A white background with blue and gray dots

Description automatically generated

**Table of Contents**

[Executive summary 3](#_Toc163040012)

[Introduction 3](#_Toc163040013)

[Data Collection 3](#_Toc163040014)

[Exploratory Data Analysis and Data Preparation 3](#_Toc163040015)

[Recommendation System Approach 4](#_Toc163040016)

[Feature 1 – Similarity between titles 4](#_Toc163040017)

[Feature 2 – Item based collaborative filtering 4](#_Toc163040018)

[Feature 3 – User based collaborative filtering 5](#_Toc163040019)

[Evaluation 5](#_Toc163040020)

[Conclusion 6](#_Toc163040021)

[Model Advantages 6](#_Toc163040022)

[Model Disadvantages 6](#_Toc163040023)

[Business Application 6](#_Toc163040024)

[Next Steps: How to improve the recommendations 7](#_Toc163040025)

# **EXECUTIVE SUMMARY**

Livraria Lello, a bookstore in Porto, is facing a decline in sales, wants to find a way to revert this trend. Our goal was to identify popular books among customers, optimize inventory, and implement personalized offers through email marketing. To do this, we built a recommendation system that integrates item-based and user-based collaborative filtering. By combining these approaches, we aimed to provide a robust recommendation tailored to individual preferences and purchase histories for each client.

The evaluation of the model revealed challenges coming from limited data volume which particularly affects model accuracy. To further improve recommendations in the future, we recommend long-term methods for data collection, such as incorporating an online platform similar to GoodReads for users to register books and ratings more easily.

# **INTRODUCTION**

Livraria Lello has experienced a significant decrease in sales over the past year, mainly due to the rising of ebooks, a trend intensified by the COVID-19 outbreak. To address this challenge, they hired to analyze the data they have collected about user preferences over the last few years to improve customer engagement and revenues.

The primary goal is to recognize what are the books with the highest level of probability to be bought in the future. With this information, we aim to optimize the company's inventory by increasing purchases of these popular titles while reducing those less demanded. Moreover, it will also help the company to create a layout strategy for the shelves to know which books to put side by side (as the company knows what are the books more probable to be bought after reading a specific book). Furthermore, to enhance customer engagement, Livraria Lello intends to implement personalized offers by email marketing tailored to individual purchasing histories and preferences.

With the aim to anticipate forthcoming consumer preferences so the library can execute the previous actions, our consulting firm decided that the best approach was to create a recommendation system.

# **DATA COLLECTION**

To make this analysis, the library gave us access to 5 dataframes. From the start we understood that we would only need the collaborative\_books\_df that has the rating customer gave to each book. *Livraria Lello* told us this information was gathered by the marketing department by reaching out via email to their customers asking them to rate from 1 (Bad) to 5 (Excellent) on books they have previously bought.

There were other datasets that looked interesting as they contained specific information about the books (# pages, author, description). However, the number of unique books was too low (97) and so we were not able to use it. The rest of the datasets were composed with mapping of clients and books titles (not relevant for the development).

# **EXPLORATORY DATA ANALYSIS AND DATA PREPARATION**

We started by pre-processing and performing EDA on the datasets to understand the data we were presented with, such as its structure and characteristics, and assess the quality. Although we performed EDA for all datasets, we are only going to present the results for the ‘collaborative\_books\_df’, since it is the only one we are going to use.

At first glance, the data presented quality standards since it had no missing values, the rating was inside the 1-5 range, no duplicate entries for each unique user or book ID and no outliers in the actual rating. Finally, we had to change some variables from integer to string.

Regarding the distribution of the variables, we found some insight regarding the average number of time a user rates books, 2.9, and the average number of user ratings for a book is 219.

The former may prove a problem in the future since we might not have enough information about users to find similarities between them. Having found this problem we believe it will be important to create a recommendation that does not focus only on users but that also includes content based on books.

# **RECOMMENDATION SYSTEM APPROACH**

With the understanding of the data, our team is now able to build an accurate recommendation system. As such, we decided to compute the similarity between book titles and perform both item and user based collaborative filtering, for each book and user. With these three different approaches to perform recommendations and predict customers ratings, we will then combine them into one strong and robust recommendation system.

For that, we created a final model that averages the result of item-based and user-based collaborative filtering techniques, and then we analyzed the possibility to sum a “bonus” that comes from the similarity between titles, leveraging the strengths of all approaches.

It is important to note that, to increase the model’s accuracy, books already rated by the user, meaning that the user has already read them, are excluded to ensure that only new recommendations are made.

## **FEATURE 1 – SIMILARITY BETWEEN TITLES**

Our initial recommendation feature consists in suggesting books based on their titles’ similarities. This is a valuable approach once titles tend to contain critical information about a book's content, theme, and genre. Thus, clients who bought, for example, ‘A History of Japan’ are recommended books with similarities to the aforementioned, like, ‘China: A History’ or, ‘A history of Japanese literature'. This is also true for trilogies for example.

By leveraging natural language processing techniques and similarity measures, we aim to provide meaningful suggestions that go according to users' reading preferences titles.

The first step involves cleaning the dataset to eliminate stop words, such as ‘the’, ‘and’ or ‘of’. This increases the precision and accuracy of the recommendation system, as the common words could have led to a higher similarity between books that are not truly related in content or theme.

With the dataset prepared, we proceed to compute the similarity between books based on their titles, tailoring recommendations to each specific user based on their purchasing history with the store. Has a person has read more than 1 book, the code will return for each book not read more than one similarity, and so we sum them. This gives strength to books that are more similar. To recommend books that are better rated, we multiplied the similarity to the average rating of the book.

This feature will return a predefined number of the best recommended books to those previously bought by a customer. So, for each client, the system assesses all books already bought and suggests 15 others with a similar title.

## **FEATURE 2 – ITEM BASED COLLABORATIVE FILTERING**

Our second recommendation prediction uses item-based collaborative filtering, that provides recommendations based on the ratings that customers with similar tastes gave to other books. By comparing the books read by a user to those read by others with similar preferences, we can determine the similarity between items and suggest books that we believe the user is likely to also enjoy.

Relying on the assumption that users who enjoy a particular book are likely to also like other books that are similar to it, we computed the similarity between books based on the rating it was given by several customers. This made it possible for the company to predict the rating a specific customer would give to each book and return the 15 books with a higher predicted rating as a recommendation for a client’s further reading.

## **FEATURE 3 – USER BASED COLLABORATIVE FILTERING**

Our third recommendation feature is based on user-based collaborative filtering, which gives recommendations based on the similarity between users tastes rather than items. Relying on the assumption that users that have liked similar items in the past will continue to have similar tastes in the future, this model will find similarities between users based on their past ratings and recommend items that similar users have liked.

In this case, we consider that two users are similar if their ratings across all books are aligned. With this in mind, it predicts books’ ratings that the user has not yet read but might like once customers of the same segment and with a similar taste also liked.

For instance, clients who have given a high rating to the book ‘A History of Japan’ are deemed similar. Consequently, if a user opts to read 'A History of Japanese Literature' and gives it a high rating, the book will then be recommended to all other users who have also read 'A History of Japan' and rated it highly.

# **EVALUATION**

To evaluate the efficacy of our recommendation system, we decided to train the model with 70% of the data and subsequently test it with the remaining 30%. As such, it assesses all books read by a customer and returns 15 books that they should also be interested in buying, based on similarities and reviews.

Since the main goal is to come up with a list of books that are interesting to the user, our approach focused on having an accurate top 15 list of books that were recommended instead of accurately predicting a rating value from 1 to 5. Therefore, a success case would be if a book that was recommended, based on the training data, was reviewed/bought by the customer in the testing data, as this simulates a purchase.

When developing the recommender system, two main aspects need to be present. Recommendations must be relevant to the customers in order to generate sales, if good recommendations are mixed in with lots of bad recommendations, the model will be inefficient, so we must maximize True Positives; and at the same time, the model must avoid as much as possible missing a good recommendation, as that would leads to unrealized sales, so we must minimize misses as well. Based on these principles, we based our evaluation on the Precision and Recall, and the F1 score which joins the previous two.

From the scores obtained, it is clear the issues with the data sparsity have affected the outcomes of our models. Based on the output of content-based and the collaborative filtering-based models, we can determine that a hybrid model that joins the user-based and item-based collaborative filtering would provide the best recommendations to the customers.

While the overall objective of the model is to recommend customers books based on similar tastes, was achieved, the accuracy results are still very underwhelming. However, the reason for poor performance is not related to how the models were built but, in fact, to the low volume interactions within the data. While analyzing the results of the system, we understood that, given the short length of customer data, it proved very difficult to find similarities between clients, affecting the overall scores of the user-based and item-based models. And for the titles approach, titles alone are not enough to capture the content of books.

Since the reviews in the data set are not chronologically annotated, the results are not yet conclusive. When rolling out the models we will be able to evaluate the model better. It would be important to do an A/B testing with the 2 possible hybrid models (one with the title similarity bonus, and the other without). To evaluate the model we would see not only how the revenue increase, but also the sales of the books that are being recommended and the number of books bought in total. We could also do a survey so see how the satisfaction of the users increased with the new recommendation.

# **CONCLUSION**

## **MODEL ADVANTAGES**

The combination of the two features in the recommendation system enhances its prediction success. By considering the similarity between items, user interactions with each item and the rating given by each user to each item, the model can generate more robust recommendations. This approach caters a wider range of user preferences and improving the overall user experience.

## **MODEL DISADVANTAGES**

The reason why we decided to introduce a hybrid recommendation system, that combines the three different recommendations given by each different single-based method, is because each of them, if used alone, brings a lot of disadvantages to the company.

For instance, when it comes to the similarity between titles, the model is not very accurate capturing the essence of the book, because the number of words is low and although 2 words have a similar meaning, they will have a 0 score if there are not the same.

Moreover, the item-based collaborative filtering might exhibit bias toward popular items, as popular books might be recommended more often due to their high similarities with other well-liked choices. Also, the system does not adapt well to evolving customer preferences, providing static recommendations based on previous ratings. Lastly, it faces the cold start problem making it challenging for the model to give recommendations of new books or users as there is no past data.

The user-based collaborative filtering has the limitation that for this system to work properly, it has to be fed with lots of ratings by each user. A company that uses this system by itself needs to assure that it has data on many users and that each user has rated many books, otherwise the system has low accuracy on the recommendations made. Moreover, this system tends to recommend more often the books that are liked by the majority, leading to possibly disregarding books that are less popular but that some users may also like.

In conclusion, the hybrid recommendation system was created to minimize disadvantages, since, it gathers all the strengths of each individual model.

## **BUSINESS APPLICATION**

This recommendation system allows the company to achieve its goals and reverse the downward trend in sales. By utilizing this model, Livraria Lello can deliver personalized book recommendations to users, enhancing their experience and satisfaction, and, consequently, boosting the company’s overall sales. This approach offers users a "If you read this book, we think you will like these" marketing strategy, recommending books that other users who also rated the mentioned book enjoyed.

Furthermore, this system helps in identifying popular books or books that are frequently read together, which can be used to manage inventory effectively and create a better layout strategy, meaning that Livraria Lello could leverage this insight to position books that are frequently read together next to each other on the bookshelves. This way, customers would be easily influenced to buy other books while searching for a specific one in a bookshelf, as it might create curiosity to customers.

Moreover, the company can also present the user with the recommended books after their purchases through email marketing and other strategies, encouraging them to swiftly proceed with their next book acquisition.

Finally, Livraria Lello can also use these results to improve their post-purchase strategy. For instance, by clustering users with similar tastes, the company can create book clubs or discussion groups based on common interests, encouraging user engagement.

For example, for user 39096, the model predicts high ratings for books like "The Monk and the Riddle: The Education of a Silicon Valley Entrepreneur," "The Complete Sherlock Holmes Volume II," and "Dust Silo 3." By identifying if more users are also predicted to buy this books Livraria Lello can stock up its inventory align with these preferences and strategically position these recommended titles in the best locations within their stores, increasing visibility and influencing customers to make additional purchases.

Furthermore, personalized email reminders can influence users to complete their books series purchases. In the example for user 39096, the company could use email marketing to remind him to acquire "Dust Silo 3", that is recommended by out model, after probably purchasing volumes 1 and 2. This targeted approach not only enhances the customer experience but also increases customer satisfaction, as users feel valued and understood by the brand.

# **NEXT STEPS: HOW TO IMPROVE THE RECOMMENDATIONS**

As for next steps on how the model can be improved over time, we advise the company to increase its data quality using long-term methods to collect more data, such as the creation of a platform similar to GoodReads. By giving points or discounts in exchange for ratings, the customers would have an incentive to keep rating books which would give more data to the company.

Moreover, it would also be beneficial to get more information about books, such as the author, price or description of the book, and more information about the customer like age, location, education.

Furthermore, the a/b testing evaluation metric for the final recommendation system should be performed with and without the bonus of the title. Also, we should do NPL algorithms in the description of the books to find similar ones.

Lastly, it would be advised to do clustering of similar clients, to create incentives for them such as discounts or clubs.

**MODEL CONSTRAINTS**

1. The books are not in order by the data they were read, difficulting the process of recommendation.
2. We have a low number of recommendations by each user, the average in 2.9, difficulting the process to get to know the client.
3. Lack of information regarding the books (Year of publishing, age target).
4. Lack of information regarding the users (age, demography, education).
5. The recommendation system is difficult to evaluate before starting to be used.

**REFERENCES**

Gaurav, P. (2023, February 14). Step-by-step content-based recommendation system. Medium.<https://medium.com/@prateekgaurav/step-by-step-content-based-recommendation-system-823bbfd0541c>

Nat, S. (2023, September 14). Overcoming data sparsity with collaborative filtering. Medium. <https://medium.com/@sruthy.sn91/overcoming-data-sparsity-with-collaborative-filtering-e9a3519376f5>

AI & Insights. (2023, June 12). Building a recommendation system. Medium. <https://medium.com/muthoni-wanyoike/building-a-recommendation-system-11b13629a96>

Pandit, D. (2023, July 7). Recommender system: User collaborative filtering. Medium. <https://medium.com/@deepapandithu/recommender-system-user-collaborative-filtering-37613f0c6a9>

Polamuri, S. (n.d.). Five most popular similarity measures implementation in Python. Data Aspirant.<https://dataaspirant.com/five-most-popular-similarity-measures-implementation-in-python/>