CCT College Dublin
Assessment Cover Page

Module Title:	Machine Learning
Assessment Title:	What factors most impact short-haul dissatisfaction?
Lecturer Name:	Dr Muhammad Iqbal
Student Full Name:	Group: Ana Isabel Nieves Barcenas Bárbara Azevedo Pereira Daniela Daia Vicente Rubio
Student Number:	2022455 - Ana Isabel Nieves Barcenas 2022310 - Bárbara Azevedo Pereira 2017207 - Daniela Daia 2022178 - Vicente Rubio
Assessment Due Date:	27 November 2022
Date of Submission:	27Th November 2022

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

What factors most impact short-haul dissatisfaction?

by,

Ana Isabel Nieves Barcenas

Bárbara Azevedo Pereira

Daniela Daia

Vicente Rubio

Higher Diploma in Science in Data Analytics for Business Strategic Thinking

Dr Muhammad Iqbal

CCT College Dublin, Ireland.

cct College Dublin
Computing • IT • Business

Abstract

This analysis is based on a classification data set research of over 120,000 airline passengers' satisfaction. It will study what factors are highly correlated to dissatisfaction with short-haul passengers.

It will describe the motivation of the chosen data, a description of the business problem and an explanation of the project goal. It will present the characterisation of the data by applying Exploratory Data Analyses (EDA), filling in the missing values, observing the outliers by plotting boxplots, and using the feature selection model to extract the influencing factors of passenger dissatisfaction by using machine learning accuracy. Afterwards, cross-validation techniques will be applied by using machine learning approaches, hyperparameters and a comparison between the chosen model. And finally, the interpretation and explanation of the results obtained based on different classification models.

Keywords: a*irlines, passenger satisfaction, machine learning models, CRISP-DM.*

Table of Contents

Abstract 3
Table of Contents 4
Table of Figures 5
Introduction 6
Business Understanding 7
Data Understanding 8
Data Preparation 17
Modelling 21
Deployment 28
Extra Contents 30
Roles and responsibilities 30
Team Project management 31
Team's challenges faced 32
Reference List 34

Table of Figures

Figure 1. CRISP-DM Methodology. 6
Figure 2. Data dictionary 8
Figure 3. Required libraries 9
Figure 4. Head and shape of the dataset. 9
Figure 5. The function .info() 10
Figure 6. Number summary (describe). 10
Figure 7. Pairplot of the dataset numerical variables 11
Figure 8. Histogram of the dataset numerical variables 12
Figure 9. Categorical variables summary statistics 12
Figure 10. Bar chart of the dataset categorical variables 13
Figure 11. Skewness distribution 13
Figure 12. Satisfaction level bar chart 14
Figure 13. Outliers boxplot 15
Figure 14. Heatmap 15
Figure 15. Finding missing values 17
Figure 16. missing values graph. 17
Figure 17. Short-distance flights 18
Figure 18. Dropping columns 18
Figure 19. Encoded data set 19
Figure 20. Sparse data check-up 19
Figure 21. Separating independent from dependent variables 20
Figure 22. Accuracy Scores with different splits 21
Figure 23 Machine learning models comparison 22
Figure 24.Random Forest Confusion Matrix 22
Figure 25.Logistic Regression Confusion Matrix 23
Figure 26. Logistic Regression Classification Report 25
Figure 27. KNN Classification Report 25
Figure 28. Random Forest Classification Report 25
Figure 29 Cross-Validation Logistic Regression 26
Figure 30.Cross Validation Random Forest 26
Figure 31. Cross Validation KNN 27
Figure 32. Hyperparameters with Cross-validation 27
Figure 33. Factors that impact the satisfaction 28
Figure 34. Plot of the factors that impact satisfaction 28
Figure 35. Roles and Responsibilities 30
Figure 36. Project Management: Trello Board 31
Figure 37.Team's effort 32

Introduction

Whether travelling to nearby destinations or even the farthest corners of the world, the keywords in choosing airlines are to travel quickly, comfortably, and safely. Airlines are constantly offering new destinations, and over time more companies have entered the market, offering passengers choices, competitiveness, and affordability.

In a highly competitive environment, the aviation industry stands to develop from a transport role to a service. Improving service quality is essential for competitiveness and ensures sustainable and healthy development. Therefore, airlines should promptly investigate passenger satisfaction and overall satisfaction with various services to understand the quality of existing services.

This report will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM provides a complete model for a Data management project. The project is divided into six phases: Business understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Shearer, 2000). Code available at *GitHub*. The life cycle is shown in Figure 1.

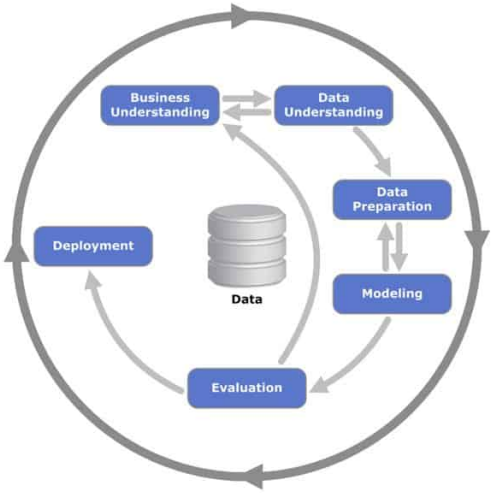

Figure 1. CRISP-DM Methodology.

Business Understanding

It is known that the airline industry is one of the fastest transportation sectors in the world; in that regard, Bart (2000) argues that traces of the strategic developments and the strategic responses of the airline players have had a profound impact on the shape and direction of the industry. These include the deregulation of the sector, the nature and extent of competition, the emergence of brand/differentiation-based competition, and airline alliance developments, strategies and their implications (Williams & Naumann, 2011).

Based on the above challenges, this study focuses on the full-service passenger information and satisfaction survey results. This analysis aims to evaluate different machine learning algorithms and determine the most suitable algorithm for classifying customer short-haul flight dissatisfaction. This analysis also aims to ascertain and highlight the most critical variables in determining customer dissatisfaction for a better insight into the issues. Finally, this study is a reference for airlines to use customer evaluation-driven service methodologies to improve their services and competitiveness.

Data Understanding

According to Han, Kamber and Pei (2011), data characterisation is a summarisation of the variables and factors of a target course of information, such as simple data summaries based on statistical measures and plots and other strategies. For this project, the dataset chosen is Airline Passenger Satisfaction, available on Kaggle, and whose information is related to customer satisfaction scores from over 120,000 airline passengers, including additional information about each passenger, their flight, and type of travel, as well as their evaluation of different factors like cleanliness, comfort, service, and overall experience.

A data dictionary is used to catalogue and communicate the structure and content of data and provides meaningful descriptions for individually named data objects (Wertz, 1993).

- **ID:** Unique passenger identifier
- **Gender:** Gender of the passenger (Female/Male)
- **Age:** Age of the passenger
- **Customer Type:** Type of airline customer (First-time/Returning)
- **Type of Travel:** Purpose of the flight (Business/Personal)
- **Class:** Travel class in the airplane for the passenger seat
- **Flight Distance:** Flight distance in miles
- **Departure Delay:** Flight departure delay in minutes
- **Arrival Delay:** Flight arrival delay in minutes
- **Departure and Arrival Time Convenience:** Satisfaction level with the convenience of the flight departure and arrival times from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Ease of Online Booking:** Satisfaction level with the online booking experience from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Check-in Service:** Satisfaction level with the check-in service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Online Boarding:** Satisfaction level with the online boarding experience from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Gate Location:** Satisfaction level with the gate location in the airport from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **On-board Service:** Satisfaction level with the on-boarding service in the airport from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Seat Comfort:** Satisfaction level with the comfort of the airplane seat from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Leg Room Service:** Satisfaction level with the leg room of the airplane seat from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Cleanliness:** Satisfaction level with the cleanliness of the airplane from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Food and Drink:** Satisfaction level with the food and drinks on the airplane from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **In-flight Service:** Satisfaction level with the in-flight service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **In-flight Wifi Service:** Satisfaction level with the in-flight Wifi service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **In-flight Entertainment:** Satisfaction level with the in-flight entertainment from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Baggage Handling:** Satisfaction level with the baggage handling from the airline from 1 (lowest) to 5 (highest) - 0 means "not applicable"
- **Satisfaction:** Overall satisfaction level with the airline (Satisfied/Neutral or unsatisfied)

Figure 2. Data dictionary

We import the required packages such as Pandas, Seaborn, Numpy, Matplotlib, Math, Missingno, Sklearn and a warning filter initial installation, allowing us to run all the analyses in our Colab notebook code. With Colab, we can import an image dataset, train an image classifier, and evaluate the model (Google Colaboratory, 2022).

In [1]:
```
#import required libraries
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import math

#Import for missing values
import missingno as msno

# Confusion Matrix
from sklearn.metrics import confusion_matrix

#Hyperparameters
from sklearn.model_selection import GridSearchCV

#Standardization
from sklearn.preprocessing import StandardScaler

# Import train_test_split function
from sklearn.model_selection import train_test_split

# Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics

#import models from scikit learn module:
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression

#Cross validation
from sklearn.model_selection import cross_val_score

# Import this library to suppress the warnings
import warnings
warnings.filterwarnings("ignore")
```

Figure 3. Required libraries

In this session, it will be imported the raw data set to find relevant information from this data by identifying the Predictor (Input) and Target (output) variables (Ray, 2019). After loading, by using the "head" function, we can see the five rows of the data set and with the ".shape" function, it is possible to see that the data size consists of 129880 rows and 24 columns (The pandas development team, 2020).

In [2]:
```
df_airline= pd.read_csv("airline_passenger_satisfaction.csv") ## Importing the dataset
df_airline.head() # looking at the dataset
```

Out[2]:

	ID	Gender	Age	Customer Type	Type of Travel	Class	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Convenience	...	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Food and Drink	In-flight Service	Se
0	1	Male	48	First-time	Business	Business	821	2	5.0	3	...	3	5	2	5	5	5	
1	2	Female	35	Returning	Business	Business	821	26	39.0	2	...	5	4	5	5	3	5	
2	3	Male	41	Returning	Business	Business	853	0	0.0	4	...	3	5	3	5	5	3	
3	4	Male	50	Returning	Business	Business	1905	0	0.0	2	...	5	5	5	4	4	5	
4	5	Female	49	Returning	Business	Business	3470	0	1.0	3	...	3	4	4	5	4	3	

5 rows × 24 columns

In [3]:
```
df_airline.shape #Looking the shape of the dataset
```

Out[3]:
```
(129880, 24)
```

Figure 4. Head and shape of the dataset.

The function info() shows more details about the dataset, such as shape, type of variables and memory used:

```
In [4]: df_airline.info() #get some information about our DataSet

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129880 entries, 0 to 129879
Data columns (total 24 columns):
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   ID                                        129880 non-null  int64
 1   Gender                                    129880 non-null  object
 2   Age                                       129880 non-null  int64
 3   Customer Type                             129880 non-null  object
 4   Type of Travel                            129880 non-null  object
 5   Class                                     129880 non-null  object
 6   Flight Distance                           129880 non-null  int64
 7   Departure Delay                           129880 non-null  int64
 8   Arrival Delay                             129487 non-null  float64
 9   Departure and Arrival Time Convenience    129880 non-null  int64
 10  Ease of Online Booking                    129880 non-null  int64
 11  Check-in Service                          129880 non-null  int64
 12  Online Boarding                           129880 non-null  int64
 13  Gate Location                             129880 non-null  int64
 14  On-board Service                          129880 non-null  int64
 15  Seat Comfort                              129880 non-null  int64
 16  Leg Room Service                          129880 non-null  int64
 17  Cleanliness                               129880 non-null  int64
 18  Food and Drink                            129880 non-null  int64
 19  In-flight Service                         129880 non-null  int64
 20  In-flight Wifi Service                    129880 non-null  int64
 21  In-flight Entertainment                   129880 non-null  int64
 22  Baggage Handling                          129880 non-null  int64
 23  Satisfaction                              129880 non-null  object
dtypes: float64(1), int64(18), object(5)
memory usage: 23.8+ MB
```

Figure 5. The function .info()

The next step consisted of obtaining the summary statistics for the numerical values in the data frame using the ".describe" function, which is responsible for generating descriptive statistics that summarise the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values (McKinney, 2017).

Descriptive Statistics

```
In [5]: df_airline.describe()
```

Out[5]:

	ID	Age	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Convenience	Ease of Online Booking	Check-in Service	Online Boarding	Gate Loca
count	129880.000000	129880.000000	129880.000000	129880.000000	129487.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000
mean	64940.500000	39.427957	1190.316392	14.713713	15.091129	3.057599	2.756876	3.306267	3.252633	2.976
std	37493.270818	15.119360	997.452477	38.071126	38.465650	1.526741	1.401740	1.266185	1.350719	1.278
min	1.000000	7.000000	31.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	32470.750000	27.000000	414.000000	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000	2.000
50%	64940.500000	40.000000	844.000000	0.000000	0.000000	3.000000	3.000000	3.000000	3.000000	3.000
75%	97410.250000	51.000000	1744.000000	12.000000	13.000000	4.000000	4.000000	4.000000	4.000000	4.000
max	129880.000000	85.000000	4983.000000	1592.000000	1584.000000	5.000000	5.000000	5.000000	5.000000	5.000

Figure 6. Number summary (describe).

The measures of dispersion evaluate how distributed the collected data are. They are standard deviation, variation and interquartile range (The pandas development team, 2020).

As it is known, a data set is very spread out when the standard deviation value is high. In this case, the standard deviation result is lower than the mean and thus is asymmetrical distribution.

The plots above show the distribution of each variable in the dataset:

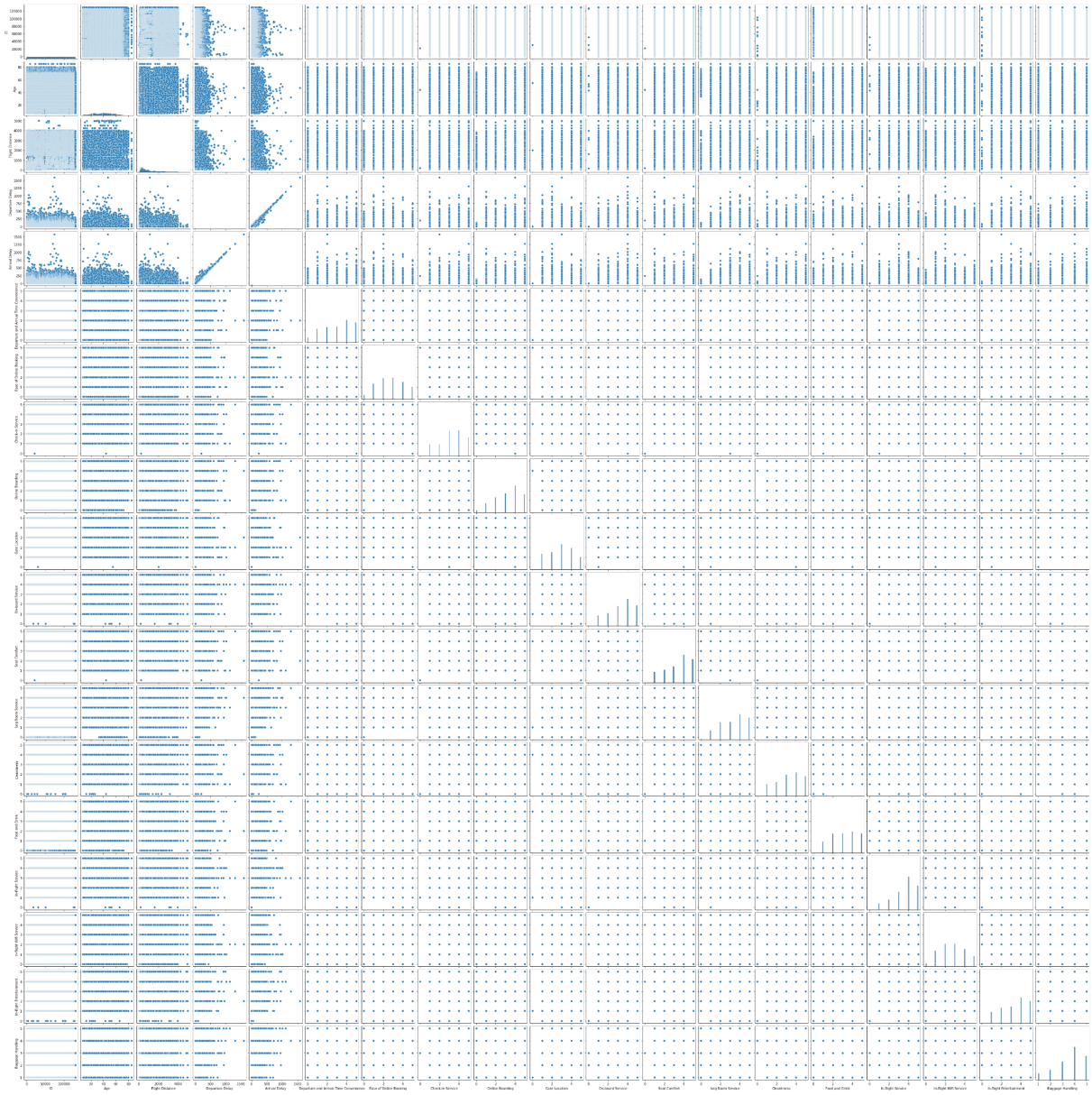

Figure 7. Pairplot of the dataset numerical variables

Figure 8. Histogram of the dataset numerical variables

Below, it can be seen the categorical summary statistics with their unique values and bar chart plot:

```
In [8]: df_airline.describe(include=object)
```

Out[8]:

	Gender	Customer Type	Type of Travel	Class	Satisfaction
count	129880	129880	129880	129880	129880
unique	2	2	2	3	2
top	Female	Returning	Business	Business	Neutral or Dissatisfied
freq	65899	106100	89693	62160	73452

Figure 9. Categorical variables summary statistics

From the bar chart below, it can be seen that most of the customer type is not travelling for the first time. They mostly travel for business and are neutral or dissatisfied regarding their overall service experience.

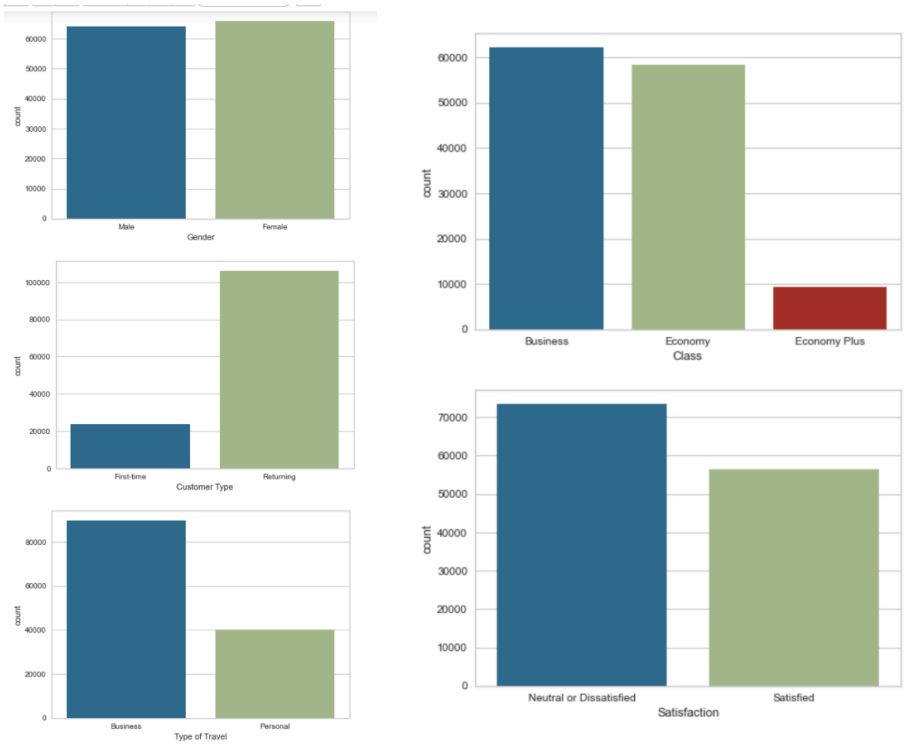

Figure 10. Bar chart of the dataset categorical variables

The function below shows the skewed distribution for the numerical variables, which is noted to likely have a negative skew.

```
In [9]: #ColObj= ['Gender', 'Customer Type','Type of Travel','Class','Satisfaction'  ]
        #for col in ColObj:
            #plt.figure(figsize=(7,5))
            #sns.countplot(x=col,data=df_airline)
            #plt.show()

In [10]: df_airline.skew()

Out[10]: ID                                        0.000000
         Age                                      -0.003606
         Flight Distance                           1.108142
         Departure Delay                           6.821980
         Arrival Delay                             6.670125
         Departure and Arrival Time Convenience   -0.332469
         Ease of Online Booking                   -0.018779
         Check-in Service                         -0.366569
         Online Boarding                          -0.456911
         Gate Location                            -0.058265
         On-board Service                         -0.421320
         Seat Comfort                             -0.485818
         Leg Room Service                         -0.348414
         Cleanliness                              -0.300926
         Food and Drink                           -0.155063
         In-flight Service                        -0.691580
         In-flight Wifi Service                    0.040465
         In-flight Entertainment                  -0.366385
         Baggage Handling                         -0.677400
         dtype: float64
```

Figure 11. Skewness distribution

The plot below shows the distribution of the variables according to their level of satisfaction classified by "Neutral" or "Dissatisfied: in blue colour and "Satisfied': in green colour:

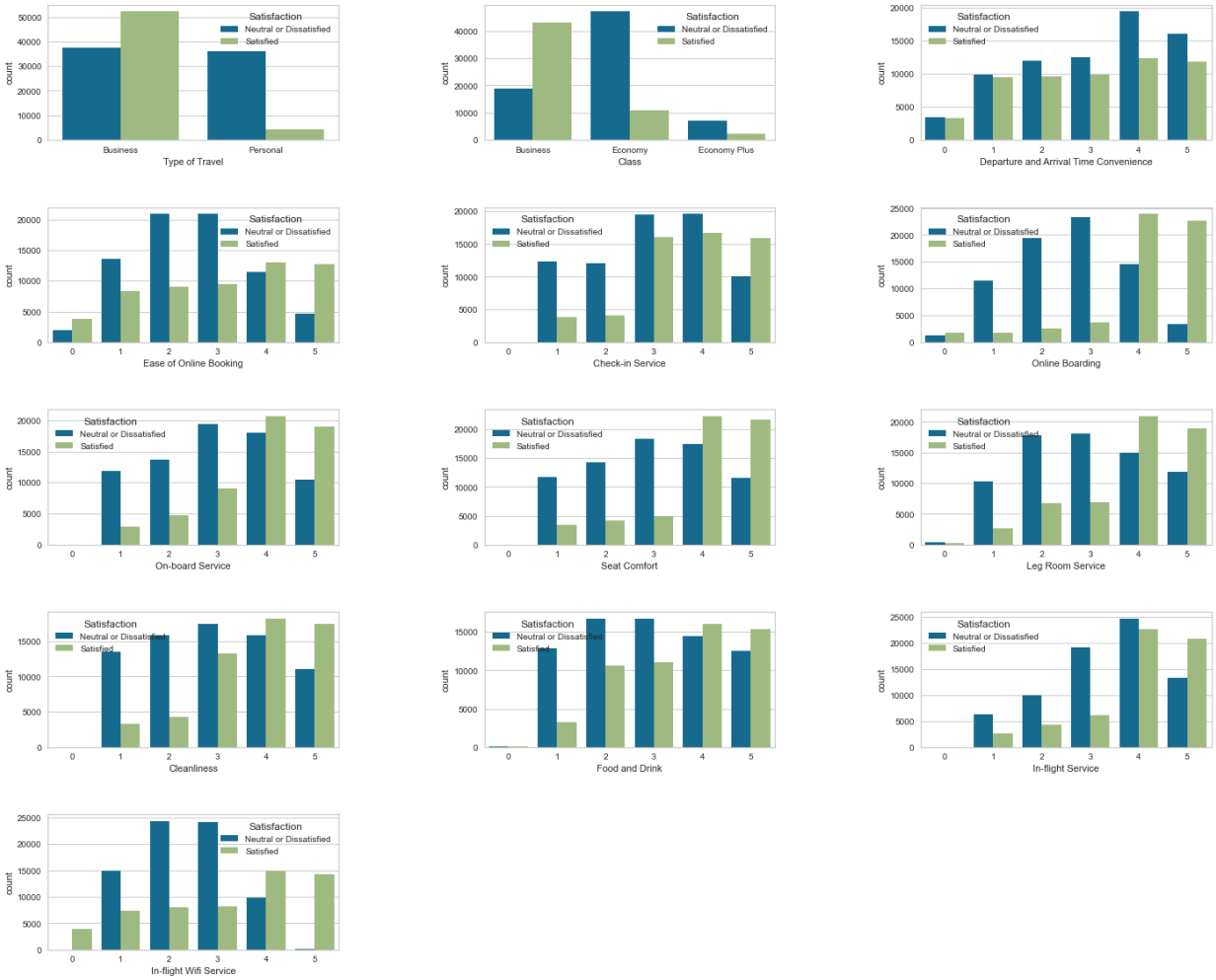

Figure 12. Satisfaction level bar chart

Here it can be seen that passengers travelling for Business propose are satisfied with the services provided and for Personal propose passengers are dissatisfied. Business class passengers are satisfied, and Economy class are not. Passengers, in general, would like to be satisfied with the arrival and departure times, the online booking system, check-in services and so on.

According to the boxplot below, it can be seen the presence of outliers in 3 variables. "Flight Distance" has short flights from 31 to long flights of 4983 nautical miles. The same happens with "Departure Delay" and "Arrival Delay": some flights are delayed by only a few minutes, while in some particular cases, it can be more than 24 hours, while some are not delayed at all.

Briefly, outliers are extreme values that are significantly from the overall pattern of values in a dataset (Ramalho, 2015). It is an observation that lies an abnormal distance from other values in a random sample from a population, and it will be discussed further about them.

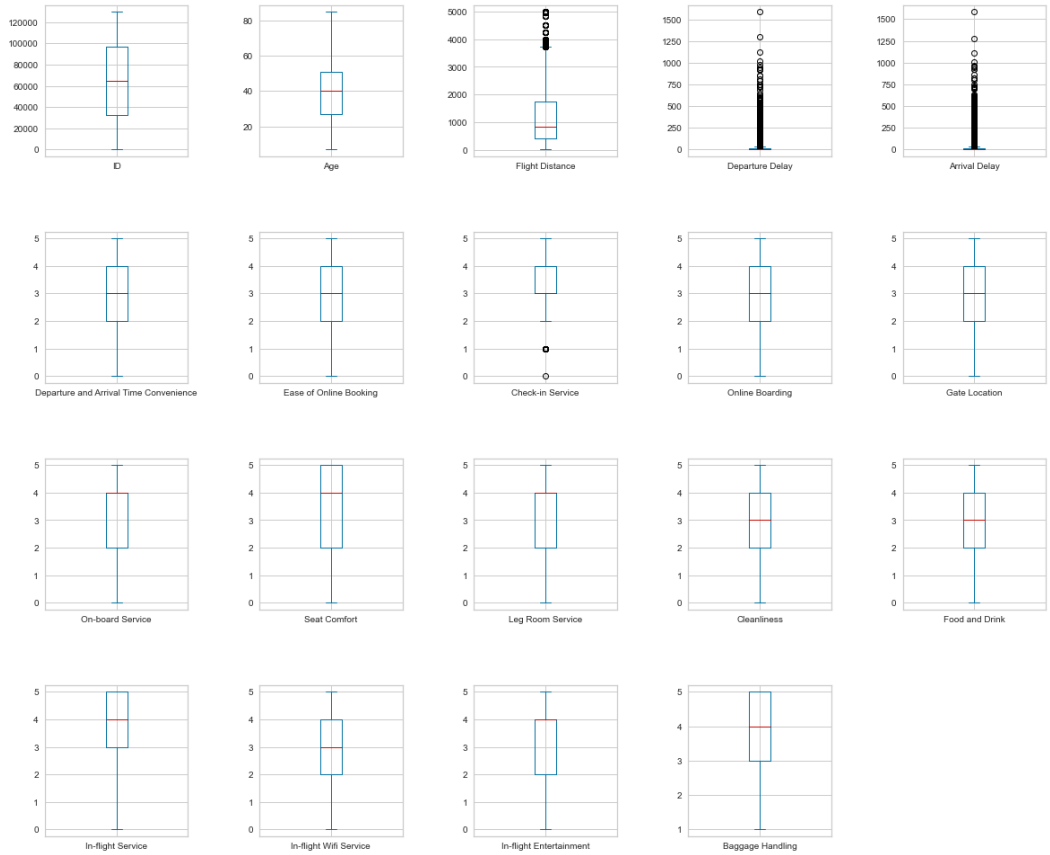

Figure 13. Outliers boxplot

Now, the correlation between the variables for this dataset will be analysed:

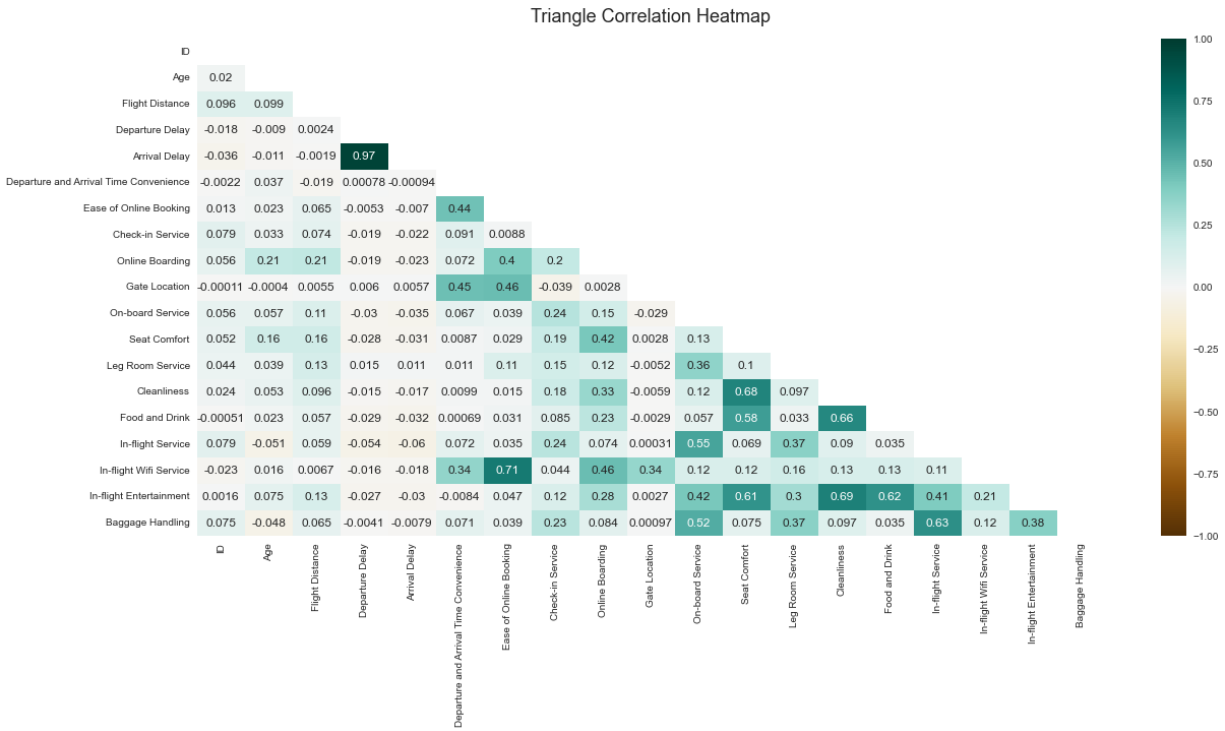

Figure 14. Heatmap

As seen in the heatmap above, there is a strong correlation between the "Arrival Delay" variable and the "Departure Delay "one. Between "In-flight-service" with the "Easy of Online Booking" and also with the "Cleanliness" and the "Seat Comfort".

Now that all the data has been presented and analysed, it is time to prepare the data for further training and application of the Machine Learning models to find the principal factors of passenger dissatisfaction with short-haul flights.

Data Preparation

Data preparation is an essential step that consists of cleaning, constructing, integrating and formatting the data set. And the first step is to identify the missing values further and analyse the best practice to deal with them (Little and Rubin, 2019).

After applying the function below, the number of 393 missing values in the "Arrival Delay" variable can be seen. It is insignificant compared to the number of data contained in our dataset. In that sense, the team decided to impute them with the median because, according to our EDA, it does not follow a normal distribution. Instead, it tends to have a positive skew (6.670125). We can also see outliers that could considerably affect our mean.

```
In [14]: df_airline.isnull().sum() #check how many values are missing (NaN)
Out[14]: ID                                        0
         Gender                                    0
         Age                                       0
         Customer Type                             0
         Type of Travel                            0
         Class                                     0
         Flight Distance                           0
         Departure Delay                           0
         Arrival Delay                           393
         Departure and Arrival Time Convenience    0
         Ease of Online Booking                    0
         Check-in Service                          0
         Online Boarding                           0
         Gate Location                             0
         On-board Service                          0
         Seat Comfort                              0
         Leg Room Service                          0
         Cleanliness                               0
         Food and Drink                            0
         In-flight Service                         0
         In-flight Wifi Service                    0
         In-flight Entertainment                   0
         Baggage Handling                          0
         Satisfaction                              0
         dtype: int64
```

Figure 15. Finding missing values

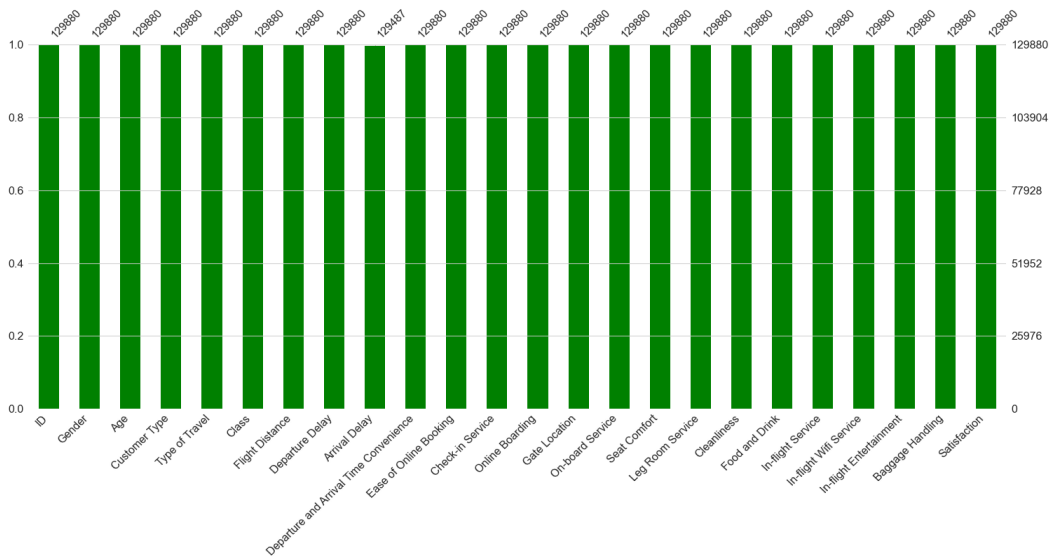

Figure 16. missing values graph.

The next step is Feature engineering. According to our scenario, we want to know which factors impact the dissatisfaction of short-distance flight passengers; that is why we need to convert our categorical values to numerical values to classify the flights according to their distance.

For this, we will rely on the aviation rules that classify them into short, medium and long distance. We will focus on short-haul flights. A short-distance trip is known to have a space of less than or equal to 800 nautical miles (InsureMyTrip,2021). After running the code below, our data set has a length of 46495 rows (before, it had 129880, as previously seen in the Data Understanding session).

```
In [18]: df_airline=df_airline[df_airline["Flight Distance"]<800]

In [19]: len (df_airline)
Out[19]: 46495

In [20]: df_airline.head()
```

Out[20]:

	ID	Gender	Age	Customer Type	Type of Travel	Class	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Convenience	...	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Food and Drink	In-flight Service
11	12	Female	27	First-time	Business	Business	421	20	21.0	2	...	2	2	5	1	1	3
12	13	Male	24	First-time	Business	Economy	453	16	30.0	2	...	2	5	4	5	5	4
30	31	Male	35	First-time	Business	Business	212	0	0.0	2	...	4	2	5	2	2	4
31	32	Male	21	First-time	Business	Economy	173	0	0.0	0	...	5	3	5	3	3	4
32	33	Female	33	First-time	Business	Business	173	22	28.0	2	...	3	2	3	2	2	5

5 rows × 24 columns

Figure 17. Short-distance flights

After that, we will drop the columns that have no relevance to this analysis: 'ID', "Gender", and "Age":

```
In [10]: df_airline=df_airline.drop(['ID',"Gender","Age"],axis=1)

In [11]: df_airline.head()
```

Out[11]:

	Customer Type	Type of Travel	Class	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Convenience	Ease of Online Booking	Check-in Service	Online Boarding	...	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Foo an Drin
0	First-time	Business	Business	821	2	5.0	3	3	4	3	...	3	5	2	5	
1	Returning	Business	Business	821	26	39.0	2	2	3	5	...	5	4	5	5	
2	Returning	Business	Business	853	0	0.0	4	4	4	5	...	3	5	3	5	
3	Returning	Business	Business	1905	0	0.0	2	2	3	4	...	5	5	5	4	
4	Returning	Business	Business	3470	0	1.0	3	3	3	5	...	3	4	4	5	

Figure 18. Dropping columns

Now is the time to encode our data set. In general, encoding is converting data from one form to another. That means if data contain a categorical variable, it just has to encode it to the numbers before fitting the data into the model (Brownlee, 2020).

- Class label: "**Satisfaction":** Neutral or Dissatisfied :0, Satisfied: 1
- **The "Customer Type"** variable is as follows: First-time :0, Returning: 1
- **The "Type of Travel"** variable is as follows: Business: 0, Personal: 1
- **"Class"** variable as Business: 0, Economy: 1, Economy Plus: 2

Now the data set is encoded. In other words, all the variables contain numerical numbers; from now, machine models can be applied.

```
In [28]: df_airline
```

Out[28]:

	Customer Type	Type of Travel	Class	Flight Distance	Departure Delay	Arrival Delay	Departure and Arrival Time Convenience	Ease of Online Booking	Check-in Service	Online Boarding	...	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Food and Drink	S
11	0	0	0	421	20	21.0	2	2	1	2	...	2	2	5	1	1	
12	0	0	1	453	16	30.0	2	2	2	2	...	2	5	4	5	5	
30	0	0	0	212	0	0.0	2	2	5	2	...	4	2	5	2	2	
31	0	0	1	173	0	0.0	0	4	3	4	...	5	3	5	3	3	
32	0	0	0	173	22	28.0	2	2	5	2	...	3	2	3	2	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
95451	1	0	0	189	5	0.0	4	4	3	4	...	4	4	4	5	3	
95454	1	0	0	239	0	0.0	5	5	4	5	...	5	4	5	5	3	
95458	1	0	2	187	1	2.0	3	3	1	2	...	1	3	1	3	3	
95461	1	0	0	677	0	0.0	2	5	3	4	...	4	4	4	5	4	
95462	1	0	0	239	0	0.0	5	5	3	5	...	4	5	4	5	2	

46495 rows × 21 columns

Figure 19. Encoded data set

But before, we will check if the data set has sparse data. According to Brownlee (2020), Sparse is a dataset with high zero values that can cause problems like over-fitting in the machine learning models and several other issues. However, our data set is not sparse, as seen below:

```
In [29]: from scipy import sparse

In [30]: sparse.issparse(df_airline) # how sparse is the dataset
Out[30]: False
```

Figure 20. Sparse data check-up

Now is the time to split the dataset into independent and dependent ("Satisfaction") variables to further train our models and find our answers.

```
In [31]:
X= df_airline.iloc[:, 0:20].values

In [32]: X
Out[32]: array([[0., 0., 0., ..., 1., 1., 4.],
                [0., 0., 1., ..., 2., 5., 4.],
                [0., 0., 0., ..., 2., 2., 5.],
                ...,
                [1., 0., 2., ..., 2., 3., 4.],
                [1., 0., 0., ..., 2., 4., 4.],
                [1., 0., 0., ..., 5., 4., 4.]])

In [33]: y= df_airline.iloc[:,20:21 ].values
         y
Out[33]: array([[0],
                [0],
                [0],
                ...,
                [0],
                [1],
                [1]])
```

Figure 21. Separating independent from dependent variables

Modelling

To have a reliable training process without failures or loss of quality, the DummyClassifier model was applied as a baseline. This baseline model was used to contextualise the results of the training models and improve understanding of the data. Although this model has low predictive power, it is considered adequate and may need to act as a guide to compare with other complex models in the ML project (Figure 22).

Raul Garreta et al. (2017) agree that testing complex models against simple models are important, and dummy estimators provide this. The usefulness of this model can be easily illustrated with a fraud example when 5% can be considered a warning. More complex models can lead to misleading accuracy without detecting the fraud at hand, and dummy models filter this kind of situation better. Before applying machine learning models, it is needed to split the data set into different percentages for training and testing to check each of its accuracies: 20%, 15%, 10% and 5%:

Split 1: With 20%, we got an accuracy of: 0.6607162060436607

Split 2: With 15%, we got an accuracy of:0.6607162060436607

Split 3: With 10%, we got an accuracy of: 0.6606451612903226

Split 4: With 5%, we got an accuracy of: 0.6606451612903226

As can be seen in the graph, It does not have a significant difference using different percentages for the test, which is why it was decided to use 20% for the test and 80% for training:

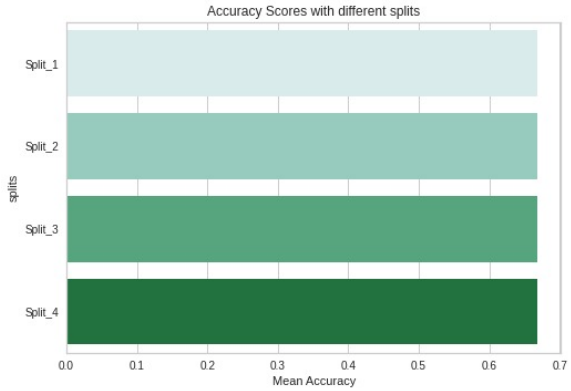

Figure 22. Accuracy Scores with different splits

Different classification algorithms were tested for the modelling. The training dataset performed well overall, showing an accuracy between 79% and 95%.

```
LR: 0.868014 (0.003814)
LDA: 0.862956 (0.004120)
KNN: 0.793454 (0.005110)
DT: 0.931878 (0.003324)
RFC: 0.953072 (0.002000)
NB: 0.855657 (0.004688)
```

Figure 23 Machine learning models comparison

After analysing the models' performance, it was decided to use the following models: RandomForest, Logistic Regression, and KNeighborsClassifier. A brief description of each model and a cross-validation score with the performance from the chosen models are given below for a better understanding:

Random Forest: Following the same supervised line of the model above, a set of trained random decision trees is built through a combination of models that increase the accuracy of the final result. According to Sarkar and Natarajan (2019), the random forest minimises overfitting because it tends to give more accurate results and does not affect the outliers.

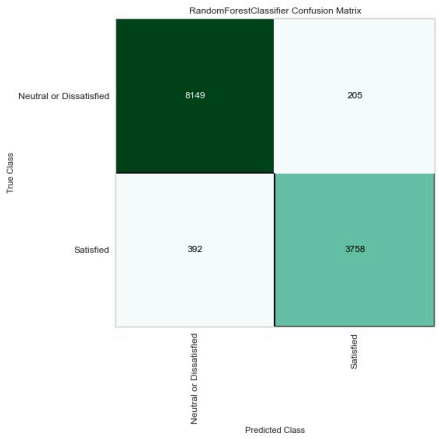

Figure 24.Random Forest Confusion Matrix

The confusion matrix tells that from 12504 surveys, 8354 surveys were predicted as "Neutral or Dissatisfied", and 403 were classified as "Neutral or Dissatisfied" when they were actually

"Satisfied" (type one error). Out of 4150 surveys that were predicted as "Satisfied", 215 were classified as "Satisfied" when in fact, they were "Dissatisfied". According to the previous sentence, Random Forest is the best model choice because it is an algorithm with the best classification, as it is possible to see that the diagonal results are higher.

Logistic Regression: According to Samir Madhavan (2015), the logistic function is handy and can take any value from negative infinity to positive infinity and return a value between 0 and 1. Therefore, it can be interpreted as a probability. Considered a primary method, it is also a fundamental resource, fast, simple, and easy to understand the results that use the sigmoid function, with values of either 0 or 1, mainly used for binary classifications.

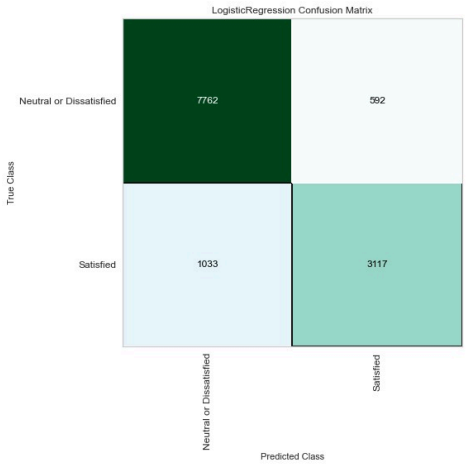

Figure 25.Logistic Regression Confusion Matrix

The confusion matrix above tells that from 12504 surveys, 8354 were predicted as "Neutral or Dissatisfied", and 1033 were classified as "Neutral or Dissatisfied" when they were actually "Satisfied" (type one error). Out of 4150 surveys that were predicted as "Satisfied", 595 were classified as "Satisfied" when in fact, they were "Dissatisfied".

K Neighbor Classifier: Also considered simple and one of the most used models for classification and regression problems. Its purpose is to use the nearest neighbours method, classifying them through the calculated distance. It is more correctly used when the data set has a low number of features because this model tends to overfit, causing the curse of dimensionality.

Chatterjee (2021) states that KNN is usually evaluated in the following aspects: ease of interpretation that can show a better visualisation for stakeholders, time to calculate the output in a way that tends to get the results quickly, and power to predict accuracy once the data set is not impacted by outliers. Chatterjee (2021) also affirms that the results can vary depending on the chosen distance measurement method. The most common distance measure for this algorithm is the Euclidean distance.

The confusion matrix tells that from 12504 surveys, 8354 were predicted as Neutral or Dissatisfied, and 1903 were classified as Neutral or Dissatisfied when satisfied (type one error). Out of 4150 surveys predicted as satisfied, 679 were classified as "Satisfied" when they were "Dissatisfied".

To improve models, two different techniques were performed: one was the standardisation of the data set, and the other was the hyperparameters to enhance the model's accuracy. As it is known, Hyperparameters control the capacity of the model to prevent overfitting.

Evaluation

Considering that the main objective of this study is to understand the factors that most impact the dissatisfaction of passengers on short air flights, then, because of the resolution of previous training, it was decided to use the Precision, Recall and F1 score classification report in each model to analyzRandome the level of the satisfaction of the passengers:

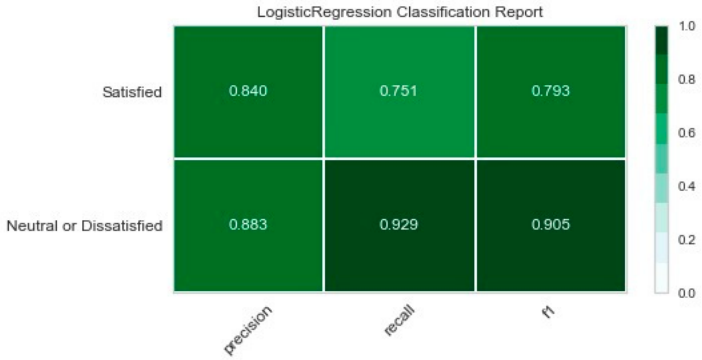

Figure 26. Logistic Regression Classification Report

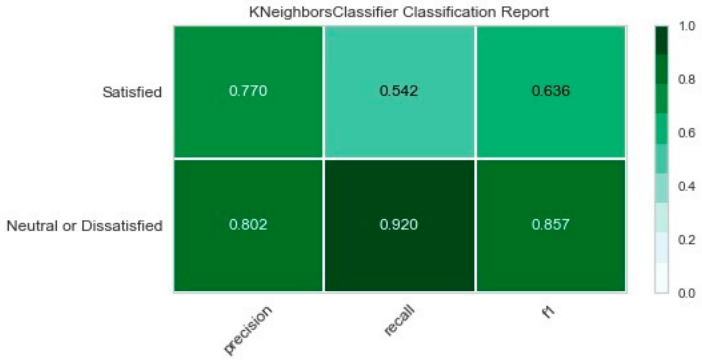

Figure 27. KNN Classification Report

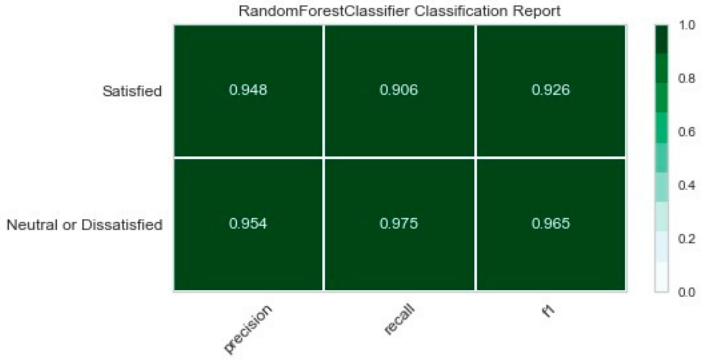

Figure 28. Random Forest Classification Report

According to the results of Precision, Recall and F1 score, it can be said that the level of dissatisfaction is higher than the satisfactory level overall. The passengers are not happy with the service provided by the aviation companies.

Further, cross-validation was applied to every model to check each model's generalisation, which demonstrates how well a trained model classifies data. According to Refaeilzadeh et al. (2009), Cross-Validation is a statistical method to evaluate and compare algorithms that are in the learning process dividing them into two parts, one part is used to train, and the other part is used to predict and validate the model.

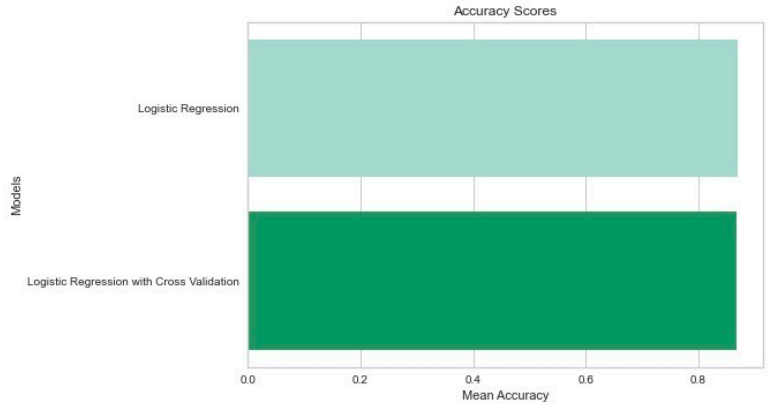

Figure 29 Cross-Validation Logistic Regression

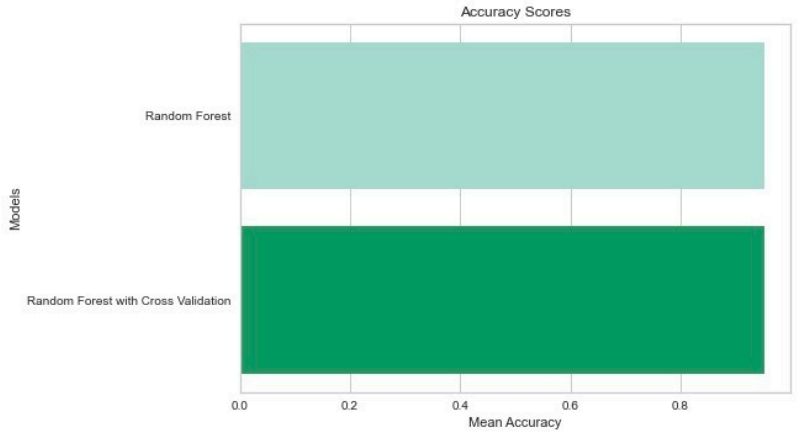

Figure 30.Cross Validation Random Forest

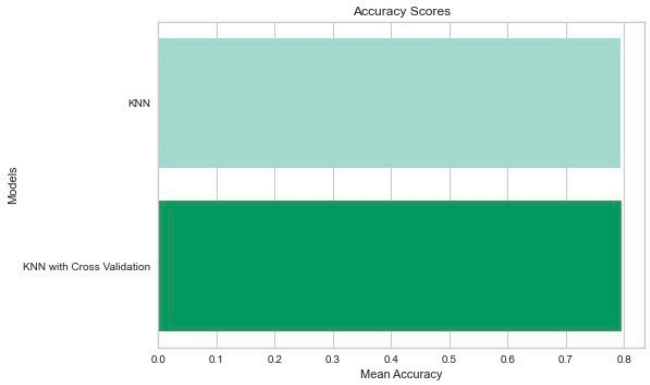

Figure 31. Cross Validation KNN

According to the graphs above, it can state that the results are close to reality. In that sense, the models are not under or overfitting.

Regardless of the accuracy results being significant in the modelling algorithms, other factors influenced the decision-making in the application of the model to the evaluation stage, such as the K Neighbors Classifier (KNN) is an algorithm sensitive to the presence of outliers, and this could cause an overfitting of the model, so it was considered not to apply it in this step and choose Random Forest algorithm.

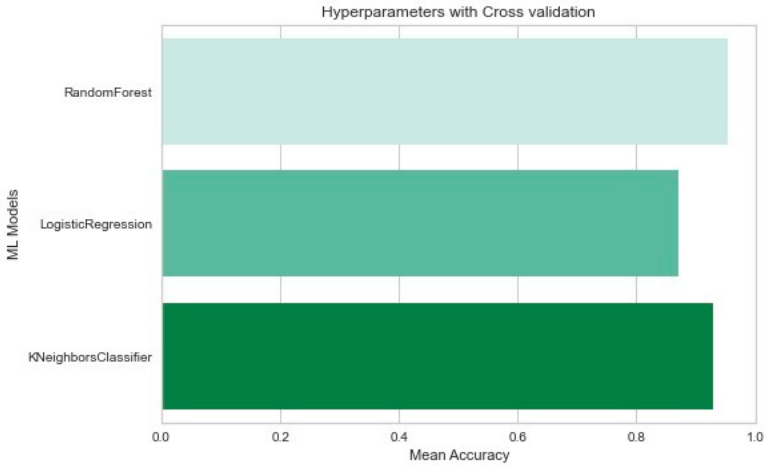

Figure 32. Hyperparameters with Cross-validation

Deployment

Some challenges were encountered during the process, such as the data set's division and the hyperparameters' application in just one code. Therefore, a more in-depth study was necessary so that there was a manual configuration of the algorithm so that the results were assertive and reached the objective of this parameter, which is the optimisation of the training models.

```
 1) In-flight Entertainment                0.274160
 2) Gate Location                          0.240379
 3) Class                                  0.115067
 4) Flight Distance                        0.060927
 5) Baggage Handling                       0.042964
 6) Type of Travel                         0.040036
 7) Check-in Service                       0.031775
 8) Online Boarding                        0.022607
 9) In-flight Wifi Service                 0.021694
10) Satisfaction                           0.021471
11) Departure Delay                        0.020289
12) Cleanliness                            0.019499
13) Leg Room Service                       0.019327
14) Seat Comfort                           0.015565
15) Ease of Online Booking                 0.011509
16) Food and Drink                         0.010695
17) On-board Service                       0.009626
18) Departure and Arrival Time Convenience 0.009604
19) Arrival Delay                          0.007516
20) In-flight Service                      0.005290
```

Figure 33. Factors that impact the satisfaction

The above and the below plots show that "In-flight Service, "Arrival Delay", "Departure and Arrival Time", "On-Board Service", "Food and Drink", "Ease of Online Booking"," Seat Confort" are the top most import factors that affect **passengers dissatisfaction**.

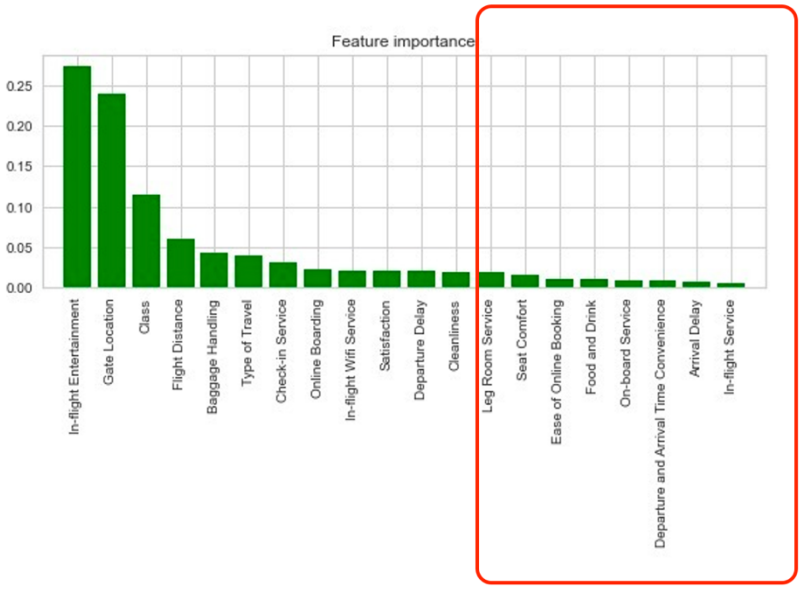

Figure 34. Plot of the factors that impact satisfaction

The analysis presents the factors containing critical information that aviation companies must improve regarding their services to be ahead of the competition. This study could be a good reference for airlines to use as a customer evaluation-driven service methodology to improve their services and competitiveness.

The results are very satisfactory to its complete development, and the group looks forward to implementing the following steps.

Extra Contents

Roles and responsibilities

A multidisciplinary team formed this research to meet the established goals and perform well even within diversities founded during its development. The roles were described as follows:

Team Member	Role	Responsibilities
Barbara	Data Researcher/Modeling undestanding/Evaluation	- Producing questions to analyse the data set;
		- Creating and managing the project in Trello;
		- Ensuring the updates from team meetings;
		- Test solutions to validade objectives;
		- Modeling Undestanding;
		- Producing and managing documentation of the report;
		- Revising final code and report compilations to ensure clarity in the reports.
		- Evaluation and Conclusions.
Daniela Daia	Business Analyst/ Data Visualisation/Evaluation	- Producing questions to analyse the data set;
		- Determining the methodology used on the project;
		- Managing the team's task in Trello;
		- Managing the development team meetings;
		- Project background development;
		- EDA understanding;
		- Producing and managing documentation of the report;
		- Revising final code and report compilations to ensure clarity in the reports.
Isabel Nieves	Project Manager /Data Analyst/Data Researcher/Modeling and improve ML models	- Producing questions to analyse with the data set;
		- Gathering data from Kaggle in the project plan;
		- Organising and cleaning the data in a specific format;
		- Managing tasks in Trello;
		- Managing the development team meetings;
		-Development of EDA;
		-Apply ML models and improve the accuracy;
		-Development of Confusion Matrix;
		- Revising final code and report compilations to ensure clarity in the reports;
		- Evaluation and Conclusions.
Vicente Rubio	Data Analyst/Data Researcher/Modeling and improve the ML models	- Producing questions to analyse with the data set;
		- Gathering data from Kaggle in the project plan;
		- Managing tasks in Trello;
		- Ensuring the updates from team meetings;
		- Split and compare of test and training;
		- Modeling the data set;
		- Development of Confusion Matrix;
		- Revising final code and report compilations to ensure clarity in the reports;
		- Evaluation and Conclusions.

Figure 35. Roles and Responsibilities

Team Project management

The team decided to work with Trello for the project management as it is a free, simple, and easy-to-use collaboration tool that enables us to organise and track the project on the board; it was a great tool to help to accomplish the task on time required, as seen in our board below:

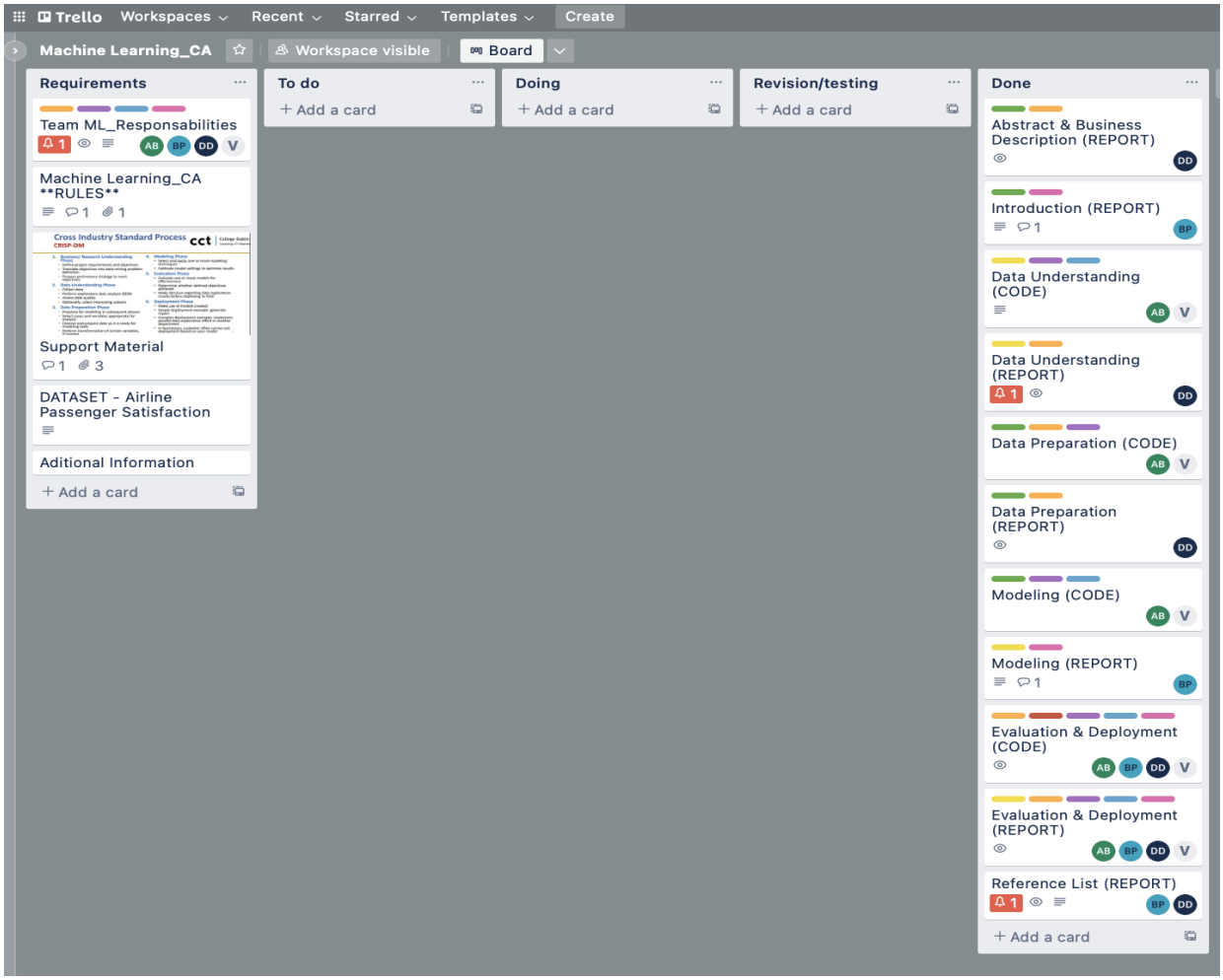

Figure 36. Project Management: Trello Board

Finally, the team's effort is represented in the pie chart below:

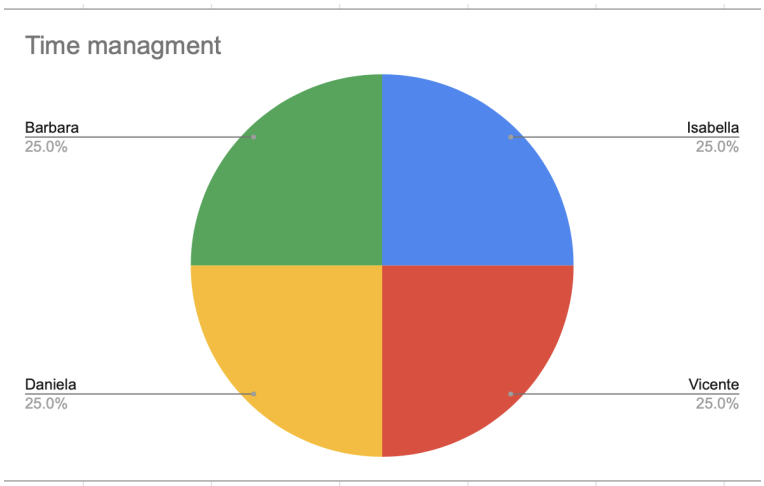

Figure 37.Team's effort

Team’s challenges faced

Ana Isabel Nieves Barcenas

Project challenges

-Finding a database that suits the needs of our problem domain.

-Find the most effective way to compare the different ML models to avoid overwork.

-How to determine which parameters I should consider improving each model to find the best possible combination and obtain higher accuracy.

Personal challenges

-Be able to organize my time to be able to carry out my responsibilities and the responsibilities I had within the team.

-Playing the project management role, it was challenging to deal with the different points of view of the team members to reach an agreement.

Bárbara Azevedo Pereira

This project had a very dedicated, intensive, and detailed development, always looking for improvements and ease of demonstration for the stakeholders. One of the important points to highlight and also one of the biggest challenges was undoubtedly the organization of teamwork,

as well as dealing with different opinions and backgrounds. In view of all the difficulties encountered along the way, the project was successfully completed, demonstrating a satisfactory result, reaching the main objective and which can be used as a basis for future analysis of problems related to the subject.

Daniela Daia

- As an individual, I faced challenges with the google collab tool for the first time using it. Also, in building the report, I needed to install Office as google docs were no longer supporting two people working simultaneously. After, Formatting the report and organizing the references was as well something really challenging.

- As a group, misinterpreting what was being discussed led to a lack of understanding and poor communication with some individuals, causing conflicts; also, working as a team, people must understand that may require negotiation and compromise.

Vicente Rubio

My challenge was the exploration and research on programming in Python. Second, it was being able to work as a team to achieve our purpose, taking into account the moods of the people and myself.

Reference List

Bart, C.K. (2000) The Relationship between Mission and Innovativeness in the Airline Industry: An Exploratory Investigation. London: Int. J. Technology Management.

Brownlee, J. (2020) Data Preparation for Machine Learning: Data Cleaning, Feature Selection and Data Transforms in Python. New York: Jason Brownlee.

Chatterjee, I. (2021) Machine Learning and Its Application: a Quick Guide for Beginners. London: Bentham Science Publishers.

InsureMyTrip (2021), Travel Tips for Long Haul Flights. InsureMyTrip Available at: https://www.insuremytrip.com/travel-advice/travel-tips/long-haul-flights/#:~:text=While%20no%20international%20standard%20definition,any%20flight%20length%20in%20between [Accessed 20 November 2022].

Google Colaboratory (2022) Tutorials. Available at: https://research.google.com/colaboratory/faq.html [Accessed 26 November 2022].

Han, J., Kamber, M. and Pei, J. (2011) Data Mining Concepts and Techniques. 3rd Edition. Burlington: Morgan Kaufmann Publishers.

IMB, (2021) IBM Documentation. Available at: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview [Accessed 20 November 2022].

Little, R., Rubin, D. (2019) Statistical analysis with missing data, third Edition. United States of America: Wiley.

McKinney, W (2017) Python for Data Analysis. 2nd ed. Beijing: O'Reilly Inc.

Madhavan, Samir (2015) Mastering Python for Data Science: Explore the World of Data Science through Python and Learn How to Make Sense of Data. Birmingham: Packt Publishing Limited.

Nair, A. (2022) Baseline Models: Your Guide for Model Building. Medium. Available at: towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d#. [Accessed 20 November 2022].

Navlani, A.(2019) Python Logistic Regression Tutorial with Sklearn & Scikit. Datacamp.com. Available at: www.datacamp.com/tutorial/understanding-logistic-regression-python.[Accessed 21 November 2022].

Navlani, A. (2018) KNN Classification Tutorial Using Sklearn Python. Datacamp.com. Available at: www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn [Accessed 22 November 2022].

Ramalho, L. (2015) Fluent Python: clear, concise and practical programming. 1st Edition. United States of America: O'Reilly Media Inc.

Raúl Garreta, Moncecchi, G., Hauck, T. and Hackeling, G. (2017) Scikit-learn: Machine Learning simplified. Birmingham, UK: Packt Publishing.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009) Cross-Validation. Available at: http://leitang.net/papers/ency-cross-validation.pdf [Accessed 26 November 2022].

Sarkar, D. and Vijayalakshmi, N. (2019) Ensemble Machine Learning Cookbook Over 35 Practical Recipes to Explore Ensemble Machine Learning Techniques Using Python. Birmingham Packt Publishing.

The panda's development team (2022) Python Data Analysis Library. [Pandas Documentation. Available at: https://pandas.pydata.org/docs/ [Accessed 22 November 2022].

Wertz, C. J. (1993) The Data Dictionary Concepts and Uses. 2nd Edition. New York: QED Information Sciences, Inc.

Williams, P. & Naumann, E. (2011) Customer satisfaction and business performance: A firm-level analysis. J. Serv. Market. Available at: https://www.researchgate.net/publication/235317930_Customer_satisfaction_and_business_performance_A_firm-level_analysis [Accessed 20 November 2022].