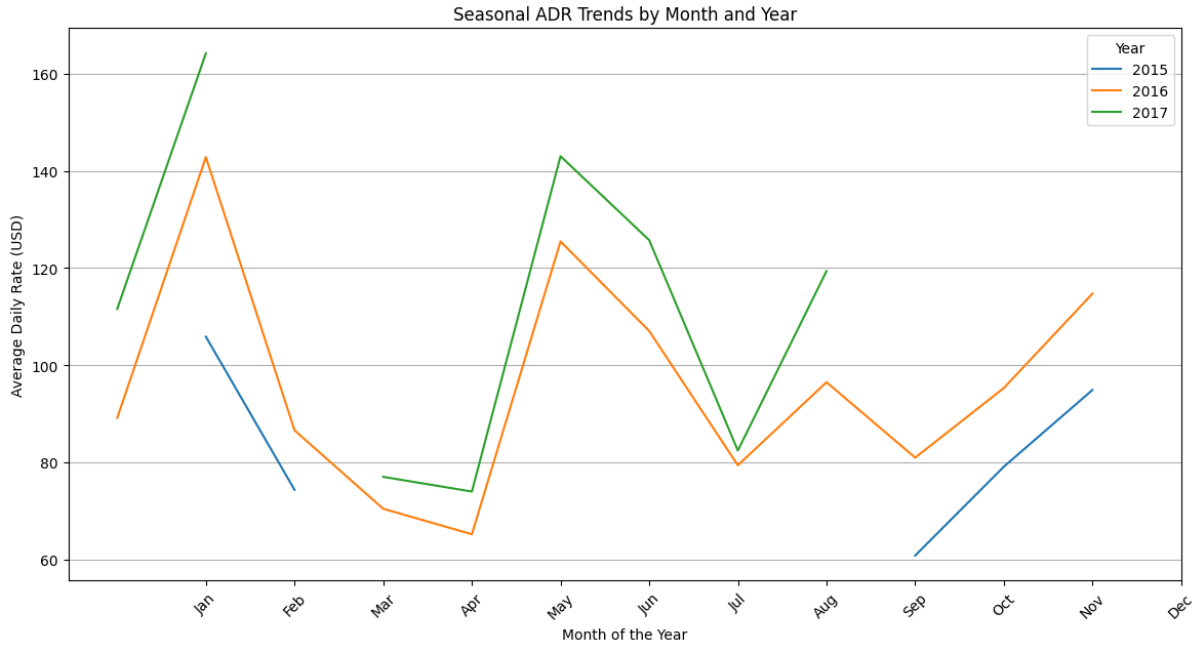


Fig - Price variation throughout the year for 2015, 2016 and 2017

Correlation Matrix Visualisation:

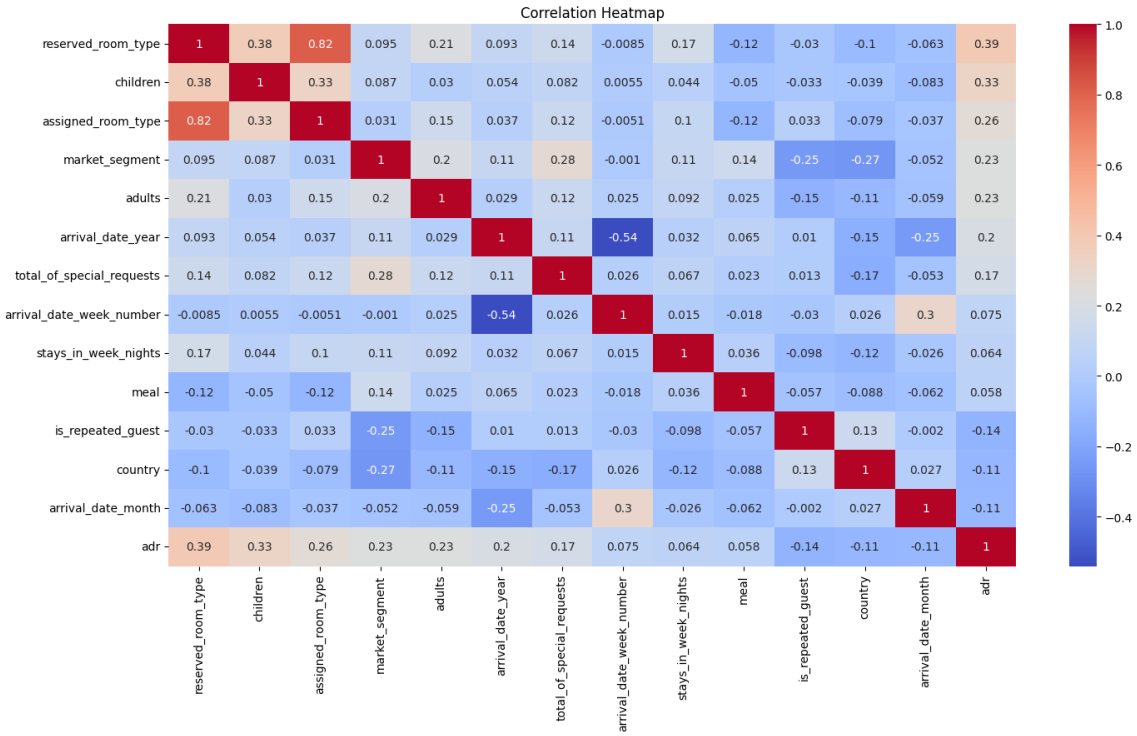


Fig- This Above image is a representation of the correlation matrix, with which we were able to choose the input variables that highly correlated with the 'adr'.

Summary of the Matrix:

- The correlation matrix provided an invaluable quantitative analysis to ascertain the linear relationships between the selected predictors and the Average Daily Rate (ADR).
- Upon close inspection, features such as 'reserved_room_type' and 'children' emerged with substantial positive correlations, suggesting a strong direct influence on the ADR.
- Conversely, 'country' and 'arrival_date_month' exhibited negative correlations, indicating an inverse relationship. Notably, temporal features like 'arrival_date_year' and 'arrival_date_week_number' presented a nuanced view, reflecting the impact of seasonal and annual market trends on pricing.
- This matrix served as a strategic tool, enabling the prioritization of variables that hold the most significant sway over room pricing, thus reinforcing the foundations for our predictive model.

Visualizing ADR against Input Variables:

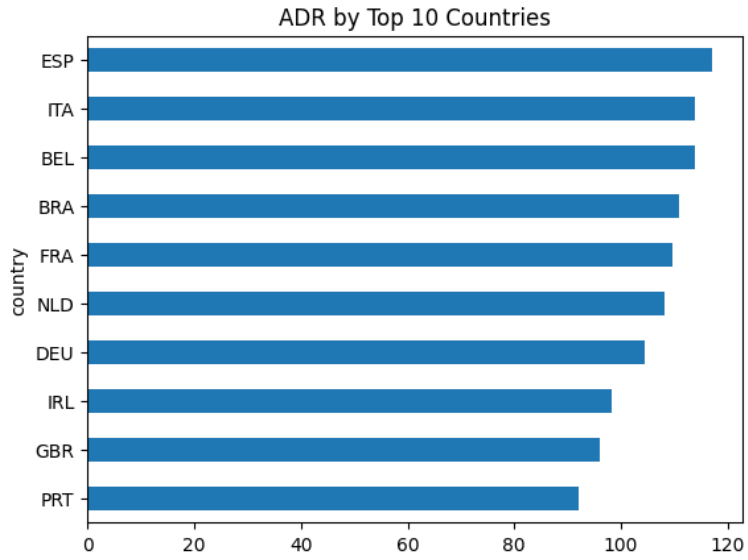


Fig - Tourists traveling from these Top 10 countries and their expenditure on the rooms.

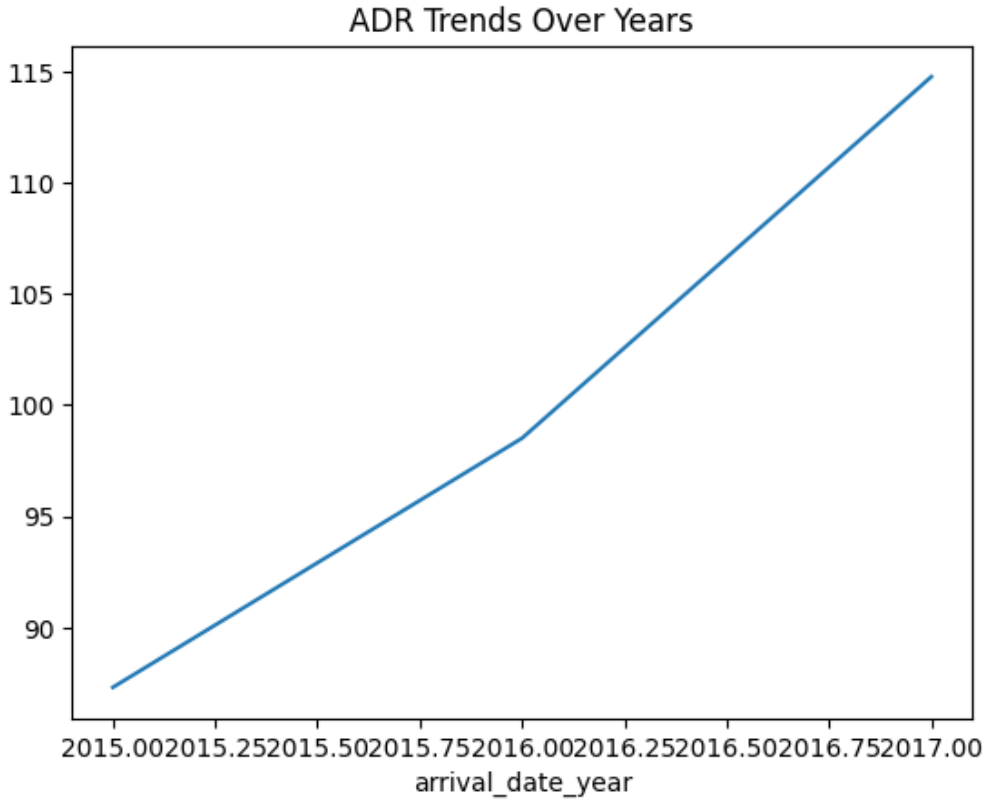


Fig - Seasonal Trend on increase in prices over the years wrt to Average Daily Rate

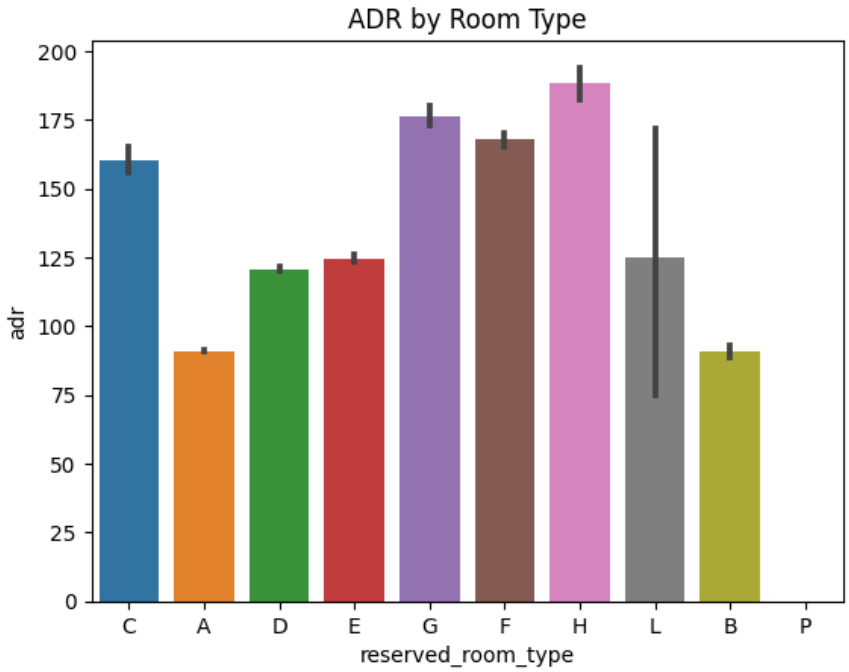


Fig - Reserved Room type's V/S adr

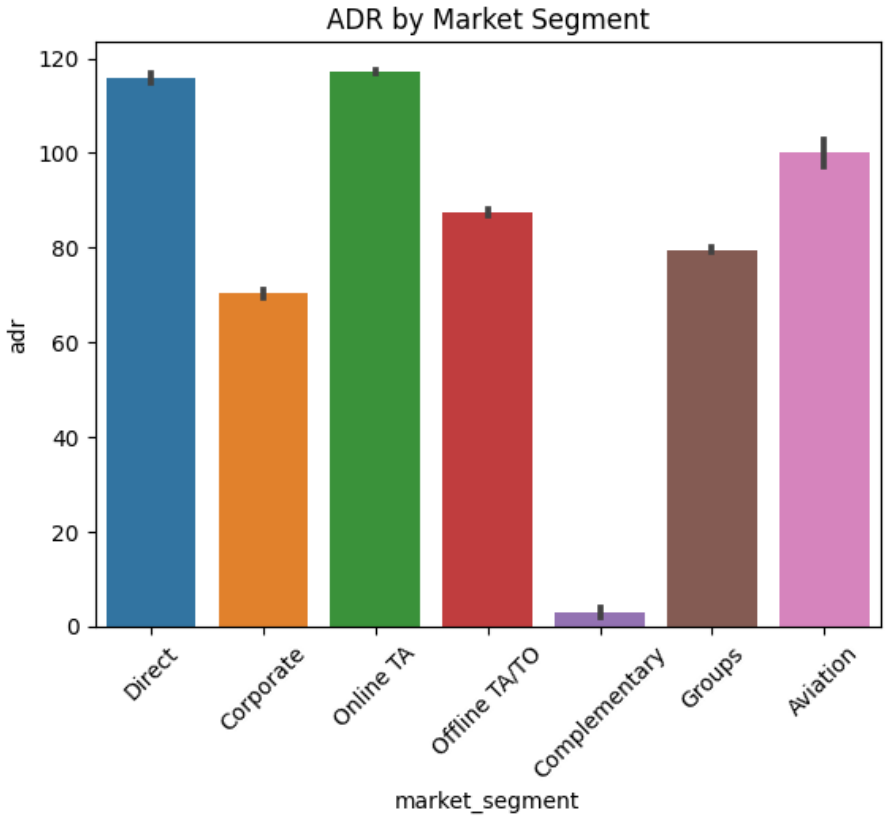


Fig - Market segment of bookings V/S adr

Encoding the Data:

- In this step, we transformed categorical data into a numerical format using Label Encoding, assigning a unique integer to each category within the features, which is necessary for the machine learning algorithms to process the input data effectively.
- This encoding method maintains the categorical nature of the data while preparing it for the correlation analysis and subsequent modeling.
- In this crucial step, we harnessed the power of encoding to seamlessly bridge the gap between categorical data and the numerical realm. Our weapon of choice was Label Encoding, a technique that bestowed each category within our features with a unique integer identity.
- This transformation is pivotal, for it equips machine learning algorithms with the ability to process the input data effectively. But why Label Encoding, you may ask? Well, it serves as the bridge that preserves the categorical essence of our data, allowing us to smoothly transition into correlation analysis and the intricate world of modeling.
- By retaining the categorical nature of our data while making it machine-friendly, we pave the way for robust analysis and modeling prowess.

Building Regression Models to predict ADR:

Model	Mean Squared Error (MSE)	R-squared
Linear Regression	1513.011	0.3318
Ridge Regression	1513.011	0.3318
Lasso Regression	1533.776	0.3226
Decision Tree	1018.325	0.5502
Gradient Boosting	897.693	0.6035
XGBoost	645.868	0.7147

Detailed Review of the Models:

Linear Regression:
MSE: 1513.01
R-squared: 0.33
Interpretation: Provides a baseline performance with a moderate R-squared value.

Ridge Regression:
MSE: 1513.01
R-squared: 0.33
Interpretation: Similar performance to Linear Regression, indicating that regularisation had little impact.

Lasso Regression:
MSE: 1533.78
R-squared: 0.32
Interpretation: Slightly worse than Linear and Ridge, potentially due to feature selection reducing model complexity.

Decision Tree Regressor:
MSE: 1005.11
R-squared: 0.56
Interpretation: Better performance than linear models, indicating that the data might have non-linear patterns.

Gradient Boosting Regressor:
MSE: 897.69
R-squared: 0.60
Interpretation: Further improvement, suggesting that ensemble methods are more effective for this dataset.

XGBoost Regressor:
MSE: 645.87
R-squared: 0.71
Interpretation: The best performance among the tested models, indicating a strong fit to the dataset.

Key Takeaways:

- ➢ The XGBoost Regressor outperforms other models in terms of both MSE and R-squared, making it the most effective model for this particular task.
- ➢ The improvement in performance from linear models to tree-based models (especially ensemble methods like Gradient Boosting and XGBoost) suggests that the relationships in your data are complex and not purely linear.
- ➢ The high R-squared value for XGBoost indicates that it captures a significant portion of the variance in the ADR, making it a reliable choice for predictions.

Next Steps:

- ➢ Will perform hyperparameter tuning for XGBoost, as fine-tuning can potentially lead to even better results.
- ➢ Cross-validation can also be used to ensure the model's stability and generalizability.
- ➢ Will also be Investigating feature importance in the XGBoost model can provide insights into which factors most heavily influence ADR.

Hyperparameter tuning with GridSearchCV

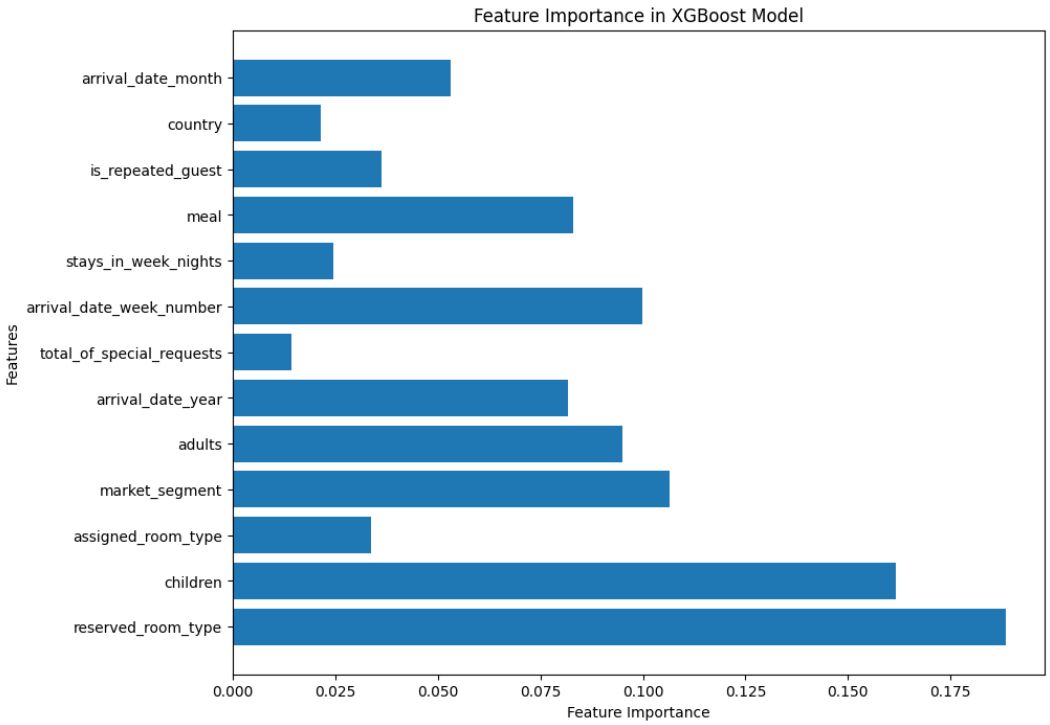
- ➢ Executed a comprehensive grid search cross-validation with 81 unique combinations across 3 folds, totaling 243 fits to identify the optimal model settings.

Optimal Model Parameters Identified:

- ➢ Learning Rate: 0.2
- ➢ Max Depth: 7
- ➢ Number of Estimators: 200
- ➢ Subsample: 0.9

Next we performed some Cross-Validation with the best parameters from the grid search:

- ➢ Cross-Validation Scores: [0.52737422 0.61701214 0.18137931 0.43330826 0.40936078]
- ➢ Mean R-squared: 0.4336869425512229
- ➢ Standard Deviation: 0.14603868246303464



Visualizing Feature Importance:

- From the bar chart, it appears that reserved_room_type and children are among the most important features in predicting ADR, followed by assigned_room_type, market_segment, and adults. This aligns with intuitive understanding—room types and the number of occupants can significantly influence the room rate.
- Key Observations: The feature importance chart is a critical tool for understanding which features most strongly influence the ADR. It can inform decision-making regarding pricing strategies and targeted marketing. The difference in feature importance suggests that focusing on room types and catering to specific market segments or family compositions could be effective strategies for revenue optimization.

Observations and Insights of Feature Importance:

Throughout this process, we've gained several insights:

- **Non-linear models**, particularly ensemble methods like XGBoost, were more effective for this dataset, indicating complex patterns in the data.
- Room types and guest composition (adults, children) are significant predictors of ADR.
- **Hyperparameter tuning and cross-validation** are essential steps to optimize model performance and ensure robustness.
- **The variability in cross-validation** scores suggests potential overfitting or the presence of outliers, which would require further investigation.

Future Scope for Regression:

Possible next steps could involve **more detailed hyperparameter tuning**, alternative feature selection methods, outlier analysis, or the use of more complex models such as neural networks if computational resources permit.

What we Learnt:

- To check the validity and source of the data.
- Understand the data with respect to each instance without any tools
- Establish and infer a pattern from basic EDA
- Delve into depth EDA after pattern recognition

Tech Stack:

- Pandas
- Numpy
- Matplotlib
- Numpy
- Polars
- Scikit Learn
- Plotly
- Stremlit

Acknowledgements:

The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

References:

1) https://www.sciencedirect.com/science/article/pii/S2352340918315191