Summary Report on Hotel Booking Analysis

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

Report Date: October 15, 2024

Table of 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.

Approach -

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.

---

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.

---

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 hotel booking data reveals critical trends that can inform strategic decisions to minimize cancellations and enhance customer satisfaction. The insights gained from this study provide a solid foundation for implementing data-driven practices in hotel management.