**Book Recommendation System**

**Saurabh Shinkar**

**Data science trainee,**

**Alma Better, Bangalore**

**Abstract:**

### During the last few decades, with the rise of YouTube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers’ articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.

### In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries).

**Problem Statement**

### Recommendation systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users.

**Data Description**:

The Book-Crossing dataset comprises 3 files.

### **Users:**

Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

### **Books:**

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-Image-URL-M, Image-URL-L), i.e., small, medium and large. These URLs point to the Amazon website.

### **Ratings:**

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.

* **Steps involved:**
* **Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. That’s what we have tried to do.

* **Making data in proper format**

We made the Date column in proper format. We removed less correlated features with the target variable. Also removed independent features which are strongly correlated and kept only one of them.

* **Analyzing each feature separately**

For numerical features we look at the distribution of each feature, through Boxplot and Distribution Plot.

* **Fitting different models**

For modeling we tried various collaborative filtering Models:

➢ SVD

➢ SVD++

➢ NMF

➢ Slope One

* **Fitting different plots**

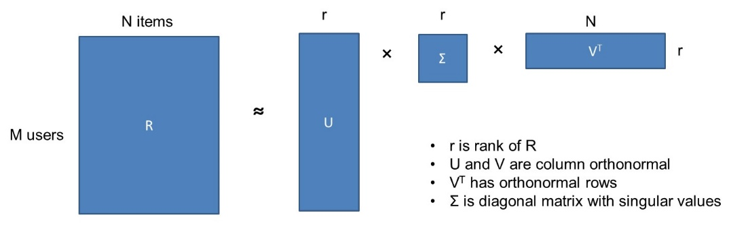
For modeling we tried various classification algorithms like:

1. **Box Plot**
2. **Bar Plot**
3. **Pie Chart**
4. **Density Plot**
5. **Count Plot**

**3. Models:**

1. **SVD:**

When it comes to dimensionality reduction, the Singular Value Decomposition (SVD) is a popular method in linear algebra for matrix factorization in machine learning. Such a method shrinks the space dimension from N-dimension to K-dimension (where K<N) and reduces the number of features. SVD constructs a matrix with the row of users and columns of items and the elements are given by the users’ ratings. Singular value decomposition decomposes a matrix into three other matrices and extracts the factors from the factorization of a high-level (user-item-rating) matrix.



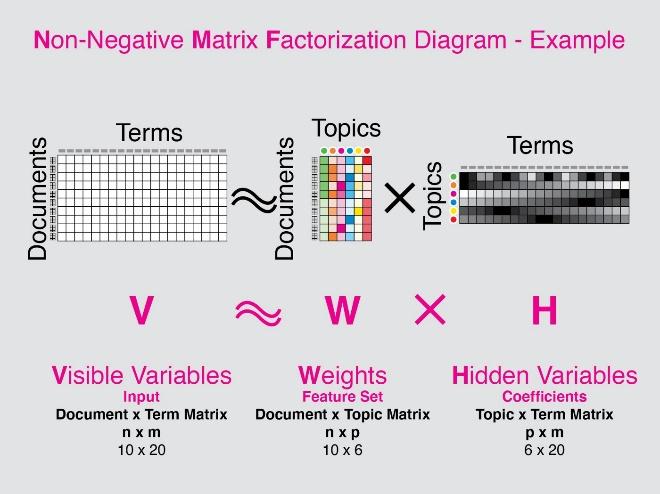
1. **SVDpp:**

The SVD++ algorithm is an extension of SVD that takes into account implicit ratings

1. **NMF:**

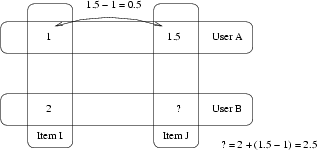
NMF is another method used for matrix factorization. Contrary to SVD, NMF decomposes the **non-negative** utility matrix R into the product of matrices *W* and *H*: Rn∗d=Wn∗rHr∗dRn∗d=Wn∗rHr∗d

Where columns in matrix Wn∗rWn∗r represent components, while matrix Hr∗dHr∗d stores the corresponding weights. More importantly, NMF introduces constraints under which: W≥0W≥0 and H≥0H≥0. The component-wise nonnegativity is a substantial difference from SVD (Gillis, 2017). Additionally, to collaborative filtering, one can find use cases of NMF in clustering, image processing, or music analysis.



1. **Slope One-**

**Slope One** is a family of algorithms used for [collaborative filtering](https://en.wikipedia.org/wiki/Collaborative_filtering), introduced in a 2005 paper by Daniel Lemire and Anna MacLauchlan. Arguably, it is the simplest form of non-trivial [item-based collaborative filtering](https://en.wikipedia.org/wiki/Item-item_collaborative_filtering) based on ratings. Their simplicity makes it especially easy to implement them efficiently while their accuracy is often on par with more complicated and computationally expensive algorithms. They have also been used as building blocks to improve other algorithms.



**Explicit Feedback Recommender Systems** These are systems where the user gives explicit feedback, usually in the form of a numeric rating for each recommendation.

Metrics used in Explicit Recommender Systems For such a system, the metrics used could be pretty similar to that used in a standard regression problem since the target is really a score that you could be predicting, and the actual score is available to measure how good the prediction is.

* **Mean Absolute Error**: Mean over all data points, absolute value of difference between actual rating and predicted rating.
* **Root Mean Square Error**: Square root of Mean over all data points, square of difference between the actual rating and predicted rating

**Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**Conclusions-**

* Wild Animus is the best-selling book
* Author Agatha Christie, William Shakespeare and Stephen King wrote most of the books
* Harlequin publication published the most books
* More than 50% readers are from USA
* Book-Ratings are negatively distributed with a median rating of 8.
* Root mean squared error of model **SVD** is 0.31 and mean absolute error is 0.21
* Root mean squared error of model **NMF** is 0.34 and mean absolute error is 0.24
* Root mean squared error of model **Slope One** is 0.39 and mean absolute error is 0.27
* **SVD++** is the **best recommendation model** with root mean squared error of 0.30 and mean absolute error of 0.20
* 