

Capstone Project

Data Science with Python – Real World Project

MovieLens Dataset Analysis

The GroupLens Research Project is a research group in the Department of Computer Science and Engineering in the University of Minnesota. The researchers of this group are involved in many research projects related to the fields of information filtering, collaborative filtering, and recommender systems. Here, we ask you to perform the analysis using the Exploratory Data Analysis technique. In particular, we want you to apply the tools of machine learning to predict the survivors of the tragedy.

- *Import the three datasets*
- *Create a new dataset [Master_Data] with following columns MovieID, Title, UserID, Age, Gender, Occupation, Rating. (Hint: (i) Merge two tables at a time. (ii) Merge the tables using two primary keys MovieID & UserID)*
- *Explore the datasets using visual representations (graphs or tables) also include your comments on following:*
 - *User Age Distribution*
 - *User rating of the movie “Toy Story”*
 - *Top 25 movies by viewership rating*
 - *Find the ratings for all the movies reviewed by for a particular user of user id = 2696*

- *Feature Engineering:*

Use column Genres:

- *Find out all the unique genres (Hint: split the data in column genre making a list and then process the data to find out only the unique categories of genres)*
 - *Create a separate column for each genre category with a one hot encoding (1 and 0) whether or not the movie belong to that genre.*
- *Determine the features affecting the ratings of any particular movie.*
- *Develop an appropriate model to predict the movie ratings*

Good Luck!