Assignment 4: Recommender Systems

Kumar Shubham (G01402581)
Deep Vora (G01388910)

Introduction:

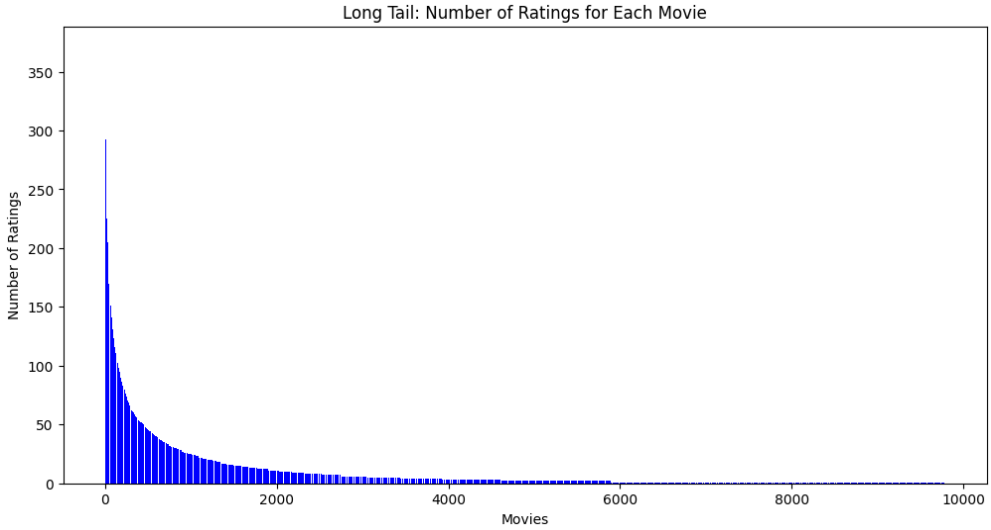
Recommender Systems are being used by many information-based companies such as Google, Netflix, Tweeter, LinkedIn, Airbnb and many others. It started in the mid of the 1990s, with the invention of Tapestry, the first Recommendation System. The idea of being valuable came from the Netflix prize contest in 2009 for 1 million dollars to improve the engine, which was the biggest leap of the techniques used. Since the late 1950s, Artificial Intelligence emerged and then machine learning came into existence such as clustering, K-nearest neighbor, Bayes network and many others.
Collaborative and content-based are the two approaches for movie recommendation. It can be also portrayed as a User-based and Item-based collaborative recommender system. In User base, the user-centered system finds the similarity of a specific user of interest with others. They consider a similar interaction on the same items or choices for estimating closeness based on a similar rating or time hovering. For example, the nearest neighbor algorithm refers to those items which have not been interacted with. Whereas the item-centered works for item similarities checking whether most users have interacted with both the items in a similar way. They are less personalized and more biased but more robust compared to user-centered techniques. Whereas the Item-to-Item-based method has been commonly used in areas such as LinkedIn (2014), YouTube (2010), Amazon (2003), etc. Item-to-Item-based system surfaces similar items from key values data. They are more interpretable as the scores of key values data are too small because of the selected threshold. The main idea is to leverage the behavior pattern and predict suggestions.

Collaborative Filtering:

Utility (User-Item) Matrix: In a recommendation-system application there are two classes of entities, which we shall refer to as users and items. Users have preferences for certain items, and these preferences must be teased out of the data. The data itself is represented as a utility matrix, giving for each user-item pair, a value that represents what is known about the degree of preference of that user for that item. So, here in the below snapshot, the value is “rating” given by the user to the movie. By using Collaborative filtering method, we will fill out all the Nulls in the below utility matrix.

movieId	1	2	3	4	5	6	7	8	9	10	...	204692	204698	204704	205054	205072	205106	205413	205499	205557	206272
userId																					
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	3.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
753	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
754	NaN	NaN	NaN	NaN	NaN	5.0	NaN	NaN	NaN	3.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
755	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
756	3.0	NaN	NaN	NaN	NaN	4.0	NaN	NaN	NaN	3.5	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
757	3.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Long Tail: The distinction between the physical and on-line worlds has been called the long tail phenomenon, and it is suggested in the figure below. The vertical axis represents popularity (Number of Ratings). The items which are Movie IDs are ordered on the horizontal axis according to their popularity.



Top 20 Movie Recommendation using ALS Algorithm:

```
+------+--------------------------------------------------------------------------------------------------------
|userId|recommendations
+------+--------------------------------------------------------------------------------------------------------
|1     |[{NOFX Backstage Passport 2}, {The Legend of Paul and Paula (1973)}, {C'est quoi la vie? (1999)}, {E
|3     |[{The Law and the Fist (1964)}, {Head Trauma (2006)}, {O Θανάσης στη χώρα της σφαλιάρας (1976)}, {De
|5     |[{Truth and Justice (2019)}, {NOFX Backstage Passport 2}, {2 (2007)}, {Adrenaline (1990)}, {Dead in
|6     |[{Adrenaline (1990)}, {Foster (2018)}, {NOFX Backstage Passport 2}, {.hack Liminality In the Case of
|9     |[{World Gone Wild (1988)}, {A Kind of America 2 (2008)}, {Civilisation (1969)}, {Truth and Justice (
|12    |[{NOFX Backstage Passport 2}, {National Theatre Live: One Man, Two Guvnors (2011)}, {Head Trauma (20
|13    |[{NOFX Backstage Passport 2}, {Adrenaline (1990)}, {Foster (2018)}, {Head Trauma (2006)}, {.hack Lim
|15    |[{NOFX Backstage Passport 2}, {The Law and the Fist (1964)}, {Dead in the Water (2006)}, {C'est quoi
|16    |[{NOFX Backstage Passport 2}, {Adrenaline (1990)}, {Head Trauma (2006)}, {Dead in the Water (2006)},
|17    |[{World Gone Wild (1988)}, {The Dragon Spell (2016)}, {The Lion of Thebes (1964)}, {Provocateur (199
|19    |[{Head Trauma (2006)}, {O Θανάσης στη χώρα της σφαλιάρας (1976)}, {Vision Portraits (2019)}, {Vergee
|20    |[{Head Trauma (2006)}, {Adrenaline (1990)}, {Foster (2018)}, {Dead in the Water (2006)}, {Truth and
|22    |[{The Miracle Maker (2000)}, {The Good Fight: The Abraham Lincoln Brigade in the Spanish Civil War (
|26    |[{The Law and the Fist (1964)}, {Dead in the Water (2006)}, {Truth and Justice (2019)}, {C'est quoi
|27    |[{World Gone Wild (1988)}, {The Lion of Thebes (1964)}, {Hello Stranger (2010)}, {The Dragon Spell (
|28    |[{Adrenaline (1990)}, {Head Trauma (2006)}, {Civilisation (1969)}, {NOFX Backstage Passport 2}, {Dea
|31    |[{Dead in the Water (2006)}, {Civilisation (1969)}, {NOFX Backstage Passport 2}, {Les Luthiers: El G
|34    |[{World Gone Wild (1988)}, {The Marquis (2011)}, {The Lion of Thebes (1964)}, {A Kind of America 2 (
|35    |[{Truth and Justice (2019)}, {Dead in the Water (2006)}, {A Kind of America 2 (2008)}, {Seeing Red:
|37    |[{NOFX Backstage Passport 2}, {Adrenaline (1990)}, {Head Trauma (2006)}, {O Θανάσης στη χώρα της σφα
+------+--------------------------------------------------------------------------------------------------------
only showing top 20 rows
```

Cross-Validation:

We used a subset (80%) of data to train the rating dataframe on 5 numFolds and test it on the rest of the data (20%) using Cross-Validator, imported from pyspark.ml library.

We passed a list of Rank ([5,10,20]) and Regularization ([0.05,0.1,0.2]) and found the best parameters, shown in the below table, alongwith the result.

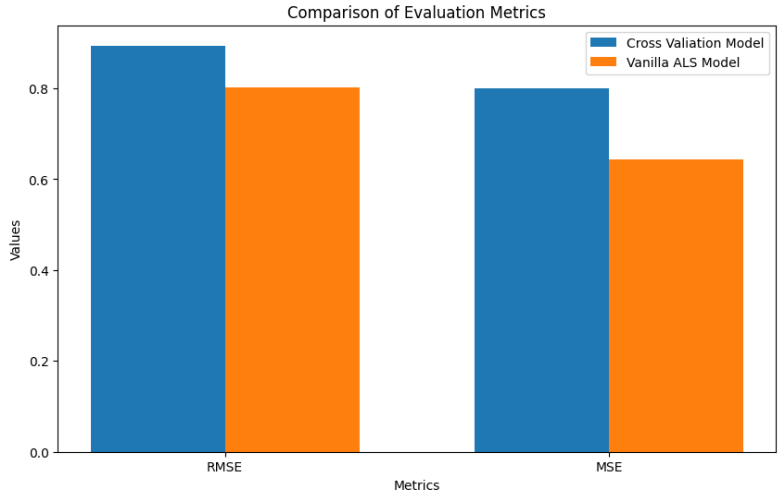
Result: **Best Parameters Found:**

RMSE	:	0.8940665670156733
MSE	:	0.7993550262551915
MAP at K	:	0.6825662149178818

Parameters	Best Value
Rank	10
Regularization	0.2

Comparing ALS model result with Cross Validation:

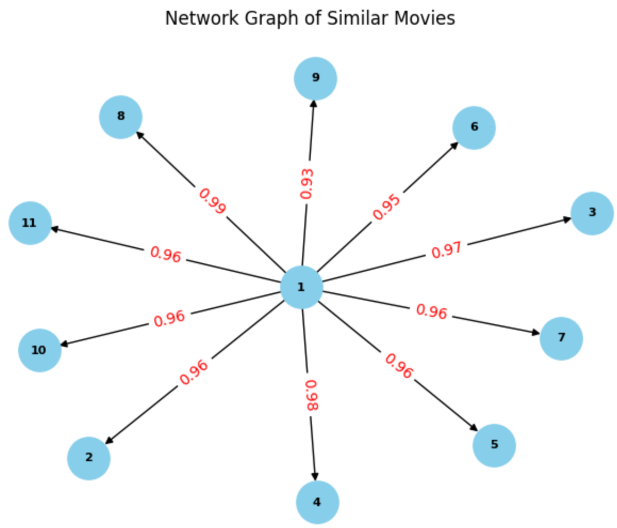
We found the vanilla ALS model gave good results as compared to the Cross Validation Model result. As the data is huge, trying with other parameters and different maxIters was not feasible. So, Assuming it may perform better than the vanilla ALS model if the cross validation performed on the right set of parameters. Here in our case, you can see the RMSE and MSE value of the ALS model is lower than the Cross Validation model which tells that the ALS model prediction is more closer and accurate.



	ALS Model	Cross Validation Model
RMSE	0.80	0.89
MSE	0.64	0.79

Item-Item Collaborative Filtering:

In this section, we will find similarities between items rather than users. To find similarity, we will use Cosine Similarity. It is a metric used to measure the similarity between the two non-zero vectors. Here, it will be used to measure the similarity between item vectors. It ranges from -1 to 1, 1 being completely similar and 0 indicates no similarity. To perform this, we will take ratings dataframe and will perform self join to get movie1, movie2, rating1, rating2 pairs. After that we will compute cosine similarity using rating1 and rating2 attributes.
Below snapshot is showing top 10 similar movies Ids for movie Id 1, along with their similarity score all of which are near to 1.



After computing the cosine similarity, we found out the top 20 recommended movies, by joining the ratings dataframe with movies dataframe and sorting the dataframe based on descending order of cosine similarity score.

Top 20 Recommended Movies and their Cosine Score:

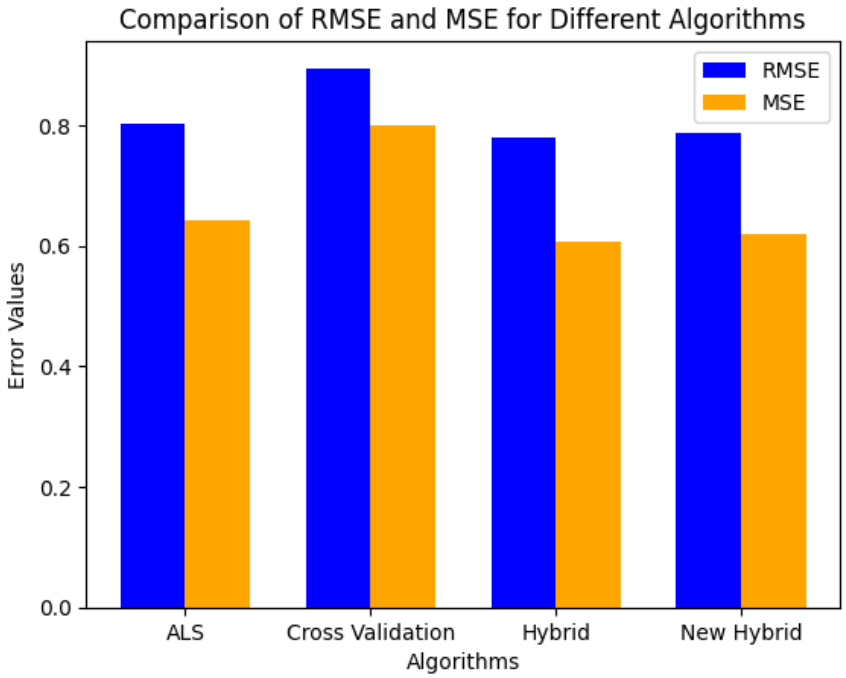
```
+------+---------------------------------------------------------------------------------------------------------------
|userId|recommendations
+------+---------------------------------------------------------------------------------------------------------------
|3     |[{Ocean's Eleven (2001), 0.9852318204565945}, {American History X (1998), 0.9847159243089344}, {American History X (1998)
|5     |[{Pulp Fiction (1994), 0.9810541878191049}, {Pulp Fiction (1994), 0.9793278126887348}, {Pulp Fiction (1994), 0.978639494
|6     |[{Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981), 0.9830718292039488}, {Goodfellas (1990)
|9     |[{Aliens (1986), 0.9814279490574713}, {Star Wars: Episode IV - A New Hope (1977), 0.9809269510296594}, {Aliens (1986), 0.
|12    |[{Departed, The (2006), 0.9878994342679043}, {Forrest Gump (1994), 0.9863951170686497}, {Forrest Gump (1994), 0.98411980
|13    |[{Forrest Gump (1994), 0.9863951170686497}, {Ocean's Eleven (2001), 0.9852318204565945}, {Twelve Monkeys (a.k.a. 12 Monke
|15    |[{Forrest Gump (1994), 0.9863951170686497}, {Forrest Gump (1994), 0.9841198002760377}, {Fargo (1996), 0.9807405873809092}
|16    |[{American Beauty (1999), 0.9833508315544977}, {Rain Man (1988), 0.980879839919869}, {Lord of the Rings: The Two Towers,
|19    |[{Forrest Gump (1994), 0.9863951170686497}, {Forrest Gump (1994), 0.9841198002760377}, {Shawshank Redemption, The (1994),
|20    |[{Harry Potter and the Chamber of Secrets (2002), 0.9871061784553216}, {Star Wars: Episode IV - A New Hope (1977), 0.980
|22    |[{Shawshank Redemption, The (1994), 0.9821413709455667}, {Shawshank Redemption, The (1994), 0.9814755766300743}, {Shawsha
|26    |[{Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981), 0.9830718292039488}, {Pulp Fiction (19
|27    |[{E.T. the Extra-Terrestrial (1982), 0.979184381316478}, {Wizard of Oz, The (1939), 0.9774371836428658}, {Wizard of Oz,
|28    |[{Terminator, The (1984), 0.9858093128569566}, {Terminator, The (1984), 0.9850631459187862}, {Back to the Future (1985),
|31    |[{Memento (2000), 0.9888893096270008}, {Seven (a.k.a. Se7en) (1995), 0.986749597982267}, {Terminator, The (1984), 0.9858
|34    |[{Braveheart (1995), 0.9812597460214314}, {Braveheart (1995), 0.9789494742101492}, {Braveheart (1995), 0.976295219176559
|35    |[{Jumanji (1995), 0.9701580436855601}, {Jumanji (1995), 0.97011627897415}]
|37    |[{Memento (2000), 0.9888893096270008}, {Good Will Hunting (1997), 0.9843186595762242}, {Good Will Hunting (1997), 0.9837
|40    |[{Good Will Hunting (1997), 0.9843186595762242}, {Good Will Hunting (1997), 0.9837153548187813}, {Good Will Hunting (199
|41    |[{Men in Black (a.k.a. MIB) (1997), 0.9813184895576078}, {Godfather: Part II, The (1974), 0.9798724526392008}, {Matrix,
+------+---------------------------------------------------------------------------------------------------------------
only showing top 20 rows
```

Hybrid Algorithm:

As you know, the Item-Item CF is often used in Hybrid Recommendation System, where multiple recommendation techniques are combined to enhance overall performance. We did the same, by using the previous ALS model with Item-Item CF and computed hybrid score. We can see a **significant improvement** in **RMSE** and **MSE** scores after implementing Hybrid Recommendation.
In our case, we tried to implement two Hybrid Recommendation System: **i. Hybrid** and **ii. New Hybrid**.
The difference between them is just that the New Hybrid model contains Linear Regression Predictions along with ALS and Item-Item CF. We found Hybrid Recommendation RMSE and MSE score closer towards actual labels

after trying with several weights. However, the New Hybrid Model is also closer to the Hybrid Model in different weights.

Approach	RMSE	MSE	Weights
ALS	0.8017801618368304	0.642851427915094	
Hybrid (ALS + Item-Item CF)	0.7794850872648431	0.6075970012682801	[0.8, 0.8]
New Hybrid (ALS + Item-Item CF + Linear Regression)	0.7877494988436883	0.620549272928482	[0.7, 0.5, 0.2]



Conclusion:

I would like to conclude by saying that the Hybrid Recommendation System is the best approach for large datasets and real world projects. Alternating Least Square (ALS) approach can capture complex relationships between users and items whereas Item-Item CF is effective for recommending items that are similar to those that a user has already liked but Not as effective for recommending items that are very different from those that a user has already liked. So here Hybrid Recommender system can overcome the limitations of each individual method and it can be more complex to implement and maintain. It is important to consider the factors such as scalability, Interpretability and cold-start problems when choosing an approach.Scalability: The approach should be able to scale to large datasets.