Capstone 3: Predicting Hotel Booking Cancellations
Kyle Reedy

Introduction:

Hotels do not want to hear that their customers are canceling their bookings. On average, according to *RevenueHub,* there is a 20% cancellation rate before the guests arrive. That is one in every five guests that arrive. That is quite a lot, and hotels definitely would want to avoid this rate. Guests canceling their bookings lower occupancy, obviously then leading the hotels to lower revenues. An empty hotel room typically results in financial loss for the day. The correlation between last-minute cancellations or no-shows and revenue is straightforward, requiring no complex mathematical understanding. Even if the room is resold later, it is likely to be at a reduced rate, further impacting revenue. That is why it will be beneficial to know when booking cancellations will be made. Utilizing a machine learning model enables the hotel to target potential order cancellations through tailored marketing efforts. A confusion matrix will also help us understand the representation of the accuracy of the models used. This proactive approach helps prevent customer churn, ultimately safeguarding against revenue loss.

Models Used for Classification

1. Logistic Regression
2. KNeighbors Classifier
3. Decision Trees
4. Random Forest
5. AdaBoost Classifier
6. Gradient Boosting Classifier
7. XGBoost Classifier

Confusion Matrix Overview:

TP (True Positive): This refers to the number of instances where the model correctly predicted that a booking would be canceled.
FP (False Positive): This represents the number of instances where the model incorrectly predicted that a booking would be canceled when it actually wasn't canceled.
FN (False Negative): This indicates the number of instances where the model incorrectly predicted that a booking would not be canceled when it actually was canceled.
TN (True Negative): This signifies the number of instances where the model correctly predicted that a booking would not be canceled.

Dataset:

The data is taken originally from the article *Hotel Booking Demand Datasets* written by Nuno Antonio, Ana de Almeida and Luis Nunes. I have downloaded the data from Kaggle (https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand). The dataset is taken

from 2 resort hotels and 2 city hotels located in the region of Algarve, Portugal's southernmost region. The location has fishing villages with sandy coves lined with hotels, villas, bars and restaurants. A great area for couples and even families that want to spend their vacation.

Here's a brief overview of the dataset: Before any data cleaning, there are 119,390 observations. The dataset encompasses booking details for both a city hotel and a resort hotel in Portugal, spanning from 2015 to 2017. It comprises 32 variables, including reservation and arrival dates (year, month, and day), duration of stay, cancellation status, the count of adults, children, and babies, available parking spaces, total special requests, and details about the agent or company responsible for the reservation, among others. Using the available dataset, we want to look into the following research questions.

Research Questions:

1. Can Cancellations be predicted with the use of machine learning from the available dataset?
2. What factors are most relevant when predicting cancellations?

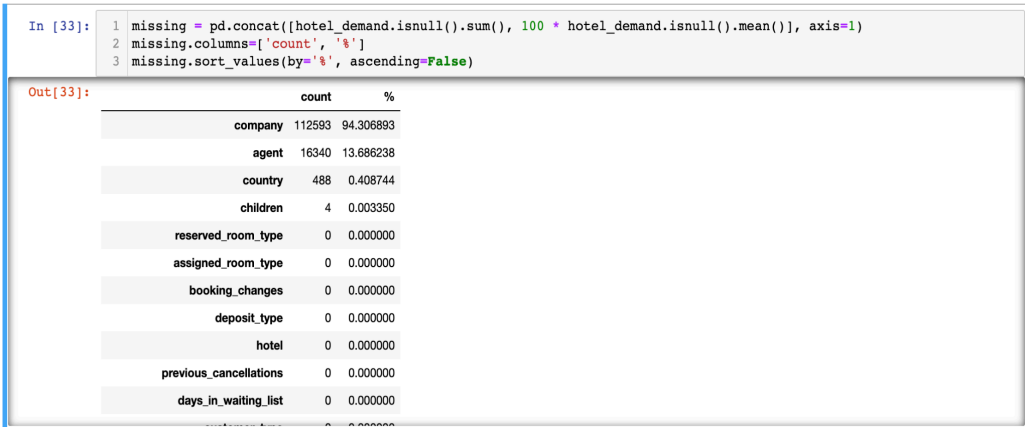
Here are the list of variables and its descriptions:

1. Hotel: Type of hotel (Resort Hotel, City Hotel)
2. is_canceled: Reservation cancellation status (0 = not canceled, 1 = canceled)
3. lead_time: Number of days between booking and arrival
4. arrival_date_year: Year of arrival
5. arrival_date_month: Month of arrival
6. arrival_date_week_number: Week number of the year for arrival
7. arrival_date_day_of_month: Day of the month of arrival
8. stays_in_weekend_nights: Number of weekend nights (Saturday and Sunday) the guest stayed or booked
9. stays_in_week_nights: Number of week nights the guest stayed or booked

10. adults: Number of adults
11. children: Number of children
12. babies: Number of babies
13. meal: Type of meal booked (BB, FB, HB, SC, Undefined)
14. country: Country of origin of the guest
15. market_segment: Market segment designation
16. distribution_channel: Booking distribution channel
17. is_repeated_guest: If the guest is a repeat customer (0 = not repeated, 1 = repeated)
18. previous_cancellations: Number of previous bookings that were canceled by the customer
19. previous_bookings_not_canceled: Number of previous bookings that were not canceled by the customer
20. reserved_room_type: Type of reserved room
21. assigned_room_type: Type of assigned room
22. booking_changes: Number of changes made to the booking
23. deposit_type: Type of deposit made (No Deposit, Refundable, Non Refund)
24. agent: ID of the travel agent responsible for the booking
25. company: ID of the company responsible for the booking
26. days_in_waiting_list: Number of days the booking was in the waiting list
27. customer_type: Type of customer (Transient, Contract, Transient-Party, Group)
28. adr: Average Daily Rate
29. required_car_parking_spaces: Number of car parking spaces required
30. total_of_special_requests: Number of special requests made
31. reservation_status: Last reservation status (Check-Out, Canceled, No-Show)
32. reservation_status_date: Date of the last reservation status

Cleaning and Wrangling Process

During the cleaning process, two variables have been taken off: agent and company.
The image below displays the percentage of missing values out of the total dataset.
The ID of the company had over 94% in missing values, plus the company responsible for the booking has almost no relevance with our predictions.

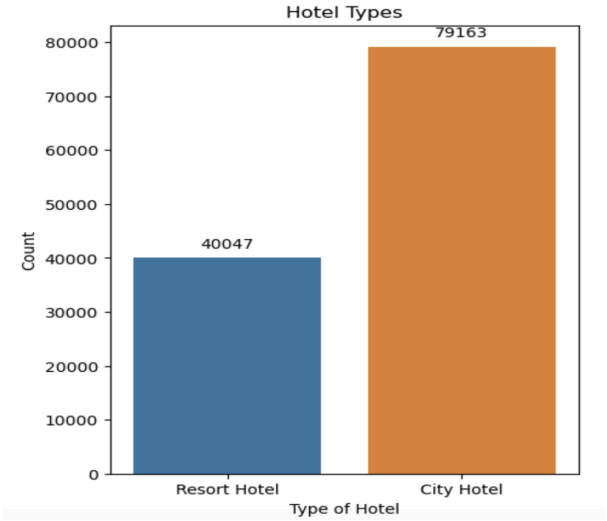


```
In [33]:
missing = pd.concat([hotel_demand.isnull().sum(), 100 * hotel_demand.isnull().mean()], axis=1)
missing.columns=['count', '%']
missing.sort_values(by='%', ascending=False)
```

Out[33]:

	count	%
company	112593	94.306893
agent	16340	13.686238
country	488	0.408744
children	4	0.003350
reserved_room_type	0	0.000000
assigned_room_type	0	0.000000
booking_changes	0	0.000000
deposit_type	0	0.000000
hotel	0	0.000000
previous_cancellations	0	0.000000
days_in_waiting_list	0	0.000000

Duplicates have been taken off with a total remaining with 119,210 observations and 30 columns. The image below shows the observations grouped by hotel: ‘City Hotel’ and ‘Resort Hotel.’



The total observations for Resort Hotel is 40,047 and the total for City Hotel is 79,163. Using the following observations, we will identify the best model for predicting booking cancellations and the most significant variables that lead to cancellations.

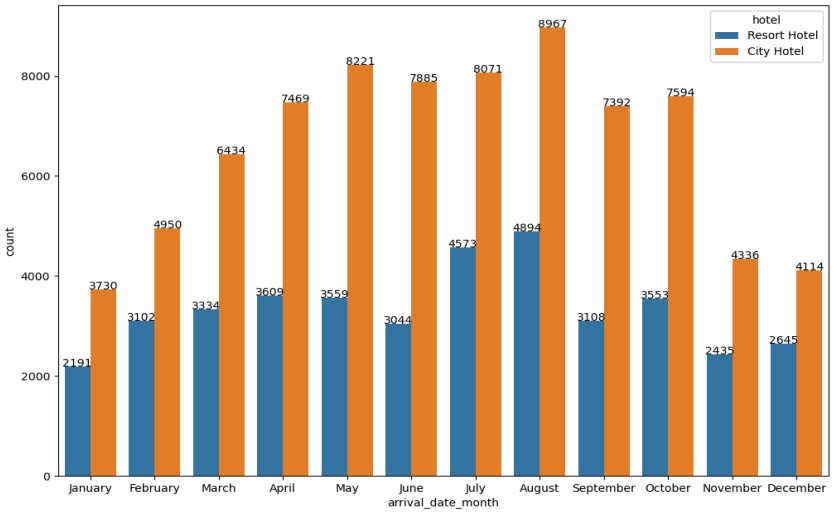
Descriptive Statistics and EDA Observations:

The top countries of origin for hotel guests are Portugal, the UK, France, Spain, and Germany. However, Portugal, along with other top countries, also experiences a significant number of cancellations. Implementing reasonable cancellation policies, particularly during off-peak seasons. It seems that guests that come from farther countries, for example, the US, have a lower cancellation rate.

	country	No of guests
0	PRT	21398
1	GBR	9668
2	FRA	8468
3	ESP	6383
4	DEU	6067
5	IRL	2542
6	ITA	2428
7	BEL	1868
8	NLD	1716
9	USA	1592

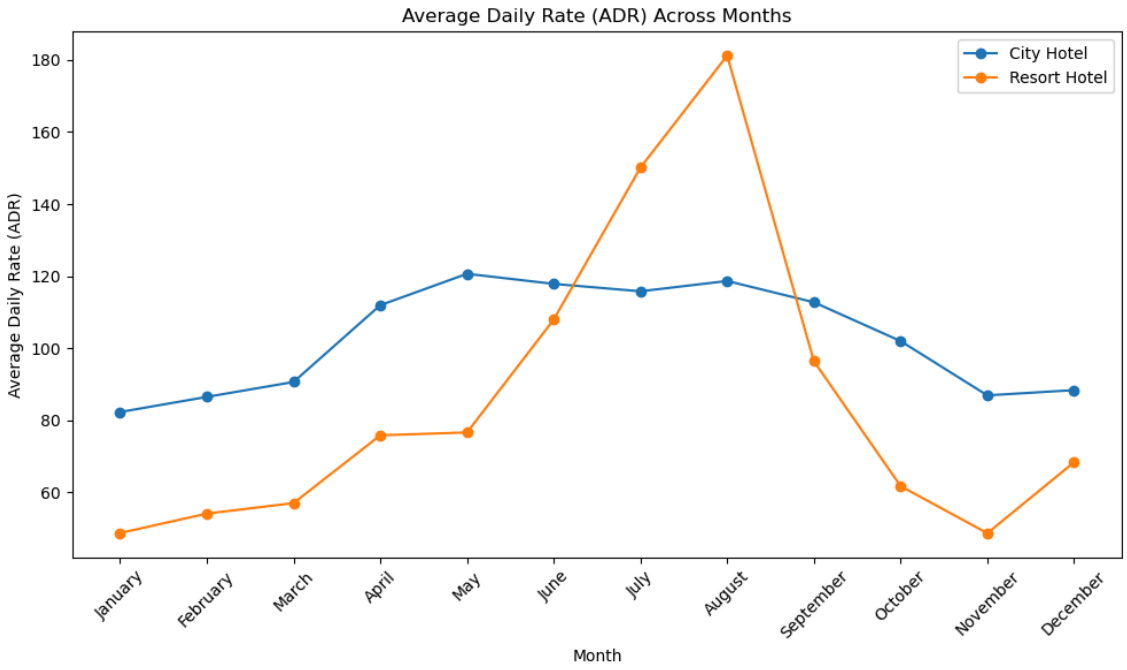
People are mostly coming from Portugal and then neighboring European countries. The US is the only non-European country in the top 10 guests by country.

Busiest Months Per Hotel



The busiest months based off of the count of arrivals are more towards the warmer months. Regardless of the type of hotel, the above graph indicates that August has the most customers for both resort hotels and city hotels. The numbers significantly decrease during the months of December and January, more likely due to the fact that those are the coldest times of the year in the region.

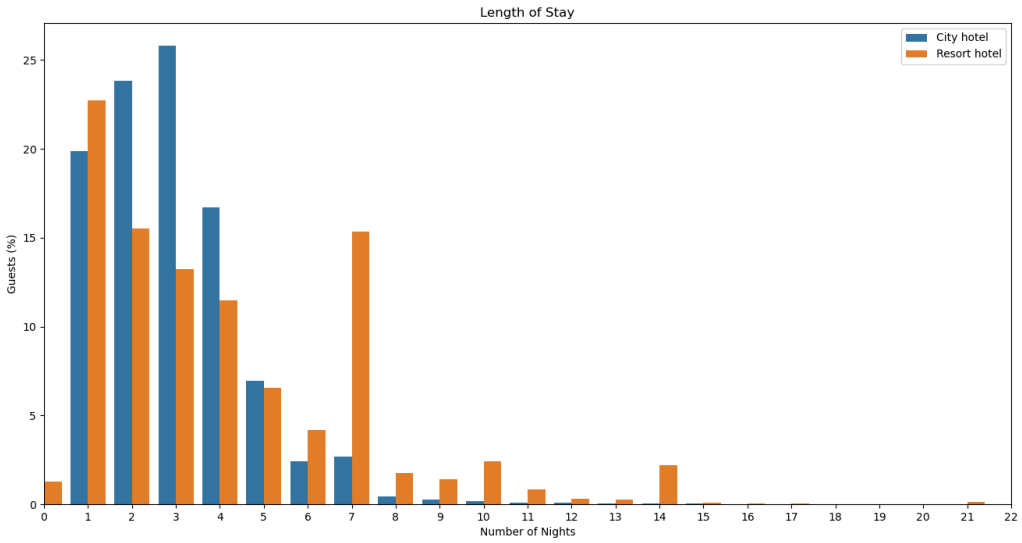
How are the Average Daily Rates for Customers per Month?



With the most busiest months being the summer time of the season, it is understandable that the average daily rates are higher during the months of June to September. Especially for the

resort hotel, there are more amenities and attractions within the hotel such as the beach or an outdoor pool which would inevitably increase the number of customers wanting to come during the warmer months.

How long Do People Stay?



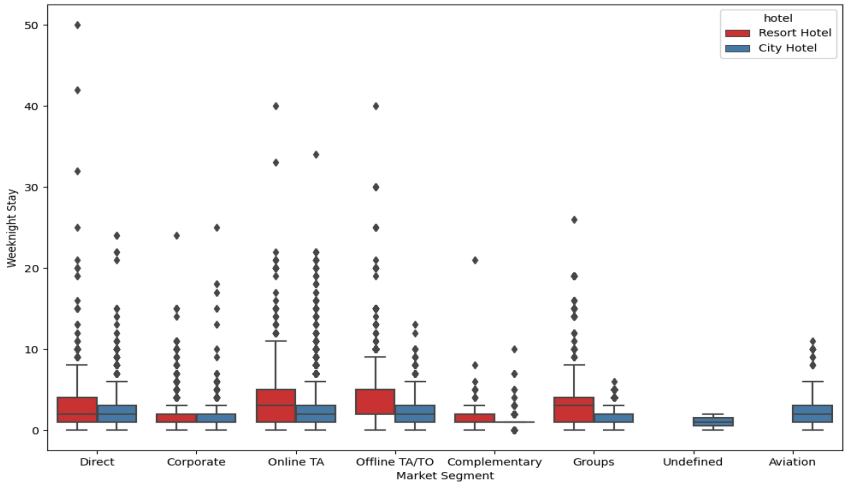
The trend we can see from this graph shows that the guests that book at the city hotel tend to stay for a shorter period. We can make a multitude of assumptions on why, such as people staying there for a business trip or people visiting this Portuguese region who are wanting to explore outside. On the other hand, you can see that over 15% of guests that stay in the resort stay for 7 nights. Another portion stayed for two weeks.

Average Key Numbers:

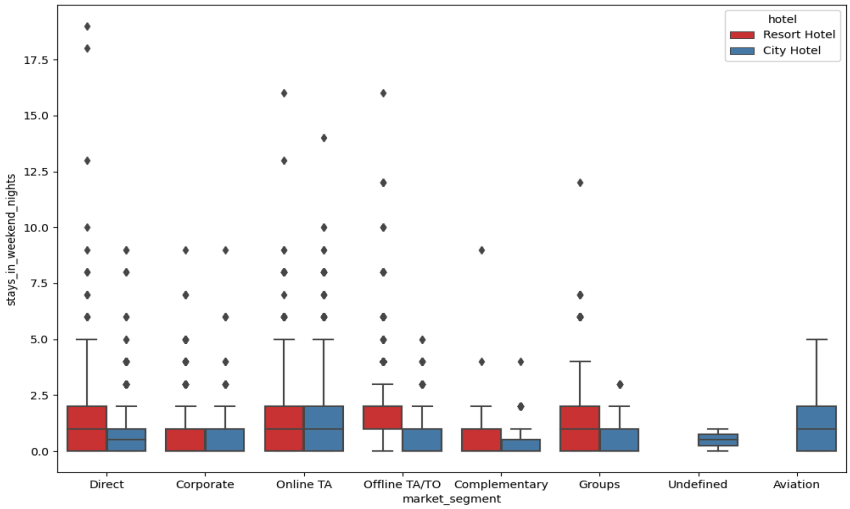
1. Guests at city hotels stay on average of 2.92 nights per stay
2. Guests at resort hotels stay on average 4.14 nights per stay

Stay (Weeknights vs Weekends) by Segments

Weeknights

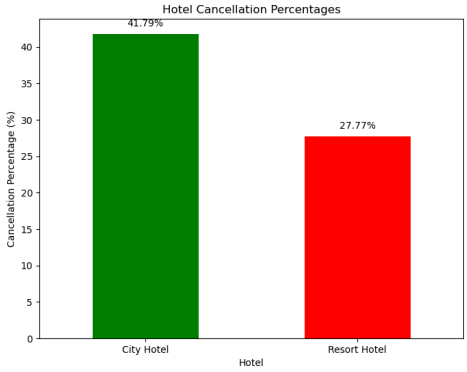


Weekends



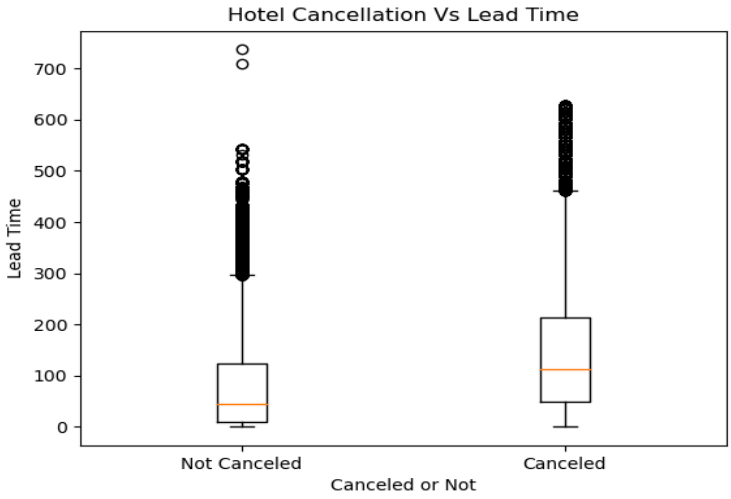
The majority of bookings in various distribution channels and market segments are facilitated by travel agencies, both online and offline. Targeting partners through these agencies' websites can be advantageous due to the high volume of visitors they attract. Among different segments, group reservations tend to have higher cancellation rates compared to non-cancelled bookings. Notably, online travel agencies (TA) and offline travel agencies/tour operators (TA/TO) show higher cancellation rates, although they are overshadowed by their non-cancelled bookings. Conversely, the 'Direct' and 'Corporate' segments exhibit the lowest cancellation rates. This suggests that cancellations are more prevalent in collective bookings, such as group reservations.

Cancellation Proportions

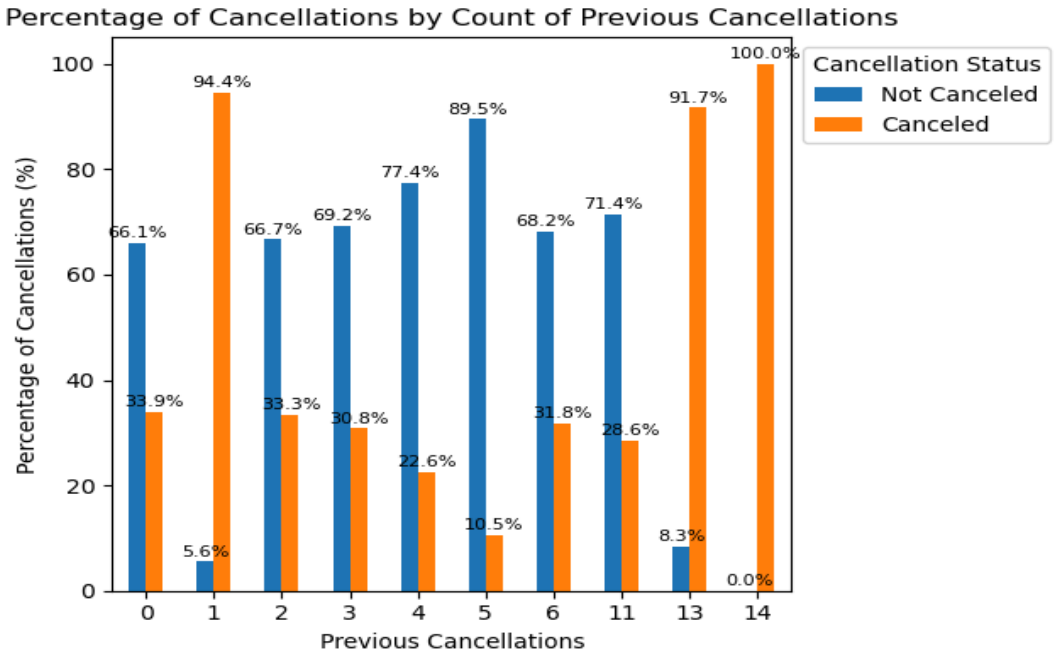


Customers tend to cancel less with resort hotels. Again, there are many assumptions one can make on why, but we can see some other graphs that show why that may be the case that resort hotels have a lower cancellation rate than city hotels. City hotels can have more single party reservations whereas the resort hotels will have more families.

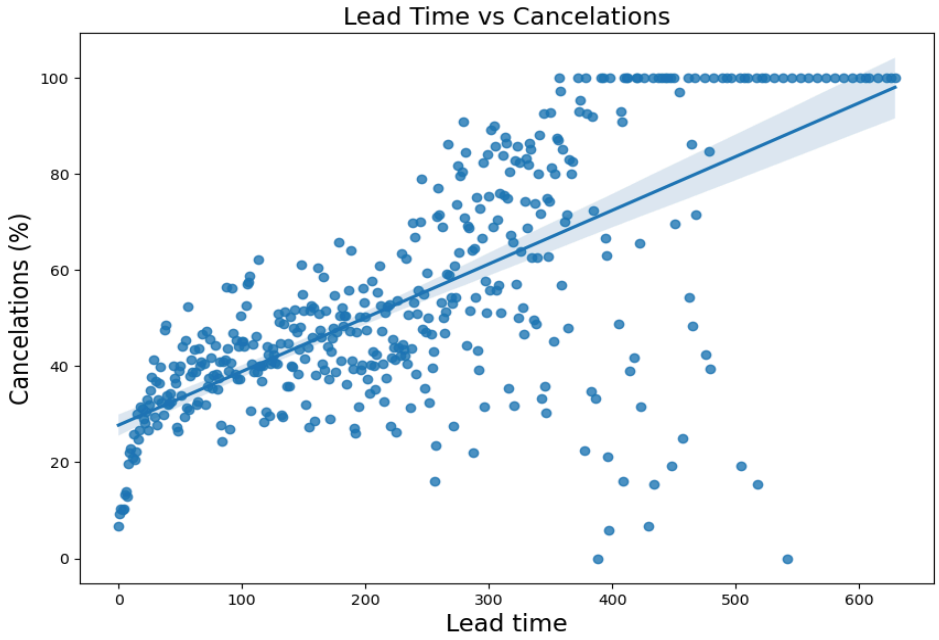
Is there a connection between the time we booked the vacation till the actual arrival date and customer cancellation?



Seeing Whether Previous Cancellations Affect Future Cancellations



From the graph displayed above, guests with previous cancellations indicate that there was 94.4% of canceled bookings. From one previous cancellation however, the percentage of cancellations decreases from 94.4% to 33.3%. It is an interesting trend to see as the most cancellations have occurred where guests have canceled previously once and 13 times. Lets see how lead time correlates with cancellations.



Looking at the graph, bookings made a few days before their arrival date are rarely canceled. On the other hand, bookings made lets say a year in advance have a high probability of cancellations.

Preprocessing

I preprocessed the data by filtering, encoding, and scaling the features to prepare them for modeling. The datasets were then divided into training and test sets, comprising 80% and 20% of the original data, respectively.

Below are the categorical variables in the dataset. Since most machine learning models only accept numeric variables, preprocessing the categorical variables becomes a necessary step. These categorical variables were converted to numbers such that the model is able to understand and extract valuable information.

```
category_cols = [col for col in df2.columns if df2[col].dtype == 'O']
category_cols
```

```
['hotel',
 'arrival_date_month',
 'meal',
 'country',
 'market_segment',
 'distribution_channel',
 'reserved_room_type',
 'assigned_room_type',
 'deposit_type',
 'customer_type',
 'reservation_status',
 'reservation_status_date']
```

Out of these variables, the following variables have been dropped:

```
category_df2.drop(['reservation_status', 'assigned_room_type', 'country'], axis=1, inplace=True)

```

```
category_df2.head()
```

	hotel	meal	market_segment	distribution_channel	reserved_room_type	deposit_type	customer_type	year	month	day
0	0	0	0	0	0	0	0	0	7	1
1	0	0	0	0	0	0	0	0	7	1
2	0	0	0	0	1	0	0	0	7	2
3	0	0	1	1	1	0	0	0	7	2
4	0	0	2	2	1	0	0	0	7	3

After making sure to change any null values and take out outliers that will possibly affect the modeling process, the next step is good to go.

Modeling

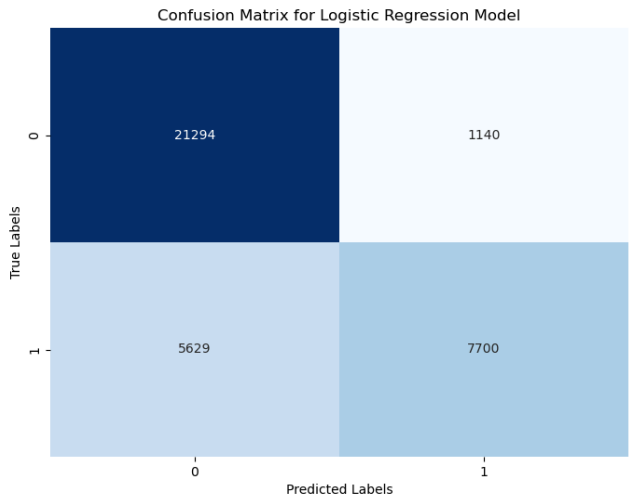
In the final step of this project, I trained multiple models using the training set and evaluated their performance by comparing their accuracy scores in predicting cancellations on the test set. To further analyze the performance, a confusion matrix is displayed.

1. Logistic Regression

Logistic regression is a widely used supervised machine learning algorithm that predicts a categorical dependent variable based on a set of independent variables. This type of regression is suitable when the dependent variable is binary, such as "Yes" or "No", "0" or "1", or "true" or "false". While logistic regression shares similarities with linear regression, its implementation differs. Linear regression is used for solving regression problems, where the goal is to predict continuous numeric values, while logistic regression is employed for classification tasks, where the goal is to classify data into distinct categories.

Here is the performance of what the logistic regression provides below:

```
Accuracy score for the Logistic Regression is: 0.8107261695047955
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.95      0.86     22434
           1       0.87      0.58      0.69     13329

    accuracy                           0.81     35763
   macro avg       0.83      0.76      0.78     35763
weighted avg       0.82      0.81      0.80     35763
```



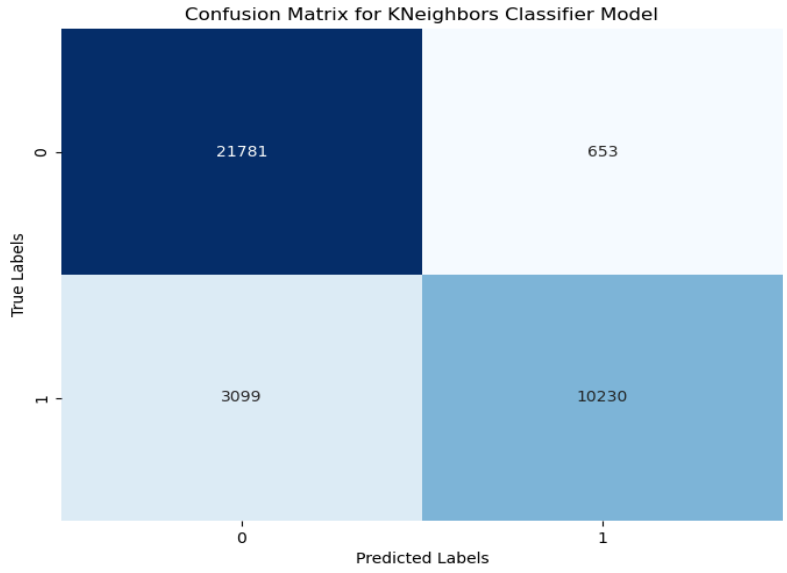
The accuracy score displays a score of 0.81 and the f1-score is approximately 0.86.
From the confusion matrix above, there are 21294+7700 correct predictions, specifically, 21294 samples are predicted correctly that they will cancel the booking. Additionally, 7700 are predicted correctly that they will not cancel the booking. 5629 canceled samples that have been incorrectly predicted as not canceled and 1140 of the not canceled samples are predicted to be canceled.

2. **KNeighbors Classifier:**

KNeighbors Classifier is a simple yet effective algorithm for classification tasks. It works by finding the 'k' nearest neighbors of a data point based on a distance metric (such as Euclidean distance) and assigns the majority class among those neighbors as the predicted class for the data point. It's a non-parametric method, meaning it doesn't make any assumptions about the underlying data distribution.

Here is the performance of what the KN Neighbors Classifier provides below:

```
Accuracy Score for the KNeighbors Classifier is : 0.8950871011939714
Classification Report :
              precision    recall  f1-score   support

           0       0.88      0.97      0.92     22434
           1       0.94      0.77      0.85     13329

    accuracy                           0.90     35763
   macro avg       0.91      0.87      0.88     35763
weighted avg       0.90      0.90      0.89     35763
```



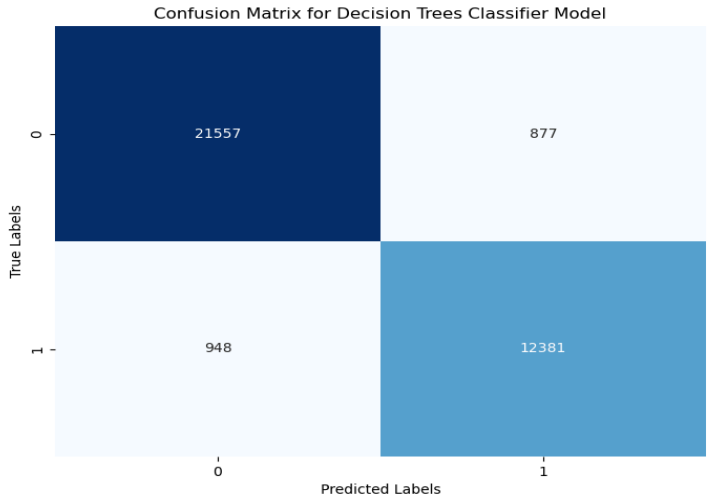
The accuracy score for the KNearest Neighbor model is 0.89 and the f1-score is 0.92. Looking at the confusion matrix, there are 21781+10230 correct predictions. 21781 canceled samples and 10230 not canceled samples are correctly predicted. 3099 canceled samples are mistakenly predicted to be not canceled and 653 not canceled samples are predicted to be canceled.

3. Decision Trees:

Decision Trees are a popular type of model for both classification and regression tasks. They work by recursively splitting the data into subsets based on features that result in the best separation of the target variable. Each split is chosen to maximize the homogeneity of the target variable within each subset. Decision Trees are easy to interpret and visualize, making them useful for understanding the decision-making process of the model.

Here is the performance of what Decision Trees provides below:

```
Accuracy Score for the Decision Tree is : 0.9489696054581551
Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.96      0.96     22434
           1       0.93      0.93      0.93     13329

    accuracy                           0.95     35763
   macro avg       0.95      0.94      0.95     35763
weighted avg       0.95      0.95      0.95     35763
```



The accuracy score for the Decision Trees model is 0.94 and the f1-score is 0.96.

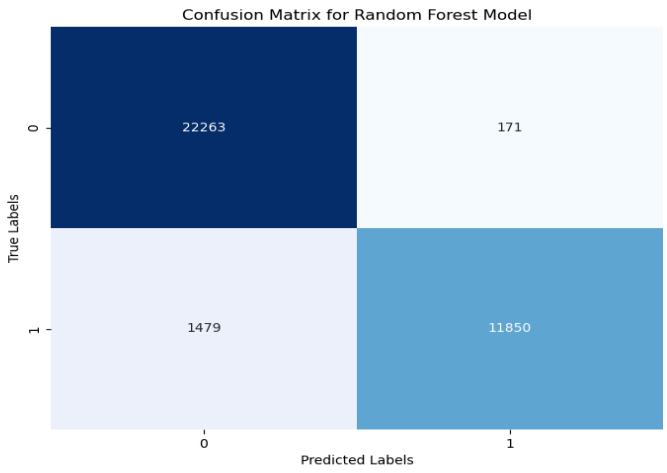
Looking at the confusion matrix, there are 21551+12381 correct predictions. Specifically, 21557 canceled samples and 12381 not canceled samples are correctly predicted. 948 canceled samples are mistakenly predicted to be not canceled and 877 not canceled samples are predicted to be canceled.

4. **Random Forest:**

Random Forest is an ensemble learning technique based on Decision Trees. It builds multiple decision trees using random subsets of the data and random subsets of the features. Each tree in the forest independently makes a prediction, and the final prediction is determined by a majority vote (classification) or averaging (regression) of the predictions from individual trees. Random Forest tends to generalize well and is less prone to overfitting compared to individual decision trees.

Here is the performance of what Random Forests provides below:

```
Accuracy score for the Random Forest Classifier is: 0.9538629309621676
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     22434
           1       0.99      0.89      0.93     13329

    accuracy                           0.95     35763
   macro avg       0.96      0.94      0.95     35763
weighted avg       0.96      0.95      0.95     35763
```



The accuracy score for the random forest model is 0.95 and the f1-score is 0.96.

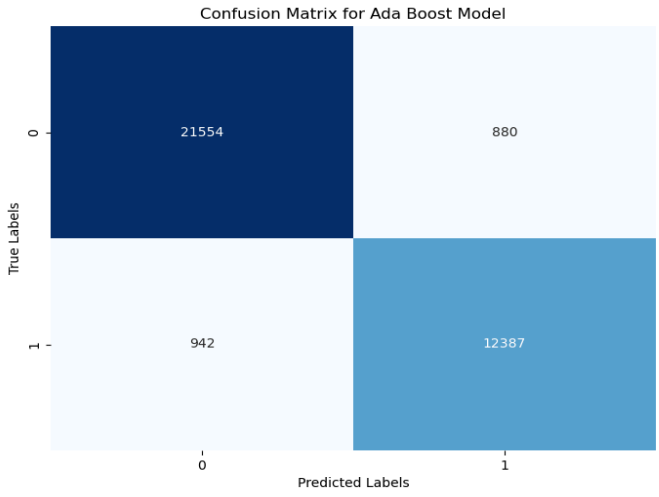
Looking at the confusion matrix, there are 22263+11850 correct predictions. Specifically, 22263 canceled samples and 11850 not canceled samples are correctly predicted. 1479 canceled samples are mistakenly predicted to be not canceled and 177 not canceled samples are predicted to be canceled.

5. AdaBoost Classifier:

AdaBoost (Adaptive Boosting) is another ensemble learning method that combines multiple weak learners (usually Decision Trees) to create a strong learner. It works by iteratively training new models, giving more weight to misclassified data points in each iteration. The final prediction is a weighted sum of the predictions from all models. AdaBoost focuses more on the hard-to-classify data points, improving the overall performance of the model.

Here is the performance of what AdaBoost Classifier provides below:

```
Accuracy score for the AdaBoost Classifier is: 0.9490534910382239
Confusion Matrix:
[[21554   880]
 [  942 12387]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96     22434
           1       0.93      0.93      0.93     13329

    accuracy                           0.95     35763
   macro avg       0.95      0.95      0.95     35763
weighted avg       0.95      0.95      0.95     35763
```



The accuracy score for the random forest model is 0.95 and the f1-score is 0.96.

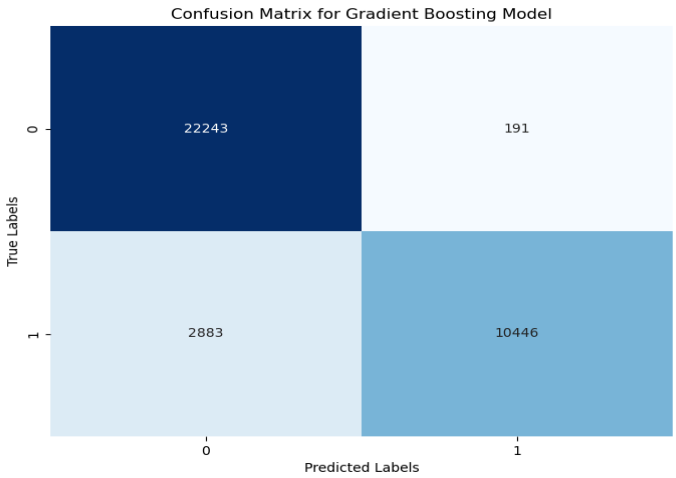
Looking at the confusion matrix, there are 21554+12387 correct predictions. Specifically, 21554 canceled samples and 12387 not canceled samples are correctly predicted. 942 canceled samples are mistakenly predicted to be not canceled and 880 not canceled samples are predicted to be canceled.

6. Gradient Boosting Classifier:

Gradient Boosting Classifier is another ensemble method that builds multiple decision trees sequentially. However, unlike AdaBoost, it doesn't adjust the weights of data points. Instead, it fits each new tree to the residual errors (the differences between the predicted and actual values) of the previous tree. By iteratively minimizing these residuals, Gradient Boosting gradually improves the model's predictive accuracy. It's a powerful and flexible technique that often achieves state-of-the-art performance in various machine learning tasks.

Here is the performance of what Gradient Boosting provides below:

```
Accuracy score for the Gradient Boosting Classifier is: 0.9140452422895171
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     22434
           1       0.99      0.89      0.93     13329

    accuracy                           0.95     35763
   macro avg       0.96      0.94      0.95     35763
weighted avg       0.96      0.95      0.95     35763
```



The accuracy score for the random forest model is 0.91 and the f1-score is 0.96.
Looking at the confusion matrix, there are 22243+10446 correct predictions. Specifically, 22243 canceled samples and 10446 not canceled samples are correctly predicted. 2883 canceled

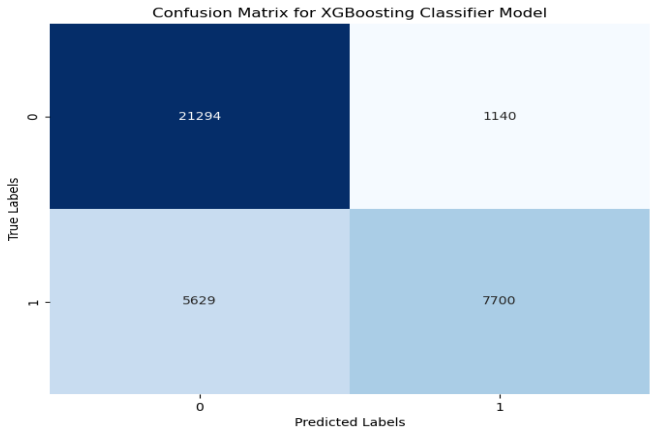
samples are mistakenly predicted to be not canceled and 191 not canceled samples are predicted to be canceled.

7. XGBoost

For the final model used, XGBoost, short for Extreme Gradient Boosting, is a powerful algorithm used in machine learning for both classification and regression tasks. It works by constructing multiple decision trees sequentially, each tree correcting the errors of the previous one. XGBoost is an ensemble learning method that employs the boosting technique, which focuses on combining weak learners to create a strong learner. The algorithm is called "gradient boosting" because it uses gradient descent optimization to minimize errors during the training process. It iteratively builds trees by learning from the mistakes of the previous trees, gradually improving its predictive performance. By combining the predictions of multiple trees, XGBoost can produce highly accurate models capable of handling complex datasets.

Here is the performance of what XGBoost provides below:

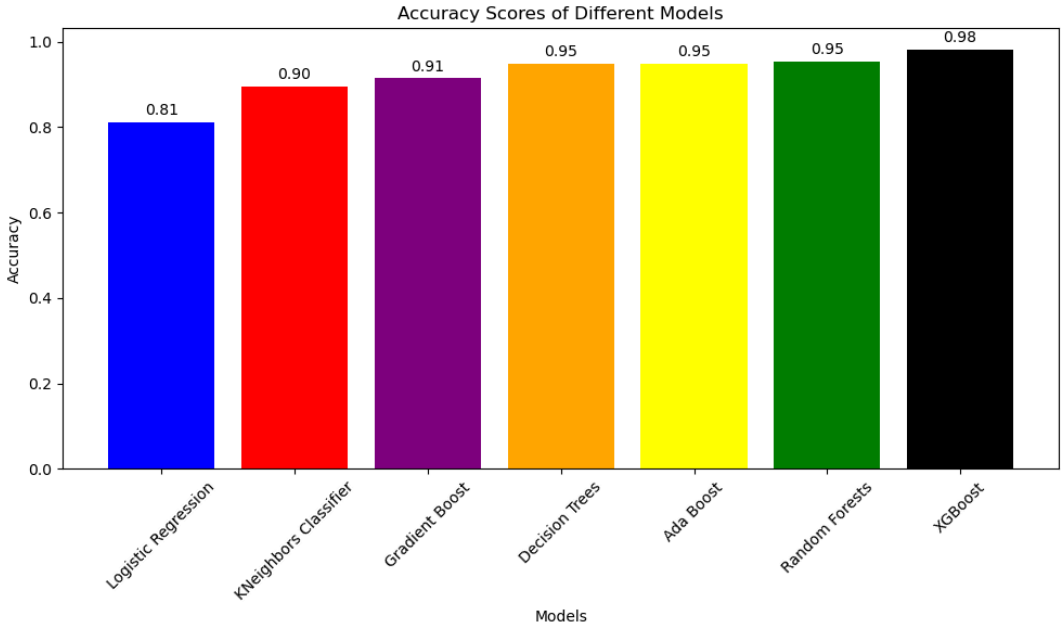
```
Accuracy score for the XGBoost Classifier is: 0.9820484858652797
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99     22434
           1       1.00      0.95      0.98     13329

    accuracy                           0.98     35763
   macro avg       0.99      0.98      0.98     35763
weighted avg       0.98      0.98      0.98     35763
```



The accuracy score for the XGBoost model is 0.98 and the f1-score is 0.99.
Looking at the confusion matrix, there are 21294+7700 correct predictions. Specifically, 21294 canceled samples and 7700 not canceled samples are correctly predicted. 5629 canceled

samples are mistakenly predicted to be not canceled and 1140 not canceled samples are predicted to be canceled.

Model Comparisons



After utilizing 7 models, XGboost had the highest score amongst the other models. The next highest have a score over 0.90 which is still high. Although each score for these models provide different perspectives, it is safe to say that improving these models can really help hotel businesses help predict cancellations made by guests. Below are the recommendations I have provided for this project.

Recommendations:

After obtaining information from the modeling process, here are the recommendations that should be considered for hotels that are affected by the cancellations by customers.

1. Focus on High-Performing Models:
   Utilize XGBoost and Random Forests for prediction tasks due to their higher accuracy scores compared to other models. These models are more reliable in identifying potential cancellations and can assist in proactive management strategies.

2. Segmentation and Targeted Marketing:
   Use market segmentation techniques to tailor marketing strategies based on customer characteristics and behavior. For instance, target repeat customers with personalized offers or incentives to encourage loyalty and reduce cancellations. Transient bookings, which include independent reservations made directly with the hotel, constitute the majority and also

experience a higher cancellation rate. Providing special deals for transient guests could help retain them and reduce cancellations. Additionally, offering extra services such as daily tours or rentals may increase guest engagement and decrease cancellations.

3. Well Planned Accommodations:

Bookings with special requests have lower cancellation rates compared to those without any special requests. Therefore, accommodating guest requests could contribute to a decrease in cancellations.