METIS®

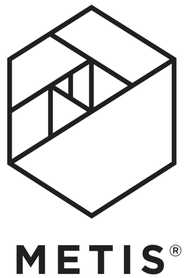
IMDB user rating prediction

TV shows

project for Metis EDA Bootcamp

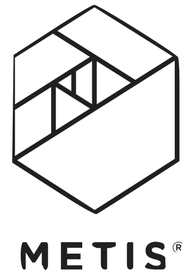
by Krystian Krystkowiak, 2022

Introduction

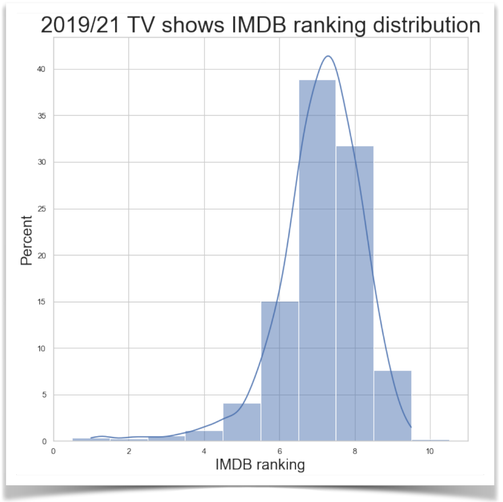


- Golden age of television 1999-now.
- To create model that can be utilised to predict how well a TV show may be received by viewers.
- GOAL: To identify factors that may help TV producers to make key decisions while production planning.

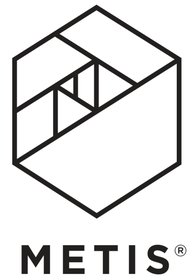
Data



- Data scraped from IMDB.
- TV shows released between 2019 and 2021 with rating count at least 1,000, to avoid vote manipulation and less popular shows.
- fake_useragent library and random intervals used to appear anonymous.
- 1335 rows of data
- Target: IMDB show ranking.



Data



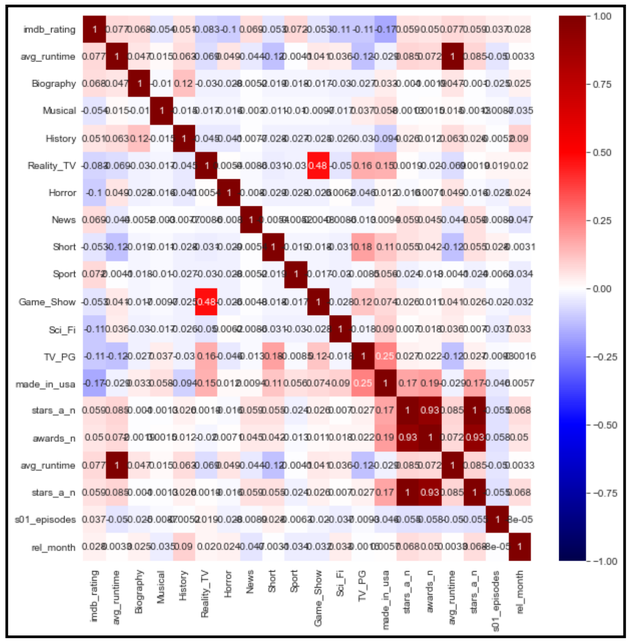
- Features: 11

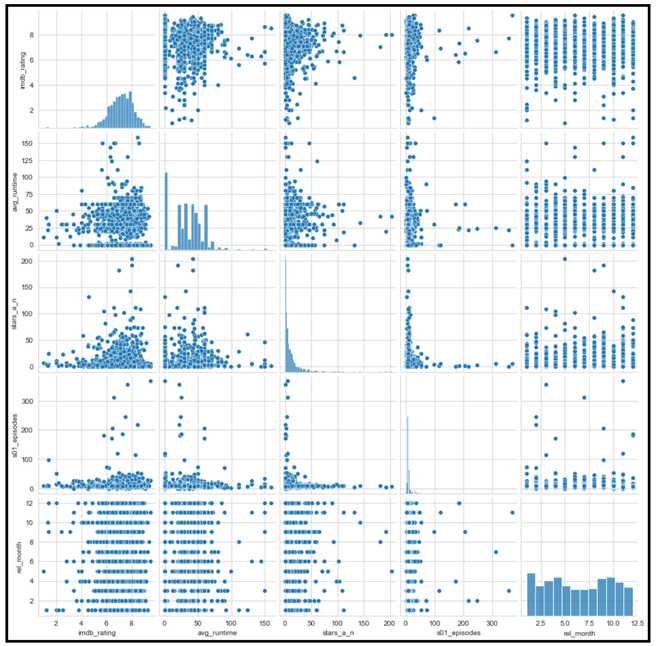
imdb_rating	title	s01_episodes	avg_runtime	genres	rel_date	certification	origin	company	creators	creators_a	stars	stars_a
5.8	The Masked Singer	10	60	[Game-Show, Music, Reality-TV]	2019-01-02	TV-PG	United States	[Smart Dog Media, Fox Alternative Entertainmen...	[]	[]	[Jenny McCarthy-Wahlberg, Ken Jeong, Nicole Sc...	[3, 5, 1]
6.2	Siempre Bruja	11	40	[Drama, Fantasy]	2019-01-01	TV-14	Colombia	[Caracol]	[]	[]	[Sofía Araujo Mejía, Angely Gaviria, Sofía Ara...	[0, 0, 0]
6.6	Tidying Up with Marie Kondo	8	40	[Reality-TV]	2019-01-01	TV-PG	United States	[Netflix, The Jackal Group]	[Marie Kondo]	[1]	[Marie Kondo, Charlotte Hervieux, Marie Iida]	[1, 0, 0]

- Numerical: s01_episodes, avg_runtime, rel_date.
- Categorical: genres, MPAA certification, origin, company, creators, stars.

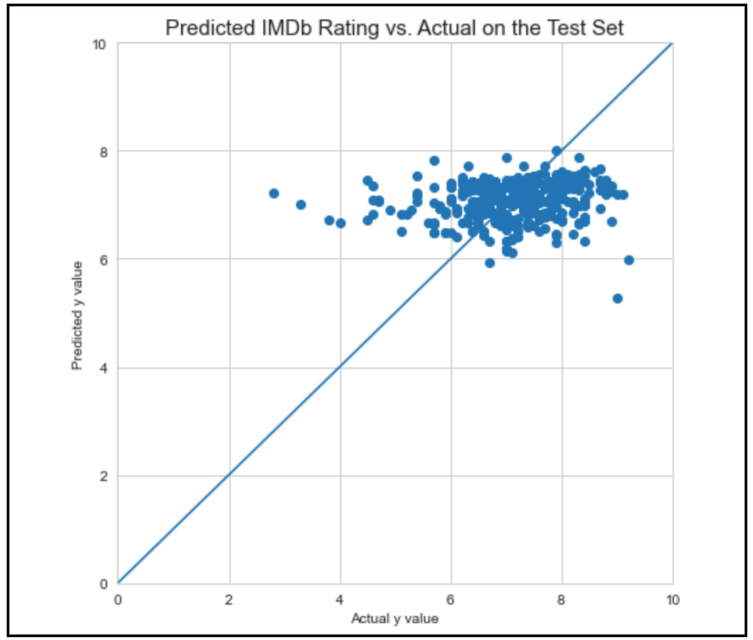
Results

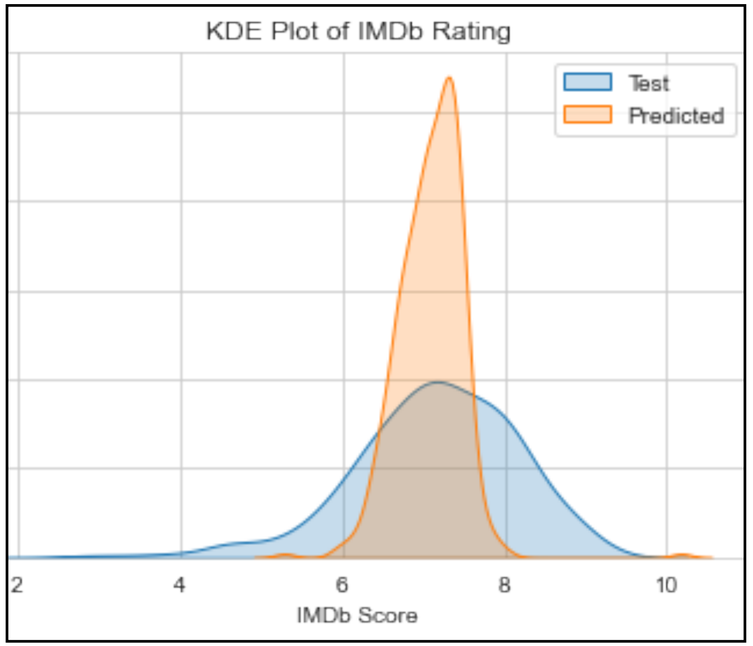
- Features not strongly correlated with the target.
- Scraped more data.
- Engineered: combination of features, counters.





Results





- Linear Regression
- 75train/25test split
- 0.096/-0.021 R-squared
- Better predictions around median.

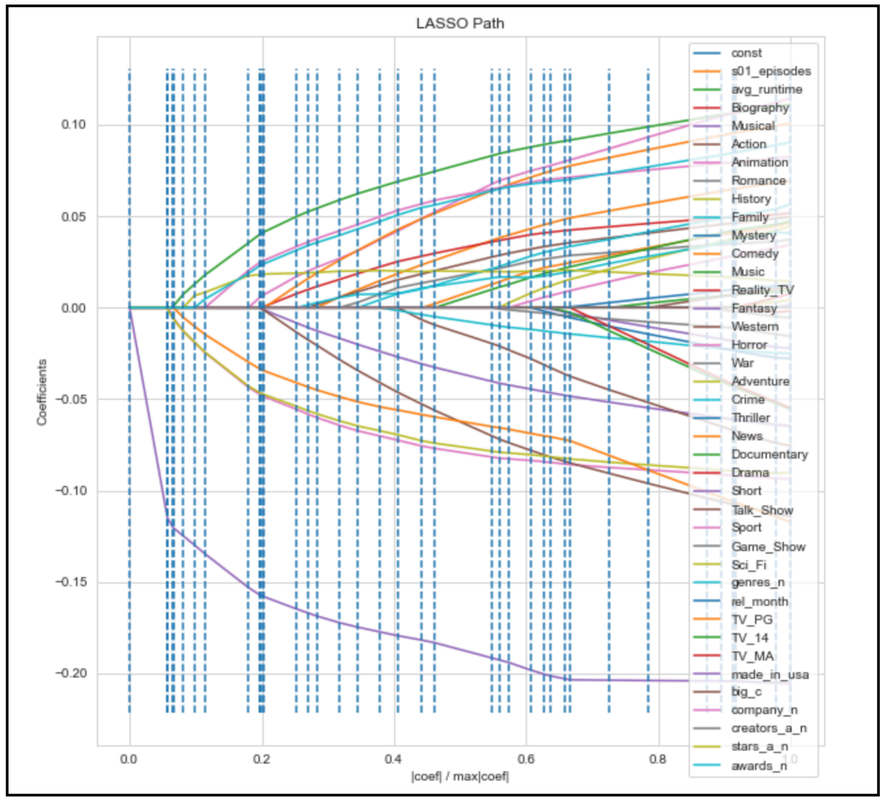
Results

- Ridge:

  MAE 0.781, R-squared 0

- LASSO:

  MAE 0.782, R-squared -



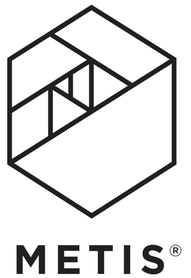
Conclusions

- Animations and shows related to News or Sport, productions with awarded main cast tend to be rated higher.
- Horrors, Sci-FI, Talk Shows, US productions and PG materials tend to be rated lower.
- The world is complicated.

Dep. Variable:	imdb_rating	**R-squared:**	0.096
Model:	OLS	**Adj. R-squared:**	0.087
Method:	Least Squares	**F-statistic:**	10.71
Date:	Tue, 12 Jul 2022	**Prob (F-statistic):**	2.01e-17
Time:	21:10:15	**Log-Likelihood:**	-1528.6
No. Observations:	1016	**AIC:**	3079.
Df Residuals:	1005	**BIC:**	3133.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.2880	0.047	155.646	0.000	7.196	7.380
avg_runtime	0.1136	0.037	3.107	0.002	0.042	0.185
Animation	0.2464	0.105	2.350	0.019	0.041	0.452
Horror	-0.5847	0.190	-3.084	0.002	-0.957	-0.213
News	3.1120	1.206	2.580	0.010	0.745	5.479
Talk_Show	-1.3391	0.492	-2.722	0.007	-2.305	-0.374
Sport	0.6547	0.255	2.568	0.010	0.154	1.155
Sci_Fi	-0.4742	0.164	-2.897	0.004	-0.795	-0.153
TV_PG	-0.3017	0.119	-2.531	0.012	-0.536	-0.068
made_in_usa	-0.4358	0.075	-5.839	0.000	-0.582	-0.289
stars_a_n	0.1002	0.035	2.887	0.004	0.032	0.168

Future Work



- TV Shows budgets are not information that producers share so widely as movie budgets.
- Could be interesting to explore actors information. Facebook, Instagram, Tweeter.
- Exploration of TOP TV shows lists to extract “success factor”.

METIS®

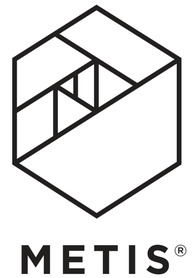
Thank you!

Questions?

project for Metis EDA Bootcamp

by Krystian Krystkowiak, 2022

Sources



- Photos: google