FINAL PROJECT

DATA SCIENCE BOOTCAMP – AIRLINE CUSTOMER SATISFACTION

CODIGO FACILITO

Topics

- Data Source
- Questions to Answer
- Conclusions

Data Source

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

kaggle

Sample of Dataset

First rows

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3
1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3
2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2
3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5
4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3
5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1180	3	4	2
6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276	2	4	2
7	96462	Female	Loyal Customer	52	Business travel	Business	2035	4	3	4
8	79485	Female	Loyal Customer	41	Business travel	Business	853	1	2	2
9	65725	Male	disloyal Customer	20	Business travel	Eco	1061	3	3	3

Questions to be Answered

- Which are the most remarkable insights in the data collected?
- How do the NULL values can influence a prediction model?
- Is it possible to predict the satisfaction variable? What information should be given?
- How to identify the more relevant features that impacts satisfaction?

Used Tools: Jupyter Notebook

Binder (mybinder.org)

Dataset Statistics

Overview

Dataset statistics

Number of variables	24
Number of observations	103904
Missing cells	310
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	53.8 MiB
Average record size in memory	542.9 B

Variable types

Numeric	18
Categorical	6

Which are the most remarkable insights in the data collected?

- Missing data only in Departure Time.
- There are delays in flights arrivals even though flights departure on time (delays on destination). Departure and Arrival average delays are 14~15 minutes. 56% of flights departure and arrive on time.
- Most of the passengers interrogated are loyal customers (82%) and business travelers (69%).
- Average age is 39 years old.
- As this data is from a US Airlines, most of the flights are continental US destinations, because flight distances.
- Feature Scoring scale from (0 to 5)

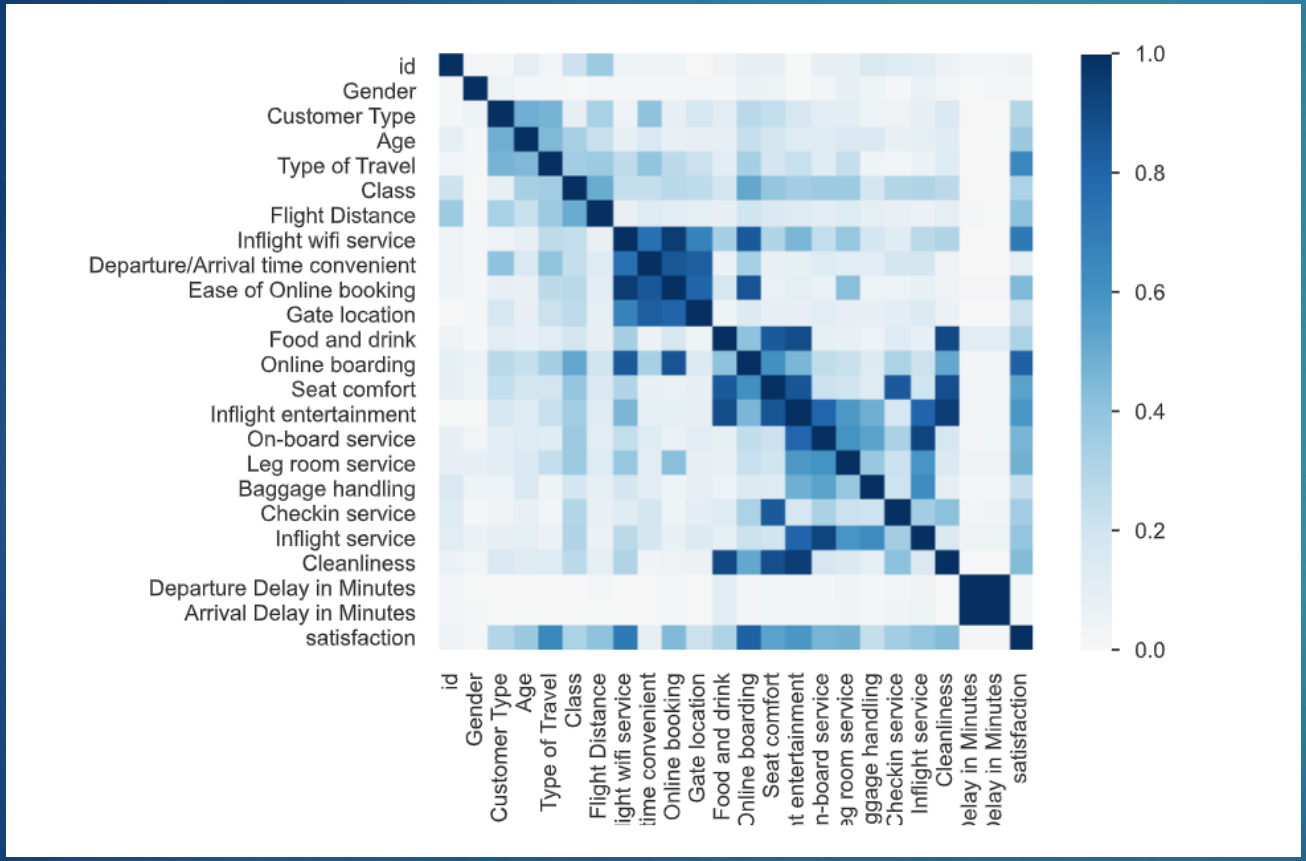
Which are the most remarkable insights in the data collected?

- Variables with the highest positive correlation: Ease of Online booking - Inflight wifi service, inflight entertainment - Cleanliness, inflight entertainment - Food & drink, Seat Comfort - Cleanliness, Departure Delay in Minutes - Arrival Delay in Minutes. Major Negative correlation between Inflight service with Departure Delay in Minutes - Arrival Delay in Minutes.
- Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking and Online boarding are the only features which have some zero scoring.
- In terms of satisfaction, 56.7% are neutral or dissatisfied and 43,3% of the people are satisfied.

How do the NULL values can influence a prediction model?

- We will use dataset **(NaN values filled with median value)** and dataset2 **(with NaN values dropped)** for our models to predict satisfaction, based on categorical and customer data and services survey scoring
- Our exercise will execute modeling based on these two dataset with different approaches to manage **NaN data** in Departure Flight column.

Is it possible to predict the satisfaction variable? What information should be given?



- From Pandas-Profiling feature, we discovered several types of correlations calculated. Phik (φk) correlation helps to combine categorical and numerical variables, with less impact on outliers.
- For the purpose of our research, that is to predict satisfaction indicator, having the correlation diagram for this, we could anticipate that satisfaction is highly related to the features:
  - Type of Travel
  - Inflight wifi service
  - Online boarding
  - Inflight entertainment
  - Seat comfort
- Let's verify if analytics models confirm this result.

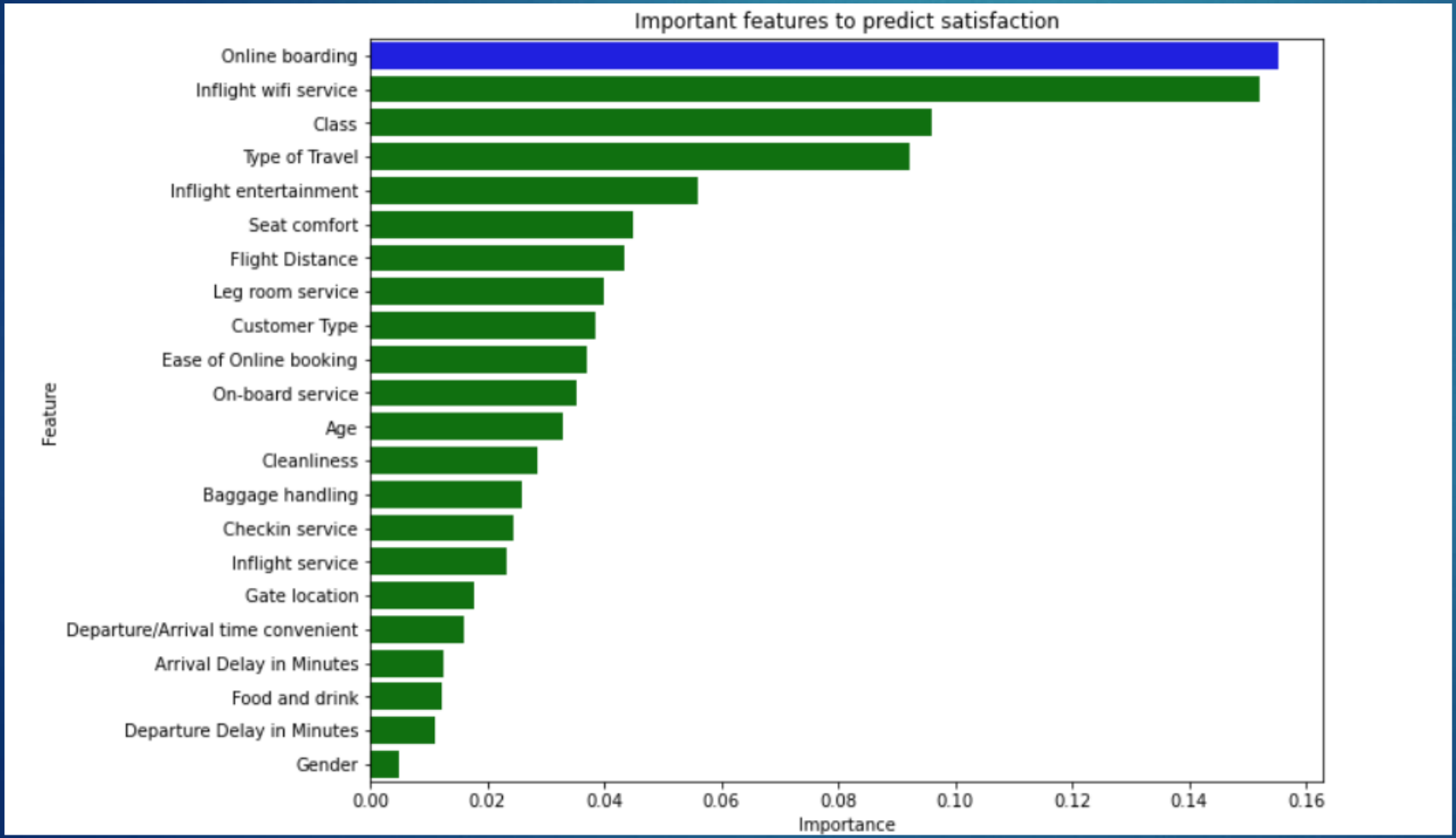
Is it possible to predict the satisfaction variable? What information should be given?

- In previous attempts we could conclude that effectiveness of the model was depending on the independent variables selected, if we only selected the features of the survey or if we included also the demographic variables (Age, Type of travel, Flight distance, etc).
- It was obtained that for the first case, our model Accuracy was lower as 53%, but including all variables could increase to 93%.
- Therefore, the model should include all relevant variable in the dataset.

Modeling Results

- A Random Forest Algorithm was suggested as we are working in a classification problem, and we require to establish which features were more relevant to predict satisfaction.
- Originally, we try to do this with a PCA, but process involved to lose original data impact and not able to identify from PCA components the main features.
- In General, both dataset (NaN values replaced by median and NaN values dropped) gave the same accuracy ~ 96%, hence both assumptions of handling NaN values were correct.

How to identify the more relevant features that impacts satisfaction?

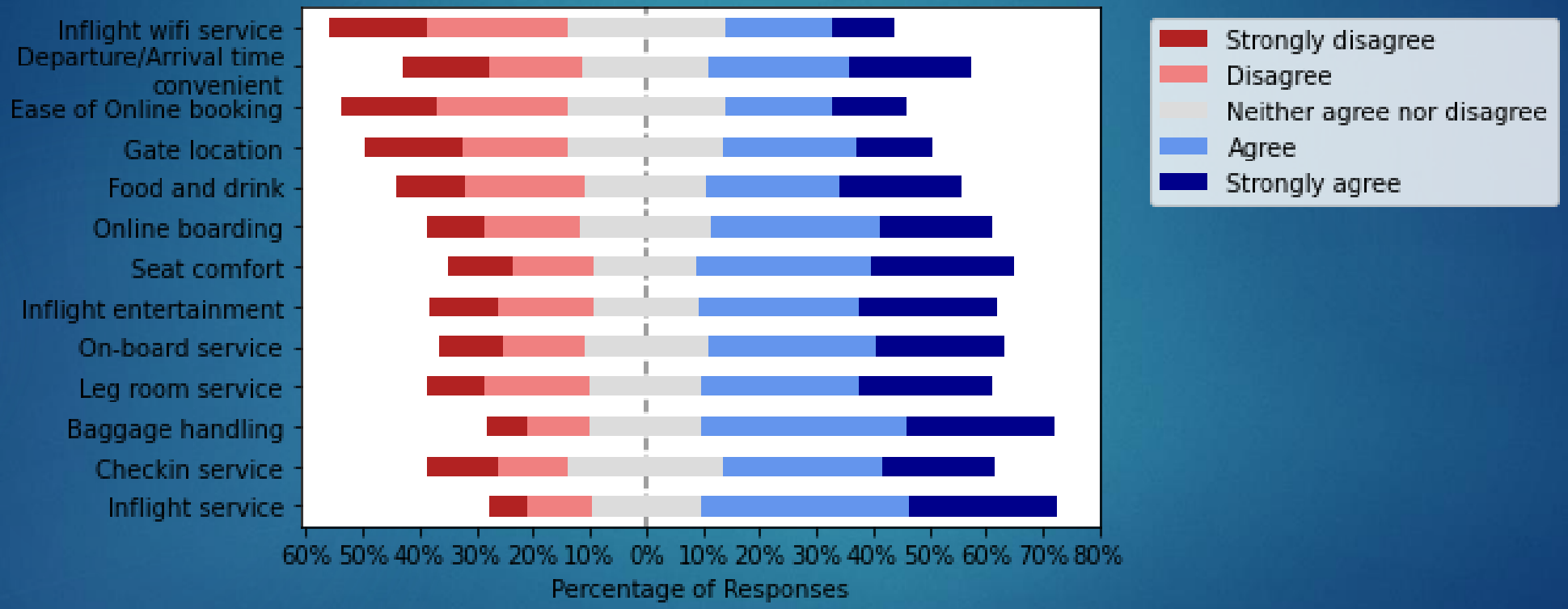


- The Random Forest Algorithm has a parameter, feature importance that calculates the statistical influence of a variable in the prediction of the target, results are shown below:

Conclusions

- Based on the feature importance, what we can conclude:
  - Airline should focus providing best quality and easy on boarding service, previous to arrive airport, and during the flight offer a best experience in inflight entertainment as well as an excellent wifi provider and having comfortable seats, primary in long-distance flights.
    - The most demanding passengers for these features are their loyalty customers, traveling for business purposes.
  - Phik Correlation gave a good approach of most important features to predict the target.
- Model to predict satisfaction is 96.2% Accurate, it can be implemented for operational usage.
- Analysis can be enhanced trying to make a clustering of the people who answered the survey, and identifying which features are relevant to each segment.

Conclusions : Results of Survey Likert Scale



Thank You!

JUAN JACOBO FIGUEROA