**Building user-based recommendation model for Amazon**

DESCRIPTION

The dataset provided contains movie reviews given by Amazon customers. Reviews were given between May 1996 and July 2014.

**Data Dictionary**  
UserID – 4848 customers who provided a rating for each movie  
Movie 1 to Movie 206 – 206 movies for which ratings are provided by 4848 distinct users

**Data Considerations**  
- All the users have not watched all the movies and therefore, all movies are not rated. These missing values are represented by NA.  
- Ratings are on a scale of -1 to 10 where -1 is the least rating and 10 is the best.

**Analysis Task**  
- Exploratory Data Analysis:

* Which movies have maximum views/ratings?
* What is the average rating for each movie? Define the top 5 movies with the maximum ratings.
* Define the top 5 movies with the least audience.

- Recommendation Model: Some of the movies hadn’t been watched and therefore, are not rated by the users. Netflix would like to take this as an opportunity and build a machine learning recommendation algorithm which provides the ratings for each of the users.

* Divide the data into training and test data
* Build a recommendation model on training data
* Make predictions on the test data

**Solution**:

We need to predict the ratings for the movies by each user. Based on the ratings, the movies can be recommended to the users. That is, if the predicted user rating for a movie is high, we can recommend that movie to the user.

For building the recommendation system, we use Surprise package’s SVD function.

Surprise is a Python scikit for building and analyzing recommender systems.

We use matrix factorization and SVD (Singular Value Decomposition) to build this recommendation system.

A screenshot of a computer

Description automatically generated with medium confidence

Notice that we have a lot of missing values in the dataframe; these are the movies that the user hasn’t rated, either because he hasn’t watched it or because he doesn’t want to watch or rate it. We use matrix factorization to predict the missing ratings.

We can think of all the ratings for movies by users as a matrix R. This matrix R can be factored into 2 smaller matrices. We break down / factorize the original matrix R into two smaller matrices so that we can recreate the matrix R by multiplying those two matrices. Once we have the smaller matrices, we can predict the rating of any movie by any user by taking a dot product of these matrices.

Using SVD, we can get these smaller matrices.

Chart

Description automatically generated

Table

Description automatically generated

Above dataframe shows the user id, movie id and the original rating by the user and the estimated rating. This is how we use SVD to predict ratings for the movies for each user. Using these ratings, we can make customized movie recommendation for the user.

Graphical user interface, text, application

Description automatically generated