Detailed Analysis of Hotel Bookings

(**Insights from Data Analysis and Predictive Modeling**)

Author: Vipul B. Khachane.
Experience: 3.2 Years as a Data Scientist.

Report Date: October 15, 2024

Introduction

- **Objective:** To analyze hotel booking data for trends in customer demographics, booking patterns, and cancellation behaviors.
- **Importance:** Understanding these trends aids in refining marketing strategies and improving customer experience.

Contents

1.Dataset Overview -

The dataset consists of **119,390 entries** across **32 features**, detailing customer demographics, booking specifics, and cancellation information. Key attributes include:

- **is_canceled**: Indicates whether the booking was canceled.
- **adr**: Average Daily Rate, a crucial metric for revenue management.

Post-exploration, the dataset was refined to **87,389 usable entries**, ensuring high data quality for subsequent analyses.

2. Data Loading :

- Loaded the hotel bookings data from a CSV file using **Pandas**.
- Imported some important libraries such as NumPy, seaborn , matplotlib.

3. Data Cleaning and Preprocessing:

Data preprocessing involved several key steps:

- **Handling Missing Values**: Imputation was employed for crucial features to ensure complete datasets.
- **Removing Duplicates**: Duplicate records were eliminated to maintain data integrity.
- **Encoding Categorical Variables**: Categorical features were encoded to facilitate analysis and modeling.

What did you know about your dataset?

After looking over the dataset,here are some following observations:

1. This Dataset contains **119390** rows and **32** columns.
2. In the data there are **31994** duplicate values, which must be dropped.
3. In the entire dataset there are nearly less null values, but some of the columns contain more than 10% of null values.
4. Agent column holds more than 20% of the missing values and can be taken care properly as it is important column and does not hold much missing values.
5. On the otherside the column Company holds more than 90% of the missing values and is of no use as it contain greater number of null values. I must drop this column afterwards while analysing the dataset.

These are some of conclusion which i have come until now after going through the dataset.

4. Exploratory Data Analysis (EDA) -

4.1 Customer Demographics

The analysis of customer demographics revealed a healthy proportion of repeated guests, emphasizing the need for loyalty programs. This segment presents opportunities for targeted marketing strategies.

4.2 Cancellation Patterns

The overall cancellation rate was determined to be **27.53%**. Monthly cancellations highlighted peak periods for cancellations, indicating varying cancellation behavior by hotel type, which suggests different risk profiles for various establishments.

4.3 ADR Trends

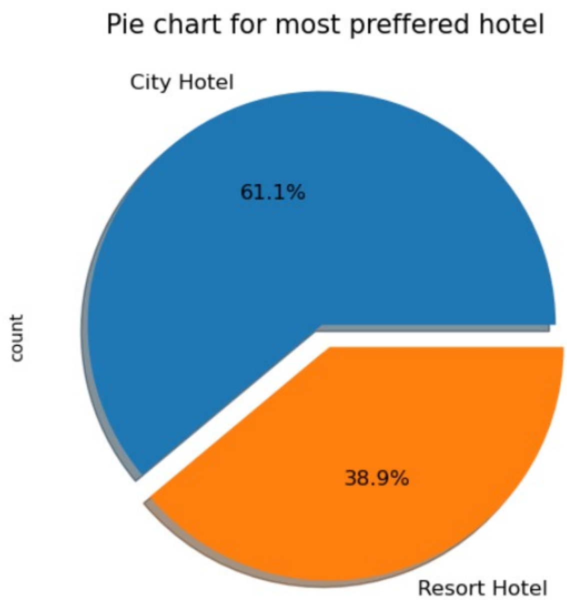
The distribution of Average Daily Rate (ADR) illustrated pricing strategies across the dataset. The analysis indicated a concentrated pricing strategy, suggesting opportunities for revenue optimization during off-peak periods.

4.4 Additional Insights

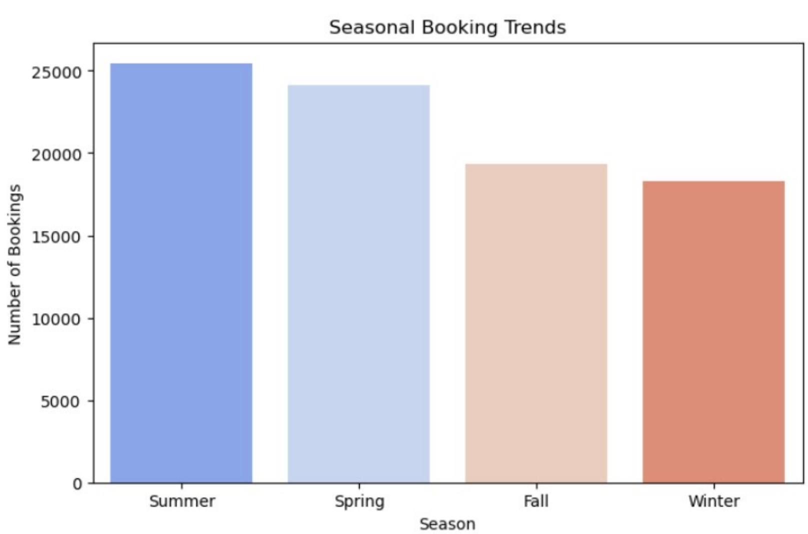
- **Booking Lead Time**: A higher lead time correlated with increased cancellations, providing actionable insights for managing future bookings.
- **Guest Composition**: Investigating the number of adults and children in bookings revealed varying cancellation patterns, emphasizing the importance of understanding guest demographics.

Visualized Graphs And Plots

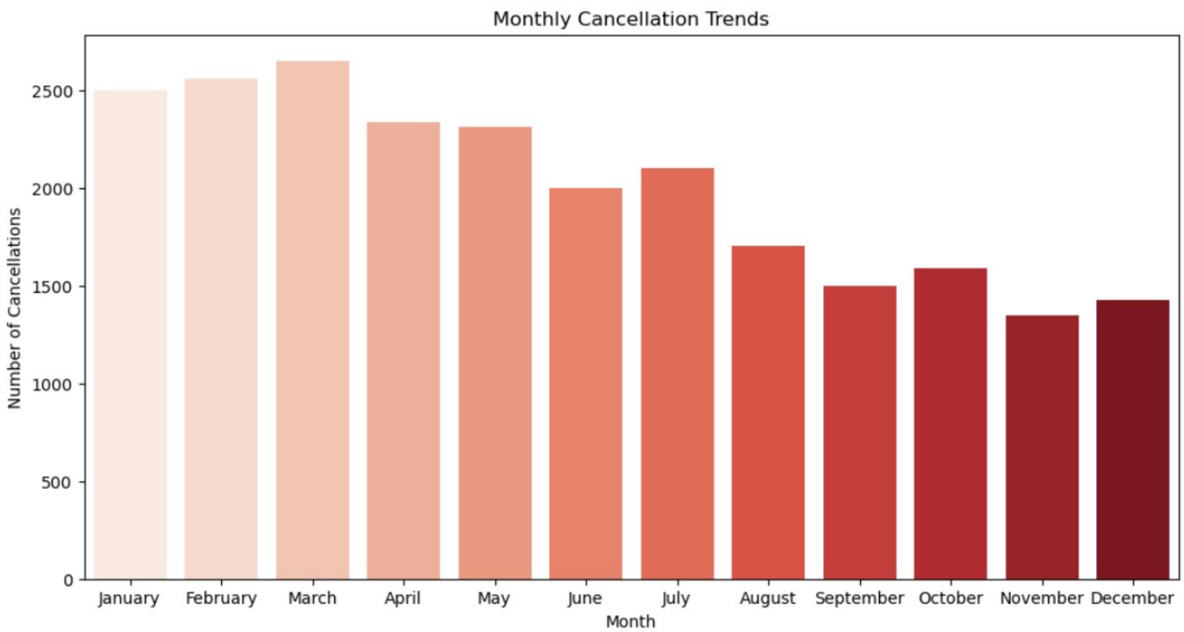
1. Pie chart for most preffered hotel-



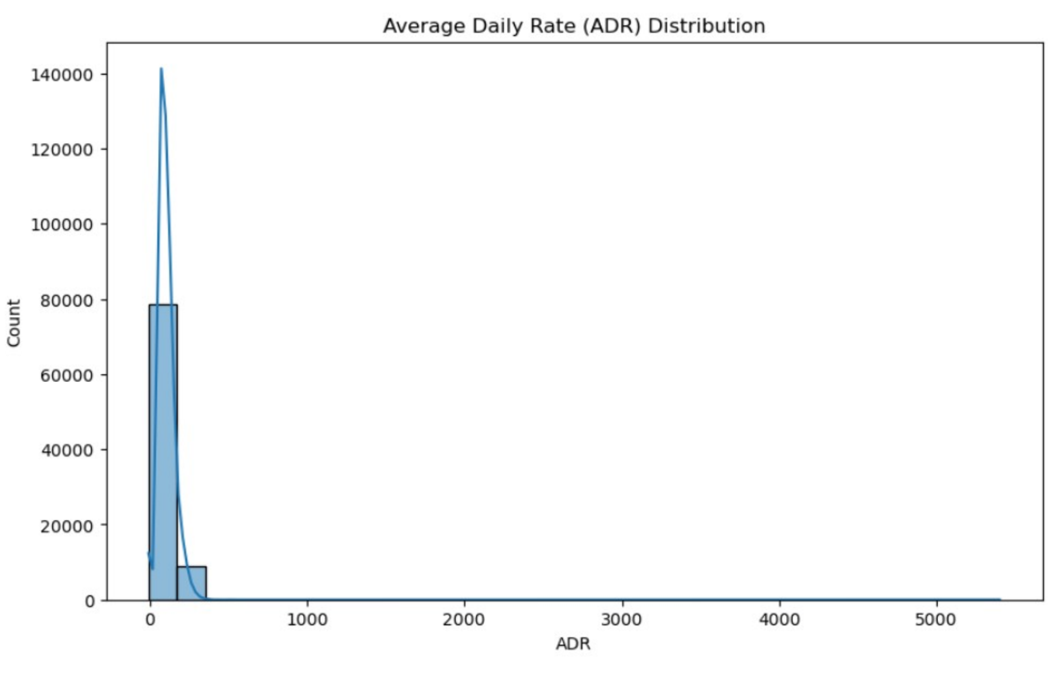
2. Plotting graph for seasonal booking trends-



3. Plotting graph for Cancellation Patterns -

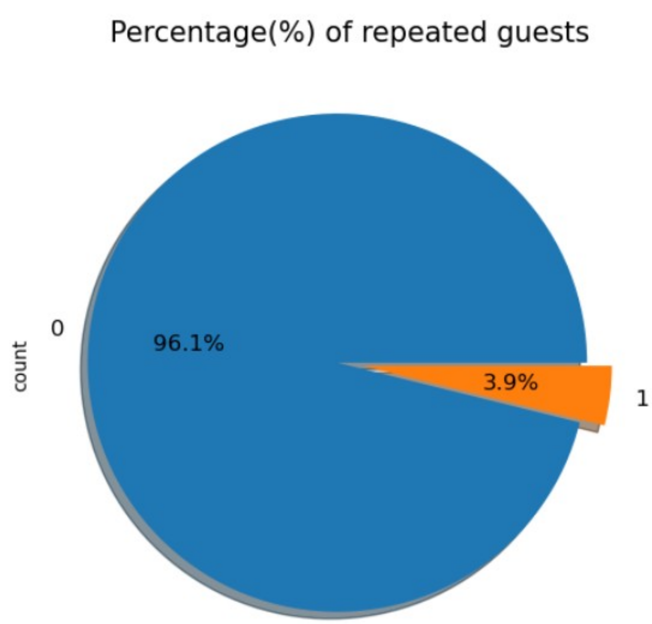


4. Plotting graph for ADR Trends -



5. Plotting the graph for Factors Affecting Cancellations -

6. Customer Demographics- Percentage of repeated guests -



5. Hypothesis Testing

Three key hypotheses were tested:

1. **Booking in Advance**: Customers booking more than 6 months in advance are more likely to cancel. A Chi-squared test showed significant results (p-value < 0.001).
2. **Weekday vs. Weekend Bookings**: Weekday bookings have a higher ADR than weekend bookings. A t-test confirmed this hypothesis with a t-statistic of 3.35 (p-value < 0.01).
3. **Special Requests and Cancellations**: A significant relationship exists between the number of special requests and cancellations, as shown by a Chi-squared test (p-value < 0.001).

These tests not only validate assumptions but also guide strategic planning.

6. Predictive Modeling -

6.1 Model Selection and Training

Two models were developed for predicting cancellations:

- **Logistic Regression**: A baseline model providing insights into linear relationships.
- **Random Forest Classifier**: An ensemble method capturing non-linear patterns.

Features Selected:

- lead_time, arrival_date_year, arrival_date_month, adults, children, hotel, meal, market_segment, deposit_type, previous_cancellations, previous_bookings_not_canceled.

The dataset was divided into training (80%) and test (20%) sets for evaluation.

6.2 Model Performance Evaluation

- **Logistic Regression**: Achieved an accuracy of **75%** but had low recall for cancellations (16%).
- **Random Forest**: Achieved an accuracy of **73%** with better precision for non-cancellations but still had low recall for cancellations (34%).

Performance Summary:

Model	Precision (Non-Cancelled)	Recall (Cancelled)	F1-score (Cancelled)	Accuracy
Logistic Regression	75%	16%	0.26	75%
Random Forest	78%	34%	0.41	73%

6.3 Insights from Model Evaluation

The models exhibited significant challenges in predicting cancellations, highlighting a need for more balanced datasets and advanced modeling techniques.

7. Recommendations

1. **Data Balancing**: Implement oversampling or undersampling techniques to improve model learning from both classes.
2. **Feature Engineering**: Investigate additional features or transformations that could enhance predictive accuracy.
3. **Hyperparameter Tuning**: Optimize model parameters using techniques like grid search, particularly for the Random Forest model.
4. **Explore Alternative Models**: Consider algorithms like XGBoost or ensemble methods for improved performance on imbalanced datasets.
5. **Focus on Evaluation Metrics**: Use ROC-AUC and confusion matrices for a comprehensive evaluation of model performance.

8. Conclusion-

- The analysis of the hotel booking dataset provides valuable insights into customer behavior, booking trends, and cancellation patterns. Key findings and real-life implications include:

- **High Cancellation Rates**: With a cancellation rate of 27.53%, this poses a significant challenge for revenue forecasting and resource allocation. Understanding the reasons behind cancellations can help mitigate losses.
- **Impact of Lead Time:** Longer lead times are associated with higher cancellation rates. This suggests that hotels can benefit from encouraging shorter booking windows through targeted promotions.
- **Booking Patterns**: Differences between weekday and weekend bookings indicate varying demand levels, suggesting that pricing strategies should be adjusted based on the day of the week to maximize occupancy.
- **Importance of Customer Segmentation**: Analyzing guest demographics and special requests can help tailor services and marketing strategies, enhancing customer satisfaction and loyalty.

Recommendations to Increase Bookings

1. **Implement Flexible Booking Policies**: Offering flexible cancellation policies may reduce booking hesitations and encourage more reservations.
2. **Targeted Promotions**: Create special offers for shorter lead times, such as discounts for last-minute bookings, to capitalize on last-minute travelers.
3. **Loyalty Programs**: Enhance loyalty programs to incentivize repeated visits, focusing on guests who have previously canceled bookings.
4. **Dynamic Pricing Strategies**: Utilize dynamic pricing based on demand forecasts to maximize revenue during peak periods while remaining competitive during off-peak times.
5. **Enhanced Customer Communication**: Improve pre-arrival communication to confirm bookings and provide personalized offers, which may reduce cancellations.


THANK YOU!