Capstone Project : Marketing-Airplane Passenger Satisfaction Prediction Using MachineLearningTechniques

Bekee Sorbarisere Yirakpoa[1] and Mercy Nwanyanwu[2]

[1,2]*Department of Computer Science*
Captain Elechi Amadi Polytechnic, Port Harcourt

Abstract.

Customer satisfaction questionnaires are a rich and strong source of information for companies to seek loyalty, customer and client retention, opti-mize resources, and repurchase products. Several advanced machine learning and statistical models have been employed to estimate the customer satisfaction score; however, there is not a single model that can yield the best result in all situations. Ensembles of regression techniques have demonstrated their effective-ness for various applications, where the success of these models lies in the con-struction of a set of single models. Iperformed an experimental study using a real dataset of 90917samples from US airline carrier 'Falcon airlines', in order verify the ben-efits of ensemble models for predicting customer satisfaction. Accordingly, in this project I evaluated the following models; Logistic Regression, Decision Tree, Bagging classifier and Random forest. The obtained results indicate that the Random forest performs better in terms of Recall and Precision.

***Keywords: Logistic** regression, customer satisfaction,Bagging classifier and Random forest.*

Introduction

Problem Description

Airline businesses globally are faced with the challenge of grounding especially due to the pandemic. To stay in business, Airline operators need to determine relative parameters that can contribute to the satisfaction of passengers.This is the dilemma of a reputed US airline carrier 'Falcon airlines'. They aim to determine the relative importance of each parameter with regards to their contribution to passenger satisfaction. To achieve this Aim a random sample of 90917 individuals who travelled using their flights is provided. The on- time performance of the flights along with the passengers' information is published in the Marketing Project-Flight data.csv file named 'Flight data'. These passengers were asked to provide their feedback at the end of their flights on various parameters along with their overall experience. These collected details are as available in the Marketing Project-Survey data.csv 'Survey data'.

In the survey, the respondents were requested to express their satisfaction or not with their overall flight experience and that is captured in the data of survey report under the variable labelled 'Satisfaction'. The need for this study by 'Falcon airlines' is to give themselves a competitive edge over other airline operators by identifying critical factors that lead to customer satisfaction.

Motivation For Project

It is obvious that at the end of the pandemic, there will be an increase demand for air travel as most persons may wish to be on vacations. Hence the need for 'Falcon airlines' to carry out this study to ascertain Passengers level of satisfaction with their services to give themselves a competitive edge over other airline .

Aim and Objectives

This project was a few weeks' efforts to develop a predictive model for Airline passenger satisfaction using data from US airline carrier 'Falcon airlines'. I hope the outcome of the project will help streamline the analysis and prediction of passenger's satisfaction for US airline carrier 'Falcon airlines' and other airlines.

The objective of this project are-

1. To understand which parameters play an important role in swaying a passenger feedback towards 'satisfied'.
2. To predict whether a passenger will be satisfied or not given the rest of the details are provided, using supervised machine learning techniques.
3. To compare different classification techniques to understand which is best suitable for this application.

No doubt the development of a framework and codes that incorporate analytics and machine learning concepts studied in the program is the goal. The success of the project is predicated on the accuracy of the classification results and the extent of analysis conducted. It is my hope that

.Working steps of Machine learning technique

Limitations of the Study

In this project, I evaluated the effectiveness of using specific supervised machinelearning techniques to address the problem of predicting passengerssatisfaction on Airline flight and survey data provided. Thelimitations of the methods applied in the course of this project study are as follows:

I used a pre-labelled dataset to train the algorithms. However, usually, it isdifficult to find labelled data and thus applying supervised machine learningtechniques may not be feasible. In such cases, the option should be to evaluate unsupervisedtechniques which were beyond the scope of this project.

Lterature Review

This part of the project seeks to review considerable literature on the subject matter. A few researchers havealso conducted literature reviews of articles published on airline passenger satisfaction and the techniques used.

the final report will serve as a benchmark for further development on this topic.

Research Methodology

The typical machine learning approach was followed in this project. The identified dataset has labelled class variable, which was used as the prediction variable in machine learning models.

Through exploratory analysis, we analysed the data set in detail and identified possible predictors.Through various visualization techniques, we observed the separation between Satisfied and neutral or non-satisfied assengers. To solve the Airline passenger satisfaction prediction problem, we experimented with a few supervisedmachine learning techniques – Logistic Regression, Decision tree, Bagging classifier and Random Forest,

Performance measures, like Confusion Matrix and Area Under Curve (AUC), wasused to compare the performance of the models.

This analysis was conducted using Python through Jupyter notebook. In-built librariesand methods were used to run the machine learning models. When needed, functionswere defined to simplify specific analyses or visualizations. The diagram below shows in detail the full process that was followed in the project

This project considers marketing survey and flight data that was mined byUS airline carrier 'Falcon airlines' only. I evaluated a few machine learning algorithm – Logistic Regression, Decision tree, Bagging classifier and RandomForest. Although the result of the study using these algorithms is good, it isnecessary to evaluate other techniques to determine which algorithm works bestfor this application. Due to the large size of data, I was limited by computation capacity to explore different other techniques

Air travel is one of the most convenient way for long distance travel at both national and international level [Park et al.,2009]. There are many airline service providers (ASPs) around the world. The competitive world motivates the airlines company to attract the customers.However, a traveller considers quite a few factors before deciding on any airline.

These points can be airfare, tour time, quantity of stoppages, number of baggage allowed, and existing customer feedback etc. Therefore, all ASPs are working in all these client service areas to enhance their facility and in-flight remedy in order to attract the customers.

It is very vital to recognize the desires and remedy level of customers i.e.customersatisfaction for the duration of the flight. Therefore, client remarks are veryimportant or any airline industry. There could be quitea possible pproaches to gather the customer feedback. The most easiest and regular way is the customer feedback structure available during the journey. However, most of the passengers do no longer show any activity in filling feedback forms. Another shortcoming of this strategy is that it may additionally or can also not have appropriate questionnaire and might also be biased on positive parameters i.e.the feedbackform mayalso only have sure unique qu estions. Other processes for purchaser feedback collection could be via online website or on-line mobile purposes of the airlines.

After the journey, an electronic mail with a link can be despatched to the passenger to request for a feedback. However, there is no guarantee of its success. Another strategy is to send a message on passenger's cellular cellphone and ask them to rate your provider (1 for negative and 5 for excellent) on certain parameters. All these standard strategies opted by means of the industry are limited to certain parameters only. The greater handy way for a passenger is to express their feedback, as they want.

Therefore, the most convenient way for the passengers to share their opinions is the social media instead of feedback form.Socialmedia pre sents a platform the place a consumer can freely categorical his feedbacks on any issues they observed all through flight. Twitter [https://en.wikipedia.org/wiki/Twitter] is one of the famous platforms worldwide. Theinformation from Twitter can be utilized to strengthen a recommender system [Abel et al.,2019]. In addition, travellers are more comfortable in sharing their views about tour experiences on Twitter.

A variety of fundamental issues influences the emotions of a passenger in air travel. These issues can be cabin crew behavior, food quality, loss of baggage, seat comfort, flight delay, airfare etc. All these issues may give upward push to each superb and bad emotion. Also, ifthere is a continuous trend of terrible tweets for an airline, then it might also put a negativeimpact to the financial growth of the airline company. Therefore, it is ital to understand the issues that provide upward e to terrible tweets so that the respective airline company can take splendid action on time. There are large number airlines operating every day to join different geographical places [www.quor a.com]. Therefore, we might so anticipate alarge number of people journeyin g every day in these flights. In addition, the number of tweets by passengers for airways would be very large. Therefore, it is a difficult assignment toextract the hidden emotion at the back of a tweet. Therefore, we required some tools and techniquesthat are in a position to take care of such a large quantity of tweet database and can provideinsights to assist airline industry

Sumary of related work

Pramod., et al 2019] in their work"a literaturec review: customer satisfaction on airline tweet susing machine learning"; presents the earlier work done by the various researchers in the field of customersatisfaction of airline tweets using machine-learning techniques such as logistic regression, SVM, KNN, randomforest,Naïve Bayesclassifier, Adaboostetc as shown below

Authors	Description	Publishing Year
T.HemakalaandS.Santhoshkumar	In this research, design a framework for sentiment analysis with opinion mining for thecase of airlines service feedback.Most available datasets ofhotel reviews are not labelled which presents many works for researchers as faras text data pre-processing task is concerned. Twitter is a SNS that has a hugedata with user posting, with this significant amount of data, it has the potentialof research related to text mining and could be subjected to sentiment analysis.The airline industry is a very competitive market, which has grown rapidly inthe past 2 decades. Airline companies resort to traditional customer feedbackforms which in turn are very tedious and time consuming. In this work, workedon a dataset comprising of tweets for 6 major Indian Airlines and performed amulti-classs entiment analysis.This approach starts with pre-processing techniques used to clean the tweets and then representing these tweets as vectors using a deep learning concept to do a phrase-level analysis.The analysis was carried out using 7 different classification strategies: DecisionTree,RandomForest,SVM,K-Nearest Neighbors, Logistic Regression, Gaussian Naïve Bayes and Ada Boost. The outcome of the test set is the tweetsentiment.	2018
GuoningHuetal.	Analyzetheopinionof19MTwitteruserstowards62popularindustries, encompassing 12,898 enterprise and consumer brands, as well as associatedsubject matter topics, via sentiment analysis of 330M tweets over a eriodspanningamonth.We find that users tend to be most positive towards manufacturing and most negative towards service industries. In addition,heytend to be more positive ornegative when interacting with brands thangenerally on Twitter. We also find that sentiment towards brands within anindustry varies greatly and we demonstrate this using two industries as usecases. In addition, we discover that there is no strong correlation between topics entiments of different industries,demonstrating that topics entiments are highly dependent on the context of the industry that they are mentioned in.We demonstrate the value of such ananalysis in order to assess the impactof brands on social media. We hope that this initial study will prove valuable for both researchers and companies in understanding users‘perceptionofi ndustries, brands	2017
	and associated topics and encourage more research in thisfield.	

YasminYashodha	This study examines the extensive strategic analysis of Air AsiaBerhad that hasenabledit tosustain its competitiveadvantageasAsia'sleading lowcostcarrier (LCC). The study demonstrates the diverse business-level, corporatelevel and competitive strategies of Air AsiaBerhad, played crucial roles in theLCC to successfully penetrate the under-served market segment of the airlineindustry within the ASEAN region. An in-depth analysis using a wide array ofacademic resources, relevant financial, legal and management resources andauthorized websites, including face-to-face interviews were used to provide amore consequential comprehension on the varied business and internationalstrategies that were implemented by AirAsiaBerhad. This research exhibitscritical analysis pertaining to the current macro environment of the aviationindustrywhichincludesthePESTELframeworkandPorter'sIn ustryAnalysis.ThecompetitiveenvironmentanalysisforAirAsiaBerha disthoroughlyscrutinisedtoexaminethedrivingdeterminantsthatattrib utedtotheorganization'scompetitiveadvantage inthe industry.	2012
M.Vadivukarassie al.	In Twitter, the customer of airline services can tweet their opinions about theirtravelled experiences in flight. So Twitter contains massive amount of data andinformation regarding airline services. These tweets are collected and exploredthe sentiments about the airline services to track customer satisfaction reportsand todiscoverlocationofthecustomer.	2018
Janet R. McColl-Kennedyetal.	Contextualized in post purchase consumption in business-to-business settings,the authors contribute to customer experience (CX) management theory andpractice in three important ways. Firs by offering a novel CX conceptualframework that integrates prior CX research to better understand, manage, andimproveCXs—comprisedofvaluecreationelements(resources,activities,context,inte actions,andcustomerrole),cognitiveresponses,anddiscreteemotionsa touchpointsacrossthecustomerjourney.Second,bydemonstratingtheu efulnessofalongitudinalCXanalyticbasedontheconceptual framework that combines quantitative and qualitativemeasures.Third, by providing a step-by-step guide for implementing the text miningapproach in practice, thereby showing that CX analytics that apply big datatechniques to the CX can offer significant insights that matter. The authorshighlightsixkeyinsightspractitionersneedinordertomanageth ircustomers' journey, through (1) taking a customer perspective, (2 identifyingrootcauses,(3)uncoveringat-risksegments,(4)capturingcustomers'emotionalandcognitiverespons es,(5)spottingandpreventingdecreasingsales, and (6) prioritizing actions to improve CX. The article concludes withdirectionsfor future research.	2018

Rida Khan and SiddhalingUrolagin	Social media today is an integr al part of people’sdaily routines and the live lihood of some. As a result, it is abundant in user opinions. The analysis of brand specific opinions can inform companies on the level of satisfaction within consumers.This research focus is on analysis of tweets related to airlines based in fourregions:Europe,India, Australia and America for consumer loyalty prediction.Sentiment Analysis is carried out using Text Blobanalyzer.	2018
AnkitaRaneand	The airline industry is a very competitive market which has grown rapidly in the past 2 decades. Airline companies resort to traditional customer feedback	2018
nandkumar	Forms which in turn are very tedious and time consuming.This is where Twitter data serves as a good source to gather customer feedback tweets and perform a sentiment analysis. In this paper, we worked on a dataset comprising of tweets for 6 major US Airlines and perform edamulti-classsentimentanalysis. This approach starts off with pre-processing techniques used to cleanthe tweets and then representing these tweets as vectors using a deep learningconcept (Doc2vec) to do a phrase-level analysis. The analysis was carried outusing7different classification strategies: Decision Tree, Random Forest, SVM, K-Nearest Neighbors, Logistic Regression, Gaussian Naïve Bayes andAda Boost. The classifiers were trained using 80% of the data and tested usingthe remaining 20% data. The outcome of the test set is the tweet sentiment(positive/negative/neutral). Based on the results obtained, the accuracies were calculated to draw a comparison between each classification approach and the overall sentiment count was visualized combining all six airlines.	
YunWanandQigangGao	In airline service industry, it is difficult to collect data about customers' feedback by questionnaires, but Twitter provides a sound data source for them to do customers entiment analysis.However,little research has beendoneinthe domain of Twitter sentiment classification about airline services. In this paper,an ensembles entiment classification strategy was applied based on Majority Vote principle of multiple classification methods, including NaiveBayes, SVM,Bayesian Network, C4.5 Decision TreeandRandom Forest algorithms. In our experiments, six individual classification approaches,and the proposed ensemble approach were all trained and tested using the samedataset of 12864 tweets, in which 10 fold evaluation is used to validate the classifiers. The results show that the proposed ensemble approach outperformstheseindividual classifiersinthisairline serviceTwitter dataset.	2015

MustafaAltinkök	This research was conducted for the purpose of analyzing the effect of the movement education program through a12-week-coordination on the development of basic motor movements of pre-school children. A total of 78students of pre-school period, 38 of whom were in the experimental group and40 of whom were in the control group, were incorporated into the study in line with theirown consent after their families had also beeninformed.	2016
M.Vadivukarassie al.	Analyzed the twitter airline dataset for finding the best and the worst airlines and alsotopredict the most common issues occurred during the airline services. Then the word clouds of negative tweets are created and also thelocation of the negatively tweeted customer is predicted and visualized using geographical analysis. Finally training and testing was done on the dataset and also compared with seven different classifierssuch as Logistic Regression classifier, KNeighbors classifier, SVC, Decision Treeclassifier,RandomForestclassifier,AdaBoostclassifierandGaussianNB.Theresultsofourexperiments demonstrate that the Random forest approach works best in real-world practice onsentimentclassificationoftweetdata.	2018
BeeYeeLiau and Pei PeiTan	The purpose of this paper is to study the consumer opinion towards the low-costairlinesorlow- cost carriers(LCCs) (the set woter msare used inter changeably) industry in Malaysia to better understand consumers' needs and to provide better services. Sentiment analysis is undertaken in revealingcurrent customers' satisfaction level towards low-cost airlines. About 10,895 tweets (data collected for two and a half months) are analyzed. Text miningtechniques are used during data pre-processing and a mixture of statistic altechniques are used to segment the customers'opinion.There sults with two different sentiment algorithms show that there is more positive than negative polarity across the different algorithms.ClusteringresultsshowthatbothK-	2016
	Me an sands phericalK-Meansalgorithms delivered similar results and the four main topics that are discussed by theconsumerson Twitter are customer service, LCC stickets promotions, flight cancellations and delays and post-booking management.	
XiangJietal.	An important task of public health officials is to keep track of health issues,such as spreading epidemics. In this paper, we are addressing the issue of spreading public concern about epidemics.Public concern about acommunicable disease can be seen as a problem of its own. Keeping track oftrends in concern about public health and identifying peaks of public concernare therefore crucial tasks. However, monitoring public health concerns is not only expensive with traditional surveillance systems,butal so suffers from limited coverage and significant delays.	2015

Methodology

This methodology served as the deliverables of the project. It describes the results each phase that wastried out and do a comparison between them to i dentify which is the best technique to address theairline passenger satisfaction prediction problem.

Eachphaseoftheprojecthasanoutputthatdescribes thefindingsinthatphase.These deliverables were used inthisfinalprojectareexplainedbelow

projectdeliverables

MethodologyPhases	ProjectDeliverables
Understandingthedataset	• Reportonthesummaryofthedatasetandeachvariable it containsalong with necessaryvisualizations
ExploratoryDataAnalysis	• Report on analysis conducted and critical findings with a full description of data slices considered • Visualizations and charts that show the differencesbetweensatisfied andneutral or non-satisfied passengers • Pythoncodeof theanalysisperformed
Modeling	• Report on the results of the different techniques tried out,iterations that were experimented with, data transformations and the detailed modelingapproach • Python code used to build machine learning models
FinalProjectReport	• Final report summarizing the work done over the course of the project, highlighting the key findings, comparing different models and identifying best model for predicting airline passenger satisfaction

Tools used

This project was entirely done using Python, and the analysis was documented in aJupyternotebook. Standard python libraries were used to conduct different analyses.These libraries are described below–

- ***sklearn***–usedformachine learningtasks
- ***seaborn***–usedtogeneratechartsandvisualizations
- ***pandas***–usedforreadingandtransformingthedata
- ***Gridsearch*** – used for model tuning

Dataset

The problem consists of 2 separate datasets: Flight data & Survey data. The flight data has information related to passengers and the performance of flights in which they travelled. The survey data is the aggregated data of surveys collected post service experience. You are expected to treat both the datasets as raw data and perform any necessary cleaning/validation steps as required

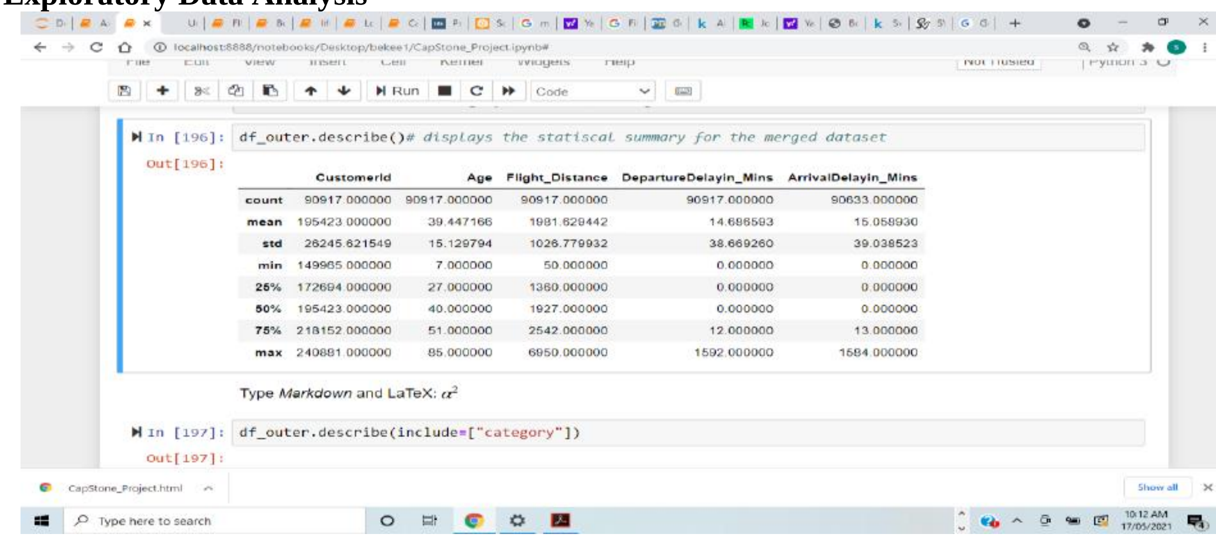
Data Information

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90917 entries, 0 to 90916
Data columns (total 24 columns):
 #   Column                                   Non-Null Count  Dtype
---------------------------
 0   CustomerId                                90917 non-null  int64
 1   Satisfaction                              90917 non-null  object
```

2 Seat_comfort 90917 non-null object
3 Departure_Arrival_time_convenient 82673 non-null object
4 Food_drink 82736 non-null object
5 Gate_location 90917 non-null object
6 Inflightwifi_service 90917 non-null object
7 Inflight_entertainment 90917 non-null object
8 Online_support 90917 non-null object
9 Ease_of_Onlinebooking 90917 non-null object
10 Onboard_service 83738 non-null object
11 Leg_room_service 90917 non-null object
12 Baggage_handling 90917 non-null object
13 Checkin_service 90917 non-null object
14 Cleanliness 90917 non-null object
15 Online_boarding 90917 non-null object
16 Gender 90917 non-null object
17 CustomerType 81818 non-null object
18 Age 90917 non-null int64
19 TypeTravel 81829 non-null object
20 Class 90917 non-null object
21 Flight_Distance 90917 non-null int64
22 DepartureDelayin_Mins 90917 non-null int64
23 ArrivalDelayin_Mins 90633 non-null float64

Objects will need to be converted to category as demonstrated with a few variable.
DATA TYPES: float64(1), int64(4), object(19)

Exploratory Data Analysis



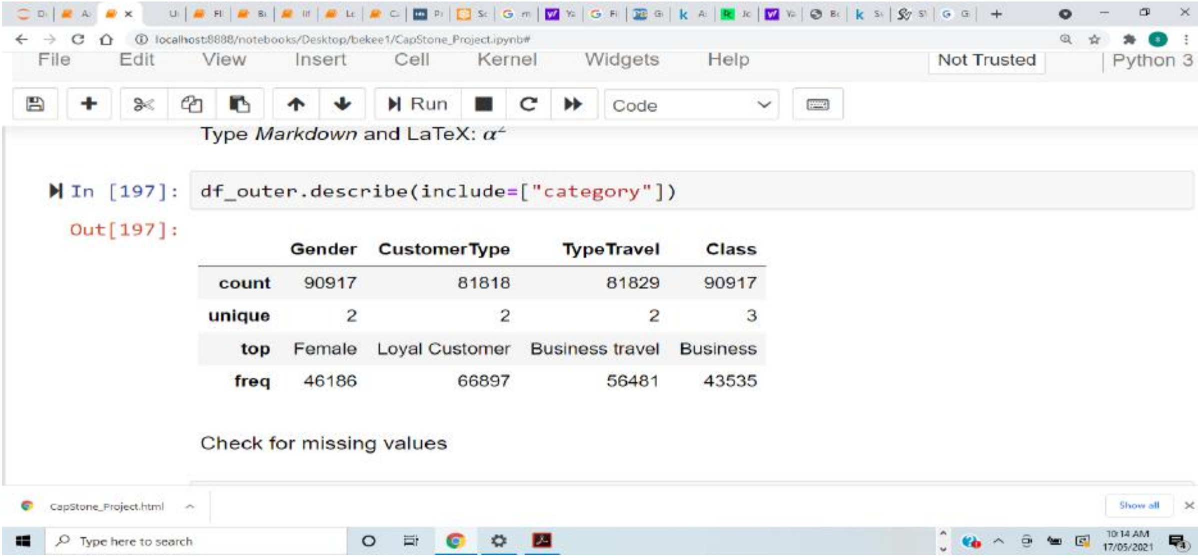
```
In [196]: df_outer.describe()# displays the statiscal summary for the merged dataset
Out[196]:
```

	CustomerId	Age	Flight_Distance	DepartureDelayin_Mins	ArrivalDelayin_Mins
count	90917.000000	90917.000000	90917.000000	90917.000000	90633.000000
mean	195423.000000	39.447166	1981.629442	14.686593	15.058930
std	26245.621549	15.129794	1026.779932	38.669260	39.038523
min	149965.000000	7.000000	50.000000	0.000000	0.000000
25%	172694.000000	27.000000	1360.000000	0.000000	0.000000
50%	195423.000000	40.000000	1927.000000	0.000000	0.000000
75%	218152.000000	51.000000	2542.000000	12.000000	13.000000
max	240881.000000	85.000000	6950.000000	1592.000000	1584.000000

Type *Markdown* and LaTeX: α^2

```
In [197]: df_outer.describe(include=["category"])
Out[197]:
```

Mean Age respondents is 39.4 and Standard deviation 15.1,mean light distance 1981.6 & Standard deviation 1026.8, mean DepartureDelayin_Mins 14.6 Std 38. ArrivalDelayin_Mins 15.1 Std 39.0



Type *Markdown* and LaTeX: α^2

```
In [197]: df_outer.describe(include=["category"])
```

Out[197]:

	Gender	CustomerType	TypeTravel	Class
count	90917	81818	81829	90917
unique	2	2	2	3
top	Female	Loyal Customer	Business travel	Business
freq	46186	66897	56481	43535

Check for missing values

Gender have unique count of 2, top being female, CustomerType have unique count of 2, top being Loyal Customer with count indication some value are missing same to TypeTravel and Class with 3 unque count

Data set

The problem consists of 2 separate datasets: Flight data & Survey data. The flight data has information related to passengers and the performance of flights in which they travelled. The survey data is the aggregated data of surveys collected post service experience. You are expected to treat both the datasets as raw data and perform any necessary cleaning/validation steps as required

Data Information

<class 'pandas.core.frame.DataFrame'>
Int64Index: 90917 entries, 0 to 90916
Data columns (total 24 columns):

#	Column	Non-Null Count	Dtype
0	CustomerId	90917 non-null	int64
1	Satisfaction	90917 non-null	object
2	Seat_comfort	90917 non-null	object
3	Departure_Arrival_time_convenient	82673 non-null	object
4	Food_drink	82736 non-null	object
5	Gate_location	90917 non-null	object
6	Inflightwifi_service	90917 non-null	object
7	Inflight_entertainment	90917 non-null	object
8	Online_support	90917 non-null	object
9	Ease_of_Onlinebooking	90917 non-null	object
10	Onboard_service	83738 non-null	object
11	Leg_room_service	90917 non-null	object
12	Baggage_handling	90917 non-null	object
13	Checkin_service	90917 non-null	object
14	Cleanliness	90917 non-null	object
15	Online_boarding	90917 non-null	object
16	Gender	90917 non-null	object
17	CustomerType	81818 non-null	object
18	Age	90917 non-null	int64
19	TypeTravel	81829 non-null	object

20 Class   90917 non-null object
21 Flight_Distance   90917 non-null int64
22 DepartureDelayin_Mins   90917 non-null int64
23 ArrivalDelayin_Mins   90633 non-null float64
Objects will need to be converted to category as demonstrated with a few variable.
DATA TYPES: float64(1), int64(4), object(19)

Data Preprocessing

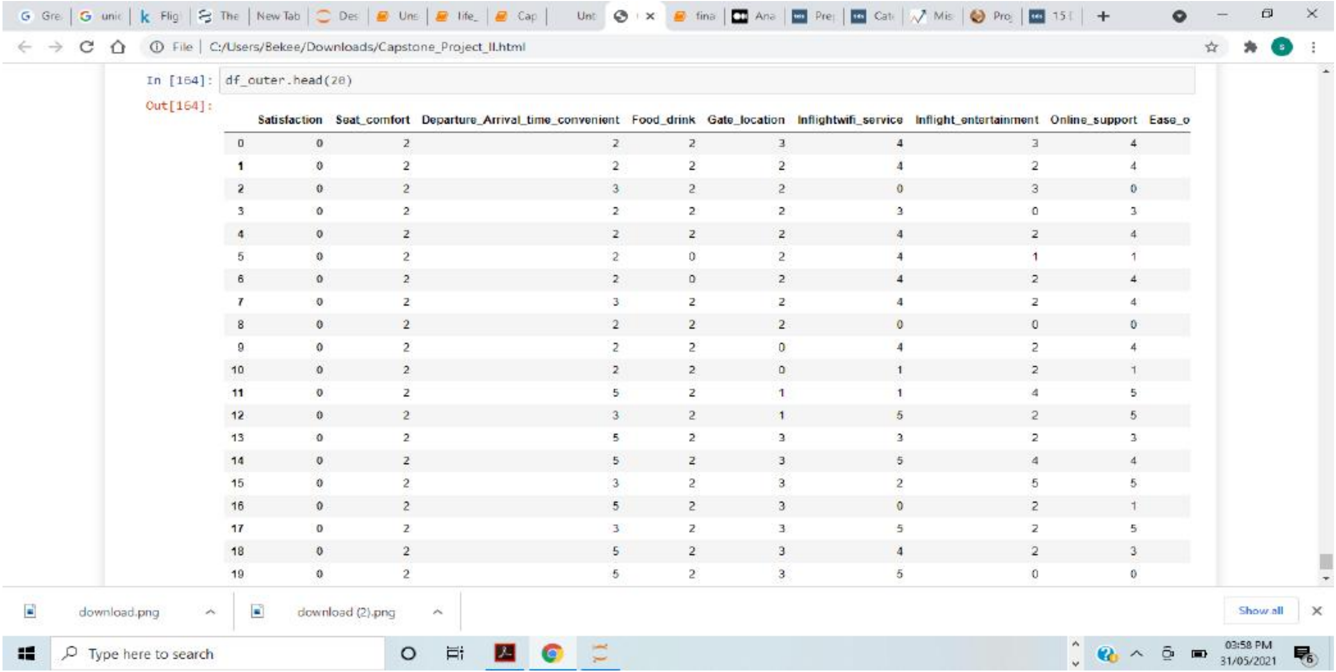
The data set contains CustomerId variable which of type int, in my opinion I feel this variable may not add any significant value to prediction variable so I decided to drop it.

```
df_outer.drop(
columns=["CustomerId"], inplace=True)
```

Fixing Categorical Variables (Data Type)

Before putting our data through models, two steps that need to be performed on categorical data is encoding and dealing with missing nulls. Encoding is the process of converting text or boolean values to numerical values for processing. This approach was adopted for 19 columns that are of type object. First the object were converted to Categories and the Ordinal values were passed as shown below;categorical_variables=df_outer.select_dtypes(exclude=["number","bool_"]).columns.tolist()# *list of categorical varaibles*

```
forcolmnincategorical_variables:
df_outer[colmn]=df_outer[colmn].astype('category')
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90917 entries, 0 to 90916
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
----------------------------
 0   Satisfaction                    90917 non-null  category
 1   Seat_comfort                    90917 non-null  category
 2   Departure_Arrival_time_convenient S  82673 non-null  category
 3   Food_drink                      82736 non-null  category
 4   Gate_location                   90917 non-null  category
 5   Inflightwifi_service            90917 non-null  category
 6   Inflight_entertainment          90917 non-null  category
 7   Online_support                  90917 non-null  category
 8   Ease_of_Onlinebooking           90917 non-null  category
 9   Onboard_service                 83738 non-null  category
 10  Leg_room_service                90917 non-null  category
 11  Baggage_handling                90917 non-null  category
 12  Checkin_service                 90917 non-null  category
 13  Cleanliness                     90917 non-null  category
 14  Online_boarding                 90917 non-null  category
 15  Gender                          90917 non-null  category
 16  CustomerType                    81818 non-null  category
 17  Age                             90917 non-null  int64
 18  TypeTravel                      81829 non-null  category
19 Class  n-null
```



In [164]: `df_outer.head(20)`

Out[164]:

	Satisfaction	Seat_comfort	Departure_Arrival_time_convenient	Food_drink	Gate_location	Inflightwifi_service	Inflight_entertainment	Online_support	Ease_o
0	0	2	2	2	3	4	3	4	
1	0	2	2	2	2	4	2	4	
2	0	2	3	2	2	0	3	0	
3	0	2	2	2	2	3	0	3	
4	0	2	2	2	2	4	2	4	
5	0	2	2	0	2	4	1	1	
6	0	2	2	0	2	4	2	4	
7	0	2	3	2	2	4	2	4	
8	0	2	2	2	2	0	0	0	
9	0	2	2	2	0	4	2	4	
10	0	2	2	2	0	1	2	1	
11	0	2	5	2	1	1	4	5	
12	0	2	3	2	1	5	2	5	
13	0	2	5	2	3	3	2	3	
14	0	2	5	2	3	5	4	4	
15	0	2	3	2	3	2	5	5	
16	0	2	5	2	3	0	2	1	
17	0	2	3	2	3	5	2	5	
18	0	2	5	2	3	4	2	3	
19	0	2	5	2	3	5	0	0	

Missing Value

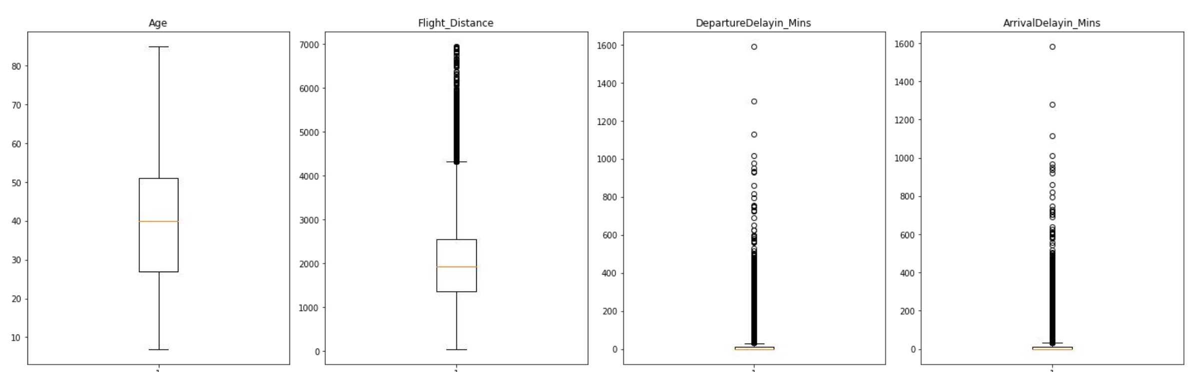
Missing value in a dataset is a very common phenomenon in the reality. Missing value correction was carried out on the data set to reduce bias and to produce data set suitable modelling. The following variables contained missing values .

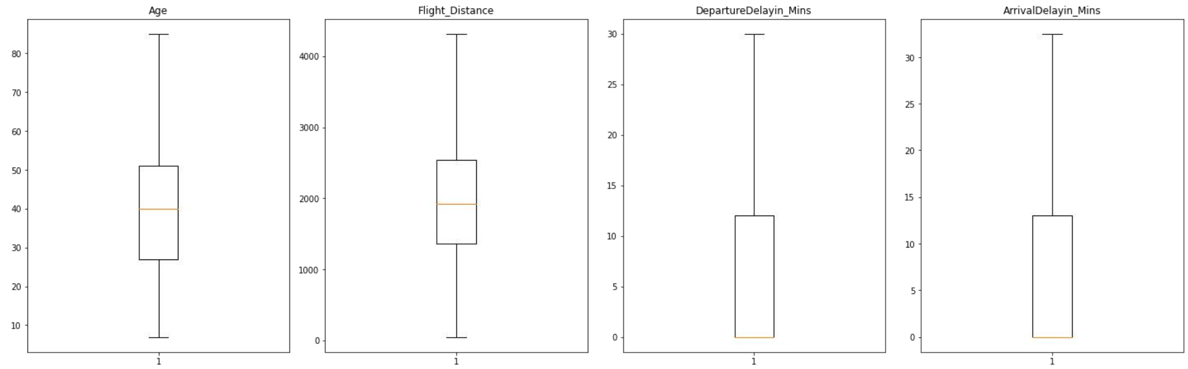
: Departure_Arrival_time_convenient 8244
Food_drink 8181
Onboard_service 7179
CustomerType 9099
TypeTravel 9088
ArrivalDelayin_Mins 284

They were corrected with the following codes for numerical and categorical respectively

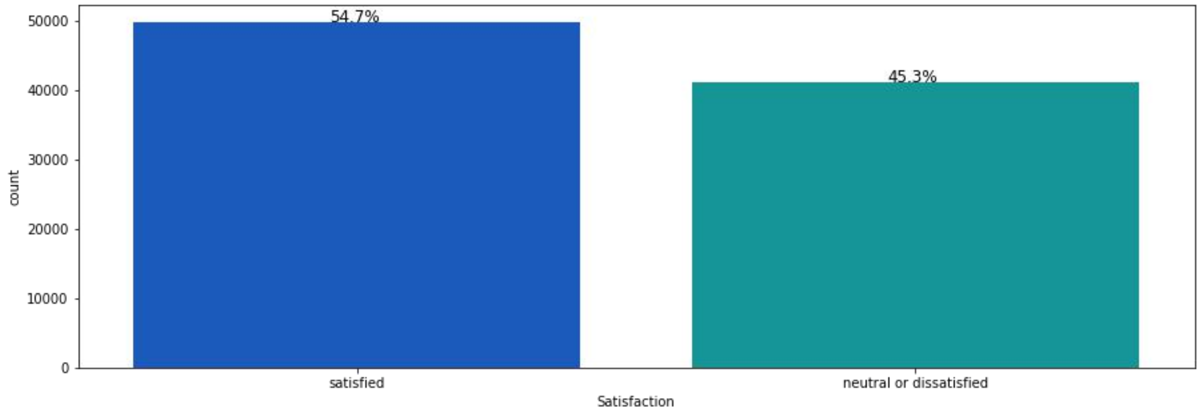
Outlier Detection & Treatment

Outlier detection and treatment was carried out on the dataset, the outcome is shown below;

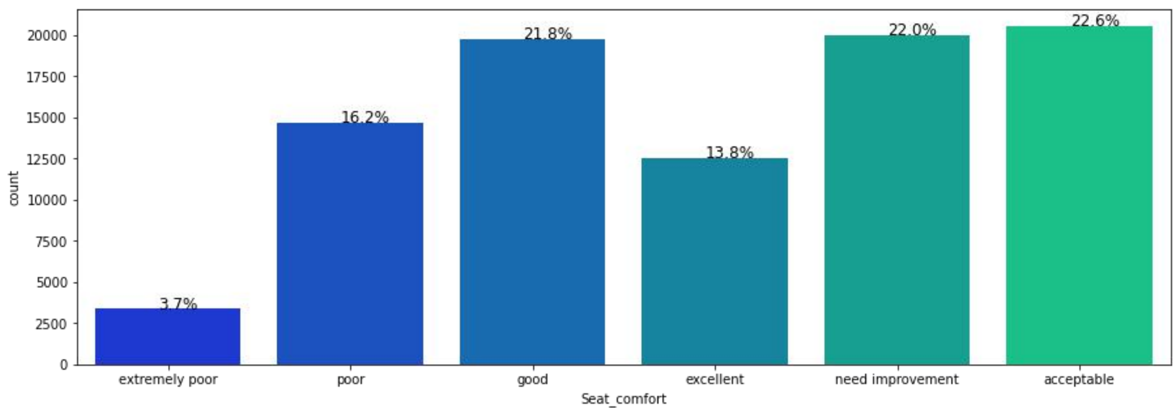




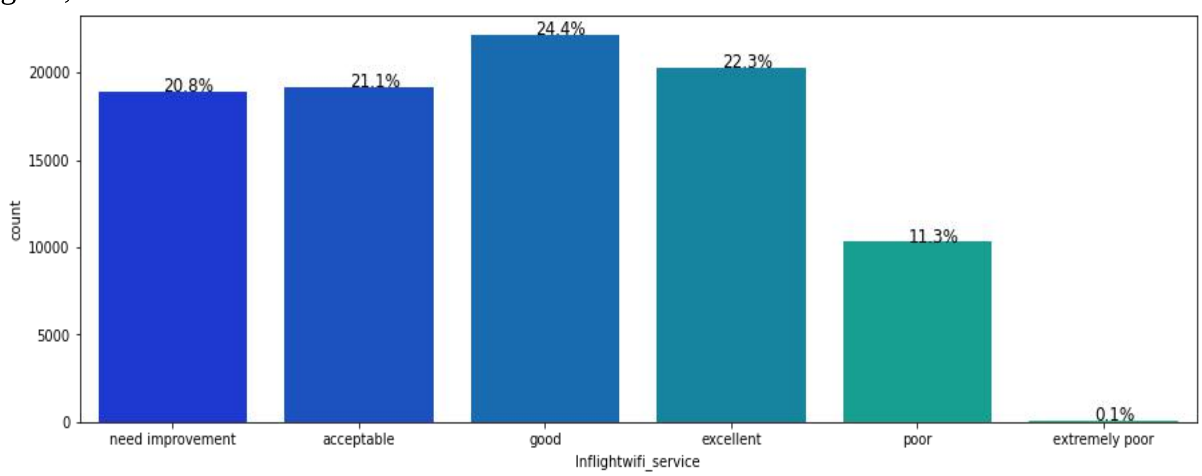
Exploratory Data Analysis



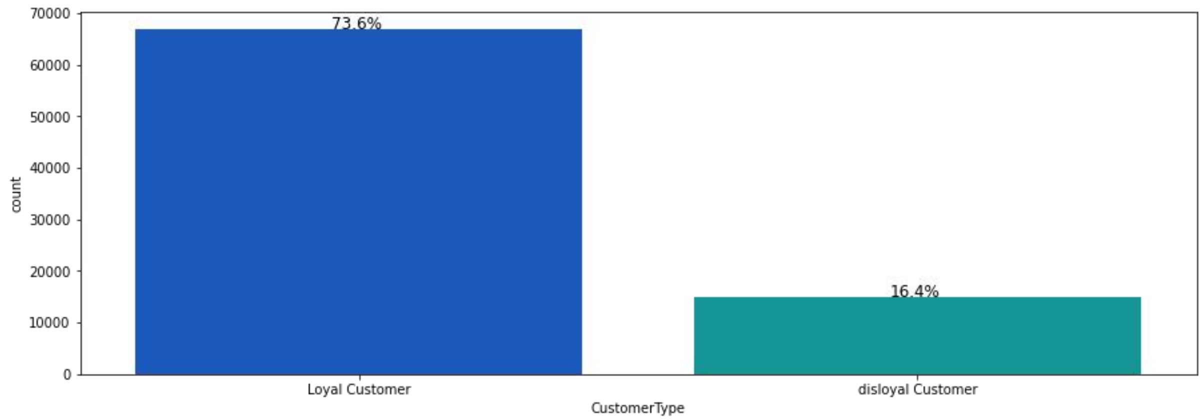
54.7 % of the passengers are satisfied while 45.3% are of the respodent fall into the group neutral or dissatisfied



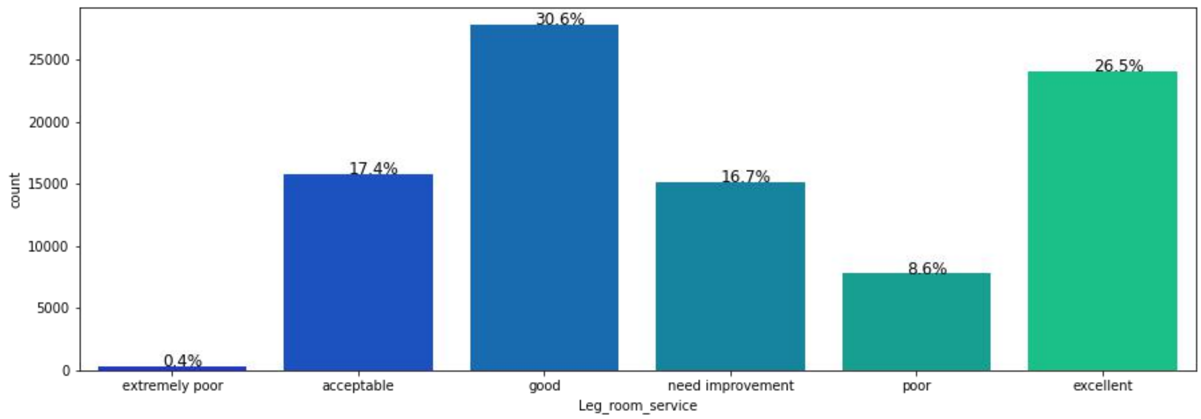
Seat_Comfort 22.6% respodent acceptable, 22% indicate seat need improvement, 21.8% indicate good, 13.8 % excellent



inflight service has six categories with 24% good, 22.3 excellent and 21%, 20.8 acceptable & need improvement repectively



About 74% the repondent are loyal customers of the



Airline Good has the has highest percentage, followed by excellent and acceptable with 17 & 16.7 respectively

Alternate Analytical Approach

The two objective of this project are-

1.To understand which parameters play an important role in swaying a passenger feedback towards 'satisfied'.

2. To predict whether a passenger will be satisfied or not given the rest of the details that are provided in the data set.

The alternate approach to this problem would be to build, test and implement a classification model.

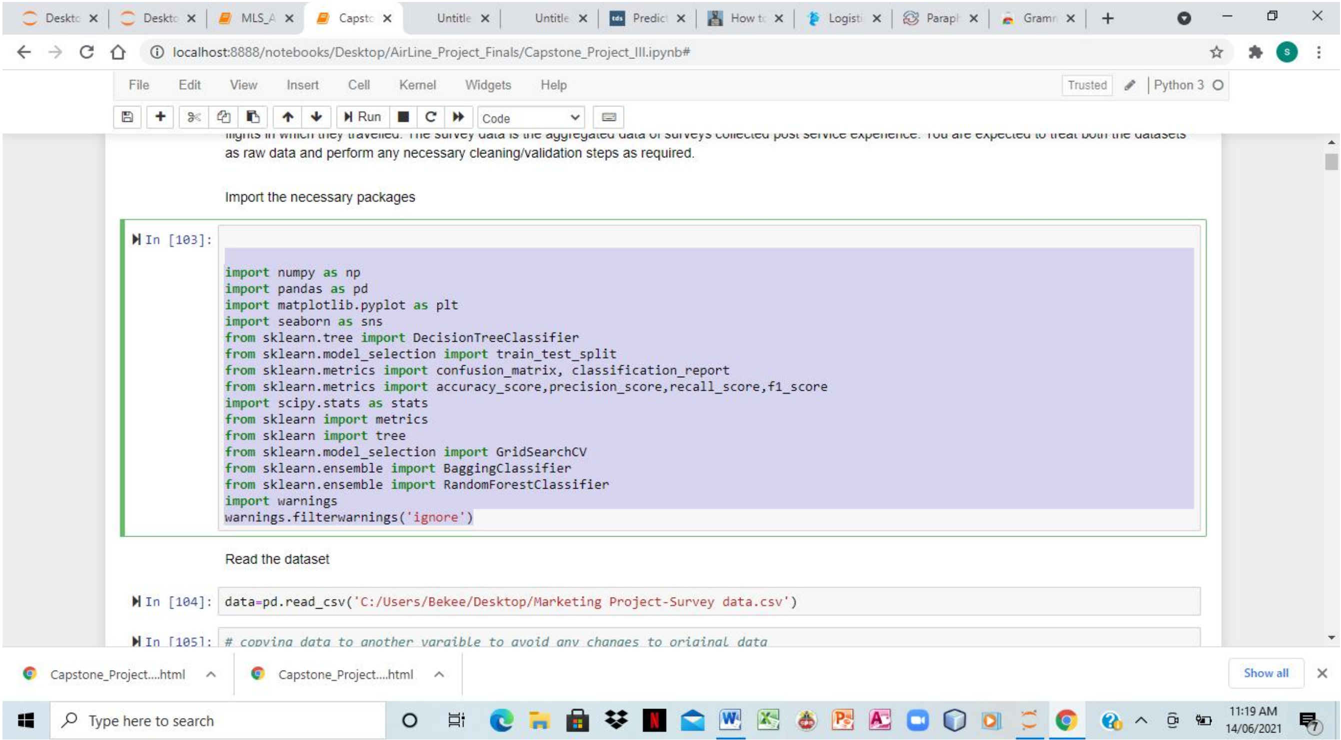
Building Model

Logistic regression is a is a type of supervised machine learning used to predict the probability of a target variable. The most common logistic regression models a binary outcome; The output of the dependent variable is represented in discrete values such as 0 and 1, true/false, yes/no, etc. In some case logistic regression can model scenarios where there are more than two possible discrete outcomes that is referred to as Multinomial logistic regression. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber Security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

With the above explanation, it is clear that logistic regression modelling technique suit the Airline passenger satisfaction prediction problem.

Importing the libraries

To build our model, the first step is to import the necessary libraries. I used the Pandas library to load in the CSV or the dataset, and Numpy to convert the data frame into arrays.



The second step was to define the target variable(Y) and the independent variables, and then split the data set into the training set and the test set. We will use the training set to train our logistic regression algorithm. Similarly, the test data set will be used to validate the logistic regression model.

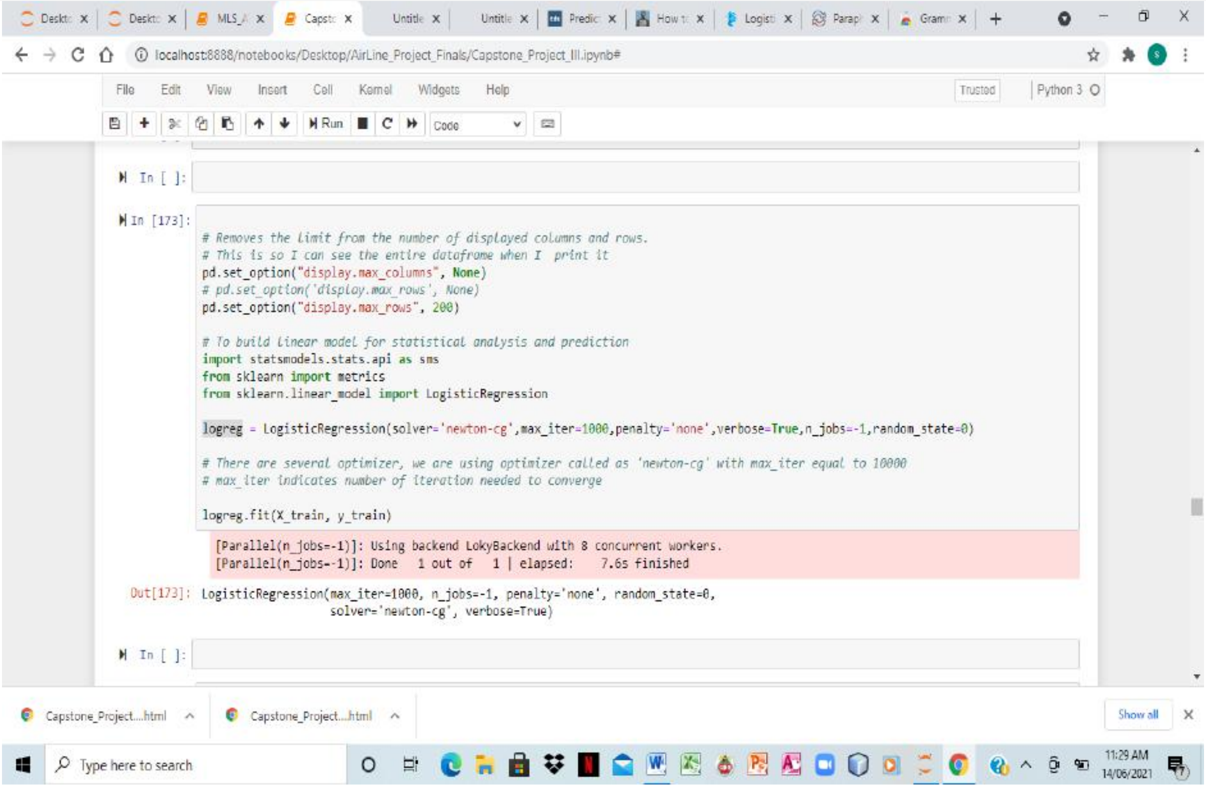
To split the data into two sets, we will use Sklearn. The train_split_function can be used and we can specify the amount of data we want to set aside for training and testing

```
From sklearn.model_selection import train_test_split
    # Splitting data into training and test set:
    X_train, X_test, y_train, y_test =train_test_split(X, y, test_size=0.3, random_state=1,stratify=y)
    print(X_train.shape, X_test.shape)
    (63641, 22) (27276, 22)
```

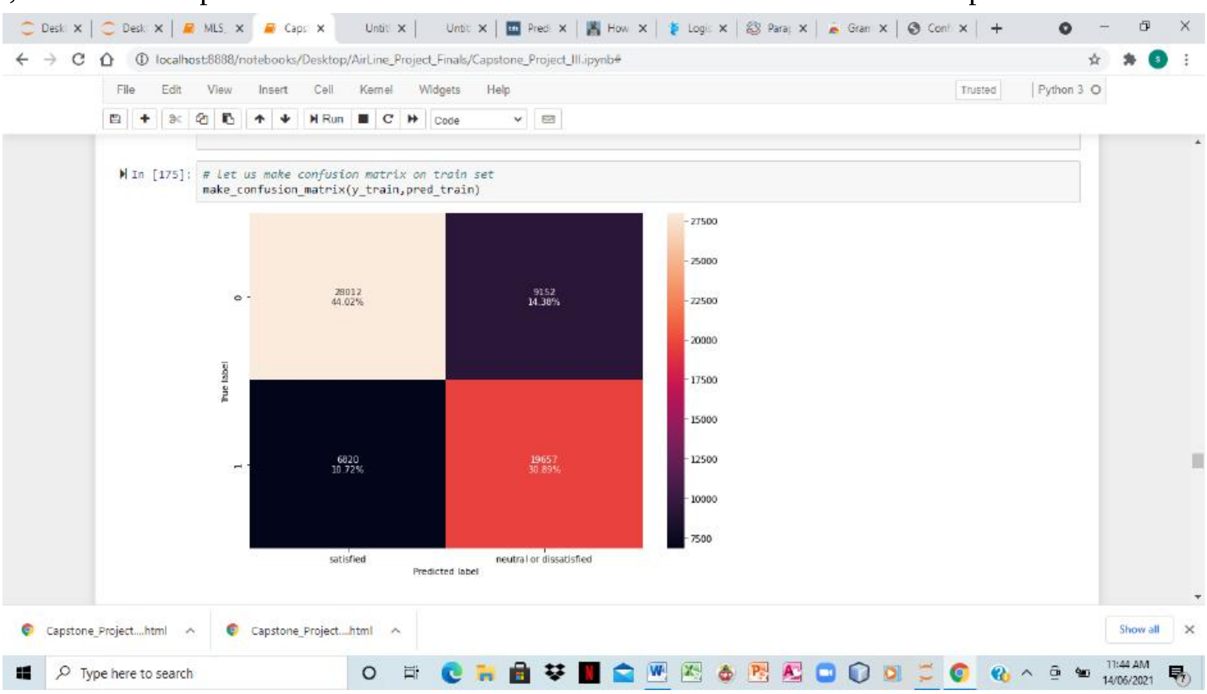
Building The Logistic Regression Mode

Next was to build the logistic regression model and fit it to the training data set. First, we will need to import the logistic regression algorithm from Sklearn

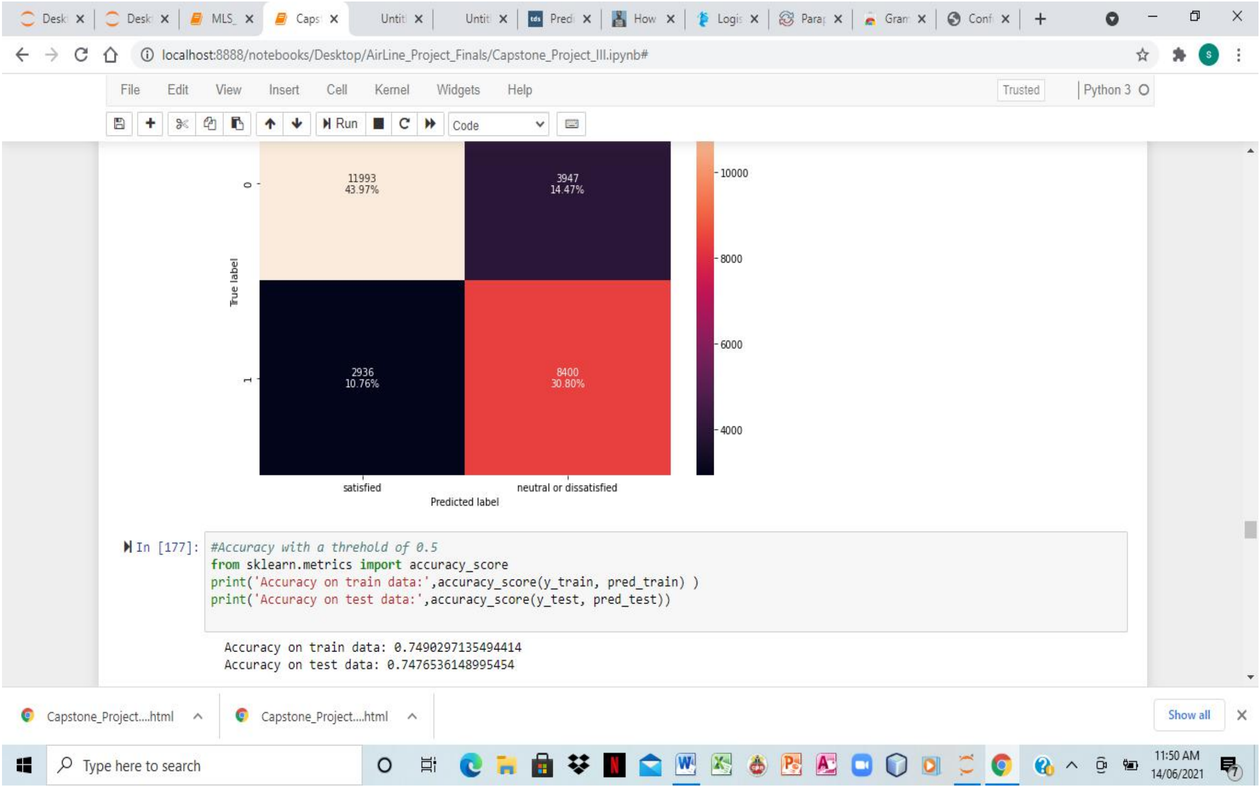
Accuracy	0.7490	ACC = (TP + TN) / (P + N)



Next, was to create predictions on the test dataset and train data with the model performance as shown

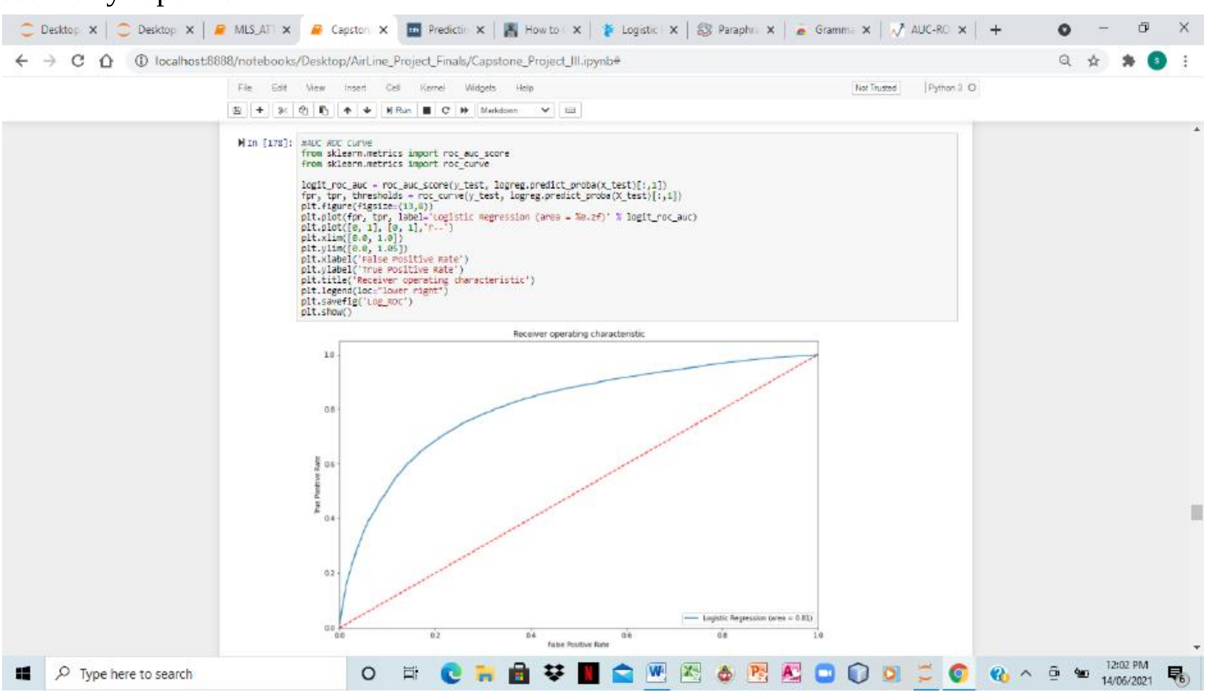


low Accuuracy of 74% and pression 75%



Next the AUC-ROC curve was used to visualize how well our machine learning classifier is performing.

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The result obtained is as below



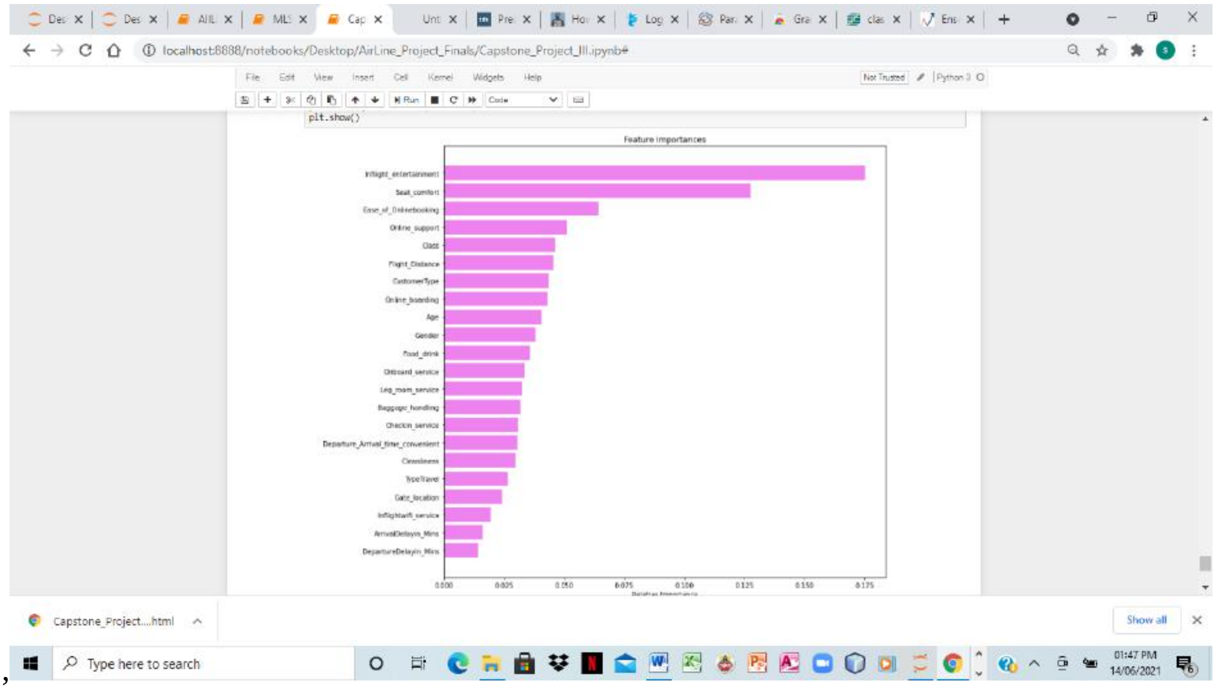
The AUC of 0.81% looks quite good performance .

Build Different Models

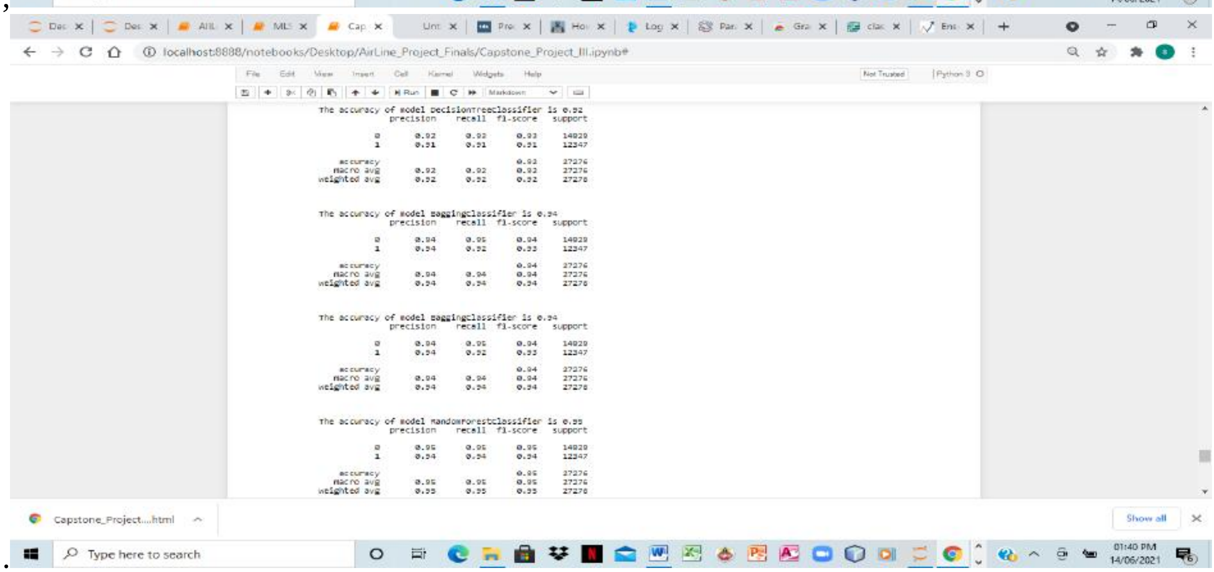
DecisionTreeClassifier, RandomForestClassifier, BaggingClassifier were used on the data set. A result of the comparison of the different models are as bellow The best model is the rf, rf_wt performance of 95% accurate prediction on Recall and Precession.

Feature Importance

The RandomForestClassifier was used to check for featre importance and the result obtained is as below



,



```
The accuracy of model DecisionTreeClassifier is 0.92
              precision    recall  f1-score   support

           0       0.92      0.92      0.93     14929
           1       0.91      0.91      0.91     12347

    accuracy                           0.92     27276
   macro avg       0.92      0.92      0.92     27276
weighted avg       0.92      0.92      0.92     27276


The accuracy of model BaggingClassifier is 0.94
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     14929
           1       0.94      0.92      0.93     12347

    accuracy                           0.94     27276
   macro avg       0.94      0.94      0.94     27276
weighted avg       0.94      0.94      0.94     27276


The accuracy of model BaggingClassifier is 0.94
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     14929
           1       0.94      0.92      0.93     12347

    accuracy                           0.94     27276
   macro avg       0.94      0.94      0.94     27276
weighted avg       0.94      0.94      0.94     27276


The accuracy of model RandomForestClassifier is 0.95
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     14929
           1       0.94      0.94      0.94     12347

    accuracy                           0.95     27276
   macro avg       0.95      0.95      0.95     27276
weighted avg       0.95      0.95      0.95     27276
```

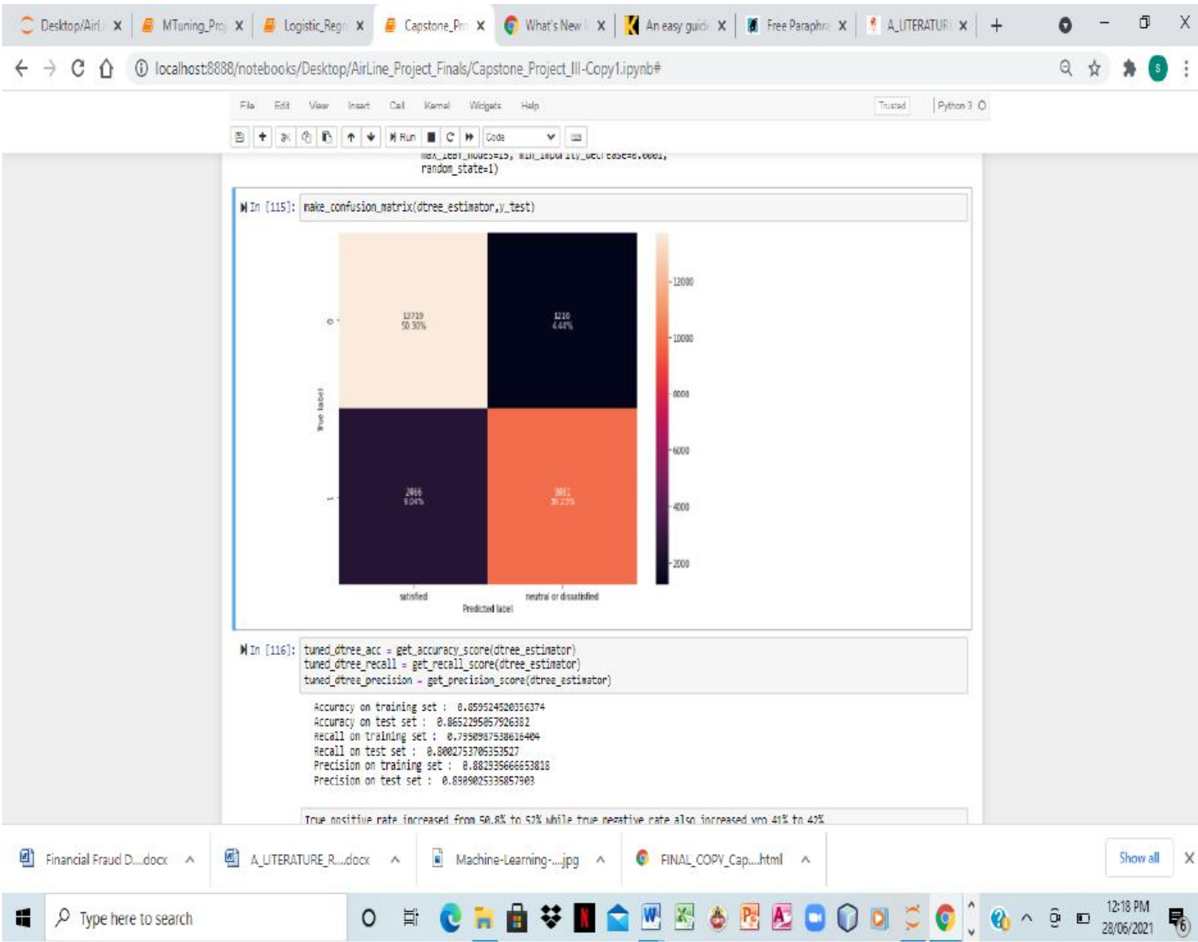
.

Inflight_entertainment is the most important feature for prediction followed by Seat_comfort ,Ease_of_Onlinebooking and Online_support.

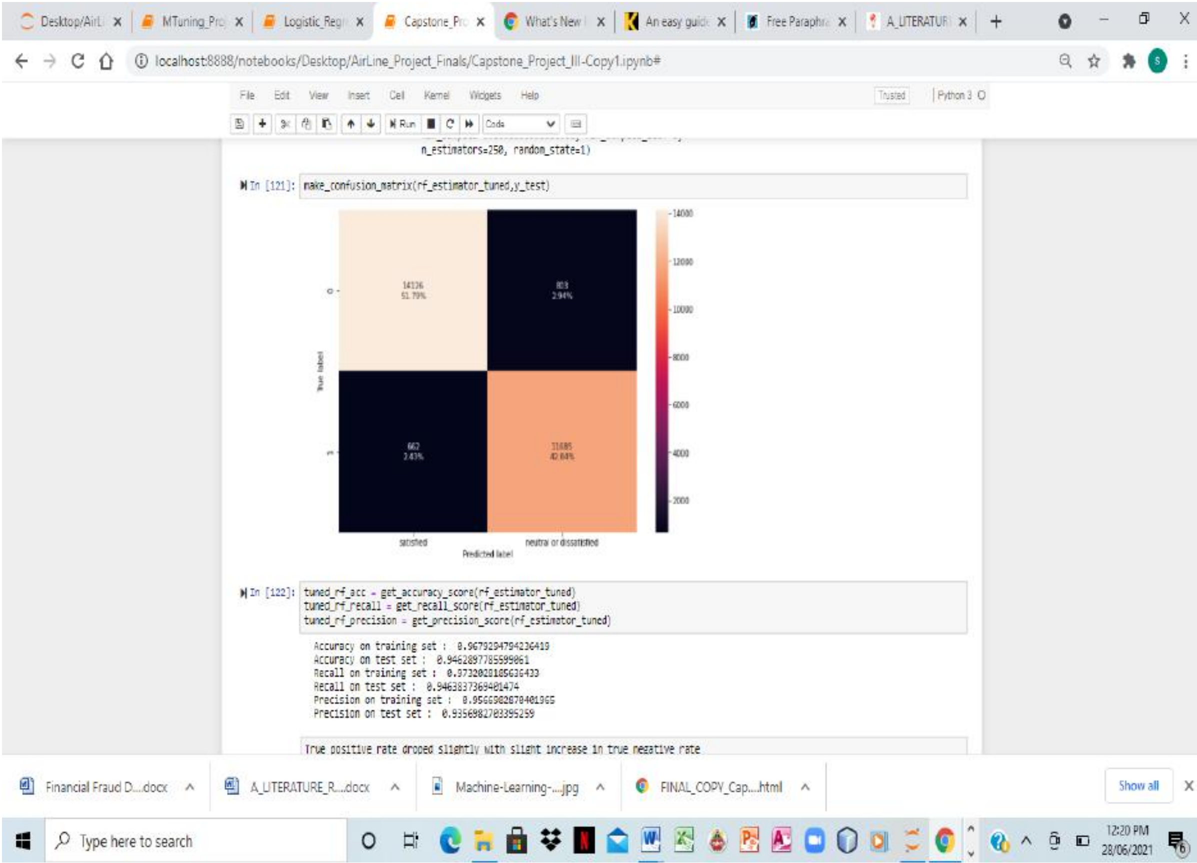
Tunned Models Compared Alongside Initial Models

The results of the comparison of the different models are as bellow.

Tunned Decision tree



TUNNED RANDOM FOREST



Model Comparison Results.

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done 1 out of 1 | elapsed: 11.9s finished
The accuracy of model LogisticRegression is 0.75

	precision	recall	f1-score	support
0	0.75	0.80	0.78	14929
1	0.74	0.68	0.71	12347
ccuracy			0.75	27276
macroavg	0.75	0.74	0.74	27276
weightedavg	0.75	0.75	0.75	27276

The accuracy of model DecisionTreeClassifier is 0.92

	precision	recall	f1-score	support
0	0.92	0.93	0.93	14929
1	0.91	0.91	0.91	12347
accuracy			0.92	27276
macroavg	0.92	0.92	0.92	27276
weightedavg	0.92	0.92	0.92	27276

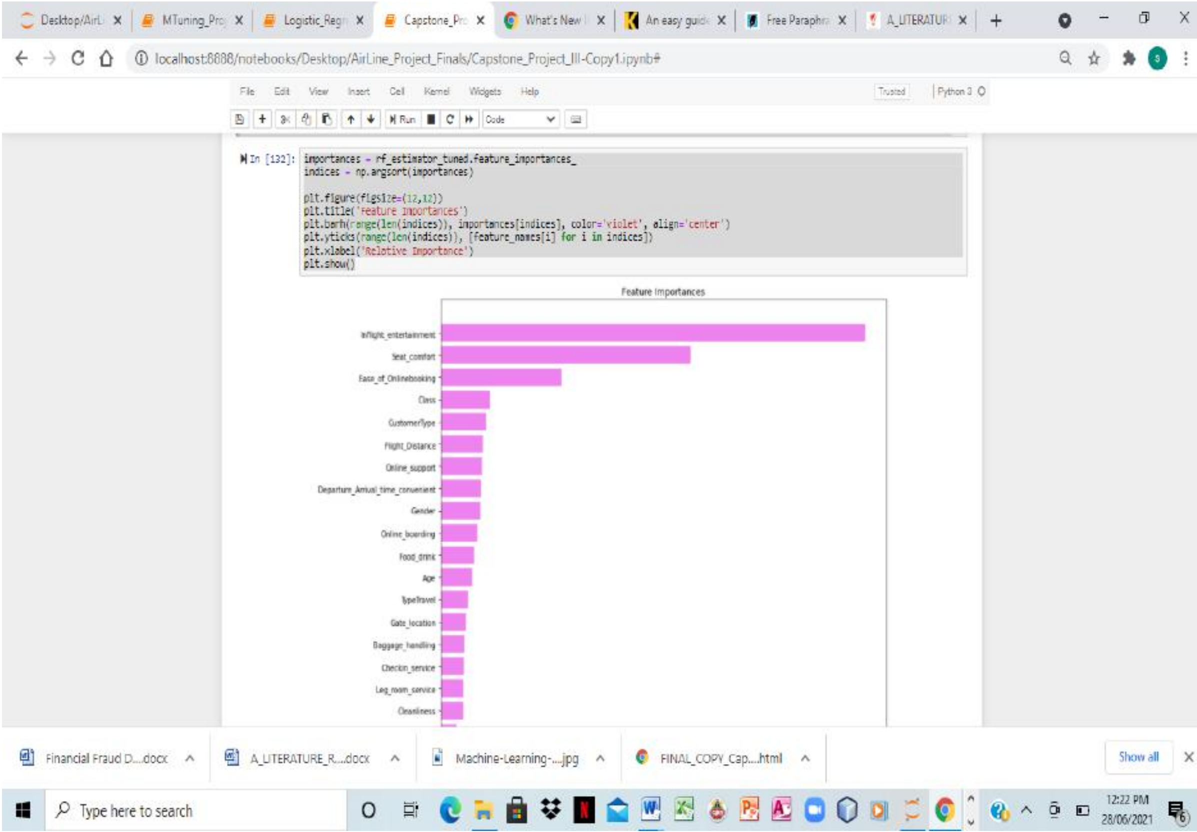
The accuracy of model DecisionTreeClassifier is 0.87

	precision	recall	f1-score	support
0	0.85	0.92	0.88	14929
1	0.89	0.80	0.84	12347
accuracy			0.87	27276
macroavg	0.87	0.86	0.86	27276
weightedavg	0.87	0.87	0.86	27276

The accuracy of model BaggingClassifier is 0.94
precision recall f1-score support

95 0.94 14929
1 0.94 0.92 0.93 12347
accuracy 0.94 27276
macroavg 0.94 0.94 0.94 27276
weightedavg 0.94 0.94 0.94 27276
The accuracy of model BaggingClassifier is 0.95
precision recall f1-score support
0 0.96 0.95 0.95 14929
1 0.94 0.95 0.95 12347
accuracy 0.95 27276
macroavg 0.95 0.95 0.95 27276
weightedavg 0.95 0.95 0.95 27276
The accuracy of model RandomForestClassifier is 0.95
precision recall f1-score support
0 0.95 0.95 0.95 14929
1 0.94 0.94 0.94 12347
accuracy 0.95 27276
macroavg 0.95 0.95 0.95 27276
weightedavg 0.95 0.95 0.95 27276
The accuracy of model RandomForestClassifier is 0.95
precision recall f1-score support
0 0.95 0.95 0.95 14929
1 0.94 0.94 0.94 12347
accuracy 0.95 27276
macroavg 0.95 0.95 0.95 27276
weightedavg 0.95 0.95 0.95 27276
The accuracy of model DecisionTreeClassifier is 0.92
precision recall f1-score support
0 0.92 0.93 0.93 14929
1 0.91 0.91 0.91 12347
accuracy 0.92 27276
macroavg 0.92 0.92 0.92 27276
weightedavg 0.92 0.92 0.92 27276
The accuracy of model RandomForestClassifier is 0.95
precision recall f1-score support
0 0.96 0.95 0.95 14929
1 0.94 0.95 0.94 12347
accuracy 0.95 27276
macroavg 0.95 0.95 0.95 27276
weightedavg 0.95 0.95 0.95 27276
The accuracy of model BaggingClassifier is 0.95
The accuracy of model RandomForestClassifier is 0.95

Randomforestclassifier Important Features



Inflight_flight entertainment is the most important variable for predicting airline passengerssatisfactionfollowed by seat_comfort, ease_of_online booking, Class, customerType and Flight_Distance.

Conclusion

A predictive classification model has been built. Given the performance in terms of Precision and Recall, the model can be deployed to identify Passengers who may not be satisfied or are indifferent to the Airline services and shall take appropriate actions to build and improve on services that drive passengers satisfaction. Factors that drive satifaction - Inflight_entertainment, Seat_comfort,Ease_of_Onlinebooking and Online_support.

Business Insights and Recommendations

I recommend that airlines should focus on improving the Inflight_entertainment experience.

In addition, airlines should also focus on Ease of Online Booking, as business passengers prioritize on ease and convenience in their travel.Finally, I hope that the model will provide a reference for airlines and be utilized for business value

.

References

https://www.researchgate.net/figure/Machine-Learning-Model-Training-Process_fig1_336020567/download retrieved 28/06/2021

Park, Y., Gates, S.: Towards real-time measurement of customer satisfaction using automat-ically generated call transcripts. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1387–296. ACM (2009)

https://en.wikipedia.org/wiki/Twitter Accessed 28 June 2021

Abel F, Gao Q, Houben G-J, Tao K (2013) Twitter-based user modeling for news recommendations. In: Rossi F (ed)

IJCAI 2013, proceedings of the 23rd international joint conference on artificial intelligence, Beijing, China, August3–9, 2013.IJCAI/AAAI.

https://www.quora.com/How-many-airplanes-fly-each-day-in-the-world retrieved 28/06/2021

Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI

International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

Leture slides and Jupyter notebook on Supervised learning and Model tuning file from Greatlearning.