Contents

I. Introduction

In recent years, airline passenger satisfaction has become increasingly vital with the rise of competitors in the industry. It has become important for airline companies to pursue and preserve customer loyalty to differentiate themselves among other competitors. In this report, I will use machine learning algorithms to predict customer's satisfaction based on a variety of attributes.

The study's aim is to develop an accurate and reliable binary classification machine learning model that can identify and forecast essential characteristics that have an impact on customer satisfaction.

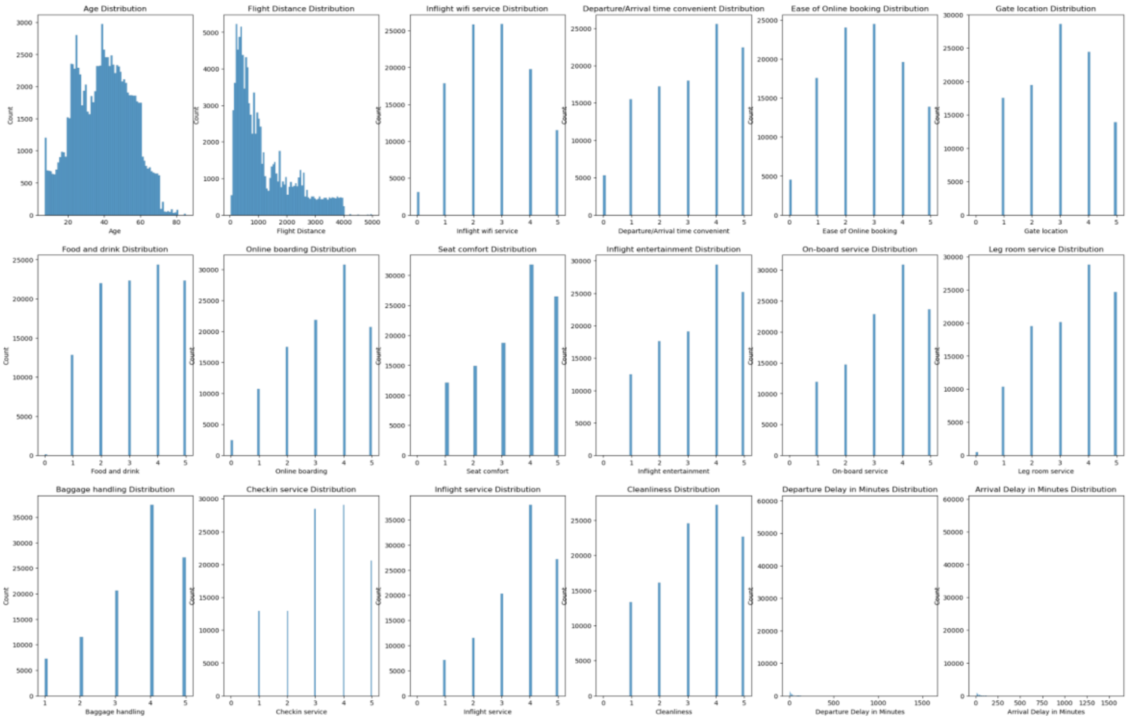
I then build machine learning models (CatBoost, SVM) to predict customer satisfaction based on a range of variables. The performance is then evaluated based on each model's accuracy score and the speed of training and prediction.

II. Exploratory Data analysis

II.1 Distribution of features and classes

Our target class is 'Satisfaction' which indicates if the passenger is satisfied ('satisfied') or not satisfied ('neutral or dissatisfied').There is slight imbalance in our target class where 56.67% passengers are neutral or dissatisfied and 43.33% are satisfied.

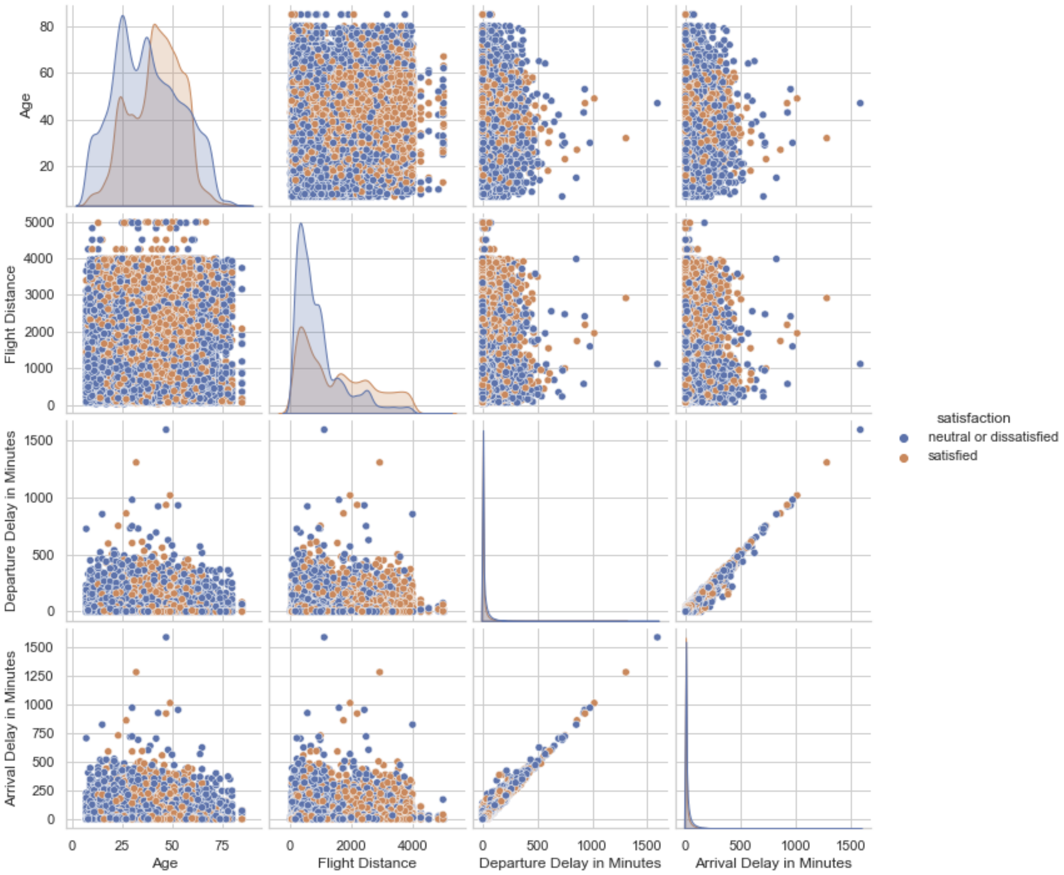
In our data, our categorical features are: 'Gender', 'Customer Type', 'Type of Travel', 'Class' while the rest are numerical.



As we can infer from the plots above, the most popular age range in our dataset is 20-50. The flight distance is right-skewed where distances are mostly less than 1500. For most features, the most frequent rate is at level 4, with the exception of “Inflight wifi service”, “Ease of online booking”,”Gate location”.

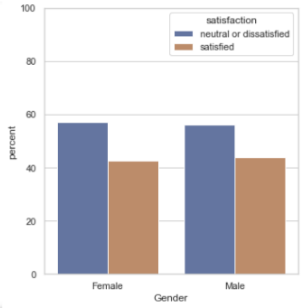
II.2 Distribution of features vs target (“Satisfaction”)

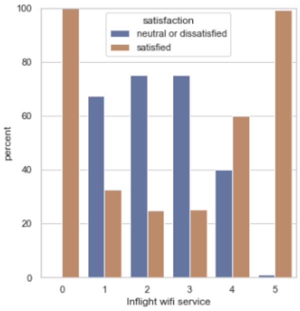
Understanding of the target distribution is extremely important to define which features will be used in our model. A detailed analysis on its distribution among different features is presented below.

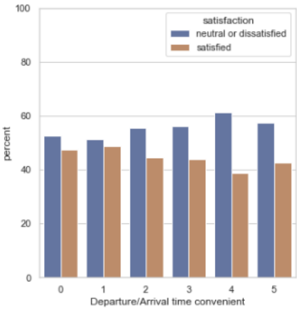


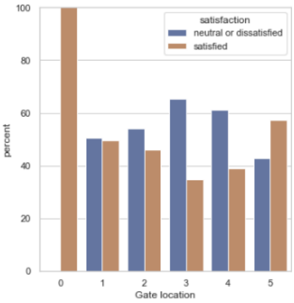
For the continuous variables, “Departure Delay in Minutes” is strongly correlated with “Arrival Delay in Minutes”. This can be interpreted as feature redundancy.

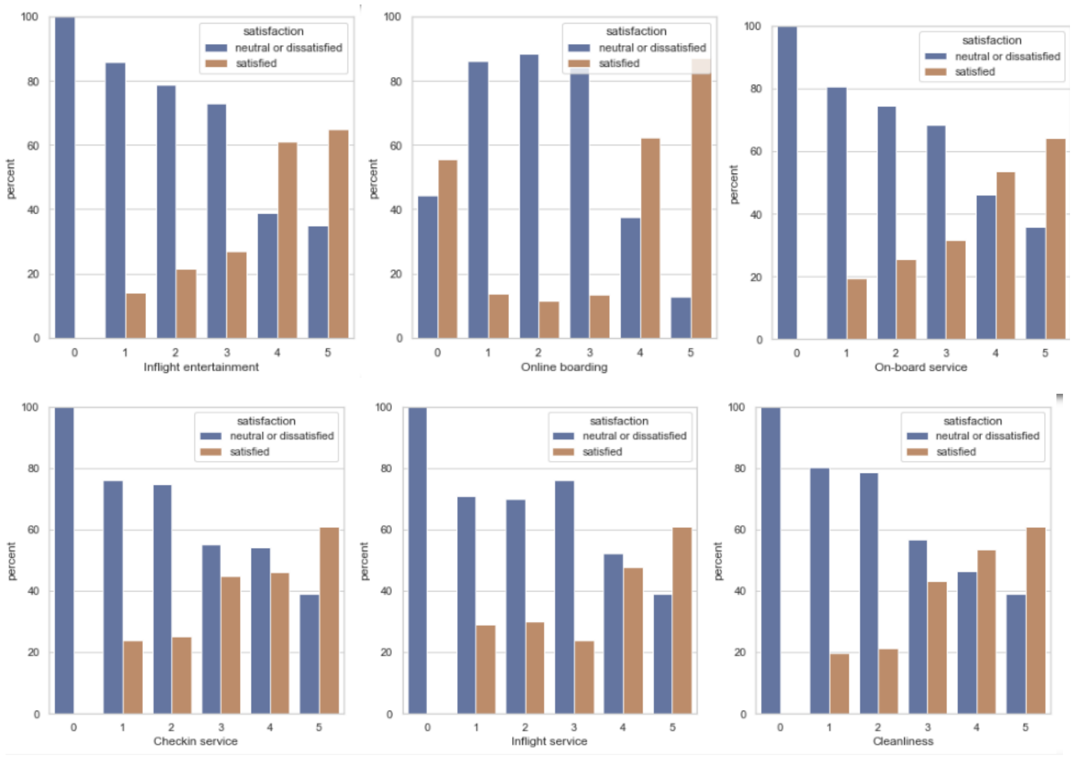
Age and Flight distance distribution are fairly normally distributed, but no different distribution when compared “satisfied” vs “neutral or dissatisfied” customers.







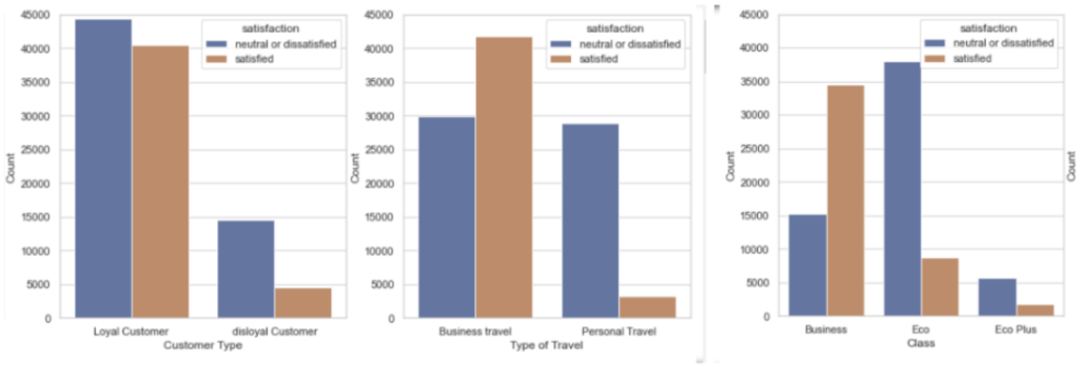




From the relative comparison of satisfaction, we can infer that:

- Gender, Departure/ Arrival time convenient, Gate Location have low correlation with the target “Satisfaction”
- Inflight wifi service, Online Boarding, Seat Comfort, Inflight entertainment, On-board service, checkin service, inflight service and cleanliness have good potential as predictors of our target.

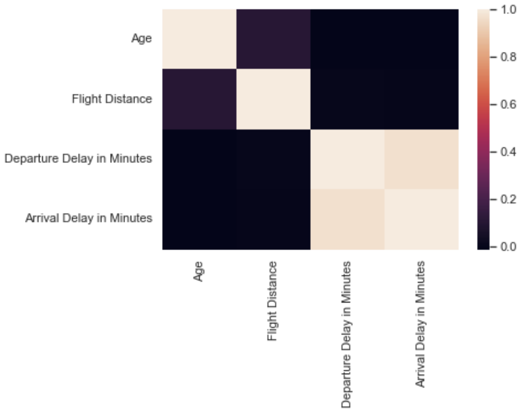
We also take into account that the data can also be skewed. Therefore, selected data in absolute values (occurrences) are presented below:

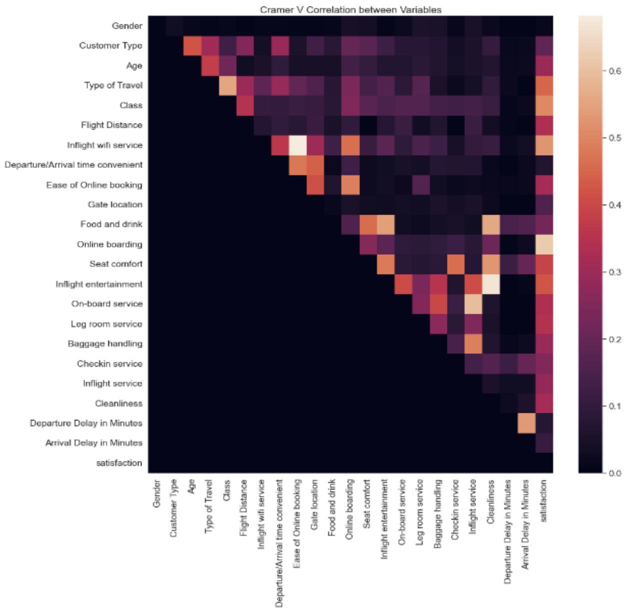


Observations:

- Most of the customers are loyal customers.
- Business travellers have a much higher rate of satisfaction than economy class travellers and around 75% of satisfied customers flew business class.
- Most of the trip's purpose was business travelling, which may correlate with the fact that most travellers flew business class.

II.3 Correlation





Observations:

- The correlation graphs (both numerical and categorical data presented on figure 7) confirm the suspicion from the pairwise analysis (bar charts and pairplot charts presented on figure 4 and 5).
- All other features present relatively low correlation to each other and may be useful for our final model.

II.4 Features Dropped and justification

The features “Arrival Delay in Minutes” will be dropped based on the fact that it is redundant and it also presented some missing values. By dropping this feature, it is possible to solve both these problems.

III. Methodology

Model used: CatBoost, KNN, SVM, Logistic Regression

1. Categorical variables (Customer type, Type of Travel, and Class) are converted into dummy variables (ones and zeroes).
2. The unnecessary features are then dropped based on section II.4, and the data is standardised.
3. SVM uses a Radial Basis Function (RBF) kernel with a default regularisation parameter (C = 1).

IV. Results

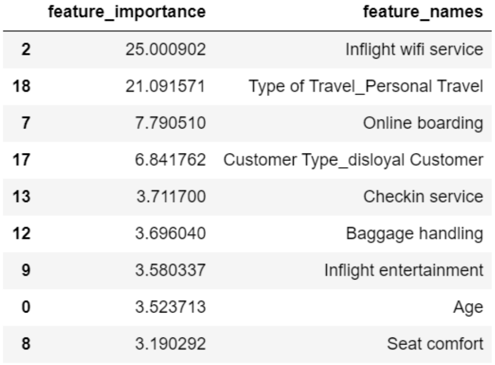
	SVM	Catboost
Accuracy on validation set	0.956	0.964
Accuracy on test set	0.955	0.963
R2_score	0.816	0.85

Catboost has the highest accuracy in both the validation and test set, however SVM also has very good results with very slight differences in these metrics.

The R2_score are both higher than 0.8 which indicates great variability is explained by the model.

V. Discussion

Top factors affecting passengers' satisfaction extracted from feature_importance method. Besides the factors airlines cannot control, the important features are: Inflight wifi service, Online boarding, checkin service, baggage handling and inflight entertainment.



	feature_importance	feature_names
2	25.000902	Inflight wifi service
18	21.091571	Type of Travel_Personal Travel
7	7.790510	Online boarding
17	6.841762	Customer Type_disloyal Customer
13	3.711700	Checkin service
12	3.696040	Baggage handling
9	3.580337	Inflight entertainment
0	3.523713	Age
8	3.190292	Seat comfort

Overall, I suggest that inflight wifi service and online boarding be invested in, in order to increase passenger satisfaction.

VI. Conclusion

The goal of our project was to develop a reliable machine learning classification model that could predict customer satisfaction based on data representing their flight experience. This was achieved through building multiple models each with its own strengths and drawbacks to determine an optimal classification model.

Key areas of importance included the data analysis stage to discover experience features that had the greatest impact on overall customer satisfaction. The most important features with high correlation to customer satisfaction but with the lowest inter-correlation with other features included: Inflight Wifi service, Online boarding, Check-in service, baggage handling, inflight entertainment.

Areas for further research could include improving quality of data in the preprocessing stage and feature selection to uncover hidden correlation between features that were dropped and the target value. Additionally, the potential use of ensemble methods to combine models and arrive at a conclusion based on majority ruling (out of the 5 models) could also improve accuracy.

VII. References

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