Machine Learning Insights Into Hotel Booking Cancellation Trends

Lynni Do

Mohini Gupta

Chai Zhang

1. Abstract

This is a supervised learning project with the main objective of classification, where we aim to predict whether a hotel booking will be canceled, and we aim to understand pricing patterns and their potential association with cancellations.

We first ran a decision tree model to predict whether a hotel booking will be canceled based on several factors. Key factors include the type of deposit (non-refundable deposits increase cancellation risk), booking lead time (shorter lead times reduce cancellations), country of origin (certain countries have higher cancellation rates), car parking requirements (higher needs might influence cancellations), and the customer's past booking behavior (previous cancellations and non-cancellations impact future actions). The model uses these factors to classify bookings as either likely to be canceled or not.

We then ran a random forest model with all the variables to see which had the greatest influence on cancellation of hotel bookings. Random forest trees can be seen as an extended and more robust version of the decision tree mode since it combines multiple decision trees and trains the data based on the subset of data. Thus, we can expect to see similar results between the two. After running the model, it showcased the most important features which could highlight the most significant predictors for hotel cancellation as deposit_type, country, and lead_time.

We ran a linear regression model with adr (average daily rate) with the following variables lead_time, adults, children, previous_cancellations, total_of_special_requests, and

is_canceled. The linear model seeked an answer to whether price was a factor that played into hotel booking cancellations to find out which types of bookings were the most expensive. The variables to be run against adr in the model were selected using a LASSO analysis. Through the linear model, we discovered that canceled bookings tend to be more expensive than non-canceled ones, and that the most expensive type of bookings were canceled ones with adults, children, and special requests.

2. Introduction

This is a supervised learning project with the main objective of classification, where we aim to predict whether a booking will be canceled. Our goal was to use the hotel booking dataset (sourced from Kaggle) to predict whether a booking will be canceled. The main clientele for our analysis are hotel management and booking sites. We wanted to identify and analyze the key factors that drive booking cancellations, in order to uncover patterns and trends that could help with decision-making within the hotel industry. In particular, we want our stakeholders to be aware of which bookings are at greatest risk for cancellations so that they can take preventative steps to reduce the overall number of canceled bookings.

The main variable used in our analysis was is_canceled, a variable that held a binary value for whether a booking was canceled (0 for non-canceled bookings and 1 for cancellations). The variables chosen to predict booking outcomes varied depending on the type of model used to analyze the data. We used three different models and algorithms for our analysis: decision tree, random forest, and linear regression. Decision tree and random forest models find the features most important for predicting the target variable. For our purposes, we wanted to find the factors most important for predicting the cancellation of a booking. Even though both models are very similar (with random forest being an improved version of decision trees), we wanted to use both models so that we could validate both models. Linear regression was used to explore how the numerical variable ADR (average daily rate) relates to other features. With linear regression, we were able to understand pricing patterns and their association with cancellations. These models were built to help uncover which bookings would be most likely to be canceled by analysing how the predictor variables affected the explanatory variables.

The decision tree model evaluates booking data to classify outcomes, either cancellation or non-cancellation, based on features like deposit type, lead time, country, and market segment. Key splits include the deposit type, with "Non Refund" bookings more likely to result in class 1 (cancellations), especially for specific countries like Türkiye and Germany when certain conditions (e.g., low previous bookings not canceled) are met. Conversely, bookings with refundable deposits and short lead times or fewer previous cancellations often lead to class 0 (non-cancellations).

Through random forest, a machine learning technique, we can identify the most influential feature which contributes to cancellation by testing all the variables. However, for the model we drop certain variables such as arrival_date, reservation_status, is_canceled, and reservation_status_date because these variables can give a direct answer to our model when they run the test. If we were to leave these variables in, it would skew our findings. Through running the random forest model, we were able to see that hotels with deposit types which were marked as 'Non Refund' were the most significant predictors of cancellation. This could be because if there is a non refund policy, individuals are more likely to stick with their original booking since they would not be getting their money back if they were to cancel. The second most significant predictor was the country Portugal, and lastly the third most significant predictor of hotel cancellation was if there was a longer lead_time.

Linear regression models weight explanatory variables in order to predict the outcome variable. In our linear regression model, the outcome variable was the average daily rate, and our explanatory variables were lead_time, adults, children, previous_cancellations, total_of_special_requests, and is_canceled. Other variables from the dataset were dropped because according to a LASSO analysis, they were not significant enough in predicting the average daily rate. We wanted to use linear regression to understand differences in pricing for bookings that were and weren't cancelled, as well as how the other explanatory variables influence pricing and cancellations. After creating the linear regression model, we discovered that cancelled bookings tended to be more expensive than non-cancelled bookings. We also found that the average daily rate for the least expensive type of canceled booking was still more expensive than the most simple type of non-canceled booking. Our findings lead us to recommend that stakeholders take price into account when identifying bookings at risk of cancellation.

3. Data

This study aims to predict hotel booking cancellations using a dataset sourced from Kaggle. The primary goal is to identify and analyze factors contributing to cancellations, providing insights into patterns and trends that can enhance decision-making in the hotel industry. The dataset comprises 36 variables describing various aspects of hotel bookings, such as booking details, customer demographics, and reservation outcomes. It includes data for two types of hotels ('Resort Hotel' and 'City Hotel') and spans the years 2015 to 2017. While some variables are self-explanatory, others are abbreviated (e.g., 'AUT' representing Austria in the 'country' column). To address this, interpretations were inferred based on context. Ambiguous or unclear variables, as well as certain data anomalies, were excluded to ensure a focused analysis. To ensure data reliability, rows with missing or unclear values for key attributes were removed (e.g., undefined meal plans, ambiguous country codes). Columns without interpretable information such as 'agent', 'company', and personal identifiers (e.g. 'name', 'email') were excluded. Outliers, such as negative ‘adr’ values, were omitted to focus on realistic business scenarios.

4. Methods

This project employs supervised learning to achieve both classification and regression objectives. For the classification task, the goal is to predict whether a hotel booking will be canceled, with a binary outcome where 1 represents cancellation and 0 represents non-cancellation. In addition, regression analysis will be used to explore how the Average Daily Rate (ADR) relates to other features, such as lead time and customer type. This analysis aims to uncover pricing patterns and investigate their potential connection to booking cancellations.

The project will utilize three models. Decision trees will help identify key variables influencing cancellations and provide a clear visualization of decision pathways leading to cancellation versus non-cancellation outcomes. Random forests will enhance predictive accuracy and reduce overfitting by training multiple decision trees on random subsets of the data, with final predictions determined by averaging or majority voting across the trees. Linear regression will be employed to analyze how ADR is influenced by features like lead time, customer type, and booking status (is_canceled). This approach will help uncover patterns in pricing, such as whether canceled bookings are associated with lower ADR, potentially reflecting trends like last-minute or discounted bookings being more prone to cancellation.

To ensure model validity, the dataset will be split into training and testing sets using an 80/20 ratio. Classification models will be evaluated using metrics such as accuracy and confusion matrix results to identify the most impactful predictors. Regression models will be assessed using R² and RMSE metrics, with the aim of achieving values indicative of strong predictive performance, such as R² values close to 1 and low RMSE values.

One challenge with linear regression models is the topic of variable selection. The dataset had several variables that could have been explanatory variables, and arbitrarily picking picking variables could have led to a biased or ineffective model. To combat this, we decided to choose our variables using LASSO. This would make sure the model is built using variables that are effective at explaining differences in the average daily rate. Additional challenges may come from external factors not captured in the dataset, such as cancellations driven by better deals elsewhere, may affect model accuracy and are outside the scope of control.

Feature engineering will play a key role in preparing the dataset for analysis. Categorical variables like meal type will be transformed into binary columns using one-hot encoding.

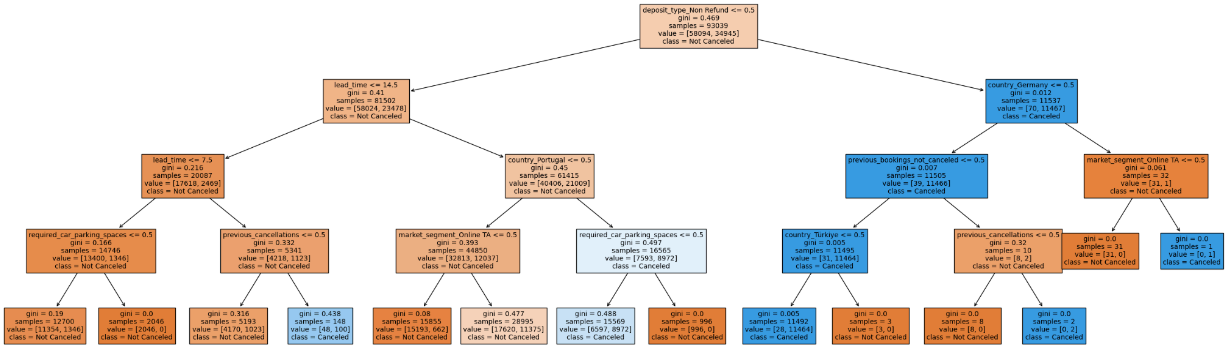
The results of this analysis will be communicated using multiple approaches. Classification outcomes will be assessed through confusion matrices, which will demonstrate the models' ability to differentiate between canceled and non-canceled bookings. For regression models, coefficients and intercepts will be interpreted to reveal how factors such as lead time and customer type influence ADR and its relationship with cancellation likelihood.

5. Results

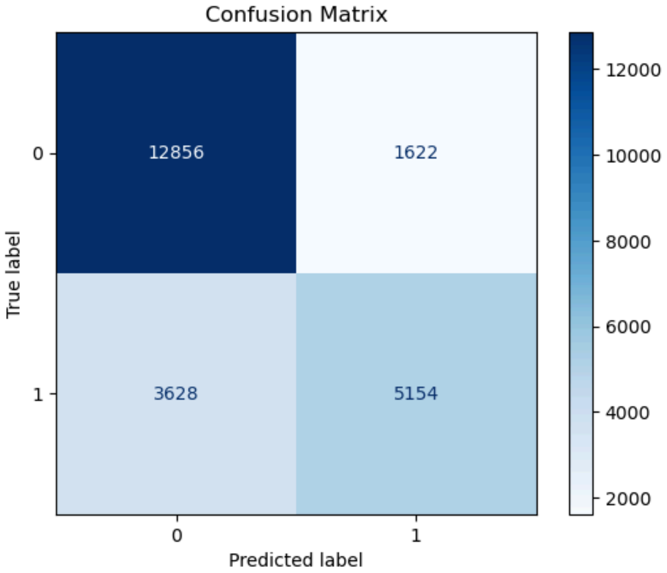
Decision Tree Model

To create the decision tree model, the dataset was preprocessed by handling missing values, dropping irrelevant columns, and encoding categorical variables using one-hot encoding. The data was then split into training and testing sets (80-20 split), and a Decision Tree Classifier with a maximum depth of 4 was trained on the training set. After training, the model was evaluated on the test set, resulting in an accuracy of approximately 0.77429. The model's structure was visualized, and key metrics like the classification report and confusion matrix were analyzed.

The resulting decision tree has a depth of 4 and contains 14 leaves. The tree's primary splits are based on the ‘deposit_type_Non Refund’ feature, followed by ‘lead_time’, ‘previous_cancellations’, ‘country_Portugal’, ‘market_segment_Online TA’, and others. The most important insights from the tree include the impact of non-refundable deposits, short lead times, and previous cancellations on the likelihood of cancellation. For example, non-refundable deposits significantly reduce the probability of cancellation, and previous cancellations increase it.


deposit_type_Non Refund <= 0.5 gini = 0.469 samples = 93039 value = [58094, 34945] class = Not Canceled
lead_time <= 14.5 gini = 0.41 samples = 81502 value = [58024, 23478] class = Not Canceled
country_Germany <= 0.5 gini = 0.012 samples = 11537 value = [70, 11467] class = Canceled
lead_time <= 7.5 gini = 0.216 samples = 20087 value = [17618, 2469] class = Not Canceled
country_Portugal <= 0.5 gini = 0.45 samples = 61415 value = [40406, 21009] class = Not Canceled
previous_bookings_not_canceled <= 0.5 gini = 0.007 samples = 11505 value = [39, 11466] class = Canceled
market_segment_Online TA <= 0.5 gini = 0.061 samples = 32 value = [31, 1] class = Not Canceled
required_car_parking_spaces <= 0.5 gini = 0.166 samples = 14746 value = [13400, 1346] class = Not Canceled
previous_cancellations <= 0.5 gini = 0.332 samples = 5341 value = [4218, 1123] class = Not Canceled
market_segment_Online TA <= 0.5 gini = 0.393 samples = 44850 value = [32813, 12037] class = Not Canceled
required_car_parking_spaces <= 0.5 gini = 0.497 samples = 16565 value = [7593, 8972] class = Canceled
country_Türkiye <= 0.5 gini = 0.005 samples = 11495 value = [31, 11464] class = Canceled
previous_cancellations <= 0.5 gini = 0.32 samples = 10 value = [8, 2] class = Not Canceled
gini = 0.0 samples = 31 value = [31, 0] class = Not Canceled
gini = 0.0 samples = 1 value = [0, 1] class = Canceled
gini = 0.19 samples = 12700 value = [11354, 1346] class = Not Canceled
gini = 0.0 samples = 2046 value = [2046, 0] class = Not Canceled
gini = 0.316 samples = 5193 value = [4170, 1023] class = Not Canceled
gini = 0.438 samples = 148 value = [48, 100] class = Canceled
gini = 0.08 samples = 15855 value = [15193, 662] class = Not Canceled
gini = 0.477 samples = 28995 value = [17620, 11375] class = Not Canceled
gini = 0.488 samples = 15569 value = [6597, 8972] class = Canceled
gini = 0.0 samples = 996 value = [996, 0] class = Not Canceled
gini = 0.005 samples = 11492 value = [28, 11464] class = Canceled
gini = 0.0 samples = 3 value = [3, 0] class = Not Canceled
gini = 0.0 samples = 8 value = [8, 0] class = Not Canceled
gini = 0.0 samples = 2 value = [0, 2] class = Canceled

The model achieved an accuracy of 77.4%, but its performance can be further analyzed using the confusion matrix:



Confusion Matrix Breakdown:

- True Negatives (TN): 12,856 - The model correctly identified 12,856 instances where cancellations did not occur.
- False Positives (FP): 1,622 - The model incorrectly predicted 1,622 cancellations when there were none.
- False Negatives (FN): 3,628 - The model failed to predict 3,628 cancellations that did occur.
- True Positives (TP): 5,154 - The model correctly predicted 5,154 cancellations.

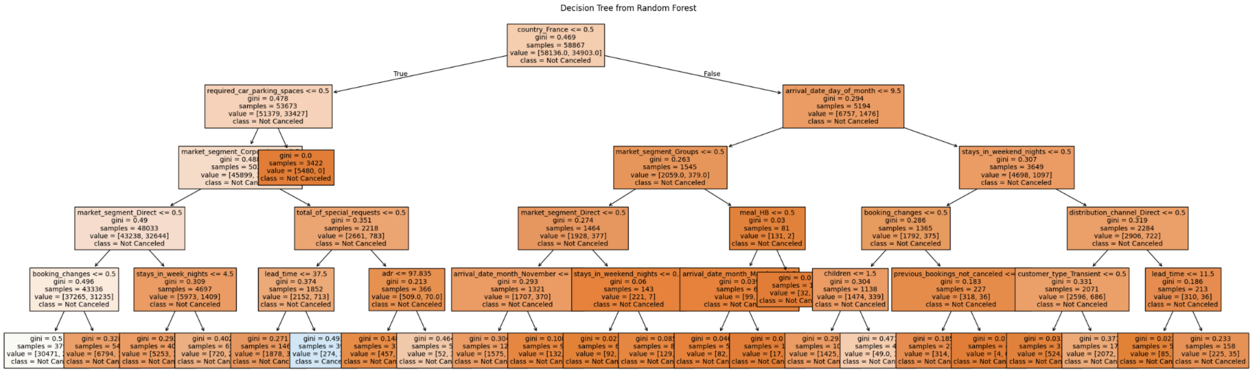
Using these values, we can calculate precision and recall, which provide more insight into the model's performance: Precision indicates the proportion of correct positive predictions out of all positive predictions. A precision of 0.7606 means that when the model predicts a cancellation, it is correct about 76.06% of the time. Recall measures the proportion of actual positives correctly identified by the model. A recall of 0.5869 suggests that the model captures about 58.69% of all actual cancellations.

These values indicate that the model has a moderate rate of false positives and false negatives. To improve the model, future work could incorporate additional features or feature engineering. Additionally, addressing class imbalance through techniques like oversampling, undersampling, or using more sophisticated metrics could enhance the model's predictive performance.

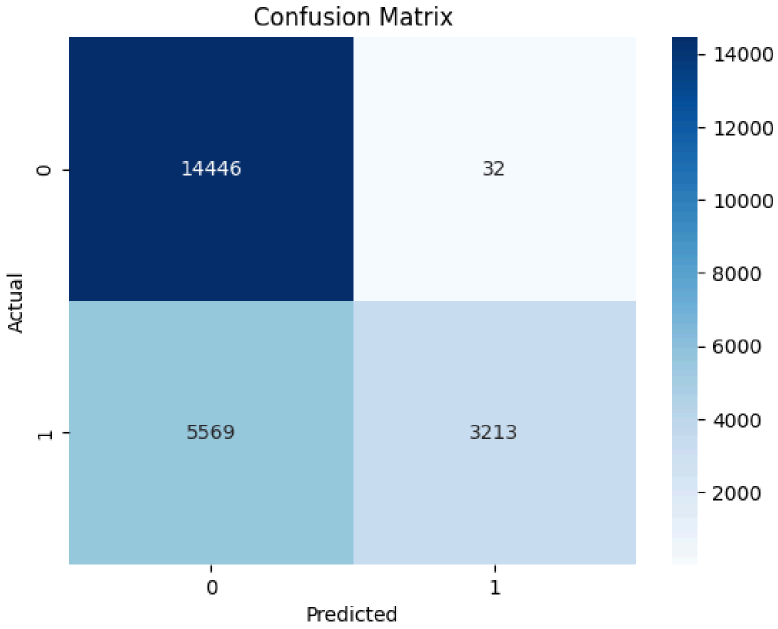
Random Forest Model

In order to construct the random tree model, we can prepare the dataset the same way we did for the decision tree model since the random tree model is a more robust version. We can encode the categorical variables the same way, as well as split the training and testing set in the 80-20 split. After the model was trained, there was an accuracy of about 0.7592 (75.92%). The model determined that the top five most influential features was deposit_type_Non Refund (0.236), country_Portugal (0.134), lead_time (0.091), previous_cancellations (0.061), and total_of_special_requests (0.058).

The results of the random forest tree had a depth of 5 and contained 22 leaves. When examining the left subtree, we can see that if the lead_time <= 37.5 days, there is a higher likelihood of the booking not being canceled. On the flip side, if there is a lead time greater than 37.5 days then it is associated with a higher likelihood of cancellation. Moreover, when examining the right subtree, we can see if the lead_time <= 11.5 days, the bookings are more likely to not be cancelled. However, when lead time exceeds 11.5 days, there is a stronger chance of cancellation. This essentially can give insight on how shorter lead times can be associated with a lower likelihood of cancellation compared to longer lead times.



Overall, the model achieved an accuracy of 75.92%. We can analyze the performance some more when looking at the confusion matrix:



Confusion Matrix Breakdown:

- True Negatives (TN): 14,446 - The model correctly identified 14,446 instances where cancellations did not occur.
- False Positives (FP): 32 - The model incorrectly predicted 32 cancellations when there were none.

- False Negatives (FN): 5,569 - The model failed to predict 5,569 cancellations that did occur.
- True Positives (TP): 3,213 - The model correctly predicted 3,213 cancellations.

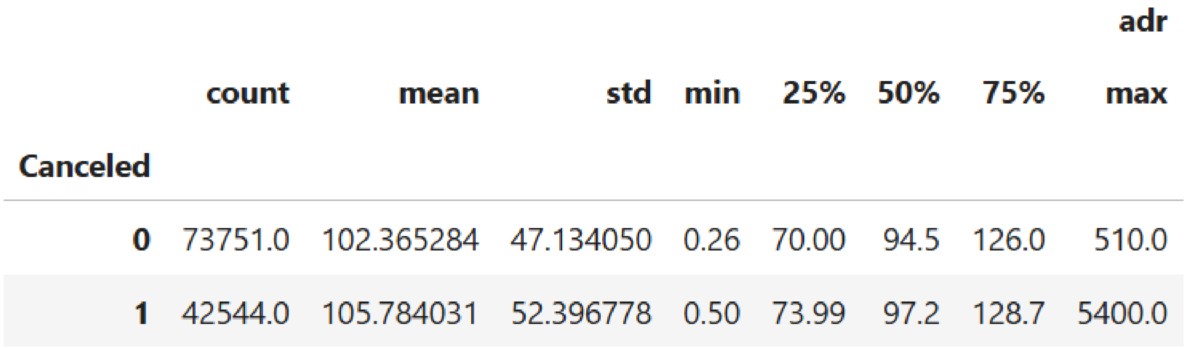
Through the confusion matrix breakdown, we can calculate the precision and recall as well. The precision for this model is 0.99 which shows us how many of the predicted canceled bookings were canceled. This is a very high precision, which can highlight how there were very few false positives. Moreover, the recall was 0.34. This is a low recall which can suggest that there are some positive cases that are overlooked. Such as some actual cancellations which are not being detected.

In order to improve the model, in the future we could increase sample size for the canceled class if possible. By increasing the sampling size we can help the model to learn patterns better. Another method we could do to improve the model is to test different parameters of the random forest. By testing different maximum depth, minimum sample leafs, or number of trees it could produce different outputs and possibly improve the recall value.

Linear Regression Model

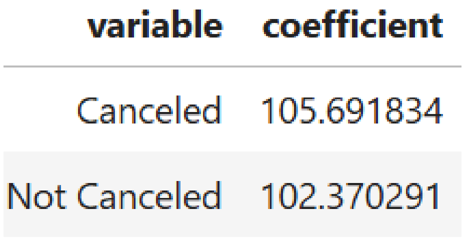
The linear regression model was created with the aim of answering the following questions: is price a factor that plays into hotel booking cancellations, and which types of bookings are the most expensive? After selecting model features with LASSO, two models were created to answer each question.

The first model was a simplified linear regression model with average daily rate being predicted by the canceled status of a booking (0 for not canceled, 1 for canceled). Preliminary data analysis found that the mean value and all percentiles values were greater for canceled bookings.



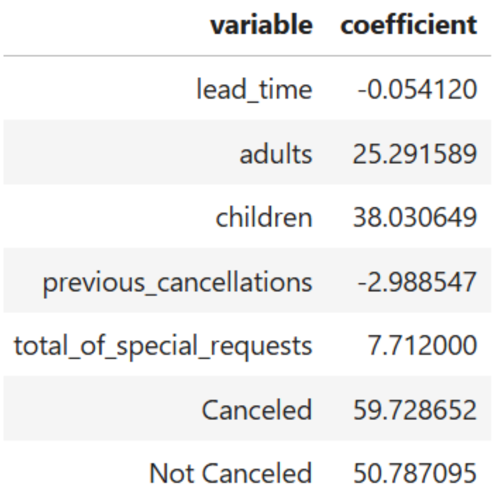
	adr							
	count	mean	std	min	25%	50%	75%	max
Canceled								
0	73751.0	102.365284	47.134050	0.26	70.00	94.5	126.0	510.0
1	42544.0	105.784031	52.396778	0.50	73.99	97.2	128.7	5400.0

These results were validated by the linear regression model.



variable	coefficient
Canceled	105.691834
Not Canceled	102.370291

From the results, we were able to answer the first question, is price a factor that plays into hotel cancellation? Based on the results from the simplified model, it appears that canceled bookings are higher on average, which could be a factor playing into their cancellations. To confirm this we decided to create an expanded model with the other explanatory variables chosen using LASSO.



variable	coefficient
lead_time	-0.054120
adults	25.291589
children	38.030649
previous_cancellations	-2.988547
total_of_special_requests	7.712000
Canceled	59.728652
Not Canceled	50.787095

The expanded model confirmed our findings from the simple model. Canceled bookings are more expensive than non-canceled ones, and the most expensive type of bookings are canceled bookings with multiple adults, children, and special requests. Interestingly, one of the cheapest types of canceled booking that can be constructed from this model is a booking for 1 adult, 14 days lead time, and previous_cancellations being true (59.73 + 25.29 - 0.054(14) -2.99 = ~$81.27) and that booking is still more expensive than the average rate for the simplest type of non canceled booking for 1 adult (50.79+25.29 = $76.08).

For selection of the best model, the R^2 and RMSE values were compared for both models. The R^2 and RMSE values for the simplified model were 0.0015 and 102.371 respectively, and for the expanded model, they were 0.231 and 41 respectively. R^2 values close to 1 and low RMSE values are preferred, so the expanded model is the better predictive model. One thing to note is that even though the R^2 value improved significantly from the simplified to the expanded model, it still is not very close to 1. An R^2 value extremely close to 1 could be due to overfitting, but with the current preferred model, there is still some room for improvement in terms of model fit.

6. Conclusion:

Decision Tree Model

The decision tree model performed reasonably well but has room for improvement. The precision value indicates that false positives are moderate, and the recall value shows that false negatives are relatively high. Here are some ways to we believe can enhance the model:

1. Handling class imbalance: The current model might be affected by class imbalance. Techniques like oversampling the minority class, undersampling the majority class, or using algorithms designed for imbalanced data could be beneficial.

2. Feature engineering: Creating new features based on existing ones (e.g., booking season) might capture more complex relationships and improve model accuracy.

By addressing these areas, the decision tree model can be made more robust and accurate, better predicting hotel cancellations and providing valuable insights for operational planning and customer management.

Random Forest Model

Overall, the random forest tree is effective at identifying key factors which can influence hotel booking cancellation. The analysis underlines how non-refundable deposits are the strongest predictors of cancellation, followed by geographic origin as Portugal and longer lead times before booking. While the random forest tree had a very high precision, it had a very low recall which can be improved upon. Specifically, we can enhance the model through these methods:

1. Collect More Data: If possible, we increase the sample size. Through increasing the sample size, we can better help the model learn the patterns and make better predictions.
2. Tuning of Parameters: To help improve the recall value, one method is setting different parameters on the random forest, through the maximum depth, minimum samples per leaf, and the number of trees

Linear Regression Model

The results from the linear regression model should be used by stakeholders to identify bookings at risk of cancellations. The model found that expensive bookings are at greatest risk of being canceled, particularly those with multiple adults, children, and special requests since those factors lead to the most expensive type of cancellation. Our suggestion to the hotel industry is to target bookings with multiple adults, children, and special requests and either incentivize remaining booked or penalizing cancellations in order to retain hotel guests.

One major limitation of the chosen linear model is that its R^2 value is low, so there is still variability in the data that the model is not accounting for. We were able to improve upon the R^2 score by expanding the model, but a lot of the variability in the data is still left unexplained. Due to this, predictions made from the model may not be accurate or a complete representation of the factors playing into changes in the average daily rate.

One point of future work for the model is to create a further expanded model in order to better explain variability in the data and improve the R^2 score. This could be done by adding in variables from the current dataset or by collecting additional data and data fields for hotel bookings that might explain variability in the data without adding too many. Collecting additional data could help prevent the model from becoming too complex and hard to understand.

7. References/Bibliography (APA Format)

Dawood, M. (2024). Hotel Booking Cancelations. Kaggle.com. https://www.kaggle.com/datasets/muhammaddawood42/hotel-booking-cancelations