Our Movie Recommendation System

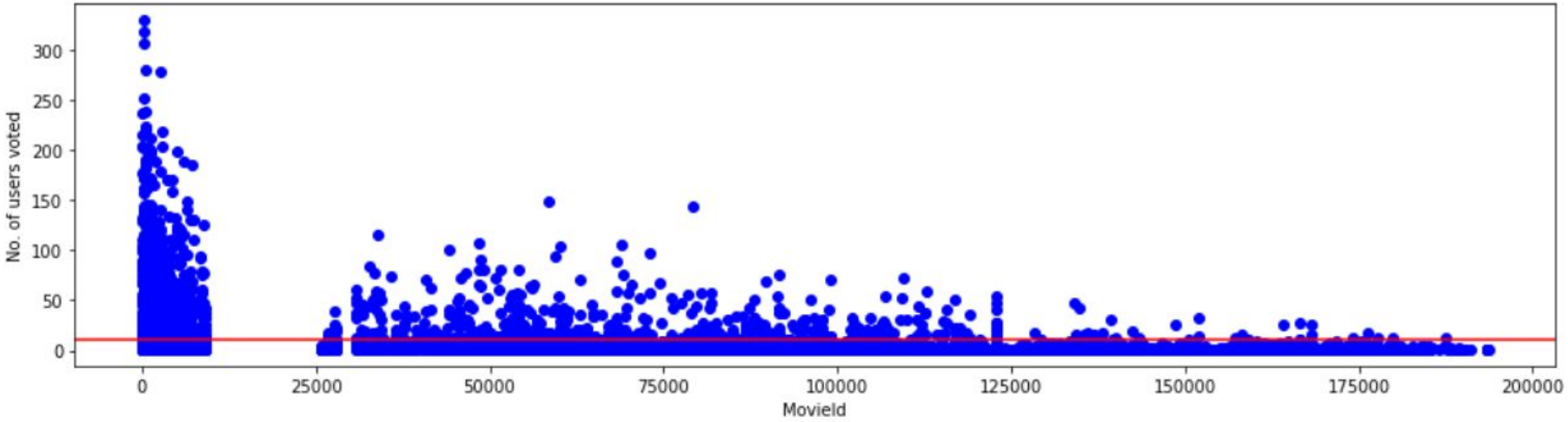
Alex Valencia
Mallory Wilson

Data Cleaning Steps

- Selected only two datasets to create our recommendation system model:
  - ‘movies’
  - ‘ratings’
- Used ‘movieId’, ‘userId’, and ‘rating’ columns from ‘ratings’ dataset.
- Used ‘movieId’ and ‘title’ columns from the ‘movies’ dataset.

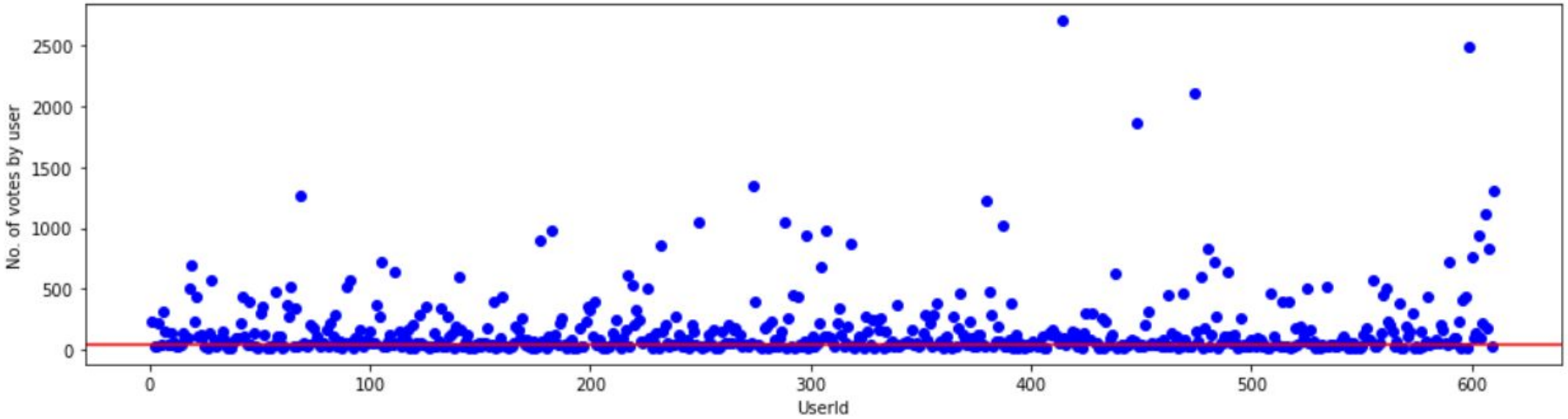
Building the Baseline Model

The Number of Users Who Voted for Each Movie



- The red line represents the minimum amount of users who rated a movie.
- For this model, each movie needs 10+ reviews to be included in the recommendation system.

How Many Times a User Voted for Movies



- The red line represents the minimum amount of votes a user has given for movies.
- For this model, each user needs to vote for 50+ movies to be included in the recommendation system.

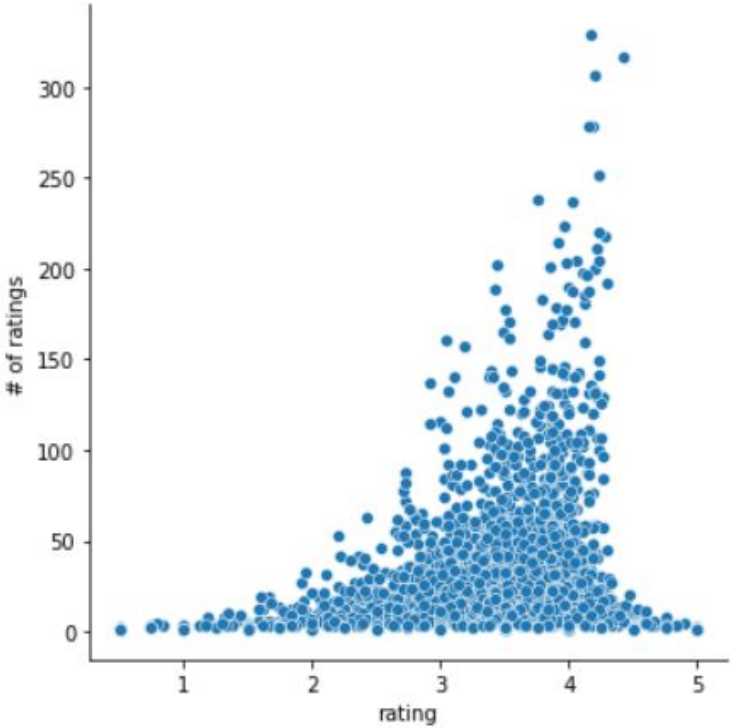
Final Steps Building Our Baseline Recommendation System Model

- CSR Matrix
  - Reduces sparsity in our final dataset.
- NearestNeighbors
  - Algorithm finds a certain number of movies that are the closest distance to your movie selection.
- Movie Recommendation Function
  - Enter a movie title to find the top 10 most similar movies based on similarity distance.
- RMSE Score: 0.8715

Building the Final Model

How the Number of Ratings Impact the Rating

- Rating on scale of 1-5
- Number of rating is the number of ratings for each movie
- As the number of ratings increase, the rating score also increases
- The higher the number of ratings seems to give a more accurate rating score



Final Steps in Creating our Final Model

- Matrix Factorization
  - Finds the user ratings, using userId and rating, for different movies and correlates the user rating with a different movie
- Only uses movies with 50 or more ratings
- Enter a movie into the function to find the 10 most correlated movies
- Final Model RMSE Score: 0.7877
- Baseline Model RMSE Score: 0.8715

Final Model Output Compared to Baseline Model Output

Baseline Model vs. Final Model: Movies Similar to Harry Potter and the Chamber of Secrets

	Title	Distance
1	Spider-Man (2002)	0.398373
2	Ice Age (2002)	0.397131
3	Harry Potter and the Half-Blood Prince (2009)	0.394569
4	Lord of the Rings: The Two Towers, The (2002)	0.391141
5	Pirates of the Caribbean: The Curse of the Bla...	0.367197
6	Pirates of the Caribbean: Dead Man's Chest (2006)	0.349314
7	Harry Potter and the Order of the Phoenix (2007)	0.346729
8	Harry Potter and the Goblet of Fire (2005)	0.265915
9	Harry Potter and the Prisoner of Azkaban (2004)	0.208909
10	Harry Potter and the Sorcerer's Stone (a.k.a. ...	0.196221

Movie Title	Distance	# of Ratings
Harry Potter and the Sorcerer's Stone (a.k.a. Harry Potter and the Philosopher's Stone) (2001)	0.884597	107
Harry Potter and the Order of the Phoenix (2007)	0.825499	58
Harry Potter and the Half-Blood Prince (2009)	0.692219	58
Harry Potter and the Goblet of Fire (2005)	0.680157	71
Harry Potter and the Prisoner of Azkaban (2004)	0.662563	93
Star Trek (2009)	0.662398	59
Pretty Woman (1990)	0.658711	135
28 Days Later (2002)	0.655416	58
How to Train Your Dragon (2010)	0.652580	53
Deadpool (2016)	0.635105	54

Baseline Model vs Final Model: Movies Similar to Inception

	Title	Distance
1	Hangover, The (2009)	0.369214
2	Iron Man (2008)	0.369175
3	Fight Club (1999)	0.367898
4	Sherlock Holmes (2009)	0.366418
5	Django Unchained (2012)	0.362976
6	Shutter Island (2010)	0.345888
7	Avengers, The (2012)	0.340302
8	Dark Knight Rises, The (2012)	0.335075
9	Inglourious Basterds (2009)	0.305288
10	Dark Knight, The (2008)	0.213876

Movie Title	Distance	# of Ratings
Full Monty, The (1997)	0.909324	56
Crow, The (1994)	0.859338	64
Grumpier Old Men (1995)	0.836421	52
Hook (1991)	0.817990	53
Desperado (1995)	0.807960	66
Interview with the Vampire: The Vampire Chronicles (1994)	0.789059	109
Maverick (1994)	0.773249	74
Leaving Las Vegas (1995)	0.753912	76
Animal House (1978)	0.739351	62
Dances with Wolves (1990)	0.734847	164

Baseline Model vs. Final Model: Movies Similar to Shrek 2

	Title	Distance
1	Star Wars: Episode III - Revenge of the Sith (...	0.422463
2	Spider-Man 2 (2004)	0.405562
3	Harry Potter and the Prisoner of Azkaban (2004)	0.392267
4	Spider-Man (2002)	0.384256
5	Ice Age (2002)	0.372899
6	Incredibles, The (2004)	0.346440
7	Finding Nemo (2003)	0.346279
8	Monsters, Inc. (2001)	0.340306
9	Pirates of the Caribbean: The Curse of the Bla...	0.333260
10	Shrek (2001)	0.302120

Movie Title	Distance	# of Ratings
Shrek (2001)	0.821400	170
Gone in 60 Seconds (2000)	0.642714	61
How to Train Your Dragon (2010)	0.639178	53
Eraser (1996)	0.635763	64
Sense and Sensibility (1995)	0.618499	67
Johnny Mnemonic (1995)	0.611842	53
Field of Dreams (1989)	0.596883	56
Monsters, Inc. (2001)	0.591414	132
Kung Fu Panda (2008)	0.579984	54
Mr. Holland's Opus (1995)	0.568374	80

Next Steps

- Our recommendation system model works well, but contains under 10,000 movies.
  - Include more movies for better performance.
- Make a recommendation system model using movie tags and genres.
  - Explore how movie tags and genres are correlated to other movies with the same tags and similar genres.
- Find the demographics of each user to better predict which movies they will want to watch next.
  - Age, Ethnicity, Geographic Location

Thank You

Alex Valencia
Email: asvalencia1688@gmail.com
GitHub: @asval211

Mallory Wilson
Email: mallorye1103@gmail.com
GitHub: @malloryewilson

Questions?

Index

Understanding the ‘ratings’ columns

userId - Users were selected at random for inclusion. Each user is represented by an id and no other information is provided.

movieId - Only movies with at least one rating or tag are included in the dataset. These movie ids are consistent with those used on the MovieLens web site (e.g., id 1 corresponds to the URL https://movielens.org/movies/1). Movie ids are consistent between ratings.csv, tags.csv, movies.csv, and links.csv (i.e., the same id refers to the same movie across these four data files).

rating - Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars).

timestamp - Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

Understanding the ‘movies’ columns

movieId - Only movies with at least one rating or tag are included in the dataset. These movie ids are consistent with those used on the MovieLens web site (e.g., id 1 corresponds to the URL https://movielens.org/movies/1). Movie ids are consistent between ratings.csv, tags.csv, movies.csv, and links.csv (i.e., the same id refers to the same movie across these four data files).

title - Movie titles are entered manually or imported from https://www.themoviedb.org/ and include the year of release in parentheses. Errors and inconsistencies may exist in these titles.

genres - Genres are a pipe-separated list, and are selected from the following:

Action, Adventure, Animation, Children’s, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, (no genres listed)

Preprocessing Steps

- Reshaped ‘ratings’ dataset in the following way:
  - Set ‘movieId’ as the index.
  - Set ‘userId’ as a column.
  - Set ‘rating’ as values within the dataset.
- Replaced missing values with 0.
- Displayed all these changes in a new dataset.