CAPSTONE PROJECT

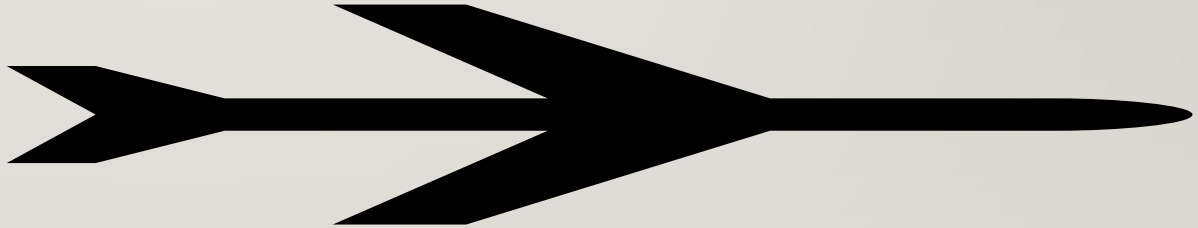
AIRLINE PASSENGER REFERRAL PREDICTION

BY

TILAK R

OBJECTIVE

- The data provided includes airline reviews from 2006 to 2019 for popular airlines worldwide with multiple choice and free text questions.
- The data is scraped in the spring of 2019. The main goal is to predict whether passengers will recommend airline to their friends.



METHODOLOGY

- Data Information
- Exploratory Data Analysis
- Feature Engineering
- Model Building
- Machine learning Models
- Conclusion

INFORMATION RELATED TO DATA

- The dataset has 16 variables of which "recommended“ is the dependent variable and the rest are independent variables.
- Data size is (131895.17)i.e. we have 131895 rows with 17 columns.
- There are many null and duplicate values in the dataset, so we need to clean the data first.
- The data set is a mixture of categorical and numerical data , so we need to organize and code the data before feeding it into the ML model.

INFORMATION RELATED TO DATA (CONTD..)

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 131895 entries, 0 to 131894
Data columns (total 17 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   airline           65947 non-null   object 
 1   overall           64017 non-null   float64
 2   author            65947 non-null   object 
 3   review_date       65947 non-null   object 
 4   customer_review   65947 non-null   object 
 5   aircraft          19718 non-null   object 
 6   traveller_type    39755 non-null   object 
 7   cabin             63303 non-null   object 
 8   route             39726 non-null   object 
 9   date_flown        39633 non-null   object 
 10  seat_comfort      60681 non-null   float64
 11  cabin_service     60715 non-null   float64
 12  food_bev          52608 non-null   float64
 13  entertainment     44193 non-null   float64
 14  ground_service    39358 non-null   float64
 15  value_for_money   63975 non-null   float64
 16  recommended       64440 non-null   object 
dtypes: float64(7), object(10)
```

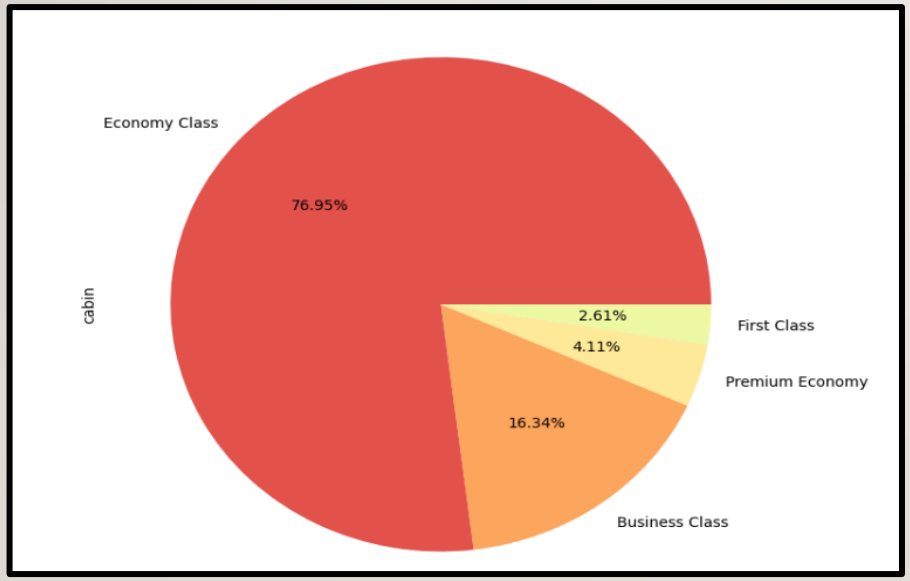
FEATURE DESCRIPTION

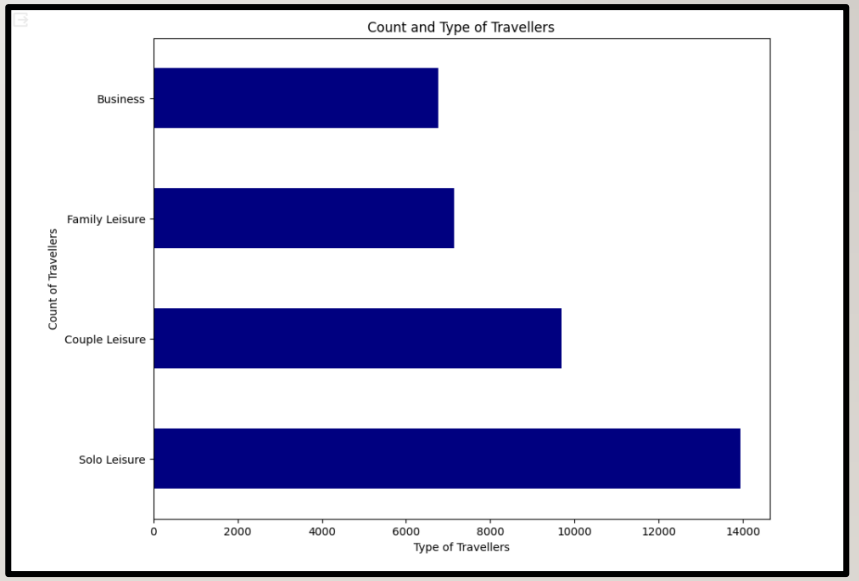
Data descriptions:

- **airline**: Name of the airline.
- **overall**: Overall point is given to the trip between 1 to 10.
- **author**: Author of the trip
- **reviewdate**: Date of the Review customer review: Review of the customers in free text format
- **aircraft**: Type of the aircraft
- **travellertype**: Type of traveler (e.g. business, leisure)
- **cabin**: Cabin at the flight date flown: Flight date
- **seatcomfort**: Rated between 1-5
- **cabin service**: Rated between 1-5
- **foodbev**: Rated between 1-5 entertainment: Rated between 1-5
- **groundservice**: Rated between 1-5
- **valueformoney**: Rated between 1-5

EXPLORATORY DATA ANALYSIS

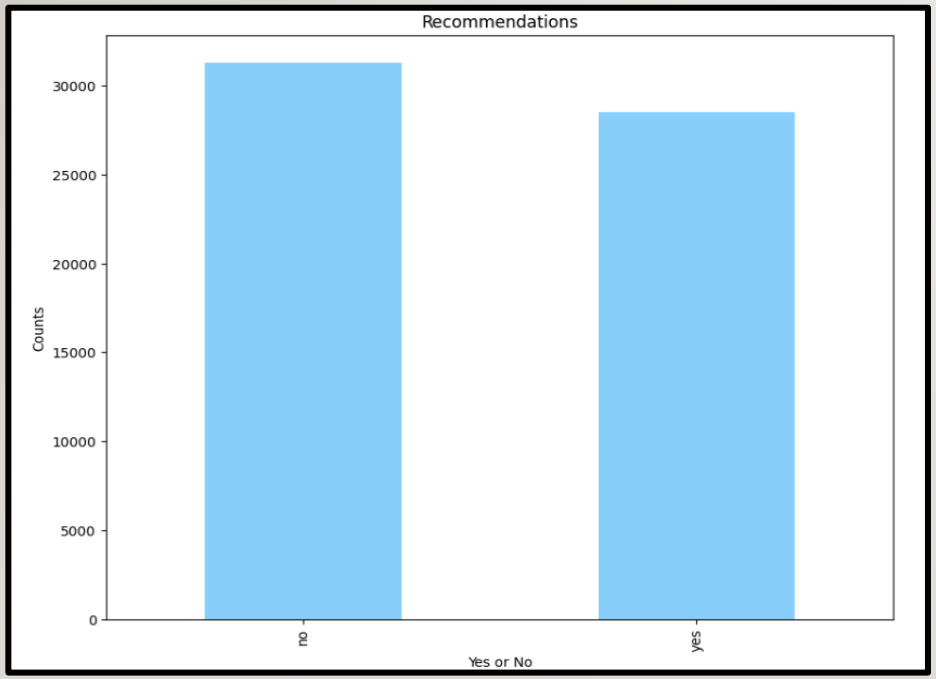
EDA for passengers class, Traveller Type

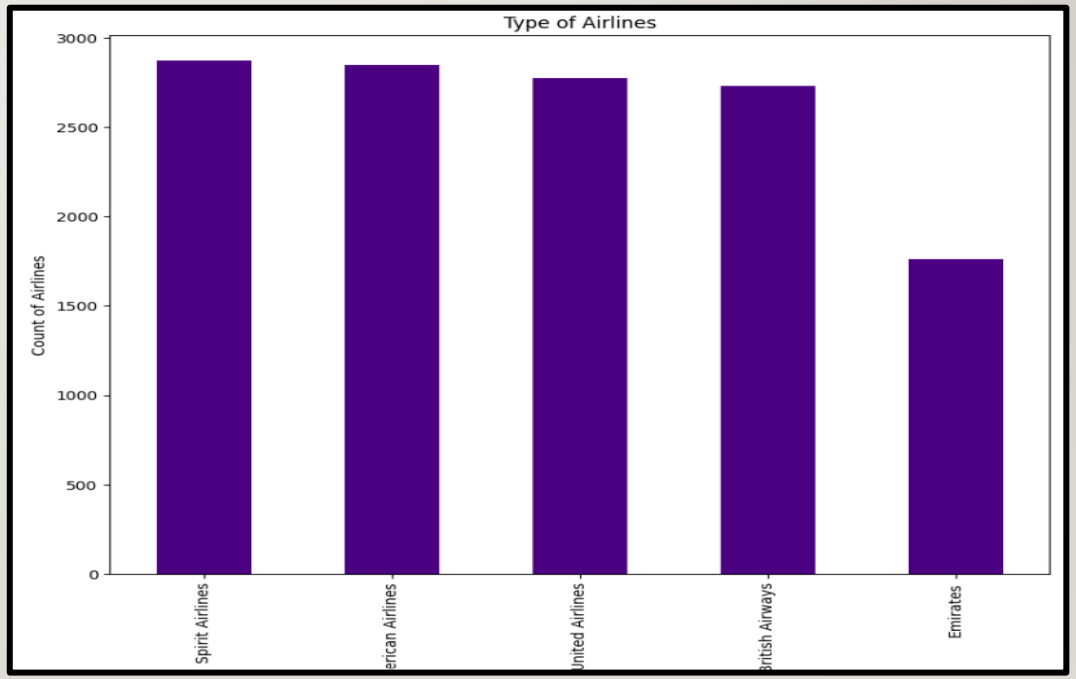




CONTINUED..

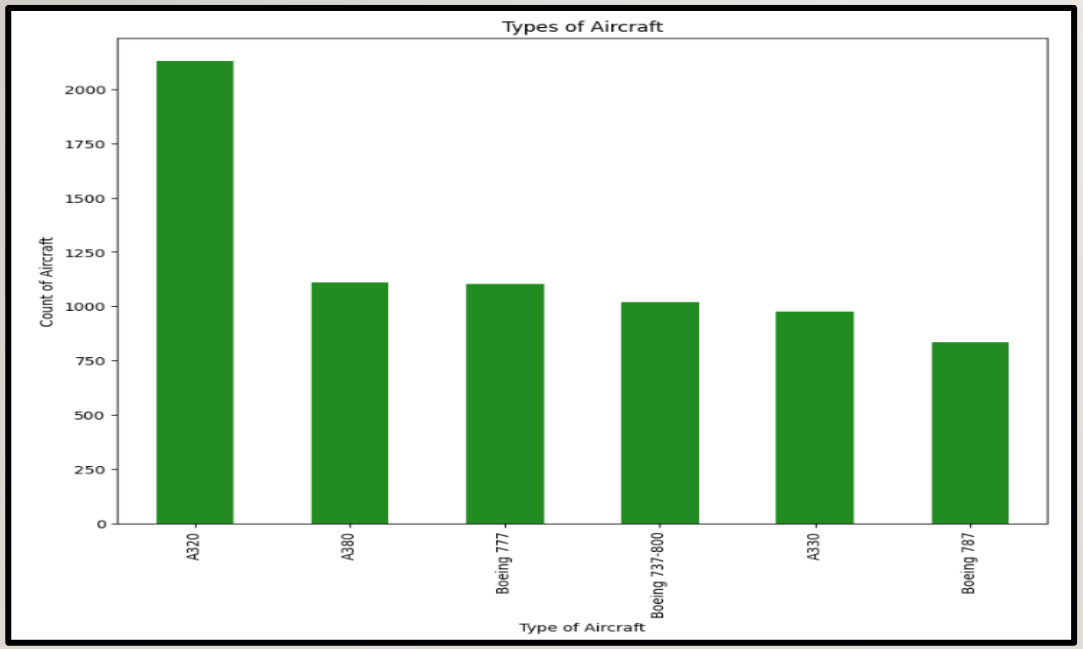
EDA for Recommendations and Airline type

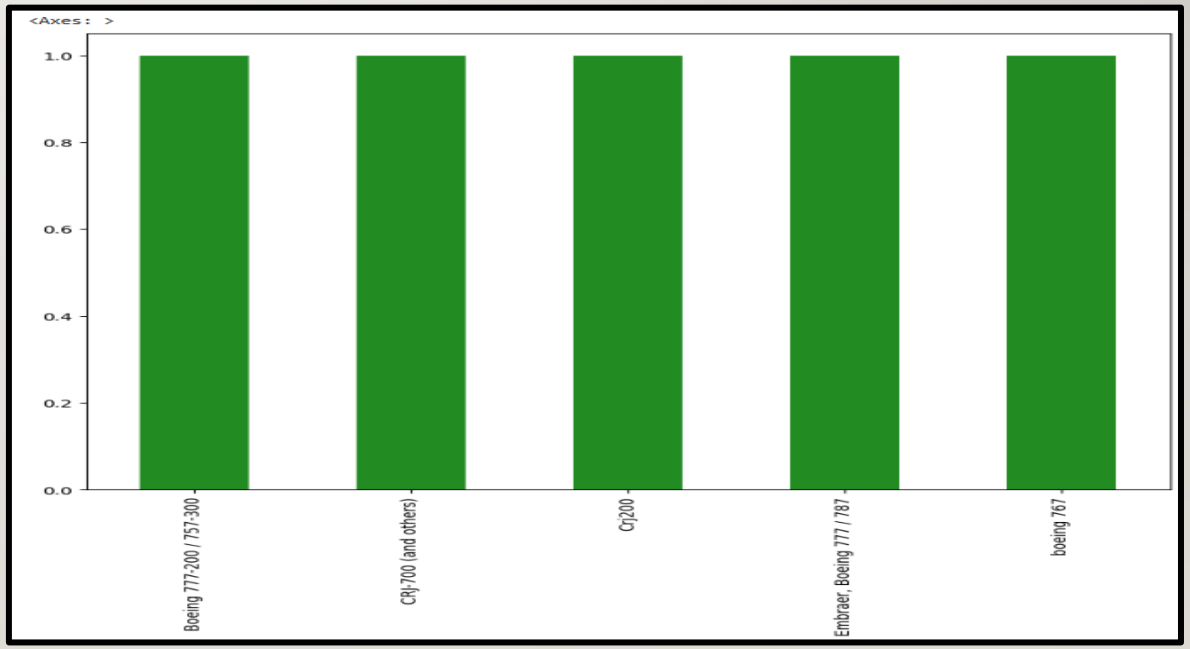




CONTINUED..

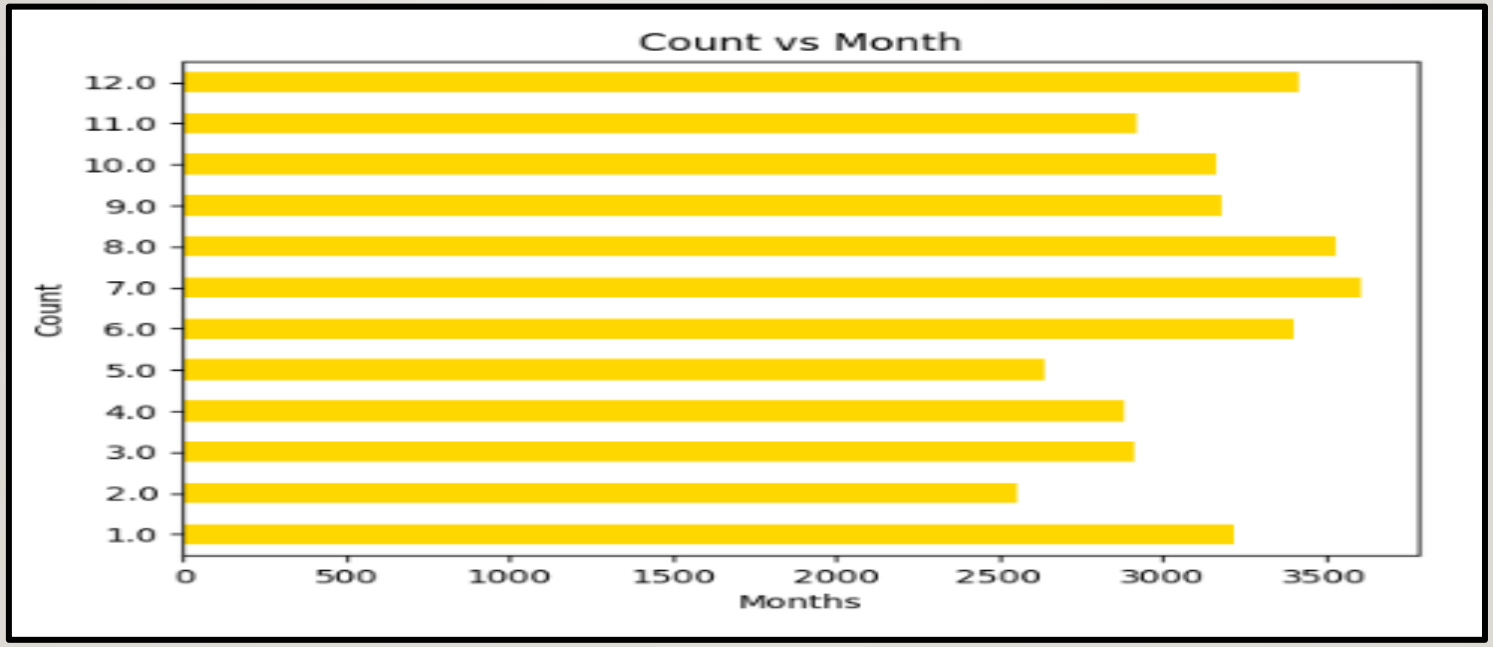
EDA for Most Frequently used and Rarely used Aircrafts



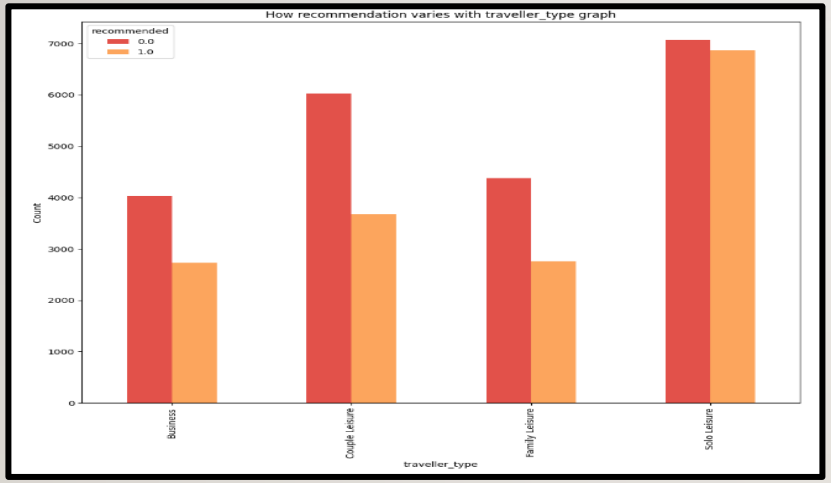


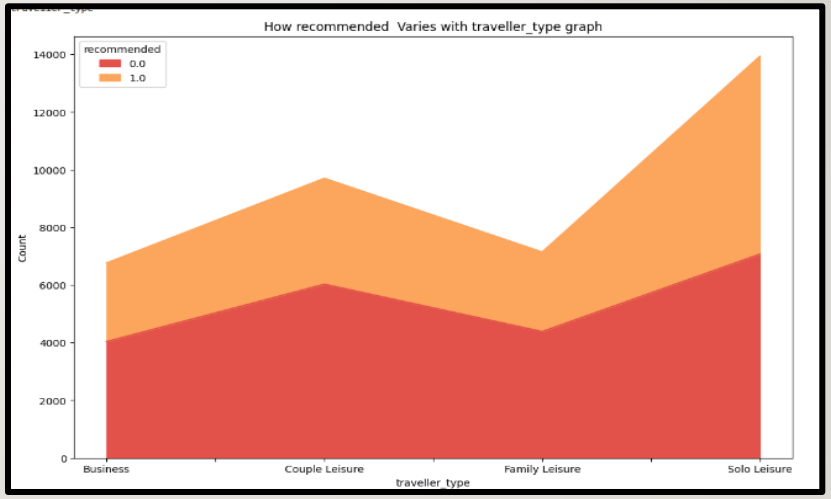
CONTINUED..

EDA for Date flown and Month Flown



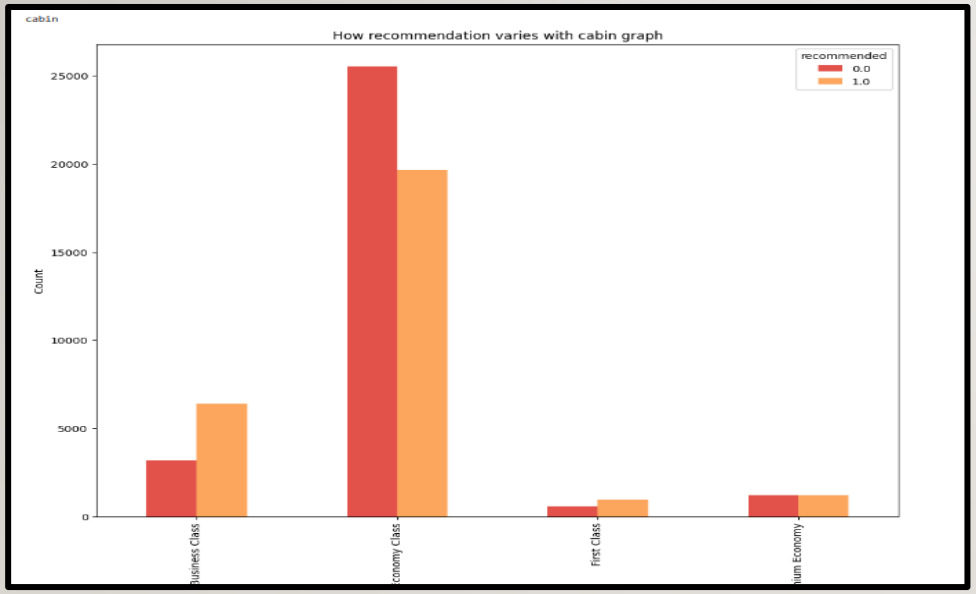
CONTINUED..

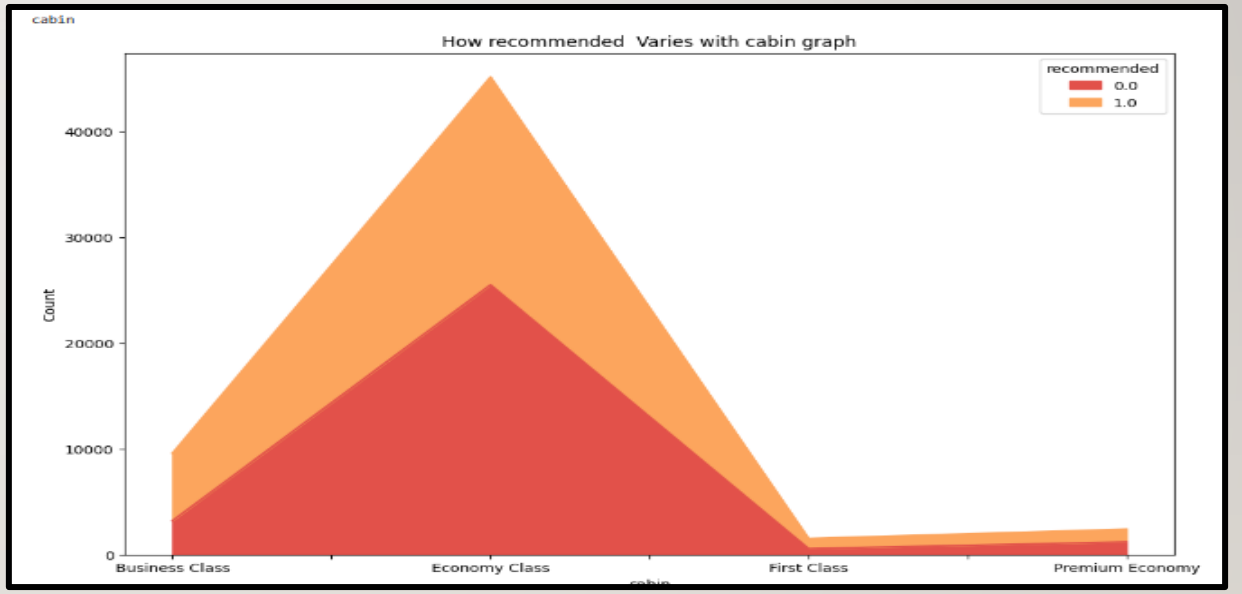




In traveler type, It is evident that people have given both 0 and 1 which we can take it as positive and negative recommendation to interpret to this to family type. positive recommendation given may be due to low price and negative recommendation may be due to lack of proper services and infrastructure. But this is just and assumption from the data we have received.

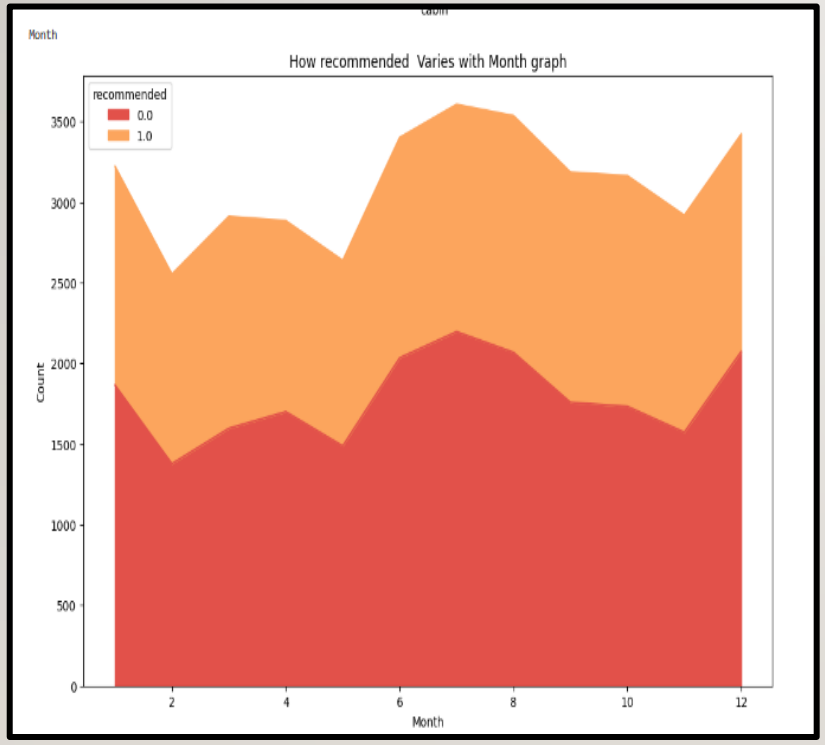
CONTINUED..

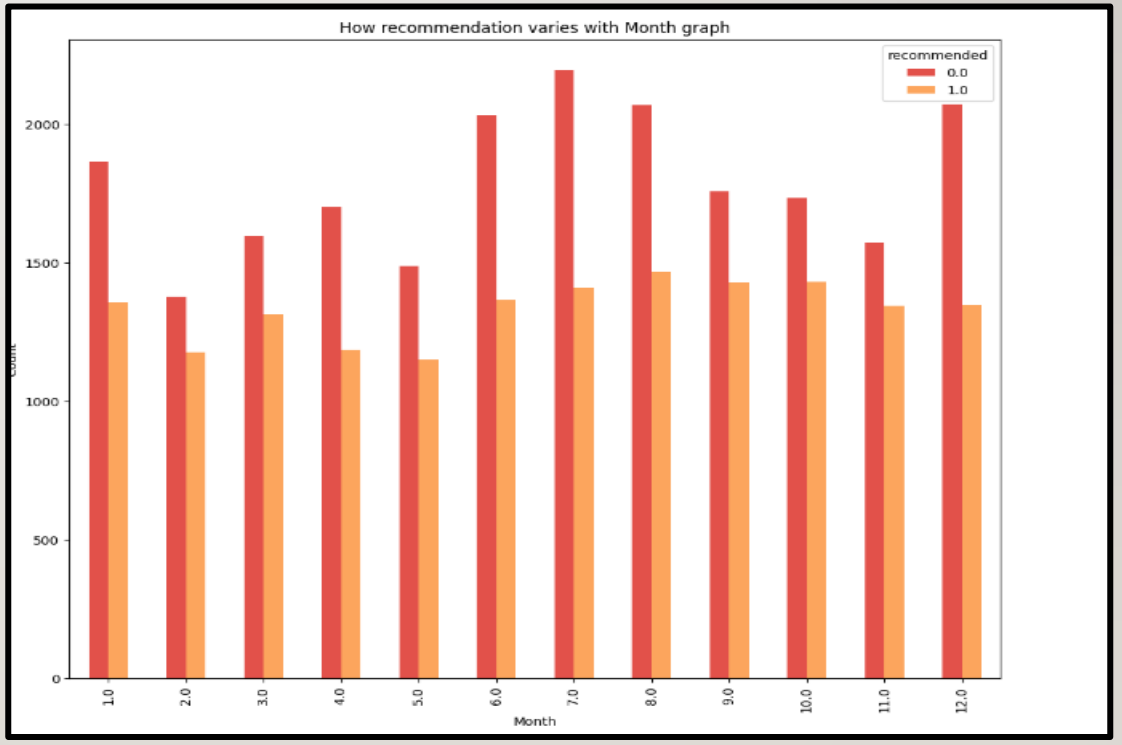




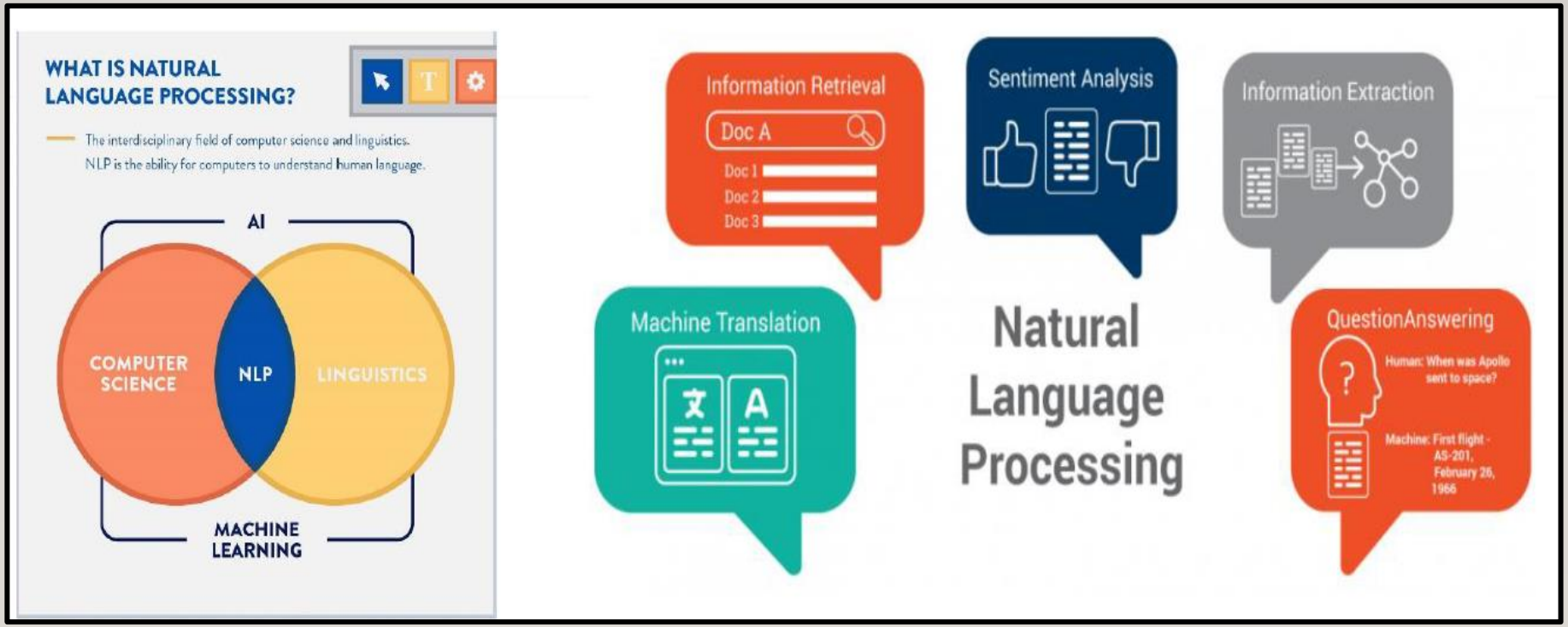
•In cabin type, It is evident that more passengers prefer Economy class cabin type out of which many of them have given negative recommendation due to lack of proper infrastructure and services and positive recommendation may be due to low prices and offers. It is also noticed that business class has highest positive recommendation which might be due to better service and infrastructure and negative recommendation due to high price.

CONTINUED..


Month
How recommended Varies with Month graph
recommended
0.0
1.0
Count
Month


How recommendation varies with Month graph
recommended
0.0
1.0
Month

NATURAL LANGUAGE PROCESSING



CONTINUEDD.

- We have used Vander sentiment in NLP so to convert sentiments in customer review into scores to have our model prediction.
- We have also created new feature numeric reviews to store sentiment score we have retrieve during sentiment analysis from customer review feature.

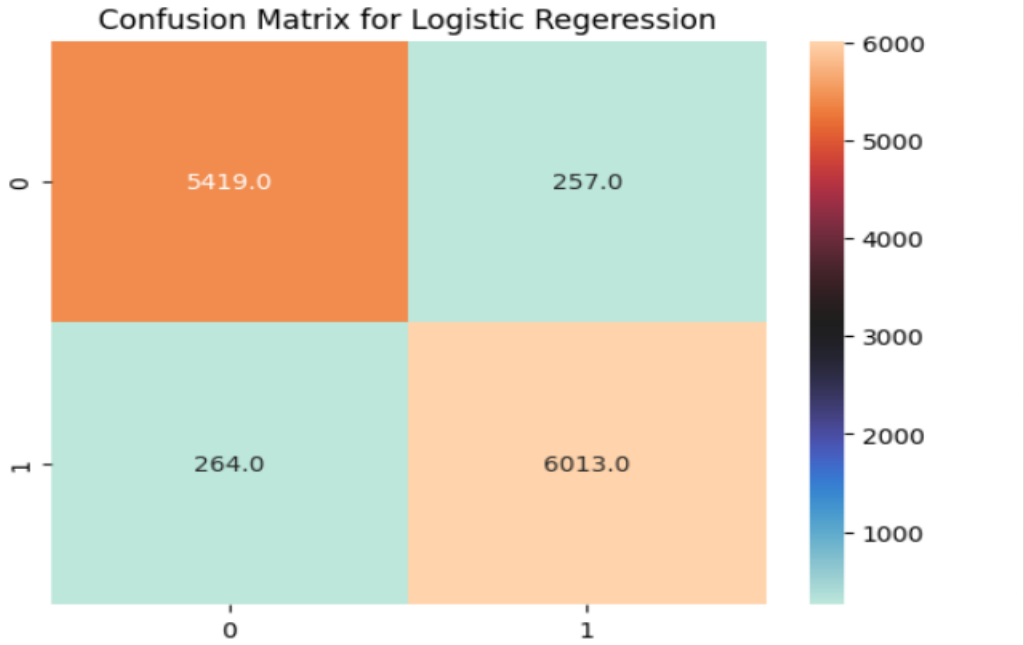
MODEL BUILDING..

- This experiment includes various models such as,

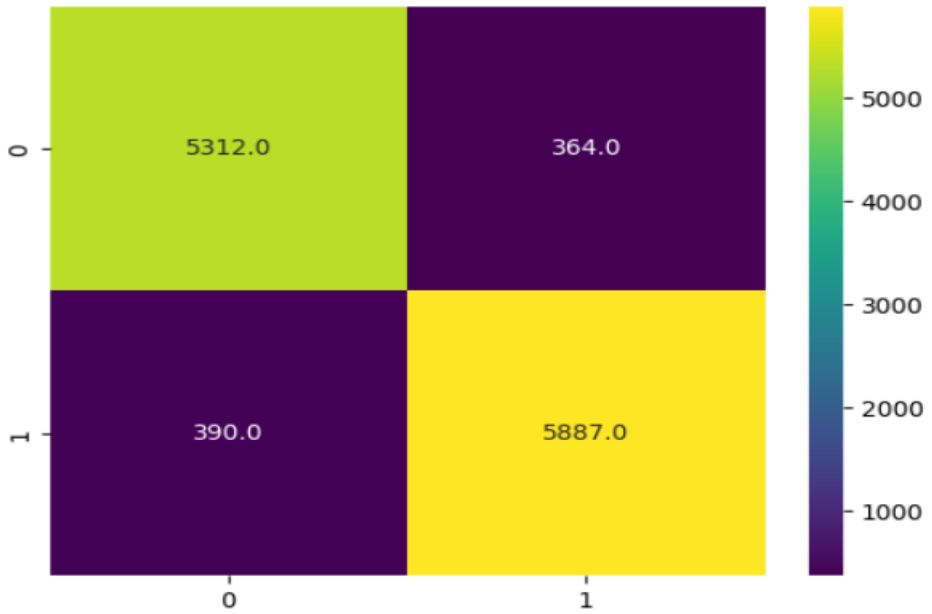
1. Random Forest
2. Logistic regression
3. Decision Tree
4. Random Forest with GridsearchCV
5. XGBOOST
6. K-Nearest Neighbour
7. K-Nearest Neighbour with GridsearchCV
8. Support vector Machine

MODEL BUILDING..

```
Accuracy score % of the model is 95.64%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```

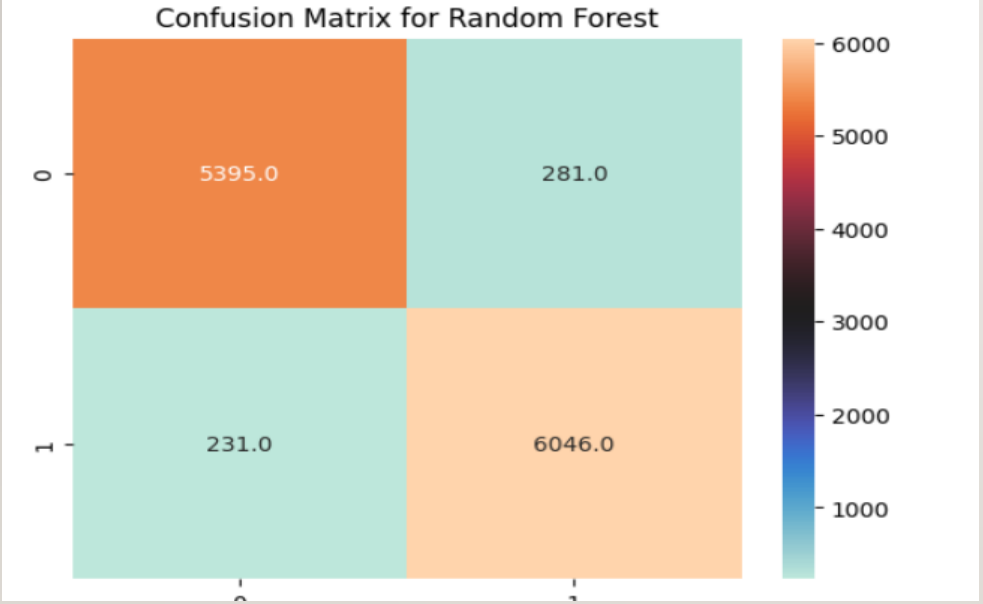



```
Accuracy score % of the model is 93.69%

Confusion matrix for Decision Tree
<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```

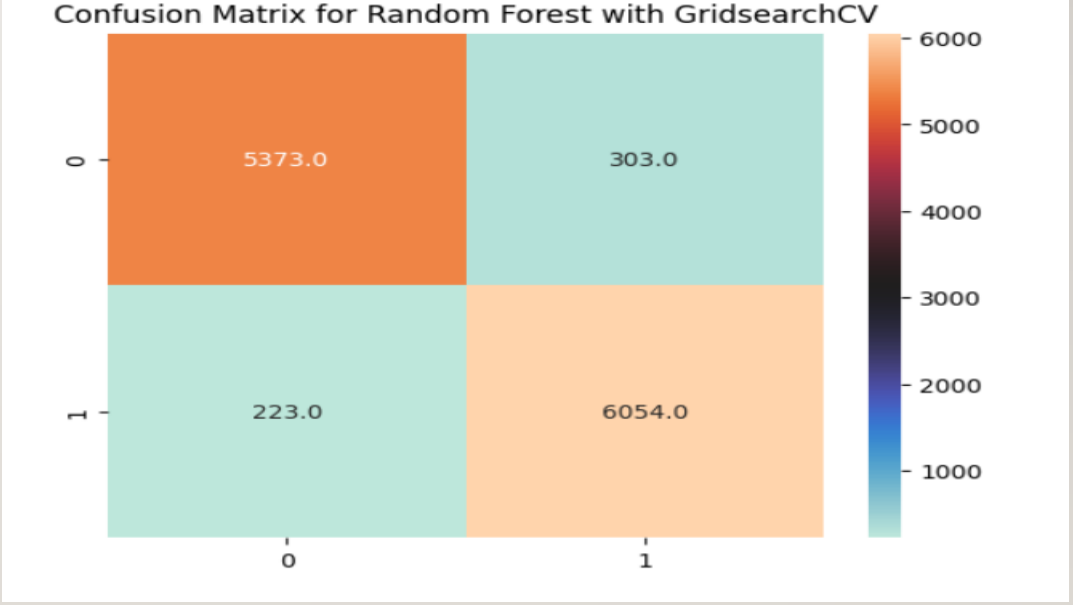


MODEL BUILDING..

```
Accuracy score % of the model is 95.72%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```

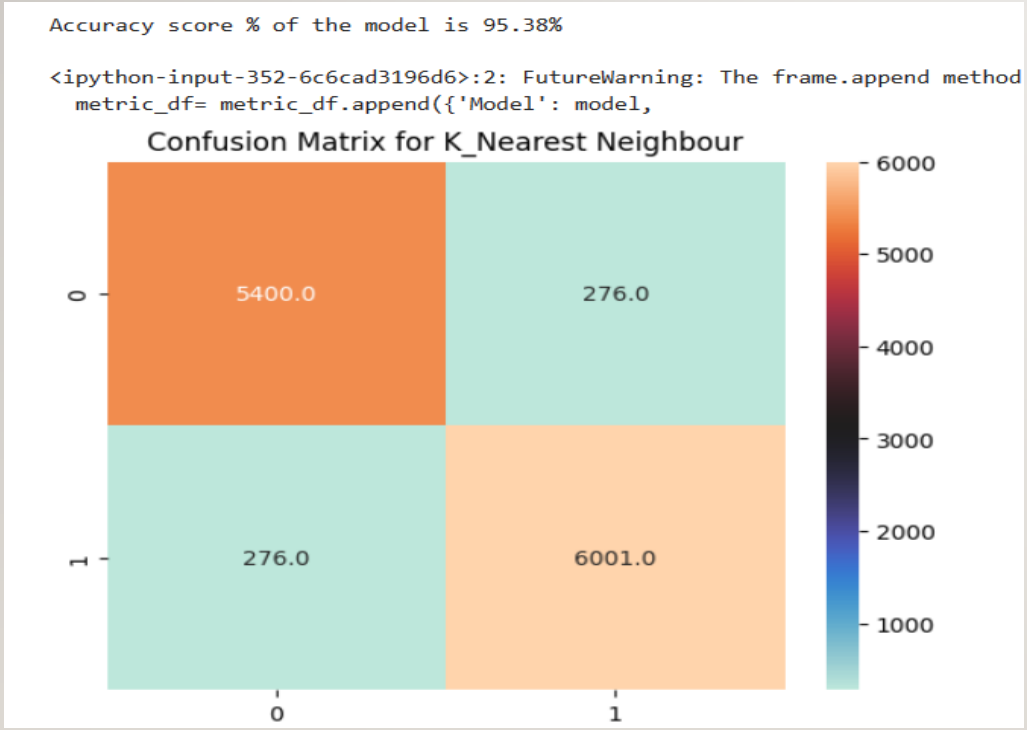



```
Accuracy score % of the model is 95.6%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```

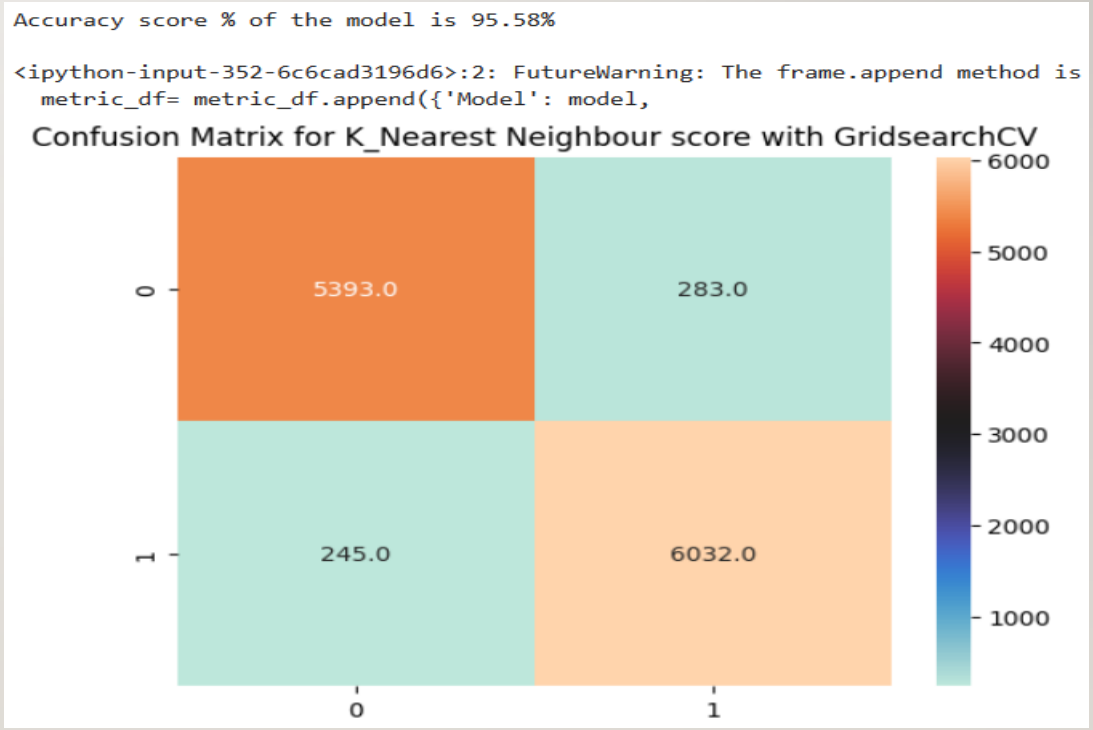


MODEL BUILDING..

```
Accuracy score % of the model is 95.38%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```

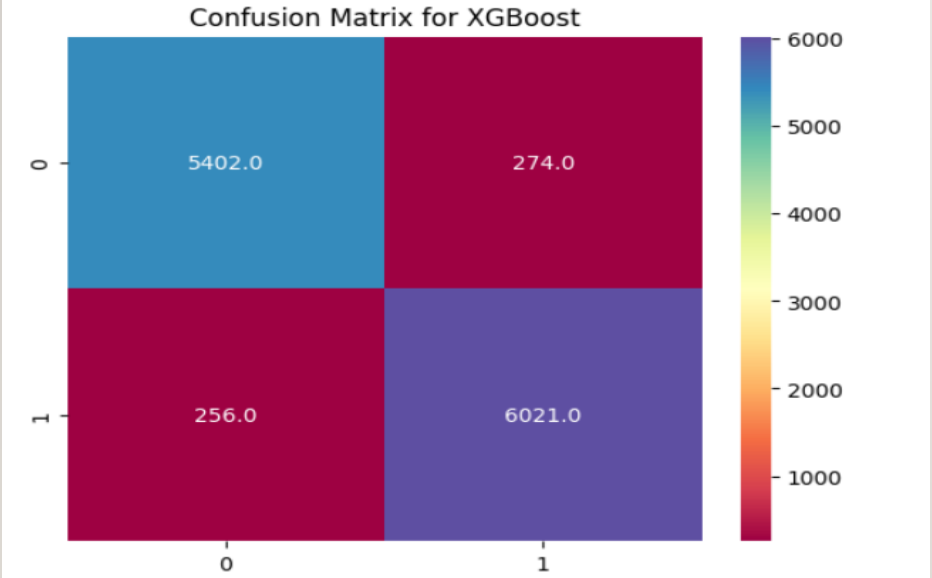



```
Accuracy score % of the model is 95.58%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method is
  metric_df= metric_df.append({'Model': model,
```

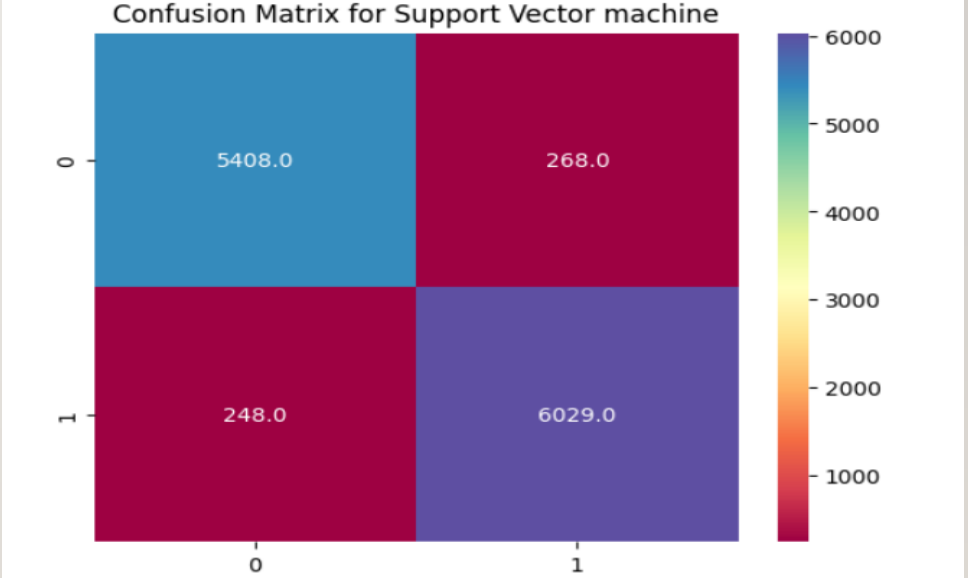


MODEL BUILDING..

```
Accuracy score % of the model is 95.57%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```




```
Accuracy score % of the model is 95.68%

<ipython-input-352-6c6cad3196d6>:2: FutureWarning: The frame.append method
  metric_df= metric_df.append({'Model': model,
```



MODEL BUILDING RESULTS

	Model	Accuracy	Precision	Recall	f1-score	roc_auc_score
0	Logistic Regression	0.956413	0.953546	0.954722	0.954133	0.956332
1	Decision Tree	0.936920	0.931603	0.935870	0.933732	0.936869
2	Random Forest	0.957166	0.958941	0.950493	0.954698	0.956846
3	Random Forest with GridSearchCV	0.955994	0.960150	0.946617	0.953336	0.955545
4	K_Nearest Neighbour	0.953819	0.951374	0.951374	0.951374	0.953702
5	K_Nearest Neighbour score with GridsearchCV	0.955827	0.956545	0.950141	0.953332	0.955555
6	XGBoost	0.955660	0.954754	0.951727	0.953238	0.955471
7	Support Vector machine	0.956831	0.956153	0.952784	0.954465	0.956637

CONCLUSION

- We can see that people have given both 1 or 0 which we will consider from now on as positive and negative recommendation so to interpret it effectively to the solo leisure. This may because of the poor infrastructure or the service received by the people and positive recommendation may be because of low price for solo. But this is approximate analysis based on the data provided.
- Also we can see that people gives the high positive recommendation to economic class in cabin. From this we can conclude that people love to travel in economic class as of low price also in same way we can see people give highest negative recommendation to economy class maybe because less infrastructure or service provided to them.
- Also we can see people have given highest positive recommendation to Business class it may be because of the quality of service provided to them in Business class and similarly negative recommendation because of high price of business class or less travelling percentage.

CONCLUSION(CONTD)

- From month vs no. of recommendation. We can see that people tents to travel most in the month of July considering the total of positive and negative recommendation combined.
- From overall vs recommended graph we can see which is perfectly understandable that negative recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10. But it is very true that highest negative recommendation has been given to overall rating of 1.0 which is really a matter of concern.

CONCLUSION (CONTD)

- In seat comfort people has given highest positive recommended to the seat of class 5 as compared to very low negative recommendation to the same. Also we can see seat of class 1 have been given highest negative recommendation as compare to its positive recommendation. Here we come to a conclusion it must be removed as early as possible.
- In cabin service rating people has given highest recommendation to rating to cabin service rating 5 as compare to its counterpart. From this we can conclude that cabin service is doing pretty good.

CONCLUSION (CONTD)

- In food and beverage rating people have given highest negative recommendation to rating 1.0 from this we can conclude that airline service has to improve their food delivery and quality service.
- In entertainment also we can see most people has given highest negative recommendation to entertainment rating 1 which shows that airline has to improve their entertainment system as well.
- In ground service also we can see most people has given highest negative recommendation to ground service rating 1 which shows that airline has to improve their ground service.

CONCLUSION (CONTD)

- In value for money also we can see most people has given highest negative recommendation to value for money rating 1 which shows that airline has to make their flight service more cost effective.
- In model Selection we can see that Random Forest has highest accuracy followed by and Support Vector Machine and Logistic Regression Model. But we can also see that recall, precision, f1-score and Roc_Auc_Score of Random Forest model combined is giving higher score than all other models hence we have chosen Random Forest Model for further prediction.