Hotel Bookings-A data driven approach to maximizing profitability

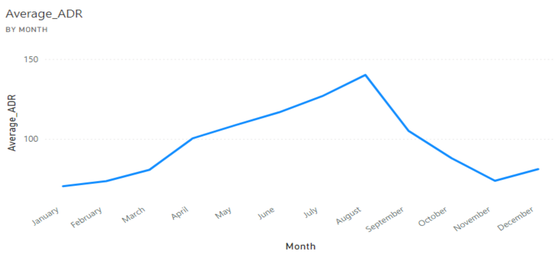
Tyler Trocchi, Natascha Banda, Nana Amoah
GSU J. Mack Robinson College of Business
Institute for Insight, Fall 2024 Sprint Project

Problem Statement

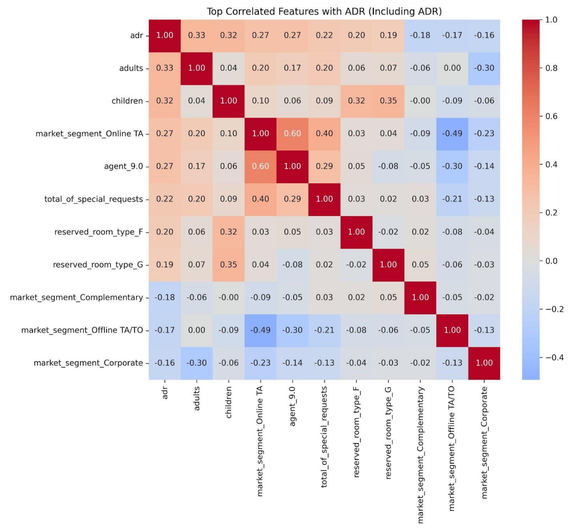
The project focuses on optimizing hotel booking demand by addressing two key areas:
Maximizing the Average Daily Rate (ADR): Identifying features influencing ADR to enhance profitability.
Cancellation Analysis: Understanding patterns and predicting cancellations to mitigate revenue loss.

Exploratory Data Analysis

One of our initial goals was the find some patterns in terms of the sales/bookings data as the year progresses. The visualization below demonstrates that the hotel is able to charge a much higher rate during the peak summer months than any other time in the year.



Another part of maximizing the average daily rate was analyzing which features could correlate most strongly to an increasing average daily rate. This correlation coefficient was creating using the non-cancelled bookings, filtering out some outliers and creating dummy variables for the categorical features.



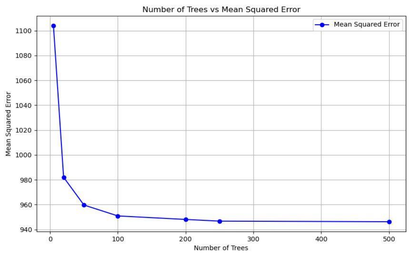
Methodology

ADR Prediction Model

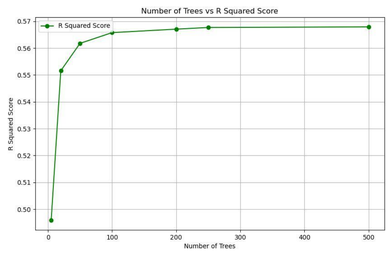
We determined that using a random forest regressor model would be best in order to predict the ADR based on the following features. We used metrics such as the Mean Squared Error and R Squared Score since this is a regressor model.

Based on these results, we can see that the model performs best when 100 trees are used. From 5 to 20 to 50 trees, the model improves at a drastic rate, but once we move from 100 to 200 trees, the rate of improvement drop off significantly.

Mean Squared Error vs Trees



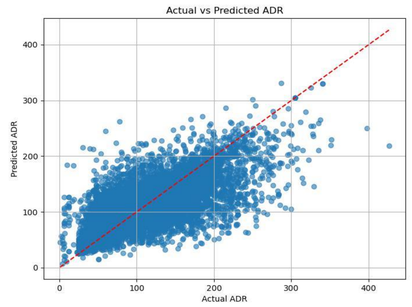
R Squared Score vs Trees



Model Evaluation

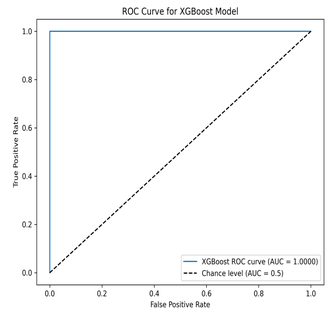
ADR Prediction Model

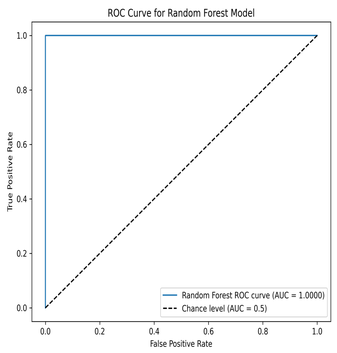
The graph below visualizes the ADR prediction model. The red line is a perfect prediction line, and the data points are the predicted values. This model predicted the ADR with great accuracy when the ADR was low, but struggled with the outliers as the values grew.



Cancellation Prediction Model

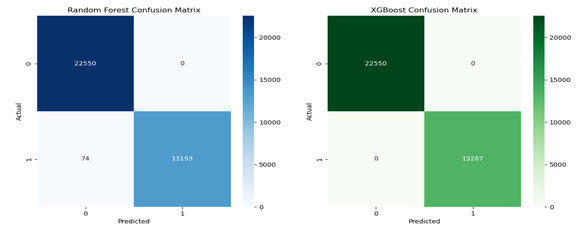
The graphs below demonstrate the false positive and true negative rate of the 2 cancellation prediction models, the Random Forest classifier and the XGBoost classifier.





Results

Cancellation Prediction Model Confusion Matrix



These confusion matrices depict how well the cancellation predictive models performed. With the blue matrix representing the random forest model and the green matrix representing the XGBoost model, we can see that both models have performed exceptionally well at predicting cancellations.

Evaluation Statistics

1. ADR Prediction:
- Mean Squared Error: 946.29
- R Squared Score: 0.568

2. Cancellation Prediction:
- Accuracy (99.45%): The model is highly accurate, correctly predicting cancellations and non-cancellations for most bookings.
- Precision (100%): Precision is perfect, indicating that every prediction of "Canceled" was correct.
- Recall (99.44%): The model successfully identified almost all actual cancellations, showing a strong ability to detect cancellations.
- F1-Score (99.72%): The high F1-score suggests a good balance between precision and recall.

Conclusion

- **Maximizing ADR**: While the model predicts ADR accurately for most cases, additional cost-related data could refine profitability optimization.
- **Cancellations**: Effective models can predict cancellations, aiding in resource planning and reducing revenue losses.