Rishabh Shrestha

Csci 543 Project 1

Airlines Customer Satisfaction Survey

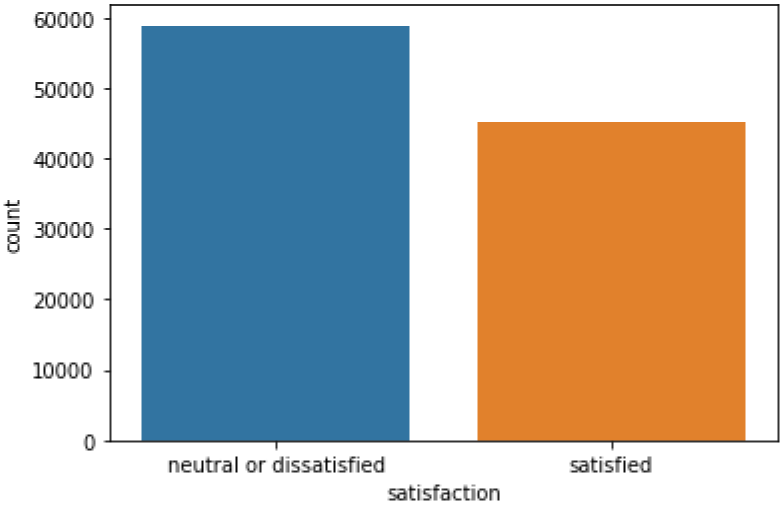
Project Overview:

In this project, I explored the dataset containing an airline passenger satisfaction survey. I have explored the features of the dataset and determined what factors are highly correlated with passenger satisfaction. I created some statistical visualizations that help to understand the dataset. And finally, I built different classifier models to predict the passenger satisfaction and carefully analyzed and tuned the hyper parameters of the model to get the best one.
The dataset contains the features like Id, Gender, Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, On-board service, Leg room service, Baggage handling, Checkin service , Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes.

Many of the features have a discrete (Likert) score of 1-5. (or 0-5) where 0 means missing, 1 is bad and 5 is excellent. The size of the training data set has 104000 instances whereas the test data set has 26000 instances. And our target variable(class) is ‘satisfaction’ which has two values- neutral or dissatisfied and satisfied.

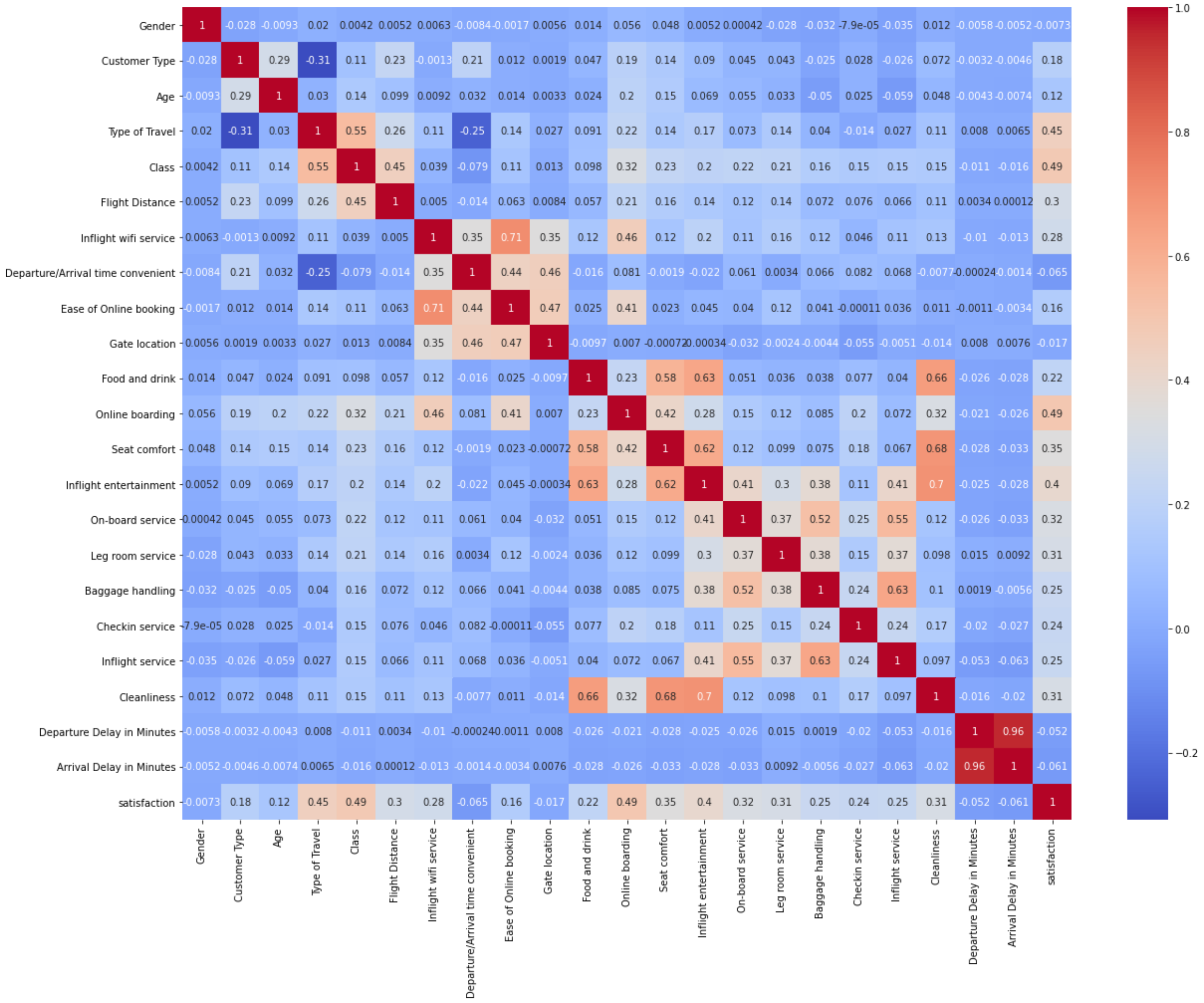
Understanding Data:

At first, I checked for missing values in the dataset and found there were 310 missing values in the 'Arrival Delay in Minutes'. So, I replaced the missing values with the median value. When the data is skewed, it is better to consider using the median value to replace the missing values. I checked for the skewness of the class labels of satisfaction in the training set. The training set has 58879 instances as neutral or satisfied and 45025 instances as satisfied which is pretty much even distribution.

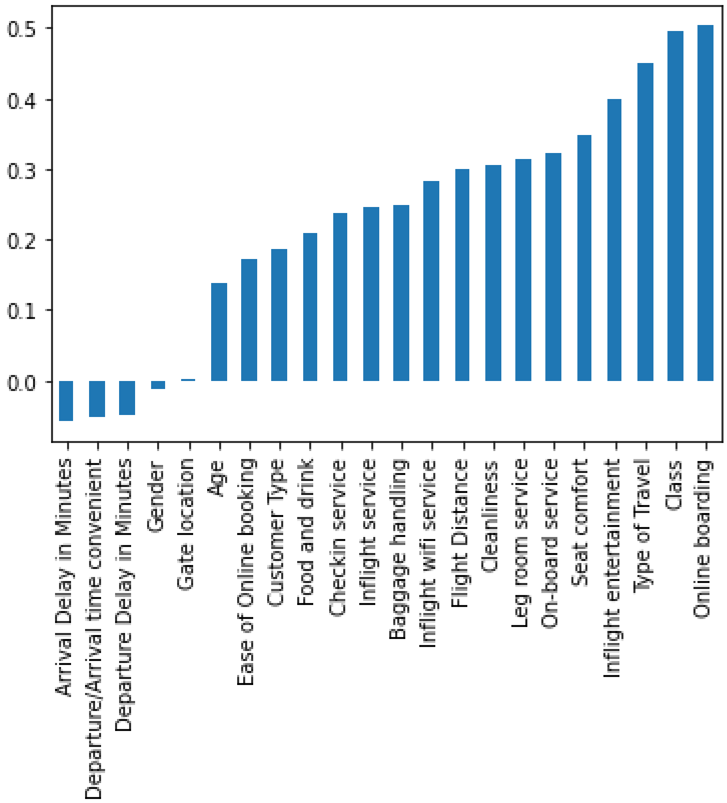


The data was preprocessed for deeper understanding of the data and to help in the creation of models. The dataset had some categorical features such as Gender as Male and Female. Similarly, Customer Type, Travel Type, Class and Satisfaction also had categorical values. So those values were encoded as 1, 0, in most cases i.e. Gender(Male=0, Female=1), Customer Type(Loyal=1, Disloyal=0), Travel Type(Personal=0, Business=1, Class(Business=2, Eco Plus=1, Eco=1) and Satisfaction(Satisfied=1, Neutral or dissatisfied=0). I dropped the feature id before proceeding further. All of these preprocessing was done for both train and test sets.

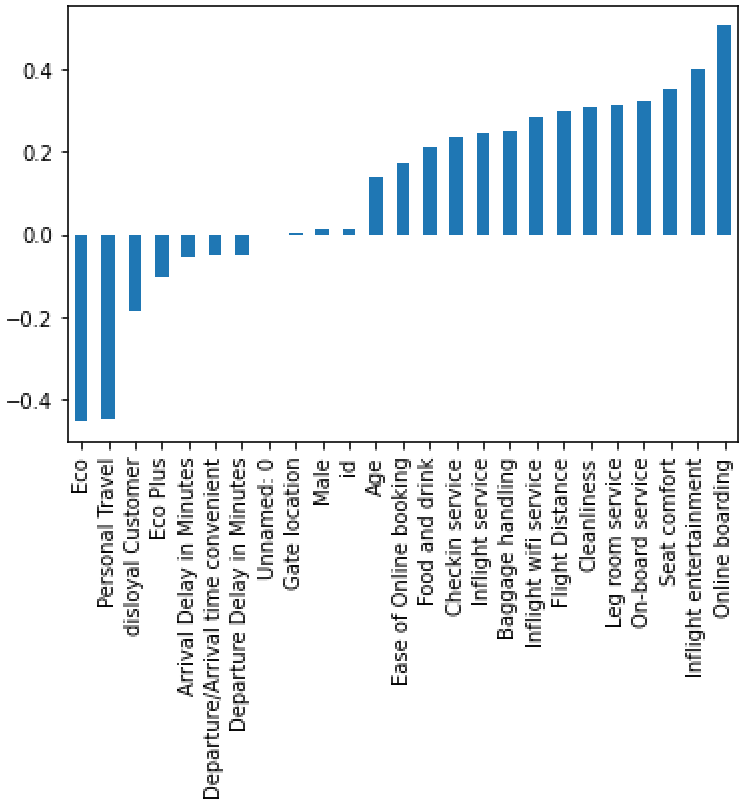
To understand which features correlate well with our target variable i.e. customer satisfaction, and which don't, a correlation heatmap was generated. The features that correlate best with the satisfaction was Online boarding, Class, Type of travel, Inflight Entertainment. The features with the high negative correlation with satisfaction were Departure/Arrival Time Convenient, Gate Location, Gender, Departure Delay in Minutes and Arrival Delay in Minutes.



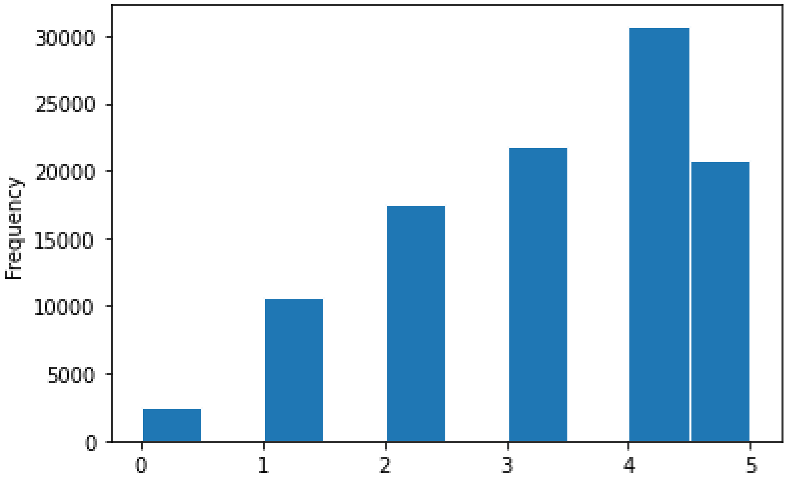
This can also be visualized from the bar chart below to see the correlation between satisfaction and other features.



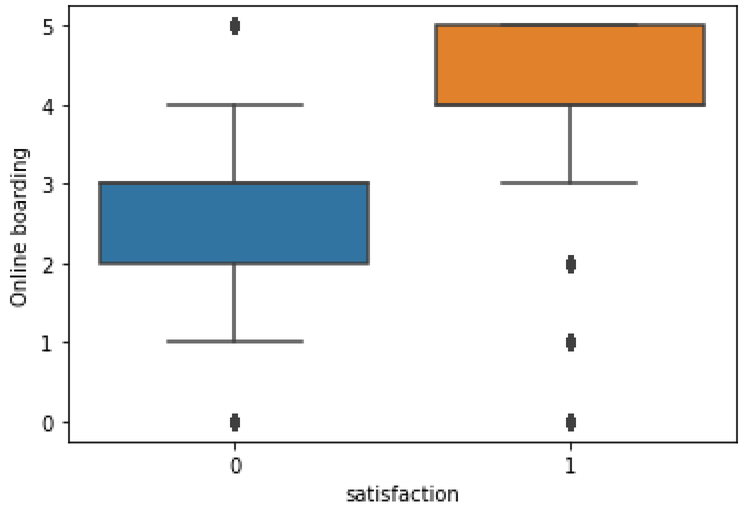
I also wanted to see what values of those variables highly negatively correlated with the satisfaction and came up with this bar graph to visualize that. Analysing the categorical variables, we came to know that People travelling in economy class and on personal travel, disloyal customers are unlikely to get satisfied.



Now we know that people choosing online boarding are the most satisfied.Let’s see how many people fall under each category of online boarding.

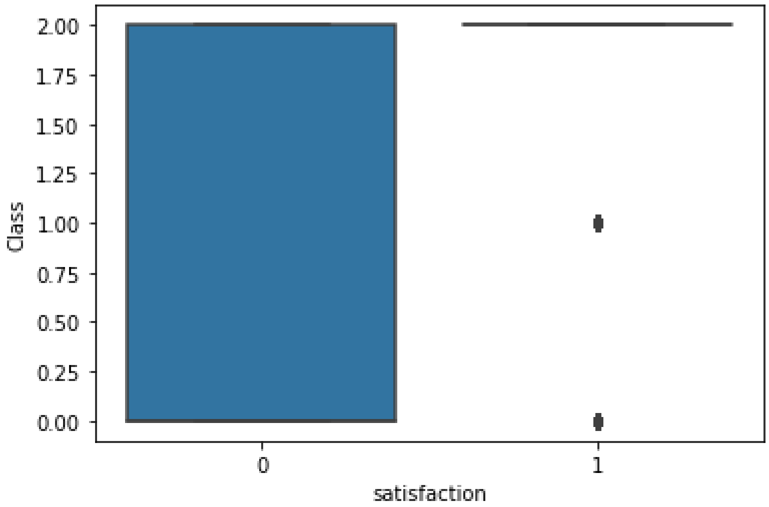


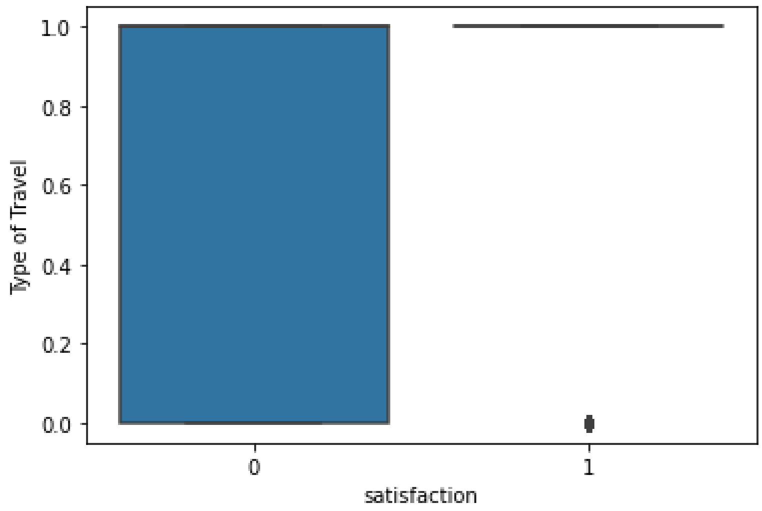
Let’s check the relation of score levels of online boarding with satisfaction.



From the above graph we can see that.the more satisfied the person is with online boarding then the higher the person tends to be satisfied.

Since the Class, Type of travel are also highly correlated with satisfaction, let’s visualize how the values are distributed.





From the above boxplots, we can see that people with Business class(1) are more likely to be satisfied and similarly, people doing Business travel(1) are more likely to be satisfied.

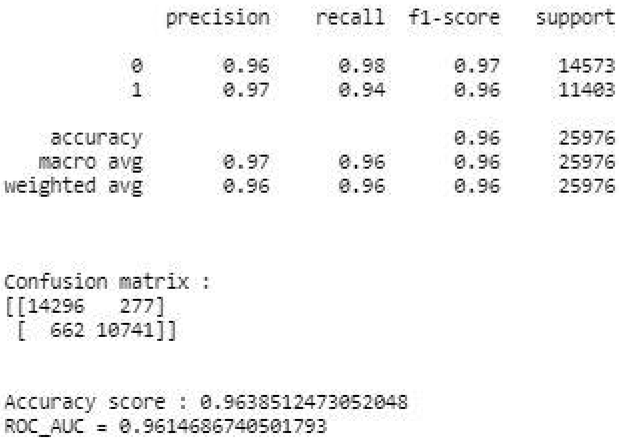
Classification Models and Performance Evaluation:

Now, to develop the Machine Learning classification models to predict the customer satisfaction, I split the dataset into X and y where y is our target variable i.e. satisfaction for both training and testing set.

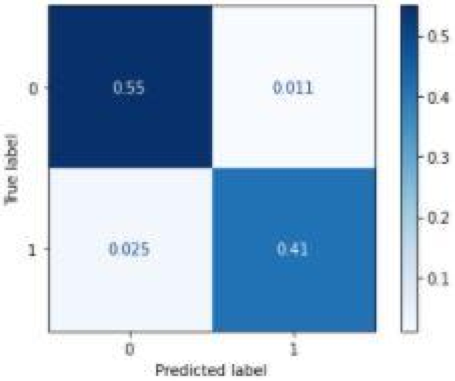
I created a few classification models and I have presented some of them here. While checking the accuracy of the models, I was able to get higher accuracy when I normalized the dataset. One of the classification models is Random Forest Classifier. I used python sci-kit learn library to build the classifier models. Using the Random Forest Classifier, I got an accuracy score of 0.9639 which is impressively high. I created the classification report with scores for precision, recall, f1-score and support to compare the models. I evaluated the performance of models with the ROC_AUC metric. This metric is good for classification of a dataset. I also created confusion matrices for our model to understand how our model is mischaracterizing predictions. The ROC_AUC of Random Forest classifier is 0.9615. I used this metric to compare the performance with other models I created and to determine which one is the best in this case of the dataset.

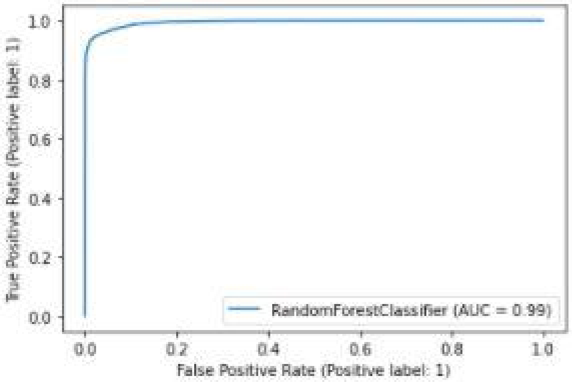
I also determined the 5-fold cross validation of the model. This helped me in determining the performance of the models on unseen data.
Cross Validation Scores: [0.96111833 0.96280256 0.96068524 0.96318753 0.96361886]
The scores look pretty uniform on all the folds of Cross Validation.

In the figure below, we can see the different metrics calculated to determine the performance of the model.



```
              precision    recall  f1-score   support

           0       0.96      0.98      0.97     14573
           1       0.97      0.94      0.96     11403

    accuracy                           0.96     25976
   macro avg       0.97      0.96      0.96     25976
weighted avg       0.96      0.96      0.96     25976


Confusion matrix :
[[14296   277]
 [  662 10741]]


Accuracy score : 0.9638512473052048
ROC_AUC = 0.9614686740501793
```

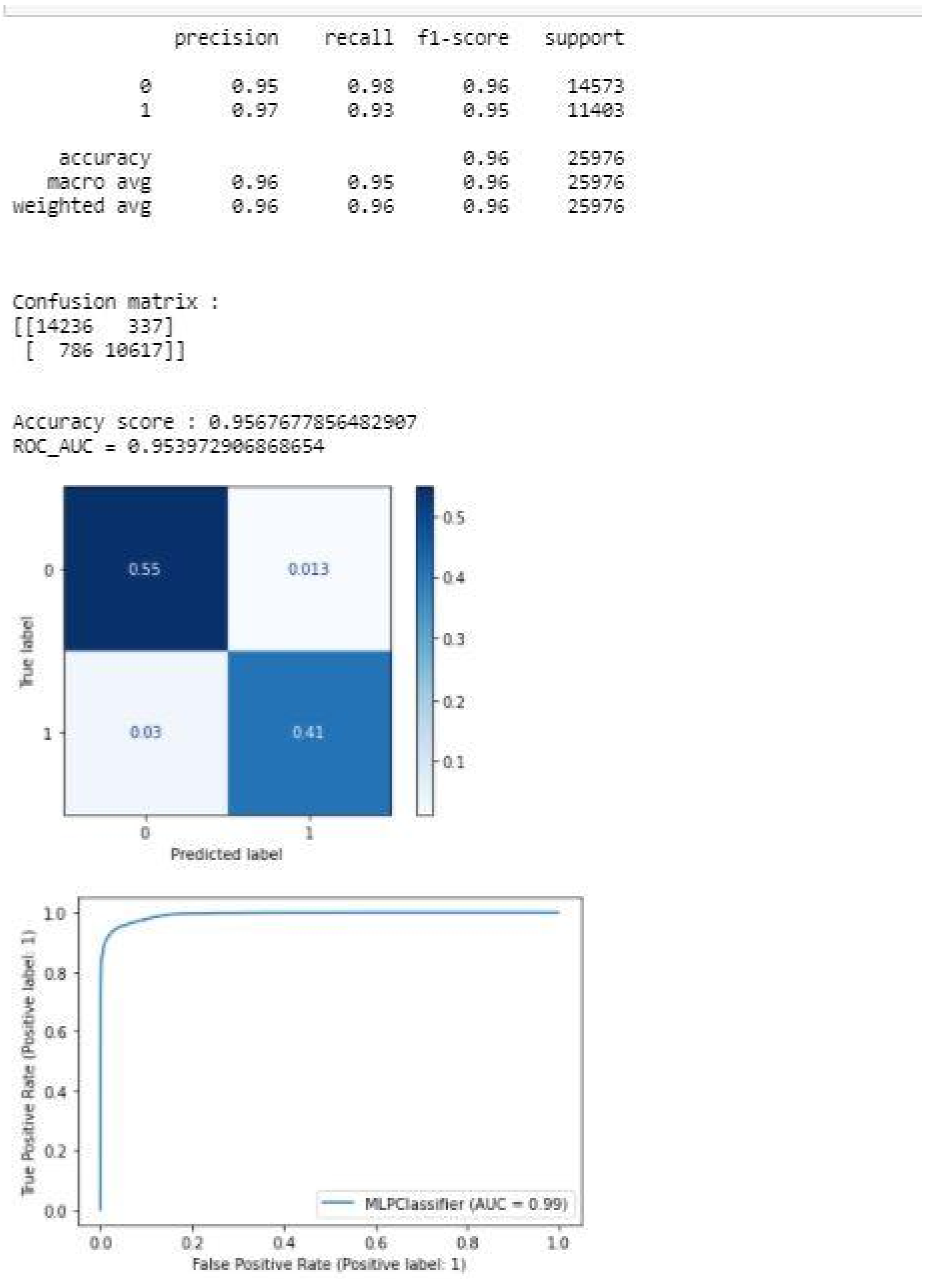




Similarly, another classification model I created was Multiple Layer Perceptron Classifier, commonly known as Neural Network. I got an accuracy score of 0.9567 and ROC_AUC as 0.9540. I determined all other metrics that calculated for Random Forest Classifier to compare with other models.

The Cross Validation Scores: [0.95558443 0.95693181 0.95659497 0.95909725 0.95866218]

We can see all the other metrics in the figure below.



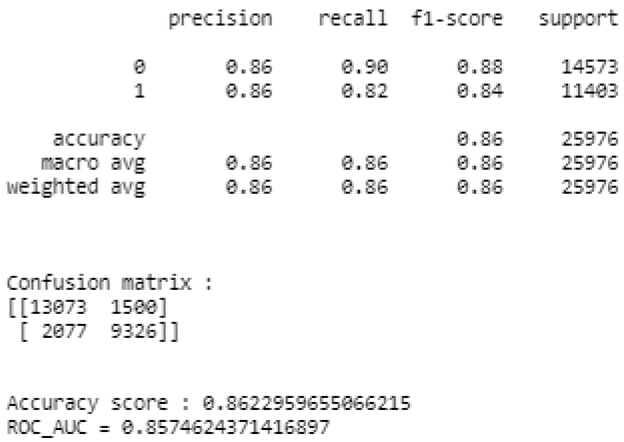
```
              precision    recall  f1-score   support

           0       0.95      0.98      0.96     14573
           1       0.97      0.93      0.95     11403

    accuracy                           0.96     25976
   macro avg       0.96      0.95      0.96     25976
weighted avg       0.96      0.96      0.96     25976


Confusion matrix :
[[14236   337]
 [  786 10617]]


Accuracy score : 0.9567677856482907
ROC_AUC = 0.953972906868654
```

0.55
0.013
0.03
0.41
0
1
True label
0
1
Predicted label
0.5
0.4
0.3
0.2
0.1
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
MLPClassifier (AUC = 0.99)

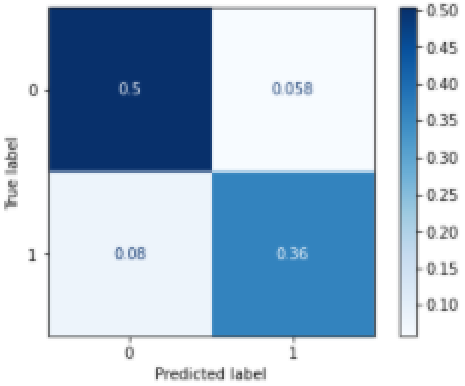
Similarly, I created Naive Bayes Classifier and received an accuracy score of 0.8623 and ROC_AUC of 0.8575.

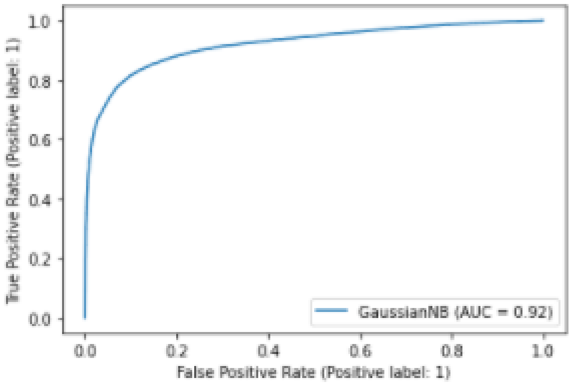
Cross Validation Scores: [0.86242241 0.86348106 0.86237428 0.86530966 0.87271415]

We can see all the other metrics below.



```
              precision    recall  f1-score   support

           0       0.86      0.90      0.88     14573
           1       0.86      0.82      0.84     11403

    accuracy                           0.86     25976
   macro avg       0.86      0.86      0.86     25976
weighted avg       0.86      0.86      0.86     25976


Confusion matrix :
[[13073  1500]
 [ 2077  9326]]


Accuracy score : 0.8622959655066215
ROC_AUC = 0.8574624371416897
```





Therefore, evaluating the performance of these three models,we can see that Random Forest Classifier outperforms all of them.

Accuracy of different models:

Random Forest 0.9638512473052048

MLPClassifier 0.9567677856482907

Naive Bayes 0.8622959655066215

ROC Area Under Curve of different models:

Random Forest 0.9614686740501793

MLPClassifier 0.953972906868654

Naive Bayes 0.857462437141689

Hyper parameter Tuning using GridSearchCV:

I did some hyper parameter tuning using GridSearchCV on our best model ie.Random Forest Classifier for two score types- precision and recall. The parameter set I used was

```
param_grid = {
    'n_estimators': [100, 200, 500],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [4,5,6,7,8],
    'criterion' :['gini', 'entropy']
}
```

With an effort to tune the hyper parameters, I somehow ended up with the scores for Random Forest Classifier models lower than I had before for some reason. So, I chose the best parameter combination with highest accuracy scores as the one I selected before ie.

n_estimators=1000,criterion='entropy', max_depth=25, min_samples_leaf= 1, min_samples_split=2, random_state=42,n_jobs=-1

Note: I tried to perform hyper parameter tuning for the MLP classifier too but the program took too long time to run and ran out of space. So, I used Google Collab's notebook service to perform this which uses TPU for computation of hyper parameter tuning however I was still not able to do with the neural network since it would take too long to run.

References:

1. **https://www.kaggle.com/nikhilsharma4/data-visualization-and-ml-for-psg-satisfaction**
2. **https://www.kaggle.com/chandrimad31/flight-passenger-satisfaction-eda-and-prediction**
3. **https://www.kaggle.com/teejmahal20/classification-predicting-customer-satisfaction**