Movie Clustering and Rating Prediction for Content Organisation

Saravanan Arumugam

Reenae Victor

Introduction

In the ever-changing entertainment landscape, there's a growing need for personalized content recommendations and efficient organization due to the expanding movie catalog.

This project focuses on developing a smart movie clustering system for improved organization and user experience. Additionally, it aims to create a precise movie rating prediction model, aligning with industry demands for tailored content delivery.

01 - Literature Review

01/ Literature Review

- 2 technical papers were chosen to get a better understanding for the implementation of our project.
- These papers were based on a movie prediction system and user preference predictions using a database.
- With the knowledge gained from these papers it helped us in implementing tasks such as hierarchical clustering, silhouette score, SSE graph and random forest (with a subset of 50% data)

02 - Data Description

02/
Data Description

IMDb Dataset Overview:

- Utilizing the Kaggle platform, we leverage the IMDb dataset for comprehensive insights into the world of movies.

Key Files in the Dataset:

- title.akas.tsv.gz: Details on localized titles, regions, languages, and more.
- title.basics.tsv.gz: Information on titles, encompassing type, primary/original titles, genres, and additional details.
- title.principals.tsv.gz: Insights into the principal cast and crew for various titles.
- title.ratings.tsv.gz: IMDb ratings and vote statistics, crucial for assessing the popularity and reception of titles.
- name.basics.tsv.gz: Comprehensive information about individuals involved in the industry, such as actors, directors, birth/death years, professions, and notable titles.

'df_name'

	nconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
0	nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0053137,tt0050419,tt0072308,tt0031983
1	nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0037382,tt0038355,tt0117057,tt0075213
2	nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department	tt0056404,tt0054452,tt0057345,tt0049189
3	nm0000004	John Belushi	1949	1982	actor,soundtrack,writer	tt0077975,tt0078723,tt0072562,tt0080455
4	nm0000005	Ingmar Bergman	1918	2007	writer,director,actor	tt0050986,tt0050976,tt0069467,tt0083922

(12936747, 6)

'df_akas'

'df_akas'

	titleId	ordering	title	region	language	types	attributes	isOriginalTitle
0	tt0000001	1	Карменсіта	UA	\N	imdbDisplay	\N	0
1	tt0000001	2	Carmencita	DE	\N	\N	literal title	0
2	tt0000001	3	Carmencita - spanyol tánc	HU	\N	imdbDisplay	\N	0
3	tt0000001	4	Καρμενσίτα	GR	\N	imdbDisplay	\N	0
4	tt0000001	5	Карменсита	RU	\N	imdbDisplay	\N	0

(37517220, 8)

'df_basics'

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
0	tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
1	tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation,Short
2	tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation,Comedy,Romance
3	tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	12	Animation,Short
4	tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy,Short

(10246265, 9)

'df_principals'

	tconst	ordering	nconst	category	job	characters
0	tt0000001	1	nm1588970	self	\N	["Self"]
1	tt0000001	2	nm0005690	director	\N	\N
2	tt0000001	3	nm0374658	cinematographer	director of photography	\N
3	tt0000002	1	nm0721526	director	\N	\N
4	tt0000002	2	nm1335271	composer	\N	\N
5	tt0000003	1	nm0721526	director	\N	\N
6	tt0000003	2	nm1770680	producer	producer	\N
7	tt0000003	3	nm1335271	composer	\N	\N
8	tt0000003	4	nm5442200	editor	\N	\N
9	tt0000004	1	nm0721526	director	\N	\N
10	tt0000004	2	nm1335271	composer	\N	\N
11	tt0000005	1	nm0443482	actor	\N	["Blacksmith"]
12	tt0000005	2	nm0653042	actor	\N	["Assistant"]
13	tt0000005	3	nm0005690	director	\N	\N
14	tt0000005	4	nm0249379	producer	producer	\N
15	tt0000006	1	nm0005690	director	\N	\N
16	tt0000007	1	nm0179163	actor	\N	\N
17	tt0000007	2	nm0183947	actor	\N	\N
18	tt0000007	3	nm0005690	director	\N	\N
19	tt0000007	4	nm0374658	director	\N	\N
20	tt0000007	5	nm0249379	producer	producer	\N

'df_ratings'

	tconst	averageRating	numVotes
0	tt0000001	5.7	2002
1	tt0000002	5.8	269
2	tt0000003	6.5	1893
3	tt0000004	5.5	178
4	tt0000005	6.2	2678

(1360168, 3)

03 - EDA

03/ EDA

- Performing Exploratory Data Analysis (EDA) is a crucial step in understanding a dataset and gaining insights into the underlying patterns and relationships.
- We have understood the data for the columns essential to us after getting rid of all the null values.
- As you can see we have created various plots in EDA to investigate the relationship between different datasets.

Histogram

GRAPHS for EDA

Correlation Matrix

Bar Plot

Time-Series Plots

Scatter-Plot

04 - Approach

04/ Approach

Rating Prediction

Clustering

05 - Result

Results

For Predictions

Model	MSE (Without PCA)	RMSE (Without PCA)	MAE (Without PCA)	R-squared (Without PCA)	Computing Time (Without PCA)
Linear Regression	1.652633621814121	1.2855479850297775	0.9965003229939857	0.16226080881355764	0.223466157913208
Lasso Regression	1.9727589610221221	1.4045493800582884	1.0963325922837333	-1.4446395161460401e-05	0.17841482162475586
Ridge Regression	1.6526323766613444	1.2855475007409662	0.996506762255127	0.1622614399959722	0.14157986640930176
Linear Regression	1.5242035746899396	1.2345863982281433	0.9335617308827272	0.22736349242212672	271.8334357738495

Model	MSE (With PCA)	RMSE (With PCA)	MAE (With PCA)	R-squared (With PCA)	Computing Time (With PCA)
Linear Regression	1.6845836042932902	1.2979150990312465	1.0072967129676207	0.14606499134547346	0.22043752670288086
Lasso Regression	1.9727589610221221	1.4045493800582884	1.0963325922837333	-1.4446395161460401e-05	0.12948131561279297
Ridge Regression	1.6845836041454756	1.2979150989743033	1.0072967123563425	0.1460649914204024	0.11056184768676758
Random Forrest	1.5288005472562631	1.236446742587914	0.9349111088239197	0.22482582828685838	180.8328583240509

Hierarchical Clustering Scatter Plot

Dendrogram with threshold 60

K-Means Clustering with PCA

06 - Conclusion

Conclusion

Linear Models and PCA Impact:
Linear Regression, Lasso, and Ridge models showed similar performance, with a slight difference in metrics.
PCA had limited impact on linear models, suggesting careful consideration of the trade-off between computational efficiency and model performance.

Random Forest Trade-off:
Random Forest outperformed Linear Regression but incurred higher computational costs.
Consider the trade-off between predictive accuracy and efficiency when deciding on model selection.

Clustering Efficiency:
K-Means proved more efficient for larger datasets compared to hierarchical clustering.
Considering the substantial time difference (200 seconds for hierarchical vs. 300 seconds for K-Means), prioritize K-Means for scalability in clustering tasks.

Future Scope:
Prediction - Include textual features using NLP to improve prediction model.
Clustering - More complex models capture relationships better.