MovieLens Recommendation System

By: Adina Steinman & Matthew Zhang

Business Problem

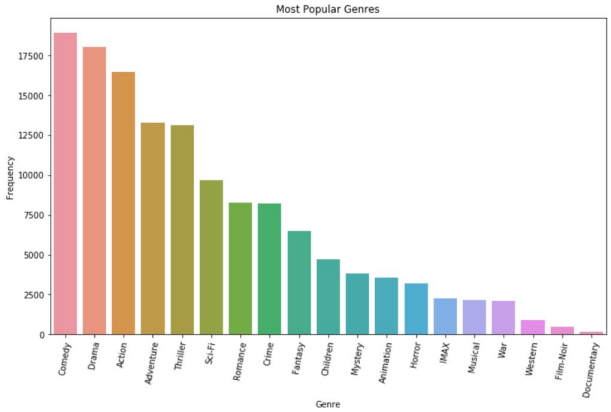
- Our team of data scientists work together on a team for a new growing movie platform looking to compete with other giants such as Netflix, Hulu, and HBO.
- Our aim is to create a unique experience for each of our customers while subtly increasing our ROI.
- We aim to achieve this by building a tailored, unique recommendation system that can effectively suggest movies to our users, in order to continue to engage them with our platform.
- We want to be able to make predictions to our existing clients, as well as use our recommendation system as a product we use to attract new users to our platform.

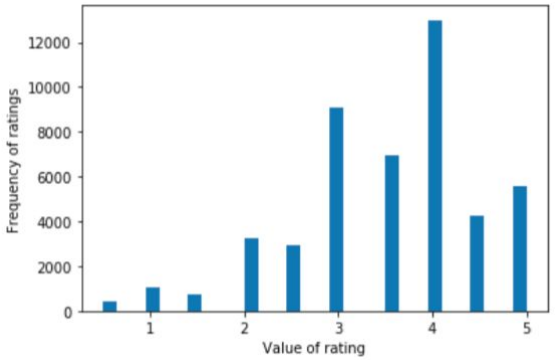
Datasets

- Our team used “MoveLens” database, developed by the GroupLens research lab at the University of Minnesota
- Initial consisted of 100,000 user ratings; then was reduced to 47,000 once both users and movies with low number of ratings were removed to improve matrix sparsity
- Final DataSet included information on userId, movieId, rating, title, genres and year

	userId	movieId	rating	title	genres	year
0	1	1	4.0	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1995
1	1	3	4.0	Grumpier Old Men (1995)	Comedy|Romance	1995
2	1	6	4.0	Heat (1995)	Action|Crime|Thriller	1995
3	1	47	5.0	Seven (a.k.a. Se7en) (1995)	Mystery|Thriller	1995
4	1	50	5.0	Usual Suspects, The (1995)	Crime|Mystery|Thriller	1995

Exploratory Data Analysis





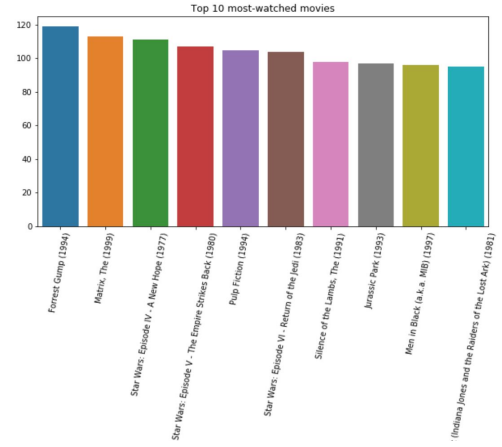
- Generated feature ‘year’ to see if certain years had higher rated films than others
- Created genre labels in order to determine the most popular genres in the dataset; top 3 were Comedy, Drama, and Action
- Analyzed the distribution of ratings: a rating of 4 was the most frequent
- Analysed the top 10 movies overall, as well as by genre, to use as a future comparison to post-model EDA to determine presence of popularity bias

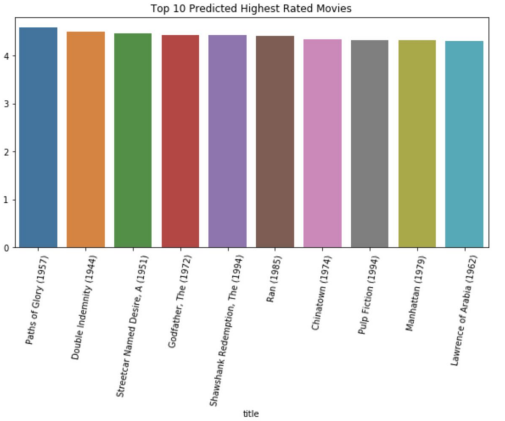
Model Results

- Several models, KNNBasic, KNNBaseline and KNNWithMeans, all performed worse than our baseline model
- SVD was our best performing model with an RMSE, MAE and CV all lower than the baseline
- SVD RMSE of 0.7897 tells us that our model estimates ratings with an error of approx 0.8, on a rating scale of 0-5

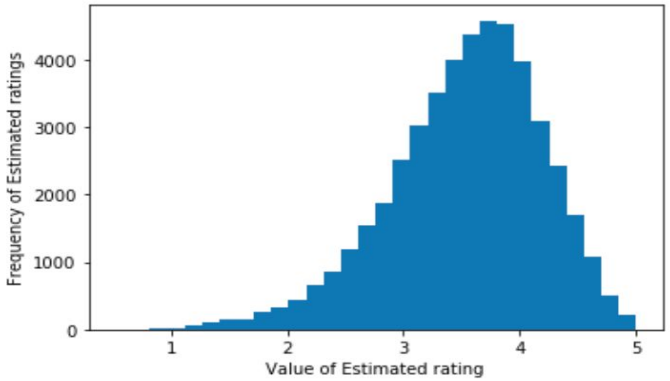
	RMSE	MAE	CV
model			
baseline	0.801115	0.612688	0.815118
svd	0.789745	0.604842	0.808279
knnbasic	0.889289	0.687625	0.902386
knnbaseline	0.804648	0.615868	0.821286
knnwm	0.813338	0.622101	0.828150

Post-Modeling EDA





- Top 10 most-watched based on estimations show similar to films from our original ratings, no popularity bias in our recommendation system
- The distribution of estimated ratings follows a similar pattern to our original ratings, with majority of ratings around a rating of 4.0



Conclusions and Future Work

- Overall, our model does a fairly decent job of estimating users' ratings, with an approximate error of 0.78
- Our model is a purely collaborative filtering model, and therefore does not address the *"cold start problem"*
  - Future work should look to incorporate aspects of a content based filtering model to address this, specifically through the use of LightFm
  - Incorporation of features, such as genres and year, into our matrix factorization models
- Validation of our models:
  - It is difficult to validate the results of our estimations; however, it would be interesting to see if users do end up watching the movies recommended to them, and if so, how they are rated

Thank You!

- MovieLens DataSet: https://grouplens.org/datasets/movielens/latest/
- Github: https://github.com/adinas94/Phase4Project
- Contact information:
  - Adina Steinman: adinasteinman@gmail.com
  - Matthew Zhang: mattzhang989@gmail.com