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

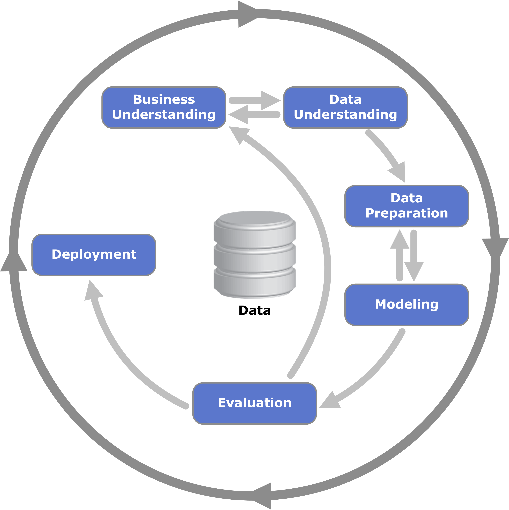
**Contents**

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 airline anonymised 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 following questions will be attempted in this assignment.

* Can Airline Customer Satisfaction be modelled with machine learning?

This project is a supervised learning problem and will attempt to perform a binary classification to predict airline customer satisfaction. This is an important first question to ask.

* Which classification techniques can best represent the data?

A range of classification machine learning models will be performed on the data. Different types of models being employed will aid in finding a model that best fits the data. Logistic Regression, decision trees, Gaussian Naïve Bayes, and Support Vector Machine Classifiers are planned. Finding the best classification model for the data improves confidence in the results and deployment.

* 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.

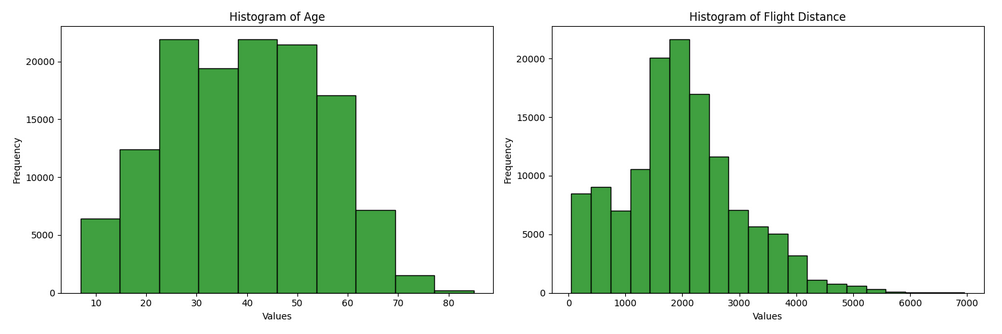


Figure 2: Histograms of continuous features.

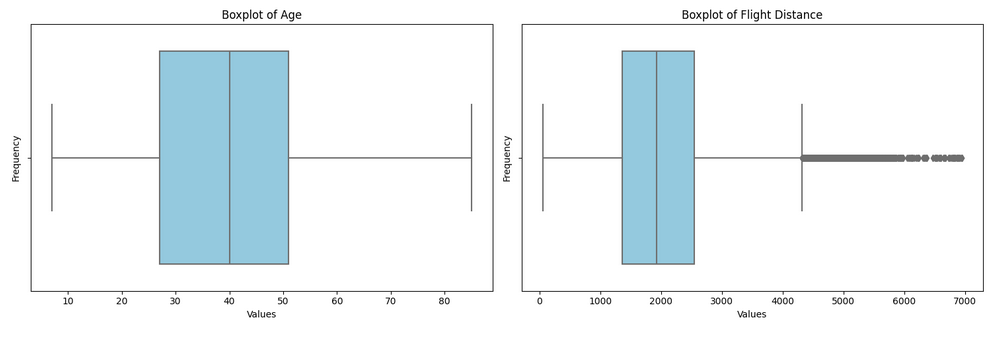
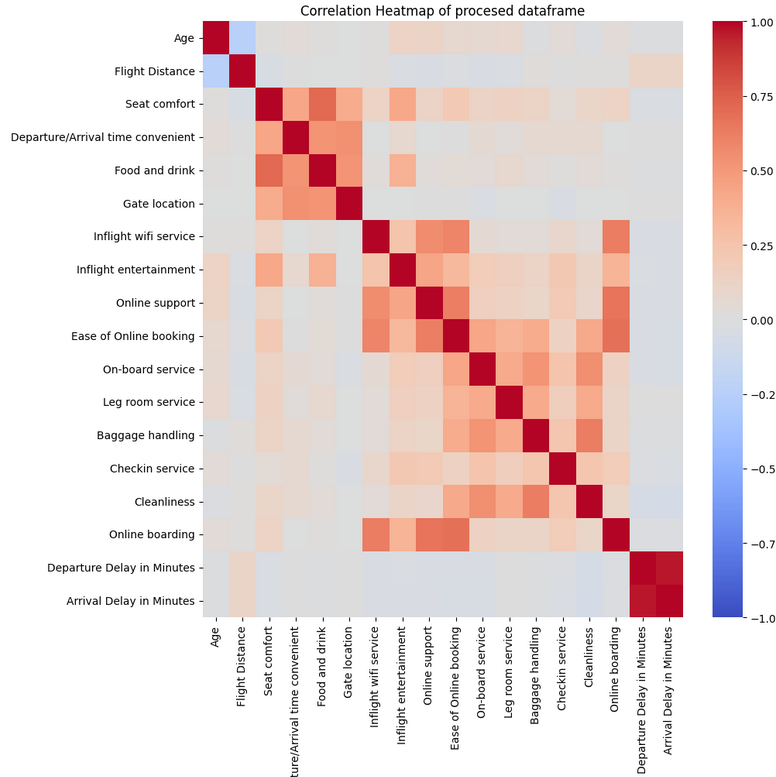


Figure 3: Boxplots of continuous features, with outliers shown for Flight Distance.

Some features are correlated, features can be engineered from some of these obvious groupings to extract signal. Two features are highly correlated and will be examined further.

****

**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 and could potentially introduce bias (Gallatin & Albon, p. 85), but the remaining sample size is large enough to make statistical inferences from.

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

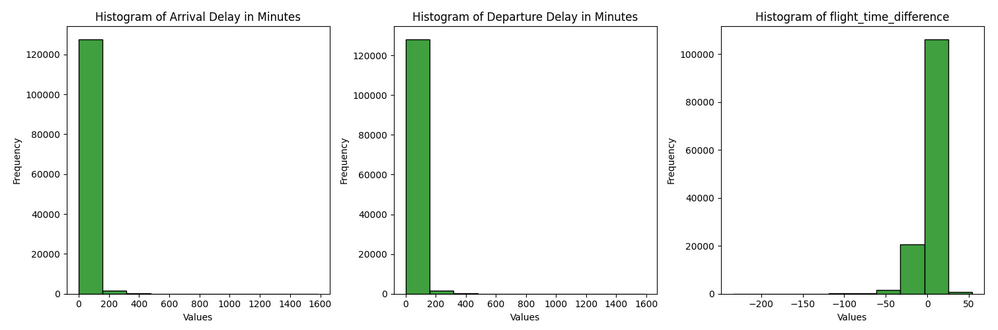
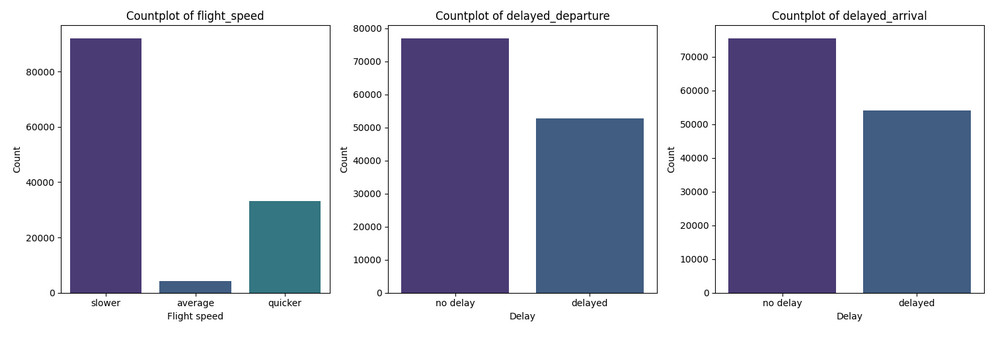


Figure 4: Signal captured in flight time difference

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.

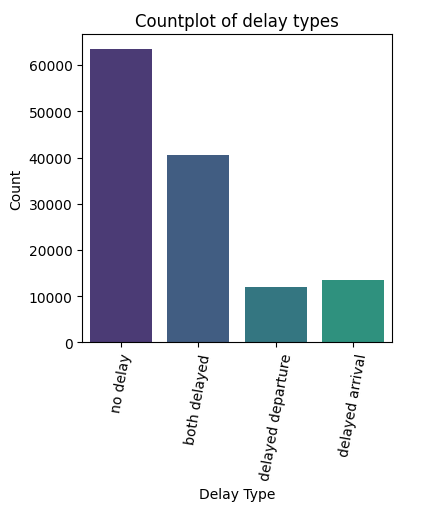


Figure 5: Countplot of delay types

Some of the ranked features are grouped and averaged, creating three new features: *‘avg\_amenities\_ratings’, ‘avg\_online\_services’,* and *‘avg\_facilities\_rating’.*

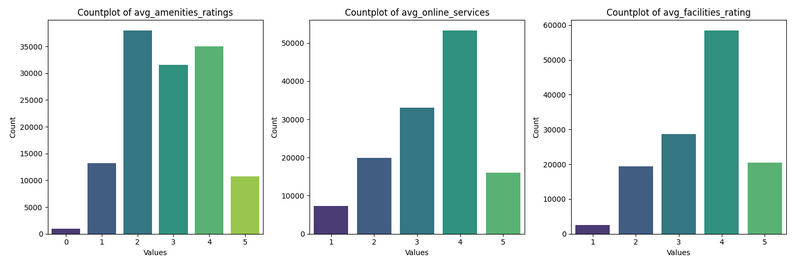


Figure 6: Engineered features of average services

The columns were renamed. The ranked columns and categories are one hot encoded. Binary columns are label encoded to avoid multicollinearity. Age is binned in decades and continuous features are scaled.

As there are many related features dimensionality reduction in the form of principal component analysis is applied.

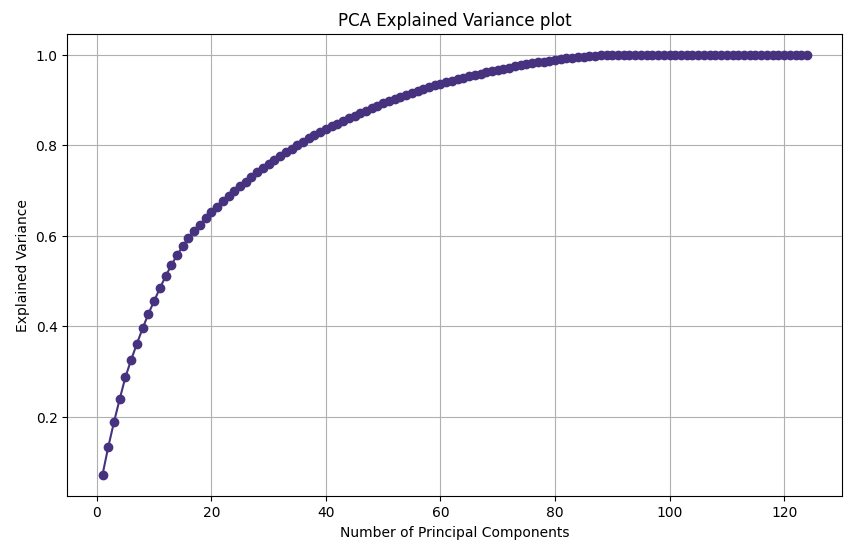


Figure 7: PCA Explained Variance plot

**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.

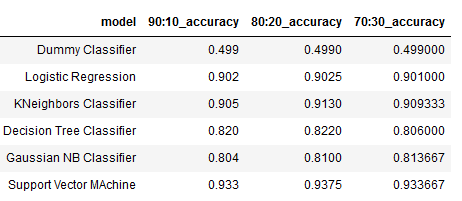


Figure 8: Accuracy results from training and test splits

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 added to the table below.

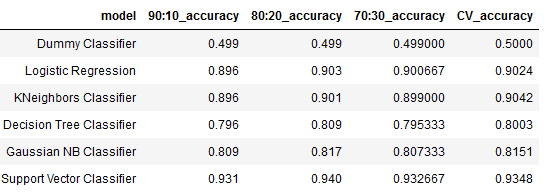
****

Figure 9: Accuracy results with cross validation accuracy

The models can model the data, with three models warranting further exploration. From the above results, it seems the 80:20 split has the highest accuracy and will be used going forward.

**2.5 Evaluation**

Three models are chosen for further evaluation: Logistic Regression, KNeighbours Classifier, and Support Vector Machine Classifier.

**2.6 Deployment**

Actual deployment is outside of the scope of this assignment, but through modelling and evaluation choices can be made to recommend a model for deployment that would be facilitate business goals.

Throughout modelling and evaluation, Support Vector Machine Classifier has been consistently out performing other models.

Grid search is performed to refine the parameters giving a model that uses the following parameters

* ‘rbf’ kernel (the default kernel in the Scikit Learn library (Scikit Learn, 2023)).
* Regularisation, C, of 10. Which is L2 Regularisation.
* ‘Auto’ kernel coefficient, gamma, as 1 / n\_features.

Dimensionality reduction was performed on the data before SVC modelling, so one way to examine which features are important to the model is to understand the features that are important to the variance of the principal component analysis.

The figures below show the top weighted, and lowest weighted features for the PCA used in the final model.

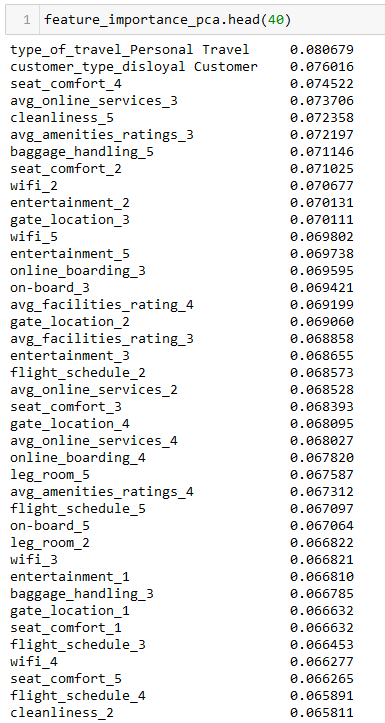
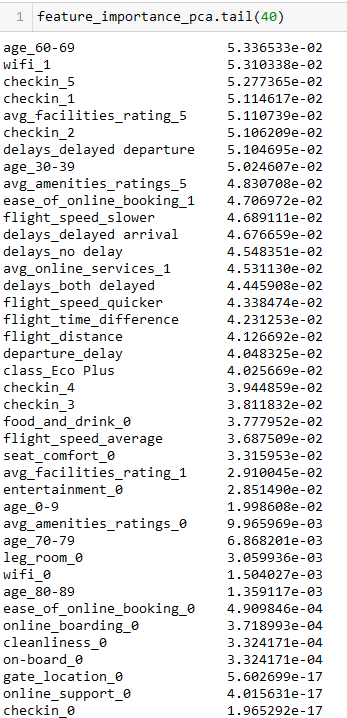
****

Figure 10: Bottom 40 features for PCA

Figure 11: Top 40 features for PCA

**3. Conclusion**

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

Of all the classification technique Support Vector Machine was the most fitting for the data with the highest accuracy, and best generalisation of the different models trialled.

Some recommendations can be made from the deployment model by examining the weights. It seems that the type of travel is the most important for customer satisfaction followed by the type of customer. This is important for business understanding.

Other engineered features scored high, attesting to their use in the model while highlighting business targets to achieve customer satisfaction. The first entry of average online services scored slightly higher than the first entry of average amenities in examining the PCA variance coefficients. Although they are similar enough to suggest to focus on achieving an average rating in them would increase customer satisfaction.

CRISP-DM was successfully followed and informed this study. Even at the last stage during deployment the iterative nature of CRISP-DM allowed insight into the business understanding and model evaluation, showing the benefits of the cyclic nature of CRISP-DM.

The business questions posed were successfully answered: Yes, airline customer satisfaction data can be modelled, and as stated, Support Vector Machine Classifier was the best model for the data, and the important features for the model were successfully identified to inform business decisions and strategies.

**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.
* Scikit Learn developers (2023) “sklearn.svm.SVC”, Scikit Learn. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> (Accessed: 26/11/23).
* 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).