Movie Recommendations

a project by:

Abdullah Adlouni
Ben Healy
Charles Dorman
David Shackelford

Project Description

Purpose

- To get a better understanding of movie ratings
  - Explore recommendation systems
- Combine multiple data sources to try and get useful information
- Find interesting data which shows relationships in what people enjoy watching

Major Questions

1. Can we make recommendations of movies based on prior customer ratings?
2. Can we create inferences based on attributes of a movie?
3. Can we observe trends in the movie industry and how people enjoy movies?

Datasets

Description of Datasets

- Sources
  - Netflix - 17,770 movies rated by 480,189 users, > 100M ratings.
  - IMDB - 7,980,307 movies/shows/shorts objects with ratings, cast info, etc.
  - TMDB - Subset of ~5k movies from IMDB, but with info on revenue and budget

Data Preparation

- **Import** data sets from csv and tsv files
- **Merge** data sets
  - Split titles into words, made all lowercase
  - Use ML model to encode titles as fixed-length vectors to allow for finding cosine-similarity
  - Merge sets based on titles' cosine similarity
- **Reduce** data
  - Drop unpopular titles in lower 30% of ratings
  - Drop low-activity users

Tools

Tools

Project Management

- Github
- AWS
- Discord

Data Analysis

- Pandas
- Numpy
- Regex
- Bash

Machine Learning

- Scikit-Learn
  - Sklearn-surprise : SVD based Recommendation
  - Cosine Similarity
- Tensorflow
  - Bert-as-Service

Results

Results - Integration

- 3 methods - 1:1, NLP + cosine similarity, algorithmic approach

	Mismatched Pairs (/100)	Unmatched Pairs	Time
1:1 Matching	0	11,213	15m
NLP + Cosine Similarity	6	6,516	3d 8h
Algorithmic	23	5,731	4h

Results - Recommendation

Fold	RMSE	MAE	Fit Time (s)	Test Time (s)
1	0.825	0.703	1045.862	432.187
2	0.824	0.701	1072.961	308.564
3	0.824	0.701	958.641	203.421
4	0.824	0.701	897.596	198.323
5	0.824	0.701	932.118	221.168
Avg.	0.824	0.701	981.436	272.73

Results - Recommendation

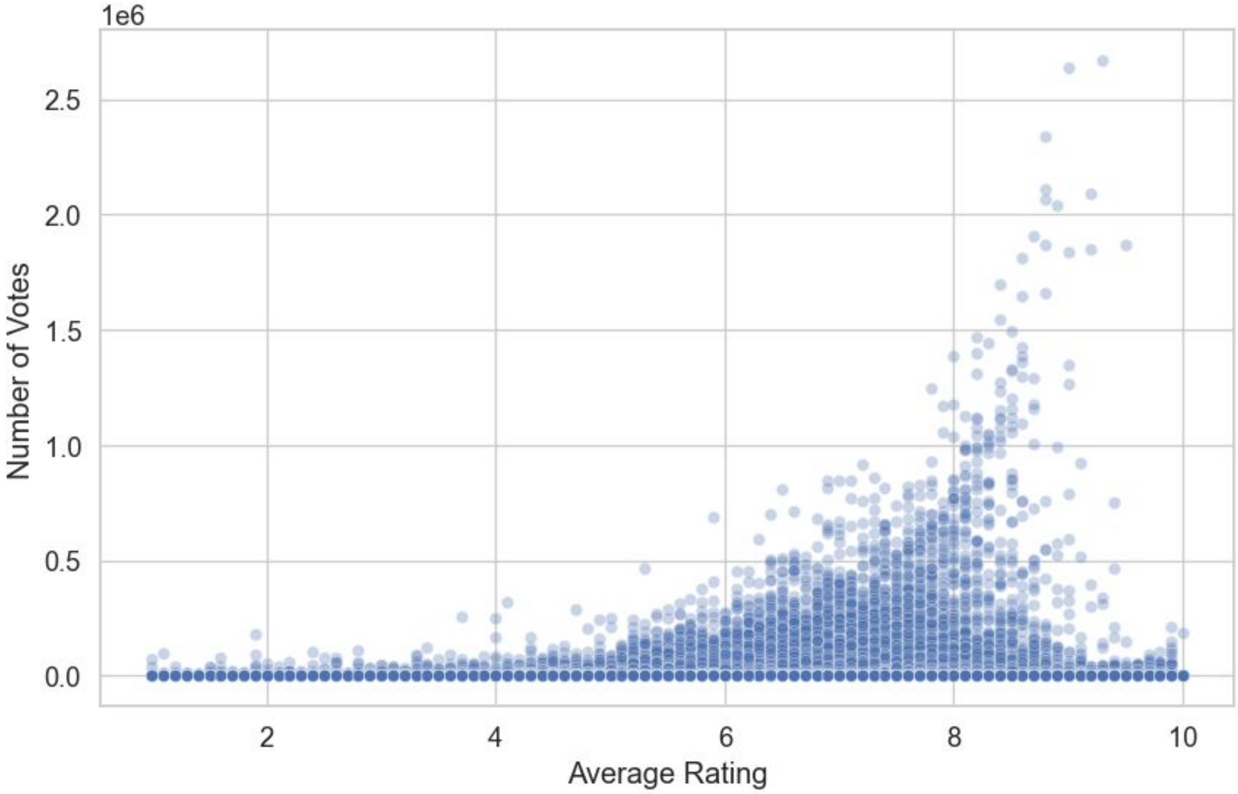
Fold	RMSE	MAE	Fit Time (s)	Test Time (s)
1	0.888	0.719	600.917	202.503
2	0.888	0.718	603.431	137.481
3	0.888	0.718	602.818	128.296
4	0.888	0.718	603.871	203.589
5	0.888	0.717	599.386	120.589
Avg.	0.888	0.718	602.085	158.491

Results - Recommendation

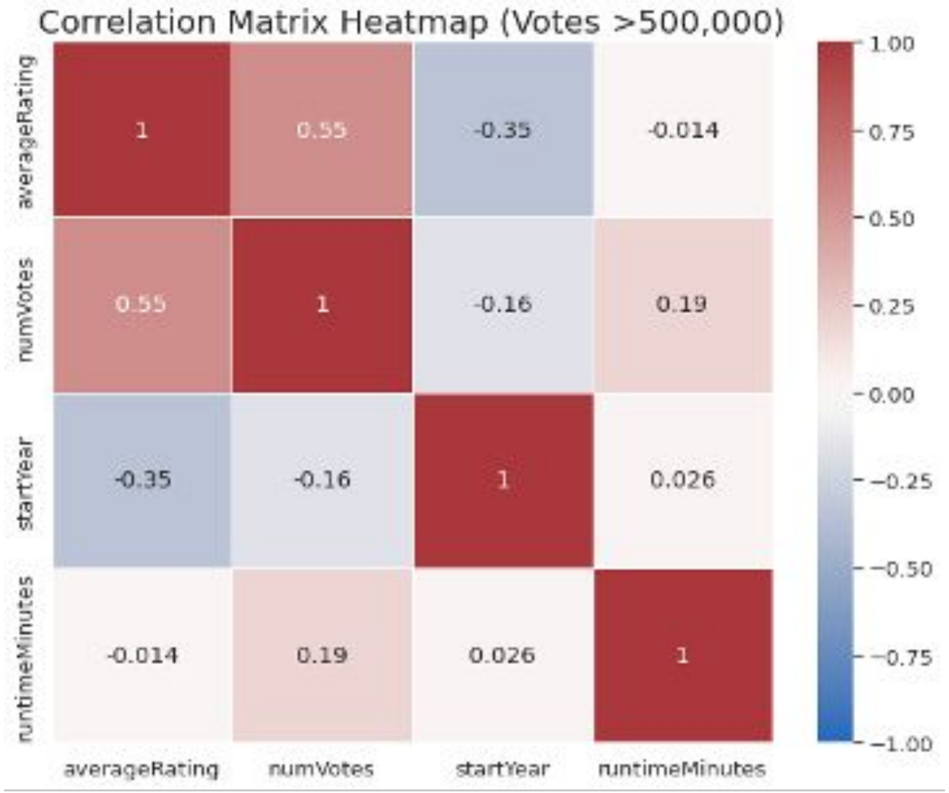
- Simple Ensemble: (Movie + Genre)/2 = Ensembled_Rating

	RMSE	MAE
Movie Recommender	0.824	0.701
Genre Recommender	0.888	0.718
Ensembled	0.749	0.603

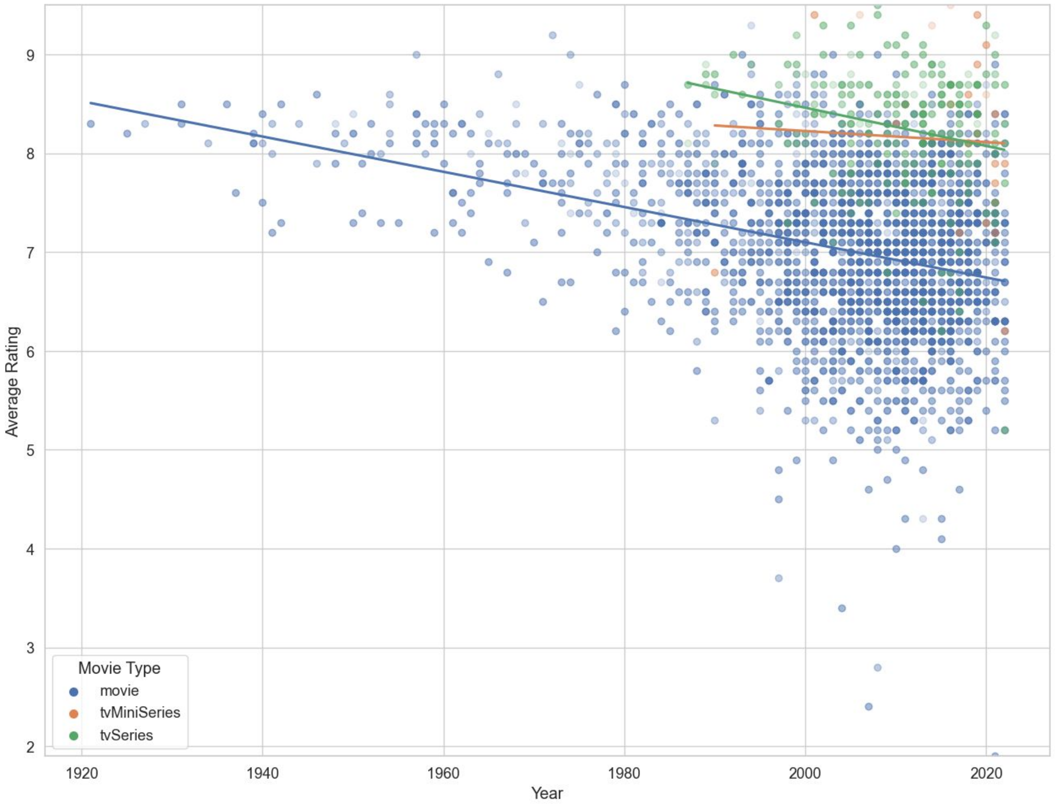
Results - Data Exploration



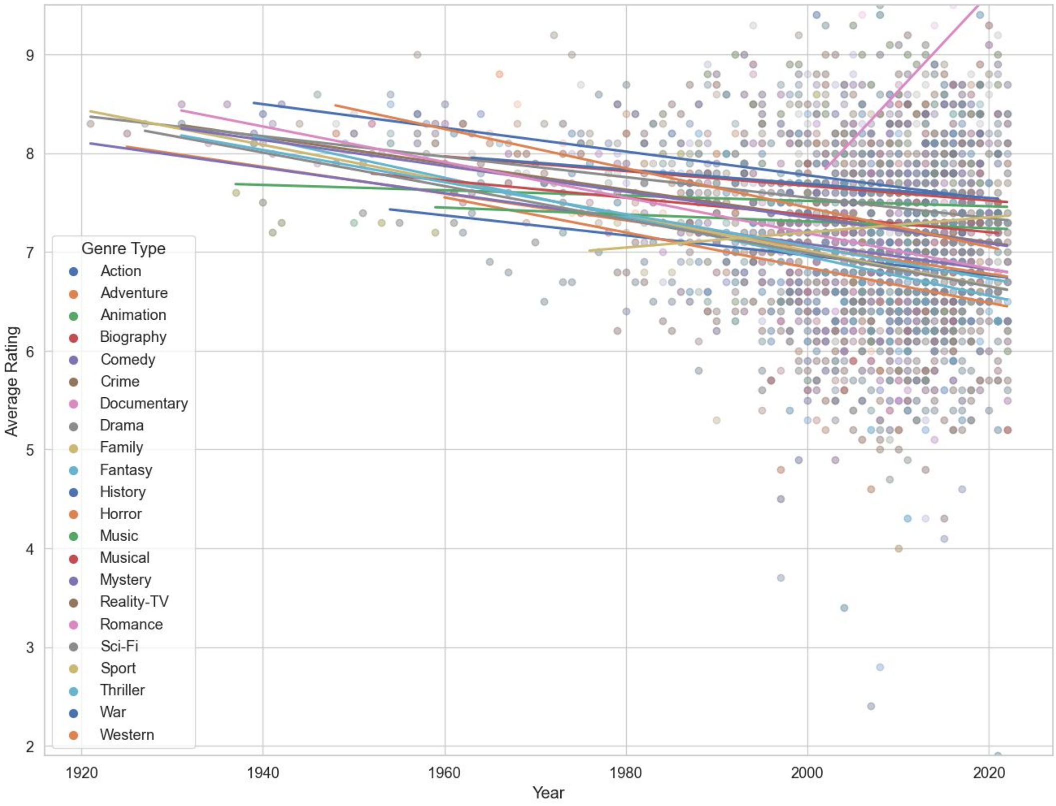
Results - Data Exploration



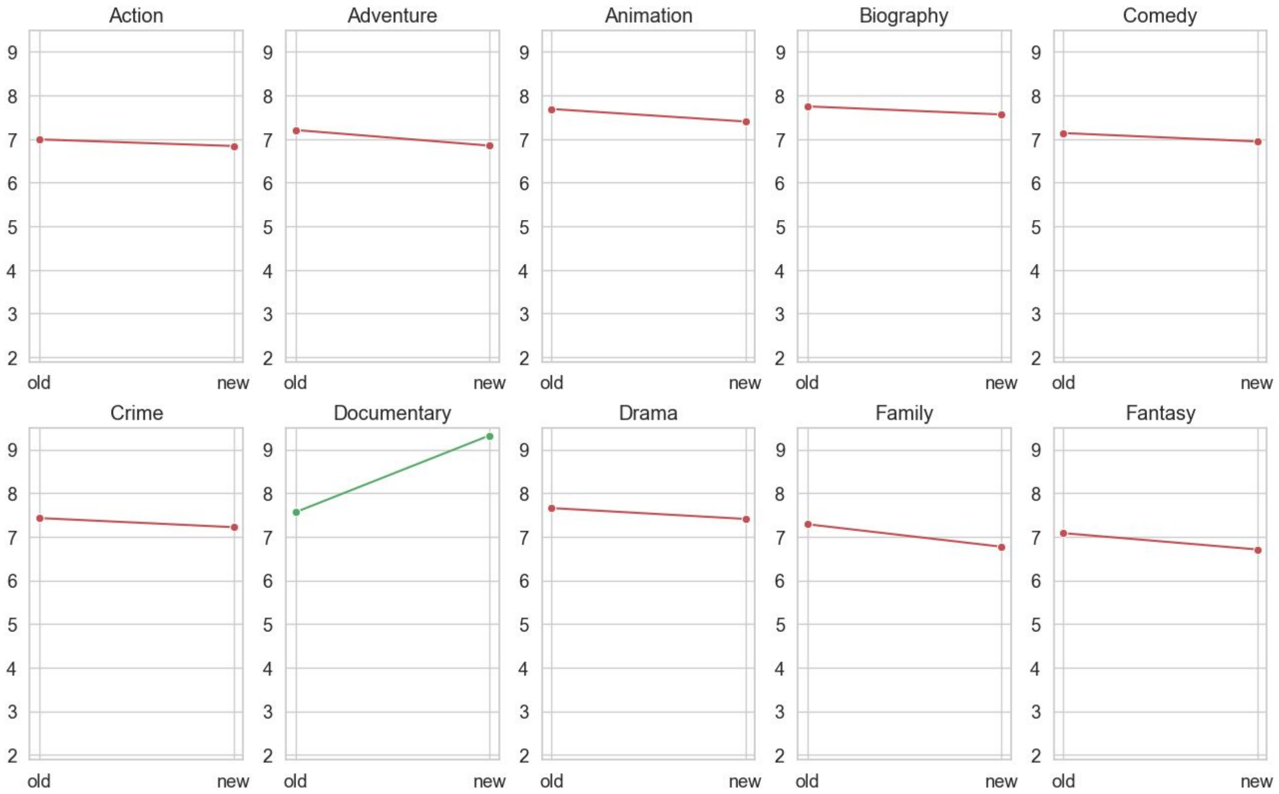
Results - Data Exploration



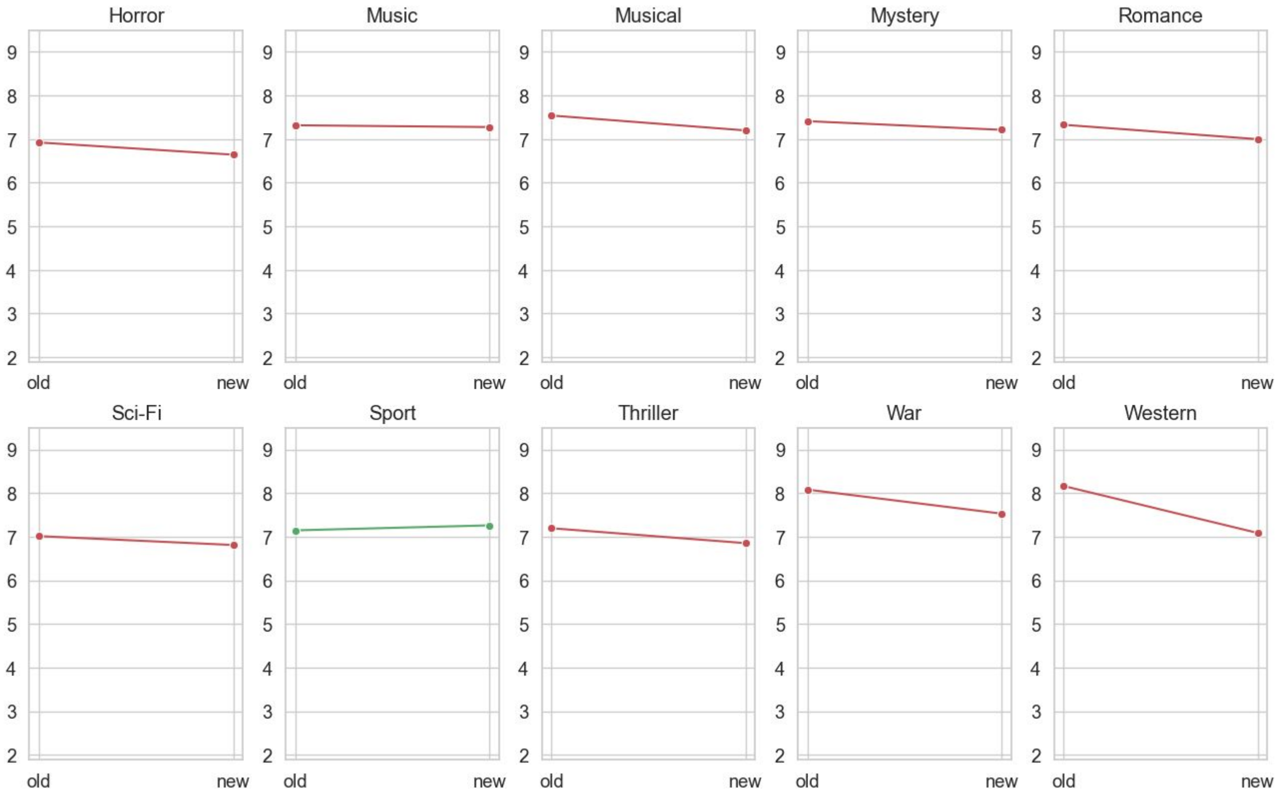
Results - Data Exploration



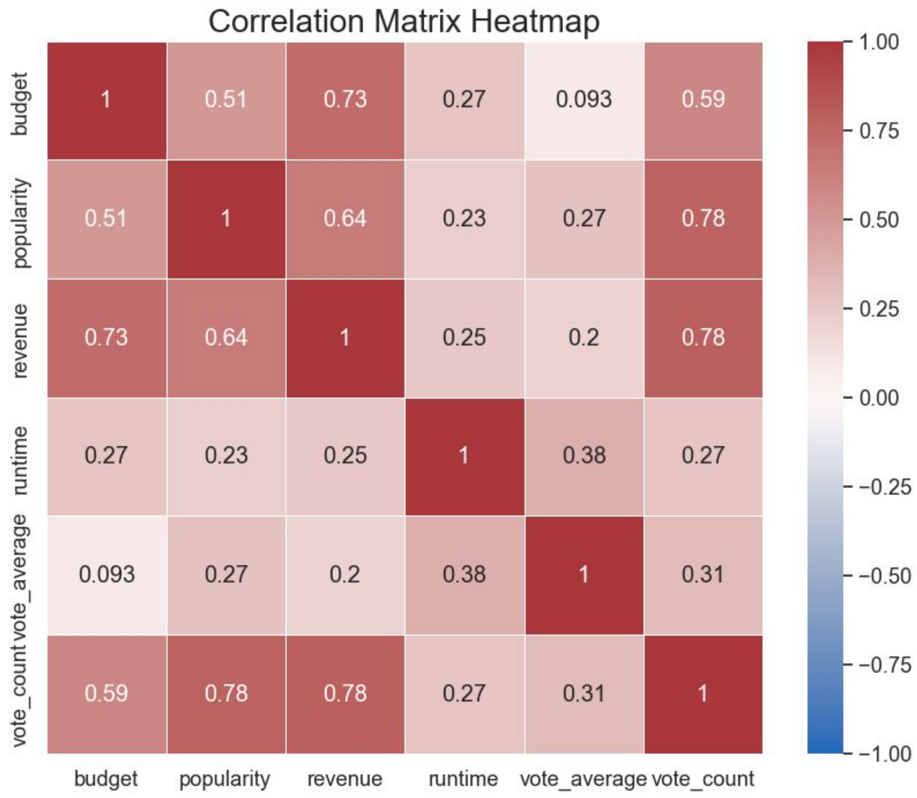
Results - Data Exploration



Results - Data Exploration



Results - Data Exploration



Knowledge Gained

Knowledge Gained - Data Analysis

Achieved improved recommendation model performance by averaging the results of multiple models (ensemble methods)

Knowledge Gained - Data Analysis

- Newer movies tend to have greater variance in ratings than older movies (survivorship bias? selection bias?)
- Documentaries are trending well (visuals becoming cheaper?)

Knowledge Gained - Application

- Credibility-weighting users' rating by correlation with target user had significant benefit
- Breaking out movies by genre before finding correlation increased predictive value minimally (each movie listed in multiple genres–overlap loses distinction?)

Knowledge Gained - Application

- Producers could respond to trends in genres of increasing popularity uncovered in the IMDB ratings
- Our results: feature-length and TV documentaries and movies and TV serials about sports.

Thank you!