**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Visualization Techniques, Machine Learning for Business |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | David McQuaid, Muhammad Iqbal |
| **Student Full Name:** | Daniela Mariano Barreto |
| **Student Number:** | 2023278 |
| **Assessment Due Date:** | 26/05/2024 |
| **Date of Submission:** | 26/05/2024 |

**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. |

Table of Contents

[**Introduction** 3](#_Toc167287255)

[Recommendation System 3](#_Toc167287256)

[Word count 4](#_Toc167287257)

[**Business Understanding** 4](#_Toc167287258)

[**Data Understanding** 4](#_Toc167287259)

[Data description 4](#_Toc167287260)

[**Data Preparation** 6](#_Toc167287261)

[**Machine learning models** 7](#_Toc167287262)

[Content filtering 7](#_Toc167287263)

[Collaborative filtering 7](#_Toc167287264)

[User-User 7](#_Toc167287265)

[Item-Item 7](#_Toc167287266)

[Evaluation 7](#_Toc167287267)

[Market Basket Analysis 7](#_Toc167287268)

[Apriori 7](#_Toc167287269)

[Frequent Pattern (FP growth) 7](#_Toc167287270)

[Evaluation 7](#_Toc167287271)

[**References** 8](#_Toc167287272)

# **Introduction**

With the advancement of the internet in people's lives, a large amount of data is generated, such as products researched, purchases made, films watched, music listened to, and more. Without a system that makes accurate recommendations most of the time, customers could get lost in the vast number of options and might even give up on making a purchase or watching a movie, which tends to impact the company's revenue. That is why many companies, such as Amazon, Netflix, and eBay, invest in recommendation systems.

## Recommendation System

Recommendation systems are algorithms designed to suggest relevant items to users based on the similarity of items or the characteristics of the user's profile (Patel, Patel and Chauhan, 2023, p.851). In other words, these algorithms can analyse user behaviour, interests, and characteristics to suggest similar products or services based on previous interactions.

This system interacts with users to learn their characteristics and preferences, storing this feedback in the recommender database that can be used for generating new recommendations for users with similar characteristics (Ricci et al., 2015, p.3; Patel, Patel and Chauhan, 2023, p.851).

According to Ricci et al. (2015, p. 5), there are several reasons to use a recommendation system in online retail businesses, such as increasing the number of items sold by tailoring to the user’s needs and wants, as well as selling more diverse items by offering items that might be hard to find, and so on. There are several techniques used to develop recommendation systems. In this project, the focus will be on Content-based and Collaborative-based filtering.

## Word count

|  |  |
| --- | --- |
| Introduction: | 0 |
| : |  |
| : |  |
| Total: |  |

# **Business Understanding**

In this project, I used the Cross Industry Standard Process for Data Mining (CRISP-DM), a helpful method for managing data mining projects and making decisions.

To develop this project, I created a business in which I had to answer some questions to help stakeholders make decisions, such as:

*Background*

A company called AnimeNow has been in the streaming sector for a decade. Recently, with the increase in the popularity of anime, the company wants to evaluate its users' satisfaction with the recommended anime.

*Business objectives*

To optimize user satisfaction and retention by enhancing the effectiveness of anime recommendations, thereby capitalizing on the growing popularity of anime within our streaming platform.

# **Data Understanding**

## Data description

*Recommendation system*

The datasets used in this project contain records about anime, a distinctive style of animated shows or movies originating from Japan that has been gaining global appreciation and recognition (Binjola, 2023, p.1). The datasets are from the Kaggle repository (www.kaggle.com, n.d.) that was first gathered from myanimelist.net API (Link below). There are two datasets: one named ‘anime.csv’, which contains 12,294 records of various anime and 7 features; the second dataset is called ‘rating.csv’ and has 7,813,737 records and 3 features. The 'rating' dataset will be referred to as 'user' to facilitate understanding that these records are from the users. The data dictionaries are presented in Tables 1 and 2.

Dataset: [Anime Recommendations Database (kaggle.com)](https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database)

**Table 1:** Data dictionary of ‘anime.csv’ dataset.

A screenshot of a computer

Description automatically generated

**Table 2:** Data dictionary of ‘rating.csv’ dataset.

A screenshot of a computer

Description automatically generated

*Market basket analysis*

In this analysis, the dataset used is from Kaggle and contains e-commerce data from an online electronics store (Kabir, n.d. p.1). It is composed by 92,250 records and 5 features, such ‘product’ and ‘transaction\_id’. The data dictionary is presented in Table 3.

**Table 3:** Data dictionary of ‘ecommerce.csv’ dataset.

A screenshot of a computer

Description automatically generated

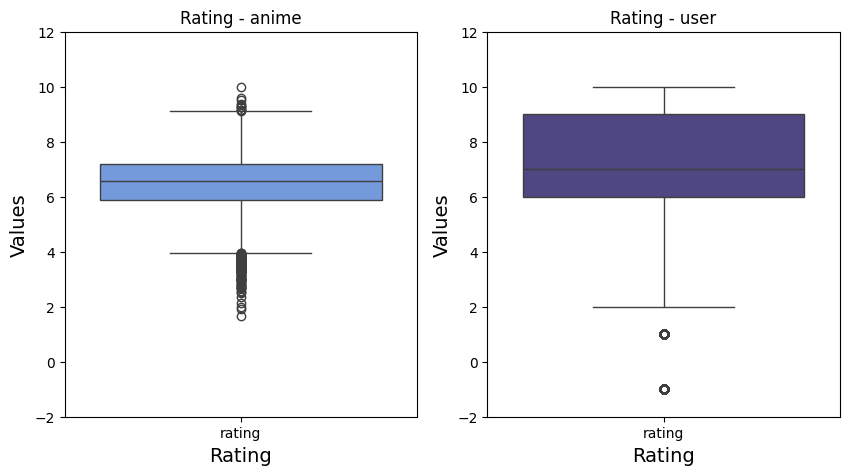
# **Data Preparation**

*Data cleaning*

Procedures such as verifying duplicates and missing values are essential for ensuring data quality when performing statistical analysis, visualization, and modelling. Without these checks, the results can be unreliable and inaccurate, potentially leading to incorrect decisions. The 'user' dataset did not contain any missing values. However, 2.25% of the 'anime' dataset was removed due to this proportion not being substantial. Regarding duplicates, only one was present and subsequently removed.

I used the 'normal\_test()' function to determine if the data were normally distributed. The result indicated that the 'rating' and 'members' variables from the 'anime' dataset are non-normally distributed, while the 'rating' variable from the 'user' dataset is approximately normally distributed.

After searching for outliers, I found that they appear to be part of the dataset, as they fall within the expected range of rating values from 0 to 10. Because of this, I have decided to continue with them. The boxplot is presented in Figure 1.



**Figure 1:** Boxplot with outliers. a) ‘rating’ from ‘anime’ dataset, b) ‘rating’ from ‘user’ dataset.

b

a

The EDA and preprocessing were conducted on the 'ecommerce' dataset to perform Market Basket Analysis. With zero missing values and no duplicates, I proceeded to remove whitespace and cast the 'Transaction ID' to a string datatype to ensure consistency during the analysis. Additionally, I utilized one-hot encoding to transform the data into a format suitable for analysis.

# **Machine learning models**

## Content filtering

This technique creates a user profile based on data provided directly (explicit feedback) or indirectly (implicit feedback) by the user, such as a rating given to a product; then this information is used to recommend more similar products or services (Patel, Patel and Chauhan, 2023, p.852).

## Collaborative filtering

The collaborative filtering is widely implemented, especially in e-commerce sites, and this technique works by identifying similarities between users and recommending items based on what similar users have liked or purchased in the past (Ricci et al., 2015, p.12,13).

### User-User

### Item-Item

## Evaluation

## Market Basket Analysis

Also known as association-rule it is a method employed to unveil customer purchase patterns by analyzing transactional data from stores. This insightful approach can yield a competitive edge for retail companies. By discerning the typical items a customer purchases, it facilitates strategic enhancements in store layouts, website design, and marketing strategies, such as promoting bundled offerings (Chen et al., 2005, p.339).

Apriori and FP-Growth are the most common algorithms for mining frequent itemsets by defining the minimum support parameter for identify the frequent itemsets.

### Apriori

### Frequent Pattern (FP growth)

## Evaluation

# **References**

Patel, D., Patel, F. and Chauhan, U. (2023). Recommendation Systems: Types, Applications, and Challenges. *International Journal of Computing and Digital Systems*, 13(1), pp.851–868. doi:https://doi.org/10.12785/ijcds/130168.

Ricci, F., Lior Rokach, Bracha Shapira and Springerlink (Online Service (2015). *Recommender Systems Handbook*. New York, Ny: Springer Us.

www.kaggle.com. (n.d.). Anime Recommendations Database. [online] Available at: https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database [Accessed 13 May 2024].

Binjola, K. (2023). Effect of Anime on Personality and Popularity of Japanese Culture. In: Lecture Notes in Electrical Engineering. [online] Proceedings of International Conference on Recent Innovations in Computing. ICRIC 2022. Singapore: Springer. Available at: https://doi.org/10.1007/978-981-99-0601-7\_50 [Accessed 17 May 2024].

Kabir, F. (n.d.). Market Basket Analysis. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/farjanakabirsamanta/analytics-case-studyecommerce [Accessed 21 May 2024].

Chen, Y.-L., Tang, K., Shen, R.-J. and Hu, Y.-H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, 40(2), pp.339–354. doi:https://doi.org/10.1016/j.dss.2004.04.009.