Optimization and Evaluation of Booking Prediction Models - INN Hotels

Course Name: Supervised Learning

Kingsley Uchenna Azinna Ukwu

Date: 14th June, 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

INN Hotels Group is experiencing a significant number of booking cancellations and no-shows, leading to revenue loss, increased operational costs, and lower profit margins. These issues are exacerbated by modern online booking channels, which have changed customer behavior, making it easier to cancel bookings at little or no cost. To mitigate these challenges, a machine learning-based solution is needed to predict which bookings are likely to be canceled, allowing the hotel to implement more effective and profitable policies for cancellations and refunds.

The goal is to develop a predictive model to identify bookings that are likely to be canceled. This will enable INN Hotels Group to take proactive measures to reduce cancellations, optimize room occupancy, and improve overall profitability.The dataset provided includes various attributes related to customer booking details, such as the number of adults and children, stay duration, meal plan type, car parking requirements, room type, lead time, arrival details, market segment, repeat guest status, previous cancellations, average price per room, special requests, and booking status.

Key Findings and Actionable Insights:

1. High Lead Time Increases Cancellation Risk-Bookings with longer lead times are more prone to cancellations. Implementing stricter cancellation policies or requiring a non-refundable deposit for bookings made well in advance can help mitigate this risk.

2. Impact of Previous Cancellations- Customers with a history of previous cancellations are more likely to cancel again. Offering incentives for these customers to complete their stays, such as discounts or loyalty points, may reduce cancellation rates.

Executive Summary

3. Repeat Guests- Repeat guests are less likely to cancel their bookings. Encouraging loyalty programs and providing personalized offers can enhance customer retention and reduce cancellations.

4. Booking Period- Certain months may have higher cancellation rates. Analyzing trends by month can help in adjusting marketing strategies and cancellation policies during peak cancellation periods.

5. Special Requests- Customers with special requests might have higher expectations and a lower tolerance for changes. Ensuring these requests are fulfilled or offering alternatives can improve customer satisfaction and reduce cancellations.

6. Meal Plans and Room Types- Different meal plans and room types may have varying cancellation rates. Tailoring offers and policies based on the meal plan and room type preferences can optimize booking confirmations.

7. Market Segment- Market segments (e.g., business vs. leisure travelers) show different behaviors. Customizing cancellation policies and offers according to market segments can help in reducing cancellations.

Predictive Model by way of a decision tree classifier was used to predict booking cancellations, considering various features such as the number of adults and children, stay duration, lead time, previous cancellations, average price per room, special requests, and booking status. The model was trained and evaluated to ensure its effectiveness in predicting cancellations.

Executive Summary

Next Steps:

- Implement Predictive Model: Deploy the predictive model in the booking system to flag potential cancellations. This will enable the hotel to take proactive measures to secure bookings and minimize cancellations.
- Revise Cancellation Policies: Based on the model's insights, revise the cancellation policies to be stricter for high-risk bookings, such as those with long lead times or multiple previous cancellations.
- Customer Engagement: Increase engagement with customers who are flagged as potential cancellers through personalized communication and offers to encourage them to honor their bookings.

Monitor and Adjust: Continuously monitor the model's performance and adjust features, policies, and strategies as needed to adapt to changing customer behaviors and market conditions.

By implementing these actionable insights and leveraging the predictive model, INN Hotels Group can significantly reduce booking cancellations, optimize occupancy, and enhance overall profitability.

Business Problem Overview and Solution Approach

INN Hotels Group is facing a significant challenge with a high number of booking cancellations and no-shows. These cancellations are primarily driven by flexible cancellation policies and modern online booking systems, which allow customers to cancel bookings easily and at low or no cost. This situation results in considerable revenue loss, increased operational costs, and reduced profit margins, particularly with last-minute cancellations.

Impacts of Booking Cancellations

i. Revenue Loss: Unsold rooms lead to direct revenue loss.

ii. Increased Distribution Costs: More spending on commissions and advertising to resell canceled rooms.

iii. Lowered Profit Margins: Often requiring price reductions to resell rooms at the last minute.

iv. Additional Human Resources: Extra effort needed to manage and accommodate guest arrangements.

v. Please mention the solution approach / methodology

Business Problem Overview and Solution Approach

Solution Approach:To address the increasing number of cancellations, INN Hotels Group requires a machine learning-based solution to predict potential booking cancellations. This will help the hotel group to identify the bookings likely to be canceled in advance, implement more effective policies for cancellations and refunds and Optimize room occupancy and overall profitability and the steps includes -

a) Data Analysis: Analyze the provided booking data to understand factors influencing cancellations and Identify significant features affecting booking status such as lead time, previous cancellations, repeat guest status, special requests, etc.

b) Feature Engineering: Encode categorical variables and Normalize numerical variables if necessary.

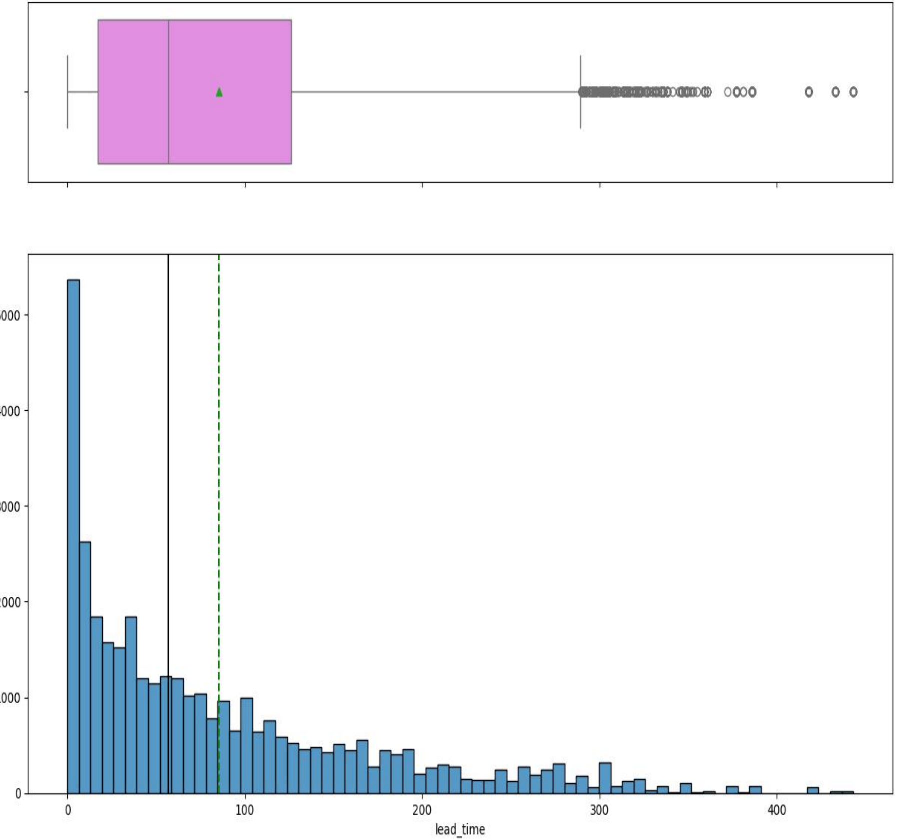
c) Model Development: To develop a predictive model using decision tree classifiers and other appropriate machine learning algorithms and to train the model on historical booking data to predict cancellations.

d) Model Evaluation: To Evaluate the model using metrics such as accuracy, precision, recall, and F1-score, Fine-tune the model to improve its predictive performance.

e) Implementation: Deploy the predictive model to flag potential cancellations in the booking system, Formulate new cancellation and refund policies based on model insights.Continuously monitor model performance and update it with new data to maintain accuracy.

By implementing this machine learning-based solution, INN Hotels Group can proactively manage booking cancellations, thus enhancing their operational efficiency and profitability.

EDA Results



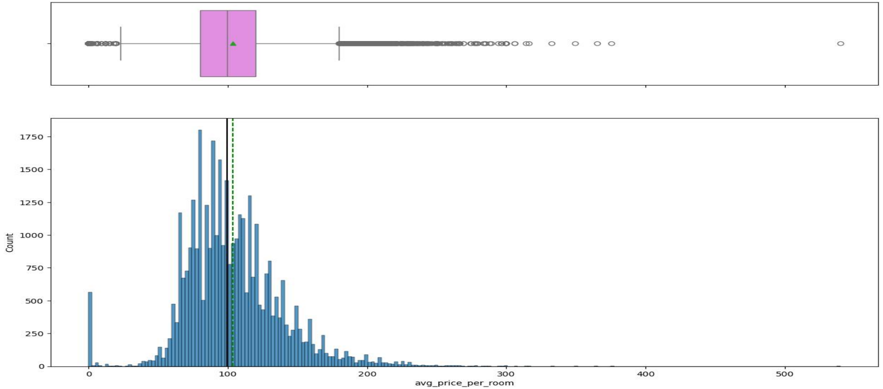
The EDA shows that there are 36275 rows and 19 columns on the data comprising of booking status, market segmentations, arrivals periods, room type reserved among others with oone of them being float, 13 integers and 5 object as per the data types. We also saw that there are some duplicated data.

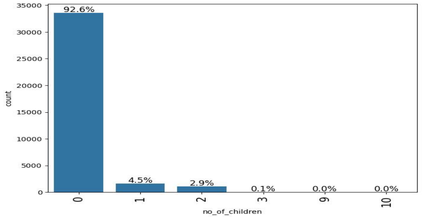
There are average of 2 adults from statistical summary, which is also true for 75% of the data but we also have 4 adults in some cases which is the most. Though there are some outliers of 10 children, most of the bookings comes withoit children.. Most of the bookings were 2 weekend nights with average being around 1 while we have 7 weekend nights as the most. for week nights we saw 17 week nights as the most though we have 2 as the avergae and 75th percentile being 3.

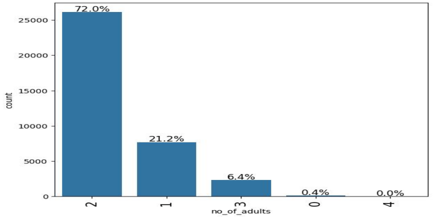
Most bookings comes withouit parking spoace request. the average lead time is 85 days which is occasioned by outilier found in the most lead time which is 443 days because 75th percentile is 126 days lead time and we have 50th percentile of the booking coming at 57 days. the outliers from the univariate analysis comes around 290 days and more on lead time.

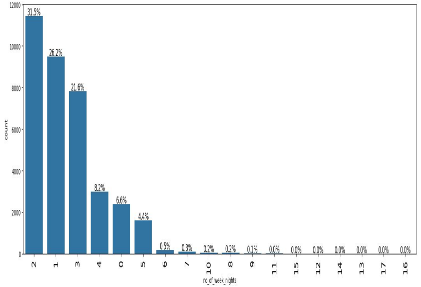
Most guests arrivale date is July and August expectedly due to summer and we have very few repeated guests from the data.

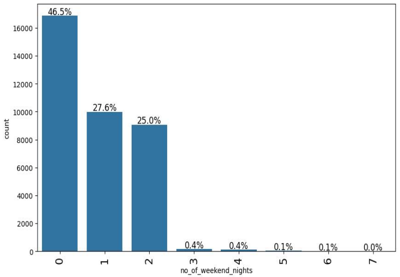
EDA Results









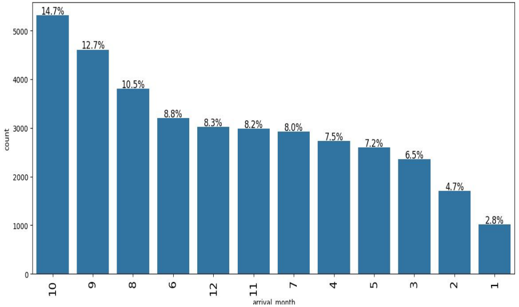


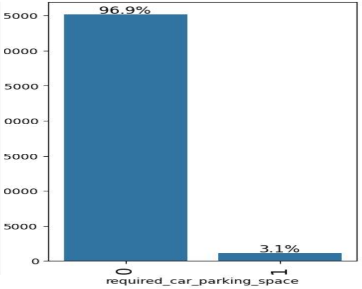
The average price per room is seem to be around 103 dollars which is pretty similar to the circa 99 dollars of the 50th percentile. Most price came under 120 dollars which is the 75th percentile, apart from few outliers which came after 180 dollars with highest price being 540 dollars- obvious outlier.

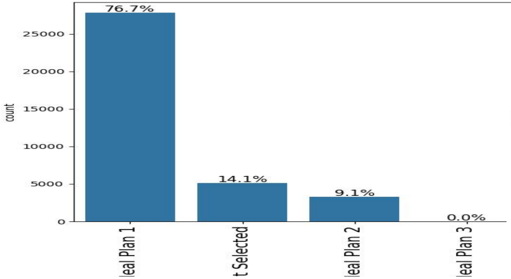
72% of the bookings came for 2 adults and 21% came for a single adults. Only 6.4% has 3 adults. Most of the bookings (over 92%) came without children, while only aboput 4% came with one child. There are average of 2 adults from statistical summary, we also have 4 adults in some cases which is the most. Though there are some outliers of 10 children, most of the bookings comes without children.

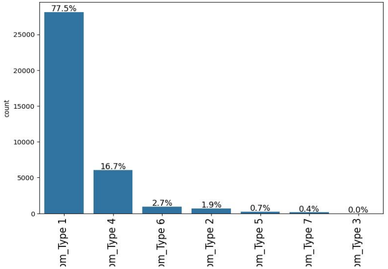
31% of the bookings were for 2 week nights, 21% 1 week nights and 22% three week nights. We saw 17 week nights as the most though we have 2 as the averge and 75th percentile being 3. For the weekend nights, 27% is for 1 weekend night while 25% is for 2 weekend night. The average is around 1 weekend night, while we have 7 weekend nights as the most. We have more bookings without weekend night at 46%.

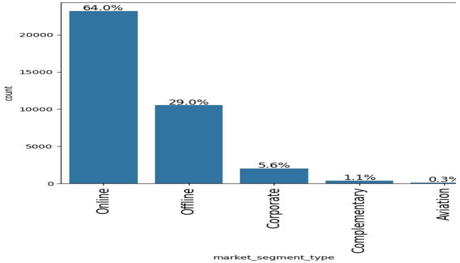
EDA Results

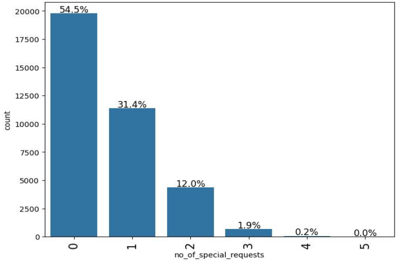












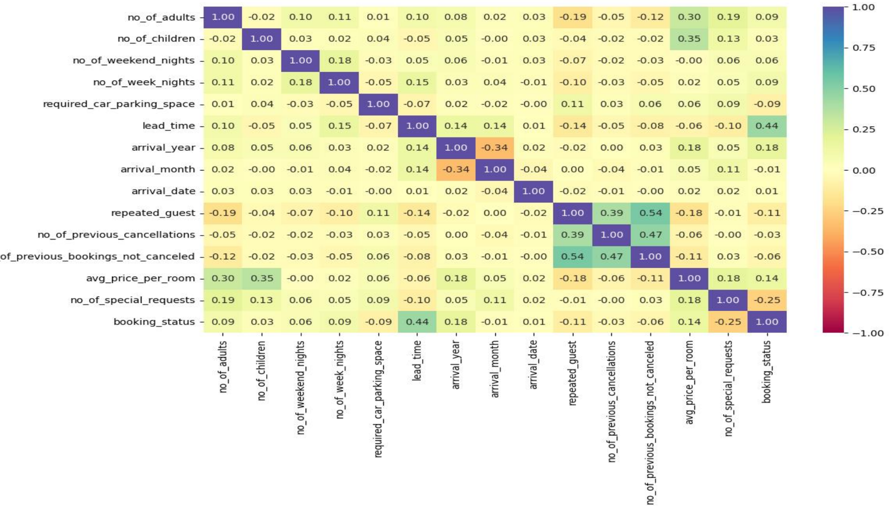
About 65% of the bookings were for July though to December with th most guests arrivale month being October at about 15% and September at 13%. we have the least booking arrival month at January and Feb which is the peak of the winter. Only about 3% of the bookings comes with Car Parking space.

Most of the bookings at almost 80%, comes with meal one plan which bed and breakfast. 14% of the bookings has no meal selections and only 9% chose the half borad meal.

77% of the bookings were also seen to room type one followed by around 17% that went for room type 4.

^4% of the bookings were online, 29% were done offline while only5.6% were actually corporate relationships. Most bookings at 55% did come with any special request, while only 31% has just one special requests.

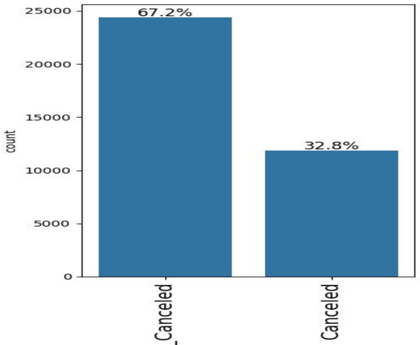
EDA Results

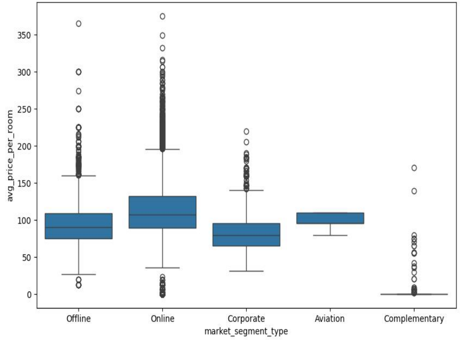


we saw frfom the correlation plot that repeated gusts mostly do not cancle their bookings and that is the most correlated variables on the table. Those that didnt cancle before are seen to sustain that as well as seen that those two variables are also correlated. No of person (children or adults) are alos very correlated to the price of the room.

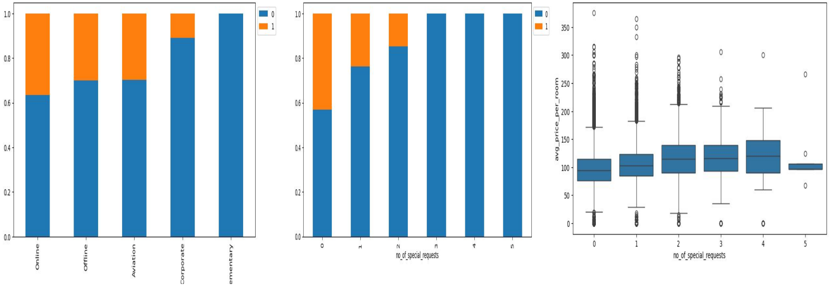
About 33% of the bookings were cancvlelled and that is quite a substansive number considering the imapct of revenue.

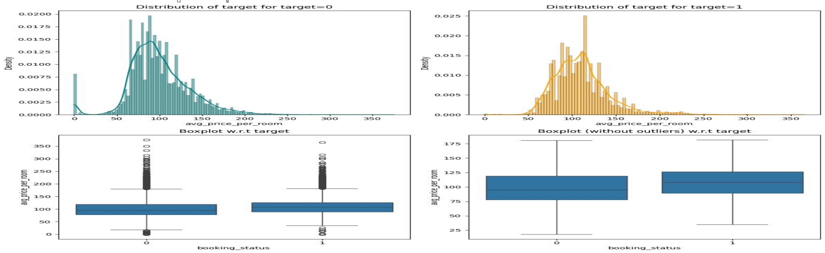
Though with strong outliers, we saw that the prices of the rooms were higher for online booking, folowed by offline bookings. while corporates enjoyed more lower booking average price rates.

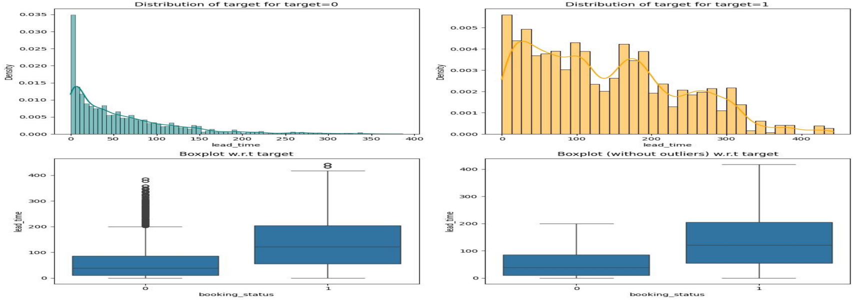




EDA Results







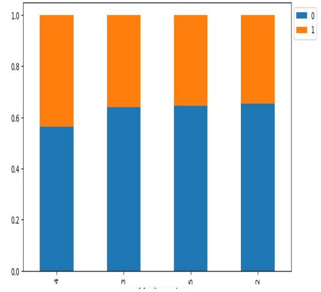
Further analysis shows that most cnacellations came from online bookings, followed closely by offline and aviation. Corpotaye is quite low while we saw almost no cancelation in *complimentary.*

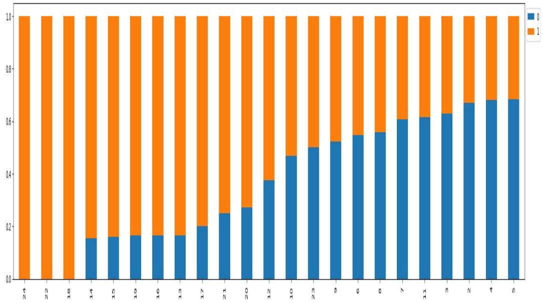
We saw the higher the special request, the least likely that there womt be any cancellations as highest cancellations came from zero special request and trended downward as the special requests increases.

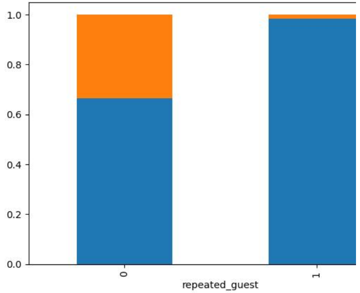
We also saw that barring the outliers, the price per room increases slightly as the special requests increases. We alsop saw that most of the cancelled rooms were of higher price

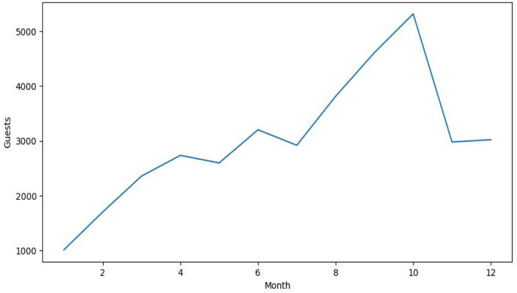
Furthermore, we saw that the higher the lead time, the more likely the canclellations from the data.

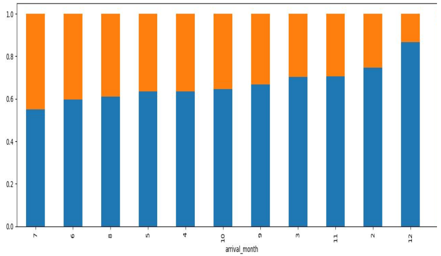
EDA Results

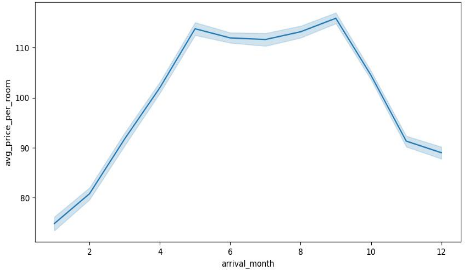












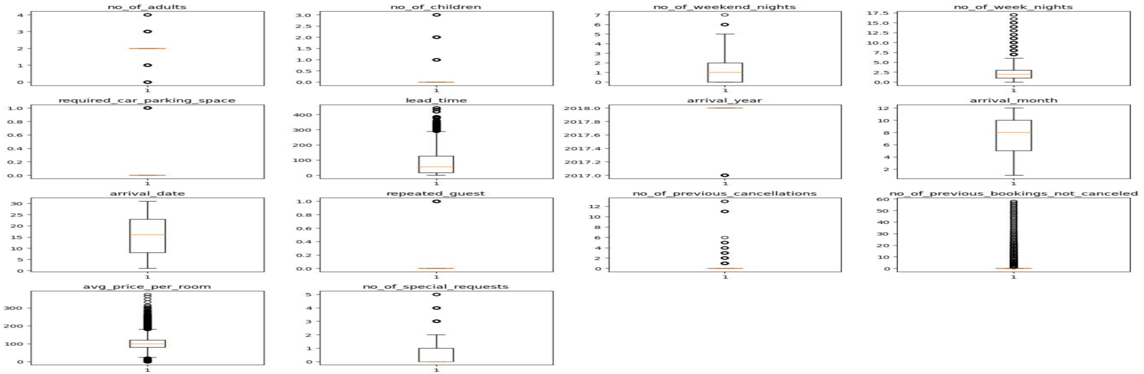
The booking status didnt realy change significantly with number of family members, But we saw that the bookings with longer total stay days tend to cancle more than bookings with fewer stay days (stay days being number of weekdays night + number of weekday nights).

We also saw that most repeating guests do not cancel their booking as much as first timers that are not repeating guests. From the line plot, we saw that the busy months starts in July through to December with busiest month is October . But then January which is the least busy month also has the least cancellations.

Over the months review also shows that average price per room rises consistently Jan though to June and stays steady at that highest place and then starts dropping from septeber till December.

Data Preprocessing

After applying the duplicated method and filtering the DataFrame to show only duplicates, the result is an empty DataFrame. This means there are no rows in the dataset that are considered duplicates based on all columns. Using ***data.isnull().sum(),*** we found that therte is also no missing value in the data.



We also saw after applying the necessary codes that only arrival month, and arrival date does NOT have outliers, all other columns had outliers.

Data Preprocessing

- Duplicate value check
- Missing value treatment
- Outlier check (treatment if needed)
- Feature engineering
- Data preparation for modeling

***Note**: You can use more than one slide if needed*

Model Performance Summary

The logistic regression model attempts to predict a binary outcome (canceled or not canceled booking) based on multiple predictors of Dependent Variable: booking_status (whether a booking is canceled or not). However, the model did not converge in the first time after multiple iterations, indicating potential issues such as multicollinearity or insufficient iterations. This requires further investigation or adjustment to the model parameters.

We then, defined a function for VIF and calculated for each feature to see those that are above 5 and treat them by dropping and checking their imapct. We will drop the predictor variables having a p-value greater than 0.05 as they do not significantly impact the target variable. so using the python codes, we build a model, check the p-values of the variables, and drop the column with the highest p-value.Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value. Repeat the above two steps till there are no columns with p-value > 0.05.

The final model then converged using the Method of Maximum Likelihood Estimation (MLE) after treating the multicollinearity with following parametrs Model Intercept (const) as 1828.8682, and the Coefficients are - no_of_adults: 0.1905, no_of_weekend_nights: 0.1429, no_of_week_nights: 0.0533, required_car_parking_space: -1.3089, lead_time: 0.0124, arrival_year: 0.9046, arrival_month: -0.0280, repeated_guest: -2.6583, no_of_previous_cancellations: 0.2101, avg_price_per_room: 0.0181, no_of_special_requests: -1.1046.

We also have Number of Observations: 25,392, Degrees of Freedom (Model): 11, Degrees of Freedom (Residuals): 25,380, Pseudo R-squared: 0.2735, Log-Likelihood: -11,691, LL-Null: -16,091, LLR p-value: 0.000

Model Performance Summary

Positive Predictors (Increase Likelihood of Cancellation):

1. Number of Adults (no_of_adults, coef: 0.1905): *More adults in the booking increase the likelihood of cancellation.*

2. Number of Weekend Nights (no_of_weekend_nights, coef: 0.1429): *More weekend nights increase the likelihood of cancellation.*

3. Number of Week Nights (no_of_week_nights, coef: 0.0533): *More week nights increase the likelihood of cancellation.*

4. Lead Time (lead_time, coef: 0.0124): *Longer lead times increase the likelihood of cancellation.*

5. Arrival Year (arrival_year, coef: 0.9046): *Bookings in later years are more likely to be canceled.*

6. Previous Cancellations (no_of_previous_cancellations, coef: 0.2101): *More previous cancellations increase the likelihood of cancellation*.

7. Average Price per Room (avg_price_per_room, coef: 0.0181): *Higher average price increases the likelihood of cancellation*.

Negative Predictors (Decrease Likelihood of Cancellation):

1. Required Car Parking Space (required_car_parking_space, coef: -1.3089): *Bookings that require car parking are less likely to be canceled.*

2. Arrival Month (arrival_month, coef: -0.0280): *Bookings in later months are less likely to be canceled.*

3. Repeated Guest (repeated_guest, coef: -2.6583): *Repeated guests are significantly less likely to cancel.*

4. Special Requests (no_of_special_requests, coef: -1.1046): *More special requests decrease the likelihood of cancellation.*

Model Performance Summary

Summary of Key Performance Metrics for Training and Test Data of All Models

Metric	Logistic Regression
Training Accuracy	0.824
Test Accuracy	0.820
Training Precision	0.792
Test Precision	0.789
Training Recall	0.752
Test Recall	0.748
Test F1-Score	0.768
ROC AUC	0.882

Summary

✓ *The logistic regression model identified key features impacting booking cancellations.*

✓ *Features such as the number of adults, weekend nights, lead time, and previous cancellations increase the likelihood of a booking being canceled.*

✓ *Features like requiring a car parking space, being a repeated guest, and making special requests decrease the likelihood of cancellation.*

✓ *The model performed well with a good balance of precision, recall, and F1-score on both training and test datasets.*

This summary helps stakeholders understand the model's behavior and effectiveness in predicting booking cancellations, facilitating data-driven decision-making to reduce cancellations.

Model Performance Summary

- Overview of the final ML model and its parameters
- Summary of most important features used by the ML model for prediction
- Summary of key performance metrics for training and test data of all the models in tabular format for comparison

***Note**: You can use more than one slide if needed*

APPENDIX

Data Background and Contents

Background: The dataset pertains to hotel booking records and includes various features that describe the booking details and customer behaviors. The primary objective is to predict the booking_status, which indicates whether a booking was canceled or not.

Data Description:The dataset contains-

1. Booking_ID: Unique identifier for each booking.no_of_adults: Number of adults in the booking.
2. No_of_children: Number of children in the booking.
3. No_of_weekend_nights: Number of weekend nights (Saturday and Sunday) included in the stay.
4. No_of_week_nights: Number of week nights (Monday to Friday) included in the stay.
5. Type_of_meal_plan: Type of meal plan chosen by the customer (e.g., Meal Plan 1, Meal Plan 2, etc.)
6. Required_car_parking_space: Indicates whether a car parking space is required (1 for yes, 0 for no).
7. 7. room_type_reserved: Type of room reserved by the customer (e.g., Room_Type 1, Room_Type 2, etc.).

8. Lead_time: Number of days between the booking date and the arrival date.

9. Arrival_year: Year of arrival.

10. Arrival_month: Month of arrival.

11. Arrival_date: Day of the month of arrival.

12. Market_segment_type: Market segment classification (e.g., Online, Offline, Corporate, etc.).

13. Repeated_guest: Indicates if the customer is a repeated guest (1 for yes, 0 for no).

14. No_of_previous_cancellations: Number of previous bookings that were canceled by the customer.

15 No_of_previous_bookings_not_canceled: Number of previous bookings that were not canceled by the customer.

16. Avg_price_per_room: Average price per day of the booking.

no_of_special_requests: Number of special requests made by the customer (e.g., high floor, extra bed).

booking_status: Target variable indicating if the booking was canceled (1 for canceled, 0 for not canceled).

Data Background and Contents

Dataset Overview:

- Number of Records: 25,392 bookings.
- Time Frame: Data spans bookings made over multiple years (includes arrival_year, arrival_month, and arrival_date).
- Booking Behavior: Includes customer preferences (e.g., meal plans, special requests), booking lead time, and previous booking behavior.
- Market Segment: Provides insight into the source of the booking (e.g., online, offline, corporate).Please mention the data background and contents

Data Preparation:

- ✓ Handling Missing Values: Checked for missing values and duplicated values
- ✓ Feature Engineering: Created relevant features to improve model performance.
- ✓ Data Encoding: Converted categorical variables into numerical values when necessary.
- ✓ Created new Data Frames when required.
- ✓ Data Splitting: Split the dataset into training and testing sets to evaluate model performance.

Analysis Objectives:

- ▪ Predictive Modeling: Develop a logistic regression model to predict booking cancellations.
- ▪ Feature Importance: Identify the key features that influence the likelihood of booking cancellations.
- ▪ Performance Evaluation: Compare the performance of different models using metrics like accuracy, precision, recall, F1-score, and ROC AUC.

Data Background and Contents

Tests Conducted to Check Assumptions of Logistic Regression

- Multicollinearity: Logistic regression assumes that the predictors are not highly correlated with each other. This can be tested using:Variance Inflation Factor (VIF)

- Linearity of Logits: Logistic regression assumes a linear relationship between the independent variables and the logit (log-odds) of the dependent variabl

Interpretation of Results Based on Coefficients and Odds:

✓ Coefficients (Log-Odds): The coefficients in logistic regression represent the change in the log-odds of the dependent variable for a one-unit change in the predictor.

✓ Odds Ratios: Converting coefficients to odds ratios helps interpret the impact of predictors. Odds Ratio = exp(coefficient): Represents the multiplicative change in the odds for a one-unit increase in the predictor. Percentage Change in Odds = (exp(coefficient) - 1) * 100: Represents the percentage change in odds.

Interpretation of Key Predictors:

- no_of_adults: Odds ratio = 1.21; for each additional adult, the odds of a booking increase by 21%.
- no_of_weekend_nights: Odds ratio = 1.15; for each additional weekend night, the odds of a booking increase by 15%.
- required_car_parking_space: Odds ratio = 0.27; needing a car parking space decreases the odds of a booking by 73%.
- lead_time: Odds ratio = 1.01; for each additional day of lead time, the odds of a booking increase by 1%.

Data Background and Contents

Comment on Model Performance

Training Performance:

Accuracy: 78.33%

Recall: 55.27%

Precision: 72.40%

F1-Score: 62.68%

Test Performance with Default Threshold:

Accuracy: 79.06%

Recall: 56.36%

Precision: 72.79%

F1-Score: 63.53%

Test Performance with Optimized Threshold (0.37):
Accuracy: 75.96%
Recall: 74.11%
Precision: 60.50%
F1-Score: 66.62%

Test Performance with Threshold (0.42):
Accuracy: 78.41%
Recall: 65.05%
Precision: 67.19%
F1-Score: 66.10%

Overall Model Performance:

The logistic regression model performs reasonably well with high accuracy and acceptable F1-scores.
There is a trade-off between precision and recall, which can be adjusted using different thresholds based on the business requirement (e.g., higher recall might be preferred for capturing more positive cases, even at the expense of precision).

Recommendations:

Threshold Adjustment: Adjusting the threshold based on the ROC curve or business needs can help balance precision and recall.
Feature Engineering: Consider additional feature engineering to improve model performance.
Model Comparison: Comparing with other models (e.g., decision trees) might provide better insights or alternatives for improvement.

The model shows that it can identify key predictors of bookings and perform well, but there's always room for optimization based on specific goals and further analysis.

Data Background and Contents

Model Building Steps for Decision Tree

Data Preparation:

- ✓ Load and Inspect Data: Load the dataset and inspect its structure and quality.
- ✓ Preprocess Data: Handle missing values, encode categorical variables, and scale/normalize features if necessary.
- ✓ Split Data: Divide the dataset into training and test sets.

Train the Model:

- ✓ Initialize Model: Create a DecisionTreeClassifier object from sklearn with parameters such as max_depth, min_samples_split, min_samples_leaf, etc.
- ✓ Fit the Model: Train the decision tree model on the training dataset using the .fit() method.

Tune the Model:

- ✓ Pruning: Use techniques like cost complexity pruning to avoid overfitting. Plot impurity vs. alpha to determine the optimal alpha value.
- ✓ Cross-Validation: Perform cross-validation to ensure the model's performance is consistent across different subsets of data.

Evaluate Model Performance:

- ✓ Predict on Test Set: Use the .predict() method to generate predictions on the test set.
- ✓ Confusion Matrix: Generate a confusion matrix to evaluate performance in terms of true positives, false positives, true negatives, and false negatives.
- ✓ Performance Metrics: Calculate accuracy, precision, recall, F1-score, and plot ROC curves if needed.
- ✓ Feature Importance: Examine feature importances to understand which features contribute most to the model's decisions.

Visualize the Model:

- ✓ Decision Tree Plot: Visualize the decision tree to understand how decisions are made. This can be done using tools like plot_tree in sklearn.

Make Predictions and Interpret Results:

- ✓ Predict: Make predictions using the test set or new data.
- ✓ Interpret Results: Analyze the model's predictions and performance metrics to make informed decisions or recommendations.

Data Background and Contents

Model Building Steps for Decision Tree

Comment on Model Performance

Training Performance:

Accuracy: 99.42%

Recall: 98.66%

Precision: 99.58%

F1-Score: 99.12%The training performance metrics indicate that the model fits the training data exceptionally well, with very high accuracy, recall, precision, and F1-score. However, such high performance on the training data might indicate overfitting, especially if the test performance is not as strong.

Test Performance:

Accuracy: 87.12%

Recall: 81.18%

Precision: 79.46%

F1-Score: 80.31%

I

Interpretation of Test Performance

✓ Accuracy: The model correctly classifies approximately 87.12% of the test samples.

✓ Recall: The model identifies 81.18% of actual positive cases, which is relatively high

✓ Precision: The model's precision is 79.46%, indicating the proportion of true positives among all predicted positives

✓ F1-Score: The F1-score of 80.31% balances precision and recall, providing a single metric to assess model performanc

Confusion Matrix Analysis:

- True Positives (TP): 3014
- True Negatives (TN): 6441
- False Positives (FP): 920
- False Negatives (FN): 508

The confusion matrix shows that the model has a high number of true positives and true negatives, but also a considerable number of false positives and false negatives. This implies that while the model is generally accurate, there is room for improvement in minimizing false positives and false negatives, particularly if these errors have significant implications for business decisions.

Data Background and Contents

Decision Rules

Root Node Decision: The first decision based on the most important feature.

Subsequent Nodes: Each decision node splits the data based on feature values, leading to different branches and ultimately to the leaf nodes where the final classification is made.

Interpretation: Each path from the root to a leaf node represents a series of conditions (rules) that lead to a particular classification. These rules can be used to understand the model's decision-making process.

Feature Importance:

Feature Importances: Feature importance measures how much a feature contributes to the prediction. Features with higher importance are more influential in the model's decision-making process.

Visualization: A bar plot of feature importances can help identify which features have the most impact.
Interpretation: Features with higher importance are those that provide the most significant information for the model's predictions

Overall, pruning techniques contribute to a more robust and generalizable model by balancing complexity and performance, ensuring that the model performs well on both training and unseen data.

Pruning Techniques and Their Impact:

Cost Complexity Pruning:

Purpose: Cost complexity pruning helps to reduce the size of the decision tree by removing nodes that have little importance, thereby simplifying the model and improving generalization.

Process: We compute a series of alpha values and their corresponding impurity measures to decide which nodes to prune.

Results: By examining the plot of impurity versus alpha, we can choose an optimal alpha value that balances model complexity and performance. This typically results in a tree with fewer nodes and branches, which can improve test performance by reducing overfitting.

Effect on Performance:

Training vs. Test Performance: Before pruning, the model might show very high performance on the training set, but lower performance on the test set. After applying pruning, we generally expect:

Reduced Overfitting: Improved generalization to unseen data.
Improved Test Metrics: Metrics such as accuracy, precision, recall, and F1-score might improve or stabilize on the test set as the model becomes less complex.

Balanced Performance: A trade-off between training and test performance, ensuring the model is not too complex but still captures the necessary patterns in the data.

Great Learning