Big Data Project Document

Passenger Satisfaction problem

Team 11

Team Members:

Omar Mohammad	Sec. 2	B.N. 08
Mohammad Jamal	Sec. 2	B.N. 15
Walid Mohammad	Sec. 2	B.N. 31
Yahia Ali	Sec. 2	B.N. 33

1. Brief Problem Description

The problem we tackle in our project is a business one. Airline companies collect a lot of data about their passengers. After each trip, passengers are asked about their overall satisfaction as well as their rating of various services. Companies want to use this data to further enhance their services to maximize satisfaction. This isn't a very straightforward task; first glances may lead to entirely inaccurate decisions as there could be hidden correlations at play. This is the task we handle in the project.

2. Project Pipeline

- Data visualization
- Data Preprocessing
- Data Splitting - Split the data into training & testing sets (70:30)
- Training 6 models on relationship between satisfaction level and the most correlated features.
- Comparing the 6 models using 10-fold cross validation
- Association Rule mining

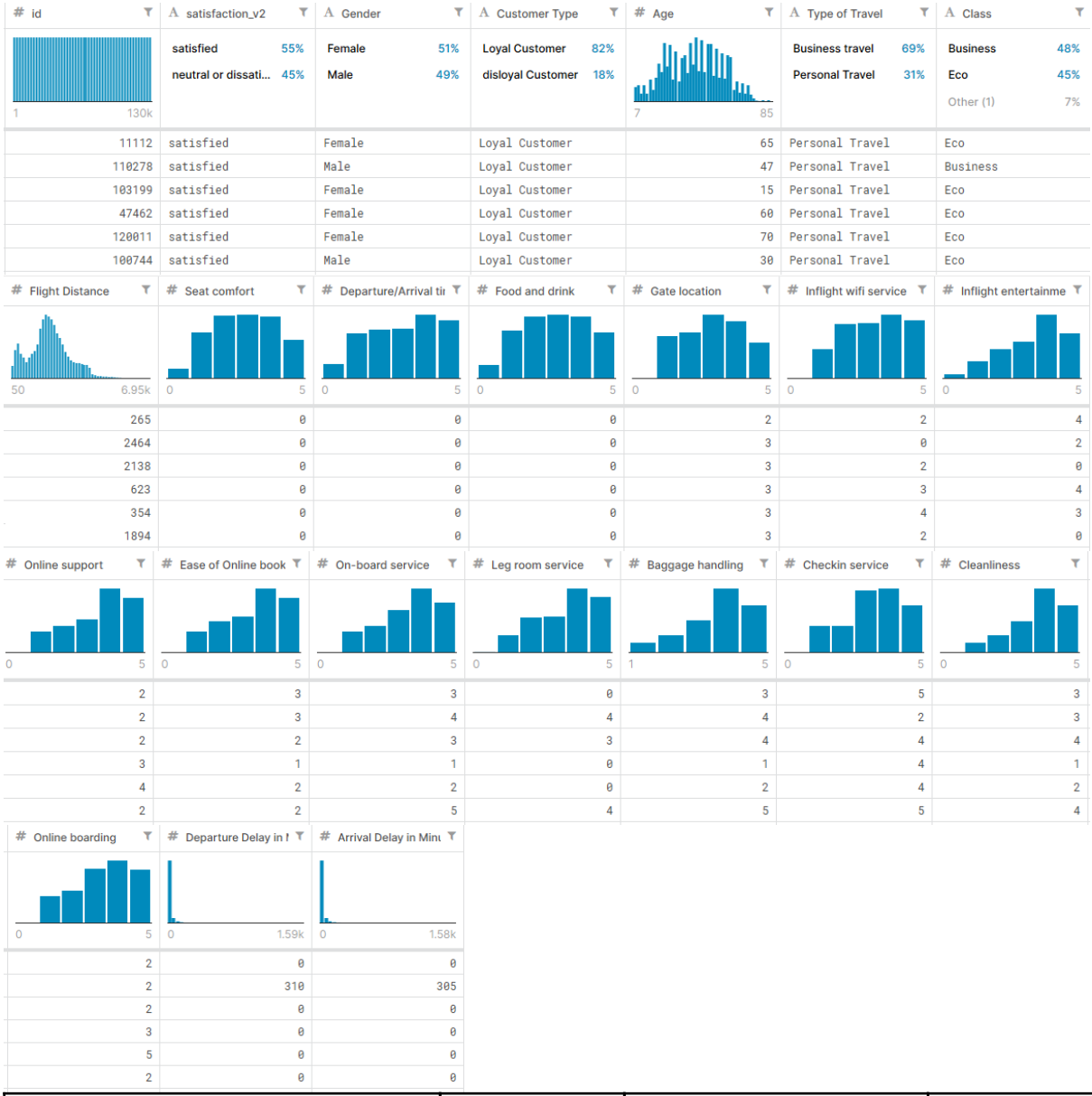
3. Analysis and Solution of the Problem

3.1 Data Preprocessing

- Convert string-valued and categorical columns to numerical .
- Striping off records with N/A values.
- Drop features that are poorly correlated with satisfaction. The important features are (Customer Type, Age, Baggage Handling, Food and drink, Seat comfort, Leg room service, Online boarding, Online support, In-flight Wi-Fi service, In-flight entertainment, On board service, Check-in service, Cleanliness, Ease of Online booking).
- Removed continuous features (when building the Naïve Bayes model).
- Before Association Rule Mining, we expanded some features to be able to convert the dataset to transactions to apply apriori algorithm. (e.g. Online_support becomes BadOnline_support, GoodOnline_support, and ExcellentOnline_support)

3.2 Data visualization

The dataset has 129,880 records with 24 columns.



id	satisfaction_v2	Gender	Customer Type	Age	Type of Travel	Class
11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco
110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business
103199	satisfied	Female	Loyal Customer	15	Personal Travel	Eco
47462	satisfied	Female	Loyal Customer	60	Personal Travel	Eco
120011	satisfied	Female	Loyal Customer	70	Personal Travel	Eco
100744	satisfied	Male	Loyal Customer	30	Personal Travel	Eco

Flight Distance	Seat comfort	Departure/Arrival ti	Food and drink	Gate location	Inflight wifi service	Inflight entertainme
265	0	0	0	2	2	4
2464	0	0	0	3	0	2
2138	0	0	0	3	2	0
623	0	0	0	3	3	4
354	0	0	0	3	4	3
1894	0	0	0	3	2	0

Online support	Ease of Online book	On-board service	Leg room service	Baggage handling	Checkin service	Cleanliness
2	3	3	0	3	5	3
2	3	4	4	4	2	3
2	2	3	3	4	4	4
3	1	1	0	1	4	1
4	2	2	0	2	4	2
2	2	5	4	5	5	4

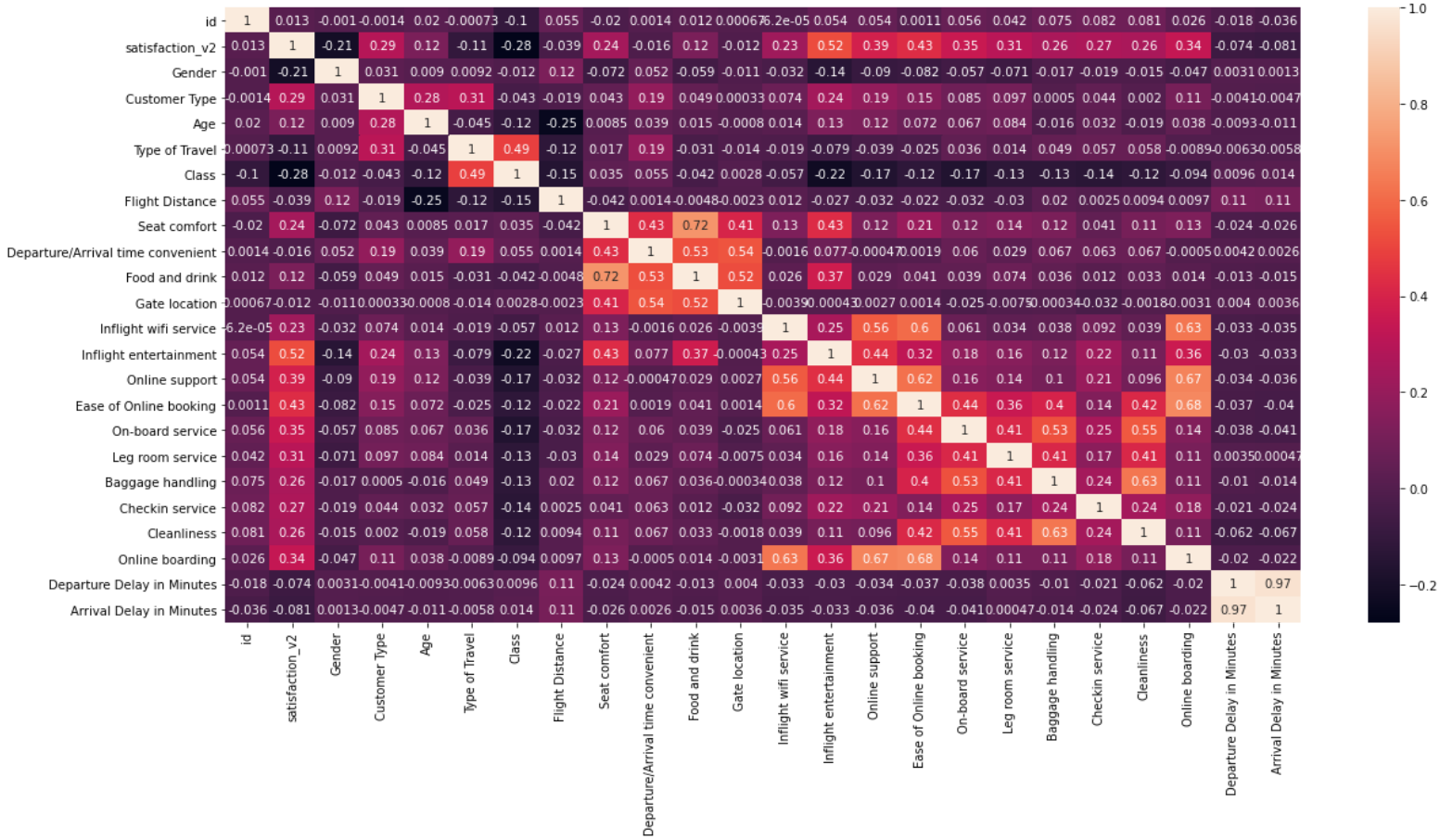
Online boarding	Departure Delay in I	Arrival Delay in Minu
2	0	0
2	310	305
2	0	0
3	0	0
5	0	0
2	0	0

Feature	Type	Feature	Type
id	*Continuous*	Inflight wifi service	Categorical
Satisfaction_v2	Categorical	Inflight entertainment	Categorical
Gender	Categorical	Online support	Categorical
Customer Type	Categorical	Ease of online booking	Categorical
Age	*Continuous*	On-board service	Categorical
Type of Travel	Categorical	Leg room service	Categorical
Class	Categorical	Baggage handling	Categorical
Flight Distance	*Continuous*	Check-in service	Categorical
Seat Comfort	Categorical	Cleanliness	Categorical
Departure/Arrival time convenient	Categorical	Online boarding	Categorical
Food and drink	Categorical	Departure Delay	*Continuous*
Gate location	Categorical	Arrival Delay	*Continuous*

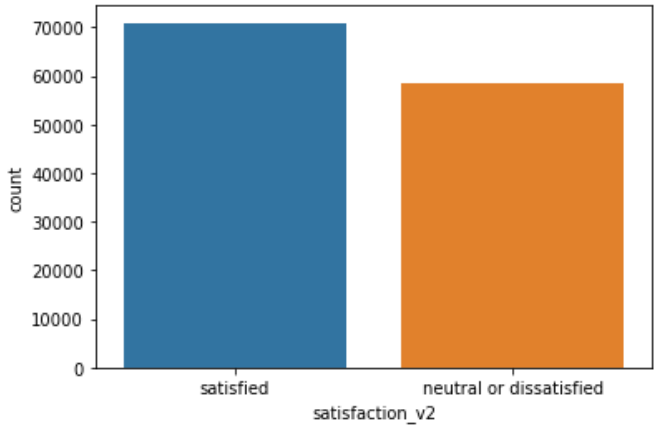
This an explanation about this dataset

- **Satisfaction:** Airline satisfaction level (Satisfied, neutral or dissatisfied)
- **Age:** The actual age of the passengers
- **Gender:** Gender of the passengers (Male, Female)
- **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)
- **Customer Type:** The customer type (Loyal customer, disloyal customer)
- **Flight distance:** The flight distance of this journey
- **Inflight wifi service:** Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
- **Ease of Online booking:** Satisfaction level of online booking
- **Inflight service:** Satisfaction level of inflight service
- **Online boarding:** Satisfaction level of online boarding
- **Inflight entertainment:** Satisfaction level of inflight entertainment
- **Food and drink:** Satisfaction level of Food and drink
- **Seat comfort:** Satisfaction level of Seat comfort
- **On-board service:** Satisfaction level of On-board service
- **Leg room service:** Satisfaction level of Leg room service
- **Departure/Arrival time convenient:** Satisfaction level of Departure/Arrival time convenient
- **Baggage handling:** Satisfaction level of baggage handling
- **Gate location:** Satisfaction level of Gate location
- **Cleanliness:** Satisfaction level of Cleanliness
- **Check-in service:** Satisfaction level of Check-in service
- **Departure Delay in Minutes:** Minutes delayed when departure
- **Arrival Delay in Minutes:** Minutes delayed when Arrival

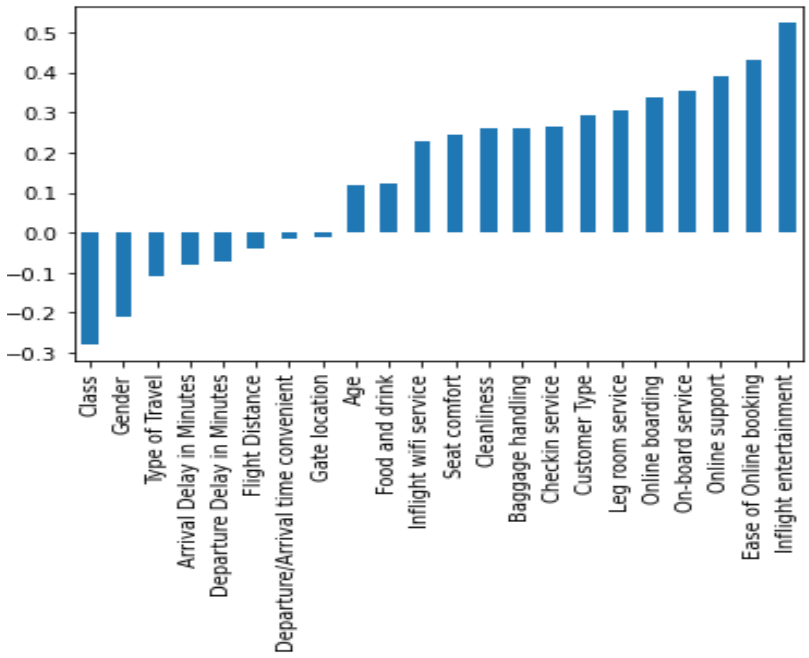
Correlation Matrix:



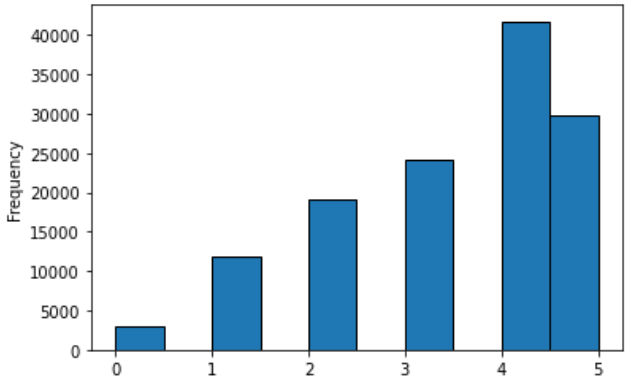
Count of satisfied and dissatisfied passengers from the dataset: 70882 vs 58605



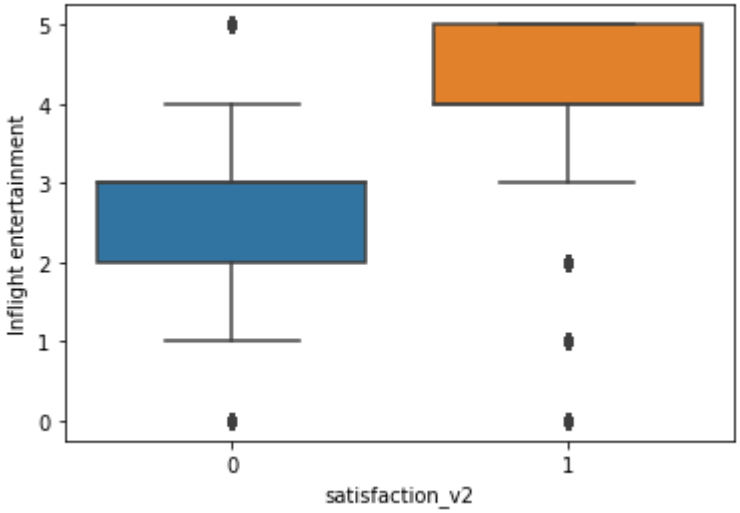
Sorted histogram of the correlated columns with satisfaction: (excluding id and satisfaction_v2)



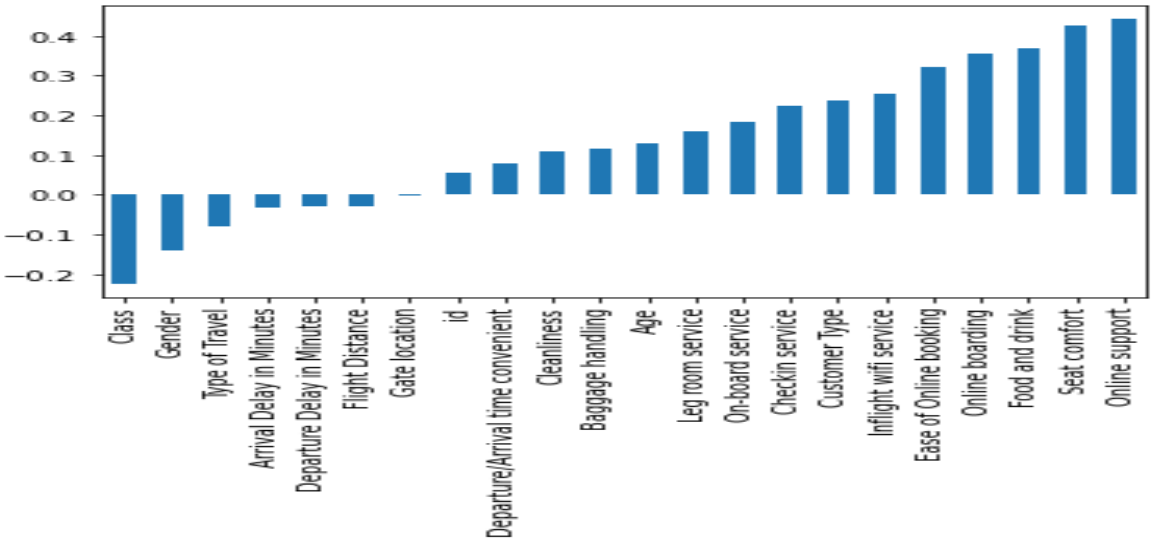
Exploring the most correlated column (<u>Inflight entertainment</u>): (41752, 29748, 24133, 19118, 11768, 2968) counts for rates (4, 5, 3, 2, 1, 0) respectively



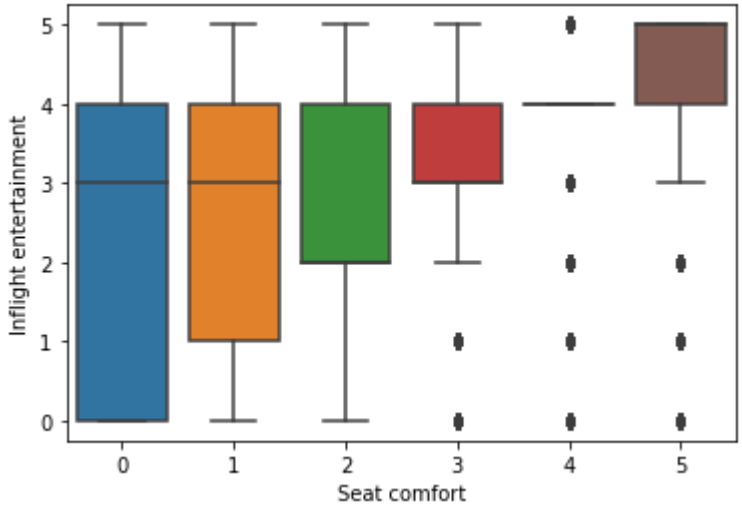
The more satisfied the person is with Inflight entertainment, the greater the chance that the person will be satisfied. The same applies for all the other correlated features.



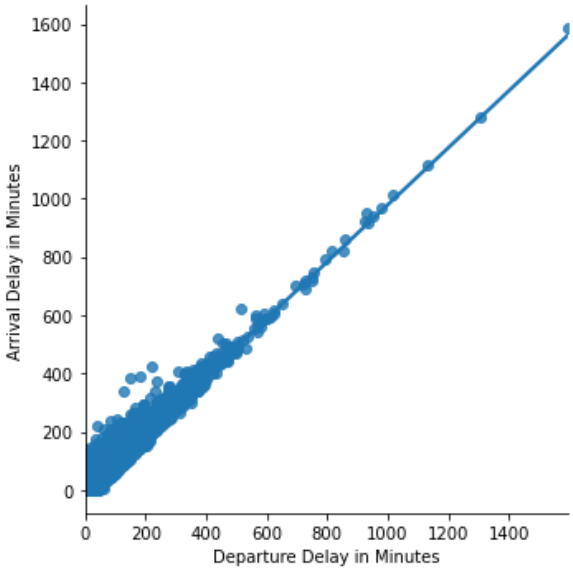
The correlation of Inflight entertainment with others:



Relation between Inflight entertainment and Seat Comfort which are highly correlated:



Relation between Departure delay and Arrival delay:



3.3 Extracting insights from data

- The most relevant factors (in descending order) to passenger satisfaction within the airline's control are:
    - In-flight Entertainment
    - Ease of online booking
    - Online support
    - On-board service
    - Online boarding
    - Leg room
    - Check-in service
    - Cleanliness
    - Baggage handling
    - Seat Comfort
    - In-flight Wi-Fi service
    - Food and drink
- People who get better seat comfort are likely to enjoy their inflight entertainment.
- Somewhat surprisingly (at least to us), ticket class doesn't affect satisfaction. Our assumption is that the expectations of passengers for each class are being met.
- Departure delay and Arrival delay have a linear relation, so we can remove one of them when predicting the passenger's satisfaction.
- Association Rule Mining: Most notably, passengers were more likely to rate the entertainment badly when the flights were longer.

3.4 Model building and training

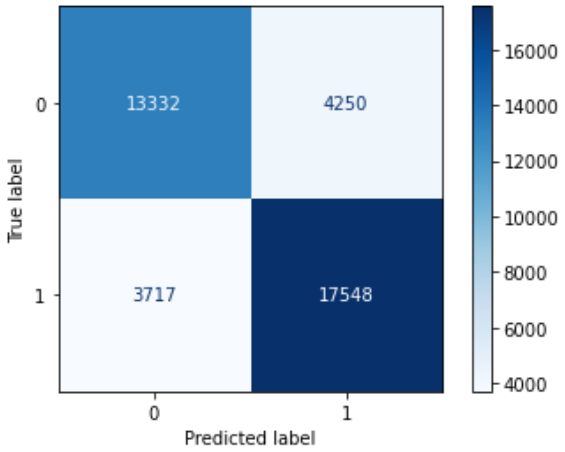
We tried 6 algorithms to train models to predict the satisfaction of the passengers using a set of highly correlated features with passenger's satisfaction:

1. Naïve Bayes: the fastest but the less accurate, need to remove the continuous features(e.g. Age)
2. Random Forest: the more accurate, medium speed
3. Decision Tree: the slowest (using max_depth=10 is the best in both time & accuracy after some trials) but it's the second in accuracy after Random Forest
4. K-Nearest Neighbours: K = 5, slow, feature engineering makes it really better (from 69% to 87% accuracy)
5. Logistic Regression: the second less accurate with average speed
6. Gradient Boosting: moderate in both accuracy and speed

4. Results and Evaluation

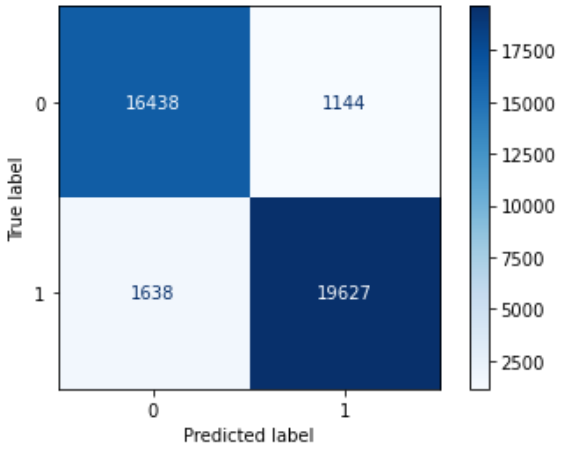
4.1 Naïve Bayes

Training set Accuracy: 79.63%
Test set Accuracy: 79.49%
Confusion Matrix:



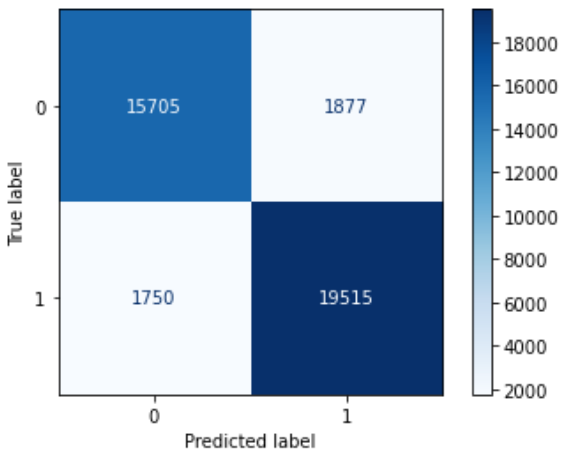
4.2 Random Forest

Training set Accuracy: 99.42%
Test set Accuracy: 92.84%
Confusion Matrix:



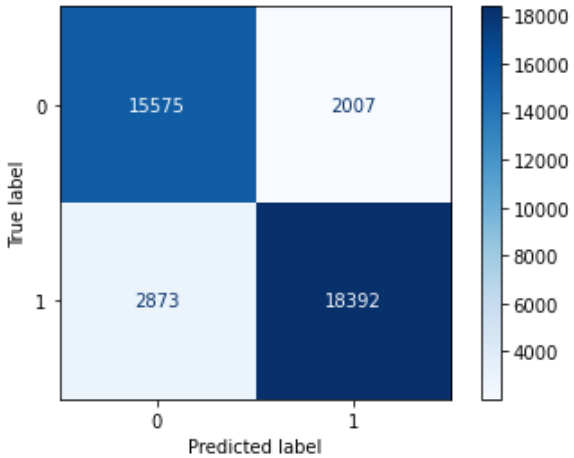
4.3 Decision Tree

Training set Accuracy: 91.50%
Test set Accuracy: 90.66%
Confusion Matrix:



4.4 K-Nearest Neighbours

Training set Accuracy: 91.29%
Test set Accuracy: 87.44%
Confusion Matrix:

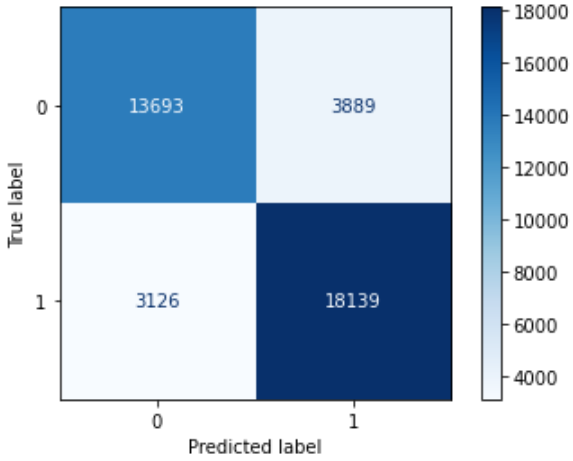


4.5 Logistic Regression

Training set Accuracy: 82.04%

Test set Accuracy: 81.94%

Confusion Matrix:

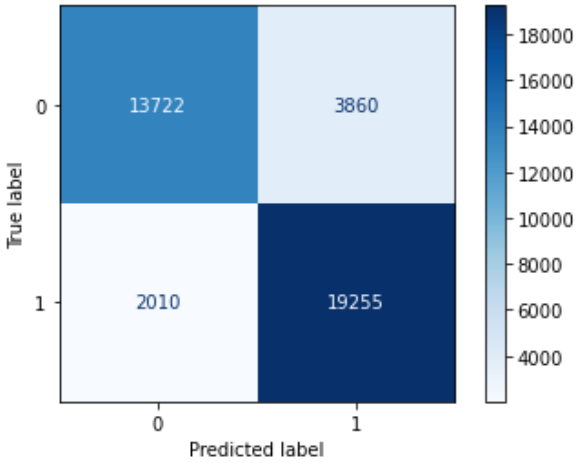


4.6 Gradient Boosting

Training set Accuracy: 85.07%

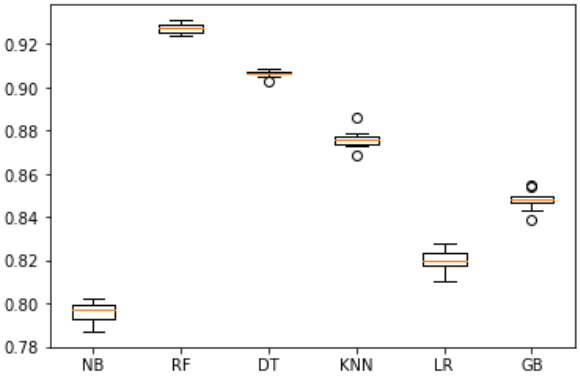
Test set Accuracy: 84.89%

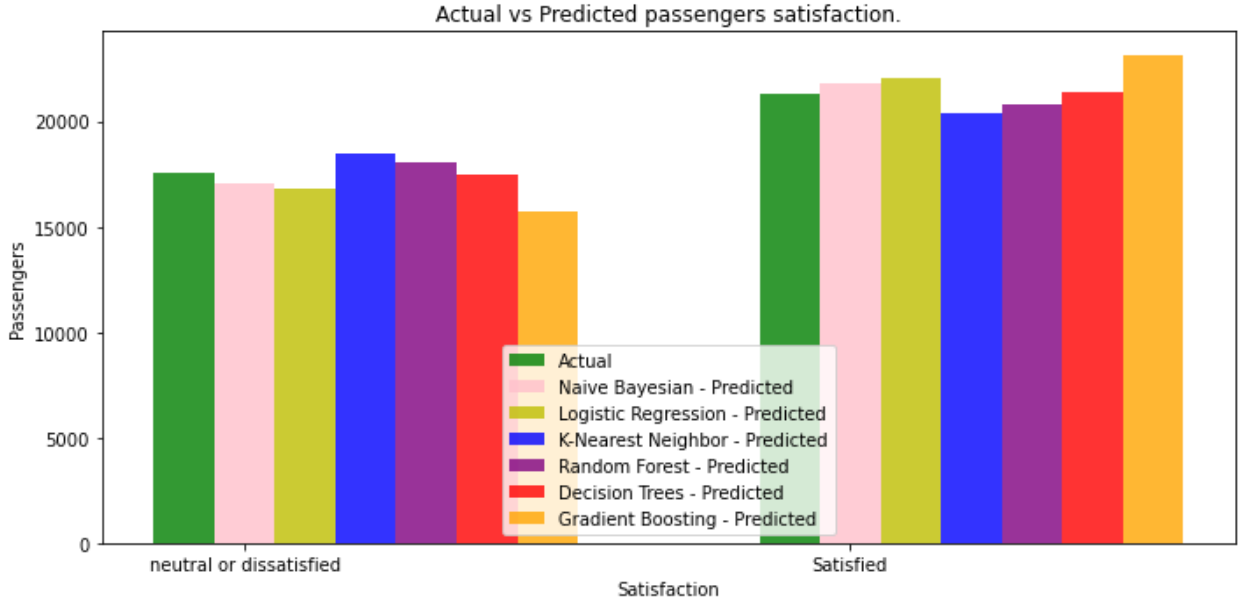
Confusion Matrix:



4.7 Algorithm Comparison

Accuracy comparison:





5. Unsuccessful trials

5.1 Association Rule Mining

- We tried to add additional features in the association rule mining phase but some lead to redundant rules such as (Personal_Travel and Business_Travel or Type_of_Travel values) and some lead to a lot of processing time because of the size of dataset and the number of columns (variables or features).
- We tried to make min_support = 0.1 but it sometimes doesn't return any rules if min_confidence > 0.5.
- We tried more than 5 combinations of (min_support, min_confidence, min_lift) to decide their current values, based on the time (some continued to more than hour and didn't finished yet!) and the result rules size and usefulness (when we included Type_of_Travel values with Class values [Eco, EcoPlus, Business], it leads to a lot of redundant rules (many versions of Eco_Class => Personal_Travel rules))

5.2 Models

- In KNN model, we tried to make the weight function to be 'distance' instead of the default 'uniform' and it leads to 100% training accuracy but it doesn't change the testing accuracy at all.
- We tried to include other variables in building the model without making feature engineering but we found that feature engineering results in more accurate predictions.
- Removing continuous features before building Naïve Bayes model results in better accuracy.
- Stratifying the data while splitting results in better accuracies.

6. Future work

A possible improvement we considered is decreasing the size of the dataset by doing Dichotomization (subsetting) and evaluating the results.