DATA ANALYTICS (UE20CS312) PROJECT REPORT

TOPIC: Airline passenger satisfaction analysis

NAME: RACHANA R

SRN: PES1UG20CS677

SECTION: 5L

OBJECTIVE:

In this project, I attempt to find the features that affect the satisfaction of the clients the most. In addition, I use different algorithms to predict that satisfaction.

Methods followed:

Dataset:

- Airline passenger satisfaction -Kaggle.

URL:

https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction?select=airline_passenger_satisfaction.csv

About data:

The dataset contains an airline passenger satisfaction survey.

- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (First time travel, Returning)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Flight distance: The flight distance of this journey
- Inflight wifi service: Satisfaction level of the inflight wifi service
- Departure/Arrival time convenient: Satisfaction level of departure/arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment
- On-board service: Satisfaction level of on-board service
- Leg room service: Satisfaction level of leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival
- Satisfaction: Airline satisfaction level (Satisfied, neutral or dissatisfaction)

1.Reading dataset

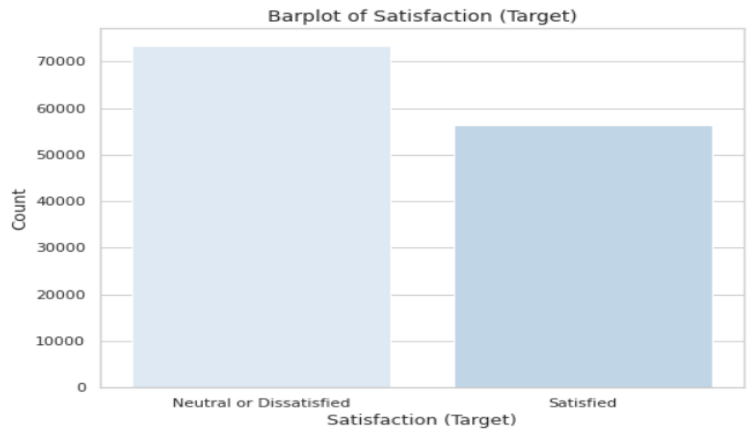
[62]:
```
df.info()
df.shape
```

```
<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
```

[62]: (129880, 24)

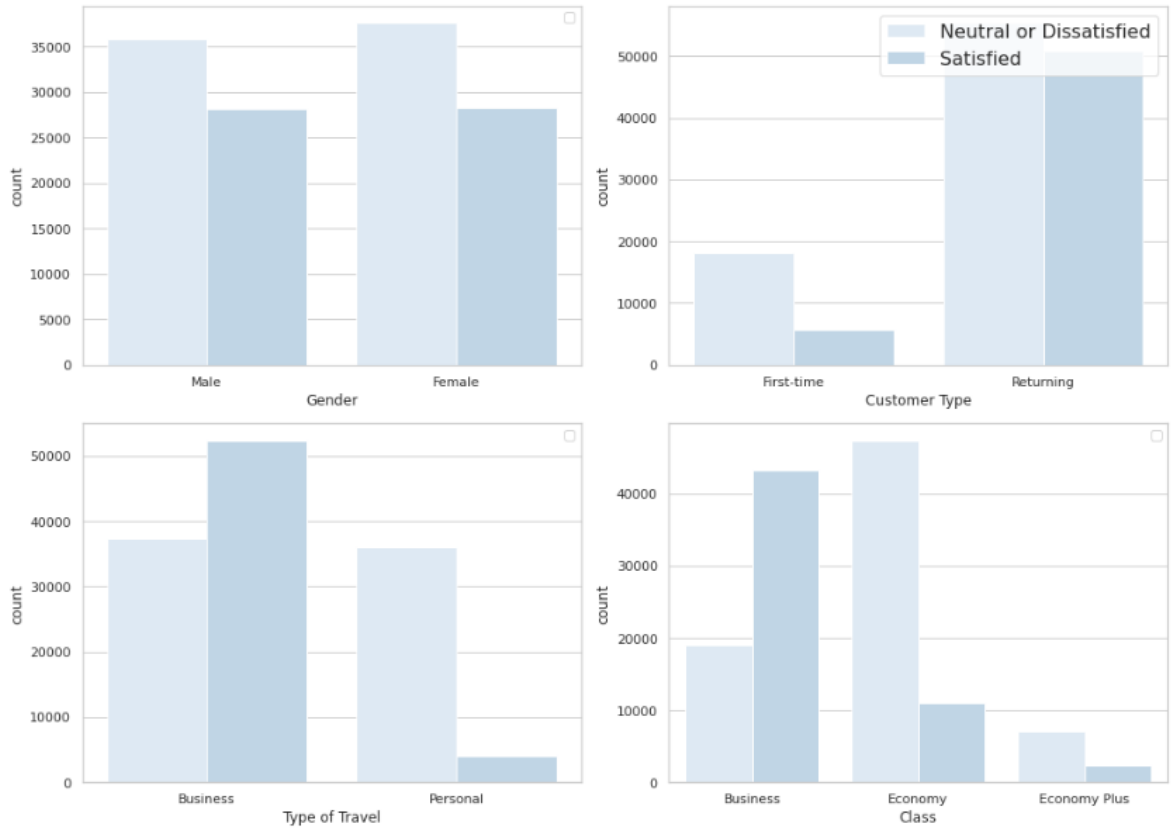
2.Data visualization

Target Variable - Satisfaction



It seems like the data is quite balanced, with slightly more neutral or dissatisfied costumers

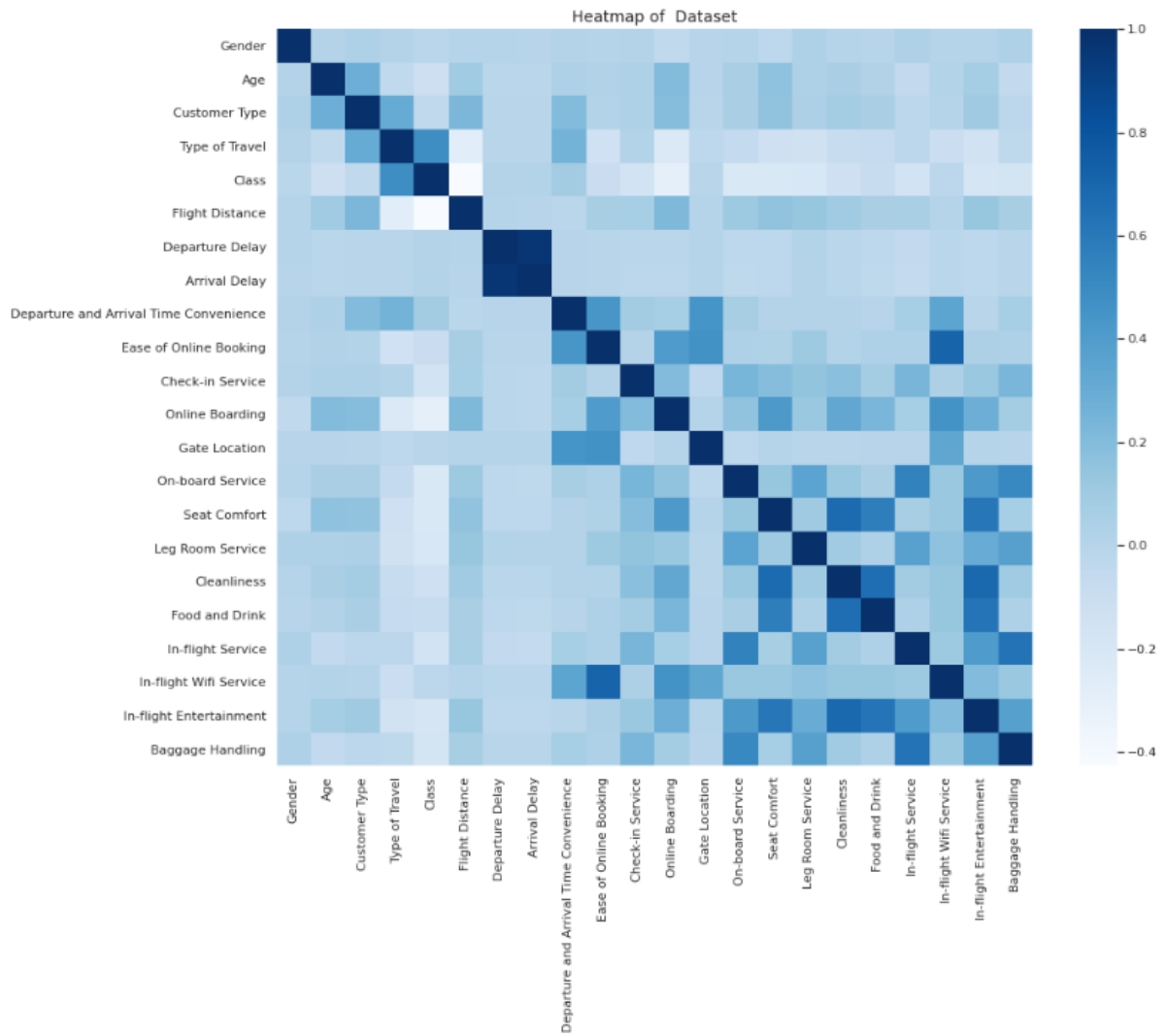
Visualizing distribution of Categorical Variables



Can see few interesting points:

- Males and females have similar satisfaction rates.
- First time travelling passengers are more dissatisfied compared to returning type passengers.
- Personal travellers are more dissatisfied compared to buisness travel.
- Eco class are more dissatisfied compared to buisness class.

Correlation Analysis:



Most features don't seem to be very correlated, except for Departure Delay and Arrival Delay which have a correlation score of 0.96.

3. Data Wrangling

In this step satisfaction column has been transformed:

- **Satisfied=1**
- **Neutral or dissatisfied=0**

Looking for Missing values

```
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
```

As the variable Arrival Delay is highly correlated with the variable Departure Delay, one of it can be dropped.
As Arrival Delay has many missing values, it can be dropped.

One-Hot Encoding

```
dummy_df.head()
```

	Age	Flight Distance	Departure Delay	Departure and Arrival Time Convenience	Ease of Online Booking	Check-in Service	Online Boarding	Gate Location	On-board Service	Seat Comfort	...	In-flight Service	In-flight Wifi Service	In-flight Entertainment	Baggage Handling	Satisfaction	Gender_Male	Customer Type_Returning	Type of Travel_Personal	Class_Economy
0	48	821	2	3	3	4	3	3	3	5	...	5	3	5	5	0	1	0	0	0
1	35	821	26	2	2	3	5	2	5	4	...	5	2	5	5	1	0	1	0	0
2	41	853	0	4	4	4	5	4	3	5	...	3	4	3	3	1	1	1	0	0
3	50	1905	0	2	2	3	4	2	5	5	...	5	2	5	5	1	1	1	0	0
4	49	3470	0	3	3	3	5	3	3	4	...	3	3	3	3	1	0	1	0	0

5 rows × 23 columns

Representing the categorical variables as binary vectors.

Scaling

	Age	Flight Distance	Departure Delay
0	0.566960	-0.370261	-0.333948
1	-0.292868	-0.370261	0.296454
2	0.103976	-0.338179	-0.386481
3	0.699242	0.716512	-0.386481
4	0.633101	2.285515	-0.386481

```
final_df.head().style.set_properties(**{'text-align': 'center'}).set_table_styles([dict(selector='th', props=[('text-align', 'center')])])
```

	Age	Flight Distance	Departure Delay	Departure and Arrival Time Convenience	Ease of Online Booking	Check-in Service	Online Boarding	Gate Location	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Food and Drink	In-flight Service	In-flight Wifi Service	In-flight Entertainment	Baggage Handling	Satisfaction	Gender_Male	Customer Type_Returning	Tr
0	0.566960	-0.370261	-0.333948	3	3	4	3	3	3	5	2	5	5	5	3	5	5	0	1	0	
1	-0.292868	-0.370261	0.296454	2	2	3	5	2	5	4	5	5	3	5	2	5	5	1	0	1	
2	0.103976	-0.338179	-0.386481	4	4	4	5	4	3	5	3	5	5	3	4	3	3	1	1	1	
3	0.699242	0.716512	-0.386481	2	2	3	4	2	5	5	5	4	4	5	2	5	5	1	1	1	
4	0.633101	2.285515	-0.386481	3	3	3	5	3	3	4	4	5	4	3	3	3	3	1	0	1	

Scaling is done inorder to standardize the independent features like age, flight distance and departure delay in the data in a fixed range.

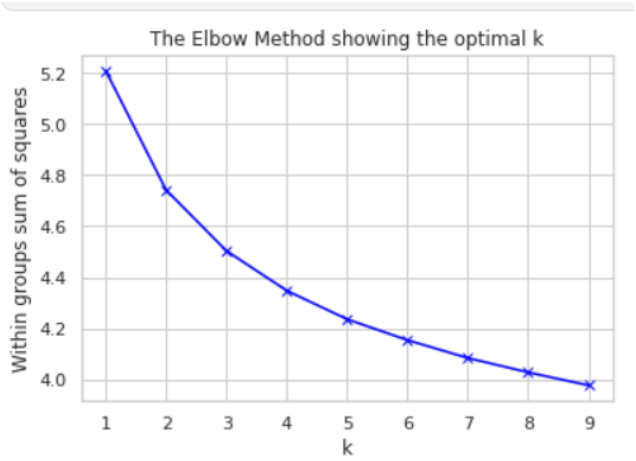
4.Feature Importance

In this section, made an attempt to find the features that affect the clients' satisfaction the most, using different methods.

4.1 Clustering Algorithm - K-Means

using the K-Means clustering algorithm to identify clusters and find important features in each cluster.

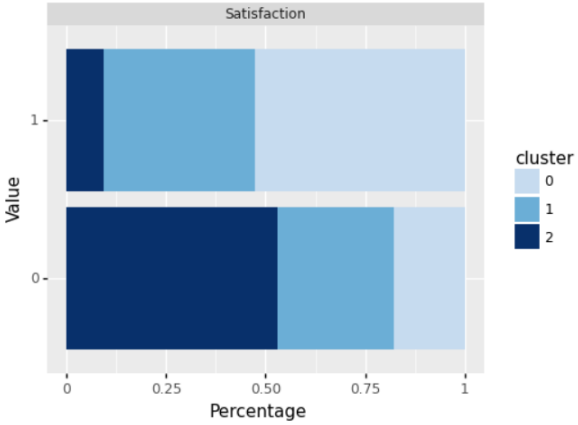
Defining the `elbow_met` function to find the optimal number of clusters `k`.

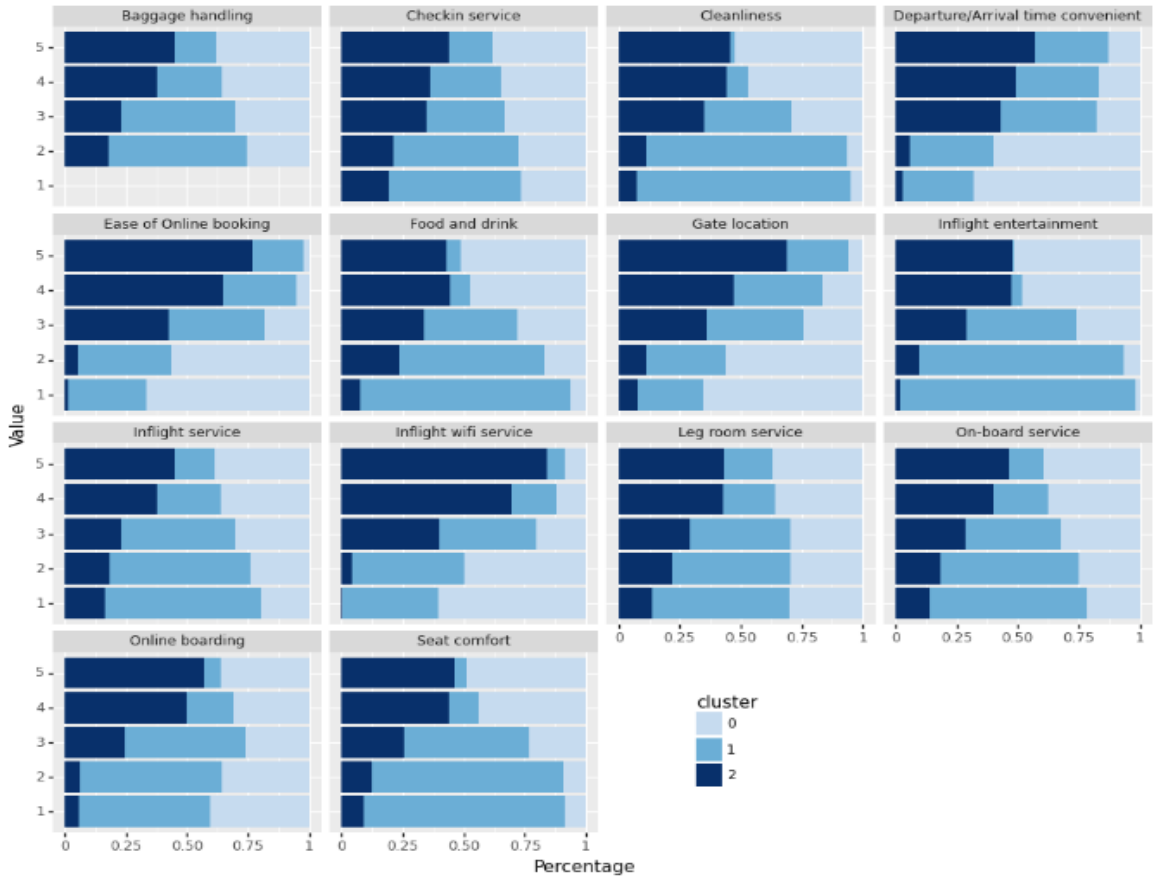


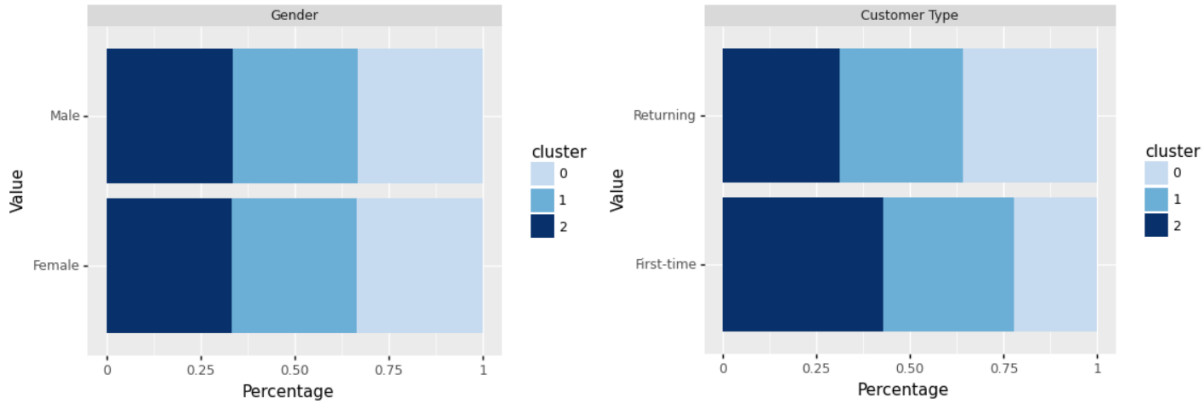
select k=3 and fit the model.

Defining a function `get_melted_clusters` to assign the observations to the clusters created by K-Means.

In the following plots, representing the different features and the percentage of observations that belong to every category, divided by cluster.




Baggage handling
Checkin service
Cleanliness
Departure/Arrival time convenient
Ease of Online booking
Food and drink
Gate location
Inflight entertainment
Inflight service
Inflight wifi service
Leg room service
On-board service
Online boarding
Seat comfort
Value
Percentage
cluster
0
1
2


Gender
Male
Female
Customer Type
Returning
First-time
Value
Percentage
cluster
0
1
2


Type of Travel
Personal
Business
Class
Economy Plus
Economy
Business
Value
Percentage
cluster
0
1
2

Conclusions

Cluster 0:

- Quite neutral satisfaction (a bit more satisfied).
- Features that had high ratings:
  - Cleanliness
  - Food and drink
  - Inflight entertainment
- Features that had low ratings:
  - Departure and Arrival time
  - Ease of Online booking
  - Gate location

Cluster 1:

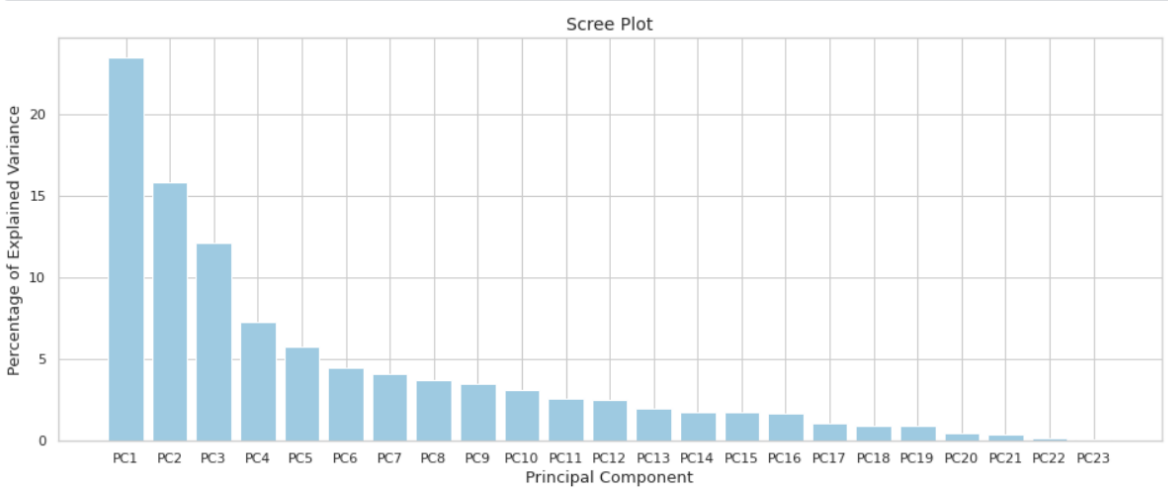
- Not satisfied with their experience.
- Mostly disloyal costumers in Eco/Eco Plus class for personal travel.
- Features that had extremely low ratings:
  - Cleanliness
  - Food and drink
  - Inflight entertainment
  - Inflight service
  - Seat comfort

Cluster 2:

- Mostly satisfied with their experience.
- Mostly loyal costumers in buisness class for buisness travel.
- Features that had extremely high ratings:
  - Departure and Arrival time
  - Ease of Online booking
  - Gate location
  - Inflight wifi service

4.2 Visualization Algorithm - PCA

1) using the PCA algorithm to find important features
2) creating scree plot to visualize the percentage of explained variance for each PC.



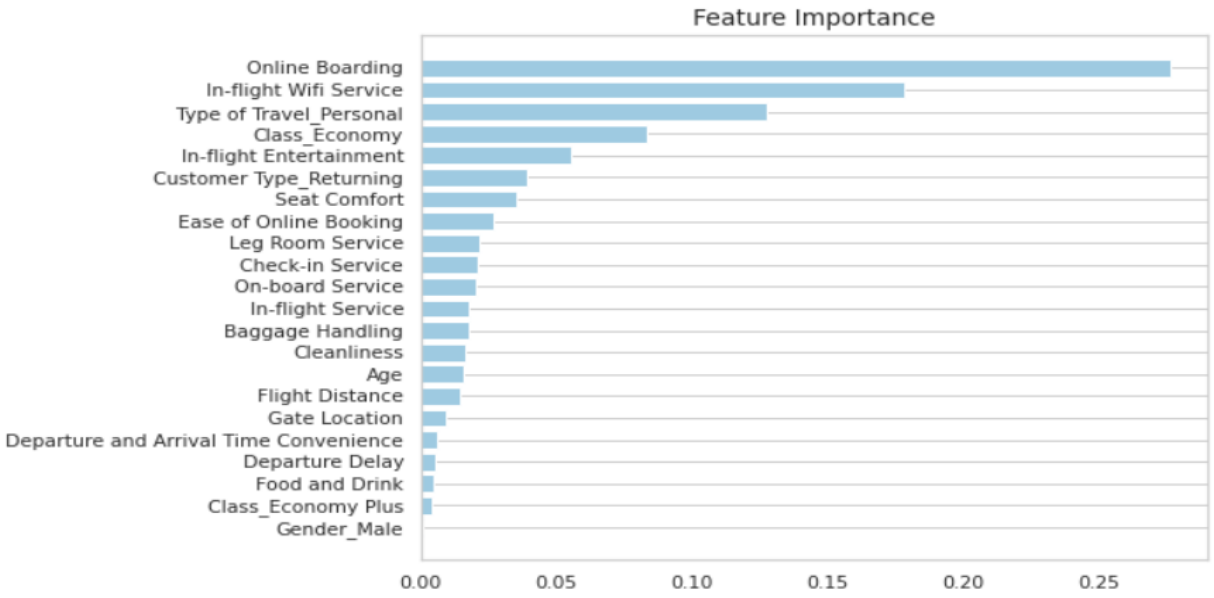
In the following table we see the 10 most important features for the first principal component.

	Score
In-flight Entertainment	-0.420072
Cleanliness	-0.358010
Seat Comfort	-0.354100
Food and Drink	-0.315279
Online Boarding	-0.314597
In-flight Wifi Service	-0.262107
On-board Service	-0.241422
Ease of Online Booking	-0.209818
Leg Room Service	-0.205480
In-flight Service	-0.200067

4.3 Classification Algorithms

fitting a Random Forest and and Ada-Boost model and find the most important features.

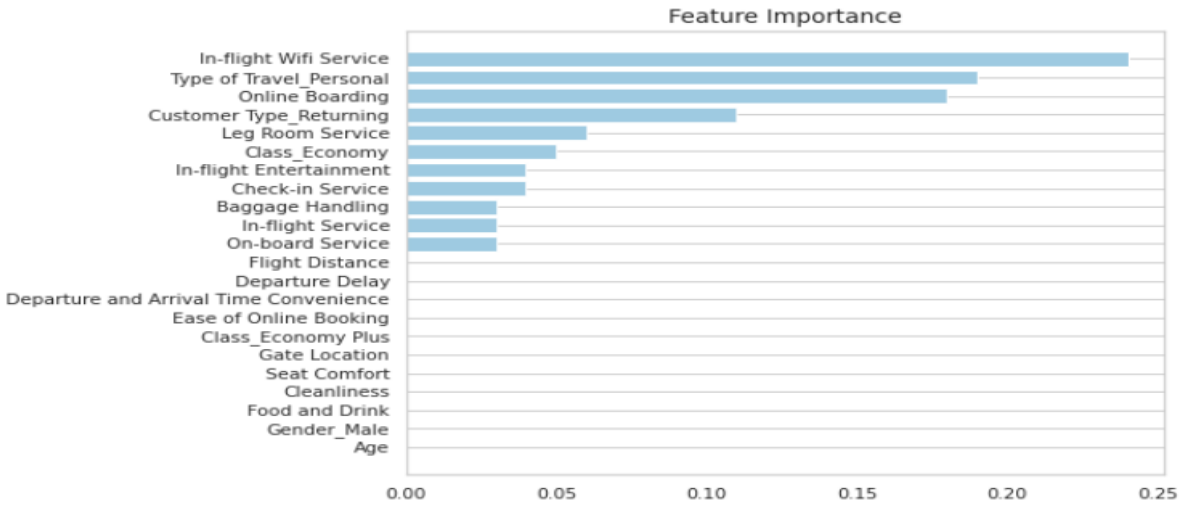
Random Forest Classifier



According to the Random Forest algorithm, the most important features are:

- Online boarding
- In-flight Wifi Service
- Type of Travel – Personal

Ada-Boost Classifier



According to the Ada-Boost algorithm, the most important features are:

- In-flight wifi service
- Type of Travel - Personal
- Online boarding

The same features as the Random Forest, but in different order.

4.4 Conclusion

According to the first two methods, K-Means clustering and PCA, the most important features are:

- Cleanliness
- Food and drink
- Inflight entertainment

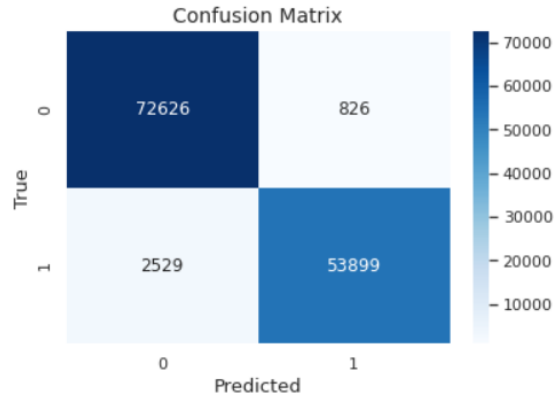
According to the classification algorithms, the most important features are:

- Online boarding
- Inflight wifi service
- Type of Travel - Personal Travel

5.Classification

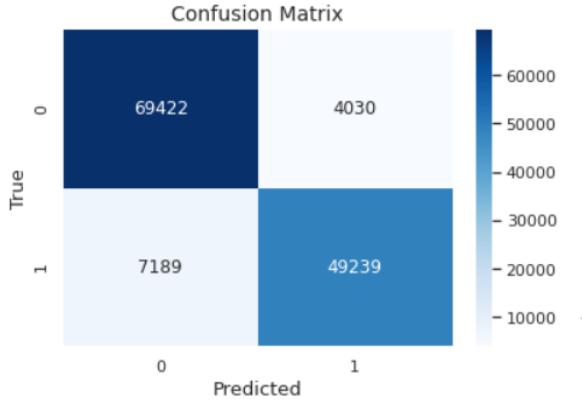
In this section, apply several classification algorithms on the data. To compare the methods, and calculate the accuracy and F1 score.

5.1 Random Forest Classifier



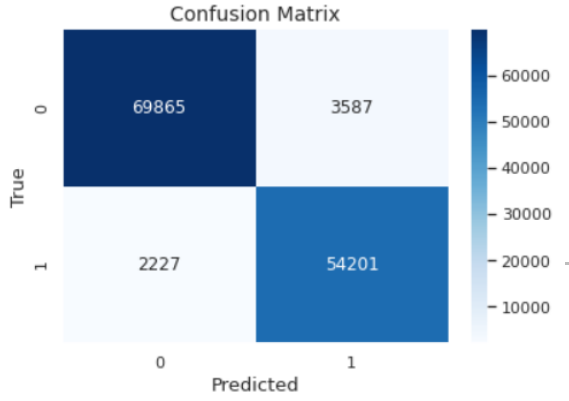
	Accuracy	F1 Score
Random Forest Classifier	0.974	0.97

5.2 Ada-Boost Classifier



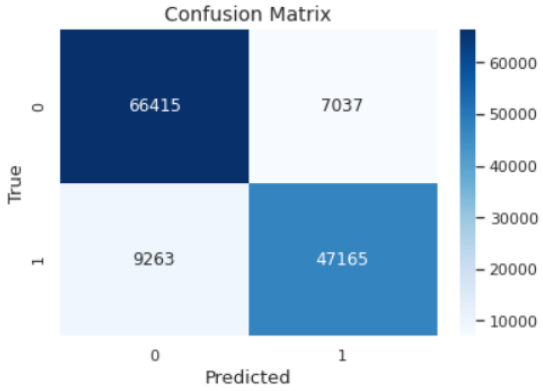
	Accuracy	F1 Score
Ada-Boost Classifier	0.914	0.898

5.3 MLP (Multi-Layer Perceptron)



	Accuracy	F1 Score
MLP	0.955	0.949

5.4 Logistic Regression



	Accuracy	F1 Score
Logistic Regression	0.874	0.853

5.5 Comparison

	Accuracy	F1 Score
Random Forest Classifier	0.974000	0.970000
Ada-Boost Classifier	0.914000	0.898000
MLP	0.955000	0.949000
Logistic Regression	0.874000	0.853000

The best predictions were given by the Random Forest algorithm, closely followed by MLP. Next, the Ada-Boost algorithm had quite good prediction, and lastly, the Logistic Regression had the worst predictions.

6.Regression

In this section create a new label that consists of the average score given by the costumers for the following features:

- Gate location
- Seat comfort
- Cleanliness

Then apply two regression algorithms on the data, using two different approaches:

1. **Drop the features** - the features that were used to create the label are dropped.
2. **Don't drop the features** - the features that were used to create the label are **not** dropped.

To compare the methods and approaches, calculate the R^2 and MSE.

6.1 New Label

create a new label that consists of the average score given by the costumers for the following features:

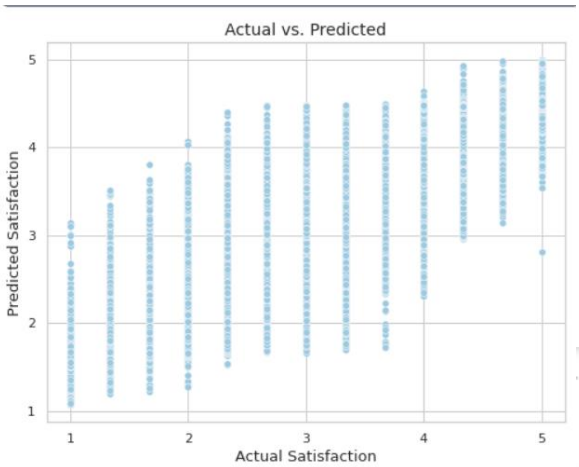
- Gate Location
- Seat Comfort
- Cleanliness

6.2 Method 1 - Drop Features

In the first method, the features that were used to create the label are **dropped.**

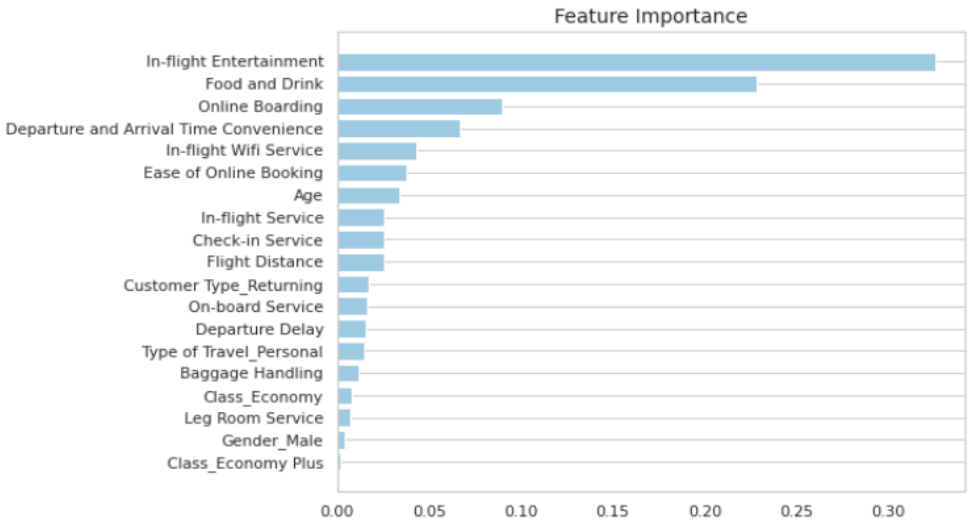
Then apply the Random Forest and Ada-Boost algorithms, and calculate feature importance.

Random Forest Regressor

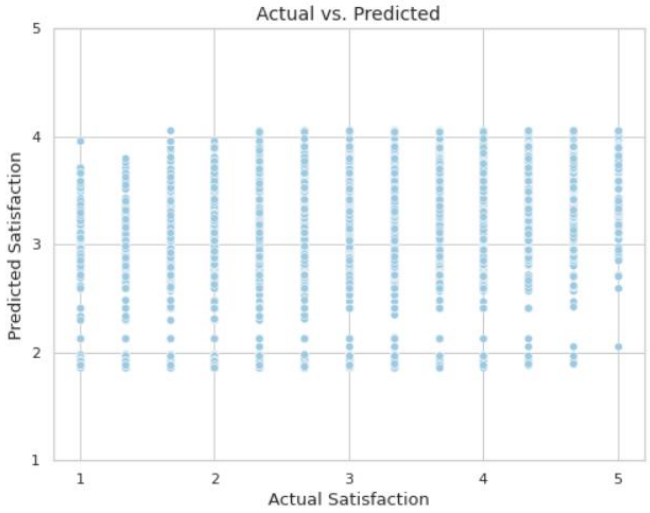


	R2	MSE
Random Forest Regressor	0.827352	0.142839

Feature Importance:

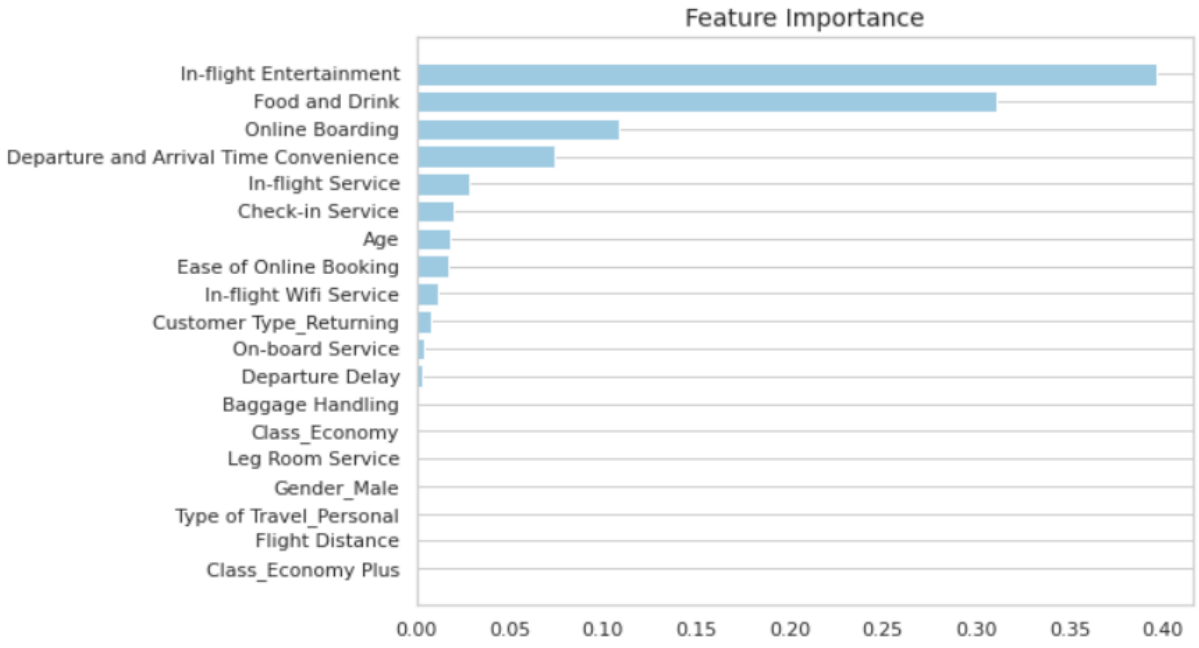


Ada-Boost Regressor



	R2	MSE
Ada-Boost Regressor	0.566494	0.358658

Feature Importance

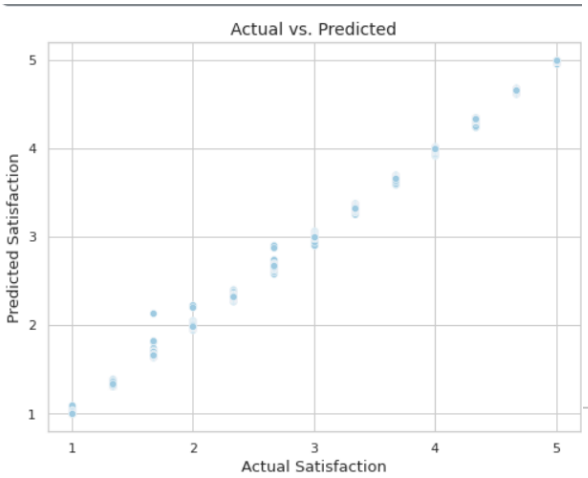


6.3 Method 2 - Don't Drop Features

In the second method, the features that were used to create the label are **not dropped.**

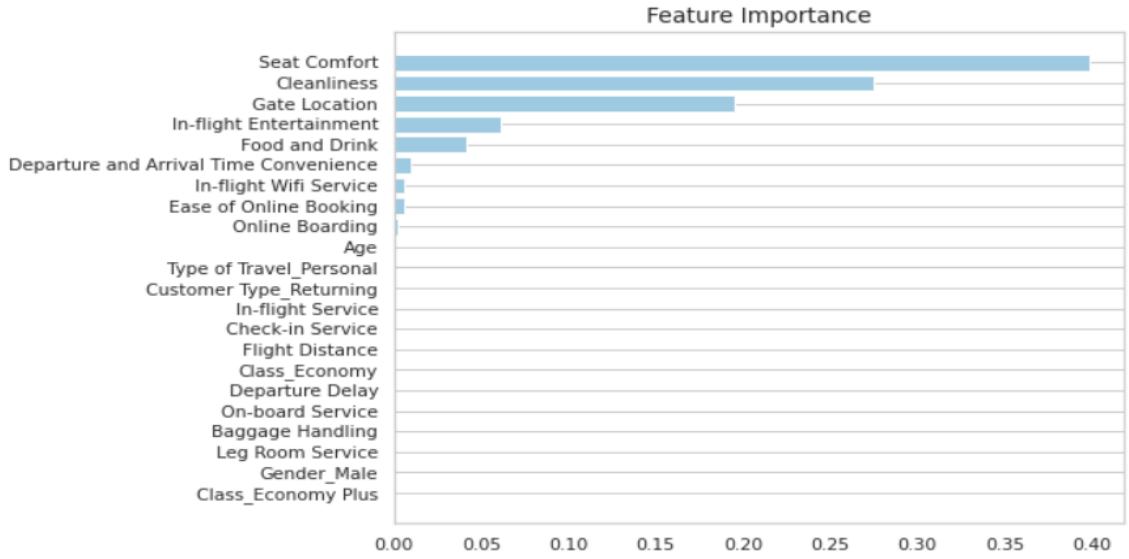
Apply the Random Forest and Ada-Boost algorithms, and calculate feature importance.

Random Forest Regressor

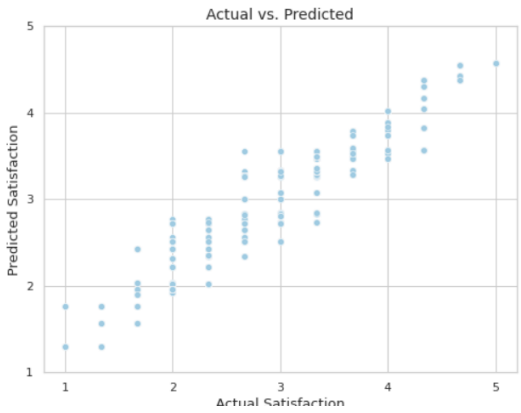


	R2	MSE
Random Forest Regressor	0.99996	0.000033

Feature Importance:

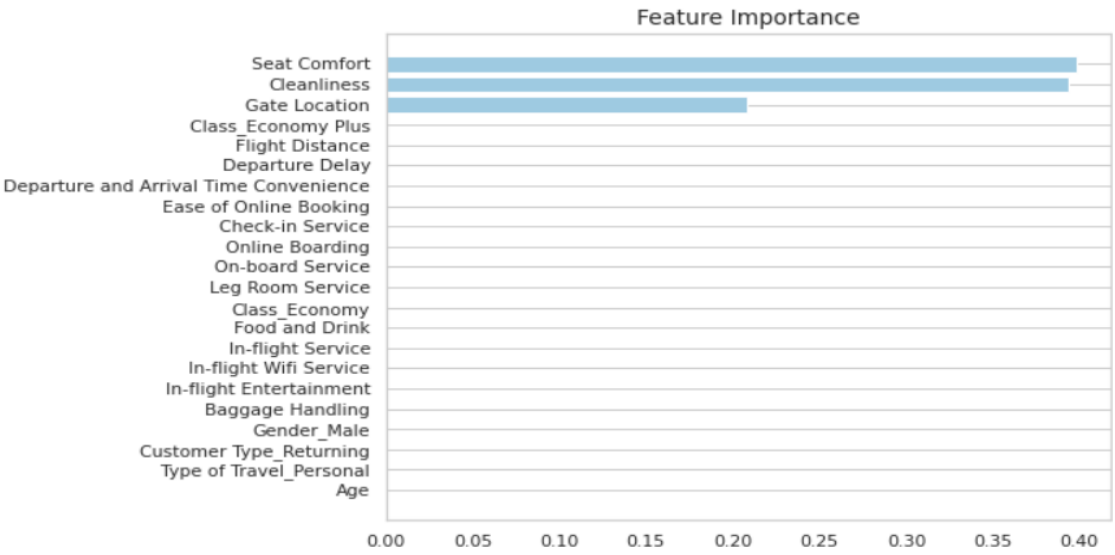


Ada-Boost Regressor



	R2	MSE
Ada-Boost Regressor	0.944418	0.045985

Feature Importance:



6.4 Comparison

The R^2 and MSE scores of the 1st and 2nd method are presented in the 1st and 2nd table, respectively.

1st method:

	R2	MSE
Random Forest Regressor	0.827352	0.142839
Ada-Boost Regressor	0.566494	0.358658

2nd method

	R2	MSE
Random Forest Regressor	0.999960	0.000033
Ada-Boost Regressor	0.944418	0.045985

- It can be observed that using the second method, the algorithms had extremely higher scores of R^2, much higher than using the first method, and also lower MSE scores.
- In terms of feature importance, using the second method, i.e. not dropping the features used to create the label, these features had the highest importance scores. In the Ada-Boost model, it seems like they were almost the only features used.

7.FINAL CONCLUSIONS

In this project, I attempted to find the features that affect the satisfaction of the clients the most. In addition, I used different algorithms to predict that satisfaction.

It is found that the most important features affecting the customer satisfaction were:

- Cleanliness
- Food and drink
- Inflight entertainment
- Online boarding
- Inflight wifi service
- Type of Travel - Personal Travel

Next, I have used four classification models to predict the satisfaction. **The best predictions were given by the Random Forest algorithm, closely followed by MLP.**

	Accuracy	F1 Score
Random Forest Classifier	0.974000	0.970000
Ada-Boost Classifier	0.914000	0.898000
MLP	0.955000	0.949000
Logistic Regression	0.874000	0.853000

Lastly, created a new label that consists of the average score given by the costumers for the following features:

- Gate location
- Seat comfort
- Cleanliness

Applied two regression algorithms on the data, using two different approaches:

1. **Drop the features** - the features that were used to create the label are dropped.
2. **Don't drop the features** - the features that were used to create the label are **not** dropped.

The R^2 and MSE scores of the first and second method are presented in the first and second table, respectively.

	R2	MSE
Random Forest Regressor	0.827352	0.142839
Ada-Boost Regressor	0.566494	0.358658

	R2	MSE
Random Forest Regressor	0.999960	0.000033
Ada-Boost Regressor	0.944418	0.045985

It has been seen **that using the second method, the features used to create the label had the highest importance scores and the algorithms gave much better predictions**, compared to the first method, where we dropped the features.