Hotel Booking Cancellation Prediction

Decision Systems

Submitted by
Sankar Kumar Palaniappan

Contents / Agenda

- Data Dictionary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Conclusions and Recommendations
- Appendix

Data Dictionary

The data contains the different attributes of **customers' booking details**. The detailed data dictionary is given below:

- **Booking_ID:** the unique identifier of each booking
- **no_of_adults:** Number of adults
- **no_of_children:** Number of Children
- **no_of_weekend_nights:** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- **no_of_week_nights:** Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **required_car_parking_space:** Does the customer require a car parking space? (0 - No, 1- Yes)
- **room_type_reserved:** Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group

Data Dictionary

- **lead_time:** Number of days between the date of booking and the arrival date
- **arrival_year:** Year of arrival date
- **arrival_month:** Month of arrival date
- **arrival_date:** Date of the month
- **market_segment_type:** Market segment designation.
- **repeated_guest:** Is the customer a repeated guest? (0 - No, 1- Yes)
- **no_of_previous_cancellations:** Number of previous bookings that were canceled by the customer prior to the current booking
- **no_of_previous_bookings_not_canceled:** Number of previous bookings not canceled by the customer prior to the current booking

How to use this deck?

- This slide deck serves as a comprehensive template for your project submission
- Within this deck, you will come across various questions that are intended to test your ability to understand data visualizations, discover patterns / insights and postulate hypothesis. Think thoroughly and provide answers to these questions
- You are encouraged to modify this deck as required, by replacing the questions with suitable answers
- Please feel free to incorporate additional points if you deem necessary

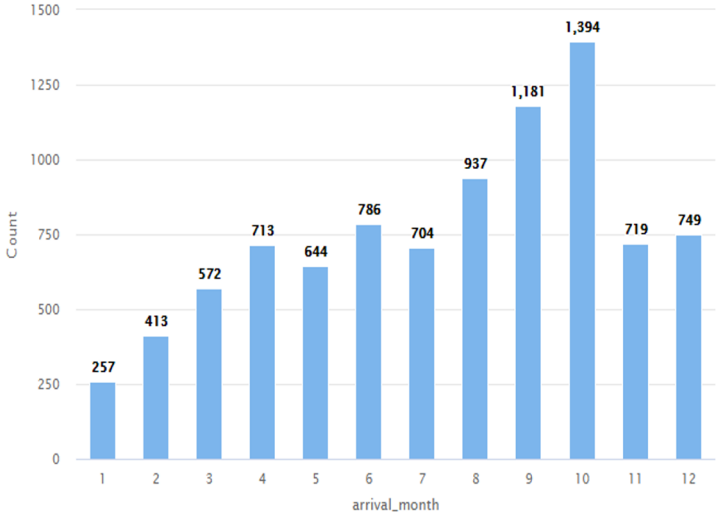
Note: The data visualizations you see in this deck are obtained from RapidMiner

Business Problem Overview and Solution Approach

The hospitality industry employs demand management strategies to navigate the unpredictable nature of consumer decisions, which directly influence revenue. Hotels typically maintain a fixed room inventory, but exceptions are made for 'blocked bookings' due to construction or conferences. To cater to diverse market demands from travel agencies, businesses, organizations, and individual consumers, the industry has established cancellation and no-show policies. However, overly strict penalties can make bookings less appealing, highlighting the delicate balance needed in these policies. Occupancy rates are pivotal for financial forecasting and revenue generation. Achieving the right occupancy level at the right time is crucial. When cancellations or no-shows surpass manageable levels, hotels face increased financial risks, shifting from consumers to the hotel itself, especially when it impacts profit margins negatively. Analytics plays an indispensable role in this scenario, examining various factors like market trends, profit margins, customer preferences, and the drivers behind decisions to cancel or no-show. Understanding these intricacies can help hotels optimize their strategies and mitigate risks. With the continuous growth of online travel bookings and digital transactions, the importance of data science in the hospitality industry is escalating. Utilizing data science can not only sustain but also boost profit margins and enhance customer satisfaction. Predictive analysis emerges as a valuable tool, with random forest modeling being particularly effective for addressing this complex business challenge. By leveraging these advanced analytics techniques, hotels can make informed decisions, improve customer retention, and optimize revenue streams in a competitive market landscape.

EDA - Univariate Analysis

- Discover effective strategies to manage and meet the rising demand in the hospitality industry.
- Utilize multivariate and bivariate analytics, along with pattern detection and trend analysis, to investigate the human and influential factors driving these fluctuations.
- With a staggering 500% increase from the first month to the tenth peak month followed by a 50% drop, understanding these variations requires thorough analytical exploration.
- It's crucial to identify correlations and coefficients of the most impactful influencers.
- To adapt to these varying demands, dynamic pricing, performance marketing, and AI-driven digital traffic strategies for consumer channels are essential.
- These approaches can help hotels optimize revenue, enhance marketing effectiveness, and cater to consumer preferences more effectively in a fluctuating market landscape.

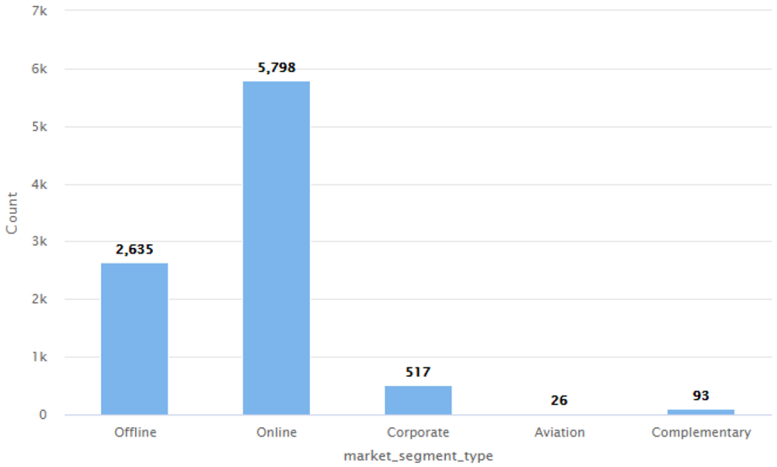


X-axis: Arrival month
Y-axis: Number of bookings
Description: Number of bookings per month

EDA - Univariate Analysis

Online bookings account for 64%, while offline bookings make up 29%, with corporate bookings comprising just over half a percent of other categories. These figures offer valuable insights into market preferences and indicate a digital-centric future. In the realm of online travel bookings, the scope extends beyond hotel rooms to include airline tickets, various advertising models, merchant models, subscription models, partnership initiatives, and peer-to-peer advertising networks. These diverse revenue streams and efficiency measures highlight the potential for enhanced digital services in booking, pricing, marketing, touring, and personalizing consumer experiences.

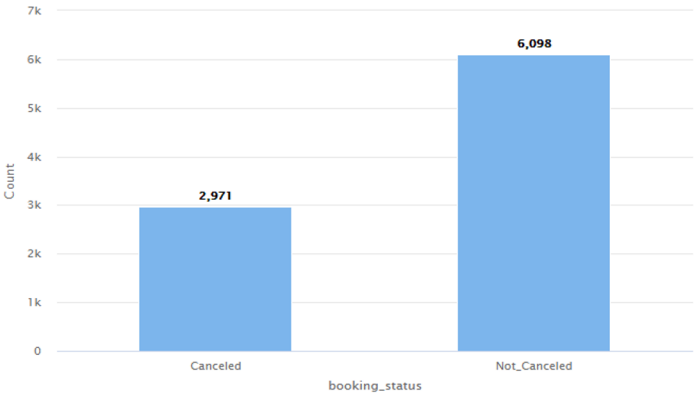
Data science can play a pivotal role in integrating traditionally offline segments. By leveraging insights gleaned from data analytics, hotels can improve service delivery, transportation options, expense accounting, and overall consumer satisfaction. Aligning innovation with data science has the potential to break down traditional offline silos, leading to greater consumer satisfaction and a more comprehensive understanding of consumer profiles. This alignment can pave the way for a more integrated and efficient approach to hospitality services in the digital age.



X-axis: Market Segment Type
Y-axis: Number of customers
Description: Number of Bookings received in each market segment

EDA - Univariate Analysis

- Marketing and sales teams strive to target the right customers, but they face challenges when booking cancellations reach 33%, which falls within the reported industry range. The revenue loss from these cancellations can be as high as 15% of net revenue or even more. Manual systems often result in a lack of integration among sales, marketing, customer service, and reservations, leading to missed insights and inaccurate revenue predictions.
- To address this issue, binary classification models that distinguish between canceled and non-canceled bookings are essential. Achieving balanced accuracy through simple confusion matrix analysis, benchmarking, hyperparameter tuning, and employing bivariate, multivariate, and logistic regression techniques on the feature variables presented can provide more accurate predictions.
- As the industry embraces digital booking and integrates with air travel, transportation, and touring services, it can mitigate risks. This shift transfers financial risks from hospitality providers to consumers while also creating opportunities for offering vouchers, tour packages, incentives, luxury experiences, and personalized services. This approach not only enhances revenue potential but also enriches consumer experiences, paving the way for a more resilient and customer-centric hospitality industry.



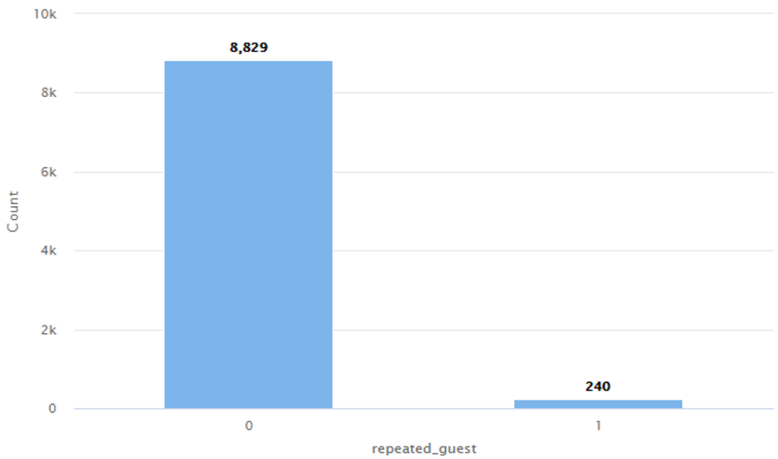
X-axis: Booking status

Y-axis: Number of bookings

Description: Count of bookings that are Cancelled v/s Not Cancelled

EDA - Univariate Analysis

- Improving data integrity is crucial, especially when channels categorize customers as 'new' or 'repeat' guests without considering their unique booking patterns. For instance, a guest switching from Booking.com to Agoda for a last-minute 50% discount might appear as a repeat guest due to non-unique profiling.
- A key reason guests don't return is often the lack of local marketing targeting local businesses, families, events, and weddings. Tailored marketing aimed at these local segments can significantly improve guest retention. Currently, 99.7% of guests have only single experiences, indicating a missed opportunity for repeat business.
- Collecting metrics on geographic location and restrictions is vital, especially to prevent illicit activities like sex and drug trafficking, which can sometimes be facilitated through bookings. To understand the contributing factors to non-returning customers, structured data alone may not be sufficient. Incorporating non-structured data and channels that interpret the consumer experience from booking to departure can provide deeper insights.
- Recognizing that consumers who experience flight delays often carry negative sentiments throughout their travel, it's crucial to address such issues promptly. Ensuring a refreshed and positive experience from booking to departure can make a significant difference in guest satisfaction and their likelihood to return.



X-axis: Repeated guests

Y-axis: Number of Guests

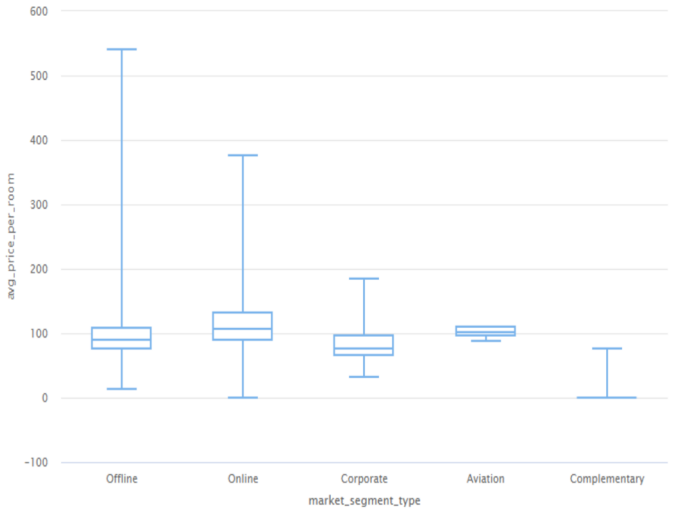
Description: Customers who have visited the hotel more than once vs customers who have visited the hotel only once

EDA - Univariate Analysis

- It's intriguing to note that despite significant variations, averages in some sectors, like offline and aviation, remain relatively consistent. Avoiding overpricing in offline bookings could potentially lead to higher average returns.

- Optimizing pricing through data science involves considering various key performance indicators (KPIs) such as forecasting, length of stay, guest type, and occupancy rates. Revenue can be predicted 40 days or more before check-in by factoring in both internal and external variables. Internal factors include historical prices, location, room sizes and types, competition, amenities, services, and holidays. External factors encompass competitor strategies, economic conditions, events, exchange rates, climate, and politics.

- For pricing optimization, regression analytics, decision trees, linear models, and recurrent neural networks are preferred methods. A price elasticity model focusing on occupancy rates serves as a foundational approach, with linear modeling as the base. Sequential additional modeling techniques can then be applied to enhance generalization. Recurrent neural networks are particularly effective for capturing temporal changes in rates over time and across seasons.



X-axis: Market segment type

Y-axis: Average price per room

Description: Price per room across various market segments

Results EDA - Univariate Analysis

Univariate analysis reveals several key insights:

- Bookings are not evenly distributed over time, with nearly 40% occurring between months 8-10, showing a staggering 500% increase compared to the base month.
- Online bookings account for 64%, aligning with the industry average, indicating the need for additional analytics to delve deeper into this trend.
- The industry cancellation rate stands at 33%. Further analysis using temporal trending KPIs and bivariate methods could provide more insights, particularly in understanding factorial influences.
- A significant 99.7% of guests book only once, suggesting the need for bivariate analysis considering location and other factors to explore rebooking opportunities.
- Price range and average present potential opportunities for optimization.

In summary, univariate analytics confirm industry knowledge without presenting surprises. Advanced analytics, exploratory analysis, and modeling are recommended for deeper insights and strategic planning.

EDA - Bivariate Analysis

- In terms of revenue from booking channels, complementary, aviation, and corporate bookings pose the least risk. Surprisingly, offline and online channels account for 69% and 64%, respectively. A z-score test may reveal that the booking channel isn't necessarily the primary factor influencing cancellations.

- Analyzing customer switching patterns across competitor markets could provide valuable insights. Factors such as flight cancellations, hurricanes, transportation issues, and disasters offer deeper insights into market segments. Bivariate analysis alone doesn't seem to pinpoint specific trends. Therefore, nonparametric tests like Wilcoxon-Kruskal-Wallis and chi-square tests are recommended for more comprehensive data collection.

- Further segmentation by location and market demographics, as well as psychological factors influencing travel choices, is essential to understand why travel occurs. Business-directed travel appears to be more consistent, suggesting a need for more detailed analysis of offline and online bookings to uncover underlying patterns.

X-axis: Market Segment type

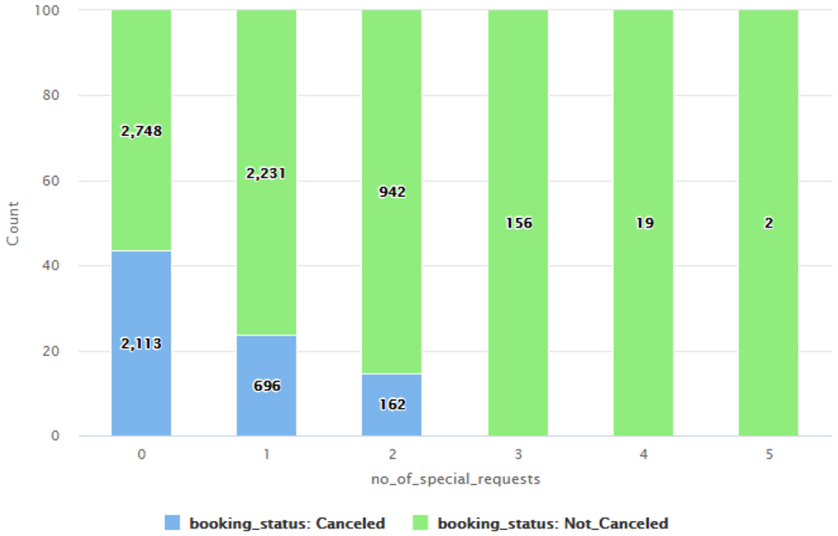
Y-axis: Number of customers

Description: Number of booking that are cancelled vs not cancelled across various market segments

EDA - Bivariate Analysis

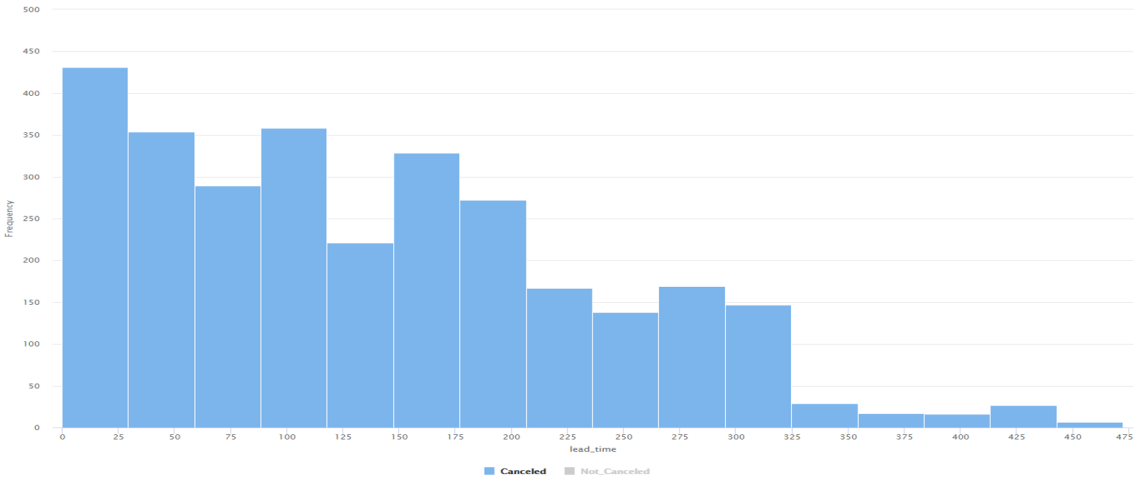
- Investigate the impact of special requests on booking cancellations and develop strategies to reduce cancellation rates associated with such requests.

- Conduct bivariate t-tests and Pearson's r correlations to assess the relationship between special requests and booking cancellations. Alternatively, Pearson's r correlation or chi-square tests can be used for inference and hypothesis testing, aiming to reject the null hypothesis of a 5% chance that no relationship exists. While the sample size might be too small for some tests, a significant chi-square value of 45.1722 and a corresponding P-value for cancellations without requests suggest that aligning touring options with bookings could mitigate risks by allowing for more personalized services, thereby shifting responsibility to the consumer.



EDA - Bivariate Analysis

- What insights can we derive from the lead time of canceled bookings, and how can we use this data to enhance our booking system and boost customer satisfaction? It's noteworthy that the histograms exhibit a right skew, particularly in the third histogram. Implementing predictive modeling by quarter or offering incentives based on booking duration could potentially stabilize risks and optimize profits.



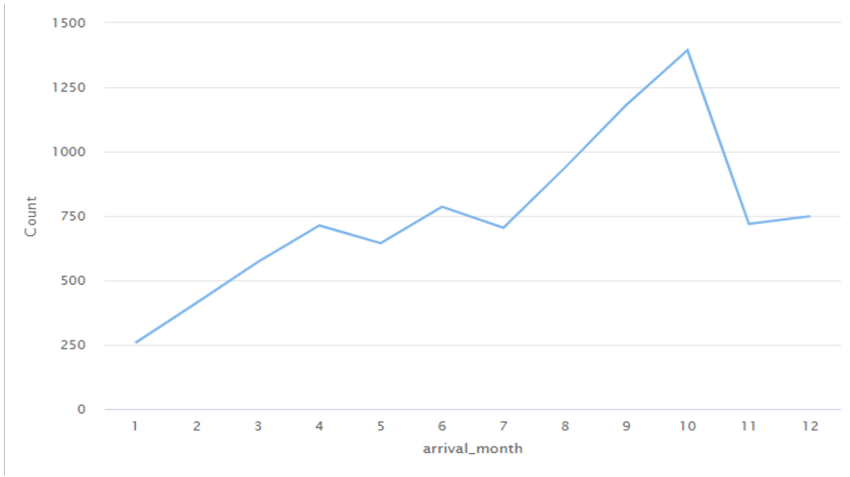
X-axis: Lead time

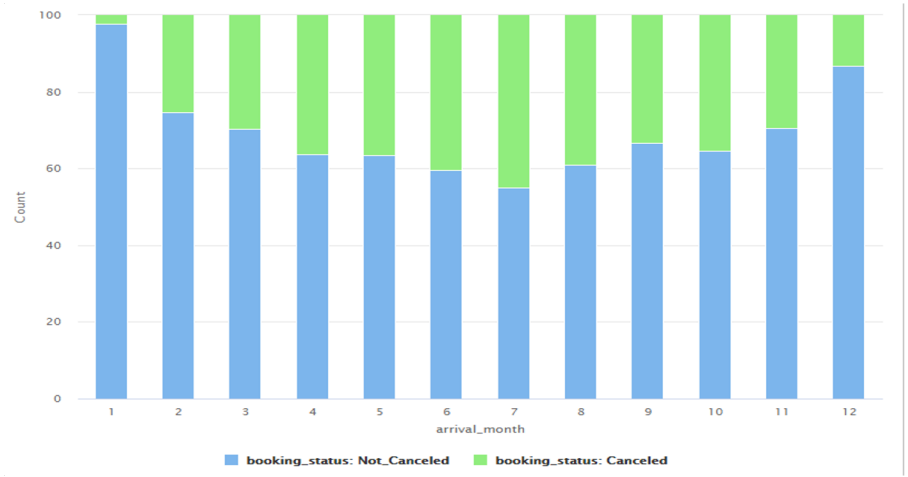
Y-axis: Number of bookings

Description: Lead time of the bookings that were cancelled

EDA - Bivariate Analysis

The complete dataset for a thorough analysis is unavailable, but a sequential increase in seven data points above the mean suggests a Shewhart's cycle pattern. This pattern hints at a potential correlation between an increased count and a higher cancellation rate. Interestingly, when the count surges between 7-10, the cancellation rate remains relatively steady. Conversely, during a declining trend, there seems to be a push towards reduced cancellations. This suggests an upper limit on the standard deviation, indicating that the data follows typical variance patterns.





Results EDA - Bivariate Analysis

Bivariate analysis reveals correlations between various factors:

- There's a correlation between the number of bookings and cancellation rates.
- Special requests and booking channels are directly correlated.
- Cancellations are closely related to booking counts, with special requests influencing booking preferences.

The frequency of these occurrences highlights opportunities to profile clients and implement new policies. One such policy could restrict further bookings for no-shows and cancellations, limiting rebooking to one or two chances. This approach aims to impose controls on consumer behavior, as current policies seem to lack such constraints, leading to unlimited cancellation and rebooking opportunities.

Model Performance Evaluation Decision Tree

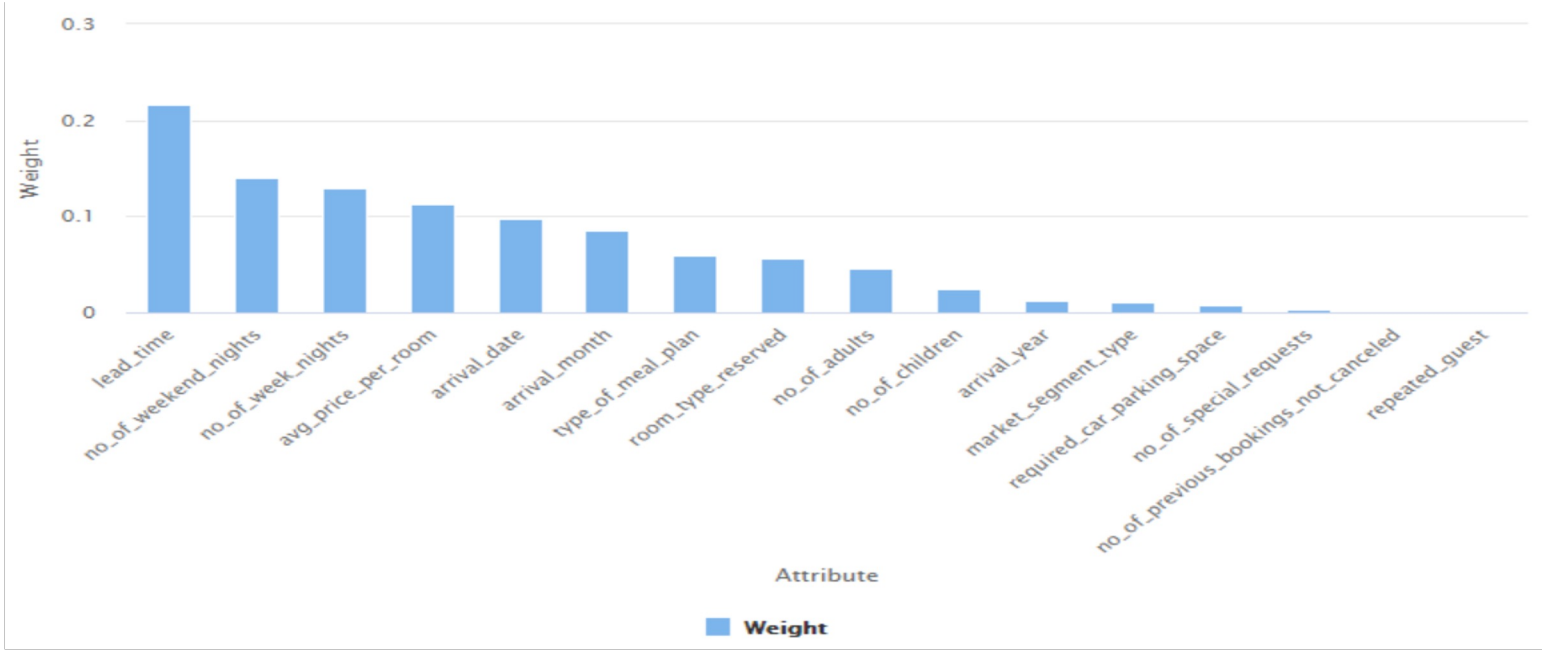
Pruning is essential for refining the model.

- Accuracy in machine learning is generally good in this instance.
- Recall measures the model's ability to identify all relevant items, and it performs better on the training set, indicating potential overfitting.
- Precision focuses on correctly predicting the target class, with the test set showing signs of overfitting.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	99.65	83.20	99.47	76.77	99.47	73.23

Model Performance Evaluation Decision Tree

- To what extent and in what ways have the attribute weights influenced the overall performance of the model?



Model Performance Evaluation Pruned Decision Tree

- Based on the evaluation metrics obtained from a pruned decision tree, do you perceive any notable enhancements in performance? What factors do you believe contributed to this improved performance compared to the previous version?

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	99.65	83.20	99.47	76.77	99.47	73.23
Decision Tree - Pruned	95.64	84.34	90.77	76.09	95.69	76.09

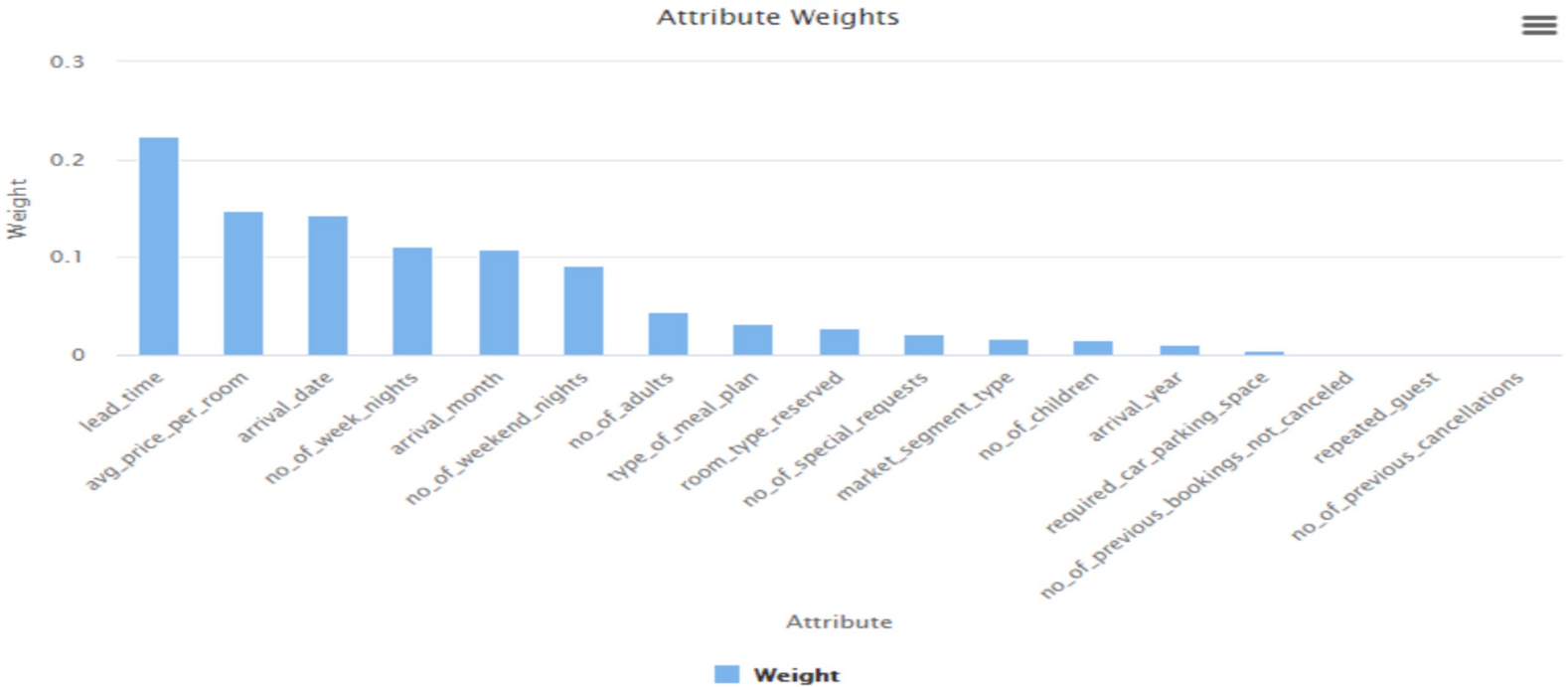
Model Performance Evaluation Random Forest

- While the Random Forest model demonstrates high accuracy and precision on both the training and test sets, it appears to have a noticeable drop in recall on the test set compared to the training set. What potential factors could contribute to this difference in recall, and how might it impact the model's overall performance and practical applications?

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest	98.47	87.94	96.44	77.10	98.87	84.71

Model Performance Evaluation Random Forest

- To what extent and in what ways have the attribute weights influenced the overall performance of the model?



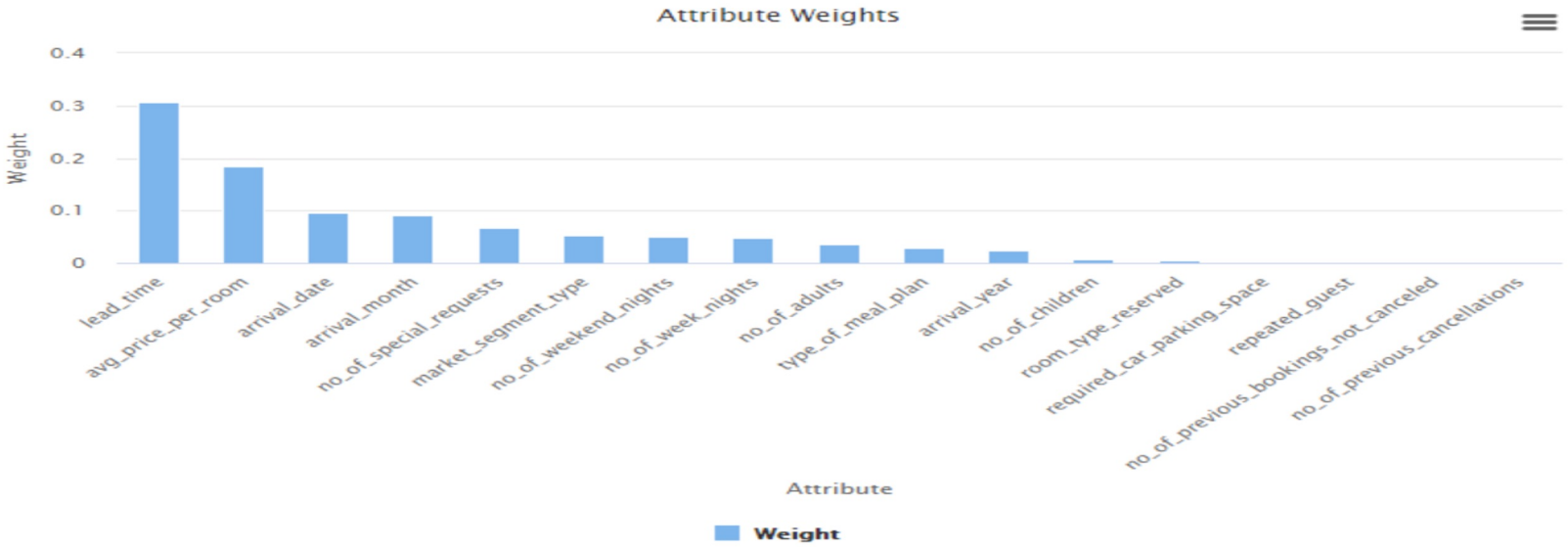
Model Performance Evaluation Pruned Random Forest

- Based on the evaluation metrics obtained from a pruned Random Forest, do you perceive any notable enhancements in performance? What factors do you believe contributed to this improved performance compared to the previous version?

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest	98.47	87.94	96.44	77.10	98.87	84.71
Random Forest - Pruned	87.32	85.04	69.86	66.22	89.09	84.77

Model Performance Evaluation Pruned Random Forest

- To what extent and in what ways have the attribute weights influenced the overall performance of the model?



Model Performance Summary

- Description of the ML model that best fits the business objective. Also state which evaluation metric (such as accuracy, recall or precision) is important to achieve the business objective and why?
- Synopsis of the key features employed by the ML model to make predictions.
- Offer suggestions and advice for the hotel based on the gathered insights.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	99.65	83.20	99.47	76.77	99.47	73.23
Decision Tree - Pruned	95.64	84.34	90.77	76.09	95.69	76.09
Random Forest	98.47	87.94	96.44	77.10	98.87	84.71
Random Forest - Pruned	87.32	85.04	69.86	66.22	89.09	84.77

Conclusions and Recommendations

- Please mention actionable insights & recommendations

Here are some helpful guidelines to consider:

- Define the hotel's target audience and understand booking patterns.
- Examine booking trends and cancellation rates over time.
- Assess the financial and operational impact of cancellations on hotel revenue.
- Review the consistency of cancellation policies across different market segments.
- Explore ways to enhance the experience for returning customers and attract new ones.

APPENDIX

Great Learning

Happy Learning !