**Machine Learning | CA1 | Louis Wilkie | sba22529**

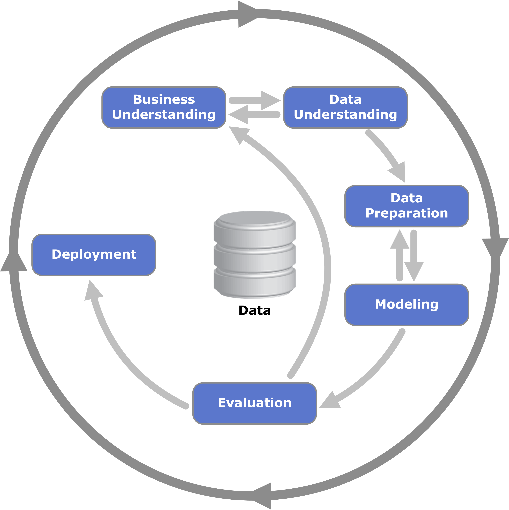
**Contents**

1. **Introduction**
2. **Results**
3. **Conclusion**
4. **References**

**1. Introduction**

This assignment will analyse an airline customer satisfaction dataset. Supervised learning machine learning models will be developed for this dataset. Different models will be built to classify (Harrison, p. 105) if an entry in the dataset is *satisfied* or *dissatisfied.*

This dataset fits under the umbrella of transport as per the assignment requirements. More specifically, the dataset contains anonymised survey responses on an anonymised airline as *Invistico Airlines.*

This assignment thus poises to answer the questions: *Can Airline Customer Satisfaction be modelled? Which classification techniques can best represent the data being analysed? Which attributes contribute most to customer satisfaction?*

This project follows the CRISP-DM framework, and the results below will be discussed through the lens of the CRISP-DM phases. This assignment has been completed by iterating through the CRISP-DM steps (Provost & Fawcett, p. 27), but the results will describe them chronologically.

Figure 1: CRISP-DM phases. source: Wikipedia

**2. Results**

**2.1 Business Understanding**

The business objectives of this project are given above as the questions the assignment will answer.

* Can Airline Customer Satisfaction be modelled with machine learning?

For this

* Which classification techniques can best represent the data?

A range of classification machine learning models will be performed on the data.

* Which attributes contribute most to customer satisfaction?

This is an important metric in understanding the data. It is important for model deployment to know which areas are related to customer satisfaction and therefor business strategies can be employed to increase customer satisfaction.

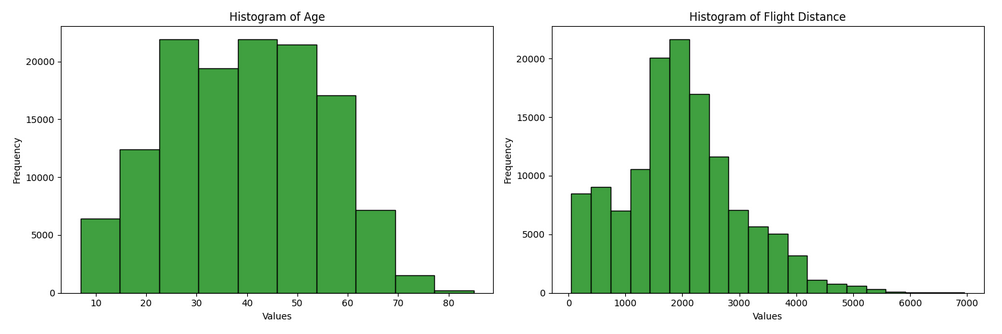
**2.2 Data Understanding**

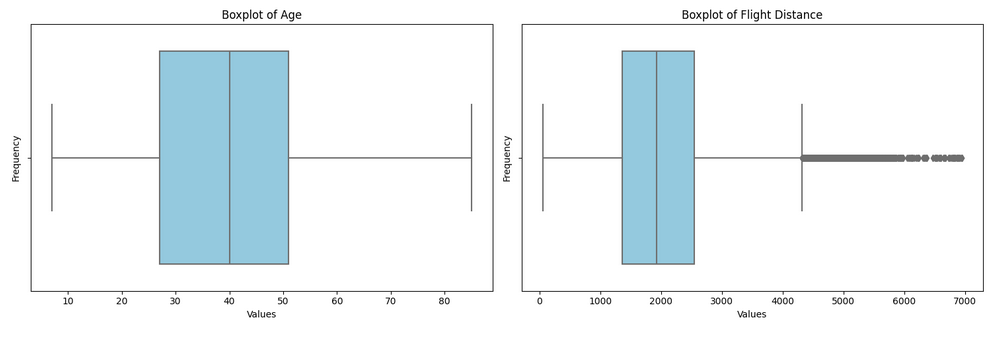
The dataset is titled “Airlines Customer Satisfaction” and is sourced from Kaggle.com, available here: <https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction>. There are 129800 entries and 23 features.

The dataset is survey responses featuring the following,

* questionnaire responses ranked 0-5
* age and gender of the respondent
* continuous variables of flight distance, departure delay, and arrival delay
* some categories describing the respondent

Plots of the continuous features are shown below. Age is normally distributed, while flight distance is skewed right.





**2.3 Data Preparation**

Data collected from surveys can face challenges from missing data (Fitzmaurice et. Al, p. 525). The *Arrival Delay in Minutes* feature had 393 null values, which amounted to 0.3% of the entries of that column. These rows were dropped from the data, which is a minor data loss, but the remaining sample size is large enough to make statistical inferences from. Removing observations with missing values can introduce bias (Gallatin & Albon, p. 85), but an imputation can introduce noise

Arrival and departure delays are highly correlated at 0.96. A new feature of flight time difference is engineered and arrival delay is dropped.

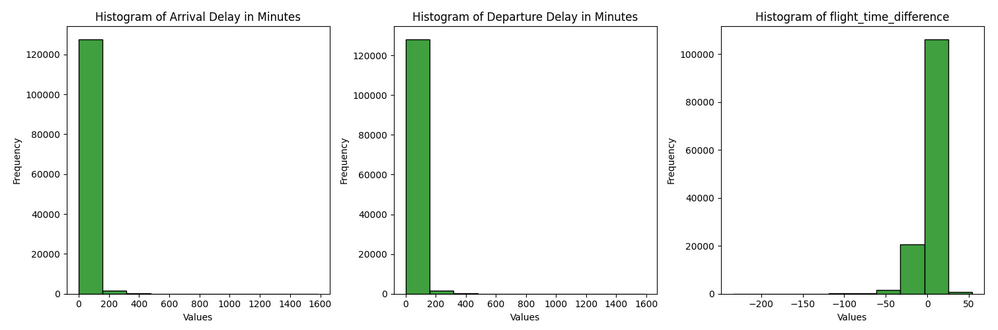
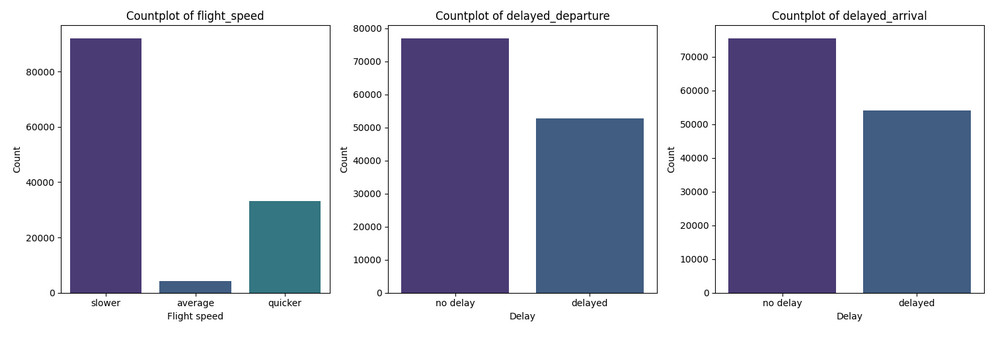


Figure : Signal captured in flight time difference

From these columns other feature can be extracted as shown below. Some variance should be expected in what is the ‘average’ flight speed, for modelling a strict definition is chosen.



The features *delayed\_departure* and *delayed\_arrival* have a significant relationship, but are not duplicated features as shown below.

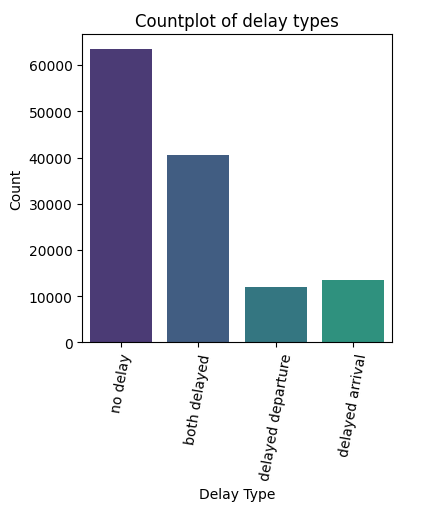


Figure : Countplot of delay types

The columns were renamed.

The ranked columns were encoded, the categories were encoded. Binary columns are label encoded to avoid multicollinearity.

Age is binned in decades.

Continuous features are scaled.

As there are many related features dimensionality reduction in the form of principal component analysis is applied. This is hoped to reduce the extent of the encoded questionnaire features.

**2.4 Modelling**

Binary classification is performed, which is a subset of supervised learning (Müller & Guido, p. 27). Several training test splits are examined; 90:10, 80:20, and 70:30. The table below shows the accuracy results from different training and testing split ratios.

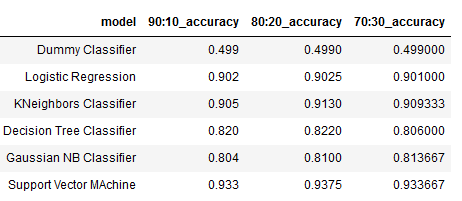


Figure : Accuracy results from training and test splits

The 90:10 split is the most accurate but it’s likely the models are being overfit.

Cross Validation is performed with the same models the results are below

**2.5 Evaluation**

Accuracy is one way for model evaluation.

Attributes that impact the model

**2.6 Deployment**

Actual deployment is outside of the scope of this assignment, but through modelling and evaluation

**3. Conclusion**

The data was successfully modelled, showing that indeed the customer satisfaction data can be modelled.

Of all the classification techniques…

**4. References**

* Fitzmaurice et. Al. (2015). “Handbook of Missing Data Methodology”. Taylor & Francis Group: Boca Raton.
* Harrison, M. (2019). “Machine Learning Pocket Reference”, O’Reilly Media: Sebastopol.
* Kaggle user sjlesharc (2019). “Airlines Customer Satisfaction”, Kaggle dataset. Available at: <https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction> (Accessed: 26/11/23).
* Müller & Guido. (2018). “Introduction to Machine Learning with Python”, ”, O’Reilly Media: Sebastopol.
* Provost & Fawcett. (2013). “Data Science for Business”, O’Reilly Media: Sebastopol.
* Wikipedia contributors (2023) “CRISP-DM process diagram.” In: Wikipedia. Available at: <https://upload.wikimedia.org/wikipedia/commons/b/b9/CRISP-DM_Process_Diagram.png> (Accessed: 26/11/23).