**Integrated CA2**

**Recommendation systems and interactive dashboard**

By

Leopoldo Rojo Romero (Student ID: 2023355)

Higher Diploma in Science in Data Analytics for Business

Data Visualisation Techniques & Machine Learning for Business

David McQuaid & Muhammad Iqbal

CCT College

Dublin, Ireland

**Table of contents**

[**Introduction** 2](#_Toc167655166)

[**Objectives of this analysis** 2](#_Toc167655167)

[**Data source** 3](#_Toc167655168)

[**Ethical considerations** 3](#_Toc167655169)

[**Content and Collaborative recommendation systems** 3](#_Toc167655170)

[**Content recommendation system** 4](#_Toc167655171)

[**Collaborative recommendation systems** 6](#_Toc167655172)

[**Market Basket Analysis** 10](#_Toc167655173)

[**Interactive Dashboard** 13](#_Toc167655174)

[**Github** 17](#_Toc167655175)

[**References** 18](#_Toc167655176)

# **Introduction**

A **recommendation system** is a technology that uses machine learning and artificial intelligence (AI) to generate product suggestions and predictive offers, these systems analyse the following types of customer data:

* Browser history
* Current purchasing behaviour
* Feedback
* Most viewed products
* Preferences
* Previous purchases
* Recently viewed items
* Search history
* Shopping carts
* Wish list

(Salesforce, 2021)

Recommendations systems are highly used by retailers and eCommerce companies due to that they aim to make recommendations to their customers by identifying products or services that are more likely to be interest to them, with that, companies are able to up sell and cross sell their products or services (Devfi, n.d.).

An **interactive dashboard** is a data and analytics visualisation that allows users to manipulate data in real time, this feature enable users to get personalized insight at a glance and facilitating a greater understanding of data (Shah, 2023).

Approximately 75% of world’s population believes that visual interpretation is the most prominent approach when understanding anything, this is why companies are seeking tools or individuals that are able to make reporting into real time interactive dashboards (Sharma, 2021).

# **Objectives of this analysis**

* Understand Content recommendation system
* Understand Collaborative filtering
  + User based Collaborative filtering
  + Item based Collaborative filtering
* Understand Market Basket Analysis
  + Market Basket Analysis with Apriori algorithm
  + Market Basked Analysis with Frequent Pattern algorithm
* Create an interactive dashboard

# **Data source**

We will approach these topics using 4 datasets from 3 sources:

For our **content recommendation system,** we will use a dataset gotten from the UCI Machine Learning Repository, this dataset provides patient reviews in specific drugs along with related conditions, with ratings going from 1 to 10, reflecting overall patient satisfaction.

For our **collaborative recommendations systems**, we will use a dataset gotten from Kaggle, this dataset provides transactions records of 1K+ Amazon product’s Ratings and Reviews as per their details listed on the official website of Amazon.

For our **market basked analysis**, we will use a dataset gotten from the UCI Machine Learning Repository, this dataset provides records from a transnational company which contains all the transactions occurring between December 2010 and December 2011.

For our **interactive dashboard**, we will use a dataset gotten from the UCI Machine Learning Repository, this dataset provides records from a transnational company which contains all the transactions occurring between December 2010 and December 2011 (same dataset used for our market basked analysis).

The datasets are publicly available in the following links:

<https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com>

<https://archive.ics.uci.edu/dataset/352/online+retail>

<https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset?resource=download>

# **Ethical considerations**

We considered the social importance of our content recommendation system dataset, as this can influence public action or inactions, for example people deciding to purchase an alternative medication for their condition(s).

Please note that we do not seek to make any medical conclusions and / or recommendations, the objective of this analysis is just to analyse and report our findings for capstone research purpose only.

Also, to prevent breach of anonymity of population from which the data was gathered, the dataset that we will utilize does not have any sensitive data.

# **Content and Collaborative recommendation systems**

Content and Collaborative recommendations systems differ based on the specific kind of information they collect and how they use it to determine the products they suggest to a customer.

A content recommendation system analyses each customer’s preferences and purchasing behaviour, with these features creates a preference profile to make suggestions of product or services, this recommendation system is the one that is usually behind of the “Since you bought this, you’ll also like this…”.

A collaborative recommendation system analyses data from multiple customers to make suggestions of what product or services will be of interest to a particular individual, in other words, takes into account the crowd’s ratings to offer products or services, “Customers that saw this product, also viewed this X-product”.

This kind of systems are great with companies that have access to a large amounts of customer data.

(Salesforce, 2021)

## **Content recommendation system**

We will approach this recommendation system with a dataset with patient records as previously mentioned, the goal is to give patients recommendations of other options of medications based on their current medication and condition.

***Dataset dictionary***

Unnamed: 0: Patient ID

drugName: Medication name

condition: Condition name

review: Patient’s review

rating: Patient’s rating from 1 to 10

date: Date of the review

usefulCount: Number of users who found the review useful

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Figure 1: Drugs dataset.

Now that we have seen our dataset, let’s check for null values.

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Figure 2: Drugs dataser null values.

Knowing that we have less than 1% of missing values in our ‘condition’ feature, and that we want to recommend a medication based on a condition, we will drop the null values.

Now, let’s create a new feature that contains ‘drug’ and ‘condition’, then drop the clumns thar we will not use (‘Unnamed: 0’, ‘review’ and ‘date’).

We are combining our ‘drug’ and ‘condition’ features to have a unique value where we can sum the ‘usefulCount’ values, to be able to see how many votes a drug – condition need to fall within the 80th quantile.

Furthermore, we will need to create a feature that contains the average rating of each drug – condition, additionally, we will add a feature of ‘keywords’ containing the data from our drug – condition feature, this ‘keywords’ feature we will be the one that we will pass to train our recommendation system.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figure 3: Drug dataset with new features.

At this stage, we can calculate the average rating of the overall records and the number of useful votes received by a drug in the 80th percentile, this gives us an average rating of 7.38 on a scale of 10 and 374 votes.

Note that we are doing this analysis with 80% since health is a sensitive topic, and we will use a mix between a good amount of votes and a good rating (above 7).

After performing these steps, this is the dataset that we are left with.

Captura de pantalla de computadora

Descripción generada automáticamente

Figure 4: Drugs dataset within 80th percentile and rating above 7.

Now, we can fit our content recommendation system and make recommendations, for instance, recommendation of medication for a person that uses ‘Acanya for Acne’.

Texto

Descripción generada automáticamente con confianza media

Figure 5: Medication recommendations.

## **Collaborative recommendation systems**

We will approach this recommendation system the User based and Item Based method, with a dataset with transaction records from Amazon as previously mentioned, the goal is to recommend a product based on the User ratings and Item ratings respectively.

***Dataset dictionary***

* product\_id: Product ID
* product\_name: Name of the Product
* category: Category of the Product
* discounted\_price: Discounted Price of the Product
* actual\_price: Actual Price of the Product
* discount\_percentage: Percentage of Discount for the Product
* rating: Rating of the Product
* rating\_count: Number of people who voted for the Amazon rating
* about\_product: Description about the Product
* user\_id: ID of the user who wrote review for the Product
* user\_name: Name of the user who wrote review for the Product
* review\_id: ID of the user review
* review\_title: Short review
* review\_content: Long review
* img\_link: Image Link of the Product
* product\_link: Official Website Link of the Product

Since our dataset has many features, we will visualise it through the .info() method.

Imagen que contiene texto, exterior, placa, calle

Descripción generada automáticamente

Figure 6: Amazon dataset .info().

Now, let’s check for null values.

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Figure 7: Amazon dataset null values.

Since we have less than 1% of null values in our ‘rating\_count’ feature and is correlated with the ‘rating’ feature, we will drop the null values.

Before training our recommendation systems, we will create a new dataset that contains the User ID, Item ID and ratings, which is the data that we need.

Texto

Descripción generada automáticamente

Figure 8: Amazon ratings dataset for recommendation systems.

Moreover, we can check the sentiment of the ratings of all the records, to see if our recommendations would be about a similar item that liked or an alternative to an item that they were okay with or did not like.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 9: Sentiment Analysis of Amazon reviews.

Now, let’s train our recommendations systems, for this, we will use the “Surprise” library by scikit.

**User based**

For the user based recommender, we need to pass ‘user\_based’ : True to the model.

Captura de pantalla de computadora

Descripción generada automáticamente

Figure 10: User based model.

Let’s print some ratings predictions of our User based model.

Texto

Descripción generada automáticamente

Figure 11: User based ratings predictions.

**Item based**

For the item based recommender, we need to pass ‘user\_based’ : False to the model.

Texto

Descripción generada automáticamente

Figure 12: Item based model.

Let’s print some ratings predictions of our Item based model.

Texto

Descripción generada automáticamente

Figure 13: Item based ratings predictons.

As shown on the above predictions, we are getting a constant rating of 4.0955 in almost every prediction, this could be caused by the fact that we were left with only 158 records when we looked of the amount of users that had rated an item.

## **Market Basket Analysis**

A Market Basket Analysis is a tool that a company can use to analyse consumer shopping behaviours and purchasing trends, understanding this type of analysis can help a business make better marketing campaigns and promotions to optimise and maximise sales, improve customer satisfaction, increase cross selling and identify other customer patterns (Indeed Editorial Team, 2023).

We will approach this recommendation system with a dataset with records from a transnational company as previously mentioned, the goal is to know which item is likely to be bought after taking another item.

***Dataset dictionary***

* InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
* StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name. Nominal.
* Quantity: The quantities of each product (item) per transaction. Numeric.
* InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
* UnitPrice: Unit price. Numeric, Product price per unit in sterling.
* CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
* Country: Country name. Nominal, the name of the country where each customer resides.

Pantalla de computadora

Descripción generada automáticamente con confianza media

Figure 14: Retail dataset.

Now, let’s check for null values.

Pantalla negra con letras blancas

Descripción generada automáticamente

Figure 15: Retail dataset null values.

As we can observe, we have 24.9% of null values in our ‘CustomerID’ feature, but we do not need this feature for our analysis, and less than 1% of null values in our ‘Description’ feature, since we need the actual description of the item for our analysis, we will drop the null values.

Note that for this analysis in particular, we only want to know the likelihood of an item being bought after taking another item, so we will not take into account any order cancellation.

Before passing the data through our algorithms, we need to set every unique item as a feature (columns) and we will use the InvoiceNo (the receipt) as our index (and not taking into account any other feature), by performing this step, we will visualise if which items where bought together (1 = bought, 0 = did not buy), here is the dataset that we were left with.

Tabla

Descripción generada automáticamente con confianza media

Figure 16: Retail dataset before passing it through the algorithms.

**Apriori algorithm**

For this algorithm, we will use the ‘confidence’ metric, this metric represents the probability of seeing the consequent in a transaction given that it also contains the antecedent.

Let’s see the results.

Captura de pantalla de computadora

Descripción generada automáticamente

Figure 17: Apriori results.

As we can see, there is a confidence of 89.44% that a individual buying a PINK REGENCY will also by a GREEN REGENCY.

**FP growth algorithm (Frequent Pattern)**

For this algorithm, we will use the ‘confidence’ metric, this metric represents the probability of seeing the consequent in a transaction given that it also contains the antecedent.

Let’s see the results.

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

Figure 18: FP growth algorithm results.

As expected, we are getting the same results as our Apriori algorithm, this is because the difference between these algorithms is mainly the time in processing the information (Anon, 2022).

# **Interactive Dashboard**

In order to create our interactive dashboard, we will use the same dataset that we used in our Market Basket Analysis, this dataset provides records from a transnational company as previously mentioned, we aim to summarise and visualise the amount of sales and amount of items sold in each European country in that period of time (December 2010 to December 2011), and how the sales were changing during that time in each country in a map.

In this case, after dropping our null vales, we will also drop the ‘InvoiceNo’, ‘StockCode’, ‘Description’ and ‘CustomerID’, here is the dataset that we were left with.

Captura de pantalla de computadora

Descripción generada automáticamente

Figure 19: Int-dash initial dataset.

Now that we have our initial dataset, we will perform the following steps, with the objective to make our dashboard using the Plotly library.

* Create a ‘Total’ feature by multiplying quantity times price.
* Create a ‘Year’ and ‘Month’ features from our ‘InvoiceDate’ feature.
* Replacing the number of the month for a string, for instance ‘01’ for ‘Jan’ (this will be only to not lose the actual year and month when we split our data into 2).
* Remove any non-European country.
* Create a ‘Code’ feature, this feature will contain the Country Code of each country, this is needed since Plotly will take these codes to highlight the right country in the map.
* Create 2 datasets, one for sales and the other for quantities (this will help us plot 2 separate bar plots, representing sales and quantities respectively).
* Create a ‘YearMonth’ feature in both datasets, this will allow us sort the dates (time) from the beginning to the end in our dashboard.

These are the datasets that we were left with.

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Figure 20: Int-dash sales dataset.

Pantalla de computadora

Descripción generada automáticamente con confianza media

Figure 21: Int-dash quantities dataset.

With these datasets, we can start making our visualisation in order to add them together in a dashboard.

First, we will create a map using the ‘choropleth’ map from Plotly, this map will show us how sales changed throughout our period of time.

* Set the initial zoom in Europe, by setting latitude=54.5260 and longitude=15.2551.
* Use our sales dataset.
* Use our ‘Code’ feature.
* Set the ‘color’ parameter with our ‘Total’ feature, this will make the colours change as the total sales change.
* Set the ‘hover\_name’, this will show as the full name of the country when we put our mouse over the Country in the map.
* Se the ‘animation\_frame’ to our ‘YearMonth’ feature, this will give us a slide bar at the bottom of the map, in based of this parameter, our map will be displaying the total sales.

Let’s see our map.

Imagen que contiene Mapa

Descripción generada automáticamente

Figure 22: Plotly map of sales.

In order to create our bar plots to visualise total sales and quantities, we will create another 2 dataset with only the ‘Country’ and ‘Total’ features respectively.

For these plots, is highly important that we set their width to half of the width of our map plot, this is because if we do not set this value, the plots will overflow even if we set then in the same row of the dashboard, and for a pleasant visualisation, we want to put them side by side.

Gráfico

Descripción generada automáticamente

Figure 23: Plotly bar of total sales.

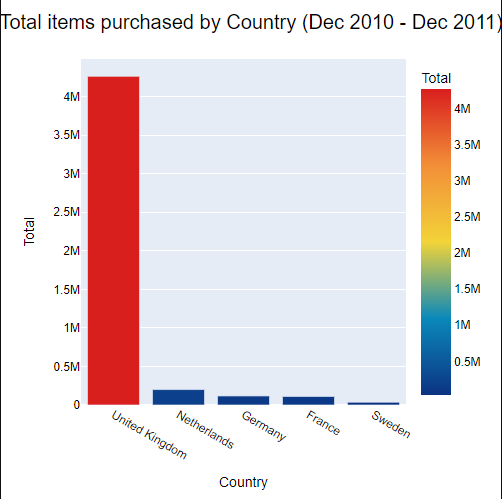


Figure 24: Plotly bar of total quantities.

To run the actual dashboard, please head to the jupyter notebook file in the Github link at the bottom of this report.

With the demographic of younger adults in mind, adults between 18 – 35 years old, we decided to highlight the map and the bar plots with the ‘portland’ palette, this palette make it fast to the viewer to identify the Country where the company sold the most by money and items, furthermore, an interactive map, allows the user to visualise every country in our data in the desired point in time.

By having a main map at the top and 2 bar plots at the bottom, the user can quickly absorb information about our data, this is relevant in a dashboard since the attention span of the younger adults is decreasing over time due to the increasing endless 15 to 30 second long videos in social media (Sijercic, 2023).

# **Github**

<https://github.com/CCT-Dublin/integrated-ca2-dvt-and-mlb-LeopoldoCCT>

# **References**

Anon, (2022). *Comparison of Apriori, Apriori-TID and FP-Growth Algorithms in Market Basket Analysis at Grocery Stores*. [online] Available at: https://www.researchgate.net/publication/368612949\_Comparison\_of\_Apriori\_Apriori-TID\_and\_FP-Growth\_Algorithms\_in\_Market\_Basket\_Analysis\_at\_Grocery\_Stores/fulltext/63f0cda419130a1a4a8d84cd/Comparison-of-Apriori-Apriori-TID-and-FP-Growth-Algorithms-in-Market-Basket-Analysis-at-Grocery-Stores.pdf.

Devfi. (n.d.). *Recommender Systems for Retail & eCommerce*. [online] Available at: https://www.devfi.com/usecases/recommender-systems-for-retail-ecommerce/ [Accessed 26 May 2024].

Indeed Editorial Team (2023). *Market basket analysis: definition and examples*. [online] Available at: https://uk.indeed.com/career-advice/career-development/market-basket-analysis.

Salesforce (2021). *What Is a Retail Product Recommendation Engine?* [online] Salesforce.com. Available at: https://www.salesforce.com/resources/articles/what-is-product-recommendation-engine/.

Shah, V. (2023). *What is an interactive dashboard? A complete overview*. [online] ThoughtSpot. Available at: https://www.thoughtspot.com/data-trends/dashboard/interactive-dashboard.

Sharma, A. (2021). *Know the Power of Interactive Dashboards and Examples*. [online] EzDataMunch. Available at: https://medium.com/ezdatamunch/know-the-power-of-interactive-dashboards-and-examples-6bb9a49cf696 [Accessed 26 May 2024].

Sijercic, A. (2023). *TikTok Effects on the Attention Span*. [online] Digital Reflections. Available at: https://medium.com/digital-reflections/tiktok-effect-on-attention-span-12211b0a06a1.