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

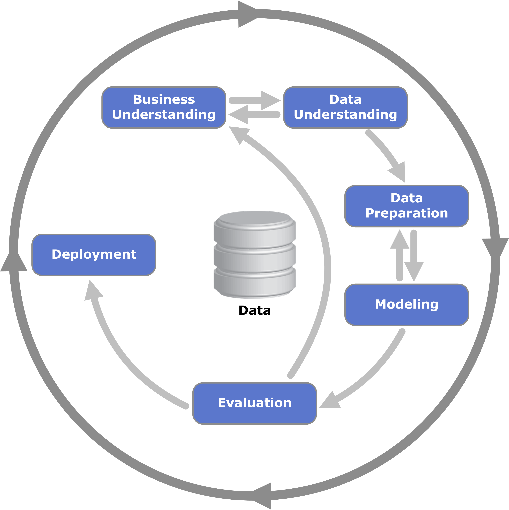
**Contents**

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. As this project

**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

**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.

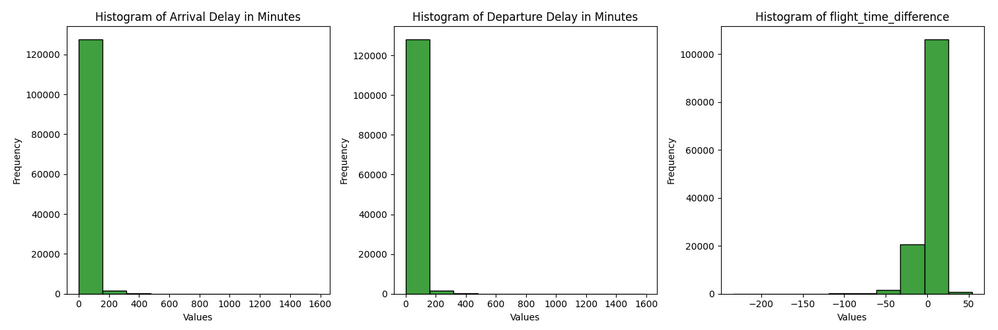


Figure : Signal captured in flight time difference

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.

**2.4 Modelling**

**2.5 Evaluation**

**3. Conclusion**

**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).
* 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).