DEPAUL
UNIVERSITY

# FINAL PROJECT REPORT

# HOTEL BOOKING DEMAND

NAWAAZ SHARIF
SYED NOOR RAZI ALI
MOHAMMED RASHIDUDDIN

DSC 478: PROGRAMMING MACHINE LEARNING APPLICATIONS

PROFESSOR BAMSHAD MOBASHER

## INTRODUCTION:

Vacations are important for everyone to clear their head and more importantly spending some quality time with their family. But as we know people do not travel the whole year and prefer travelling in good weather so that they can enjoy outdoors such as summer.

It is important for the hotel owners to know when they have peak number of guests and when they do not, so that they can make necessary arrangements to boost their revenue and make sure their guests have a good time staying at their hotel.

It is also important for the guests to book a hotel where they can enjoy and see that the hotel or resort have everything that they want to do on their vacation and plan accordingly or business men or clients likely to book the city hotel as they are much cheaper when compared to the resort hotel but doesn't provide much amenities when compared to the resort.

In this project, we will be targeting two variables – "hotel" and "is_cancelled" and implement machine learning (ML) algorithms on these variables and check how good they perform and how well they predict, calculating the accuracy from classification report and get confusion matrix to see how many rows the model predicted correct or wrong.

## CONTRIBUTION:

- NAWAAZ SHARIF: Exploratory Analysis, Data Visualization, KNN, Decision Trees.
- SYED NOOR RAZI ALI: Data Selection and Transformation Multi Nominal bayes and SVM.
- MOHAMMED RSHIDUDDIN: Data Cleaning, Random Forest, Stochastic Gradient Descent and NLP.

## DATASET OVERVIEW:

The dataset that we have done our project on is "Hotel Booking Demand" obtained from https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand.

This dataset has all the information required for booking a hotel for tourists when they plan to go for a vacation. The hotels are classified as resort hotel and city hotel. The dataset contains a total of 32 columns which has the attributes as reservation_date, reservation_status, length_of_stay, arrival_date_year, stays_in_weekend, stays_in_week_night, is_cancelled, agent, days_in_waiting_list, is_repeated_guest, and more. From this data we can see a lot of patterns – when people want to visit, which guest comes more frequently, which will help the owners/managers increase their profit by attracting guests by giving offers when there are chances of more guests travelling or visiting.

Our dataset comprises of 32 columns/attributes and 119391 rows.

## DATA CLEANING AND TRANSFOMATION:

Dropping the columns “company”, “reservation_status_date” and “reservation_status”, as company is the ID of the company that made the booking, reservation_status_date is a date that we won’t be using in our analysis, and reservation_status is same as is_cancelled.
Please refer Fig 1 for sum of the null values in the dataset.

Replacing the missing values in columns - “children” with 0.0, “country” with “unknown” which will act as a dummy variable as country is a categorical variable, and “agent” with 0.

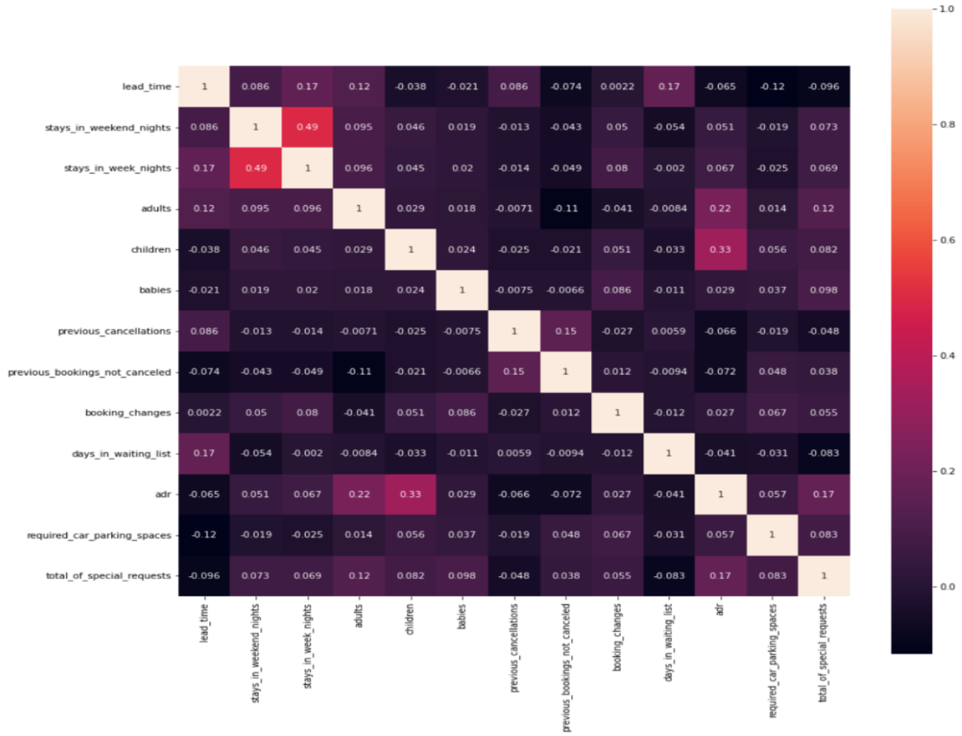
For the column “meal”, undefined and SC mean the same meal, so we will replace all undefined values in meal with SC.

There are some rows in our dataset that have “adults”, “children” and “babies” values as 0, so we will be dropping them as it does not make sense to process it.

## EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION:

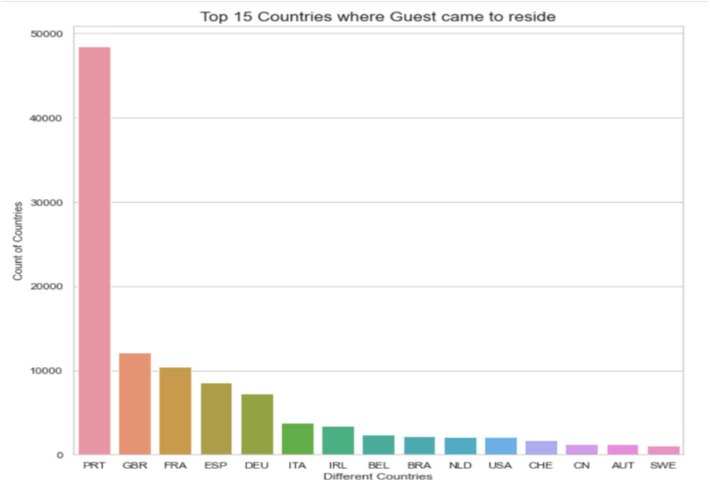
### Correlation Matrix

Before getting to implement our objective, we are making visualizations so that we get a basic idea of what our dataset has and what relation we have between our variables.



| | lead_time | stays_in_weekend_nights | stays_in_week_nights | adults | children | babies | previous_cancellations | previous_bookings_not_canceled | booking_changes | days_in_waiting_list | adr | required_car_parking_spaces | total_of_special_requests |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| lead_time | 1 | 0.086 | 0.17 | 0.12 | -0.038 | -0.021 | 0.086 | -0.074 | 0.0022 | 0.17 | -0.065 | -0.12 | -0.096 |
| stays_in_weekend_nights | 0.086 | 1 | 0.49 | 0.095 | 0.046 | 0.019 | -0.013 | -0.043 | 0.05 | -0.054 | 0.051 | -0.019 | 0.073 |
| stays_in_week_nights | 0.17 | 0.49 | 1 | 0.096 | 0.045 | 0.02 | -0.014 | -0.049 | 0.08 | -0.002 | 0.067 | -0.025 | 0.069 |
| adults | 0.12 | 0.095 | 0.096 | 1 | 0.029 | 0.018 | -0.0071 | -0.11 | -0.041 | -0.0084 | 0.22 | 0.014 | 0.12 |
| children | -0.038 | 0.046 | 0.045 | 0.029 | 1 | 0.024 | -0.025 | -0.021 | 0.051 | -0.033 | 0.33 | 0.056 | 0.082 |
| babies | -0.021 | 0.019 | 0.02 | 0.018 | 0.024 | 1 | -0.0075 | -0.0066 | 0.086 | -0.011 | 0.029 | 0.037 | 0.098 |
| previous_cancellations | 0.086 | -0.013 | -0.014 | -0.0071 | -0.025 | -0.0075 | 1 | 0.15 | -0.027 | 0.0059 | -0.066 | -0.019 | -0.048 |
| previous_bookings_not_canceled | -0.074 | -0.043 | -0.049 | -0.11 | -0.021 | -0.0066 | 0.15 | 1 | 0.012 | -0.0094 | -0.072 | 0.048 | 0.038 |
| booking_changes | 0.0022 | 0.05 | 0.08 | -0.041 | 0.051 | 0.086 | -0.027 | 0.012 | 1 | -0.012 | 0.027 | 0.067 | 0.055 |
| days_in_waiting_list | 0.17 | -0.054 | -0.002 | -0.0084 | -0.033 | -0.011 | 0.0059 | -0.0094 | -0.012 | 1 | -0.041 | -0.031 | -0.083 |
| adr | -0.065 | 0.051 | 0.067 | 0.22 | 0.33 | 0.029 | -0.066 | -0.072 | 0.027 | -0.041 | 1 | 0.057 | 0.17 |
| required_car_parking_spaces | -0.12 | -0.019 | -0.025 | 0.014 | 0.056 | 0.037 | -0.019 | 0.048 | 0.067 | -0.031 | 0.057 | 1 | 0.083 |
| total_of_special_requests | -0.096 | 0.073 | 0.069 | 0.12 | 0.082 | 0.098 | -0.048 | 0.038 | 0.055 | -0.083 | 0.17 | 0.083 | 1 |

From the above correlation matrix, we can see that few features have correlation of above 0.1 like ADR, adult children and the babies, the correlation of ADR and adults is of 0.2 which is low correlated, and the correlation of ADR and children is 0.33 which is the moderately correlated and the correlation of ADR, and babies is 0.029. The highest correlation is of stays_in_week_nights to stays_in_weekend_nights of 0.49 which is highly correlated when compared to all the variables.



From the above bar chart, we can infer that most of the guests came from Portugal to reside in the hotel as we progress, we can see a gradual decrease in count with a difference of 40,000 when compared to the top country and the second country. The following countries where the guests came from is Great Britain, France, Spain, and Germany with the count similar to 10,000 guests.

## MACHINE LEARNING ALGORITHMS PERFORMED:

Before implementing ML algorithms, we converted some of our columns that had labels as values to numerical form using LabelEncoder and performed min_max_scaler to normalize the data.

We have performed ML algorithms on two different target variables:

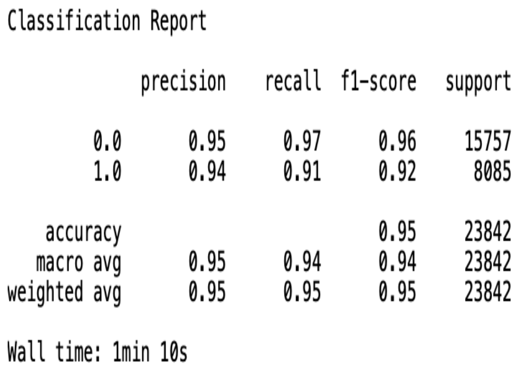
1. hotel
2. is_cancelled

Types of ML algorithms performed are:

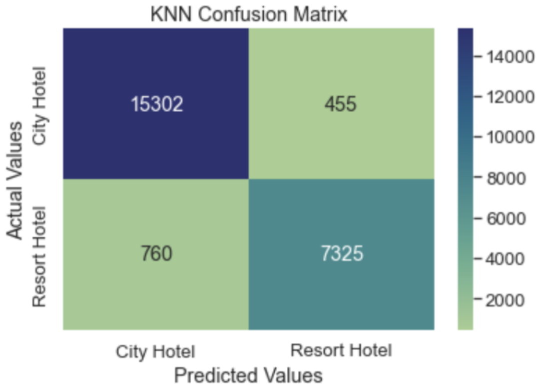
1. K-Nearest Neighbor (KNN)
2. Decision Tree
3. Multinomial Naïve Bayes
4. Support Vector Machine (SVM)
5. Random Forest
6. NLP Neural Network
7. Stochastic Gradient Descent

**Target Variable: Hotel**

1. **KNN:**



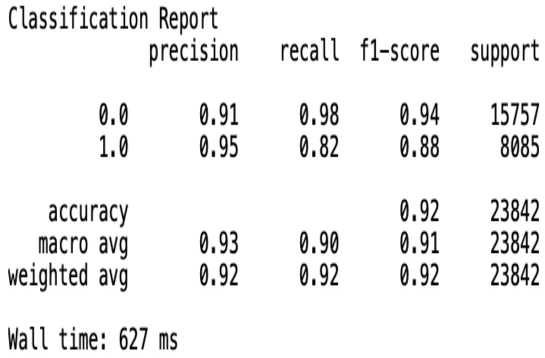
```
Classification Report

              precision    recall  f1-score   support

         0.0       0.95      0.97      0.96     15757
         1.0       0.94      0.91      0.92      8085

    accuracy                           0.95     23842
   macro avg       0.95      0.94      0.94     23842
weighted avg       0.95      0.95      0.95     23842

Wall time: 1min 10s
```



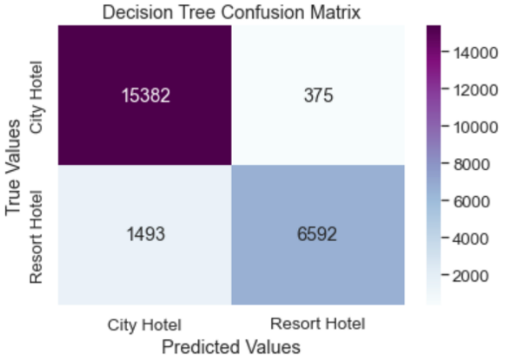
When we take the KNN as the classification metric and hotel as the target variable, we can see an accuracy of 95 percentage that means that the model was able to predict which hotel it was whether it was a city hotel, or a resort hotel, and it was able to achieve an accuracy of 95%.

Also, from the above confusion matrix we can infer that most of the hotels were predicted as city hotel since in our data said we had majority of the data points as of city hotel.

2. **Decision Tree:**

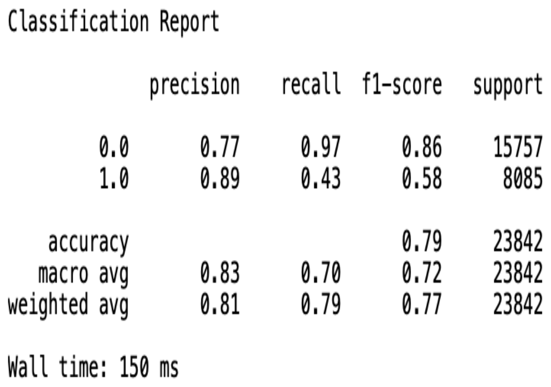


```
Classification Report
              precision    recall  f1-score   support

         0.0       0.91      0.98      0.94     15757
         1.0       0.95      0.82      0.88      8085

    accuracy                           0.92     23842
   macro avg       0.93      0.90      0.91     23842
weighted avg       0.92      0.92      0.92     23842

Wall time: 627 ms
```

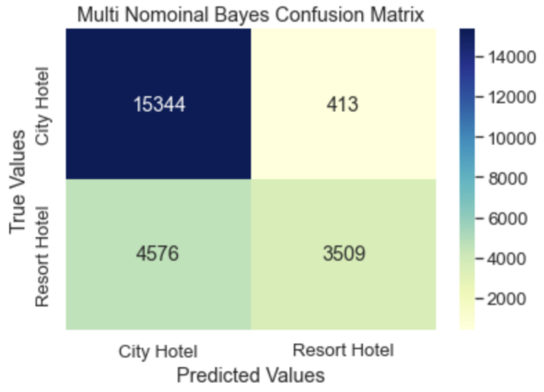


We got on accuracy of 92 percentage, while using decision tree classifier, which tells us that the model was able to predict between the resort hotel and the city hotel, and it is less when compared to KNN classifier.
Also from the above confusion matrix, we can infer that around 1500 observations were predicted wrong, which is a little less when compared to the KNN classifier.
Please refer Fig 2 in Appendix for Decision Tree for hotel as the target variable.

### 3. Multinomial Naïve Bayes:



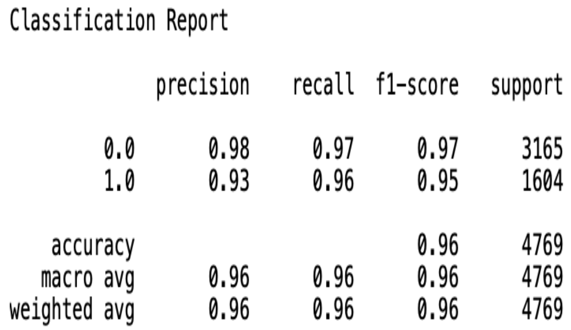
```
Classification Report

              precision    recall  f1-score   support

         0.0       0.77      0.97      0.86     15757
         1.0       0.89      0.43      0.58      8085

    accuracy                           0.79     23842
   macro avg       0.83      0.70      0.72     23842
weighted avg       0.81      0.79      0.77     23842

Wall time: 150 ms
```



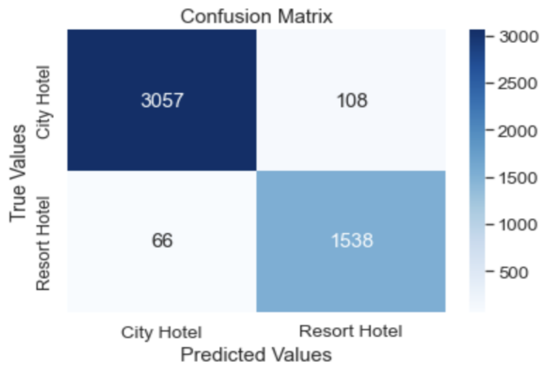
For the multi nominal bias, we can infer that the accuracy score is very less when compared to all the classification models. We have an accuracy of 79 percentage.

Also, for the confusion matrix for the multi nominal bayes, we can also see that around 5000 of the observations were labeled incorrectly.

### 4. SVM:



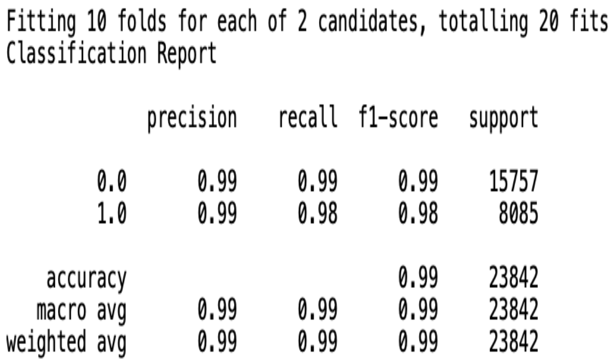
```
Classification Report

              precision    recall  f1-score   support

         0.0       0.98      0.97      0.97      3165
         1.0       0.93      0.96      0.95      1604

    accuracy                           0.96      4769
   macro avg       0.96      0.96      0.96      4769
weighted avg       0.96      0.96      0.96      4769
```



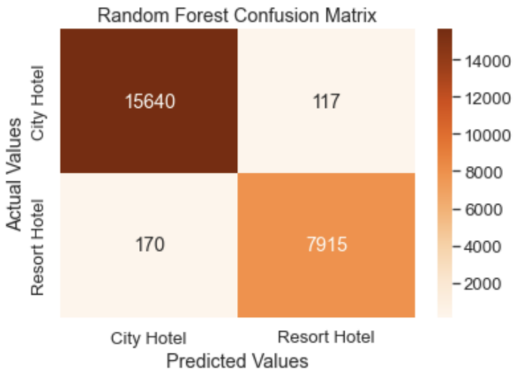
For the SVM we can see and accuracy of 96 percentage which infers that most of the hotels were predicted accurately about 96 percentage.

Also from the confusion matrix, we can infer that around 150 observations were labeled incorrectly, and we see a least observation among all the classification models, which were labeled incorrectly.

### 5. Random Forest:



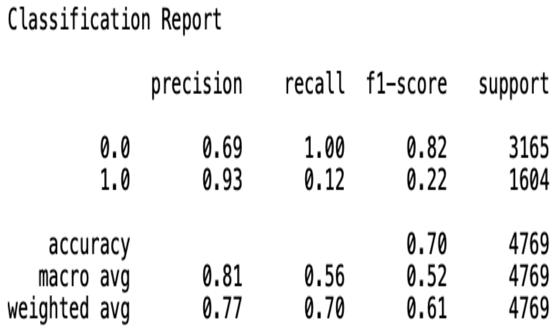
```
Fitting 10 folds for each of 2 candidates, totalling 20 fits
Classification Report

              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99     15757
         1.0       0.99      0.98      0.98      8085

    accuracy                           0.99     23842
   macro avg       0.99      0.99      0.99     23842
weighted avg       0.99      0.99      0.99     23842
```



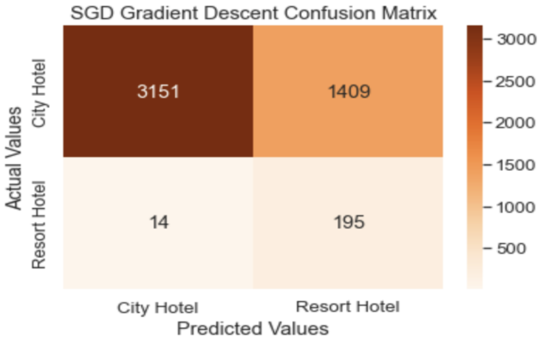
For the random forest, we can see a highest among all the models with an accuracy of 99 percentage, which says that most of the observations were predicted accurately about 99 percentage between the resort hotel, and the city hotel.

Also from the above confusion matrix, we can infer that around 300 observations labeled incorrectly as resort and city hotel.

### 6. Gradient Descent:

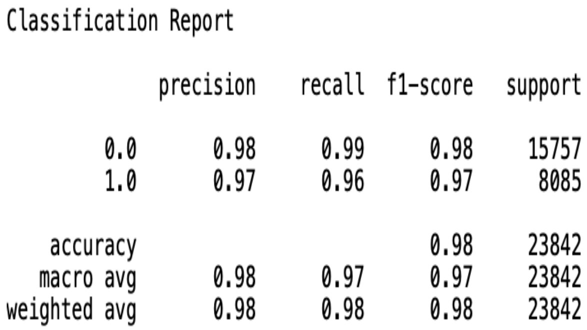


```
Classification Report

              precision    recall  f1-score   support

         0.0       0.69      1.00      0.82      3165
         1.0       0.93      0.12      0.22      1604

    accuracy                           0.70      4769
   macro avg       0.81      0.56      0.52      4769
weighted avg       0.77      0.70      0.61      4769
```

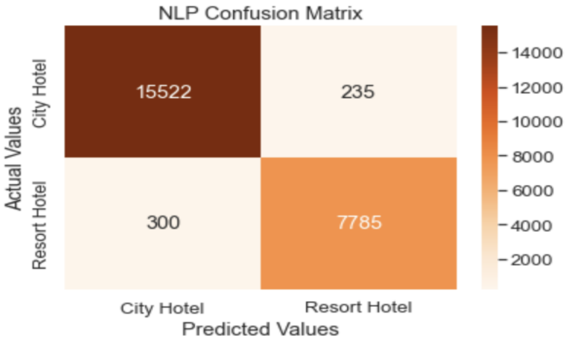


From the above SGD classification model, we can say that the accuracy is of 70 percentage, which is nearly less when compared to rest of the models, and we see that agent as well as ADR are the most contributed coefficient of SGD.

### 7. NLP Neural Networks:

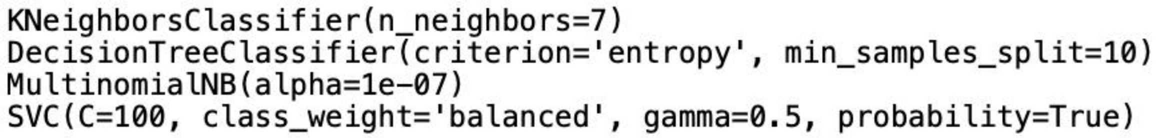


```
Classification Report

              precision    recall  f1-score   support

         0.0       0.98      0.99      0.98     15757
         1.0       0.97      0.96      0.97      8085

    accuracy                           0.98     23842
   macro avg       0.98      0.97      0.97     23842
weighted avg       0.98      0.98      0.98     23842
```

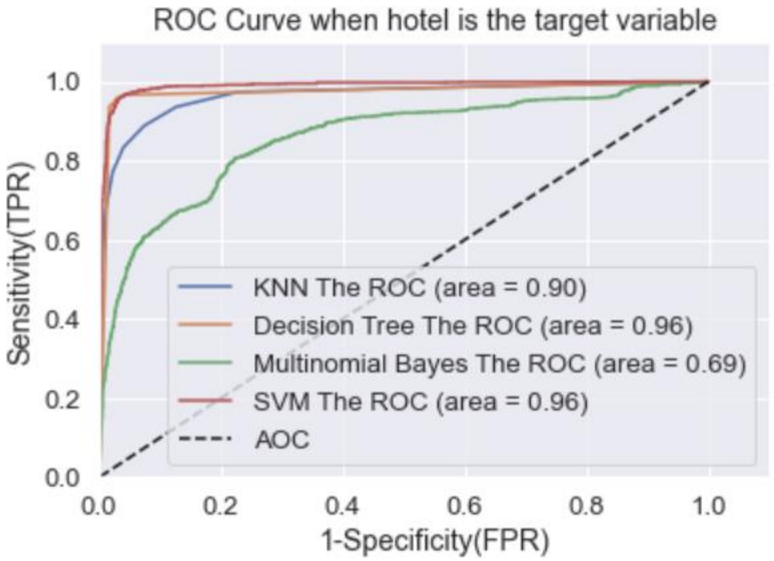


We also see the highest score of 98% when we take neural networks is the classification motor with an accuracy of 98 percentage on with the RMSE value of 0.14.

**ROC Curve:**



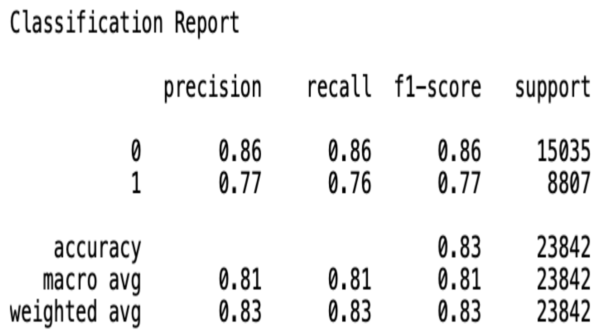
```
KNeighborsClassifier(n_neighbors=7)
DecisionTreeClassifier(criterion='entropy', min_samples_split=10)
MultinomialNB(alpha=1e-07)
SVC(C=100, class_weight='balanced', gamma=0.5, probability=True)
```



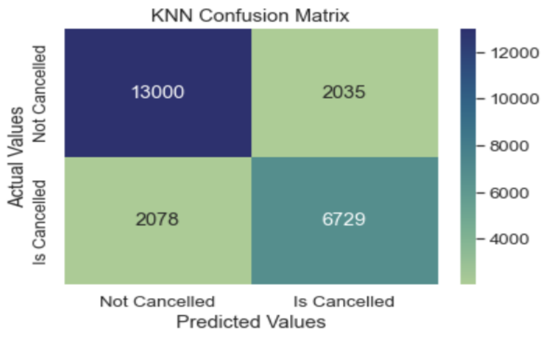
The above ROC graph shows us the performance of different models applied with "hotel" as the target variable. KNN, Decision Tree and SVM performed good as they have high AUC, whereas Multinomial Naive Bayes has low AUC. For our dataset we can choose both Decision tree and SVM as they both have an accuracy of 0.96.

**Target Variable: is_cancelled**

1. **KNN:**

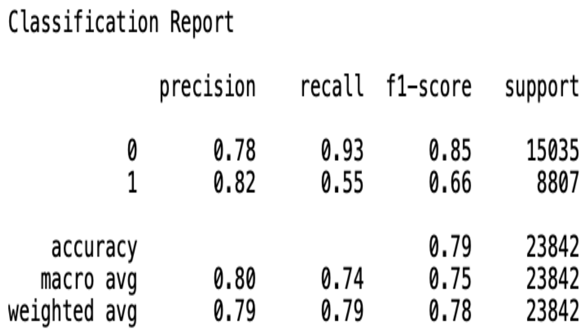


```
Classification Report

              precision    recall  f1-score   support

           0       0.86      0.86      0.86     15035
           1       0.77      0.76      0.77      8807

    accuracy                           0.83     23842
   macro avg       0.81      0.81      0.81     23842
weighted avg       0.83      0.83      0.83     23842
```

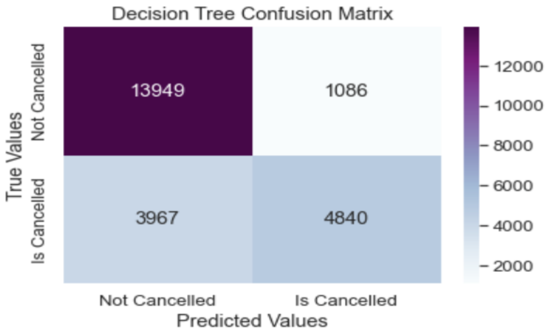


When we consider is canceled as our target variable and apply our classification model of KNN. We get an accuracy of the 83 percentage, and we also see that the similar number of is canceled and not cancelled observations are labeled incorrectly.

2. **Decision Tree:**

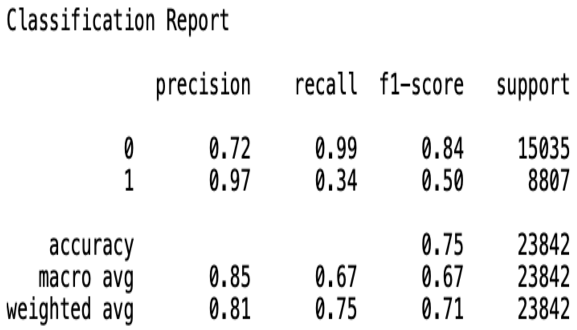


```
Classification Report

              precision    recall  f1-score   support

           0       0.78      0.93      0.85     15035
           1       0.82      0.55      0.66      8807

    accuracy                           0.79     23842
   macro avg       0.80      0.74      0.75     23842
weighted avg       0.79      0.79      0.78     23842
```

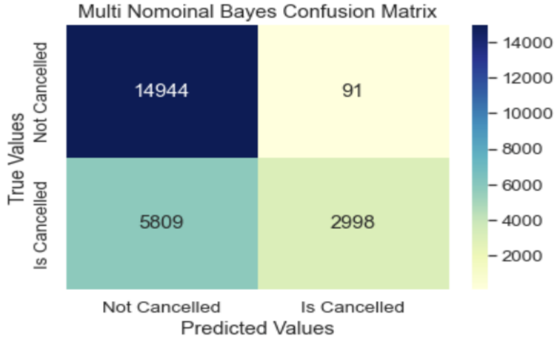


We see an accuracy of 79 percentage for the decision tree, which is a little bit less when compared to our previous model KNN we have also visualized the tree with a depth of 5, and we also infer that around 5000 observations well label incorrectly for is canceled and not canceled. Please refer Fig 3 in Appendix for Decision Tree for is_cancelled target variable.

### 3. Multinomial Naïve Bayes:

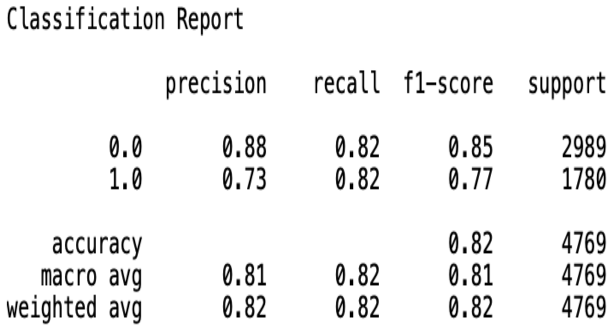


```
Classification Report

              precision    recall  f1-score   support

           0       0.72      0.99      0.84     15035
           1       0.97      0.34      0.50      8807

    accuracy                           0.75     23842
   macro avg       0.85      0.67      0.67     23842
weighted avg       0.81      0.75      0.71     23842
```

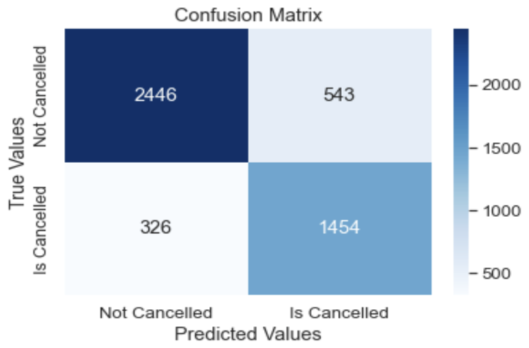


For multi nominal bayes, we also see a least accuracy when compared to the rest of the models of 75 percentage, and we see a highest amount 6000 of the observations which were labeled incorrectly.

### 4. SVM:

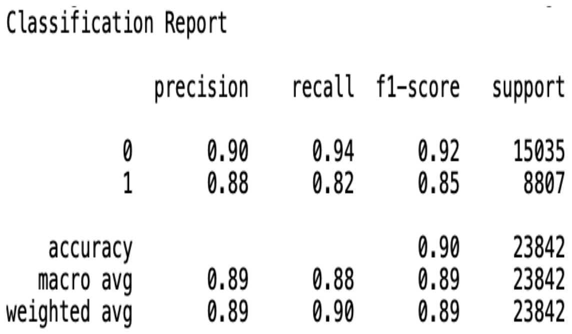


```
Classification Report

              precision    recall  f1-score   support

         0.0       0.88      0.82      0.85      2989
         1.0       0.73      0.82      0.77      1780

    accuracy                           0.82      4769
   macro avg       0.81      0.82      0.81      4769
weighted avg       0.82      0.82      0.82      4769
```

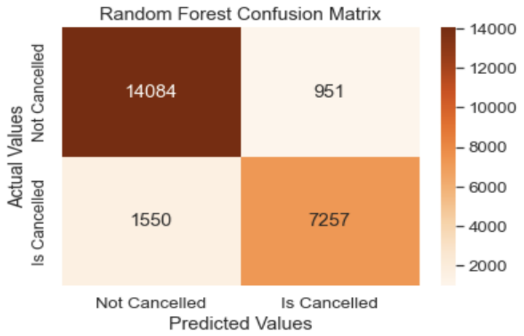


For the SVM when we consider, it’s canceled as our target variable, we see an accuracy of 82 percentage, and we see the least number of observations which were labeled incorrectly of thousand records.

### 5. Random Forest:

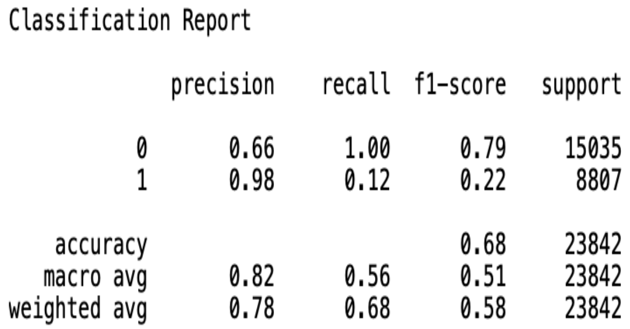


```
Classification Report

              precision    recall  f1-score   support

           0       0.90      0.94      0.92     15035
           1       0.88      0.82      0.85      8807

    accuracy                           0.90     23842
   macro avg       0.89      0.88      0.89     23842
weighted avg       0.89      0.90      0.89     23842
```

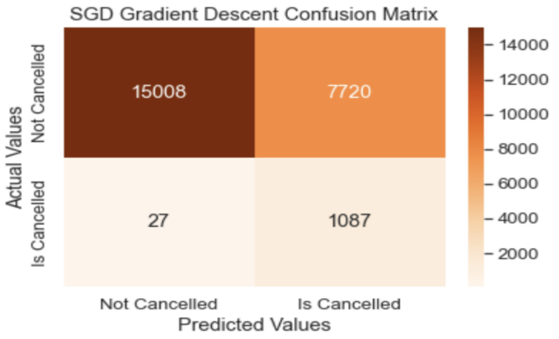


The best model which has perform accurately when we consider is canceled as the target variable and for the random forest, we achieved with a highest accuracy of 90 percentage and from the confusion matrix we see around 2500 of observations were labeled incorrectly, and most of labels were what product it has is canceled when compared to not canceled and this is an improvement when compared to the rest of the classification models.

### 6. Gradient Descent:

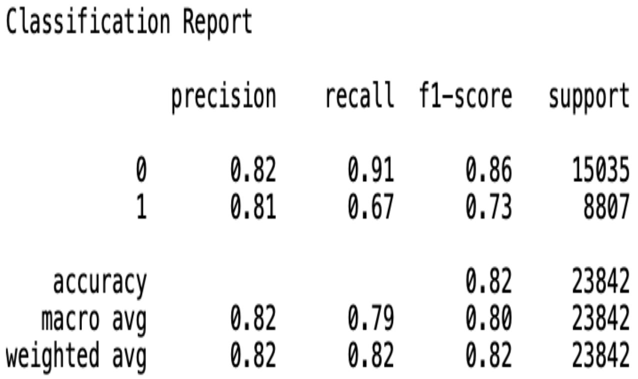


```
Classification Report

              precision    recall  f1-score   support

           0       0.66      1.00      0.79     15035
           1       0.98      0.12      0.22      8807

    accuracy                           0.68     23842
   macro avg       0.82      0.56      0.51     23842
weighted avg       0.78      0.68      0.58     23842
```

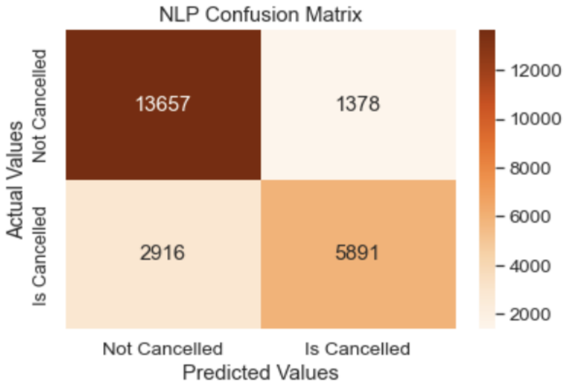


For the gradient descent we see a least accuracy of 68 percentage, and we also see the contributing coefficient of SGD which are deposit type and total of special request.

### 7. NLP Neural Network:

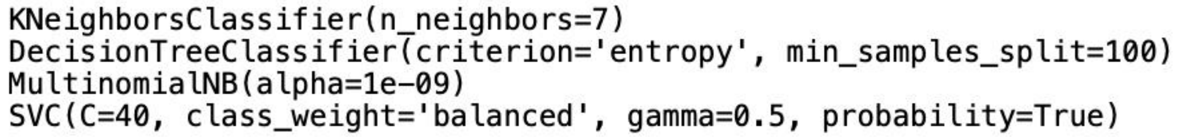


```
Classification Report

              precision    recall  f1-score   support

           0       0.82      0.91      0.86     15035
           1       0.81      0.67      0.73      8807

    accuracy                           0.82     23842
   macro avg       0.82      0.79      0.80     23842
weighted avg       0.82      0.82      0.82     23842
```

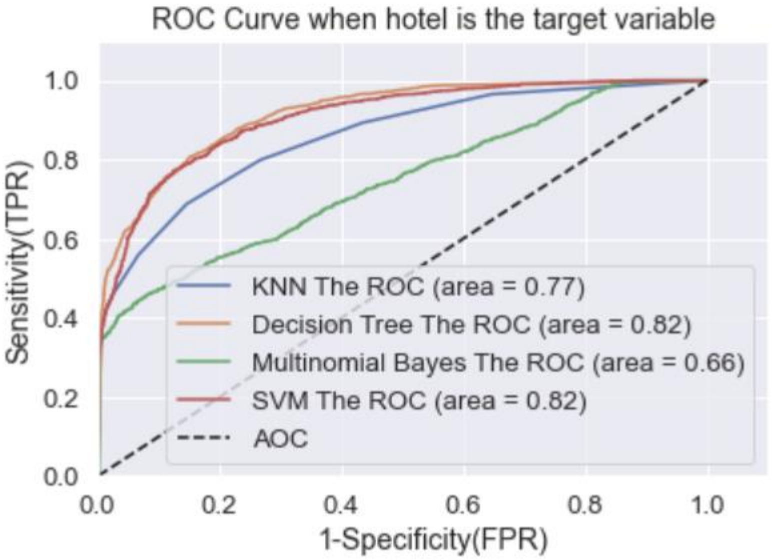


When we run neural network on our data set and considering is cancelled as our target variable, we achieved an accuracy of 82 percentage, which infers that the model was able to predict 83 percentage of is canceled and not canceled records from the data set.

**ROC Curve:**



```
KNeighborsClassifier(n_neighbors=7)
DecisionTreeClassifier(criterion='entropy', min_samples_split=100)
MultinomialNB(alpha=1e-09)
SVC(C=40, class_weight='balanced', gamma=0.5, probability=True)
```



The above ROC graph shows us the performance of different models applied with "is_cancelled" as the target variable. Decision Tree and SVM performed good as they have high AUC, whereas KNN and Multinomial Naive Bayes has low AUC. For our dataset we can choose both SVM and Decision tree as they both have an accuracy of 0.82.

## CONCLUSION:

Overall, the models we created performed good on both the target variables – hotel and is_cancelled.

With “hotel” as the target variable:

1. Random Forest performed best with the Accuracy of 0.99, RMSE value of 0.109 and score of 0.9884.
2. NLP has the accuracy of 0.98, RMSE value of 0.149
3. SVM has the accuracy of 0.96, RMSE value of 0.191 and score of 0.9588.
4. KNN has the accuracy of 0.95, RMSE value of 0.225 and score of 0.946.
5. Decision Tree has the accuracy of 0.92, RMSE value of 0.279 and score of 0.980.
6. Multinomial Naive Bayes has the accuracy of 0.79, RMSE value of 0.457 and score of 0.7911.
7. Gradient Descent has the accuracy of 0.70, RMSE value of 0.321 and score of 0.5556.

With “is_cancelled” as the target variable:

1. Random Forest performed best with the Accuracy of 0.90, RMSE value of 0.32 and score of 0.894.
2. KNN has the accuracy of 0.83, RMSE value of 0.415 and score of 0.946.
3. NLP has the accuracy of 0.82, RMSE value of 0.424.
4. SVM has the accuracy of 0.82, RMSE value of 0.426 and score of 0.828.
5. Decision Tree has the accuracy of 0.79, RMSE value of 0.460 and score of 0.8589.
6. Multinomial Naive Bayes has the accuracy of 0.75, RMSE value of 0.497 and score of 0.753.
7. Gradient Descent has the accuracy of 0.68, RMSE value of 0.39 and score of 0.342.

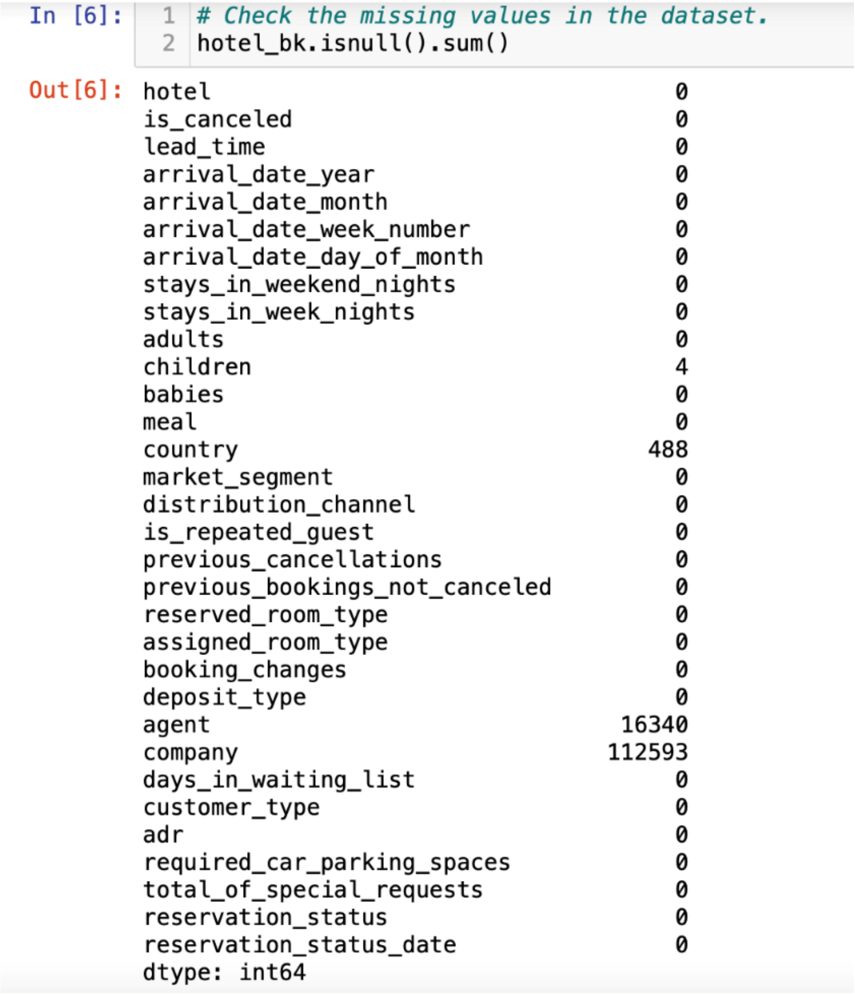
## **FUTURE SCOPE:**

In the future to improve the models to be more productive and informative, we can add more hotels of that city or state or country. All the hotel and resort owners can use this model to see when and what additional preparations need to be done so that they can accommodate more guests, which will be helpful for them to generate more revenue.

# **Appendix:**

Data Cleaning

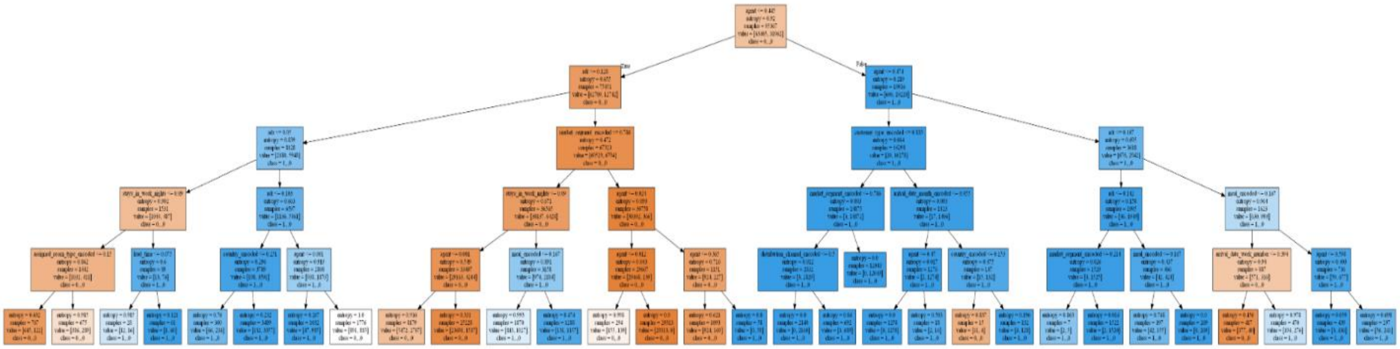
Fig 1:



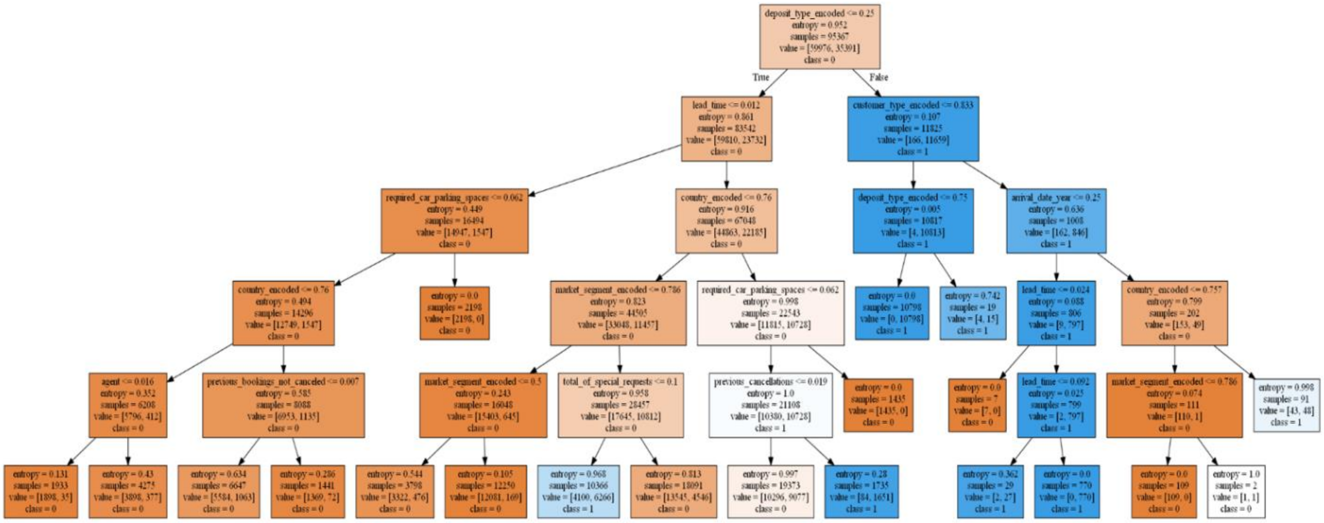
```
In [6]:  # Check the missing values in the dataset.
         hotel_bk.isnull().sum()

Out[6]:  hotel                                  0
         is_canceled                            0
         lead_time                              0
         arrival_date_year                      0
         arrival_date_month                     0
         arrival_date_week_number               0
         arrival_date_day_of_month              0
         stays_in_weekend_nights                0
         stays_in_week_nights                   0
         adults                                 0
         children                               4
         babies                                 0
         meal                                   0
         country                              488
         market_segment                         0
         distribution_channel                   0
         is_repeated_guest                      0
         previous_cancellations                 0
         previous_bookings_not_canceled         0
         reserved_room_type                     0
         assigned_room_type                     0
         booking_changes                        0
         deposit_type                           0
         agent                              16340
         company                           112593
         days_in_waiting_list                   0
         customer_type                          0
         adr                                    0
         required_car_parking_spaces            0
         total_of_special_requests              0
         reservation_status                     0
         reservation_status_date                0
         dtype: int64
```

Decision Tree when using Hotel as the target variable.

Fig 2:



Decision Tree when using is_cancelled as the target variable.
Fig 3:



## Jupyter File:

DSC 478
Final_Project.ipynb

## HTML File:

DSC 478
Final_Project.html