**CCT College Dublin**

**Assessment Cover Page**

| **Module Title:** | Machine Learning |
| --- | --- |
| **Assessment Title:** | What factors most impact short-haul dissatisfaction? |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Group:  Ana Isabel Nieves Barcenas  Bárbara Azevedo Pereira  Daniela Daia  Vicente Rubio |
| **Student Number:** | 2022455 - Ana Isabel Nieves Barcenas  2022310 - Bárbara Azevedo Pereira  2017207 - Daniela Daia  2022178 - Vicente Rubio |
| **Assessment Due Date:** | 27th November 2022 |
| **Date of Submission:** | 27Th November 2022 |

**Declaration**

| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |
| --- |

## What factors most impact short-haul dissatisfaction?

****

### by,

Ana Isabel Nieves Barcenas

Bárbara Azevedo Pereira

Daniela Daia

Vicente Rubio

**Higher Diploma in Science in Data Analytics for Business Strategic Thinking**

Dr. Muhammad Iqbal

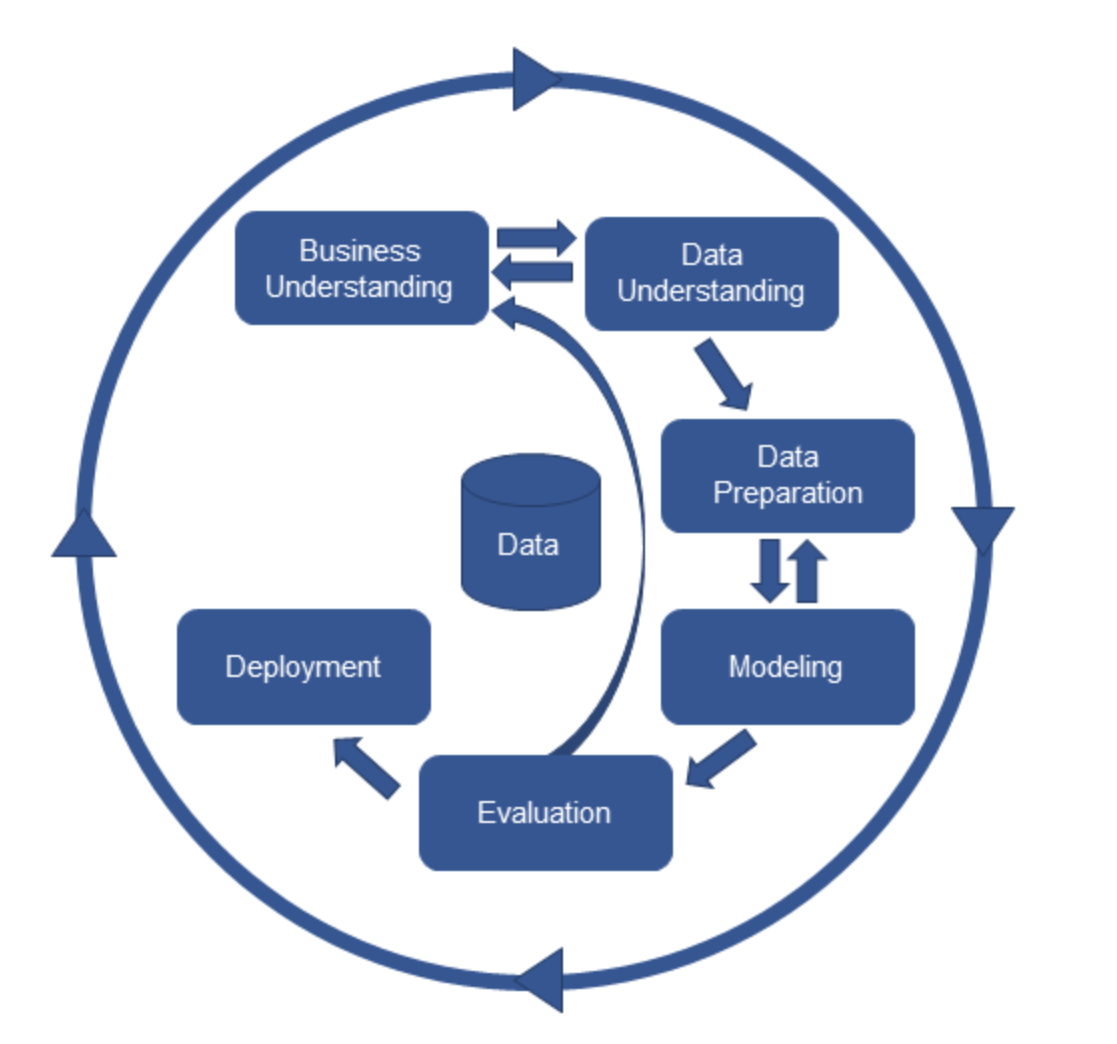
CCT College Dublin, Ireland.

**Abstract**

This analysis is based on a binary classification data set research made by over 120,000 airline passengers satisfaction. It will study what factors are highly correlated to dissatisfaction with short-haul passengers. The analysis includes additional information about each passenger, their type of flight and type of travel, and their evaluation of different factors like cleanliness, comfort, service, and overall experience.

It will be described the motivation of the chosen data, a description of the business problem and an explanation of the project goal. It will be presented the characterization of the data by applying Exploratory Data Analyses (EDA), filling in the missing values, observing the outliers by plotting boxplots, and the use of the feature selection model to extract the influencing factors of passenger dissatisfaction by using Principal Component Analyses (PCA). Afterwards, cross-validation techniques will be applied by using machine learning approaches, hyperparameters and a comparison between the chosen model. And finally, the interpretation and explanation of the results obtained based on different classification models.

This report will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, going through the stages: Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation and Deployment. Code available at [*GitHub*](https://github.com/IsabelNieves/Machine-Learning-CA.git)*.*

**

Keywords: a*irlines, passenger satisfaction, machine learning models, CRISP-DM.*

**Table of Contents**

Introduction

[Business Understanding 4](#_heading=h.1fob9te)

[Data Understanding 6](#_heading=h.ihv636)

[Data Preparation 9](#_heading=h.32hioqz)

[Modelling 15](#_heading=h.1hmsyys)

[Evaluation 25](#_heading=h.41mghml)

[Deployment 25](#_heading=h.2grqrue)

[Reference List 26](#_heading=h.vx1227)

**Table of Figures**

Figure 1. Shape and info of the dataset 6

Figure 2. Head of the dataset 7

Figure 3. 5 number summary (describe) 7

[Figure 4. Dropping and renaming columns. 9](#_heading=h.3fwokq0)

[Figure 5. Filling missing values with interpolate method. 9](#_heading=h.1v1yuxt)

[Figure 6. Calculating outliers. 10](#_heading=h.4f1mdlm)

[Figure 7. Creating a consolidated dataset for Dublin. 10](#_heading=h.2u6wntf)

[Figure 8. Boxplots per Dublin area 11](#_heading=h.19c6y18)

Figure 9. Boxplot for consolidated Dublin dataset 12

[Figure 10. Histogram for the original dataset (per area) 13](#_heading=h.3tbugp1)

[Figure 11. Histograms per Dublin area 13](#_heading=h.28h4qwu)

[Figure 12. Line graph per Dublin area 14](#_heading=h.1hmsyys)

[Figure 13. Line graph for Dublin consolidated dataset 15](#_heading=h.nmf14n)

[Figure 14. Decomposed series for Dublin consolidated dataset 16](#_heading=h.37m2jsg)

[Figure 15. Dickey-Fuller test for Dublin consolidated dataset 16](#_heading=h.1mrcu09)

[Figure 16. Fitting the best ARIMA model into the training set for Dublin consolidated](#_heading=h.46r0co2) [dataset 18](#_heading=h.46r0co2)

[Figure 17. Residual Diagnostic for Dublin consolidated dataset 18](#_heading=h.2lwamvv)

[Figure 18. House prices with forecast for Dublin consolidated dataset 19](#_heading=h.111kx3o)

[Figure 19. Difference between predicted and actual prices for Dublin consolidated](#_heading=h.3l18frh) [dataset 20](#_heading=h.3l18frh)

[Figure 20. Chi-squared test for Dublin consolidated dataset 20](#_heading=h.206ipza)

[Figure 21. Line graph for Dublin 01 dataset 21](#_heading=h.4k668n3)

[Figure 22. Decomposed series for Dublin 01 dataset. 21](#_heading=h.2zbgiuw)

[Figure 23. Dickey-Fuller test for Dublin 01 dataset 22](#_heading=h.1egqt2p)

[Figure 24. Fitting the best ARIMA model into the training set for Dublin 01 dataset. 22](#_heading=h.3ygebqi) [Figure 25. Residual Diagnostic for Dublin 01 dataset. 23](#_heading=h.2dlolyb)

[Figure 26. House prices with forecast for Dublin 01 dataset 24](#_heading=h.sqyw64)

[Figure 27. Difference between predicted and actual prices for Dublin 01 dataset 24](#_heading=h.3cqmetx)

[Figure 28. Chi-squared test for Dublin 01 dataset 24](#_heading=h.41mghml)

# Introduction

Whether travelling to nearby destinations or even to the farthest corners of the world, the key words in choosing airlines are to travel quickly, comfortably, and safely. Airlines are constantly offering new destinations, and over time more companies have entered the market, offering consumers choices, competitiveness, and affordability. The point is that affordable prices will sometimes go in hand with the quality of service. It is significant to assess passenger satisfaction regularly so that airlines can identify strengths and weaknesses, improve services and increase passenger retention.

In a highly competitive environment, the aviation industry stands to develop from a transport role to a service. Improving service quality is essential for competitiveness and ensures sustainable and healthy development. Therefore, airlines should promptly investigate passenger satisfaction and overall satisfaction with various services to understand the quality of existing services.

# Business Understanding

For this analysis, it will be used the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. The goal for the first phase is to understand the business and its needs, meaning the objectives and requirements for this project.

It is known that the airline industry is one of the fastest transportation sectors in the world; in that regard, Bart (2000) argues that traces of the strategic developments and the strategic responses of the airline players have had a profound impact on the shape and direction of the industry. These include the deregulation of the sector, the nature and extent of competition, the emergence of brand/differentiation-based competition, and airline alliance developments, strategies and their implications.

Furthermore, this industry represents important economic outcomes for national and international economies. In that sense, airlines have been introducing this service to international markets by contributing to global economic growth, where tourism plays a relevant role in the industry by enhancing competition and facilitating destinations (Williams & Naumann, 2011).

Based on the above challenges, this study focuses on the full-service passenger information and satisfaction survey results. This analysis aims to evaluate different machine learning algorithms and determine the most suitable algorithm for classifying customer short-haul flight dissatisfaction. This analysis also aims to ascertain and highlight the most critical variables in determining customer dissatisfaction for a better insight into the issues. Finally, this study is a reference for airlines to use customer evaluation-driven service methodologies to improve their competitiveness.

# Data Understanding

Data set available at: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>

### Data Dictionary

* **ID:** Unique passenger identifier
* **Gender:** Gender of the passenger (Female/Male)
* **Age:** Age of the passenger
* **Customer Type:** Type of airline customer (First-time/Returning)
* **Type of Travel:** Purpose of the flight (Business/Personal)
* **Class:** Travel class in the airplane for the passenger seat
* **Flight Distance:** Flight distance in miles
* **Departure Delay:** Flight departure delay in minutes
* **Arrival Delay:** Flight arrival delay in minutes
* **Departure and Arrival Time Convenience:** Satisfaction level with the convenience of the flight departure and arrival times from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Ease of Online Booking:** Satisfaction level with the online booking experience from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Check-in Service:** Satisfaction level with the check-in service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Online Boarding:** Satisfaction level with the online boarding experience from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Gate Location:** Satisfaction level with the gate location in the airport from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **On-board Service:** Satisfaction level with the on-boarding service in the airport from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Seat Comfort:** Satisfaction level with the comfort of the airplane seat from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Leg Room Service:** Satisfaction level with the leg room of the airplane seat from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Cleanliness:** Satisfaction level with the cleanliness of the airplane from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Food and Drink:** Satisfaction level with the food and drinks on the airplane from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **In-flight Service:** Satisfaction level with the in-flight service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **In-flight Wifi Service:** Satisfaction level with the in-flight Wifi service from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **In-flight Entertainment:** Satisfaction level with the in-flight entertainment from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Baggage Handling:** Satisfaction level with the baggage handling from the airline from 1 (lowest) to 5 (highest) - 0 means "not applicable"
* **Satisfaction:** Overall satisfaction level with the airline (Satisfied/Neutral or unsatisfied)

# Data Preparation - Isabel

# Modelling -

# Evaluation - all (daniela copied from her ryanair project)

It is known that organizations are entities in continuous and sequential process which involves effective strategic management within the whole organization in order to gain economic growth and maintain competitiveness over rivals. In that sense, Strategic Management by data analyse develop a set of suitable strategies for every area in the company to integrate all areas and members of the firm to achieve their goals and build differentiation and competitive advantage in the stock market.

In that sense, organizations should invest on different or related industries in order to gain market share also to spread risks they could face, to obtain the maximum return of investments. With that regard, Ryanair is suggested to invest on related diversification where Ryanair already run their operations. In order to realize economic benefits from the optimal utilization of capacity, or from the recombination of capabilities from multiple, related sectors.

Furthermore, it is vital in an organization to create their own organisational culture as blueprint of the company, in order to generate internal integration among all members of the organisation. As well as, to create identity with the personnel by following common goals that will bring better outcomes to everybody. In that regard, Michael O’Leary has been having an enormous responsibility by running one of the most successful firms in the airline industry, by performing transformational leadership which is developed according to the philosophy of Ryanair which spread an ambitious, creative and competitive workplace, where employees feel engaged with the way this organisation operates.

Finally, any organization should follow a strategy to achieve the organization’s goals, that strategy could be share and adapted within the organization, depending the area involved. In that context, it is possible to perform a set of strategies simultaneously through operational effectiveness that will create a unique and valuable position to the organization.

# Deployment - all

The analyses results show that the xxxxx model selects a feature subset containing xxx variables. In the classification prediction model, the random forest after xxxxx feature selection shows the best classification performance. Finally, combined with the four important variables extracted by xxx and logistic regression, further discussion is carried out, and suggestions are given for airlines to improve passenger satisfaction.

# Reference List - all

Porter, M. E., 1985. Competitive Advantage, New York: Free Press.

Bart, C.K., 2000. The Relationship between Mission and Innovativeness in the Airline Industry: An Exploratory Investigation. Int. J. Technology Management. 20 (1), pp. 475-489.

Williams, P. & Naumann, E. Customer satisfaction and business performance: A firm-level analysis. *J. Serv. Market.* <https://doi.org/10.1108/08876041111107032> (2011).

Brockwell, P.J. and Davis, R.A. (1987). *Time Series Theory and Methods*. Springer.

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection and Data Transforms in Python*. Jason Brownlee.

Chatfield, C. and Xing, H. (2019). *The Analysis of Time Series: an Introduction with*

### *R*. 7th ed. CRC Press.

Kirchgssner, G. (2014). *Introduction to Modern Time Series Analysis*. Springer.

Ladley, J. (2016). *Mastering and Managing Data Understanding*. [online] CIO. Available at: https:/[/www.cio.co](http://www.cio.com/article/238649/mastering-and-managing-data-)m[/article/238649/mastering-and-managing-data-](http://www.cio.com/article/238649/mastering-and-managing-data-) understanding.html [Accessed 19 May 2022].

McKinney, W. (2017). *Python for Data Analysis*. 2nd ed. Beijing: O’Reilly Inc.

The pandas development team (2022). *pandas: Python Data Analysis Library*. [online] pandas. Available at: https://pandas.pydata.org/ [Acc