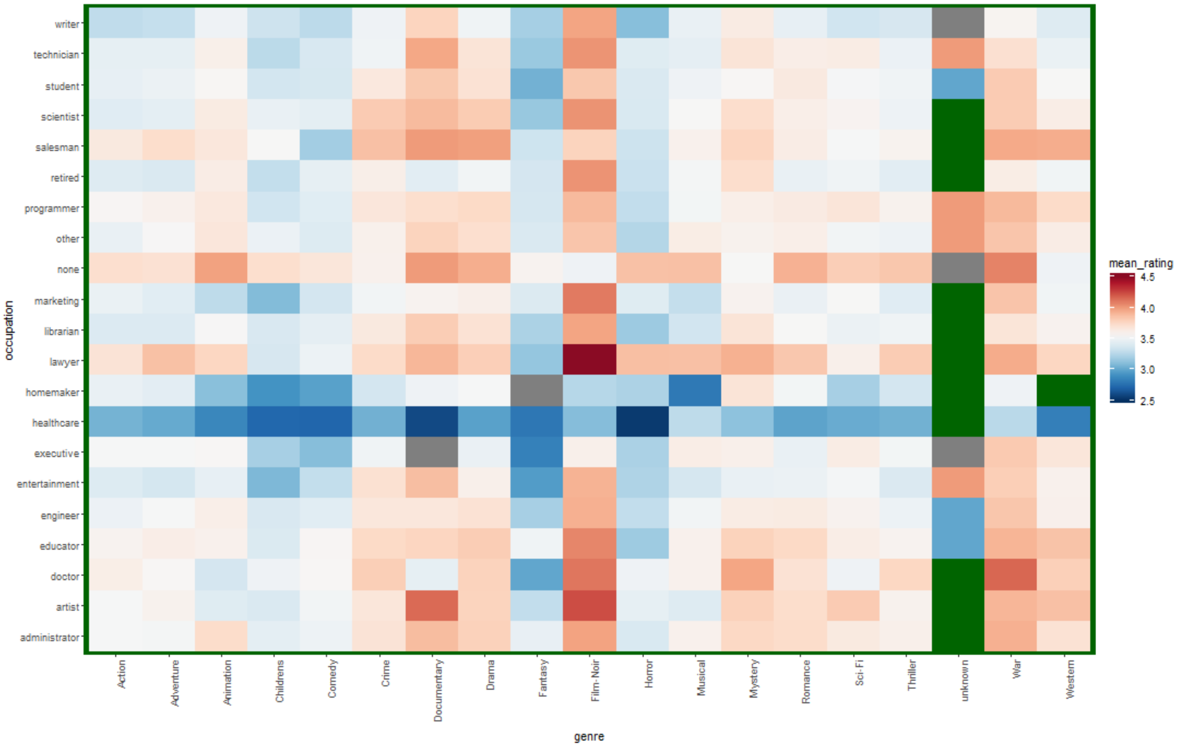
Recommender Systems Study Design

Kejia Shi, Li Luo, Yanan Huo, Yaqi Zhou

1. Introduction

Recommender systems change the way of providing a static experience in which users search for and potentially buy products. It increases interactions by identifying recommendations autonomously for individual users based on past purchases and searches, and on other users' behavior.

Given the 100K dataset from MovieLens, we are interested in predicting ratings for users and creating personalized recommendations for movies. Since we have information about movies and users in various datasets, we first clean and merge all data from MovieLens for further analysis.



The heat map above displays mean ratings for different movie genres and different occupations. Green tiles represent missing data. We can see that people whose occupation is health care tend to give lower scores. Therefore, there may exist rating bias, which makes normalizing data appropriate. Besides, it is evident that most occupations like film noir and documentary. Fantasy films and horrors don't get good ratings among most of the occupations.

This figure implies that demographic factors, like occupation, as well as genre, can be potentially powerful variables for rating prediction.

2. Key question

The question of primary interest is to predict ratings using algorithms in the `recommenderlab` package and other machine learning approaches and then compare these methods.

3. Plan of approach

1). Data visualization and relationship exploration

- Is there any relationship between `rating` and information about films, e.g. `genre`, as well as demographic variables, e.g. `age, occupation, gender` and `zip_code`?

2) Approach 1: Rating prediction using `R/recommenderlab` package

- How can we predict ratings using algorithms based on rating matrix?

3) Approach 2: Rating prediction using other machine learning approaches

- If there exists potential powerful factors in part 1), how can we utilize these associations to predict ratings for users?

4). Evaluation of rating predictions and method comparison

- After constructing models, we will evaluate and compare them based on prediction accuracy and interpretability.

4. Data source

MovieLens 100K dataset: http://files.grouplens.org/datasets/movielens/ml-100k/

5. Reference

1. Justin Chu. *cleanMovieLensData.R*. Retrieve from https://github.com/JustinChu/STAT545A_MovieStats/blob/master/cleanMovieLensData.R