Project:
Movie Recommendation System

Author Name: Ganesh R Bajaj

Problem Statement

To Build Model that can Recommend Similar Movies to a user which he/she can find interesting.

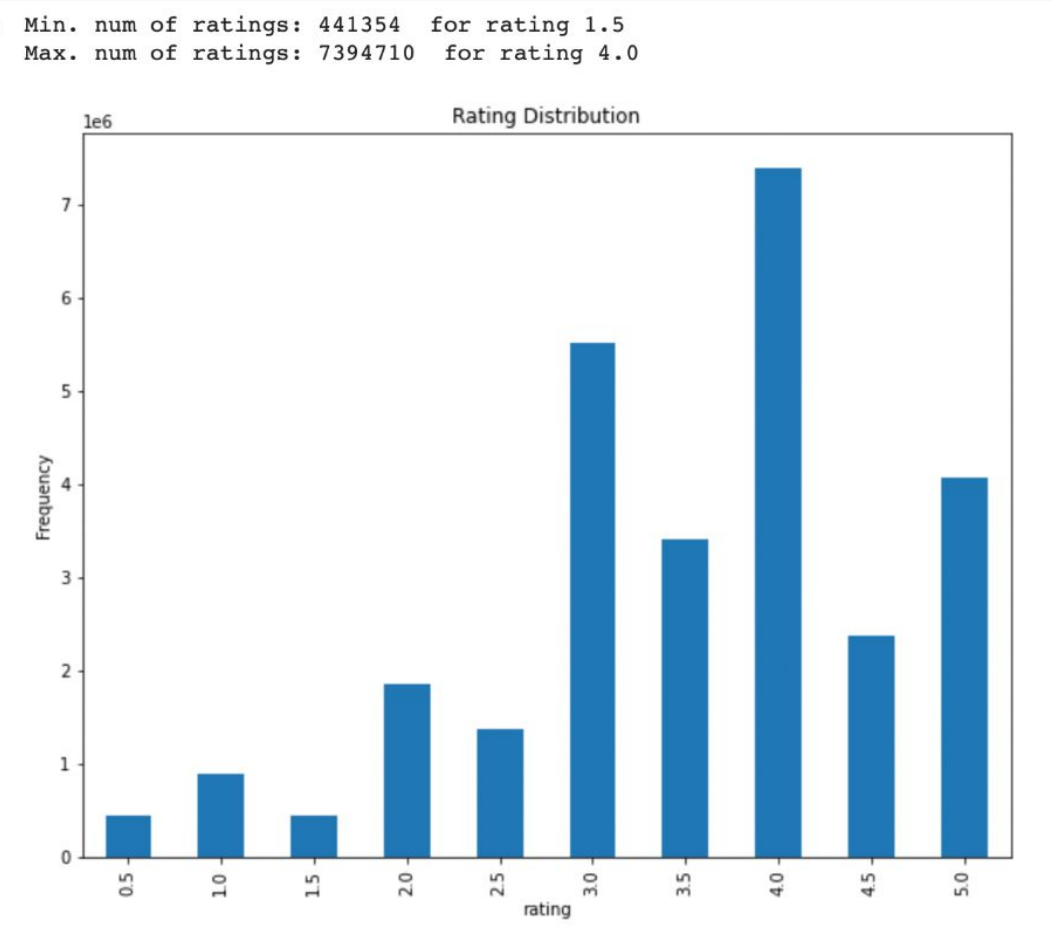
Dataset Summary

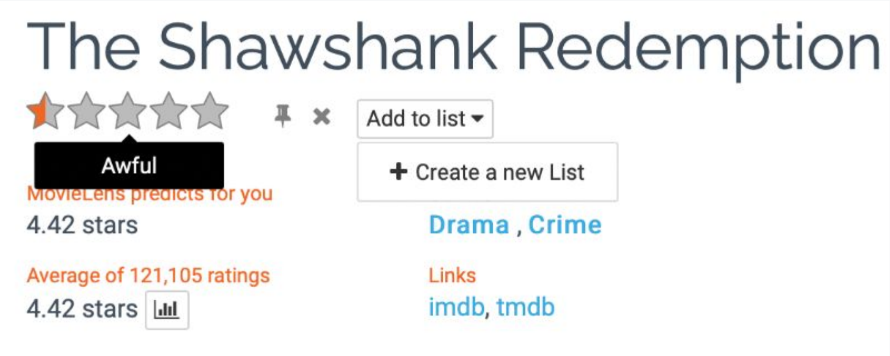
- ❏ Dataset → ML-Latest from MovieLens, a recommendation service.
- ❏ Total users → 283228
- ❏ Total movies → 58098
- ❏ Ratings → 27.7M ratings on a 5-star scale with half-star increments.
- ❏ Tags → 1.1M tags generated by users.
- ❏ Period → January 09, 1995 and September 26, 2018.
- ❏ Dataset also has tag-genome. It's a dense matrix and each movie in the genome has a relevance score for every tag in the genome.

  These are ML algorithm generated scores based on user-contributed content.

Exploratory Data Analysis (EDA)

Distribution of ratings (target variable)





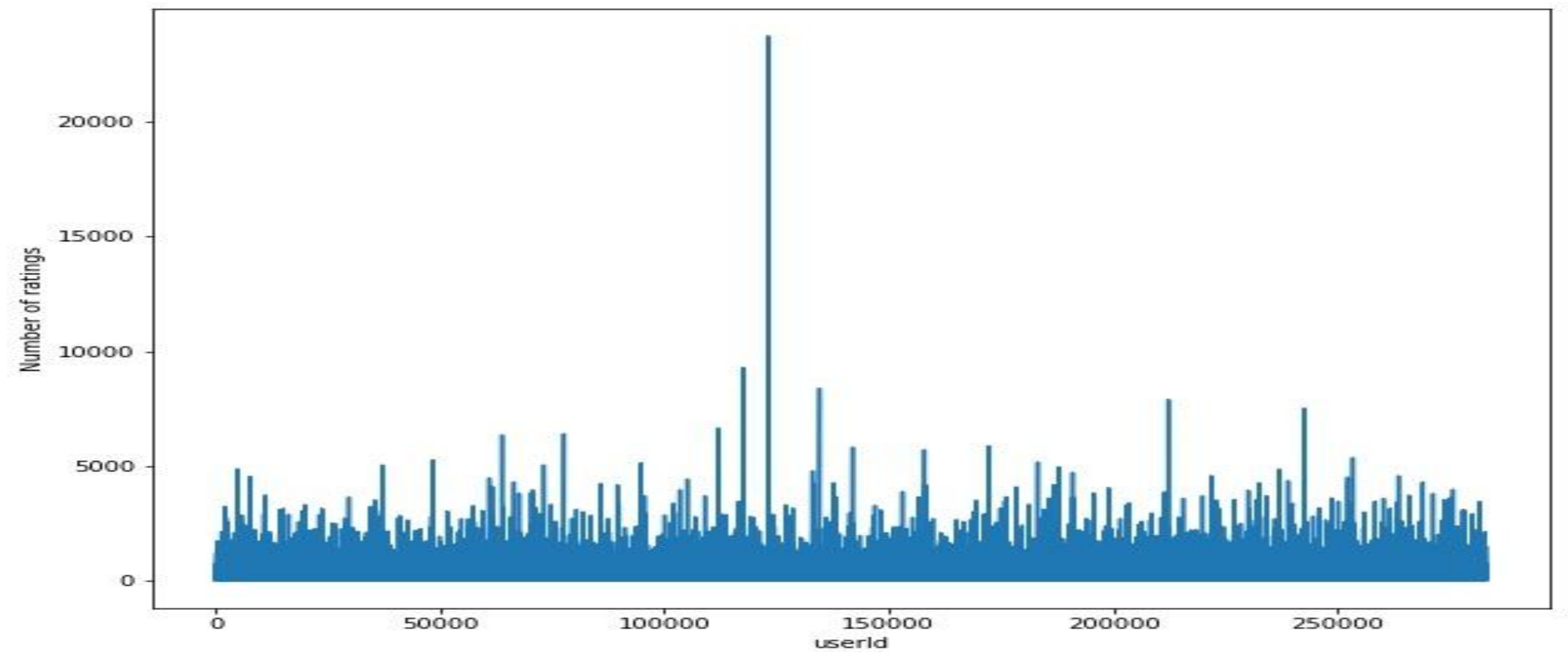
From https://movielens.org/

- Min Rating that a user can give is 0.5 (Awful) and maximum is 5.0 (Must Watch)

- The distribution is not exactly normal i.e left skewed.
- Most ratings are at least 3 star.

Exploratory Data Analysis (EDA)

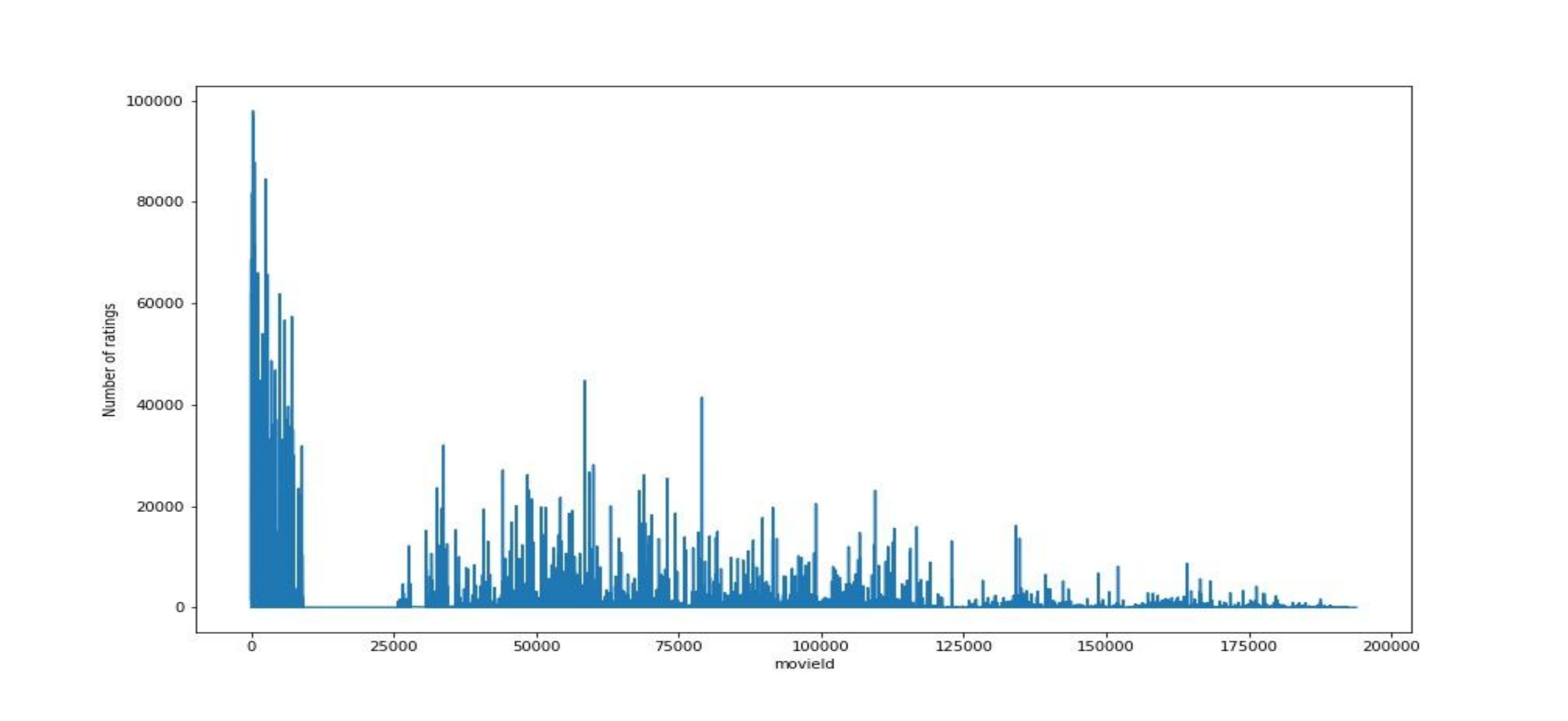
Number of ratings per user



- ❏ The ratings matrix is Sparse because each user only rates a small part around(2000) of total available movies(58098)

Exploratory Data Analysis (EDA)

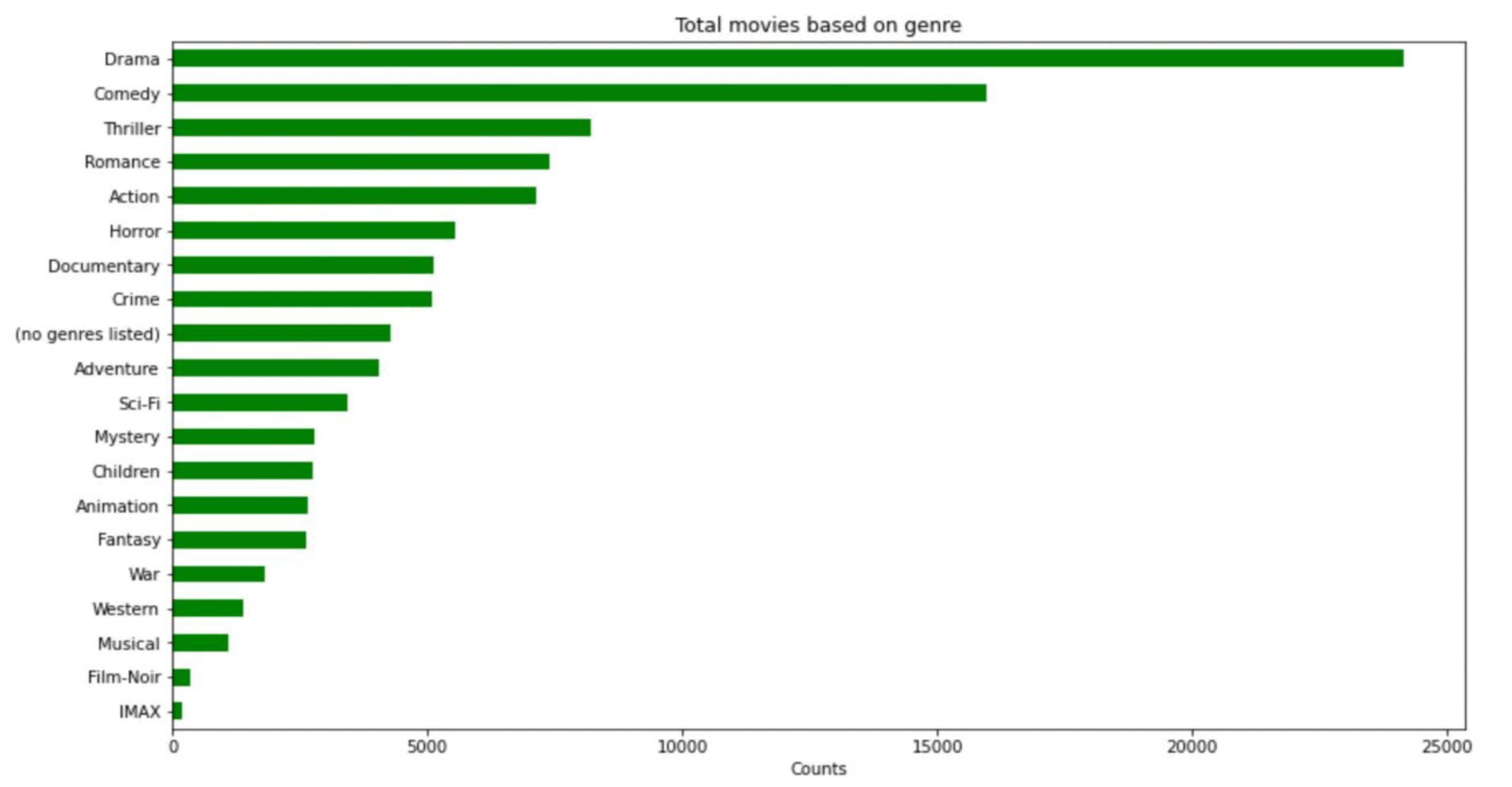
Number of ratings for each movie



- Some movies are rated and watched by tens of thousands of users while others are rated by few thousands times.
- There are around 4000 movies which are not rated at all.

Exploratory Data Analysis (EDA)

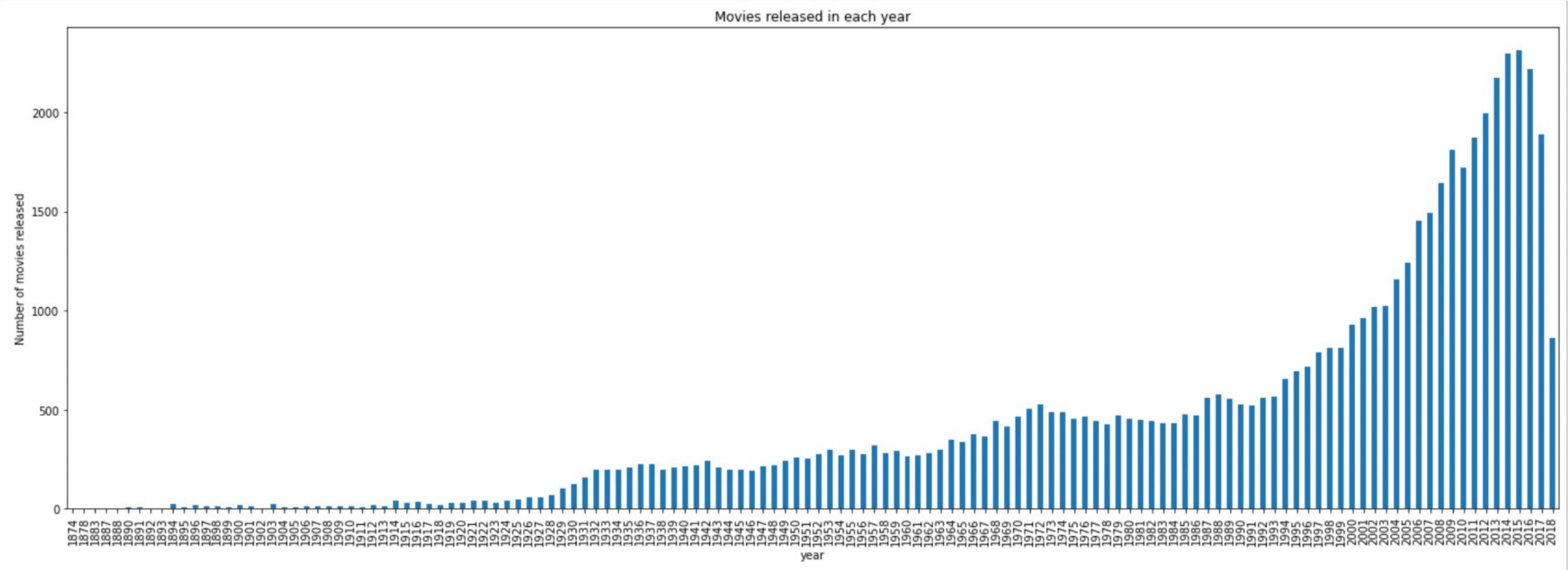
Most common Genre



- Drama and Comedy are the most common genres of movies followed by Thriller, Romance and others.
- There are around 5000 movies having no genre.

Exploratory Data Analysis (EDA)

How many movies got released in each year ?



- The number of movies released is more or less constantly increasing from 1874 to 2015.
- However after 2015 there is a decrease in total movies constantly till 2018. The data may be insufficient or there may be actually a decline in total movies released after 2015.

Exploratory Data Analysis (EDA)

Tags Data Analysis

- ❑ There are 1128 tag relevance scores (ML algorithm - generated) for every movie in genome.
- ❑ But these are available only for 13000 movies (5%) out of 58000 movies.
- ❑ Only 19k users from 280k users have given tags to movies after watching.
- ❑ It's strange but some users have given tags but not ratings.
- ❑ **Overall, the tags data is very sparse and cannot be fully utilised for the recommendation models.**

Movie Recommendation System

Baseline Model

- ❑ The model always predict the average rating 2.5.
  - RMSE on test data → 1.5
  - MAPE on test data → 42%

Reference Model

- ❑ Netflix Movie Recommendation Contest Winning Model

Rank	Team Name	Best Test Score	% Improvement
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos			
1	BellKor's Pragmatic Chaos	0.8567	10.06
2	The Ensemble	0.8567	10.06
3	Grand Prize Team	0.8582	9.90

Movie Recommendation System

❑ Collaborative filtering

Recommend items to you based on ratings of users who gave similar ratings as you.

No Domain knowledge is required.

Limitation → **Cold start problem**. How to

- Rank new items that few or no users have rated ?
- Show reasonable items to new users who have rated few or none items ?

Model needs to be **retrained** for every new user and movie.

❑ Content-based filtering

Recommend items to you based on features of user and item to find a good match.

Uses side information about users and items to get feature vector → suffer from cold start problem **to a lesser extent.**

- Item: Genre, year, average rating, location, actors,.....
- User: Demographics(age, gender, location), average rating per genre,……

No need of retraining model for every new user or movie.

Domain knowledge is required.

Collaborative Filtering Model - Feature Engineering

Feature Scaling

- **Rating (Target) is normalized using MinMaxScaler between -1 and 1.**

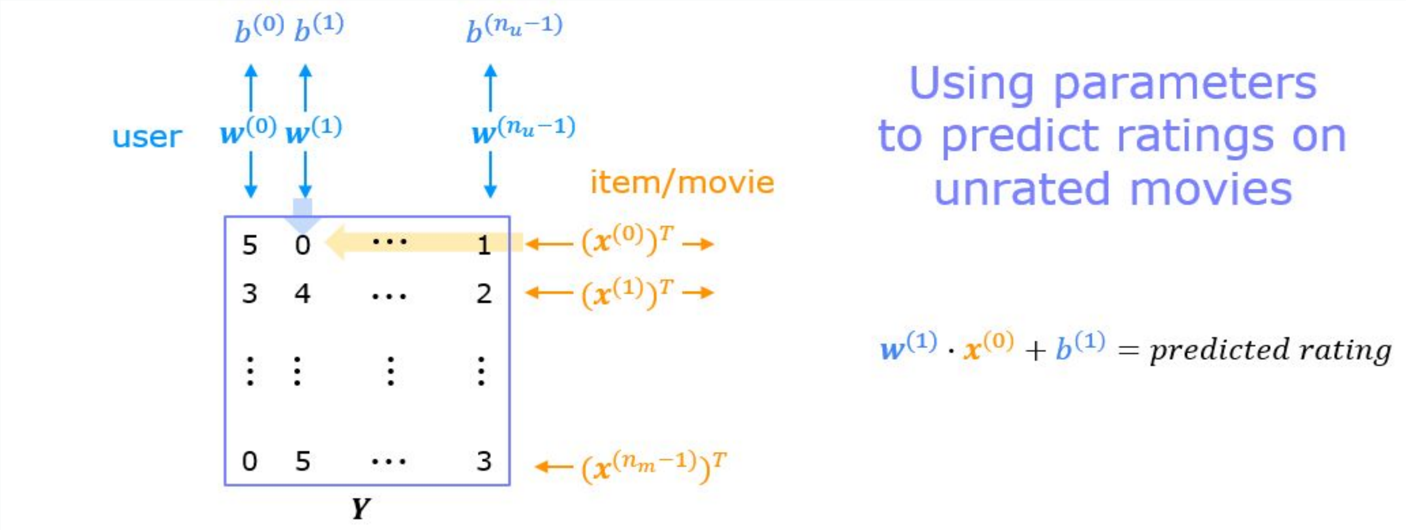
  This helps for :

  - Optimization algorithm to **run bit faster.**
  - Recommending reasonable items for **users who have rated no movies or very small number of movies.**

    i.e New users will be recommended movies based on minimum or mean rating of movies.

Collaborative Filtering - Architecture

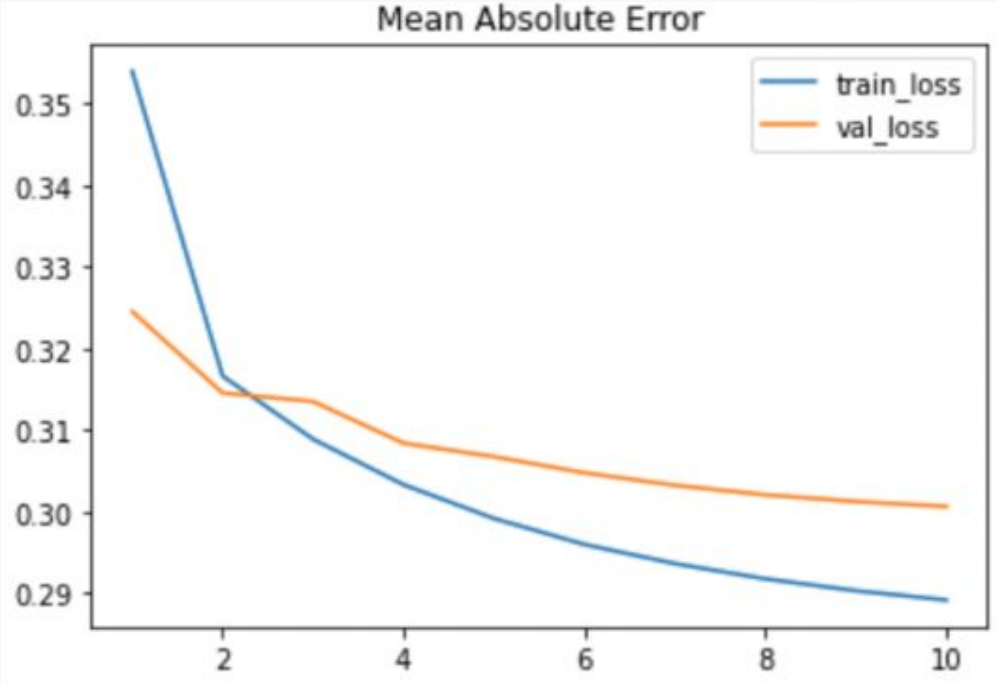
Collaborative Filtering ⇒ Learning Feature Matrices Through Ratings.



Inputs : UserId and MovieId
Output : Rating

- Model → Keras model with Two Embedding layers X and W for user and movie.
- Embedding Size = 32
- **W - Feature Matrix** for User of size (283228, 32).
- **X - Feature Matrix** for Item/Movie of size (58098, 32).
- **b - Bias** term for User of size (283228, 1)
- Trainable Parameters → 11,263,758
- Optimizer → Adam
- Loss Function → Mean Absolute Error
- Metrics → Mean Squared Error, RMSE

Collaborative Filtering Model - Training and Evaluation



Training Details:

Training size : (26654716 ,2)
Validation size : (543973,2)
Test size : (554754,2)
Epochs : 10
Batch Size : 1024

Model Training:

- MAE training loss → 0.289
- MAE validation loss → 0.300
- MAE test loss → 0.29
- RMSE → 0.876
- MAPE → 29.8%

Model Evaluation :

- RMSE → 0.88
- MAPE → 30.8 %

❏ Loss function MAE is very close for train, validation and test set ⇒ **Model is not overfitting the dataset.**

❏ RMSE is much better than the baseline model and comparable to the reference model.

❏ After training, model is able to predict **reasonable rating with an error rate of 0.88.**

Collaborative Filtering Model - Results

Top 5 movies rated by the User id 1007

	userId	movieId	rating	title	genres
0	1007	3844	5.0	Steel Magnolias (1989)	Drama
1	1007	3405	5.0	Night to Remember, A (1958)	Action|Drama
2	1007	2599	5.0	Election (1999)	Comedy
3	1007	539	5.0	Sleepless in Seattle (1993)	Comedy|Drama|Romance
4	1007	356	5.0	Forrest Gump (1994)	Comedy|Drama|Romance|War

Top 10 Movie Recommendations

```
Top10 movie recommendations for user 1007:
```

	index	movieId	title	genres	rating_pred
0	35617	142667	Black River (2001)	Sci-Fi|Thriller	4.998482
1	36938	145763	Il ricco, il povero e il maggiordomo (2014)	Comedy	4.995584
2	35505	142428	Winter Meeting (1948)	Drama|Romance	4.994129
3	18089	90112	First Love (1939)	Comedy|Musical	4.992551
4	19260	95064	House of the Rising Sun (2011)	Action|Crime|Drama|Thriller	4.932662
5	33071	136706	Cinema of Vengeance (1994)	(no genres listed)	4.932565
6	3060	3146	Deuce Bigalow: Male Gigolo (1999)	Comedy	4.928058
7	31523	133097	Orgasmo (1969)	(no genres listed)	4.917220
8	36171	143974	Home (2011)	Drama	4.913219
9	22181	105844	12 Years a Slave (2013)	Drama	4.911044

Content-Based Filtering Model - Feature Engineering

Extracting Contents

- ❑ Movie Features:
  - Year
  - Overall average rating
  - Genre present or not (1 or 0) - for all 20 genres

- ❑ User Features
  - Average rating per genre for each of 20 genres.

Content-Based Filtering Model - Feature Engineering

Feature Imputation

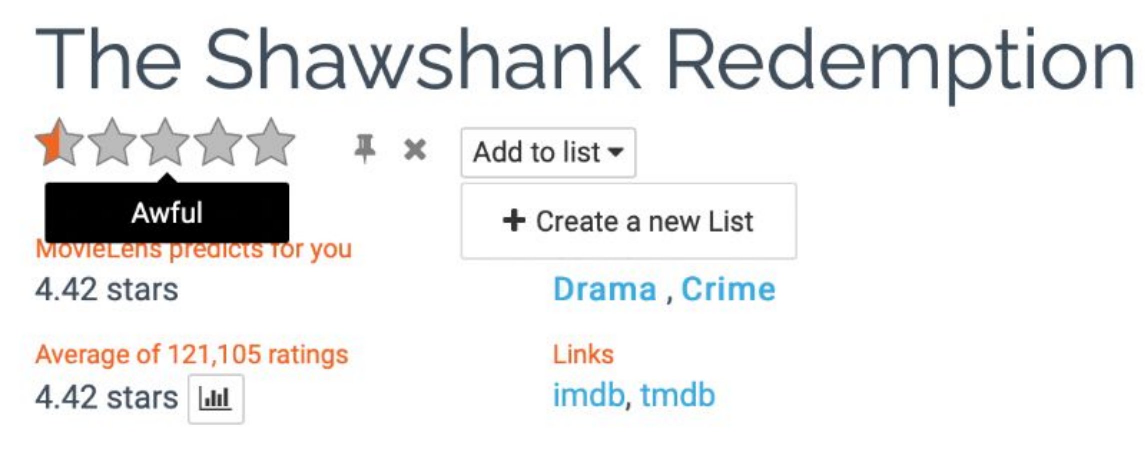


Figure from MovieLens website

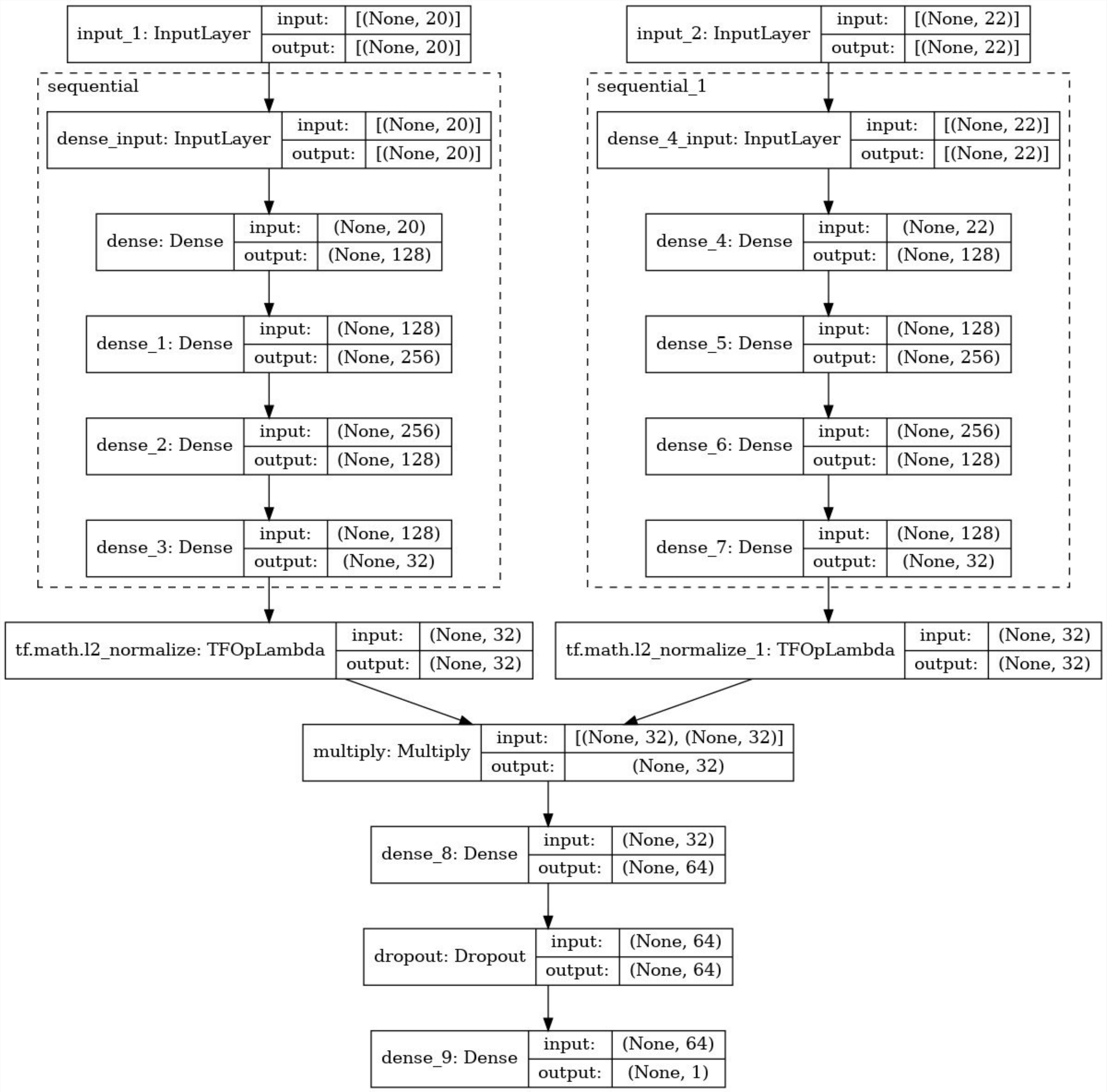
- ❑ Minimum rating that a user can give is 0.5 i.e half star or awful on MovieLens website.
- ❑ So for calculating average rating per genre, rating is imputed as zero if the user has not watched that particular genre.

Content-Based Filtering Model - Feature Engineering

Feature Scaling

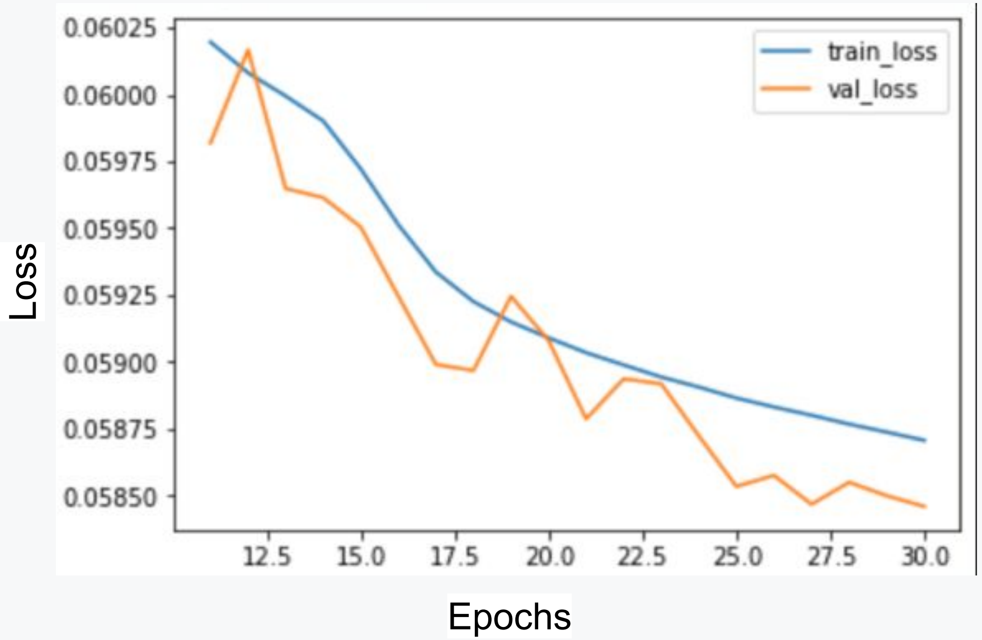
- ❑ Movie Features
  - Year and Avg. rating – Standard Normalization between 0 and 1
  - Genres – One hot encoding
- ❑ User Features
  - Scaled using Maximum Absolute Scaler between 0 and 1
  - This helps with the sparse data i.e not every user will watch every movies.

Content Based Filtering - Model Architecture



- Model → Neural Network
- Optimizer → Adam
- Loss Function → Huber Loss
- Metrics → MAPE, RMSE
- Parameters → 147,905

Content Based Filtering - Model Training & Evaluation



Training Details:

Training size : (26926391)
Validation size : (271983)
Test size : (555069)
Epochs : 30
Batch Size : 512 for 1st 10 epochs
1024 for next 20 epochs

Model Training:

- Huber training loss → 0.0587
- Huber validation loss → 0.0585
- Huber test loss → 0.0588

Model Evaluation :

- RMSE → 0.77
- MAPE → 24.8 %

Same For both training and test data

- ❏ Loss function MAE is very close for train, validation and test set ⇒ **Model is not overfitting the dataset.**
- ❏ RMSE is much better than the baseline model and comparable to the reference model.
- ❏ After training, model is able to predict **reasonable rating with an error rate of 0.77.**

Content Based Filtering - Results

1.Recommending Movies for Existing User

- ❏ Predict ratings that the user shall give to a movie using model.
- ❏ Sort the ratings and recommend top N movies.

Top 5 movies rated by user 1007

	userId	movieId	rating	timestamp	title	genres
0	1007	3844	5.0	974691715	Steel Magnolias (1989)	Drama
1	1007	3405	5.0	974689464	Night to Remember, A (1958)	Action|Drama
2	1007	2599	5.0	974693558	Election (1999)	Comedy
3	1007	539	5.0	974690177	Sleepless in Seattle (1993)	Comedy|Drama|Romance
4	1007	356	5.0	974693821	Forrest Gump (1994)	Comedy|Drama|Romance|War

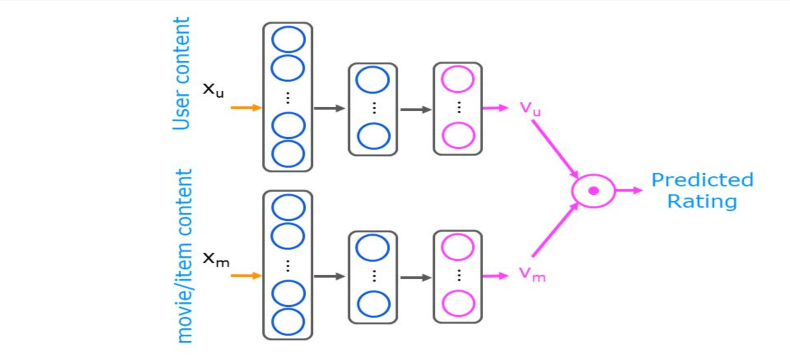
10 movie recommendations for user 1007:

	movieId	title	genres	rating
0	179649	Horse Crazy (2001)	Children|Drama	4.994435
1	140385	A Horse for Danny (1995)	Children|Drama	4.992630
2	115017	Christmas Memory, A (Truman Capote's 'A Christ...	Children|Drama	4.991246
3	140391	Pistol: The Birth of a Legend (1991)	Children|Drama	4.990455
4	125287	Heidi (2005)	Children|Drama	4.987551
5	149328	Dreamkeeper (2003)	Children|Drama	4.985034
6	162378	A Thousand Men and a Baby (1997)	Drama	4.984047
7	147035	Baile Perfumado (1997)	Drama	4.984047
8	72235	Between the Devil and the Deep Blue Sea (1995)	Drama	4.981792
9	57038	To the Left of the Father (Lavoura Arcaica) (2...	Drama	4.981567

Content Based Filtering - Results

2.Recommending Movies For New User

1. Get the new user interest (rating per genre), if any else 0 value.
2. Give the user and movie content to model to predict rating



User Interest:

```
new_no_genres_listed = 0
new_Action = 5
new_Adventure = 4.5
new_Animation = 0
new_Children = 0
new_Comedy = 0
```

Top 10 movie recommendations for given User:

	movieId	title	genres	rating
0	166297	Bagi, the Monster of Mighty Nature (Taishizen ...	Action|Adventure|Animation	4.975091
1	55995	Beowulf (2007)	Action|Adventure|Animation|Fantasy|IMAX	4.968926
2	93766	Wrath of the Titans (2012)	Action|Adventure|Fantasy|IMAX	4.965480
3	106072	Thor: The Dark World (2013)	Action|Adventure|Fantasy|IMAX	4.961287
4	86332	Thor (2011)	Action|Adventure|Drama|Fantasy|IMAX	4.956300
5	86880	Pirates of the Caribbean: On Stranger Tides (2...	Action|Adventure|Fantasy|IMAX	4.953849
6	101112	Oz the Great and Powerful (2013)	Action|Adventure|Fantasy|IMAX	4.939826
7	115669	Young Detective Dee: Rise of the Sea Dragon (D...	Action|Adventure|Drama|Fantasy|Mystery|IMAX	4.933150
8	95475	Dragon Ball Z: Cooler's Revenge (Doragon bôru ...	Action|Adventure|Animation	4.931520
9	100469	Chinese Zodiac (Armour of God III) (CZ12) (2012)	Action|Adventure|IMAX	4.931441

➔ We can see that the recommended movies are based on new user’s interest w.r.t genres

Content Based Filtering - Results

3.Recommending Similar Movies

- ❏ Given movie name recommend other similar movies.
- ❏ Compare the feature vector of given movie with every other movie.
- ❏ Similarity Function → Cosine Similarity (1 being most similar and 0 being least similar).

```
Similar movies to Toy Story (1995) with genre Adventure|Animation|Children|Comedy|Fantasy are:
```

	movieId	title	genres	similarity
0	4886	Monsters, Inc. (2001)	Adventure|Animation|Children|Comedy|Fantasy	0.998993
1	3114	Toy Story 2 (1999)	Adventure|Animation|Children|Comedy|Fantasy	0.998060
2	166461	Moana (2016)	Adventure|Animation|Children|Comedy|Fantasy	0.987920
3	95311	Presto (2008)	Animation|Children|Comedy|Fantasy	0.987907
4	4016	Emperor's New Groove, The (2000)	Adventure|Animation|Children|Comedy|Fantasy	0.987678
5	6377	Finding Nemo (2003)	Adventure|Animation|Children|Comedy	0.987375
6	189591	Jungle Emperor Leo (1997)	Adventure|Animation|Children|Comedy	0.986690
7	72356	Partly Cloudy (2009)	Animation|Children|Comedy|Fantasy	0.986599
8	134853	Inside Out (2015)	Adventure|Animation|Children|Comedy|Drama|Fantasy	0.985408
9	192225	Redwall The Movie (2000)	Animation|Children|Comedy|Fantasy	0.983926

⇒ We can see the suggested movies have similar Genres to the given movie.

Models Summary

Model Name	RMSE		MAPE	
	Training	Test	Training	Test
Baseline Model	1.48	1.48	41.84%	41.7%
Collaborative Filtering	0.877	0.888	29.81%	30.26%
Content-based filtering	0.771	0.772	24.87%	24.85%

- ➢ The model definitely outperforms the naive model which predicts average rating for all movies.
- ➢ Content-based filtering is better than collaborative filtering since it takes into account other informations about the user and movies and also works for newly added users or movies.

Further Improvements

- ❑ More information can be extracted for both movies and users that can help content-based models learn better embeddings.
  - User information → Age, gender, location, ..etc.
  - Movie information → location, actors, language, …etc
- ❑ Most common user tags can be analysed to create some useful user content.
- ❑ Tag scores can be obtained for remaining 45000 movies which can be very useful content for movie.