Capstone Project

Airline Passenger Referral Prediction

By

Ajinkya Jumde

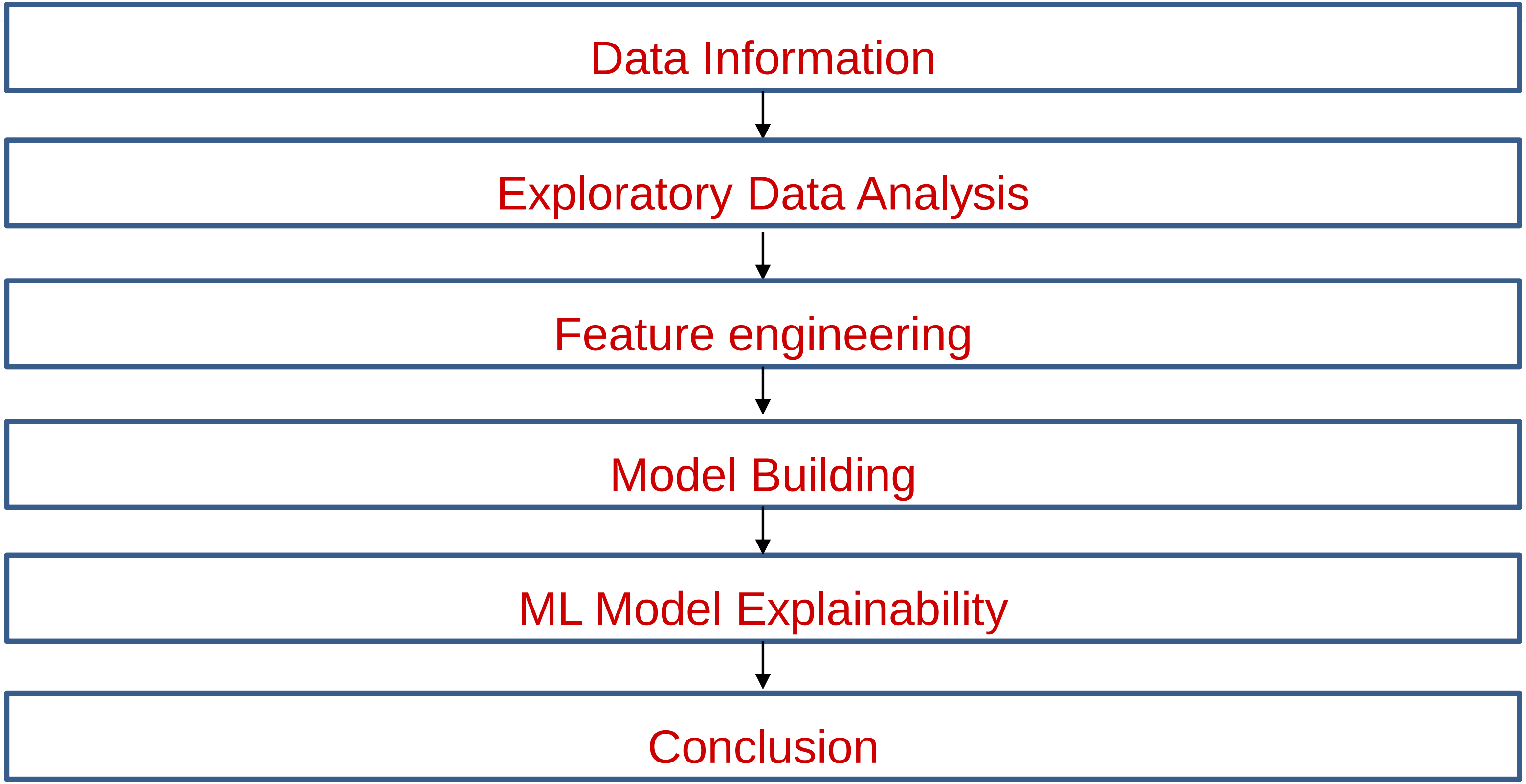
Shahfaissal I Dharwad

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 retrieved in the spring of 2019. The main goal is to predict whether passengers will recommend the airline to their friends.

Methodology

The process from getting the data to drawing the conclusion is as follows:



Data Information

Exploratory Data Analysis

Feature engineering

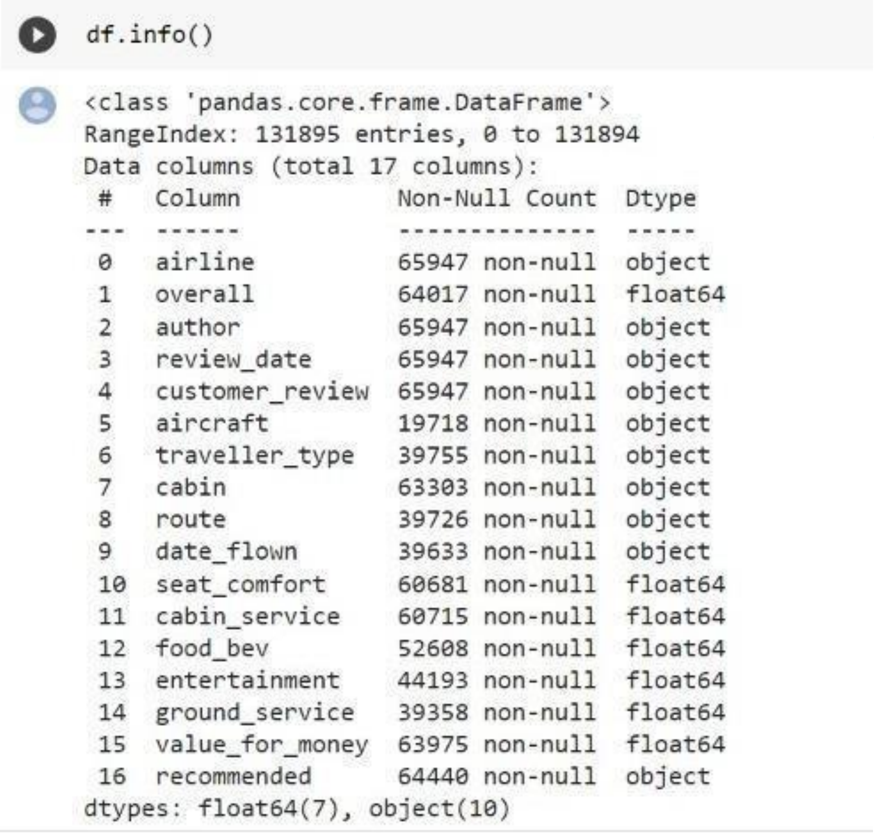
Model Building

ML Model Explainability

Conclusion

Data Insights…

- The data set 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 dataset is a mixture of categorical and numerical data, so we need to organize and code the data before feeding it into the ML model.



```
df.info()
```

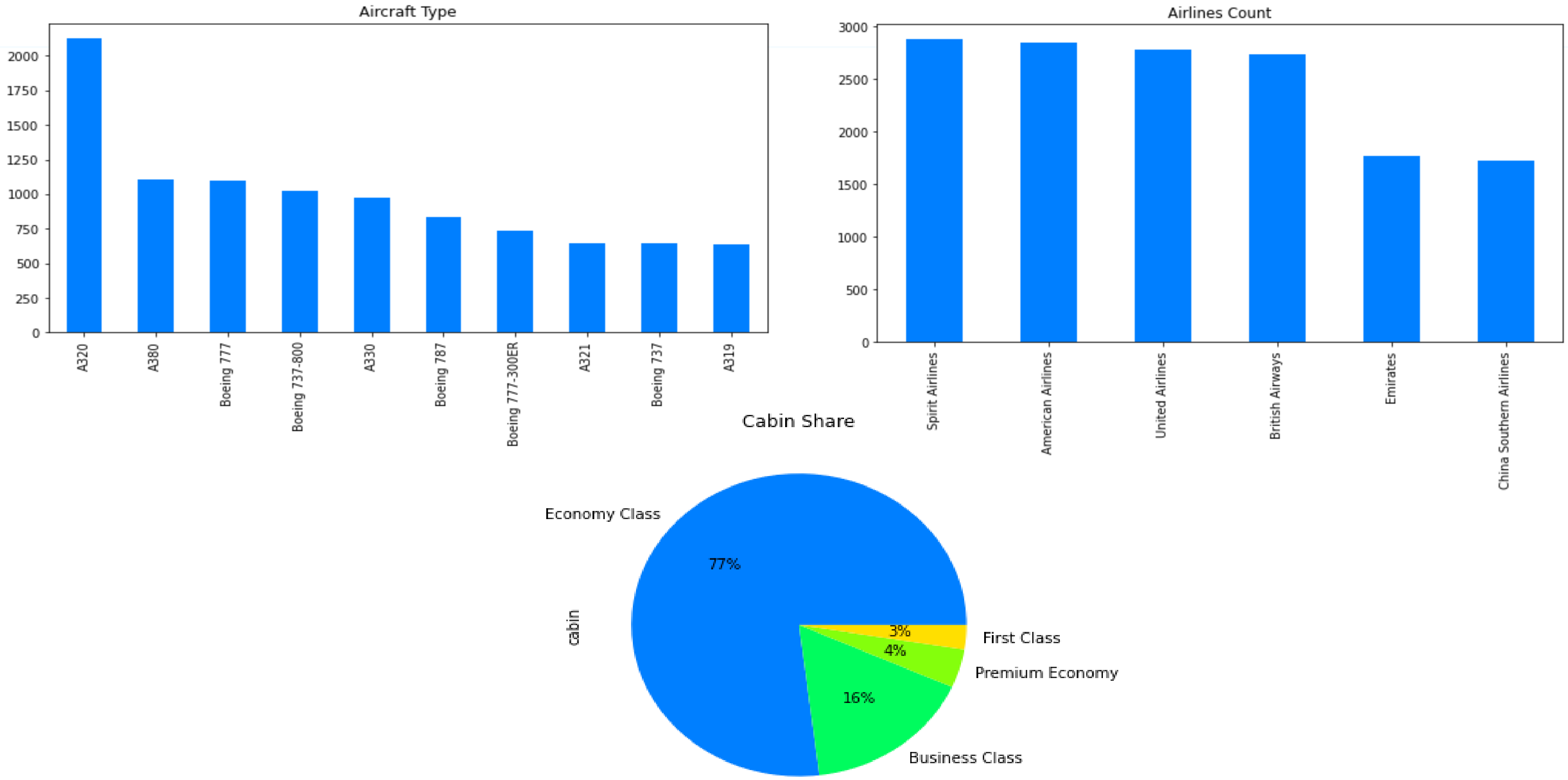
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 131895 entries, 0 to 131894
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   airline          65947 non-null  object
 1   overall          64017 non-null  float64
 2   author           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:-

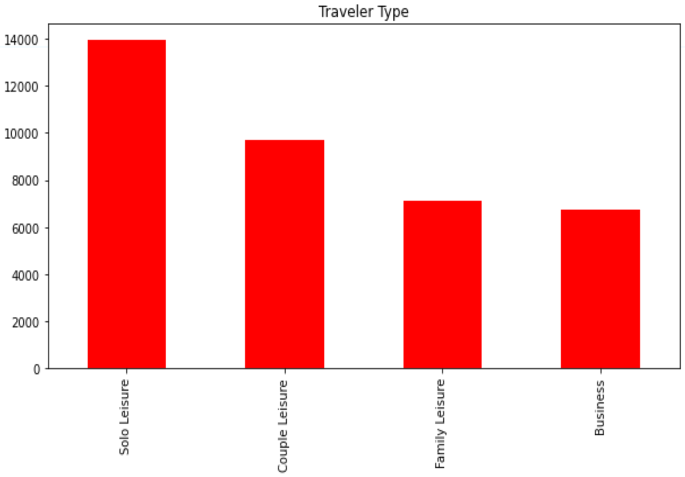
Airline: Name of the airline.
overall: Overall point is given to the trip between 1 to 10.
author: Author of the trip
Review date: Date of the Review customer review: Review of the customers in free text format
Customer Review: Feedback shared by the customers
Aircraft: Type of the aircraft
Traveler Type: Type of traveler (e.g. business, leisure)
Cabin: Cabin
Flight date: Date on which The flight has flown
Route: Route taken by flight
Seat comfort: Rated between 1-5
cabin service: Rated between 1-5
Food-Bev: Rated between 1-5 entertainment: Rated between 1-5
Ground service: Rated between 1-5
Value for money: Rated between 1-5
Recommended: The passenger has referred his friend or not.

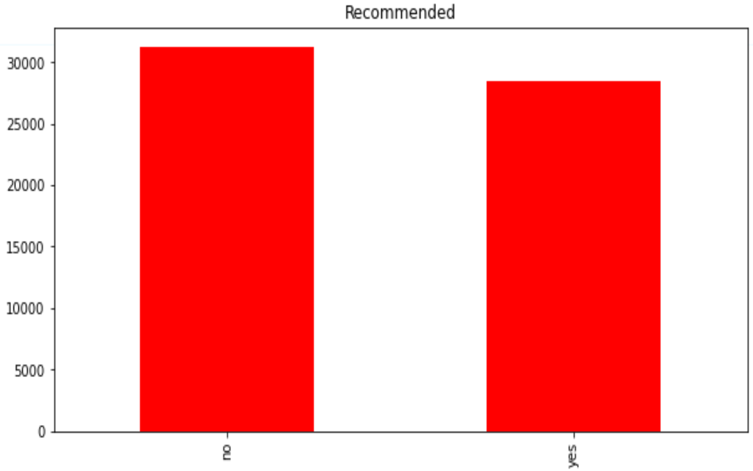
Exploratory Data Analysis

EDA for Passenger class, Airline company, Aircraft carrier



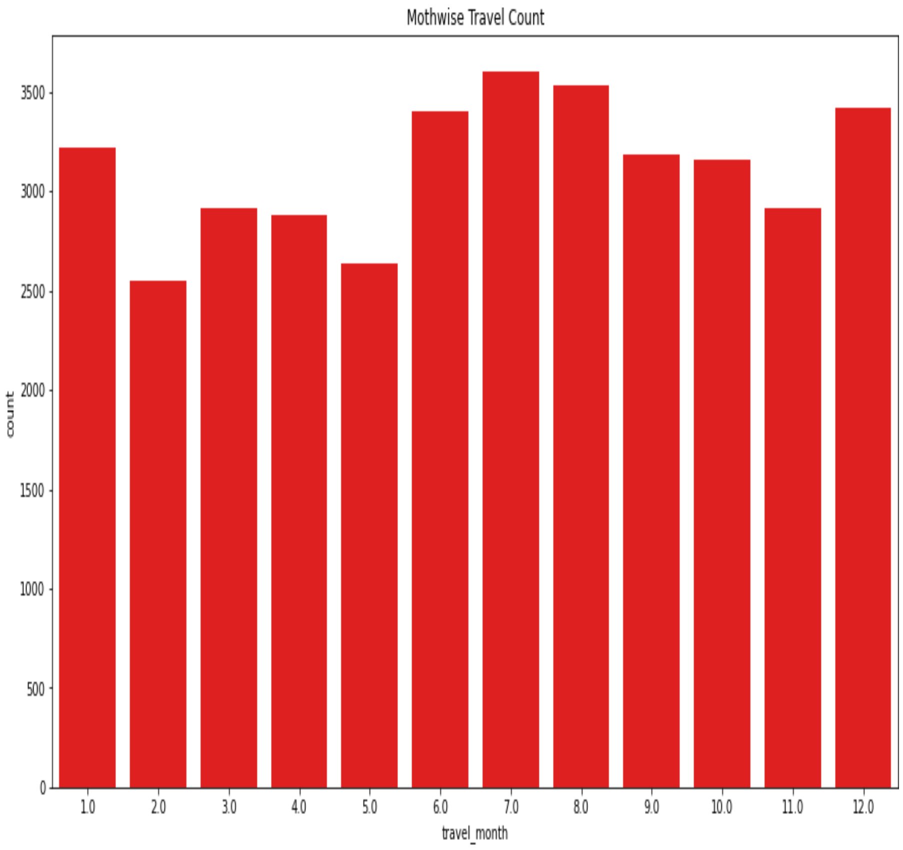
Exploratory Data Analysis





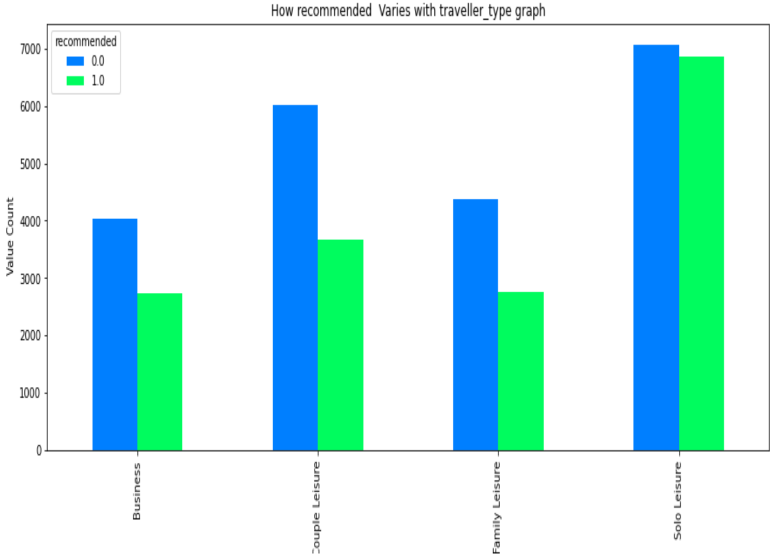
- We can see that there are 4 classes present in the Traveler type function. We can also notice that Solo Leisure has the highest value. From this we can conclude that most people who travel by air travel alone. Next comes college and then family. A very small percentage of people prefer to fly for business.
- In the recommended graph, we see that the dependent element "recommended" has balanced data in its Yes and No classes.

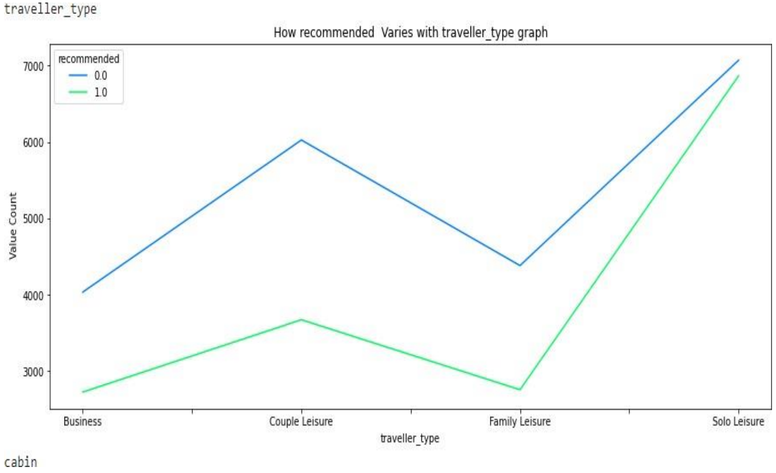
Exploratory Data Analysis



Here we can see that people flown most frequently in the month of July and least frequently in month of February.

Exploratory Data Analysis



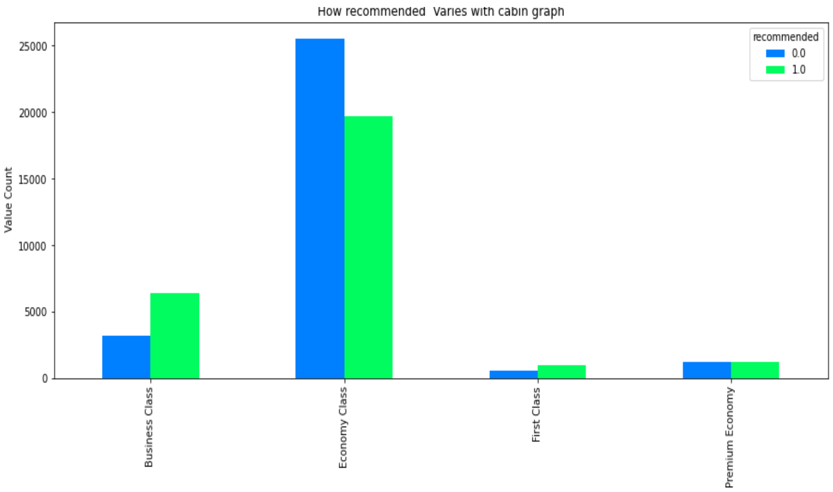

traveller_type

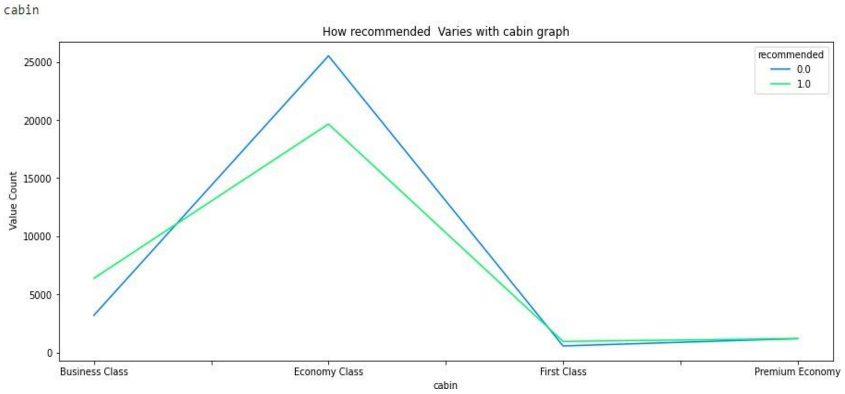
cabin

Variation of Traveller type feature with recommendation:

- We see that people gave both 1 and 0, which we will consider as positive and negative recommendations from now on, so that we can effectively interpret it for solo leisure. This could be due to poor infrastructure or services people are getting, and positive referrals could be due to the low cost of a solo. However, this is an approximate analysis based on the data provided.

Exploratory Data Analysis

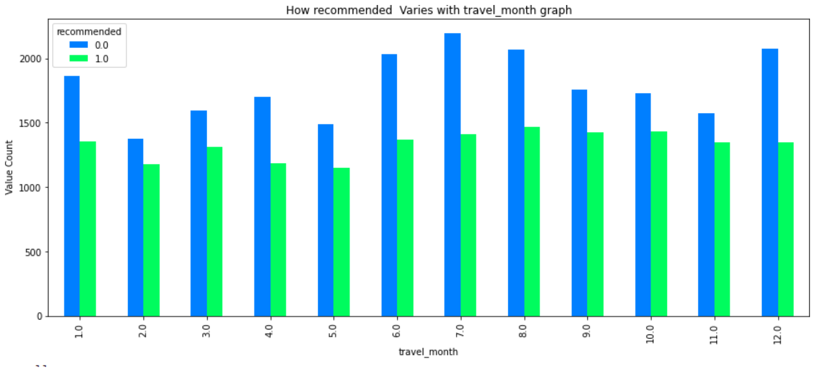




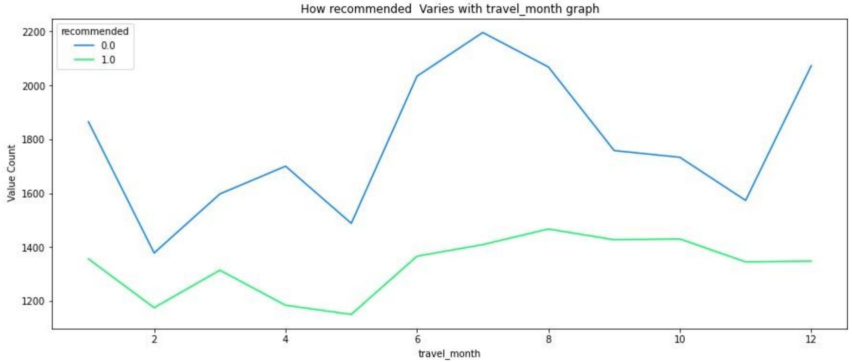
Variations of the Traveler type feature with a cabin:

- we see people giving the economy class cabin highly positive recommendations. From this we can conclude that people like to travel in economy class for a low price and in the same way we can see that people give the highest negative recommendation to economy class, maybe because it provides them with less infrastructure or services
- We can also see that people gave the highest positive recommendation to Business Class, this may be because of the quality of service provided to them in Business Class, and similarly negative recommendations due to the high price of Business Class or less percentage of travel

Exploratory Data Analysis



travel_month

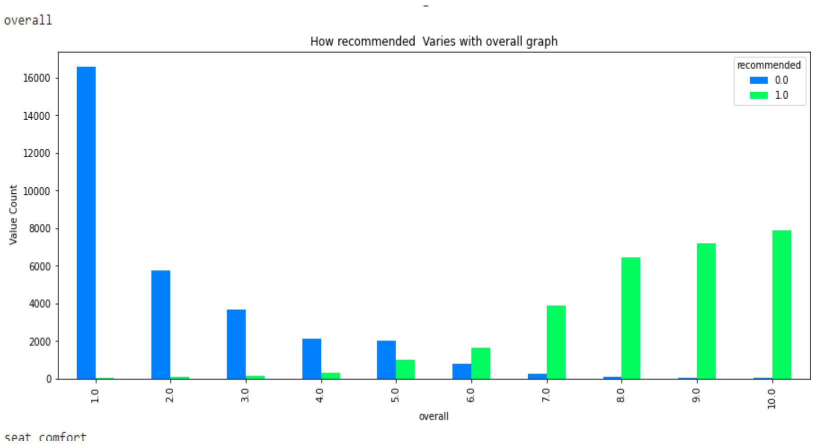


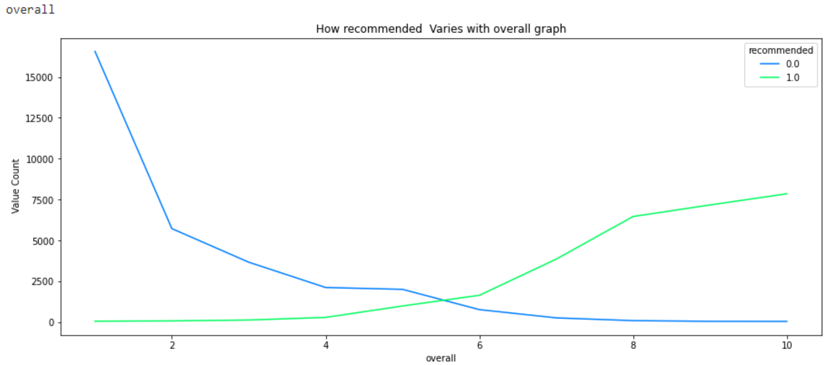
Variation of Traveller type feature with Travel Month:

- 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.

- In month we cannot see any preferable trend but here we can conclude people tent to travel highest during the month of July.

Exploratory Data Analysis



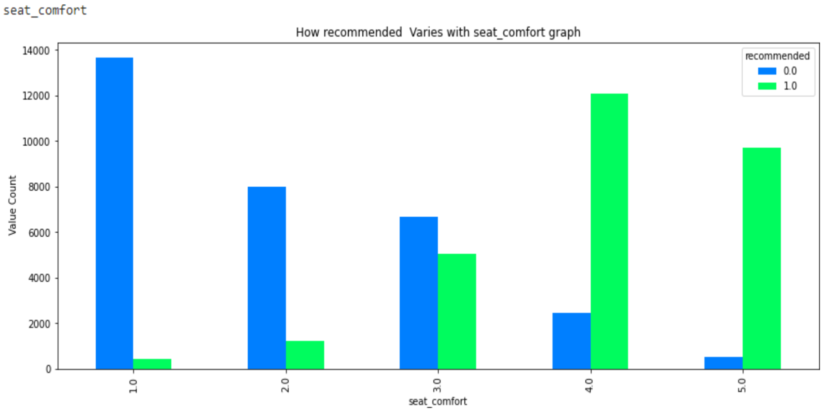


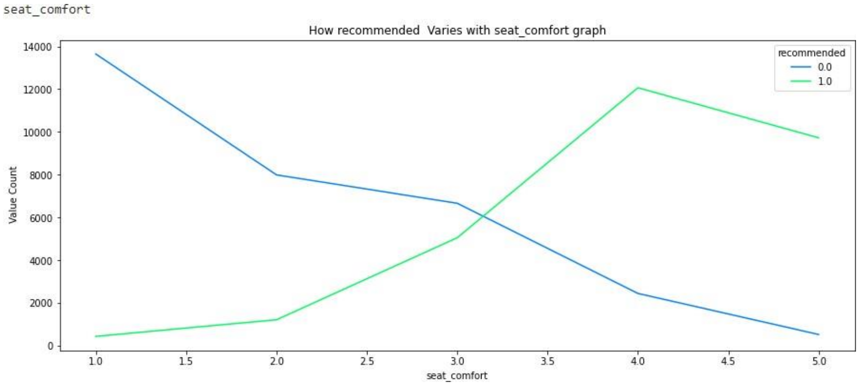
Variation of Traveller type feature with overall rating:

- From the graph of overall rating vs. we see the recommended one which is quite understandably negative the recommendation was given to an overall rating of 1.0 and a high positive recommendation was given to an overall rating of 10. But it is very true that the highest negative recommendation was given to an overall rating of 1.0 which is really disturbing.

- In the overall evaluation, we can experience very good insights, which are also regular. We can see how the positive recommendation increases with the overall rating and also the negative recommendation decreases as well.

Exploratory Data Analysis



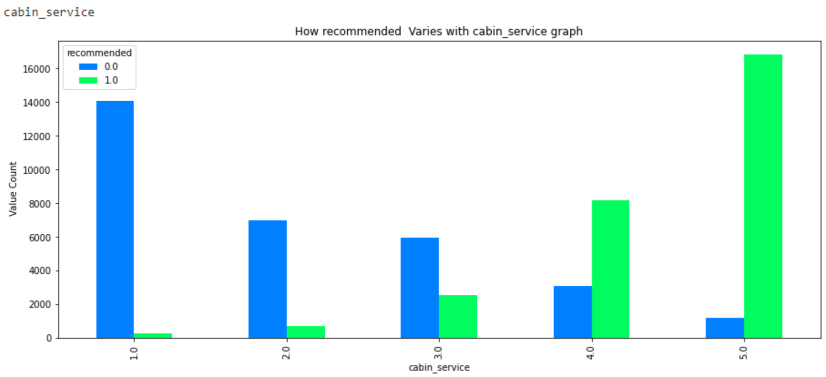


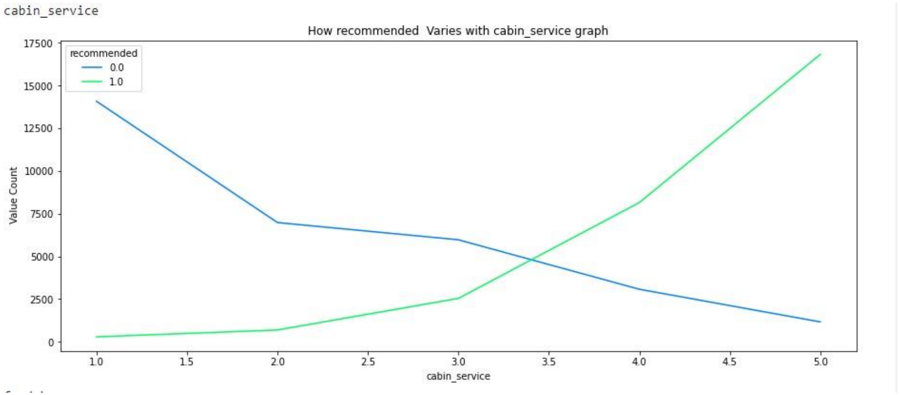
Variation of Traveller type feature with seat comfort :

- In the comfort of the seats, people gave the highest positives recommended for a class 5 seat compared to very low negative recommendation for the same. We can also see that the class 1 seat received the highest negative recommendation compared to its positive recommendation. Here we conclude that it must be removed as soon as possible.

- In the seating comfort we can see how the positive recommendation increases with the overall rating and also the negative recommendation on the same decreasing also we can intersection in the seating comfort rating 3.0 where we can see similar positive and negative recommendations

Exploratory Data Analysis



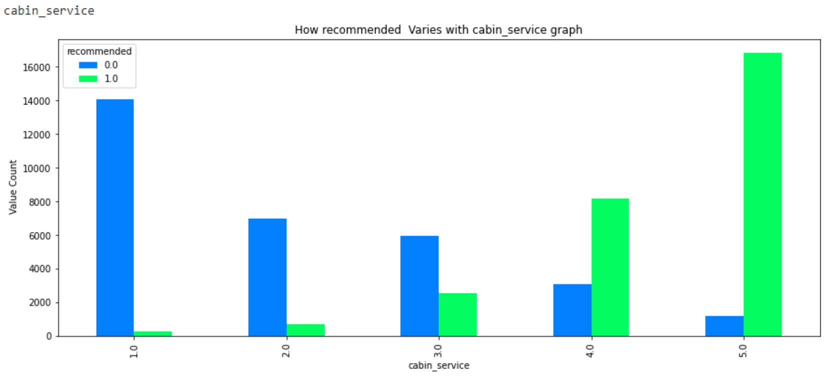


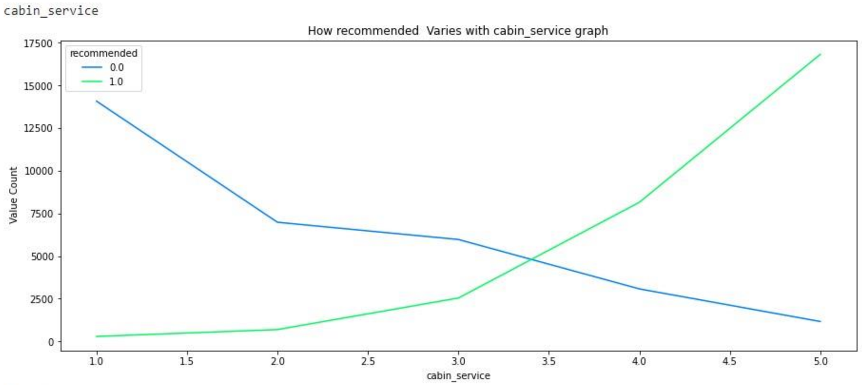
Variation of Traveller type feature with Cabin Service :

- People rated the cabin service the highest rating recommendation for cabin service rating 5 as compare with its counterpart. From this we can conclude that the cabin service is doing quite well.

- In cabin crew we can see the same as the positive recommendation increases with the overall rating and also the negative recommendation on the same declines we can also see the intersection in the cabin crew rating of 3.5 where we can see similar positive and negative recommendations

Exploratory Data Analysis



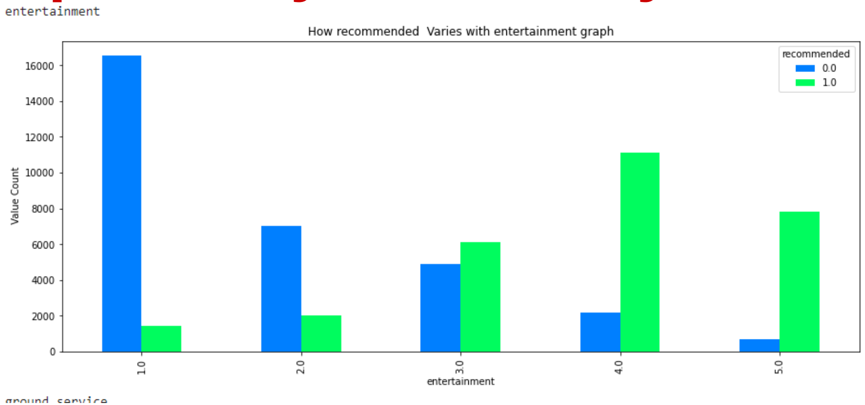


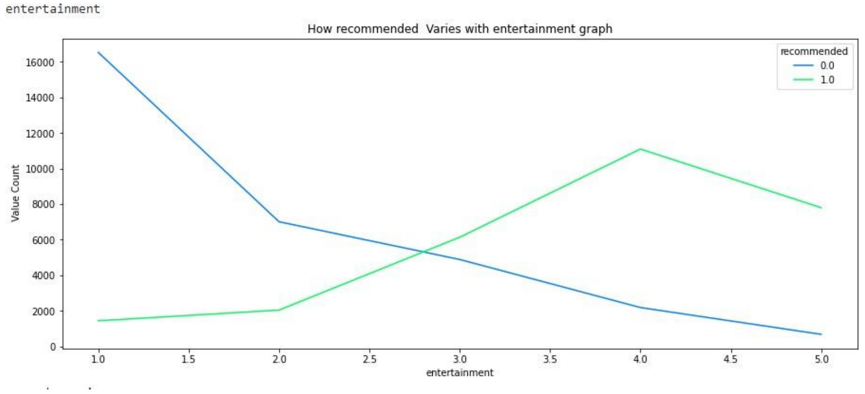
Variation of Traveller type feature with Food Bev :

- People rated food and drinks the highest negative recommendation for a rating of 1.0 from this we can concluded that airline services need to improve food delivery and service quality.

- In food services we can see the same as positive recommendations increase with the overall rating and also negative recommendations on the same declines also we can see an intersection in the food service rating near 3.0 where we can see similar positive and negative recommendations.

Exploratory Data Analysis





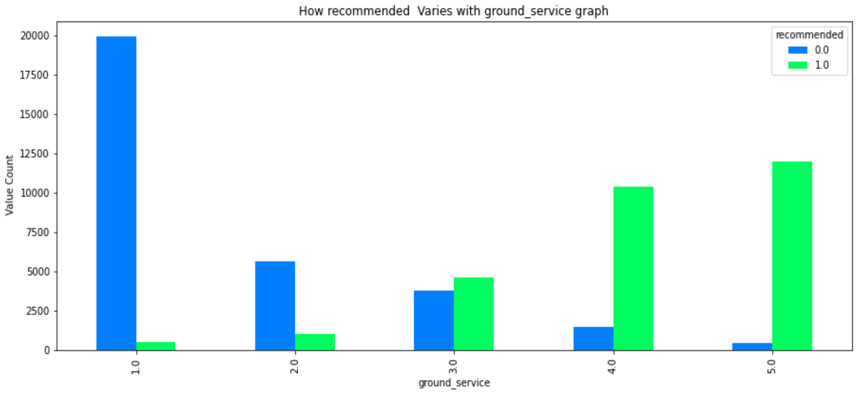
Variation of Traveller type feature with Entertainment:

- In the field of entertainment, we also see that most people have given highest negative recommendation for entertainment rating 1, which shows that the airline also needs to improve its entertainment system.

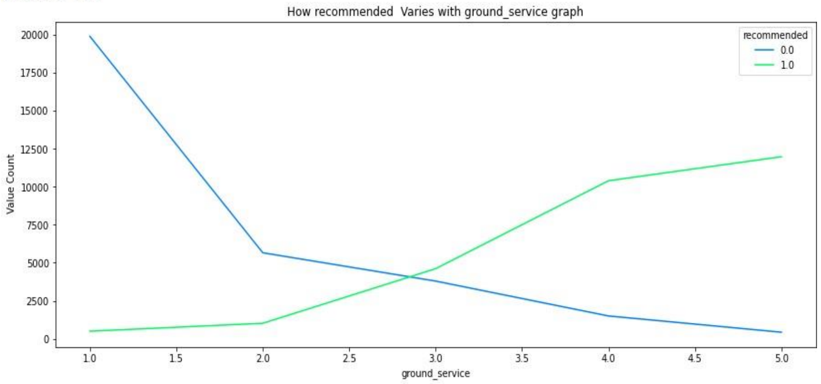
- Even in the area of Entertainment service we can see the same positive recommendation increases with the overall rating and also the negative recommendation on the same declines we can also see the intersection in the Entertainment service rating between 2.5 and 3.0 where we can see similar positive and negative recommendations.

Exploratory Data Analysis

ground_service



ground_service

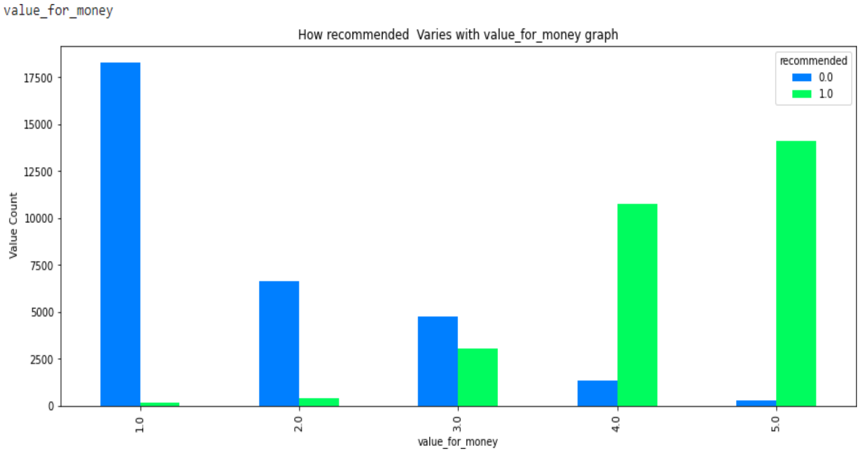


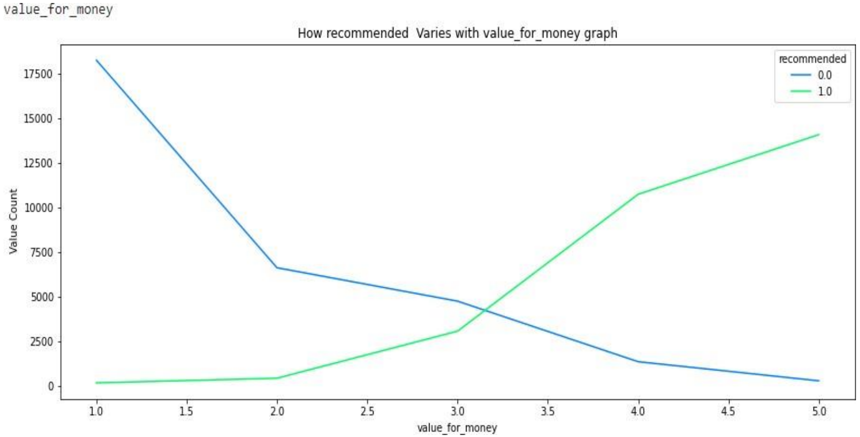
Variation of Traveller type feature with Ground Service:

- In the ground service we also see that most people gave the highest rating of negative recommendation for ground services 1 which shows that the airline needs to improve its ground services.

- In Ground service we can also see the same as the positive recommendation increases with the overall rating and also the negative recommendation on the same decreases also we can see the intersection in the Ground service rating close to 3.0 where we can see similar positive and negative recommendations

Exploratory Data Analysis



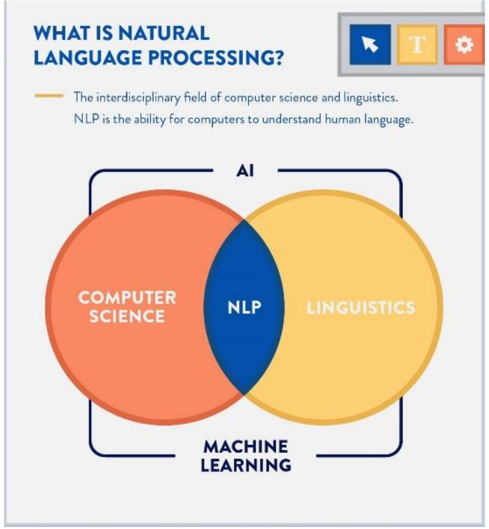


Variation of Traveller type feature with Value for Money:

In the ground service we also see that most people gave the highest rating of negative recommendation for ground services 1 which shows that the airline needs to improve its ground services.

In Ground Service we can also see how the positive recommendation increases with the overall rating and also the negative recommendation on the same declines also we can cross the Ground Service rating near 3.0 where we can see similar positive and negative recommendations.

NLP(Natural Language Processing:



WHAT IS NATURAL LANGUAGE PROCESSING?

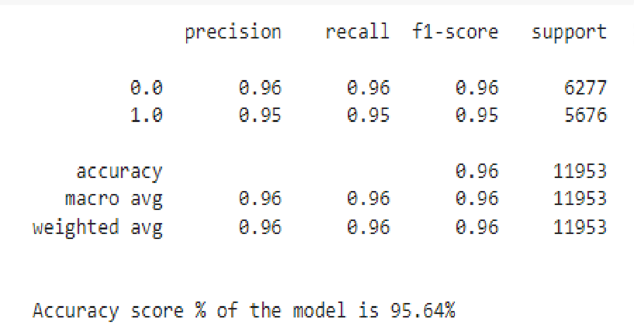
The interdisciplinary field of computer science and linguistics. NLP is the ability for computers to understand human language.

Information Retrieval
Doc A
Doc 1
Doc 2
Doc 3
Sentiment Analysis
Information Extraction
Machine Translation
文
A
Natural Language Processing
QuestionAnswering
?
Human: When was Apollo sent to space?
Machine: First flight - AS-201, February 26, 1966

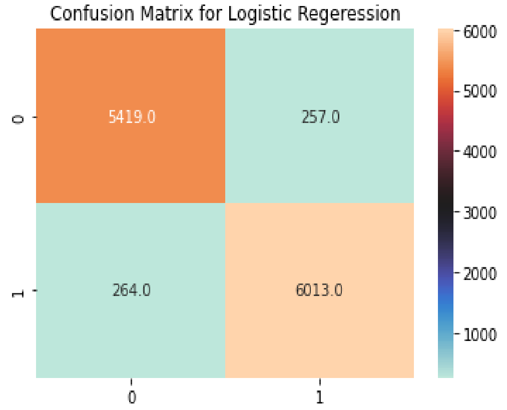
- We have used vander sentiment in NLP so to convert sentiments in customer review into score so to have our model prediction.

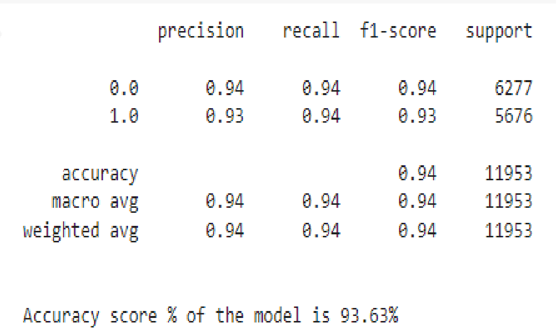
- We have also created new feature numeric review so to store sentiment score we have retrieved using sentiment analysis from customer review feature.

Model Building:

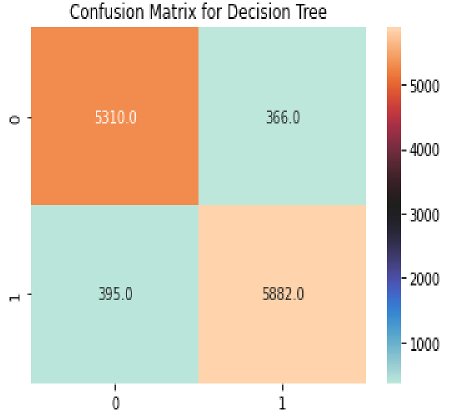


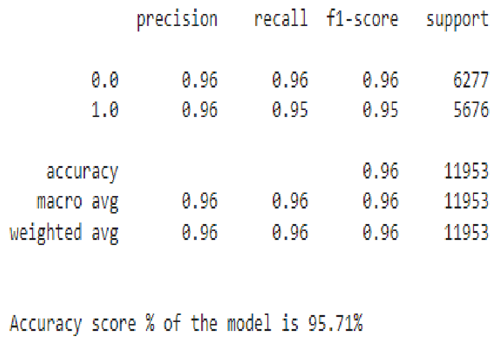
```
              precision    recall  f1-score   support

         0.0       0.96      0.96      0.96      6277
         1.0       0.95      0.95      0.95      5676

    accuracy                           0.96     11953
   macro avg       0.96      0.96      0.96     11953
weighted avg       0.96      0.96      0.96     11953


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



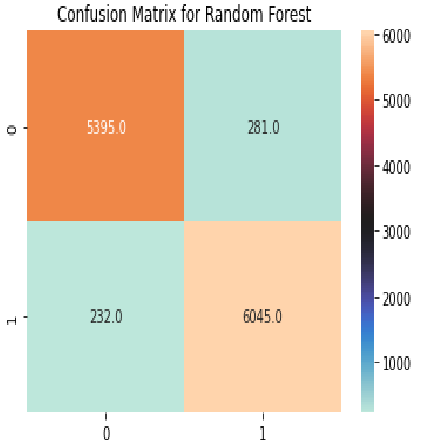


```
              precision    recall  f1-score   support

         0.0       0.94      0.94      0.94      6277
         1.0       0.93      0.94      0.93      5676

    accuracy                           0.94     11953
   macro avg       0.94      0.94      0.94     11953
weighted avg       0.94      0.94      0.94     11953


Accuracy score % of the model is 93.63%
```



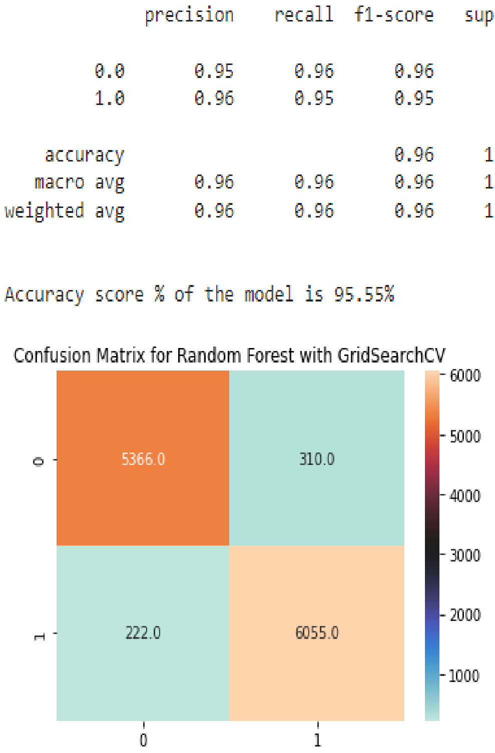


```
              precision    recall  f1-score   support

         0.0       0.96      0.96      0.96      6277
         1.0       0.96      0.95      0.95      5676

    accuracy                           0.96     11953
   macro avg       0.96      0.96      0.96     11953
weighted avg       0.96      0.96      0.96     11953


Accuracy score % of the model is 95.71%
```



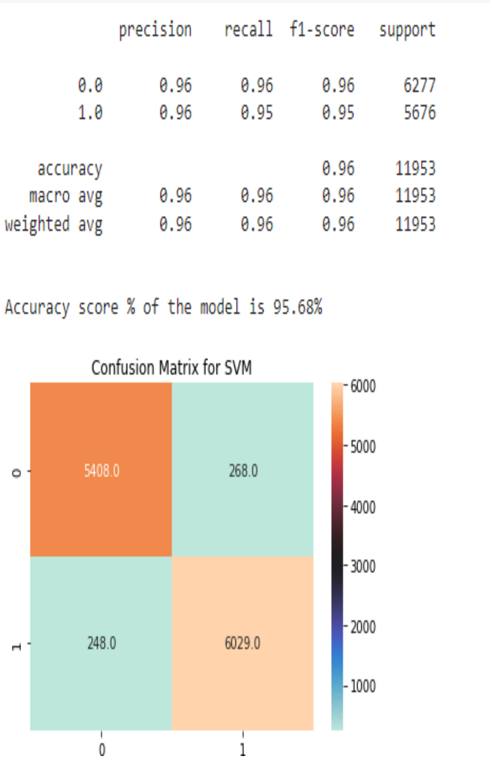
Evaluation metrics for Logistic, Decision tree, Random forest

Model Building(Continued….)



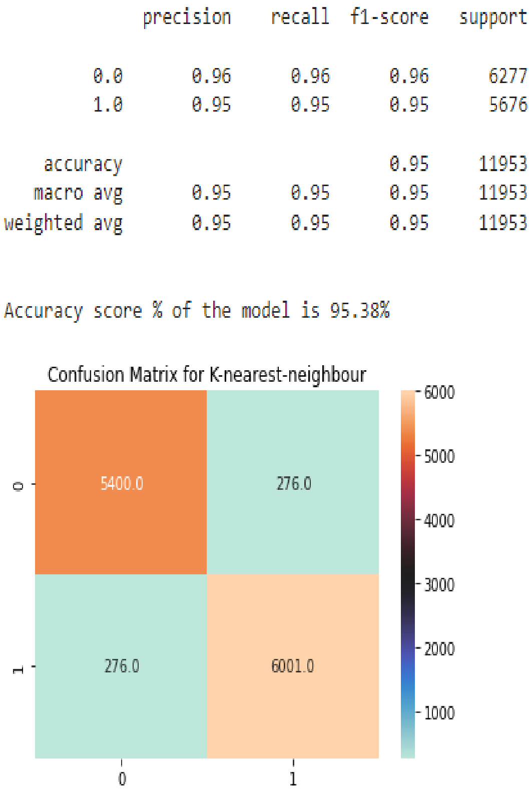
```
              precision    recall  f1-score   support

         0.0       0.95      0.96      0.96      6277
         1.0       0.96      0.95      0.95      5676

    accuracy                           0.96     11953
   macro avg       0.96      0.96      0.96     11953
weighted avg       0.96      0.96      0.96     11953


Accuracy score % of the model is 95.55%
```

Confusion Matrix for Random Forest with GridSearchCV



```
              precision    recall  f1-score   support

         0.0       0.96      0.96      0.96      6277
         1.0       0.96      0.95      0.95      5676

    accuracy                           0.96     11953
   macro avg       0.96      0.96      0.96     11953
weighted avg       0.96      0.96      0.96     11953


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

Confusion Matrix for SVM

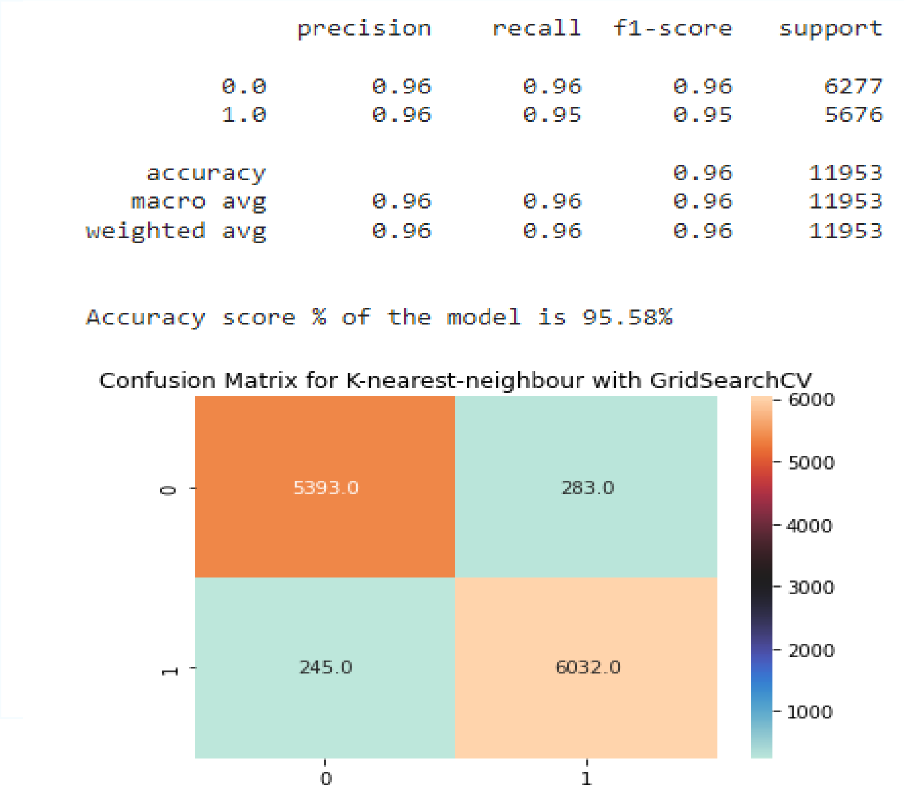


```
              precision    recall  f1-score   support

         0.0       0.96      0.96      0.96      6277
         1.0       0.95      0.95      0.95      5676

    accuracy                           0.95     11953
   macro avg       0.95      0.95      0.95     11953
weighted avg       0.95      0.95      0.95     11953


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

Confusion Matrix for K-nearest-neighbour

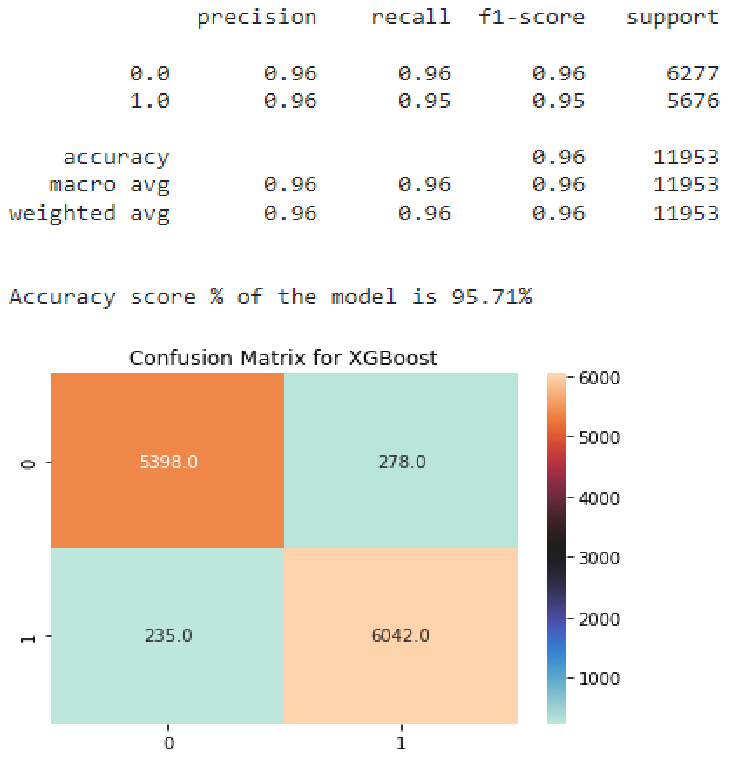
Evaluation metrics for GridsearchCV, SVM, K-nearest neighbour

Model Building(Continued….)



```
              precision    recall  f1-score   support

         0.0       0.96      0.96      0.96      6277
         1.0       0.96      0.95      0.95      5676

    accuracy                           0.96     11953
   macro avg       0.96      0.96      0.96     11953
weighted avg       0.96      0.96      0.96     11953


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

Confusion Matrix for K-nearest-neighbour with GridSearchCV



```
              precision    recall  f1-score   support

         0.0       0.96      0.96      0.96      6277
         1.0       0.96      0.95      0.95      5676

    accuracy                           0.96     11953
   macro avg       0.96      0.96      0.96     11953
weighted avg       0.96      0.96      0.96     11953

Accuracy score % of the model is 95.71%
```

Confusion Matrix for XGBoost

Evaluation metrics for K-nearest neighbour with GridsearchCV, XGBoost

Model Building(Continued….)

	Model	Accuracy	Recall	Precision	F1-score	Roc_auc_ score
0	Logistic Regression	0.956413	0.954722	0.953546	0.954133	0.956332
1	Decision Tree	0.936167	0.935166	0.930738	0.932947	0.936119
2	Random Forest	0.957835	0.950846	0.959979	0.955390	0.957500
3	Random Forest with GridSearchCV	0.956078	0.946441	0.960486	0.953412	0.955617
4	SVM	0.956831	0.952784	0.956153	0.954465	0.956637
5	K-nearest-neighbour	0.953819	0.951374	0.951374	0.951374	0.953702
6	K-nearest-neighbour with GridSearchCV	0.955827	0.950141	0.956545	0.953332	0.955555
7	XGBoost	0.957082	0.951022	0.958282	0.954638	0.956792

Conclusion:

- **We see that people gave both 1 and 0, which we will consider as positive and negative recommendations from now on, so that we can effectively interpret it for solo leisure. This could be due to poor infrastructure or services people are getting, and positive referrals could be due to the low cost of a solo. However, this is an approximate analysis based on the data provided.**
- **We can also see that people highly recommend economy class in the cabin. From this we can conclude that people love to travel in economy class for a low price and in the same way we can see that people give the highest negative recommendation to economy class, maybe because it provides them with less infrastructure or services. We can also see that people gave the highest positive recommendation to Business class, this may be because of the quality of service provided to them in Business class, and similarly negative recommendations due to the high price of business class or less percentage of travel.**
- **From the month vs. recommendation. We see that people tend to travel the most in the month of July**
- **the total number of positive and negative recommendations combined.**
- **From the graph of total vs. recommended we can see which is totally understandable that the negative recommendation was given to an overall rating of 1.0 and a highly positive recommendation was given to an overall rating of 10. But it is very true that the highest negative recommendation was given to an overall rating of 1.0 which is really disturbing.**
- **When it comes to seat comfort, people gave the highest positive recommendation to the Class 5 seat, compared to a very low negative recommendation for the same. We can also see that the class 1 seat received the highest negative recommendation compared to its positive recommendation. Here we conclude that it must be removed as soon as possible.**

Conclusion:

- In the cabin service rating, people gave the highest rating recommendation to the cabin service rating of 5 compared to that
- counterpart. From this we can conclude that the cabin service is doing quite well.
- In the food and beverage rating, people gave the highest negative recommendation to a rating of 1.0, from which we can conclude that the airline needs to improve its food delivery and service quality.
- In the area of entertainment, we can also see that most people gave the highest negative recommendation to the entertainment rating of 1, which shows that the airline also needs to improve its entertainment system.
- In the ground service we can also see that most people gave the highest negative recommendation to the ground service rating of 1, which shows that the airline needs to improve its ground service.
- In terms of value for money, we can also see that most people gave the highest negative recommendation to the value for money rating of 1, which shows that the airline needs to make its flight services more cost-effective.
- In the model selection, we can see that the Random Forest model and the XGBoost Model have the same high model accuracy with a score of 0.957082, but we can also see that the recall, precision, f1-score and roc_auc_score of the XGBoost model together give a higher score than the Random Forest from which we chose the XGBoost Model for further prediction.

Thank you