**Hotel Booking Trends and Cancellation Prediction: A Comprehensive Analysis**

**Author**: Pankaj Patil  
**Date**: 20/09/2024

**1. Problem Statement**

The goal of this project is to predict hotel booking cancellations and provide actionable insights to improve hotel operations, pricing strategies, and customer satisfaction.

Hotel cancellations significantly impact revenue and operational efficiency. Overbooked rooms can lead to dissatisfied customers, while frequent cancellations cause uncertainty in resource allocation. This analysis aims to understand the factors influencing cancellations and develop a predictive model to mitigate their impact.

**2. Background**

Cancellations affect various aspects of hotel management, including:

* **Revenue Management**: Unanticipated cancellations can result in revenue losses.
* **Resource Allocation**: Frequent cancellations disrupt staffing and inventory management.
* **Customer Satisfaction**: Overbooking or last-minute cancellations frustrate customers.

This project seeks to:

* Analyze historical booking data to identify key factors affecting cancellations.
* Build a machine learning model to predict cancellations.
* Provide recommendations to optimize hotel operations and reduce cancellations.

**3. Objectives**

1. **Understand Booking Patterns**: Analyze features like lead time, customer demographics, and seasonality to identify booking trends.
2. **Predict Cancellations**: Develop a predictive model using machine learning algorithms to forecast cancellations.
3. **Actionable Insights**: Recommend strategies for improving pricing, marketing, and customer segmentation based on analysis.

**4. Data Overview**

**Data Overview**:

**1. Booking Information:**

* **hotel**: Type of hotel (e.g., City or Resort).
* **is\_canceled**: Target variable indicating if the booking was canceled.
* **lead\_time**: Number of days between booking and arrival.
* **arrival\_date\_year**: Year of arrival.
* **arrival\_date\_month**: Month of arrival.
* **arrival\_date\_week\_number**: Week of the year for arrival.
* **arrival\_date\_day\_of\_month**: Day of the month for arrival.
* **booking\_changes**: Number of changes made to the booking.

**2. Customer Information:**

* **adults**: Number of adults.
* **children**: Number of children.
* **babies**: Number of babies.
* **is\_repeated\_guest**: Whether the customer is a repeat guest.
* **previous\_cancellations**: Number of previous cancellations by the customer.
* **previous\_bookings\_not\_canceled**: Number of previous bookings that were not canceled.
* **customer\_type**: Type of customer (e.g., Transient, Contract, etc.).

**3. Stay Information:**

* **stays\_in\_weekend\_nights**: Number of weekend nights booked.
* **stays\_in\_week\_nights**: Number of weeknights booked.
* **days\_in\_waiting\_list**: Number of days the booking was on a waiting list.

**4. Room Information:**

* **reserved\_room\_type**: Type of room reserved by the customer.
* **assigned\_room\_type**: Type of room assigned upon arrival.
* **total\_of\_special\_requests**: Number of special requests made by the customer.

**5. Financial and Transaction Information:**

* **adr**: Average Daily Rate (ADR) per room.
* **deposit\_type**: Type of deposit paid (e.g., No Deposit, Refundable, Non-Refundable).
* **required\_car\_parking\_spaces**: Number of car parking spaces required by the customer.

**6. Geographic Information:**

* **country**: Country of origin of the customer.

**7. Booking Channel Information:**

* **market\_segment**: Market segment (e.g., Direct, Corporate, Groups).
* **distribution\_channel**: Channel through which the booking was made (e.g., Direct, TA/TO).
* **agent**: ID of the travel agency that made the booking.
* **company**: ID of the company making the booking.

**8. Reservation Status Information:**

* **reservation\_status**: Final status of the reservation (e.g., Check-Out, Canceled, No-Show).
* **reservation\_status\_date**: Date the reservation was updated to its final status.

**5. Data Cleaning and Preprocessing**

**5.1 Handling Missing Values**

* **Children column**: Filled missing values with the mode.
* **Country and company columns**: Filled missing values with "Not Known."
* **Agent column**: Replaced missing values with 0 to indicate no agent involvement.

**5.2 Encoding Categorical Variables**

* **Label Encoding**: Applied to features like hotel type and deposit type.
* **One-Hot Encoding**: Applied to features such as arrival date and market segment to prevent multicollinearity.

**5.3 Feature Engineering**

* Converted **reservation\_status\_date** to datetime format, extracting year, month, and day.

**5.4 Imbalanced Data Handling**

* Applied **SMOTE** to balance the dataset. After SMOTE, the cancellation rate was adjusted to 50%, ensuring a robust training dataset.

**6. Exploratory Data Analysis (EDA)**

**1. Country Insights:**

**Portugal (PRT):**

* **Insight:** Portugal has the highest number of bookings, driven by proximity and domestic travel.
* **Recommendation:** Offer localized promotions, loyalty programs, and packages for domestic travelers to maintain and increase bookings.

**United Kingdom (GBR):**

* **Insight:** UK customers exhibit a higher cancellation rate compared to other countries.
* **Recommendation:** Introduce flexible booking options, targeted check-in reminders, and special incentives (discounts, exclusive offers) to reduce cancellations.

**France (FRA):**

* **Insight:** French customers book steadily throughout the year, with no significant peaks or dips.
* **Recommendation:** Maintain consistent marketing efforts year-round rather than focusing solely on seasonal campaigns.

**Spain (ESP):**

* **Insight:** Spanish customers tend to favor shorter weekend stays.
* **Recommendation:** Offer attractive weekend packages or flash sales targeting Spanish travelers, with amenities suited for short stays.

**Germany (DEU):**

* **Insight:** German customers generally have longer lead times, booking well in advance.
* **Recommendation:** Capitalize on this behavior with early bird discounts and advanced booking rewards.

**2. Seasonal Trends in Bookings:**

* **Peak Months:** Bookings peak between June and August, aligning with summer vacation periods.
* **Low Months:** Bookings dip between November and January, reflecting a post-holiday lull.

**3. Hotel Type Preferences:**

**City Hotels:**

* **Insight:** City Hotels receive more bookings than Resort Hotels.
* **Factors:**
  + **Business Travelers:** Preference for proximity to business hubs and amenities.
  + **Short Stays:** Higher demand for shorter trips, possibly for work or quick getaways.
* **Cancellation Rate:** Higher cancellation rate due to flexible options and spontaneous bookings.
* **Average Lead Time:** Longer lead times due to advance planning for business events.

**Resort Hotels:**

* **Insight:** Resort Hotels are preferred for longer, leisure-oriented stays.
* **Booking Patterns:** Resort bookings are often more spontaneous with shorter lead times.

**4. Pricing and Average Daily Rate (ADR):**

* **Insight:** The ADR for City and Resort Hotels is nearly the same, indicating a consistent pricing strategy across both.
* **Recommendation:** Consider seasonal pricing strategies during peak seasons to maximize revenue.

**5. Room Type Preferences:**

* **Insight:** Room Type A is the most preferred, followed by D, E, and F.
* **Recommendation:** Ensure the availability of popular room types, especially Room Type A, to meet customer demand.

**6. Customer Segments:**

* **Transient Customers:** Represent the highest proportion of bookings, indicating a preference for short-term stays.
* **Transient Party:** Typically small groups or families, second most common segment.
* **Contract Customers:** Least represented, indicating fewer corporate or long-term bookings.

**7. Booking Options:**

* **No Deposit:** Majority of bookings are made without upfront payments.
* **Non-Refundable:** Some guests prefer non-refundable bookings in exchange for lower rates.
* **Refundable Deposits:** Least common due to higher associated costs.

**8. Booking Changes:**

* **Insight:** Most bookings have zero changes, suggesting customer satisfaction with initial reservations.
* **Recommendation:** Monitor the small number of bookings with changes to identify potential improvement areas.

**9. Booking Channels:**

* **Travel Agents/Tour Operators (TA/TO):** Highest volume of bookings.
* **Direct Bookings:** Significant share, possibly driven by incentives or personalized service.
* **Corporate Bookings:** Least represented, suggesting fewer business travelers.

**10. Parking Preferences:**

* **Insight:** Majority of bookings do not require parking spaces, indicating that most guests rely on alternative transportation.

**11. Cancellations and No-Shows:**

* **Cancellation Rate:** Over 40,000 bookings were canceled, pointing to a significant rate that needs attention for revenue management.
* **No-Shows:** Less frequent, indicating that most guests arrive for their bookings.

**12. Cancellation Patterns by Customer Type:**

* **Insight:** Transient customers exhibit a higher cancellation rate, suggesting less commitment to reservations.
* **Recommendation:** Implement retention strategies like personalized offers or reminders to reduce cancellations.

**13. Yearly Booking Trends:**

* **2016:** Consistent and high bookings throughout the year, indicating stable demand.
* **2015:** Fluctuating demand, potentially due to external factors.
* **2017:** High bookings until May, followed by a decline, possibly due to seasonal travel patterns.

**14. Waiting List Analysis:**

* **Insight:** Only 3% of customers are on the waiting list, suggesting most bookings are confirmed without queuing.

**15. Lead Time and Cancellations:**

* **Non-Canceled Customers:** Tend to book with longer lead times, indicating a higher commitment to reservations.
* **Canceled Customers:** Lead time distribution is uniform, suggesting cancellations occur irrespective of booking timing.

Based on the Exploratory Data Analysis (EDA) provided, several variables are likely to influence booking cancellations. **Customer Type** plays a significant role, as transient customers exhibit the highest cancellation rates compared to contract or group bookings, and non-repeated guests are more likely to cancel than repeated guests. **Hotel Type** also impacts cancellations, with city hotels showing higher rates than resort hotels, likely due to the nature of stays, which are often business-related or short-term.

**Booking Flexibility** is another key factor; bookings made with flexible cancellation options tend to have higher cancellation rates, whereas non-refundable bookings, although less common, may decrease the likelihood of cancellations. The **Lead Time** associated with bookings is relevant as well, with shorter lead times correlating to lower cancellation rates, while longer lead times may lead to cancellations due to changes in plans. Notably, canceled bookings exhibit nearly uniform lead times, indicating that cancellations can occur at any time.

The **Booking Channel** influences cancellation patterns too, as direct bookings and travel agency/online travel operator bookings show differing tendencies. **Room Type** may affect cancellations, particularly if there are mismatches between requested and assigned room types. Additionally, **Booking Changes** are significant; bookings that undergo multiple changes are more likely to result in cancellations compared to those that remain stable.

Furthermore, **Booking Deposits** matter; bookings without deposits tend to have higher cancellation rates, while non-refundable bookings suggest higher commitment levels. The **Travel Purpose** is a factor as well, with business-related trips, often associated with city hotels, potentially leading to increased cancellation rates due to shifting travel plans. Lastly, **Seasonal Trends** indicate that off-peak seasons, such as from November to January, may experience higher cancellation rates due to lower travel demand and holiday-related changes.

**7. Hypothesis Testing**

**7.1 Hypothesis 1: Relationship Between Lead Time and Cancellations**

* **Null Hypothesis (H0)**: There is no relationship between the lead time of a booking and its likelihood of being canceled.
* **Alternative Hypothesis (H1)**: There is a relationship between the lead time and cancellation likelihood.

**Chi-Square Test Results:**

* **Chi-square Test Statistic**: 5321.73
* **P-value**: 0.0

**Conclusion**: The extremely low p-value (< 0.05) leads to the rejection of the null hypothesis. This indicates that **lead time** significantly influences the likelihood of cancellations. Specifically, bookings made well in advance (long lead times) are more likely to be canceled.

**7.2 Hypothesis 2: Difference in Average Daily Rate (ADR) Between Weekday and Weekend Bookings**

* **Null Hypothesis (H0)**: There is no significant difference in ADR between weekday and weekend bookings.
* **Alternative Hypothesis (H1)**: There is a significant difference in ADR between weekday and weekend bookings.

**T-test Results:**

* **T-test Statistic**: -21.70
* **P-value**: 3.73e-104

**Conclusion**: Since the p-value is extremely low, we reject the null hypothesis. There is a significant difference between weekday and weekend ADRs, with **weekday bookings having higher ADR** than weekend bookings. This likely reflects pricing strategies aimed at capturing business travel during weekdays.

**8. Predictive Modeling**

**8.1 Model Development**

Two machine learning models were built to predict cancellations:

**1. Logistic Regression**:

* **Accuracy**: 98.61% – This means that out of all the predictions made, 98.61% were correct. Essentially, the model is highly reliable in identifying cancellations.
* **Precision**: 99.88% – This indicates that when the model predicts a booking will be canceled, it is correct 99.88% of the time. In other words, it rarely makes mistakes in its positive predictions.
* **Recall**: 97.33% – This shows that the model correctly identifies 97.33% of all actual cancellations. While it captures most cancellations, there are still a few it misses.
* **AUC (Area Under the Curve)**: 0.9809 – This value suggests that the model is excellent at distinguishing between canceled and non-canceled bookings, performing very well overall.

**2.Random Forest**:

* **Accuracy**: 99.72% – This algorithm is even more accurate than Logistic Regression, correctly predicting 99.72% of cases.
* **Precision**: 100% – When this model predicts a cancellation, it is always correct, meaning it has no false positives.
* **Recall**: 99.43% – This model also performs exceptionally well in identifying actual cancellations, missing only a small fraction.
* **AUC**: 0.9999 – This score indicates outstanding performance, demonstrating the model's ability to differentiate between canceled and non-canceled bookings with almost perfect accuracy.

**Recommended Strategies for Model Improvement:**

1. **Generalization**: Although high cross-validation scores indicate good generalization, evaluate the model's performance on a separate test set to confirm its ability to handle truly unseen data.
2. **Data Distribution**: Ensure that the data used for training, validation, and real-time predictions comes from the same distribution. Monitor for shifts in data, such as seasonal effects or changes in customer behavior, which could impact model performance.
3. **Feature Importance**: Regularly assess the stability of the features contributing to the model's predictions. If new features or data types emerge, be prepared to retrain the model accordingly.
4. **Model Complexity**: Prefer simpler models that are less prone to overfitting. If your model is overly complex, it may perform well on training data but struggle with new inputs.
5. **Latency and Scalability**: For real-time applications, evaluate the model's computational efficiency. Ensure that prediction times are acceptable for real-time needs.
6. **Evaluation Metrics**: In addition to accuracy, utilize metrics like precision, recall, and F1 score to gain a comprehensive understanding of model performance, especially in the context of imbalanced datasets.
7. **Regular Monitoring**: After deployment, continuously monitor the model’s performance. Implement tools to track accuracy and key metrics, and be ready to retrain the model as new data becomes available.

**8.2 Feature Importance**

The top 5 features influencing cancellations:

1. **Reservation Status** - **Importance: 0.6729**  
   The reservation status is the most significant factor influencing cancellations. Certain statuses are strongly associated with higher cancellation rates, indicating that the nature of the reservation plays a crucial role in predicting whether it will be canceled.
2. **Deposit Type** - **Importance: 0.0474**  
   The type of deposit made by customers also influences their likelihood of cancellation. This suggests that refundable and non-refundable deposits can affect customer behavior, with non-refundable options potentially reducing the chances of cancellation.
3. **Total of Special Requests** - **Importance: 0.0366**  
   The number of special requests made by guests correlates with cancellation likelihood. This may reflect guest satisfaction or specific needs, where a higher number of requests could indicate a more personalized experience that influences their decision to cancel.
4. **Lead Time** - **Importance: 0.0341**  
   Lead time, or the period between booking and arrival, affects cancellation rates. Generally, longer lead times may lead to higher cancellation rates, as earlier bookings are more susceptible to changes in plans.
5. **Reservation Status Month** - **Importance: 0.0187**  
   The month in which a reservation is made has a minor impact on cancellation behavior. This could be related to seasonal trends in booking patterns, affecting customer commitment based on the time of year.

**9 Operational Insights and Recommendations for Hotel Management**

1. **Seasonal Promotions**: Increase room rates or offer exclusive packages during high-demand summer months to maximize revenue.
2. **Off-Peak Incentives**: Attract customers during low-demand months (e.g., November to January) with discounts and bundled offers.
3. **Tailored Marketing Strategies**: Develop country-specific marketing campaigns based on seasonal trends and booking behaviors to enhance customer engagement.
4. **Targeted Marketing for City Hotels**: Focus marketing efforts on convenience and specialized packages for business travelers.
5. **Resort Hotel Packages**: Promote longer stay packages, wellness retreats, and family deals to attract more visitors to Resort Hotels.
6. **Flexible Booking Policies**: Implement flexible rebooking options or incentives for non-refundable bookings to reduce cancellations.
7. **Value-Added Services**: Offer complimentary breakfast, spa services, or recreational activities to enhance customer satisfaction, particularly in Resort Hotels.
8. **Direct Booking Enhancement**: Strengthen direct booking channels through loyalty programs or exclusive online offers to increase customer commitment.
9. **Corporate Partnerships**: Target local businesses for corporate bookings, potentially increasing occupancy rates in that segment.
10. **Monitoring and Adjusting Policies**: Continuously monitor cancellation rates and customer feedback to adjust policies as needed, focusing on flexibility to enhance satisfaction.
11. **Targeted Communication**: Engage with customers who have previously canceled to understand their motivations and encourage them to commit to bookings.
12. **Forecasting and Planning**: Use insights from booking trends to inform inventory management and demand forecasting, ensuring capacity meets customer demand without extensive waiting lists.

**10. Conclusion**

This project aimed to analyze and predict hotel cancellations using a comprehensive dataset, applying various statistical and machine learning techniques to derive actionable insights.

Through Exploratory Data Analysis (EDA), we uncovered significant patterns and trends in booking behaviors, customer demographics, and cancellation rates. Key findings included:

* A higher cancellation rate for city hotels compared to resort hotels, with transient customers showing the highest likelihood of cancellations.
* Seasonal trends in bookings and average daily rates (ADR) highlighted the need for tailored marketing strategies during peak months.
* The analysis revealed that customers who booked further in advance had a greater tendency to cancel, suggesting a potential area for targeted interventions.

To address the challenge of imbalanced data, we employed SMOTE, which successfully balanced the cancellation classes, enhancing the predictive performance of our models.

In terms of model performance, the Random Forest model outperformed Logistic Regression, achieving an average accuracy of 99.93%, with excellent precision and recall metrics. These results indicate the model's reliability in predicting cancellations, providing a robust tool for hotel management to mitigate potential revenue losses.

Based on our findings, we recommend that hotel management consider implementing dynamic pricing strategies, enhancing customer segmentation efforts, and focusing on improving customer engagement for transient bookings. Moreover, leveraging predictive analytics can facilitate better operational planning and resource allocation, ultimately improving profitability.

In summary, the insights and predictive capabilities derived from this project can empower hotel management to make informed, data-driven decisions, fostering enhanced customer experiences and optimizing operational efficiency.