**Contents**

1. **Introduction**
2. **Exploratory data analysis**
3. **Distribution of features and classes**
4. **Distribution of features vs target (“Satisfaction”)**
5. **Correlation**
6. **Features dropped and justification**
7. **Methodology**
8. **General overview**
9. **Model specifications**
10. **Results**
11. **Discussion**
12. **Conclusion**
13. **References**
14. **Introduction**

In recent years, airline passenger satisfaction has become increasingly vital with the rise of competitors in the industry. It has become crucial for airline companies to pursue and preserve customer loyalty to differentiate themselves among other competitors. In this report, we 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. We begin by analysing the dataset, observing distributions in the features. Additionally, we performed feature selection by eliminating features with low correlation to the target variable.­­­­­­ We then build five different machine learning models, including XGBoost, CatBoost, SVM, Logistic Regression, and KNN. The performance is then evaluated based on each model’s accuracy score and the speed of training and prediction.

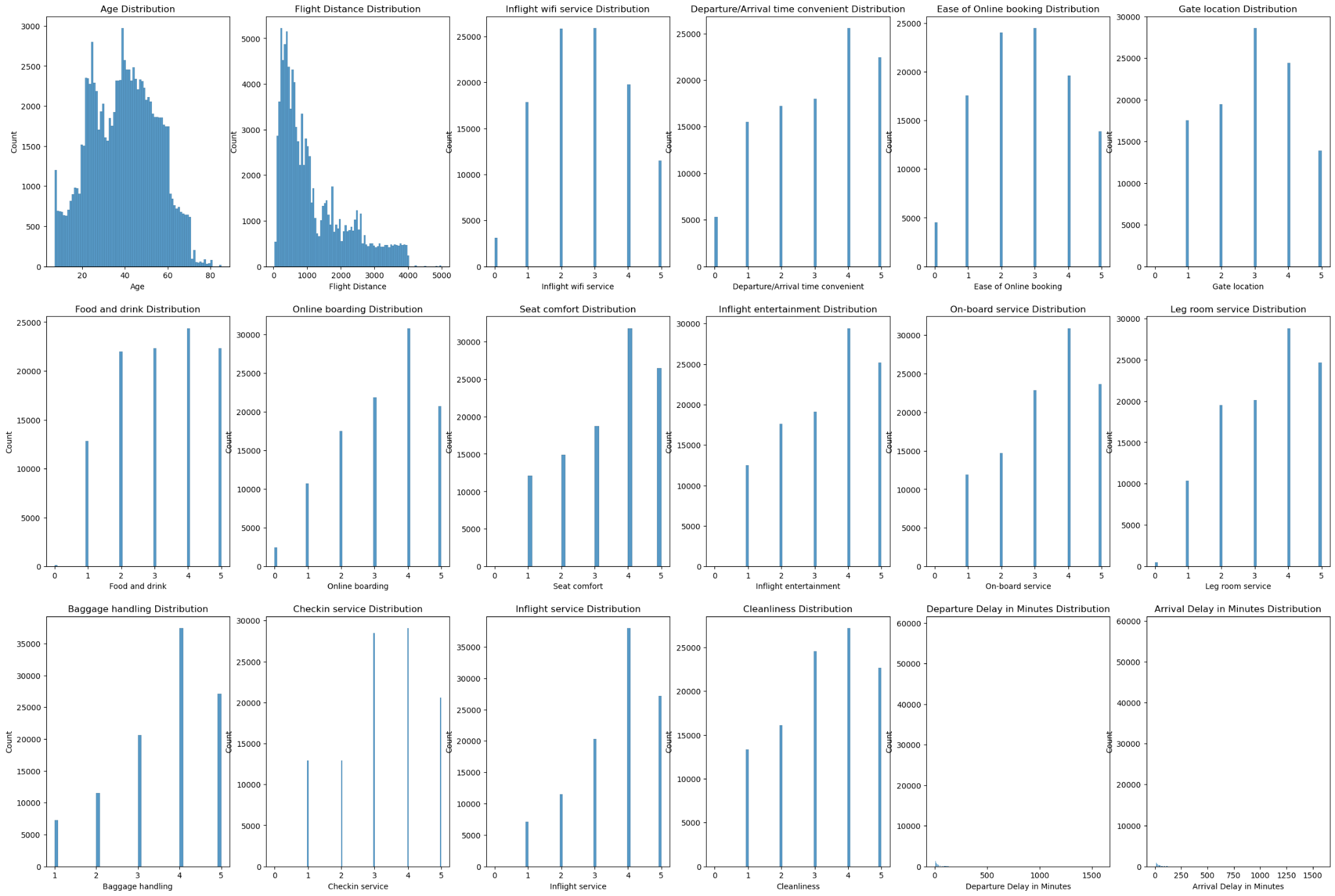
Finally, we present our results and discuss the key findings and major conclusions, including the features used and dropped, the best performing models, and applications. This study provides insight and suggestions for airline firms to focus on specific elements that could significantly boost customer’s satisfaction.

1. **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.

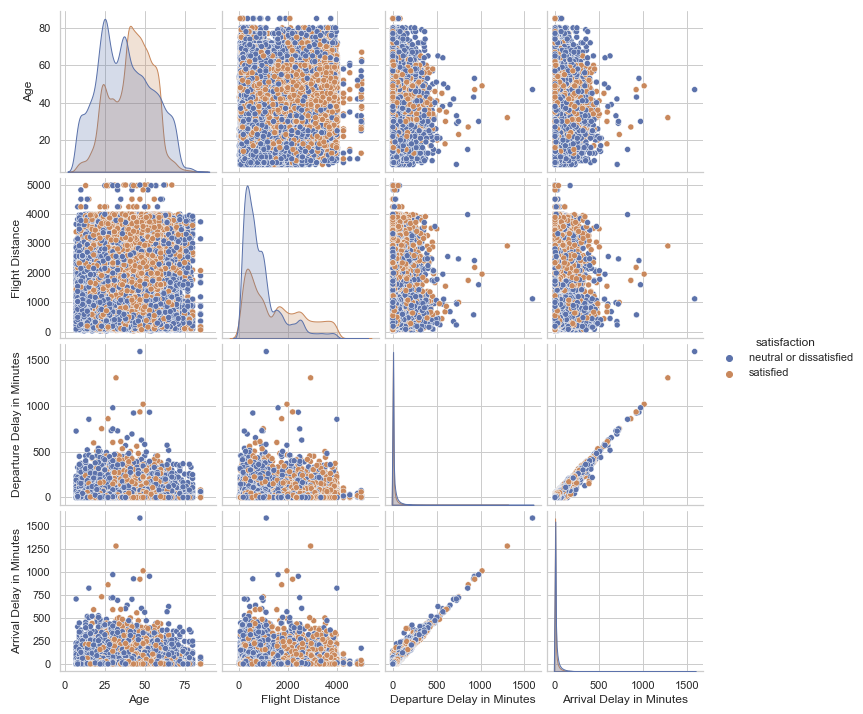
In our data, our categorical features are: ‘Gender’, ’Customer Type’, ’Type of Travel’, ’Class’ while the rest are numerical.



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”.

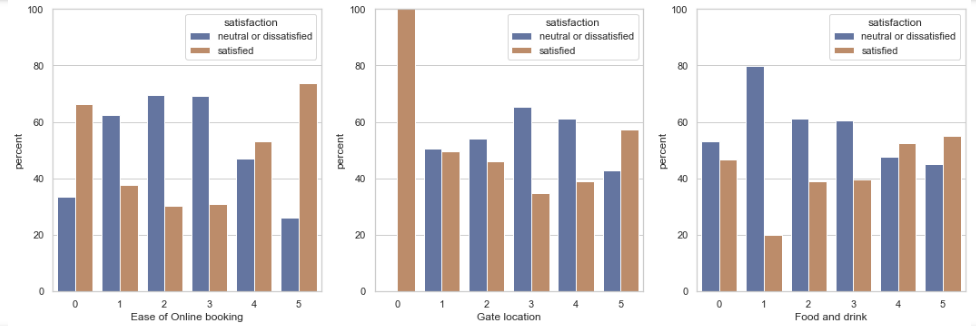
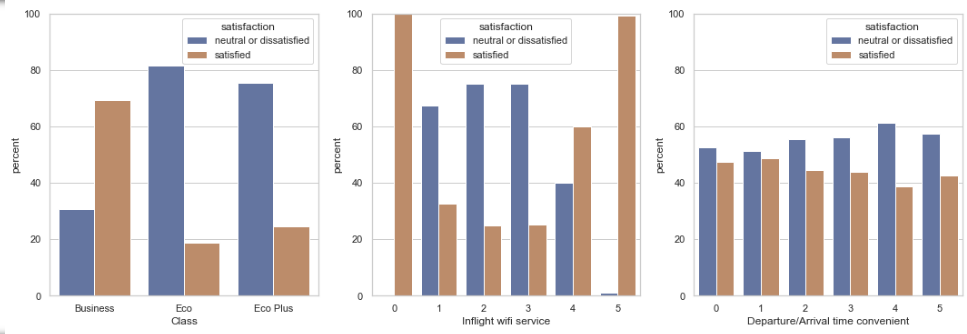
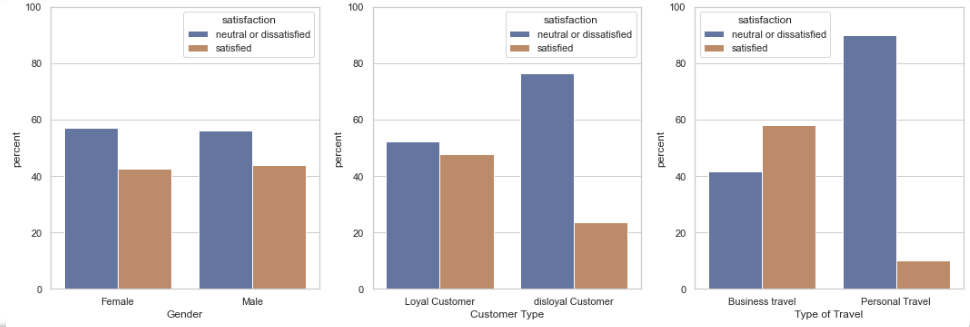
*II.2 Distribution of features vs target (“Satisfaction”)*

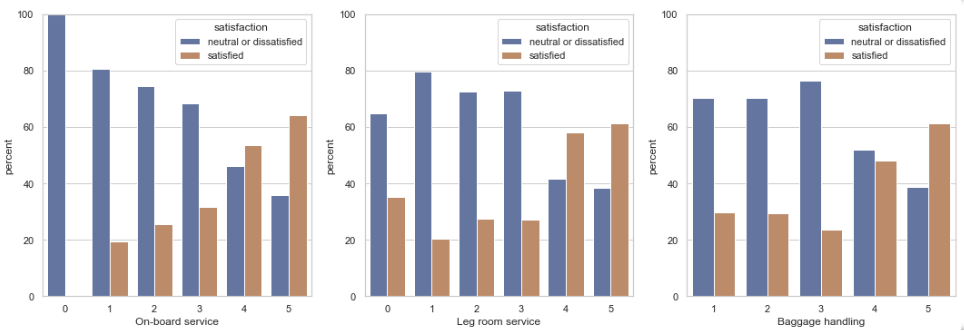
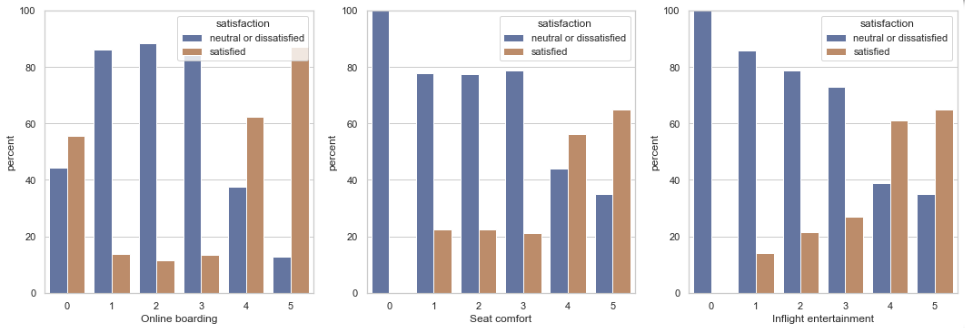
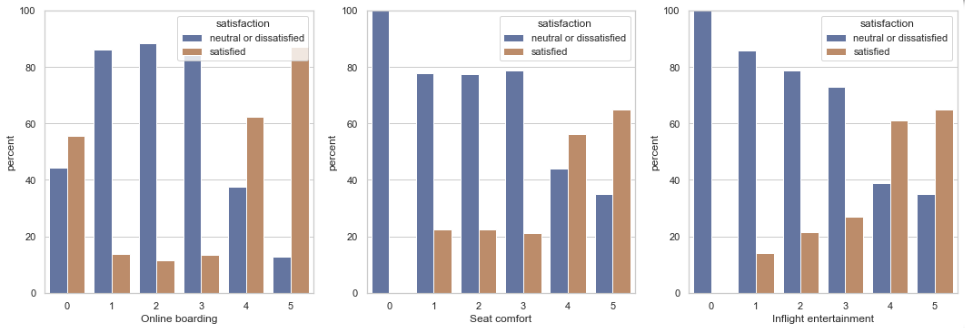
Understanding of the target distribution is extremely important to define which features will be used in our model. A detailed analysis on its distribution among different features is presented below.

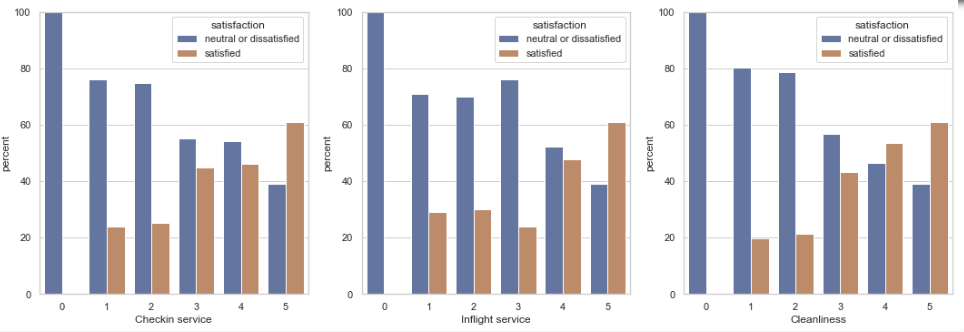


For the continuous variables, “Departure Delay in Minutes” is strongly correlated with “Arrival Delay in Minutes”. This can be interpreted as feature redundancy.

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



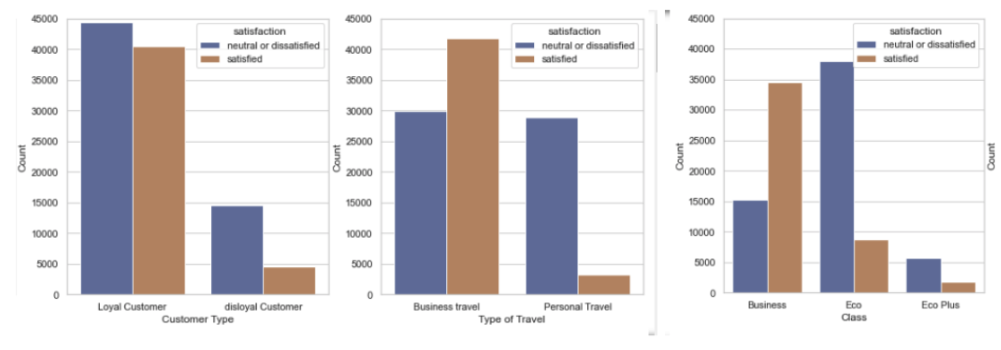




From the relative comparison of satisfaction, we can infer that:

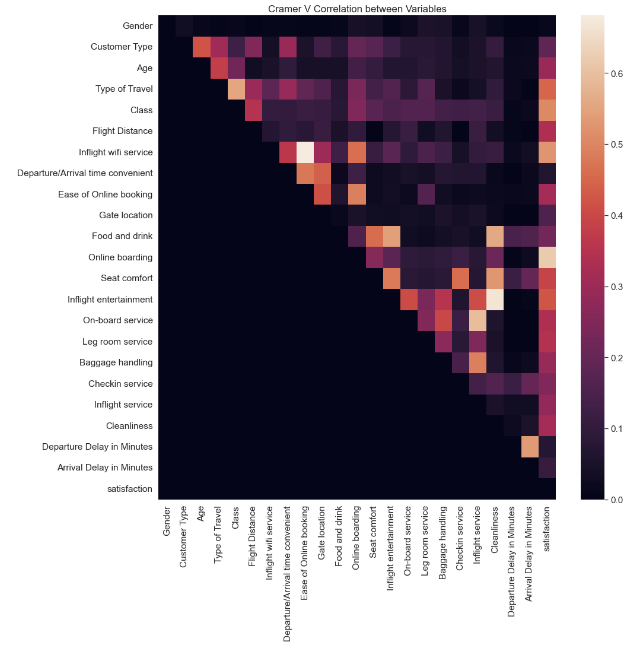
* Gender, Departure/ Arrival time convenient, Gate Location have low correlation with the target “Satisfaction”
* Inflight wifi service, Online Boarding, Seat Comfort, Inflight entertainment, On-board service, checkin service, inflight service and cleanliness have good potential as predictors of our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

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

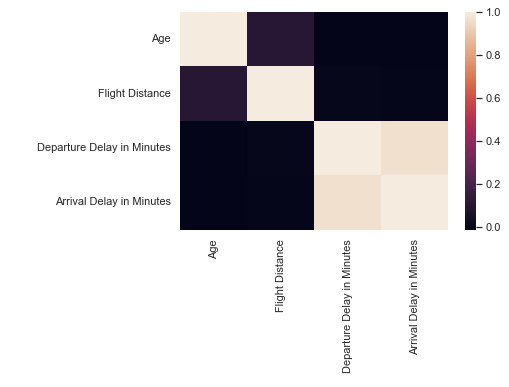


Observations:

* 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.

1. **Methodology**

*III.1 General overview:*

We divided the modelling process into 3 levels:

(1) Baseline: We used all features in the dataset besides “Arrival Delay in Minutes” (justified above).

(2) Feature Selection: We use correlation and heatmap to analyse correlation between features and target variable, we decided to drop columns: ‘Gender’,’Gate Location’, ‘Departure/Arrival time convenient’ and build our level 2 model

(3) Adding Features: We add 2 features: ‘Age Group’ (Kids: 7-13, Teen: 13-20, Young Adult: 20-30, Middle Age: 30-55, Senior: 55-85), ‘Flight Dist’ (Avg: 31-1000, Long: 1000-3000, Very Long:3000-4983)

*III.2 Model-specific:*

***XGBoost***

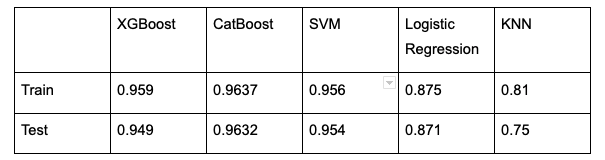
The following steps were used to build the model:

1. Data Pre-processing: Features drop presented on section II.4 only (not required standardisation and normalisation).
2. Hyperparameters tuning approach:
   1. Find the ideal Learning Rate (step of the gradient descent) and the number of trees. ⇒ Dictate the training speed of our model and how many iterations (trees) will be performed.
   2. Tune XGB hyperparameters: gamma(minimum loss threshold for splitting a tree), max\_depth (Maximum depth of a tree), min\_child\_weight (Minimum sum of instance weight needed in a child), subsample (Subsample ratio of the training instances.), colsample\_bytree (subsample ratio of columns when constructing each tree). ⇒ They are the most important parameters as this model usually tends to overfit.
   3. Check for regularisation parameters: alpha (L1), lambda (L2).
3. Parameters definition process: there is no unique direction for all datasets on tuning hyperparameters. For that reason, a grid search, which iteratively compares models with most of the hyperparameters combinations, needs to be used.

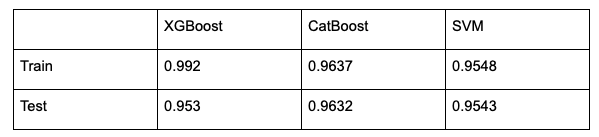
**CatBoost, KNN, SVM, Logistic Regression**

1. Categorical variables (Gender, 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).
4. **Results**

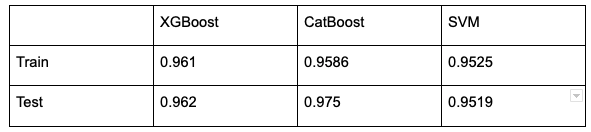
On our (1) Baseline: We ran 5 models - XGBoost, CatBoost, SVM, Logistic Regression and KNN. Since XGBoost, CatBoost, and SVM performed considerably better, we proceeded with the three models on the next processes.



On our (2) Feature Selection, we removed uncorrelated features with the target variables and obtain the following results:



Finally we continued with (3) Adding Features and got better results on Catboost but not SVM:



On the Baseline table, Catboost has the highest accuracy in both the validation and test set, both XGBoost and SVM have very similar results. The accuracy of XGboost on the Feature Selection table is much greater in the validation set compared to both CatBoost and SVM while the values of testing set accuracy have very small differences. Finally, in the Adding Features phase, XGBoost again, has the highest accuracy for the validation set and CatBoost has the highest accuracy in the test set. Hence, the final model we chose was XGBoost for its consistent accuracy and speed.

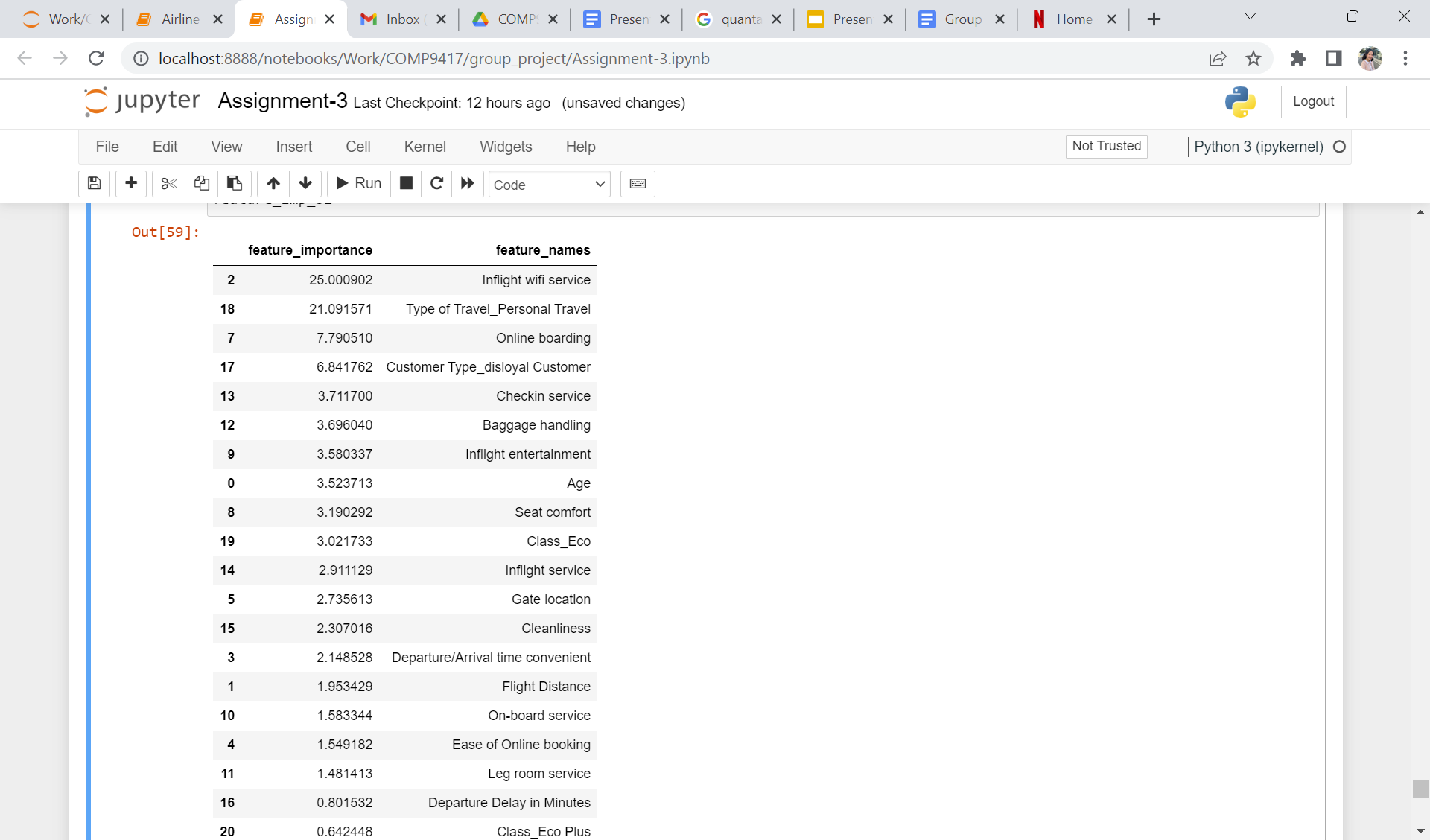
1. **Discussion**

*V.1 Comparison among models*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **XGBoost** | **CatBoost** | **SVM** | **Logistic Regression** | **KNN** |
| Features | Good interpretability of data, fairly fast training and easy to implement.  Robust to outliers, but easier to overfit if there is not enough available data. | | Difficult to interpret due to high and extremely slow training in high dimension space. Therefore feature creation and grid searching can be a problem for tuning the model. | Difficult to interpret due to high dimensional space.  Simple model, cannot “adapt” to predict more complex events. | Difficult to interpret and performs badly in high dimensional space (curse of dimensionality).  As dataset grows, becomes extremely slow (needs to compute distance to all points every prediction) |
| General Performance | Good Accuracy, Fast prediction.  But can easily overfit | | Good Accuracy,  Slow prediction. | Medium accuracy, Fast prediction | Medium accuracy, Extremely Slow prediction |

*V.2 Insights*

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.



We analysed feature importance for different groups of passengers based on their type of travel and class in order to gain insights into their factors of satisfaction in different groups. The top factors are as below (excluding factors airlines cannot control):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of Travel** | | **Class** | | |
| **Personal Travel** | **Business Travel** | **Eco** | **Eco Plus** | **Business** |
| - Inflight wifi service  - Ease of online booking  - Departure Delay  - Online boarding | - Inflight wifi service  - Online boarding  - Checkin service  - Seat comfort | - Inflight wifi service  - Online boarding  - Baggage handling | - Inflight wifi service  - Online boarding  - Baggage handling | - Inflight wifi service  - Online boarding  - Seat comfort |

Overall, we suggest that inflight wifi service and online boarding be invested in, in order to increase passenger satisfaction. In terms of segmentation, passengers who opt for “Business” options would care more about service and comfort while the remaining segmentation of passengers will put emphasis on convenience (on-board or check-in service, baggage handling).

1. **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. This proved true for different types of travel and classes of flight. Thus our studies suggested that inflight wifi service and online boarding are key areas of improvement.

Throughout our methodology stage for binary classification, the XGBoost, CatBoost and SVM proved to have the best results in binary classification (satisfaction | dissatisfaction) however the KNN model's results were based on a 4-class dataset (Satisfied, Neutral, Neutral-Dissatisfied, Dissatisfied) and proved to have reasonable results (75% accuracy on the test set).

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.

1. **References**

[**https://xgboost.readthedocs.io/en/stable/parameter.html**](https://xgboost.readthedocs.io/en/stable/parameter.html)

**Data source -** [**https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction**](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction)