MovieLens Recommendation System - Report

Uday Adusumilli

18 February, 2022

Contents

1 Executive Summary

The purpose for this project is creating a recommender system using MovieLens dataset.

The version of movielens dataset used for this final assignment contains approximately 10 Milions of movies ratings, divided in 9 Milions for training and one Milion for validation. It is a small subset of a much larger (and famous) dataset with several millions of ratings. Into the training dataset there are approximately **70.000 users** and **11.000 different movies** divided in 20 genres such as Action, Adventure, Horror, Drama, Thriller and more.

After a initial data exploration, the recommender systems builted on this dataset are evaluated and choosen based on the RMSE - Root Mean Squared Error that should be at least lower than **0.87750**.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

For accomplishing this goal, the **Regularized Movie+User+Genre Model** is capable to reach a RMSE of **0.8628**, that is really good.

2 Exploratory Data Analysis

2.1 Inital data Exploration

The 10 Millions dataset is divided into two dataset: `edx` for training purpose and `validation` for the validation phase.

The `edx` dataset contains approximately 9 Millions of rows with 70.000 different users and 11.000 movies with rating score between 0.5 and 5. There is no missing values (0 or NA).

edx dataset

Users	Movies
69878	10677

Missing Values per Column

The features/variables/columns in both datasets are six:

- **userId** `<integer>` that contains the unique identification number for each user.
- **movieId** `<numeric>` that contains the unique identification number for each movie.
- **rating** `<numeric>` that contains the rating of one movie by one user. Ratings are made on a 5-Star scale with half-star increments.
- **timestamp** `<integer>` that contains the timestamp for one specific rating provided by one user.
- **title** `<character>` that contains the title of each movie including the year of the release.
- **genres** `<character>` that contains a list of pipe-separated of genre of each movie.

	x
userId	0
movieId	0
rating	0
timestamp	0
title	0
genres	0

userId	movieId	rating	timestamp	title	genres
1	122	5	838985046	Boomerang (1992)	Comedy|Romance
1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
1	231	5	838983392	Dumb & Dumber (1994)	Comedy
1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi

First 6 Rows of edx dataset