Ensemble based Movie Recommendation System and Evaluation

Sanket Badhe (smb535) Abhishek Bhatt (ab2083) Siva Harshini Dev Bonam (sdb202) Soumya Anthareddy (sa1607)

Instructor:

Prof. Yongfeng Zhang

DATASET

MovieLens https://grouplens.org/datasets/movielens/

- 1M
 - Users 6000
 - Movies 4000
 - Ratings 1 million
 - Train set 800193 ratings
 - Test set 200016 ratings
- Full Latest(27M)
 - Users 280,000
 - Movies 58,000
 - Ratings 27,000,000
 - Train set 22207659 ratings
 - Test set 5545785 ratings

Exploratory Data Analysis (EDA)

- Analysed on 27753444 x 24 data
- No missing values
- No redundancies
- Assumes data authenticity and so no incorrect values.
- Correlations calculated using Pearson correlation method.
 - Children-Animation (0.64)
 - Action-Adventure (0.33)
 - Musical-Animation (0.32)
 - Action-SciFi (0.32)
 - Children-Musical (0.31)

- Thriller-Mystery (0.3)
- 0 ...
- o Comedy-Drama (-0.27)
- Comedy-Thriller (-0.34)
- None of columns appears to be normally distributed. All are independent variables positively skewed.
- Action, Adventure, Crime, Fantasy and War are 5 significant genres that have ratings.

Number of ratings in each bucket

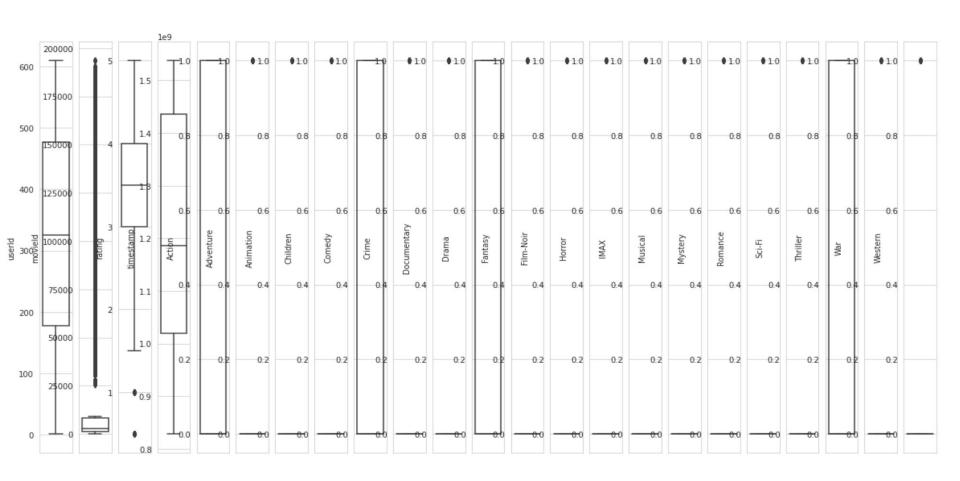
Top 10 movies based on the total number of ratings

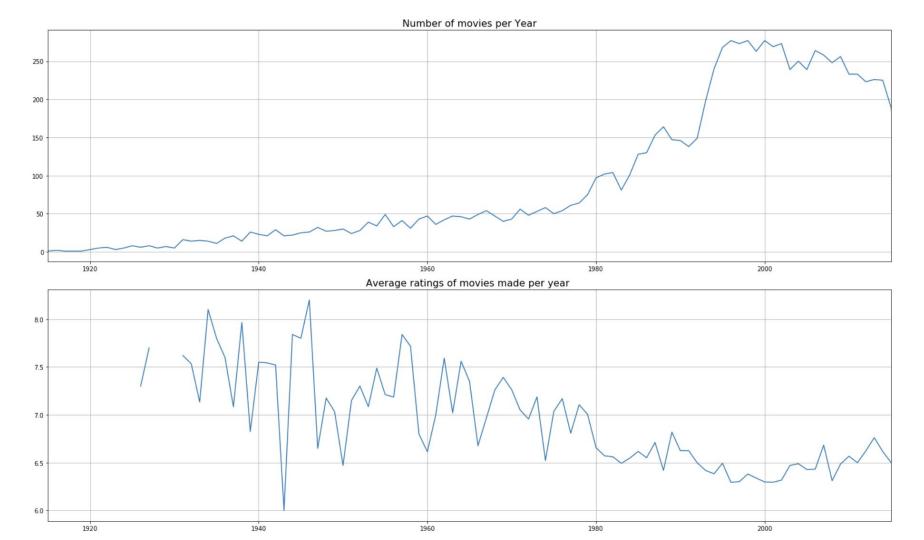
Ratings	Count	Movie title	Count
Ratings 4.0 3.0 5.0 3.5 4.5 2.0 2.5 1.0	Count 7394710 5515668 4071135 3404360 2373550 1850627 1373419 886233	Shawshank Redemption, The (1994) Forrest Gump (1994) Pulp Fiction (1994) Silence of the Lambs, The (1991) Matrix, The (1999) Star Wars: Episode IV - A New Hope (1977) Jurassic Park (1993) Schindler's List (1993)	97999 97040 92406 87899 84545 81815 76451 71516
0.5 1.5	442388 441354	Braveheart (1995) Toy Story (1995)	68803 68469

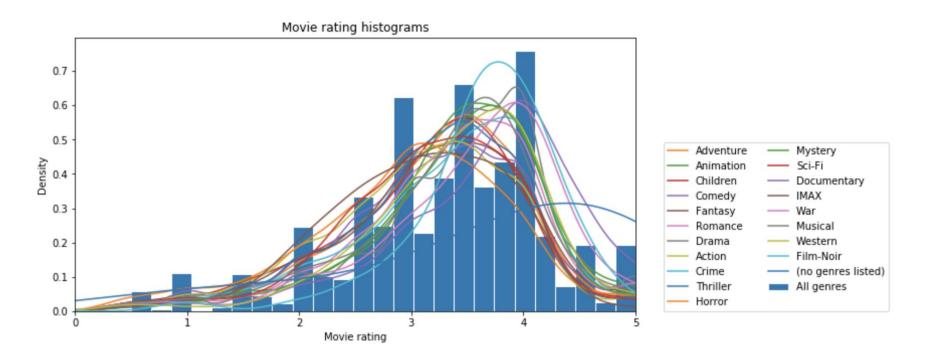
Pearson correlation of the ratings.

rating -	1	0.13	0.067	0.066	0.037	0.036	0.034	0.031	0.023	0.012	0.011	0.0038	0.0022	-0.0037	-0.0041	-0.0045	-0.0058	-0.02	-0.027	-0.034	-0.049	-0.065	-0.089
Drama -	0.13	1	0.081	0.19	-0.001	0.018	-0.13	-0.074	-0.067	-0.056	-0.0089	-0.25	0.11	-0.15	-0.021	-0.04	-0.019	-0.19	-0.15	-0.24	0.0034	-0.13	-0.27
Crime -	0.067	0.081	1	-0.079	0.14	0.13	-0.09	-0.048	-0.02	-0.083	-0.047	-0.18	-0.16	-0.13	-0.016	0.28	0.013	-0.14	-0.12	0.099	-0.015	-0.05	-0.09
War -	0.066	0.19	-0.079	1	-0.021	-0.062	-0.053	-0.0092	-0.024	-0.039	0.012	-0.051	0.01	-0.066	-0.027	-0.074	-0.018	-0.087	-0.068	0.068	-0.011	-0.06	-0.082
Film-Noir -	0.037	-0.001	0.14	-0.021	1	0.16	-0.025	-0.01	-0.019	-0.019	-0.011	-0.04	-0.022	-0.028	-0.023	0.088	-0.014	-0.028	-0.03	-0.041	0.0011	-0.022	-0.071
Mystery -	0.036	0.018	0.13	-0.062	0.16	1	-0.053	-0.03	0.0022	-0.058	-0.03	-0.093	-0.065	-0.027	-0.0019	0.3	0.017	0.01	-0.062	-0.054	-0.003	0.1	-0.17
Animation -	0.034	-0.13	-0.09	-0.053	-0.025	-0.053	1	-0.028	0.089	0.32	-0.03	0.22	-0.011	0.23	0.075	-0.15	0.079	-0.035	0.64	-0.096	-0.012	-0.067	0.09
Documentary -	0.031	-0.074	-0.048	-0.0092	-0.01	-0.03	-0.028	1	-0.017	-0.0073	-0.015	-0.059	-0.051	-0.04	0.057	-0.066	0.026	-0.05	-0.034	-0.068	0.0092	-0.03	-0.055
IMAX -	0.023	-0.067	-0.02	-0.024	-0.019	0.0022	0.089	-0.017	1	0.051	-0.023	0.18	-0.06	0.12	0.23	-0.0081	0.15	0.15	0.071	0.15	-0.0011	-0.03	-0.12
Musical -	0.012	-0.056	-0.083	-0.039	-0.019	-0.058	0.32	-0.0073	0.051	1	-0.02	0.012	0.083	0.12	-0.053	-0.12	-0.027	-0.081	0.31	-0.12	-0.00093	-0.031	0.055
Western -	0.011	-0.0089	-0.047	0.012	-0.011	-0.03	-0.03	-0.015	-0.023	-0.02	1	0.03	-0.025	-0.048	-0.017	-0.059	-0.04	-0.03	-0.033	0.036	-0.0098	-0.033	-0.029
Adventure -	0.0038	-0.25	-0.18	-0.051	-0.04	-0.093	0.22	-0.059	0.18	0.012	0.03	1	-0.15	0.26	0.047	-0.05	0.052	0.24	0.23	0.33	-0.012	-0.12	-0.08
Romance -	0.0022	0.11	-0.16	0.01	-0.022	-0.065	-0.011	-0.051	-0.06	0.083	-0.025	-0.15	1	0.032	-0.074	-0.17	-0.048	-0.16	-0.043	-0.19	-0.0013	-0.11	0.2
Fantasy -	-0.0037	-0.15	-0.13	-0.066	-0.028	-0.027	0.23	-0.04	0.12	0.12	-0.048	0.26	0.032	1	0.027	-0.14	0.061	-0.077	0.3	-0.025	0.0042	0.0034	0.054
movield -	-0.0041	-0.021	-0.016	-0.027	-0.023	-0.0019	0.075	0.057	0.23	-0.053	-0.017	0.047	-0.074	0.027	1	-0.01		0.077	-0.014	0.072	0.0068	-0.019	-0.037
Thriller -	-0.0045	-0.04	0.28	-0.074	0.088	0.3	-0.15	-0.066	-0.0081	-0.12	-0.059	-0.05	-0.17	-0.14	-0.01	1	-0.016	0.13	-0.19	0.26	-0.0064	0.16	-0.34
timestamp -	-0.0058	-0.019	0.013	-0.018	-0.014	0.017	0.079	0.026	0.15	-0.027	-0.04	0.052	-0.048	0.061		-0.016	1	0.06	0.015	0.048	0.097	0.012	-0.018
Sci-Fi -	-0.02	-0.19	-0.14	-0.087	-0.028	0.01	-0.035	-0.05	0.15	-0.081	-0.03	0.24	-0.16	-0.077	0.077	0.13	0.06	1	-0.073	0.32	0.0031	0.059	-0.2
Children -	-0.027	-0.15	-0.12	-0.068	-0.03	-0.062	0.64	-0.034	0.071	0.31	-0.033	0.23	-0.043	0.3	-0.014	-0.19	0.015	-0.073	1	-0.15	-0.018	-0.084	0.12
Action -	-0.034	-0.24	0.099	0.068	-0.041	-0.054	-0.096	-0.068	0.15	-0.12	0.036	0.33	-0.19	-0.025	0.072	0.26	0.048	0.32	-0.15	1	-0.0068	-0.042	-0.25
userld -	-0.049	0.0034	-0.015	-0.011	0.0011	-0.003	-0.012	0.0092	-0.0011	-0.00093	-0.0098	-0.012	-0.0013	0.0042	0.0068	-0.0064	0.097	0.0031	-0.018	-0.0068	1	0.03	-0.0079
Horror -	-0.065	-0.13	-0.05	-0.06	-0.022	0.1	-0.067	-0.03	-0.03	-0.031	-0.033	-0.12	-0.11	0.0034	-0.019	0.16	0.012	0.059	-0.084	-0.042	0.03	1	-0.11
Comedy -	-0.089	-0.27	-0.09	-0.082	-0.071	-0.17	0.09	-0.055	-0.12	0.055	-0.029	-0.08	0.2	0.054	-0.037	-0.34	-0.018	-0.2	0.12	-0.25	-0.0079	-0.11	1
	rating -	Drama -	Orime -	War -	Film-Noir -	Mystery -	Animation -	Documentary -	IMAX -	Musical -	Westem -	Adventure -	Romance -	Fantasy -	movield -	Thriller -	timestamp –	Sci-Fi -	Children -	Action -	userld -	Horror -	Comedy -

Significant genres that contribute to good ratings.







PROBLEM STATEMENT

Movie Recommendations

- Top N recommendations for each user
 - Input : User ID
 - Output: Top N movies based on decreasing order of predicted ratings
- Evaluation
 - Relevance
 - Precision
 - Recall
 - F Measure
 - Prediction Accuracy
 - RMSE
 - MAE

APPROACH

Ensemble Learning

- Ensemble learning is an approach that combines different predictive models in order to improve the results obtained by individual models.
- Good balance between bias and variance
- Reduce Noise and avoid overfitting
- We can create hybrid model of Collaborative as well as content based approach.

Ensemble Rating = Mean(Ratings from different recommendation model)

Content Based Approach

- list of genres for a movie as content.
- created features and vectorized each movie using TF-IDF.
- Found out cosine similarity between movies user has already rated and movies we need to be recommended
- For predicting rating for the movie i (movie_i), we computed the weighted average rating with all the movies that the user has rated(movie_u)
- Restrict just with the movie with positive cosine similarity.
- If no positive cosine similarity movie available we consider mean of movie_u as the rating for movie_i

Collaborative Filtering - Neural Network

- Used the two embedding layers to represent the users and the movies as our input.
- The interactions (rating matrix) is the target variable.
- Merge both the layer using concatenation layer.
- We then try to minimize the loss via back-propagation and optimize the embedding layers

Layer (type)	Output	Shape	Param #	Connected to
Movie (InputLayer)	(None,	1)	0	
User (InputLayer)	(None,	1)	0	
Movie-Embedding (Embedding)	(None,	1, 8)	431120	Movie[0][0]
User-Embedding (Embedding)	(None,	1, 5)	1416145	User[0][0]
FlattenMovies (Flatten)	(None,	8)	0	Movie-Embedding[0][0]
FlattenUsers (Flatten)	(None,	5)	0	User-Embedding[0][0]
dropout_1 (Dropout)	(None,	8)	0	FlattenMovies[0][0]
dropout_2 (Dropout)	(None,	5)	0	FlattenUsers[0][0]
concatenate_1 (Concatenate)	(None,	13)	0	dropout_1[0][0] dropout_2[0][0]
FullyConnected-1 (Dense)	(None,	100)	1400	concatenate_1[0][0]
FullyConnected-2 (Dense)	(None,	50)	5050	FullyConnected-1[0][0]
FullyConnected-3 (Dense)	(None,	20)	1020	FullyConnected-2[0][0]
Activation (Dense)	(None,	1)	21	FullyConnected-3[0][0]

Collaborative Filtering - SVD and ALS

- Singular Vector Decomposition (SVD):
 - o a matrix decomposition technique
 - used for dimensionality reduction, noise reduction and also compression.
 - o popular because any real matrix can be decomposed -> more stable than eigen decomposition.
 - A real matrix is decomposed into three matrices, say U, Sigma, V^T,
 - where U and V are composed of orthonormal columns
 - Sigma is a diagonal matrix with singular values, which help us deduce the most significant correlated properties.
- Alternating Least Squares(ALS):
 - a matrix factorization technique.
 - o runs in a parallel fashion,
 - o minimizing two loss functions alternatively.
 - It first fixes the user matrix and runs gradient descent with item matrix,
 - it fixes the item matrix and runs gradient descent with user matrix.
 - o simple, scalable and works on even sparse data.

Collaborative Filtering - NMF

- Discovers latent factors in utility matrix
- Maps users and movies to a k-dimensional concept space
- Intuitively : Clustering the columns of the utility matrix
- H gives the cluster membership, i.e., if H $kj \ge H$ ij for all ik, this suggests that the input data vj belongs to kth cluster
- W gives the cluster centroids, i.e., the kth column gives the cluster centroid of kth cluster

 components

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_k \end{bmatrix} \quad W = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_k \end{bmatrix} \quad H = \begin{bmatrix} h_1 \\ h_2 \\ \dots \\ h_k \end{bmatrix}$$

$$x_{i} = \begin{bmatrix} w_{i1} & w_{i2} & \dots & w_{ik} \end{bmatrix} \times \begin{bmatrix} h_{1} \\ h_{2} \\ \dots \\ h_{k} \end{bmatrix} = \sum_{j=1}^{k} w_{ij} \times h_{ij}$$

$$w_{i} : \text{ weights}$$

Collaborative Filtering - Slope One

- Simple Linear predictor, improved efficiency relative to Linear Regression
- Precompute the average difference between the ratings of one item and another for users who rated both

The prediction \hat{r}_{ui} is set as:

$$\hat{r}_{ui} = \mu_u + \frac{1}{|R_i(u)|} \sum_{j \in R_i(u)} \operatorname{dev}(i, j),$$

where $R_i(u)$ is the set of relevant items, i.e. the set of items j rated by u that also have at least one common user with i. dev(i,j) is defined as the average difference between the ratings of i and those of j:

$$\operatorname{dev}(i,j) = rac{1}{|U_{ij}|} \sum_{u \in U_{ij}} r_{ui} - r_{uj}$$

Collaborative Filtering - KNN with Baseline

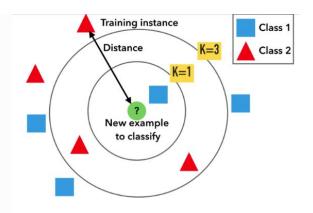
- KNN relies on item feature similarity
- Returns top K nearest neighbor movies based on Pearson Correlation "distance" between the target movie and every other movie
- Baseline estimate added to KNN to model for biases in rating

The prediction \hat{r}_{ui} is set as:

$$\hat{r}_{ui} = b_{ui} + rac{\sum\limits_{v \in N_i^k(u)} ext{sim}(u,v) \cdot (r_{vi} - b_{vi})}{\sum\limits_{v \in N_i^k(u)} ext{sim}(u,v)}$$

or

$$\hat{r}_{ui} = b_{ui} + rac{\sum\limits_{j \in N_u^k(i)} ext{sim}(i,j) \cdot (r_{uj} - b_{uj})}{\sum\limits_{j \in N_u^k(i)} ext{sim}(i,j)}$$



Collaborative Filtering - CoClustering

- Simultaneous clustering along the rows and columns of the utility matrix
- Each user and item assigned to cluster and co-cluster
- Final rating depends on the average rating of user cluster and movie cluster

The prediction \hat{r}_{ui} is set as:

$$\hat{r}_{ui} = \overline{C_{ui}} + (\mu_u - \overline{C_u}) + (\mu_i - \overline{C_i}),$$

RESULTS

1M DATASET

Model	RMSE	MAE	PRESICION	RECALL	FMeasure
NeuralNet	0.95	0.71	0.67	0.67	0.67
ContentBased	0.94	0.71	0.66	0.66	0.66
SVD	0.87	0.68	0.68	0.68	0.68
NMF	0.91	0.72	0.66	0.66	0.66
CoClustering	0.91	0.71	0.67	0.67	0.67
ALS	1.25	1.16	0.62	0.62	0.62
KNN	0.89	0.70	0.67	0.67	0.67
Ensemble	0.87	0.68	0.68	0.68	0.68

Full Latest(27M) Dataset

Model	RMSE	MAE	Precision	Recall	FMeasure	
NeuralNet	0.90	0.66	0.83	0.83	0.83	
SVD	0.78	0.59	0.85	0.85	0.85	
NMF	0.87	0.67	0.82	0.82	0.82	
CoClustering	0.89	0.69	0.83	0.83	0.83	
ALS	1.29	1.01	0.77	0.77	0.77	
Ensemble	0.80	0.61	0.84	0.84	0.84	

CONCLUSION

CONCLUSION

 Ensemble techniques balanced the bias variance trade-off and provide better result then base learner.

 Using ensemble learning we can combine different types of algorithms such as Content based, Collaborative filtering etc.

FUTURE POSSIBILITIES

 We can improve our model by using weighted average of the rating rather than taking simple average.

 Our result might improve more if we incorporate learning to rank based approached in base model.

REFERENCES

- 1. N. S. Altman. 1992. An Introduction to Kernel and Nearest-Neighbor Non-parametric Regression. ☐ e American Statistician 46, 3 (1992), 175–185. h☐p://www.istor.org/stable/2685209
- 2. Luis M. [de Campos], Juan M. Fern and Ez-Luna, Juan F. Huete, and Miguel A.Rueda-Morales. 2010. Combining content-based and collaborative recommen-dations: A hybrid approach based on Bayesian networks. International Journal of Approximate Reasoning 51, 7 (2010), 785 799. DOI: hp://dx.doi.org/hps://doi.org/10.1016/j.ijar.2010.04.001
- 3. Chris Ding, Xiaofeng He, and Horst D. Simon. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. Inin SIAM International Conference on Data Mining.
- 4. Arthur Fortes and Marcelo Manzato. 2014. Ensemble Learning in RecommenderSystems: Combining Multiple User Interactions for Ranking Personalization. WebMedia 2014 Proceedings of the 20th Brazilian Symposium on Multimedia and the Web(11 2014), 47–54. DOI:h□ p://dx.doi.org/10.1145/2664551.2664556
- 5. Simon Funk. 2006. Net □ix Update: Try □is at Home. (2006). h□ps://si□er.org/~simon/journal/20061211.html
- 6. Salma Ghoneim. 2019.Accuracy, Recall, Precision, F-ScoreSpeci-□city, which to optimize on?(2019).h□ ps://towardsdatascience.com/accuracy-recall-precision-f-score-speci□city-which-to-optimize-on-867d3f11124
- 7. [7]G´erard Govaert and Mohamed Nadif. 2008. Block clustering with Bernoullimixture models: Comparison of di□erent approaches.Computational Statistics &Data Analysis52, 6 (February 2008), 3233–3245. h□ ps://ideas.repec.org/a/eee/csdana/v52y2008i6p3233-3245.html
- 8. GroupLens. 2019. MovieLens. (2019). hps://grouplens.org/datasets/movielens/
 - [... Please refer the project report for the complete set of references]

THANK YOU

A&Q