# Collaborative Filtering-based Recommendation System using Matrix Factorization

## (Group 13)

| Name | ASU ID | Email-ID | Role |
| --- | --- | --- | --- |
| Visaj Nirav Shah | 1225369210 | vshah47@asu.edu | • Model Building/ Training<br>• Model Evaluation and Testing<br>• Project Coordinator |
| Darsh Vaidya | 1225514966 | dvaidya5@asu.edu | • Model Evaluation and Testing<br>• Literature Review |
| Malvi Mungalpara | 1225184103 | mmungalp@asu.edi | • Data Cleaning<br>• Data Analysis<br>• Research |
| Meet Kishorbhai Lakhani | 1225636854 | mlakhan3@asu.edu | • Data Cleaning<br>• Data Analysis |
| Vibavari Gurunathan | 1225412305 | vnolas25@asu.edu | • Literature Review<br>• Documentation |
| Rohit Rajesh Rughwani | 1223264705 | rrughwan@asu.edu | • Model Building/ Training<br>• Model Evaluation and Testing<br>• Documentation |

## Problem Description

Data is growing at an exponential rate and so are the choices available online. Humans are faced with endless information daily; therefore, it is important that users find relevant results and are not lost in an overwhelming sea of data. Not finding relevant information time and again can discourage users from using a particular platform and lead to frustration. Thus arises the need for some system that prioritizes and delivers pertinent data.

Recommender systems deal with information overload and give personalized suggestions to the users by analyzing their current interests and product selection pattern from a large amount of dynamically generated data according to the user's behavior with respect to various items. They help them access the content they are interested in and sometimes would never have searched for.

Recommender systems are being used in abundance in e-commerce and entertainment industries like Netflix, Spotify, Amazon, etc. to enhance customer experience, gain competitive advantage, retain customers and drive sales. The use of recommender systems reduces the browsing time of the user and improves user satisfaction and engagement with the product. Since recommender systems work in real-time, the recommendations can change according to the changing habits of the consumer.

The goal of this project is to build a recommender system based on real-world Netflix data on user preferences for different content items.

## Preliminary Plan

<table>
<tr><th></th><th>Week 1</th><th>Week 2</th><th>Week 3</th><th>Week 4</th><th>Week 5</th><th>Week 6</th><th>Week 7</th></tr>
<tr><td>Insight</td><td colspan="2">Phase 1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Literature Survey</td><td></td><td colspan="2">Phase 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Data Analysis</td><td></td><td></td><td colspan="2">Phase 1</td><td></td><td></td><td></td></tr>
<tr><td>Model Building/Training</td><td></td><td></td><td colspan="4">Phase 2</td><td></td></tr>
<tr><td>Testing</td><td></td><td></td><td></td><td></td><td colspan="2">Phase 3</td><td></td></tr>
<tr><td>Evaluation</td><td></td><td></td><td></td><td></td><td></td><td>Phase 3</td><td></td></tr>
<tr><td>Report</td><td></td><td></td><td></td><td></td><td></td><td colspan="2">Phase 3</td></tr>
</table>

***Figure 1: Gantt Chart - Milestones***

## Phase 1

### Insight

The initial stage of the project requires brainstorming and research about the problem statement. Already a lot of research work has been carried out for recommendation systems. Our main agenda is to go through all the theoretical concepts related to recommendation systems. The different methods that exist for recommendation systems are content-based filtering and collaborative filtering. All the team members will carry out research for the problem statement, bring in their own ideas, and discuss individual goals. Further, team members will examine ideas with each other and compare different methods for the recommendation system like the one used by Netflix. Another main aspect of the research phase will include finding an appropriate dataset. Since good data results in a good model, we are planning to get data like real-world data and is extensive enough to train the model over it. After finalizing the data, it will be analyzed and pre-processed so that it can be fed into the model.

### Literature Survey

A literature survey is conducted to read papers relevant to recommendation systems and learn about ideas, research work, shortcomings, and future scopes. Recommender systems are mainly classified into three types: content-based, collaborative filtering, and hybrid approach. Content-based filtering works by creating profiles of products/users and recommending products based on previous user experiences. Collaborative filtering works on the similarity between various users. It tries to recommend products/items based on the data of a group of people that shared similar interests in the past. The hybrid technique is an amalgamation of two or more techniques, resulting in better results and overcoming the limitations of each model. Even though recommendation systems are prominent today, there are several limitations to them, some of which are a cold start, scalability, sparsity, and many more. We will try to address some of these issues in our model.

### Data Analysis

We will be using the “Netflix Prize Dataset" for our project that was provided by Netflix for their competition. The training data consists of more than 17 thousand movies data. For each, we have the customer id, their rating, and the date on which they rated the movie. It contains more than 4 lakh of user data. Rating is on an integral scale of 1 to 5. Apart from this, there is a separate file that contains metadata for movies. Since the dataset is real-world data, it would require pre-processing and cleaning. Data cleaning includes removing inconsistent and inept data. As part of data pre-processing, we will analyze the available data and come up with matrices that will help in deriving useful results. Exploratory Data Analysis takes account of a summary of data and other measures of the model. For data analysis, we will be using Python libraries like Pandas and NumPy and extract the useful features for the model.

# Phase 2

**Model Development and Training**

As discussed in the previous section, most of the recommendation systems can be classified into content filtering-based, collaborative filtering-based, and a combination of the two - Hybrid. There are pros and cons to both methods and here is a brief comparison -

- Collaborative filtering requires much less labeled data compared to the content filtering method.
- Content filtering cannot expand its recommendations beyond the users' existing tastes.
- Collaborative filtering can also leverage the power of implicit feedback in addition to explicit feedback, which is not possible in content filtering.

Hence, collaborative filtering is a much better method for building a recommendation system.

There are two more types of models possible in collaborative filtering - Neighborhood models and Latent Factor models. As the name suggests, Neighborhood models take into account the quality/preferences of similar items/users. In essence, like-minded individuals discover similar content. Latent Factor models focus on discovering hidden (latent) relationships between users and items which can correctly model the item-user affinity [1]. Matrix Factorization, the model we are building, is a type of latent factor model.

Matrix Factorization is a latent factor model that models user-item interactions as inner products in a joint latent factor space [2]. It is also called an embedding model. From a given feedback matrix A, the model tries to learn user (U) and item (V) embedding matrices. Row i in U denotes the embedding for user i and Row j in V denotes the embedding for item j. The embeddings U and V are learned such that UVT is as close as possible to A [3]. To avoid overfitting the model, we can add a regularization parameter in the error/ cost function.

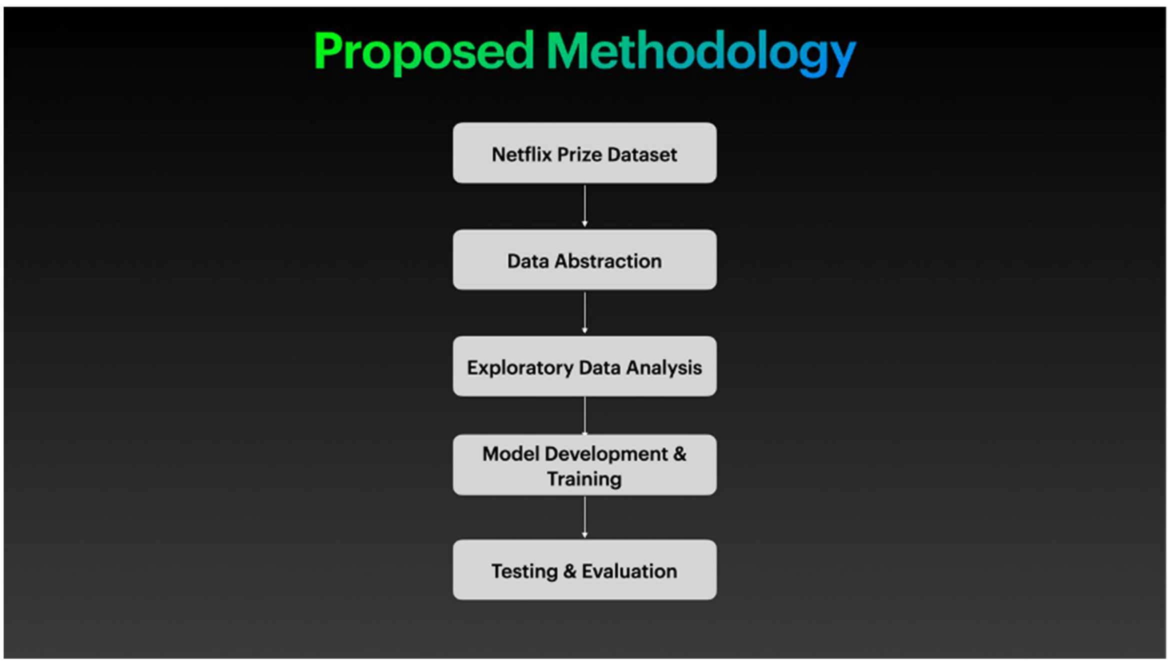
## Phase 3

### Testing and Evaluation

For testing and evaluation, we are exploring different performance metrics that can be used. There is the usual squared error metric, but in addition to that, there are metrics like MAP@K and MAR@K. Also, when it comes to recommendation systems, there are also different aspects to be tested like coverage and personalization. We will try to cover as many of these as possible in our results.

### Report

This part will involve combining all the work done for the project in a single coherent document. It will be extensively detailed with the process, results, and conclusions. The report will follow all the guidelines and formatting styles, as required by the Course Instructor.




## References

[1] (2011). Latent Factor Models and Matrix Factorizations. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_887

[2] Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," in Computer, vol. 42, no. 8, pp. 30-37, Aug. 2009, doi: 10.1109/MC.2009.263.

[3] https://developers.google.com/machine-learning/recommendation/collaborative/matrix