Rishabh Shrestha

Csci 543 Project 2

Hotel Reservation Cancellation Prediction(Resort Hotel)

Executive Summary

Hotel Reservation Cancellation Prediction

Understanding Data:

In this project, I explored a dataset of hotel reservation cancellation. I have explored the features of dataset using different statistical analysis and visualization techniques. And the finally created prediction models for cancellation and then tuned the hyper parameters of the classifiers to get the best parameters set.

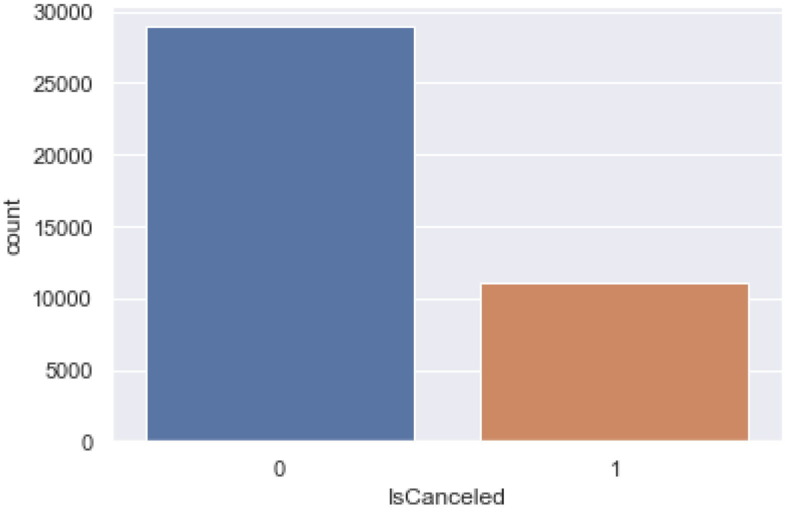
The dataset I am using is of a Resort Hotel. The dataset originally has 31 features such as IsCanceled, LeadTime, ArrivalDateYear, ArrivalDateMonth, ArrivalDateWeekNumber, ArrivalDateDayOfMonth, StaysInWeekendNights, StaysInWeekNights, Adults, Children, Babies, Meal, Country, MarketSegment, DistributionChannel, IsRepeatedGuest, PreviousCancellations, PreviousBookingsNotCanceled, ReservedRoomType, AssignedRoomType, BookingChanges, DepositType, Agent, Company, DaysInWaitingList, CustomerType, ADR, RequiredCarParkingSpaces, TotalOfSpecialRequests, ReservationStatus, ReservationStatusDate.

Exploratory Data Analysis:

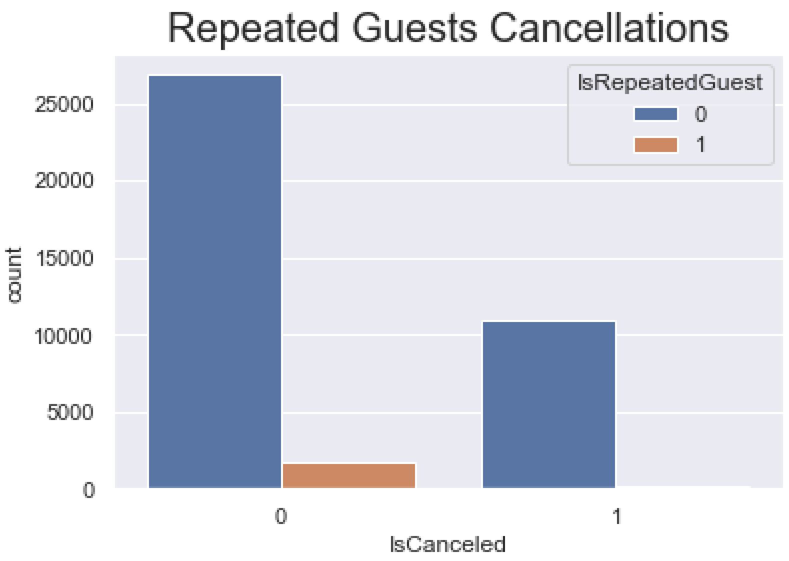
1. Label Distribution
2. Cancellations by Repeated Guests
3. Bookings by Month
4. Cancellations by Month
5. Customer Type Distribution
6. Cancellations by Assigned Room Type

7. Top Dates with Highest Cancellation

1. Label Distribution: The target variable is IsCanceled with two values 0 and 1. Probably 0 means No and 1 means Yes. There are 40060 instances with 28938 of them being 0 and 11122 being 1. Below is the bar chart of the distribution of the labels in the dataset. The distribution is pretty uneven which is a challenge while comparing the performance of different prediction models. We will look into that later.

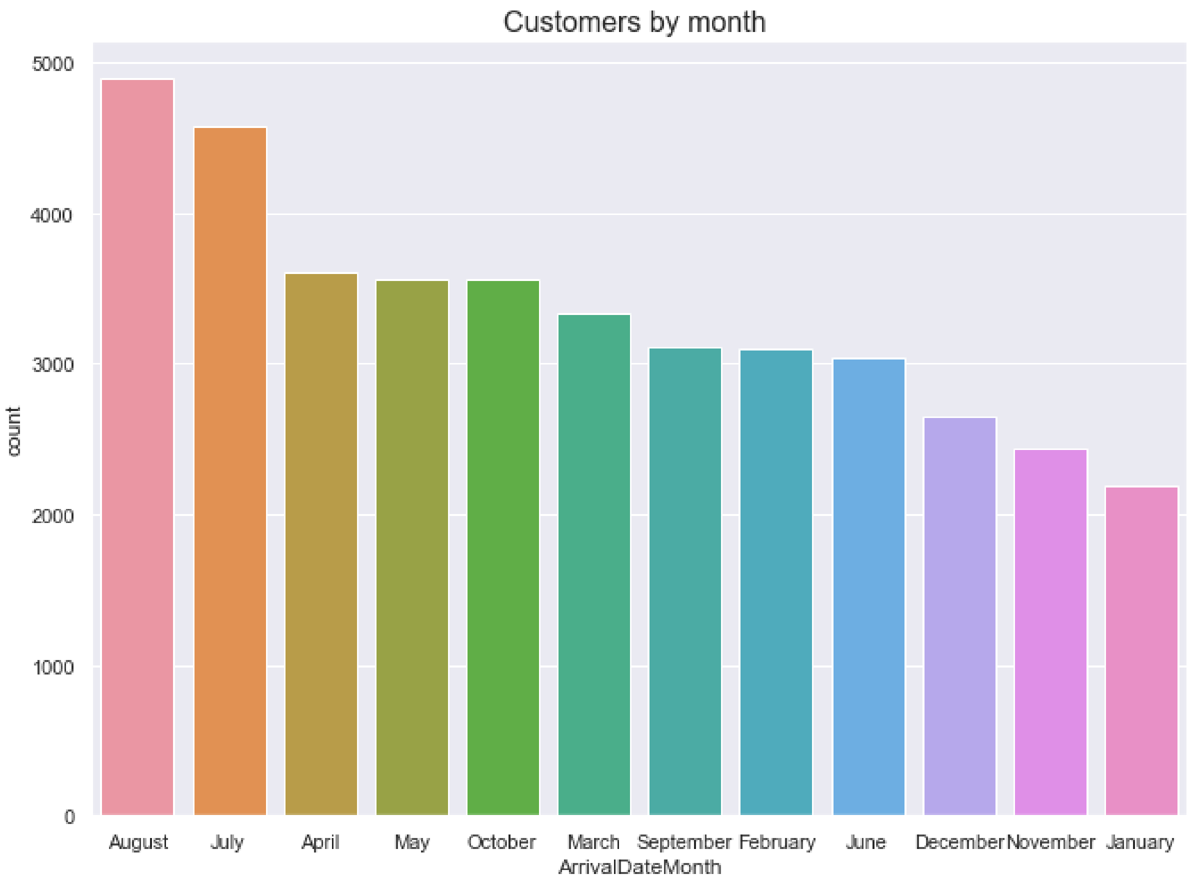


2. Cancellations by Repeated Guests: Below is the bar chart showing the Repeated Guests cancellation. We can see that repeated guests tend to not cancel their reservations as opposed to non-repeated guests who tend to have more cancellations.



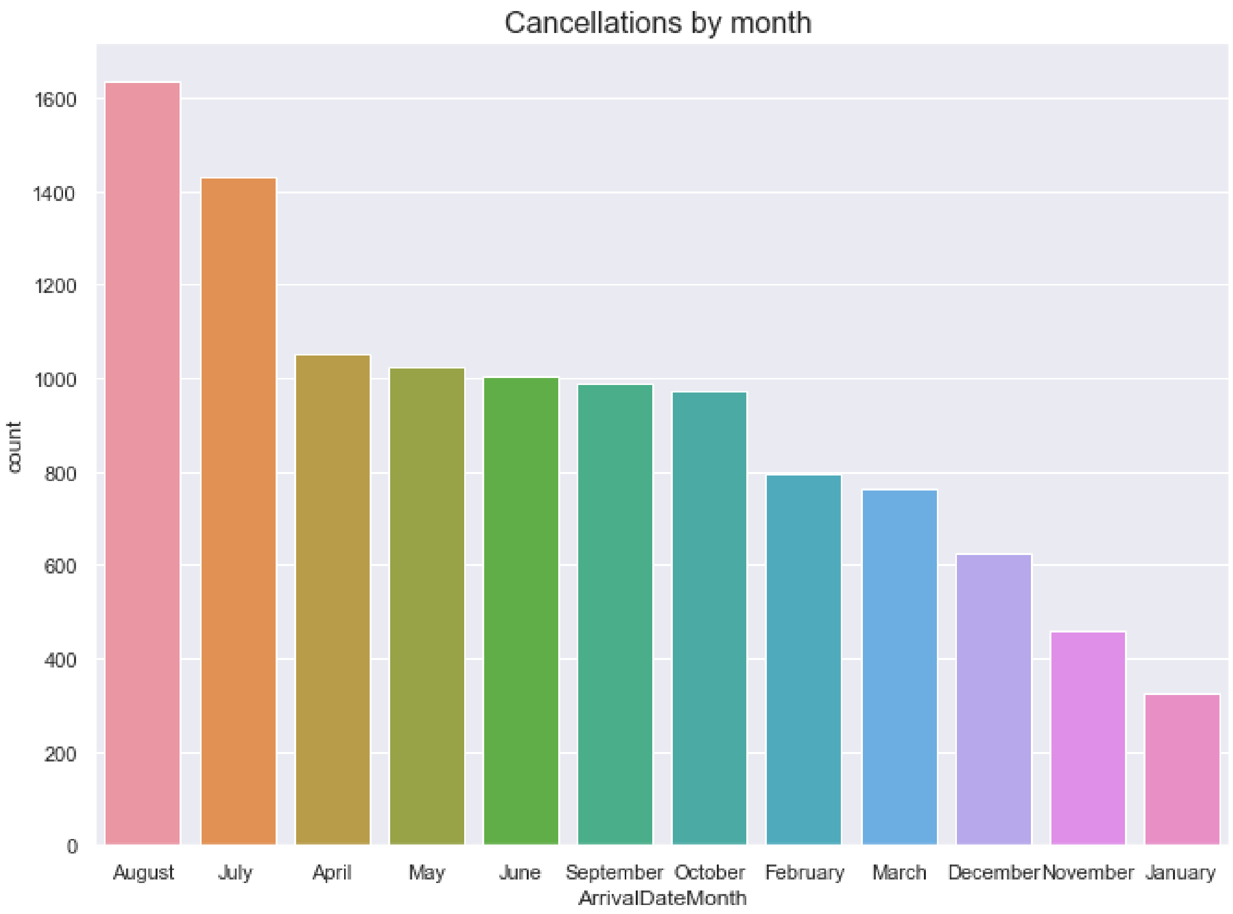
3. Bookings by Month

The graph below shows the number of customers by month. According to the bar chart, August is the busiest month and January is the month with the least number of bookings. It seems like people tend to go on vacation a lot during the Fall season which is very likely.

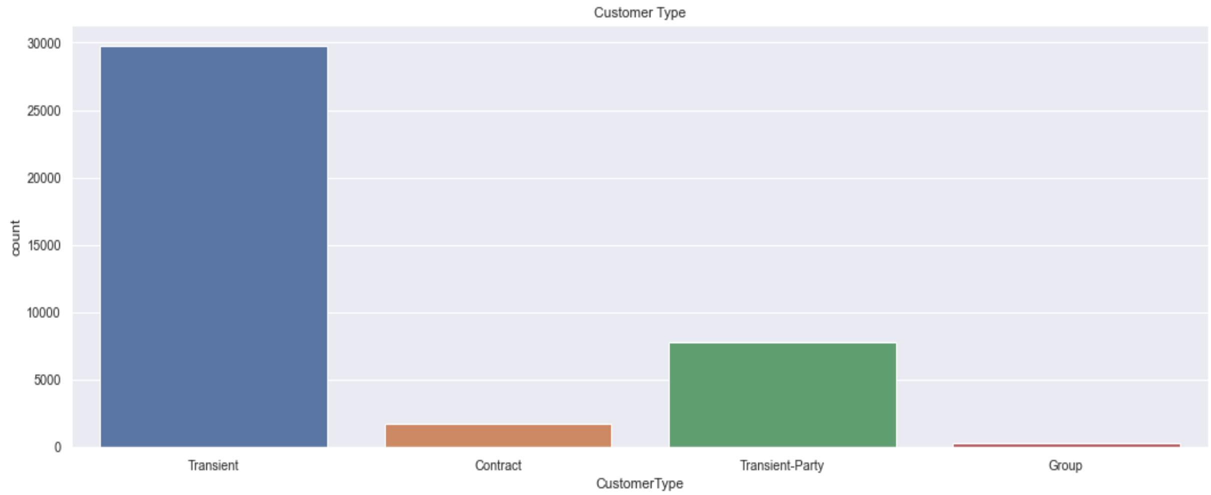


4. Cancellations by Month:

From the graph below and the previous graph, we can conclude that the month with highest and lowest bookings has the highest and lowest cancellations respectively. August, July, April are the top three months with the highest number of bookings and cancellations. January, November and December are the months with the lowest number of booking and cancellations so the ratio seems to be directly proportional.

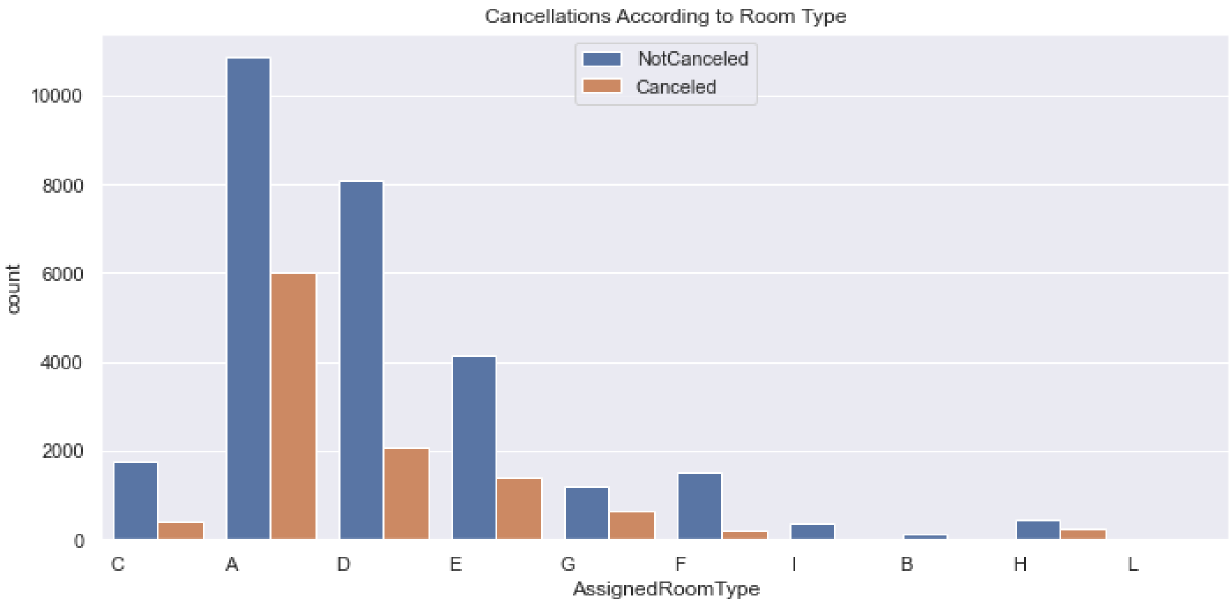


5.Customer Type Distribution: The bar chart below shows the Customer Type of the hotel. Transient type of Customer tend to have the highest bookings and Group type have the lowest number of bookings.



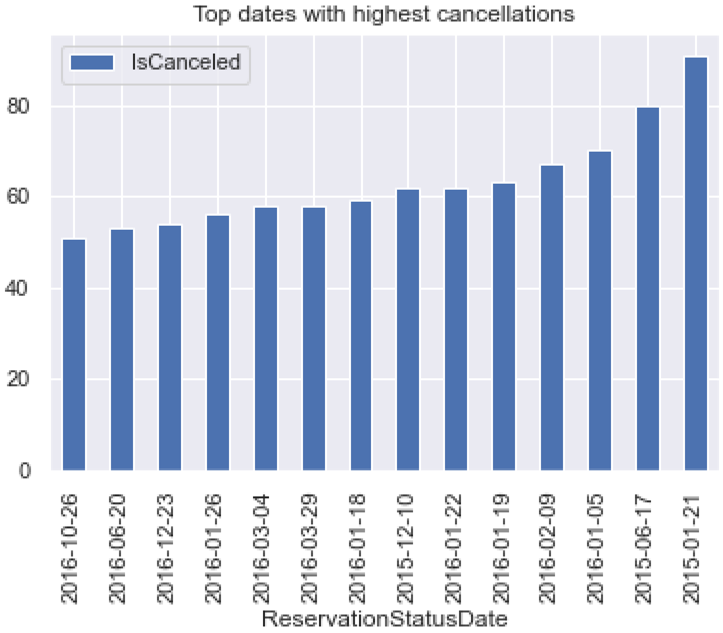
6. Cancellations by Assigned Room Type:

Similarly, we can see in the bar chart below that the cancellations of Room Type A is highest with Type B being the lowest.



7. Top Dates with Highest Cancellation:

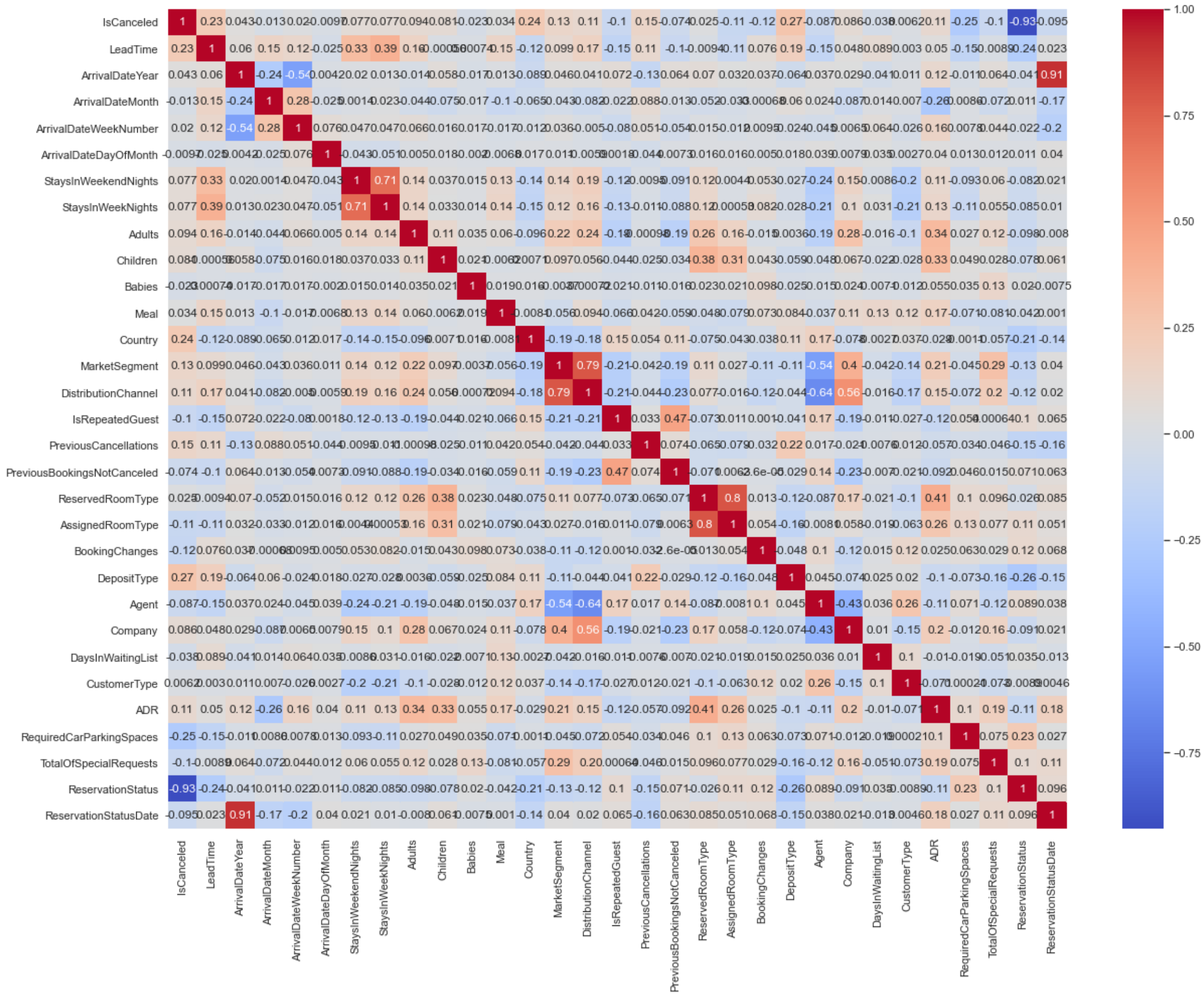
I did some exploration about the cancellation on a particular date. Below graph shows all the cancellation dates with more than 50 cancellations on a single day. 2015-01-21 has the highest number of cancellations with around 90 cancellations.



Preprocessing:

I did some preprocessing of the data before creating any Machine Learning prediction models. I found around 464 null values in Country so I removed the rows with null values in Country since it was very low and insignificant.

I encoded all the features using Label Encoder which gave a numerical representation to all the categorical variables. I used that to find correlations between features especially with isCanceled. The graph below is a heatmap showing the correlation between features after doing label encoding.



After that, I divided the training and testing set into 805 and 20% respectively and then trained and tested a Random Forest Classifier model which gave me an accuracy of around 99%. I had

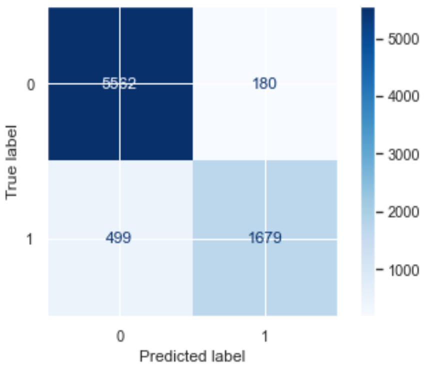
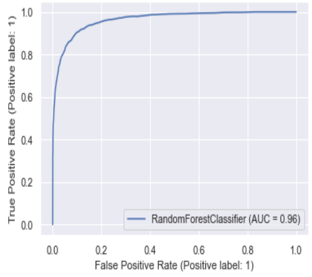
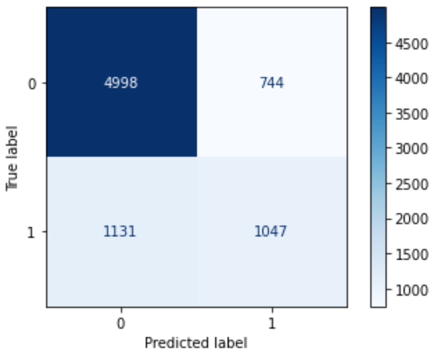
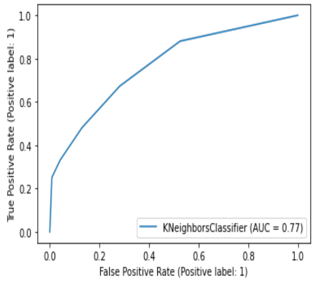
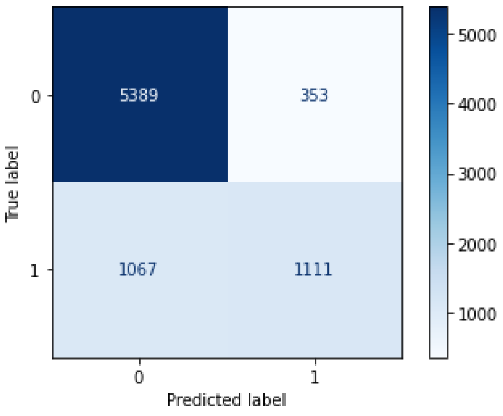
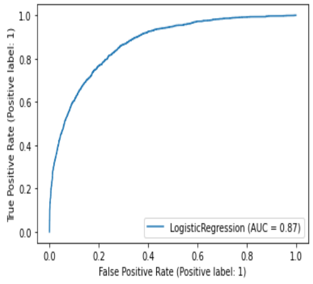
Reservation Status as one of my features in the training and testing set which is directly correlated with the IsCanceled variable. So I dropped that column to train and test my model and then came with an accuracy of around 94%.

However, sometimes using a Label encoder can be faulty since the prediction model might read some features as continuous variables instead of categorical variables that are encoded into random numbers. It has the disadvantage that the numeric values can be misinterpreted by algorithms as having some sort of hierarchy/order in them.

So to handle that problem, I instead used the One Hot Encoder to encode just the categorical features. Before building any model, I first dropped three columns Company, Agent, Reservation Status. Agent had more than 80% of the values as NULL so I thought it would be better to just drop the column. Similarly, Agent had more than 20% of the values as NULL and had very low correlation with our target variable IsCanceled, so instead of dropping any rows to disturb other columns, I just dropped the Agent Column. Similarly, I dropped ReservationStatus because it was highly and directly correlated with the IsCanceled which would make the prediction very obvious. So to make the prediction somewhat interesting, I dropped that column too. After that, I did one hot encoding of the categorical features- 'ArrivalDateMonth','Meal', ‘MarketSegment', DistributionChannel','ReservedRoomType',AssignedRoomType','DepositType', 'CustomerType', 'ReservationStatusDate'. I used this to build prediction models.

Building ML models:

I divided the training and testing set into 80-20%. After that, I built prediction models using three different types of classifiers and evaluated their performance to choose the best model. The three classifiers were- Random Forest Classifier, K-Neighbours Classifier, and Logistic Regression.
I evaluated the performance of models by comparing the accuracy, confusion matrix, F1- score and ROC curves. Accuracy is less relevant for an imbalanced classification problem since the distribution of data labels was pretty uneven. So, I used other metrics to evaluate that represents the data better, thus is a fair evaluation.

Models	Accuracy	Confusion Matrix	F1- score	ROC AUC
Random Forest	91.4 %		0= 94 % 1= 83%	
KNeighbours	76.3 %		0=0.84 1=0.53	
Logistic Regression	82.1%		0=0.88 1=0.61	

Thus, evaluating the models using above metrics, I chose RandomForest as the best model.

Hyperparameter Tuning:

I did hyperparameter tuning on my chosen RandomForestClassifier model. I used the below parameter grid to determine the best parameter set using GridSearchCV with 3-fold cross validation.

```
  param_grid = {
    'n_estimators': [200,400, 500],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [4,5,6,7,8],
    'criterion' :['gini', 'entropy']
}
```

Tuning hyper-parameters for precision, I came up with:

Best parameters set found on development set:

```
{'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'n_estimators': 500}
```

Tuning hyper-parameters for recall:0

Best parameters set found on development set:

```
{'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'n_estimators': 500}
```

Similarly, I did hyperparameter tuning on my second best model- LogisticRegression.. I used the below parameter grid to determine the best parameter set using GridSearchCV with 3-fold cross validation.

```
 param_grid = {
    'solver':['newton-cg', 'lbfgs', 'liblinear'],
    'max_iter':[100,50,200]
}
```

Tuning hyper-parameters for precision, I came up with:

Best parameters set found on development set:

```
{'max_iter': 200, 'solver': 'newton-cg'}
```

```
Tuning hyper-parameters for recall:
Best parameters set found on development set:
{'max_iter': 200, 'solver': 'newton-cg'}
```

Summary of Findings:

Some of the important finding were:

It was found that the repeated guests are less likely to cancel their reservations as opposed to not repeated guests. Similarly, August and January have the highest number of bookings and cancellations respectively. Transient customers have the highest number of bookings. Room A tends to be canceled the most and room B the least. 2015-01-21 had the highest number of cancellations among all other dates. And, in general the top 10 cancellation dates are in 2015 and 2016. After evaluating different prediction models, Random Forest Classifier is the most effective in predicting whether a customer is likely to cancel or not given different features.