Final Project Report

Airlines Review Dataset

Mark Gopani
Warren Fernandes

Introduction:
Data mining covers a variety of techniques to discover inherent but meaningful patterns from huge amount of data. Using such patterns, some models can be constructed to facilitate business processes such as decision-making, investment planning, marketing strategy development and so on.
Aviation is an ever changing, fast moving, dynamic industry that requires airlines to constantly be innovating and pre-empting events in order to remain competitive. Over the years the airlines in the industry have been challenged to satisfy their customers while facing competition from within the industry. There are several factors which give a some airlines an upper edge over others and gives them customer recognition.

Objective and Business Questions:
Airline companies in current competitive business environment can improve suitable strategies to maintain the highest level of customer satisfaction and provide high quality service. In this study, we want to answer "what factors are dominantly essential while customers are rating the services provided by company?" We investigate ratings given by customers after their flight experiences. Over given a couple of ratings for each flight, we try to understand what factors are more important in the view of customers. We group the factors and draw models for each group.
We look at questions like Which regions need to be improved for which airlines. Which factors most influence a flight to be recommended.
Review text analysis: Which words have influenced reviews to be positive or negative more. How does polarity of a review affect the recommendation for a flight.

Data Description:
Skytrax is a UK based consultancy which runs an airline and airport review and ranking site. Their mission is to improve the customer experience for airlines and airports across the world. The dataset consists of individual reviews left by verified customers of most of the major airlines across the world. The dataset is scraped from Skytrax (https://www.airlinequality.com)
An individual review contains the following 20 attributes:

1. airline name
2. link
3. title
4. author
5. author country
6. date
7. content
8. aircraft
9. type traveller
10. cabin flown
11. route
12. overall rating
13. seat comfort rating
14. cabin staff rating
15. food beverages rating
16. inflight entertainment rating
17. ground service rating
18. wi-fi connectivity rating
19. value money rating
20. recommended

The original data has more than 41k rows of individual flight reviews, data includes airline reviews from 2006 to 2019 and is scraped at Spring 2019.

The content of the dataset looks like following:

	airline_name	link	title	author	author_country	date	content	aircraft	type_traveller	cabin_flown	route	overall_rating	seat_comfort_ratin
0	adria-airways	/airline-reviews/adria-airways	Adria Airways customer review	D Ito	Germany	2015-04-10	Outbound flight FRA/PRN A319. 2 hours 10 min f...	NaN	NaN	Economy	NaN	7.0	4
1	adria-airways	/airline-reviews/adria-airways	Adria Airways customer review	Ron Kuhlmann	United States	2015-01-05	Two short hops ZRH-LJU and LJU-VIE. Very fast ...	NaN	NaN	Business Class	NaN	10.0	4
2	adria-airways	/airline-reviews/adria-airways	Adria Airways customer review	E Albin	Switzerland	2014-09-14	Flew Zurich-Ljubljana on JP365 newish CRJ900. ...	NaN	NaN	Economy	NaN	9.0	5
3	adria-airways	/airline-reviews/adria-airways	Adria Airways customer review	Tercon Bojan	Singapore	2014-09-06	Adria serves this 100 min flight from Ljubljan...	NaN	NaN	Business Class	NaN	8.0	4

cabin_staff_rating	food_beverages_rating	inflight_entertainment_rating	ground_service_rating	wifi_connectivity_rating	value_money_rating	recommended
4.0	4.0	0.0	NaN	NaN	4.0	1
5.0	4.0	1.0	NaN	NaN	5.0	1
5.0	4.0	0.0	NaN	NaN	5.0	1
4.0	3.0	1.0	NaN	NaN	4.0	1

Data pre-processing and Preparation:

Some pre-processing was performed on the data before running the model.
Dropping N/As:
Initially we see a high percentage of N/A values in some of the columns of the dataset:

Count of each attribute:	
airline_name	41396
link	41396
title	41396
author	41396
author_country	39805
date	41396
content	41396
aircraft	1278
type_traveller	2378
cabin_flown	38520
route	2341
overall_rating	36861
seat_comfort_rating	33706
cabin_staff_rating	33708
food_beverages_rating	33264
inflight_entertainment_rating	31114
ground_service_rating	2203
wifi_connectivity_rating	565
value_money_rating	39723
recommended	41396

Percentage of N/As in each attribute:	
airline_name	0.000000
link	0.000000
title	0.000000
author	0.000000
author_country	3.843367
date	0.000000
content	0.000000
aircraft	96.912745
type_traveller	94.255484
cabin_flown	6.947531
route	94.344864
overall_rating	10.955165
seat_comfort_rating	18.576674
cabin_staff_rating	18.571843
food_beverages_rating	19.644410
inflight_entertainment_rating	24.838149
ground_service_rating	94.678230
wifi_connectivity_rating	98.635134
value_money_rating	4.041453
recommended	0.000000

We drop the columns with high proportion of N/A values (>90%)along with some other irrelevant columns from the dataset. The columns dropped are: link, title, author, aircraft, type traveller, route and ground service rating.
We also drop wi-fi connectivity rating after performing analysis on it.

Generating country codes:
We introduce a new column named CODE which contains the three letter codes of the country name fetched from column ‘author country’.

	author_country	CODE
0	Germany	DEU
1	United States	USA
2	Switzerland	CHE
3	Singapore	SGP
4	Poland	POL
...	...	...
41391	United Kingdom	GBR
41392	Belgium	BEL
41393	Ireland	IRL
41394	Czech Republic	None
41395	United Kingdom	GBR

We take a look at some statistics for the columns wifi_connectivity_rating and inflight_entertainment_rating.
We create two new columns with binary values for having (1) and not having (0) wi-fi and inflight entertainment.

wifi_connectivity_rating	inflight_entertainment_rating	has_wifi	has_entertainment
NaN	0.0	0	1
NaN	1.0	0	1
NaN	0.0	0	1
NaN	1.0	0	1

We found that a very small percentage of airline reviews had wifi_connectivity_rating associated with them. Therefore we will further explore the airlines that provide wifi and inflight entertainment.

```
Percentage of reviews that feature airlines with WiFi: 1.3648661706445067 %
Percentage of reviews that feature airlines with in-flight entertainment: 75.16185138660741 %
```

```
1 df.has_wifi.value_counts()
```

```
0    40831
1      565
Name: has_wifi, dtype: int64
```

```
1 df.has_entertainment.value_counts()
```

```
1    31114
0    10282
Name: has_entertainment, dtype: int64
```

Next we take the content column which contains the review text and do some pre-processing on the text.
We store the reviews list in a new variable. We have 29170 reviews to begin with.

```
1 reviews_list = df_new['content'].copy()
2 reviews_list.shape
```

```
(29170,)
```

First we find and remove any duplicate reviews that exist in the dataset.
We find a total of 19 reviews that are duplicates in the dataset and then remove them.
n_reviews in the following figure shows the number of times a review was found repeatedly in the dataset.
We are left with 29151 reviews after removing duplicates on which we will perform some more pre-processing for sentiment analysis.

	content	n_reviews
22247	Rating : 10/10 Cabin Flown Economy Value for M...	3
346	26 April 2014 Verona-Monaco. LH 434 26 April 2...	2
16418	LAX-SFO: Simple flight made easier by a relaxe...	2
12347	I flew from Chicago O'Hare to Dublin and from ...	2
6559	Flew Delhi-Chennai-Coimbatore in December with...	2
22251	Rating : 9/10 Cabin Flown Economy Value for Mo...	2
6647	Flew FRA-MNL-FRA. First leg delayed (3 hrs) be...	2
2028	Aug 26 UA 1683 SFO EWR. On time departure and ...	2
23655	Sao Paulo - Amsterdam - London and back. Their...	2
7553	Flew Sydney - Beijing - London. Both planes we...	2
17905	London City-New York JFK via Shannon on A318 b...	2
28242	We travelled Air Canada and Air Canada Rouge t...	2
19014	Manchester to Perth business class. Experience...	2
13659	I purchased a flight with American Airline to ...	2
4795	Delayed by 10 hrs found out accidentally after...	2
22263	Reading some previous reviews and as a very ne...	2

Our next step is to take in account the reviews and assign them a polarity score based on sentiment.
We use the open source sentiment analysis package VADER which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER is built on social media text but it is in general applicable to other domains, including customer reviews. VADER is based on a lexicon (vocabulary) that is validated by multiple human judges according to a well-defined and standard procedure. Each word in the lexicon is associated with a sentiment valence, consisting of two properties, polarity and intensity. The polarity describes if the text is positive/negative. The intensity describes how much the text is positive/negative.
The output of the sentiment analysis is a series of scores, namely "compound", "pos", "neu" and "neg". The compound score is normalized between -1 (extremely negative) and 1 (extremely positive). We add the compound score of our analysis as a new column in our data frame to later include in our analysis models.

abin_flown	overall_rating	seat_comfort_rating	cabin_staff_rating	food_beverages_rating	value_money_rating	recommended	has_wifi	has_entertainment	polarity
Economy	7.0	4.0	4.0	4.0	4.0	1	0	1	0.7351
Business Class	10.0	4.0	5.0	4.0	5.0	1	0	1	0.8777
Economy	9.0	5.0	5.0	4.0	5.0	1	0	1	0.9497
Business Class	8.0	4.0	4.0	3.0	4.0	1	0	1	0.9228

Manipulating the review scores:
We will classify the overall rating into one of the 3 categories: Positive, Neutral and Negative and add it as another attribute in our data frame. We then create a two dummy variables to satisfy the new column.

recommended	has_wifi	has_entertainment	polarity	pos_neu_neg_review_score_neg	pos_neu_neg_review_score_neu	pos_neu_neg_review_score_pos
1	0	1	0.7351	0	0	1
1	0	1	0.8777	0	0	1
1	0	1	0.9497	0	0	1
1	0	1	0.9228	0	0	1

Filtering stop words:
Along with NLTK stop words, we need to take care of words which indicate the airline names therefore we create a list of all airline names in the dataset, we also create a list of words which could be linked with airline names like ‘airways’ ‘air’ ‘air lines’ etc. We also create another list of all the additional possible stop words that can be added to our list.

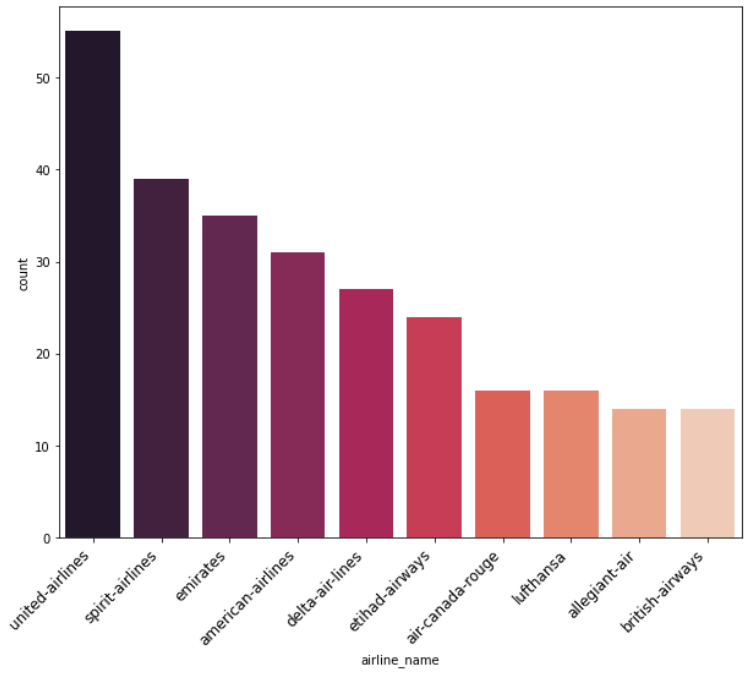
Data cleaning steps:

- Convert all characters in the review text to lower case
- Remove the punctuation and tokenize each customer review into a list of individual words.
- Remove words that contain numbers
- All stop words should are filtered out as they do not affect the meaning of the sentence.
- Remove empty tokens
- Parts Of Speech tagging the text, which allows to identify the role of each word in the sentence, according to the categories noun, verb, adjective, adverb and others. This is needed for a correct lemmatization of the words in the review text
- Lemmatize text and bring the words to their "standard" form.
- Remove words with only one letter

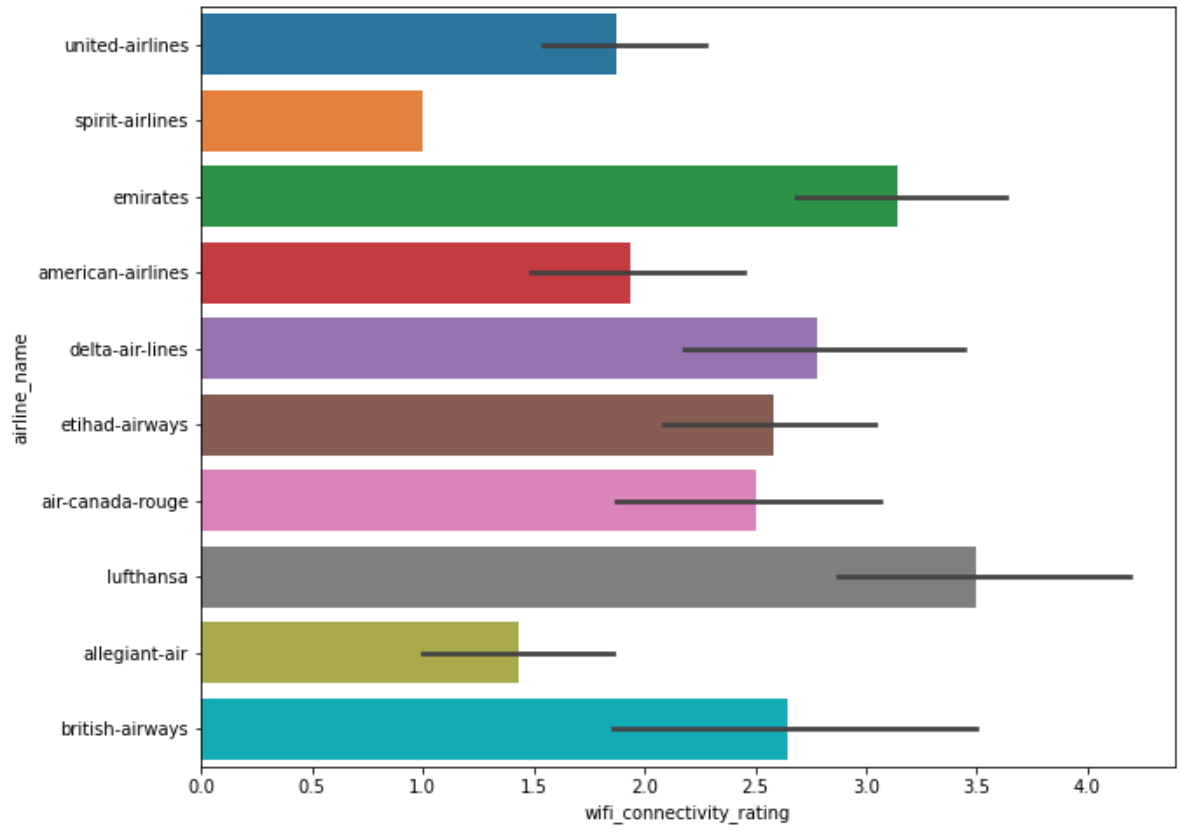
After performing the data cleaning steps on the text, we store it as a column called review_clean.

Exploratory Data Analysis:

We look at top 10 airlines which provide wi-fi connectivity:



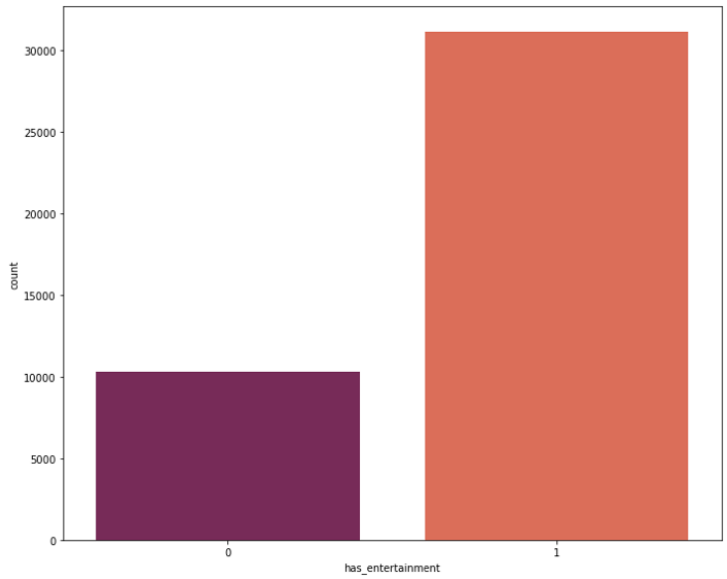
Next we look at the top 10 airlines with best wi-fi connectivity ratings:



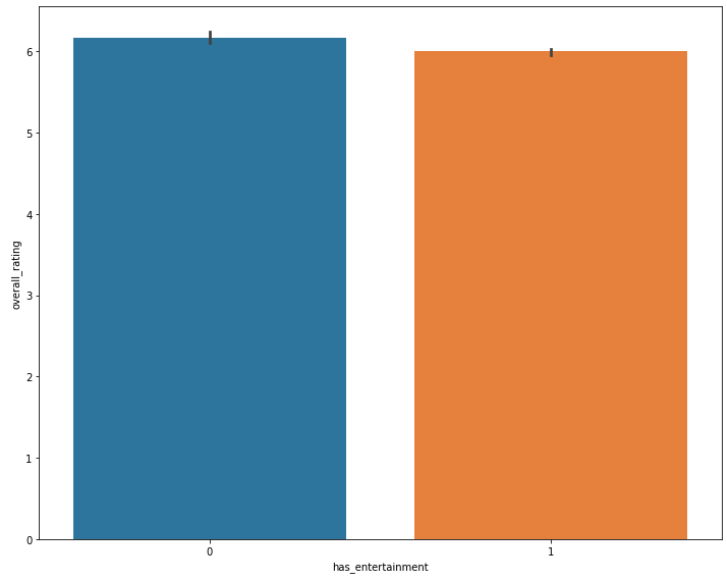
Lufthansa has the best wi-fi connectivity rating with 3.5 score out of 5 and next is Emirates with a rating just greater than 3 out of 5.

This way we can provide insights on airlines and their competitors with respect to wi-fi connectivity satisfaction of the customers.

We look at the number of flights which have inflight entertainment:

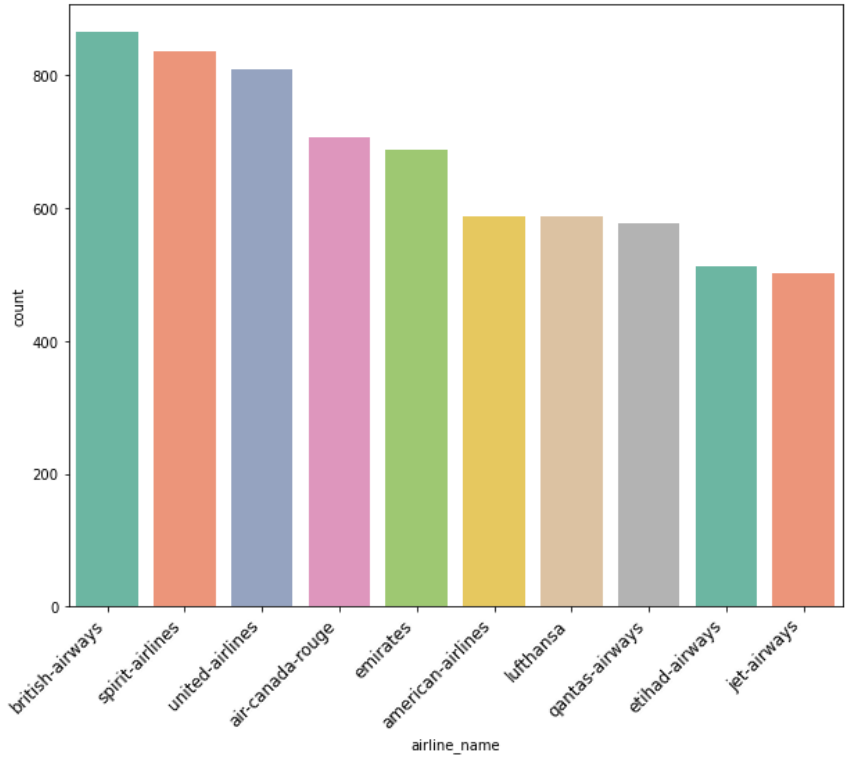


Next we look at overall rating for inflight entertainment:

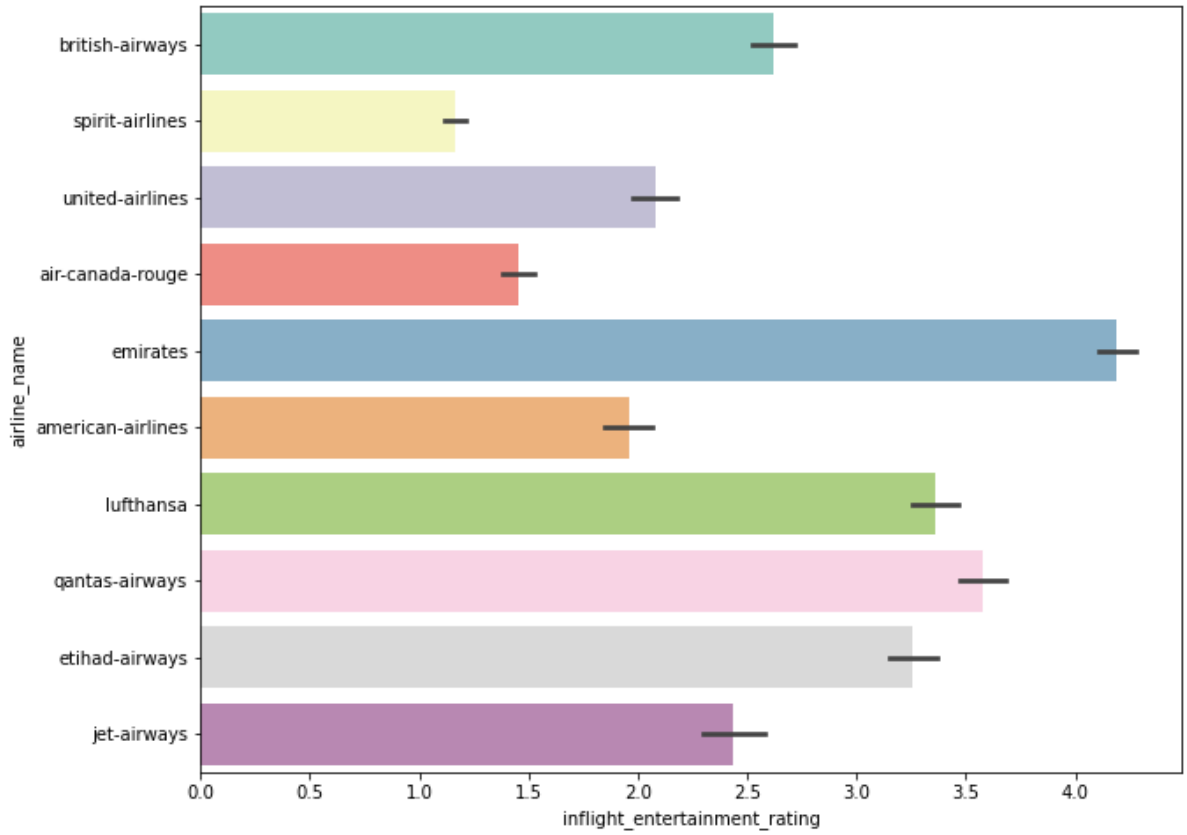


It can be seen that having in-flight entertainment does not affect overall rating of the flight by too much.

We look at top 10 airlines which provide inflight entertainment:



Next we look at top 10 airlines with best inflight entertainment ratings:



Emirates has the best in-flight entertainment rating followed by Qantas airways and then by Lufthansa.

Data analysis:

In this part, we perform Linear regression, Random Forest Classifier and Decision Tree Models:

For our initial analysis, we use the data frame with the following columns as our Independent variables for our model:

Seat comfort rating, cabin staff rating, food beverages rating, value money rating, has entertainment.

	seat_comfort_rating	cabin_staff_rating	food_beverages_rating	value_money_rating	has_entertainment
0	4.0	4.0	4.0	4.0	1
1	4.0	5.0	4.0	5.0	1
2	5.0	5.0	4.0	5.0	1
3	4.0	4.0	3.0	4.0	1
4	4.0	2.0	1.0	2.0	1

We then train and test the regression model and get the following result:

Dep. Variable:	recommended	**R-squared:**	0.659
Model:	OLS	**Adj. R-squared:**	0.659
Method:	Least Squares	**F-statistic:**	7893.
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	0.00
Time:	16:34:29	**Log-Likelihood:**	-3328.9
No. Observations:	20419	**AIC:**	6670.
Df Residuals:	20413	**BIC:**	6717.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4395	0.010	-45.552	0.000	-0.458	-0.421
seat_comfort_rating	0.0529	0.002	24.433	0.000	0.049	0.057
cabin_staff_rating	0.0913	0.002	43.441	0.000	0.087	0.095
food_beverages_rating	0.0176	0.002	10.006	0.000	0.014	0.021
value_money_rating	0.1505	0.002	65.734	0.000	0.146	0.155
has_entertainment	-0.0172	0.008	-2.110	0.035	-0.033	-0.001

Omnibus:	1311.463	**Durbin-Watson:**	1.968
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5562.802
Skew:	0.164	**Prob(JB):**	0.00
Kurtosis:	5.536	**Cond. No.**	43.3

65% of change in recommendations is explained by change in seat comfort rating, cabin staff rating, food beverages rating, value money rating and having inflight entertainment.

The decision tree model gives a score of 91.69% and the following metrics:

```
[[3076  409]
 [ 318 4948]]
              precision    recall  f1-score   support

           0       0.91      0.88      0.89      3485
           1       0.92      0.94      0.93      5266

    accuracy                           0.92      8751
   macro avg       0.91      0.91      0.91      8751
weighted avg       0.92      0.92      0.92      8751
```

while random forest gives a score of 91.02%.

```
[[3105  380]
 [ 405 4861]]
              precision    recall  f1-score   support

           0       0.88      0.89      0.89      3485
           1       0.93      0.92      0.93      5266

    accuracy                           0.91      8751
   macro avg       0.91      0.91      0.91      8751
weighted avg       0.91      0.91      0.91      8751
```

We then run the models with our added features. We added Polarity, pos_neu_neg_review_score_neg, pos_neu_neg_review_score_pos, pos_neu_neg_review_score_neu.
The training data looks like the following:

	seat_comfort_rating	cabin_staff_rating	food_beverages_rating	value_money_rating	polarity	pos_neu_neg_review_score_neg
37205	1.0	1.0	1.0	2.0	-0.4497	1
40222	1.0	1.0	2.0	2.0	-0.5611	1
30756	4.0	5.0	4.0	5.0	-0.5145	0
14273	5.0	5.0	4.0	5.0	0.8847	0
28798	5.0	5.0	5.0	5.0	0.9368	0
...	...	...	...	...	...	...
30575	2.0	4.0	2.0	1.0	-0.7143	1
7273	5.0	5.0	5.0	4.0	0.9643	0
1263	4.0	5.0	5.0	5.0	0.9957	0
22328	4.0	4.0	3.0	3.0	0.9067	0
34097	4.0	5.0	5.0	5.0	-0.6915	0

20405 rows × 8 columns

We then train and test the regression model and get the following result:

Dep. Variable:	recommended	**R-squared:**	0.823
Model:	OLS	**Adj. R-squared:**	0.823
Method:	Least Squares	**F-statistic:**	1.356e+04
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	0.00
Time:	19:35:30	**Log-Likelihood:**	3330.2
No. Observations:	20405	**AIC:**	-6644.
Df Residuals:	20397	**BIC:**	-6581.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2316	0.005	44.572	0.000	0.221	0.242
seat_comfort_rating	0.0076	0.002	4.792	0.000	0.005	0.011
cabin_staff_rating	0.0134	0.002	8.269	0.000	0.010	0.017
food_beverages_rating	0.0027	0.001	2.128	0.033	0.000	0.005
value_money_rating	0.0329	0.002	17.692	0.000	0.029	0.037
polarity	0.0472	0.003	17.093	0.000	0.042	0.053
pos_neu_neg_review_score_neg	-0.3025	0.003	-115.930	0.000	-0.308	-0.297
pos_neu_neg_review_score_neu	0.0601	0.004	16.425	0.000	0.053	0.067
pos_neu_neg_review_score_pos	0.4740	0.005	102.644	0.000	0.465	0.483

Omnibus:	2932.380	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	37014.058
Skew:	0.244	**Prob(JB):**	0.00
Kurtosis:	9.580	**Cond. No.**	1.68e+15

The R-squared has increased from 0.659 to 0.823 indicating that the added features give a better result.

The decision tree shows improved result:

```
[[3079  324]
 [ 338 5005]]
              precision    recall  f1-score   support

           0       0.90      0.90      0.90      3403
           1       0.94      0.94      0.94      5343

    accuracy                           0.92      8746
   macro avg       0.92      0.92      0.92      8746
weighted avg       0.92      0.92      0.92      8746
```

Random forest model also shows improved result:

```
[[3256  147]
 [ 350 4993]]
              precision    recall  f1-score   support

           0       0.90      0.96      0.93      3403
           1       0.97      0.93      0.95      5343

    accuracy                           0.94      8746
   macro avg       0.94      0.95      0.94      8746
weighted avg       0.94      0.94      0.94      8746
```

To better understand the factors, we change our model to predict the factors:
We predict the value money rating first, the Independent factors are

	seat_comfort_rating	cabin_staff_rating	food_beverages_rating	has_entertainment
0	4.0	4.0	4.0	1
1	4.0	5.0	4.0	1
2	5.0	5.0	4.0	1
3	4.0	4.0	3.0	1
4	4.0	2.0	1.0	1

We then train our model and use the regression model:

Dep. Variable:	value_money_rating	**R-squared:**	0.638
Model:	OLS	**Adj. R-squared:**	0.638
Method:	Least Squares	**F-statistic:**	8973.
Date:	Wed, 22 Dec 2021	**Prob (F-statistic):**	0.00
Time:	00:55:38	**Log-Likelihood:**	-26061.
No. Observations:	20405	**AIC:**	5.213e+04
Df Residuals:	20400	**BIC:**	5.217e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5904	0.029	20.091	0.000	0.533	0.648
seat_comfort_rating	0.4233	0.006	71.852	0.000	0.412	0.435
cabin_staff_rating	0.3773	0.006	64.443	0.000	0.366	0.389
food_beverages_rating	0.1336	0.005	25.149	0.000	0.123	0.144
has_entertainment	-0.2876	0.025	-11.435	0.000	-0.337	-0.238

Omnibus:	511.271	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	992.039
Skew:	-0.176	**Prob(JB):**	3.81e-216
Kurtosis:	4.021	**Cond. No.**	37.4

Next we predict the overall rating using the same independent factors, the regression model give us the following result:

Dep. Variable:	overall_rating	**R-squared:**	0.709
Model:	OLS	**Adj. R-squared:**	0.709
Method:	Least Squares	**F-statistic:**	1.244e+04
Date:	Wed, 22 Dec 2021	**Prob (F-statistic):**	0.00
Time:	00:55:40	**Log-Likelihood:**	-40247.
No. Observations:	20405	**AIC:**	8.050e+04
Df Residuals:	20400	**BIC:**	8.054e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7181	0.059	-12.193	0.000	-0.834	-0.603
seat_comfort_rating	0.8442	0.012	71.494	0.000	0.821	0.867
cabin_staff_rating	1.0337	0.012	88.092	0.000	1.011	1.057
food_beverages_rating	0.3089	0.011	29.012	0.000	0.288	0.330
has_entertainment	-0.5383	0.050	-10.679	0.000	-0.637	-0.440

Omnibus:	2516.069	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21305.797
Skew:	0.296	**Prob(JB):**	0.00
Kurtosis:	7.971	**Cond. No.**	37.4

We see from our models that the model predicting value money rating has a positive constant coefficient while the model predicting the overall rating has a negative constant coefficient

Challenges:

A challenge we faced in the project was in analysing the review text, the implementation of NLP with the pre-processing steps on the reviews was a tedious task.

Another challenge we faced with review text is the implementation of tokenization.

We faced a problem with Installing geopanads and Descartes libraries in the Jupyter platform.

We could not implement topic modelling in our project which would help us better analyse the review text with categories.

Conclusion:

In this project we have developed and evaluated different models to get a better understanding of customer satisfaction with different airlines involved. We helped predicting the overall rating of the airlines. Our prediction model for value money rating helped us to identify whether a flight was worth it or not. Text mining helped us find keywords which would highly influence the review text in being positive or negative. Further refinement is possible by examining other variables and their relationships.