Jumia Book Recommendation System

Business Understanding

Overview

What is a Recommendation System?

A Recommendation System is a type of artificial intelligence that provides personalized suggestions to customers. Previously, people would often rely on recommendations from friends to make purchasing decisions. However, with the rise of recommendation systems, companies like Google and Youtube can use data such as search history, watch history, and purchase history to suggest products or content that the customer may be interested in.

Reasons for a Recommendation System?

Over the past few decades, the emergence of web services such as Youtube, Amazon, and Netflix has led to the increasing prevalence of recommender systems in our daily lives. These systems suggest relevant items to users based on their preferences, from products to buy on e-commerce websites to content to view in online advertising.

By using a recommendation system, companies can generate substantial revenue and gain a competitive advantage over comptetiors by having a high customer retention. These systems are now so sophisticated that they can even provide recommendations to new customers who visit a site for the first time. They can suggest trending or highly rated products as well as items that are likely to generate the most profit for the company.

Types Of Recommendation System

A recommendation system is usually built using 3 techniques which are content-based filtering, collaborative filtering, and a combination of both.

1. Content-Based Filtering

The algorithm recommends a product or items that have similar characteristics to the ones that the user has already shown an interest in.. In simple words, In this algorithm, we try to find items that look alike.

2. Collaborative-based Filtering

Collaborative based filtering recommender systems are based on past interactions of users and target items. In simple words here, we try to search for the look-alike customers and offer products based on what his or her lookalike has chosen.

Problem Statement

As a rapidly growing online retailer, Jumia has a vast collection of books. Most self-published authors sell 250 books or less, regardless of how many different books they write. Traditionally published books sell around 3,000 copies on average, with only 250 of those sales in the first year. It's rare that books sell over 100,000 copies and even rarer to sell more than a million (How Many Books Do You Need To Publish To Make Money? - Letter Review, 2022).

An attributing factor to the poor sales of potentially good books in the public's eye is that there are too many books that major recommenders can read which can also be fueled by people sticking to specific authors rather than exploring due to the fear of reading a potentially different genre than what their tastes are accustomed to.

To address this issue, Jumia can implement an A.I model that uses a collaborative filtering technique. By doing so, they will not only recommend books based on the customers' ratings but also suggest books that are similar to the ones that the customer has shown an interest in.

Objectives

General Objective:

The general objective is to enable Jumia to effectively provide its customers with book recommendations that are tailored to the users.

Specific Objectives:

- To develop a machine learning model that can identify books that match the user's reading preferences by accurately predicting the rating a user will give to books they haven't read based on their previous ratings.
- To identify factors influencing a user's book preferences by analyzing the relationship amongst the individual features on the dataset
- To evaluate the performance of the recommendation system using appropriate metrics and compare it with other models.

Research Questions

i. How can we accurately match users with books that they will enjoy?

ii. What are the most important factors in determining a user's book preference?

iii. How can we measure the effectiveness of the book recommendation system?

Success Criteria

The success criteria we will follow depends on the predictive accuracy of the recommendations. This means we will rate how close the estimated ratings are to genuine use ratings, which is a measure used for evaluating non-binary ratings (e.g. 1-10 scale). Since selling books is crucial for a platform that is in business, this is the best metric we decided to use.

The two metrics that we will use are Mean Squared Error (M.S.E) and Root Mean Squared Error (R.M.S.E) due to the fact the rating scale is the same throughout.

Data Understanding

The Book Recommendation Data used in this project is from Kaggle. The data contained three files:

- **Users.csv**

  Contained the users. Some of the features in this file include user IDs (User-ID) which have been anonymized and map to integers. Demographic data such as (Location, Age) was also provided where it was available. Otherwise, these fields contained null values.

- **Books.csv**

  Contained books. Books were identified by their respective International Standard Book Number(ISBN).Some content-based information given (Book-Title, Book-Author, Year-Of-Publication, Publisher), was obtained from Amazon Web Services. In the case of several authors, only the first was provided. URLs linking to cover images were also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon web site.

- **Ratings.csv**

  Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

The number of books and users in the data were too many. As we had observed earlier, the dataset has 270,151 books and 92,106 unique users. The book ratings in the ratings dataset are a combination of implicit and explicit. The implicit ratings had a higher number of books than explicit which is understandable.This means there might have been users who registered to the website, viewed a few books but never read any of them or users who have read books but never rated them. This type of users cannot be relied on in the creation of the recommender system. Therefore the ratings were filtered to include only the explicit ratings

The three files were then merged together on their common columns to create one main dataset for easier data processing.

Data Preparation

Preparing the data from modeling was accomplished by:

i. Removing irrelevant columns

The removed columns were:

- ***Book-Title***: ISBN and Book-Title represent the same thing, hence one of the columns can be dropped
- ***Year-Of-Publication*****:** The year the book was published might have little effect on the model
- ***Image-URL-S, Image-URL-M, Image-URL-L*****:** These columns represent the URL to the book covers and will have no effect on the model

NB: The columns Book-Author and Publisher were kept since some readers may have the habit of reading books from a particular publisher or author because that's their niche. Hence recommending books from these authors/publishers might be beneficial.

ii. Dealing with duplicates

Duplicates were checked for using the *User-ID* and *International Standard Book Number(ISBN)* since we did not want users to rate the same book more than once.

The data was observed to have no duplicates.

iii. Dealing with missing values

Three columns were found to have missing values. That is the *Author, Publisher* and *Age* columns. For the *Author* and *Publisher* columns, the missing values were searched for online and replaced appropriately. For the *Age* column, the median was used to replace missing values since it is less affected by outliers which had been earlier observed in this column.

iv. Splitting location to represent the user country only

The values in the *Location* column were represented by the *city, state* and *country* of the user. These values were narrowed down to only the *country* of the user. The *clean_country* function in the *dataprep* module was then used to clean up these values further and replace the country names with the official country name. Missing values that resulted from this process were filled with the value "*Unknown*".

v. Checking the data types

The datatypes of all columns were correct except for the *Age* column whose data type was *float*. This was changed to *integer*.

Modeling

The modeling of this project consists of multiple recommender system implementations:

- **Popularity Based Recommendation system:**

  This focuses on getting the most positively rated books amongst the most viewers possible. It is a simple implementation, however, in most scenarios where knowing the users' data and books' information is not possible, this can apply due to the fact that it still will end up recommending books that will fit a good majority of the readers who rate.

- **Model-Based Collaborative Filtering Recommendation system:**

  We will be using the matrix factorization with SVD(Singular Value Decomposition) to ensure the latent features in our dataset will compare books and the users who rate them and produce a matrix that can produce estimates of books to better recommend them to potential readers.

Evaluation

Recommender systems have become an essential tool for online businesses to provide personalized recommendations to their users. The performance of a recommender system is often evaluated based on various metrics, such as precision, recall, F1-score, and so on. However, some types of recommender systems, such as the popularity-based ones, lack a performance metric as they provide recommendations solely based on the popularity of the items.

Despite this limitation, the popularity-based recommender can be used as an initial model when individual user data is unavailable. For instance, in situations where a new user signs up for a service, and the system has no data on their preferences, a popularity-based recommender can provide a general sense of the most popular or commonly viewed items. This approach can help the system to learn more about the user's preferences and provide more personalized recommendations later on.

On the other hand, a book recommender system with an RMSE of 1.47 indicates that the system's predicted ratings deviate from the actual ratings by an average of 1.47 stars. While this may seem like a significant difference, it is relatively low compared to the range of possible ratings. For example, if the rating scale is from 1 to 5 stars, an RMSE of 1.47 implies that the system's recommendations are accurate within about one star, on average. This relatively low RMSE value suggests that the system is providing fairly accurate recommendations.

It is also noteworthy that the statement mentions the removal of implicit data. Implicit data refers to the data that the system gathers indirectly, such as the user's browsing history or search queries. While this data can be useful, it may also introduce biases and noise into the recommendations. Therefore, removing implicit data can improve the accuracy and fairness of the system's recommendations.

Overall, the evaluation suggests that the book recommender system is performing quite well, especially after accounting for the removal of implicit data. However, it is essential to keep in mind that the performance of a recommender system is not solely determined by a single metric, and other factors such as user satisfaction and system scalability should also be considered.

Conclusion and Recommendations

A book recommender system that takes into account the individual preferences of each user is more likely to be effective than one that provides generic recommendations. Collaborative filtering, which involves analyzing the behavior and preferences of similar users to make recommendations, was used for building book recommender systems. The effectiveness of a book recommender system is heavily dependent on the quality of the data. For the model to run optimally the data fed to it should be accurate, relevant, and up-to-date. The performance of a book recommender system was evaluated using the Root Mean Square Metric(RMSE). This metric was used to identify areas for improvement and optimize the system for better performance.

Recommendations

- Ensure data quality. Ensure that the data used to train and test the model is accurate, relevant and up-to-date
- Incorporate more types of data, such as book genres and authors, to enhance the recommendation system's effectiveness.
- Regularly evaluate the performance of the book recommender system using appropriate metrics to identify areas for improvement and optimize the system for better performance.
- Develop a user interface that is intuitive and user-friendly, to improve user engagement and satisfaction.

References

How Many Books Do You Need To Publish To Make Money? - Letter Review (2022). Retrieved from:https://letterreview.com/how-many-books-to-publish-to-make-money/#:~:text=Most%20self%2Dpublished%20authors%20sell,sell%20more%20than%20a%20million.