FINAL

Observation in the Study

Each observation represents a single hotel booking record with various attributes such as booking lead time, stay duration, market segment, and whether or not the booking was canceled.

Type of Learning and Prediction Goal

This is a supervised learning project with the main objective of classification, where we aim to predict whether a booking will be canceled (binary outcome: 0 or 1). 1 will represent cancellation while 0 represents non-cancellation. In addition, we will use regression analysis to explore how the numerical variable ADR (average daily rate) relates to other features. This helps us understand pricing patterns and their potential association with cancellations.

Models and Algorithms

Decision Trees

Decision trees divide the dataset based on feature values, isolating important features to predict the target variable. Decision trees will be used to identify which variables significantly impact cancellations, allowing us to see the decision pathways leading to cancellation vs. non-cancellation.

Random Forest

Random forests improve upon decision trees by training multiple trees on random data subsets, reducing overfitting and improving accuracy. The final prediction is determined by the average or majority vote of all trees. Decision trees are single models that may overfit, especially with complex data.

Linear Regression

Linear regression analyzes relationships between ADR and predictor variables. We explore relationships involving ADR as the outcome variable, with features like lead time, customer type and is_canceled as predictors. This will allow us to understand whether ADR is affected by features related to cancellations, offering insights into pricing patterns for bookings with higher or lower cancellation risks. For example, by including is_canceled as a predictor, the model can capture differences in ADR associated with cancellations. If canceled bookings frequently show lower ADR, this might suggest that last-minute or discounted bookings have a higher risk of cancellation.

Testing for Success and Ensuring Model Validity

We plan on using an 80/20 train-test data split on our models so that we can evaluate each of their accuracy. We also plan on validating models by carefully controlling our variables to ensure that we don't over or underfit our models. For models with "success metrics" like R^2 and RMSE, we will be checking to see if those metrics align with what is expected for a successful model. For example, we will be checking for R^2 values close to 1 and low RMSE values.

Potential Weaknesses and Failures

We suspect that there may be model bias if canceled bookings occur less often than non-canceled bookings. If we notice model bias in our data, then we will fix it by adjusting class balance. Another potential weakness is that our model will have poor R^2 and RMSE values. We would first try to deal with this by seeing if adding in another explanatory variable would help the model make more sense. If this approach fails, we would come to the conclusion that linear regression is not optimal for predicting behaviors in bookings. Other weaknesses and model failures could come from factors out of our control that the data doesn't account for, such as clients choosing to cancel bookings due to them finding better deals elsewhere.

Feature Engineering

One-Hot encoding

This converts categorical variables, like Hotel or Meal type into binary columns. Some variables like Meal type have more than 2 unique categories, we will separate each meal type as a separate column. The dataset that we are working with has 4 categories for Meal type–BB (bed and breakfast), HB (half board), FB (full board), and SC (self-catering), each category has a value of 1 if that meal type applies to a booking and 0 otherwise.

Feature scaling

Standardizing numerical variables, such as lead_time to improve model performance by bringing all numeric data into a comparable range.

Data features

We can extract features such as day of the week or month to capture seasonal trends that may influence cancellation rate.

Results

We will use a few different methods to communicate our results and whether our analysis plans were successful or not. We will be using a confusion matrix to prove whether or not our models have the ability to successfully predict canceled vs non-canceled bookings. Feature importance will be used to visualize feature importance scores to identify the top drivers of cancellations. After using our test-train split, we'll be able to communicate our model's validity, as well as determine if we need to make adjustments. For the linear regression model, we will be communicating important findings by interpreting the coefficients and intercepts in relation to understanding what drives hotel booking cancellations.

https://github.com/chaizhang/hotel_booking_cancellations/blob/main/wrangling.ipynb

Question: Using the hotel booking dataset, we want to predict whether a booking will be canceled. We aim to identify and analyze the key factors that drive booking cancellations, helping to uncover patterns and trends that can inform decision-making within the hotel industry.

DRAFT

- What is an observation in your study?
    - An observation represents a single hotel booking record with various attributes such as booking lead time, stay duration and market segment.
- Are you doing supervised or unsupervised learning? Classification or regression?
    - This is a supervised learning problem aimed at classification, where the objective is to predict whether a booking will be canceled (binary outcome 0 or 1). There are opportunities for us to use regression, such as with the adr (average daily rate) and seeing its interactions with other numerical variables.
- What models or algorithms do you plan to use in your analysis? How?
    - Decision Trees:
        - Purpose: Decision trees split the dataset into branches based on feature values to predict the target variable.
        - Usage: Decision trees will be used to identify which features (like lead time or meal type) are most important for predicting cancellations.
        - Note: Decision trees are single models that may overfit, especially with complex data.
    - Random Forests:
        - Purpose: Random forests combine multiple decision trees to reduce overfitting and improve prediction accuracy. Each tree in the forest is trained on a different random subset of the data, and the final prediction is based on the average or majority vote of all trees.
        - Usage: Random Forests will be used to enhance accuracy and provide a more reliable model by averaging the outcomes of several decision trees, reducing the chance of overfitting.
    - Linear regression:
        - We can set one of our numerical variables (such as average daily rate) to our outcome variable, and then, using the is_cancelled variable along with any other variables we want, we can see the relationship between the rate and whether the customer canceled.
- How will you know if your approach "works"? What does success mean?
    - Train-Test Split for Accuracy: We will perform an 80/20 train-test split, training on 80% of the data and testing on 20% to evaluate model accuracy.

    - Metric for Success: Accuracy will be our main metric to assess how well the model correctly classifies canceled vs. non-canceled bookings.
- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?
    - Imbalance: We may encounter more non-canceled than canceled bookings, potentially leading to model bias. If accuracy appears biased, we'll consider adjusting the class balance.
    - Outside variables such as customers finding a better deal else where can determine if they cancel.
    - One potential weakness is that our model will have poor R^2 and RMSE values. We would first try to deal with this by seeing if adding in another explanatory variable would help the model make more sense. If this approach fails, we would come to the conclusion that linear regression is not optimal for predicting behaviors in bookings.
- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?
    - One-Hot Encoding: Converting categorical variables (like Hotel or Meal) into binary columns, allowing models to work with categories numerically.
    - Is_canceled is already in the dummy variable format, so for the linear regression model, we would just need to ensure that any other categorical variables used are also in a dummy variable format.
    - In our EDA, we did a correlation analysis on our numerical variables, and we came to the conclusion that our variables are not strongly correlated, so PCA would not be very helpful.
    - Feature Scaling: Standardizing numerical values (such as Lead_time) to similar ranges, improving model performance in some algorithms.
    - Date Features: Deriving new features from Arrival_date, such as day of the week or month, to capture potential seasonal trends in cancellations.
- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.
    - Confusion Matrix: We'll use a confusion matrix to show the model's ability to correctly predict cancellations versus non-cancellations.
    - Feature Importance: For Random Forests, we'll visualize feature importance scores to identify the top drivers of cancellations.
    - Model Accuracy: After the train-test split, model accuracy will indicate how well the approach succeeded or where it might need adjustment.
    - RMSE and R^2: Check for low RMSE scores and R^2 values for model validity
    - Model intercepts and coefficients: Understand how outcomes for hotel bookings change based on different combinations of predictor variables

Lynni

Paragraph 1

- What is an observation in your study?
    - Hotel booking record
- Are you doing supervised or unsupervised learning? Classification or regression?
    - Supervised learning since we want to label the data
    - Classification since we want to predict the categorical outcome rather than see change over time
- What models or algorithms do you plan to use in your analysis? How?
    - Logistic Regression
        - Supervised machine learning algorithm that does binary (yes or no) classification by predicting the probability of an outcome
    - Decision tree
        - Supervised that is used for classification and regression
        - Predict target variable by learning and making inferences from data
- How will you know if your approach "works"? What does success mean?
    - Set baseline which we want to meet such as metrics to measure
        - Measure the precision of our data
    - Determine if our data is underfitting or overfitting
    - Data does well on unseen data
- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?
    - If we use a complex model, then it can lead to overfitting the training data
    - Outside variables such as customers finding a better deal else where can determine if they cancel

Paragraph 2:

- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?
    - Could possibly one-hot encode some categorical variables to be numerical; this way the model can work work categorical variables
    - Variables when did EDA did not seem correlated
    - We can consider making a separate variable that can factor in other outside factors to better link correlation

Paragraph 3:

- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like

R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.

- We can use R^2 and RMSE to ensure for accuracy
- We can use confusion matrix to see if we can predict the cancellation vs non-cancellation rates

Chai

- What is an observation in your study?
    - An observation represents a single hotel booking record with various attributes such as booking lead time, stay duration and market segment.
- Are you doing supervised or unsupervised learning? Classification or regression?
    - This is a supervised learning problem aimed at classification, where the objective is to predict whether a booking will be canceled (binary outcome 0 or 1)
- What models or algorithms do you plan to use in your analysis? How?
    - Decision Trees:
        - Purpose: Decision trees split the dataset into branches based on feature values to predict the target variable.
        - Usage: Decision trees will be used to identify which features (like lead time or meal type) are most important for predicting cancellations.
        - Note: Decision trees are single models that may overfit, especially with complex data.
    - Random Forests:
        - Purpose: Random forests combine multiple decision trees to reduce overfitting and improve prediction accuracy. Each tree in the forest is trained on a different random subset of the data, and the final prediction is based on the average or majority vote of all trees.
        - Usage: Random Forests will be used to enhance accuracy and provide a more reliable model by averaging the outcomes of several decision trees, reducing the chance of overfitting.
- How will you know if your approach "works"? What does success mean?
    - Train-Test Split for Accuracy: We will perform an 80/20 train-test split, training on 80% of the data and testing on 20% to evaluate model accuracy.
    - Metric for Success: Accuracy will be our main metric to assess how well the model correctly classifies canceled vs. non-canceled bookings.
- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?
    - Imbalance: We may encounter more non-canceled than canceled bookings, potentially leading to model bias. If accuracy appears biased, we'll consider adjusting the class balance.
    - Feature Correlation: Many features might correlate. If performance suffers, we'll consider dimensionality reduction techniques like PCA.
- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?
    - One-Hot Encoding: Converting categorical variables (like Hotel or Meal) into binary columns, allowing models to work with categories numerically.
    - Feature Scaling: Standardizing numerical values (such as Lead_time) to similar ranges, improving model performance in some algorithms.

    - Interaction Features: Creating new features from combinations of existing features (e.g., interaction between Market_segment and Is_repeated_guest) that could be relevant to cancellations.
    - Date Features: Deriving new features from Arrival_date, such as day of the week or month, to capture potential seasonal trends in cancellations.
- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.
    - Confusion Matrix: We'll use a confusion matrix to show the model's ability to correctly predict cancellations versus non-cancellations.
    - Feature Importance: For Random Forests, we'll visualize feature importance scores to identify the top drivers of cancellations.
    - Model Accuracy: After the train-test split, model accuracy will indicate how well the approach succeeded or where it might need adjustment.

Mohini

- What is an observation in your study?

In our study, each observation describes the behaviors and actions of different hotel guests. We intend to use these observations to predict guest behavior.

- Are you doing supervised or unsupervised learning? Classification or regression?

We will be using supervised learning. Our data is already structured, and we have a goal in mind for the model to reach. Our models will primarily be classification based, Our primary predictor for knowing if guests canceled or not is a categorical variable, and most of the other variables are also categorical. There are opportunities for us to use regression, such as with the adr (average daily rate) and seeing its interactions with other numerical variables.

- What models or algorithms do you plan to use in your analysis? How?

One type of model we plan to use is a linear regression model. We can set one of our numerical variables (such as average daily rate) to our outcome variable, and then, using the is_cancelled variable along with any other variables we want, we can see the relationship between the rate and whether the customer canceled.

- How will you know if your approach "works"? What does success mean?

We will know if our approach is valid/if the model is valid by checking the R^2 and RMSE values. We'll be trying to make a model that has an R^2 value close to 1 and a low RMSE score. If these criteria are met then our model is successful, and we'll be able to draw conclusions from it.

- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?

One potential weakness is that our model will have poor R^2 and RMSE values. We would first try to deal with this by seeing if adding in another explanatory variable would help the model make more sense. If this approach fails, we would come to the conclusion that linear regression is not optimal for predicting behaviors in bookings.

- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?

Is_canceled is already in the dummy variable format, so for the linear regression model, we would just need to ensure that any other categorical variables used are also in a dummy variable format.

In our EDA, we did a correlation analysis on our numerical variables, and we came to the conclusion that our variables are not strongly correlated, so PCA would not be very helpful.

- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.

Comparison of R^2 and RMSE. We can