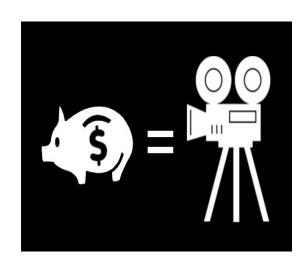
Domestic Revenue Prediction Model for Movies

Pramila Chaudhary

BACKSTORY



 Prediction model for an Investment Firm

 Interested in Knowing the Domestic Revenue of movies before release

Data Scraped

From: https://www.boxofficemojo.com

Target: Total Domestic Revenue

For: Movies (2011 - 2020)

Scraped Data points: 2000

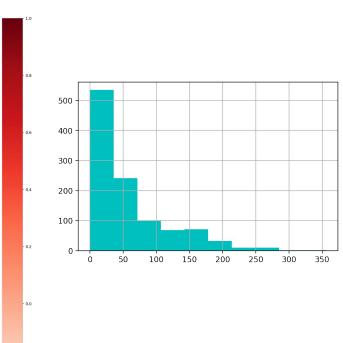
Features Scraped : 12

Process:

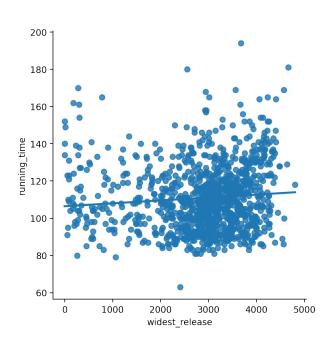
- 1. Web Scrape
- 2. EDA to select features
- 3. Regression Analysis
- 4. Prediction

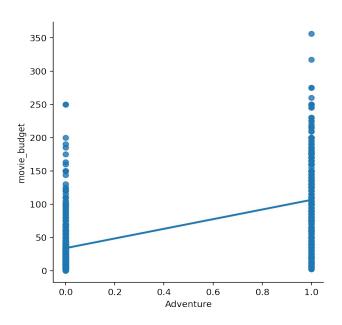
Movie budget

domestic_gross -	1				0.046	0.054	0.047		0.35			-0.0096		-0.05	-0.18		0.1	-0.052	-0.11	-0.03	0.099	-0.12	-0.081			-0.11		
movie_budget -	0.64	1	0.39	0.55	0.14	0.092	0.078		0.59	0.18	-0.13	-0.08	-0.11	-0.075	-0.26	0.19	0.28	-0.04	-0.23	-0.11	0.067	-0.19	-0.14	0.38	-0.041	-0.13	-0.0041	0.059
running_time -	0.31		1	0.088	-0.15	0.019	-0.085		0.059	-0.31	0.19	-0.31	0.042	-0.074	0.25	-0.29	-0.13		-0.26	-0.008	-0.077	-0.096	0.00091		0.079	0.0085		0.12
widest_release -	0.59	0.55	0.088	1	0.22	0.07	0.023				-0.12	0.079	-0.044	-0.079	-0.3			-0.098	-0.012	-0.075	0.0029	-0.037	-0.06		-0.028	-0.016	-0.054	0.042
rel_year -	0.046		-0.15		1	0.46	0.025	0.11	0.11	0.061	-0.26	0.017	-0.038	0.023	-0.25	0.097	0.11	-0.14	0.11	-0.027	0.017	0.071	-0.0063	0.11	-0.026	0.055	-0.086	-0.077
rel_month -	0.054	0.092	0.019	0.07	0.46	1	0.069	-0.031	0.045	0.078	-0.064	0.042	-0.022	0.038	-0.043	0.083	0.044	-0.079	0.013	0.028	0.083	-0.012	-0.067	-0.0047	0.061	-0.013	-0.028	-0.032
G -	0.047	0.078	-0.085	0.023	0.025	0.069	1	-0.072	0.11	0.25	-0.026	0.074	-0.041	-0.008	-0.019	0.18	0.058	-0.019	-0.037	0.091		-0.035	-0.038	-0.047	0.11	-0.064	-0.016	-0.0098
Action -	0.2				0.11	-0.031	-0.072	1	0.32	-0.12	-0.13	-0.25	0.15	-0.068	-0.26	-0.21	0.0013	0.001	-0.15	-0.14	-0.15	-0.13	-0.25		-0.031	0.25	0.095	0.063
Adventure -	0.35	0.59	0.059		0.11	0.045	0.11	0.32	1	0.4	-0.16	0.052	-0.21	-0.061	-0.31	0.42		-0.04	-0.17	-0.095		-0.15	-0.14	0.35	-0.053	-0.17	-0.027	0.019
Animation -	0.15		-0.31	0.19	0.061	0.078	0.25	-0.12	0.4	1	-0.091	0.35	-0.084	-0.028	-0.17	0.72	0.36	-0.048	-0.064	-0.011	0.28	-0.094	-0.076	0.029	0.028	-0.2	-0.055	-0.0037
Biography -	-0.062	-0.13	0.19	-0.12	-0.26	-0.064	-0.026	-0.13	-0.16	-0.091	1	-0.15	-0.042	-0.025	0.32	-0.055	-0.14	0.29	-0.11		-0.021	-0.11	-0.065	-0.14		-0.13		0.0035
Comedy -	-0.0096	-0.08	-0.31	0.079	0.017	0.042	0.074	-0.25	0.052	0.35	-0.15	1	0.0038	-0.047	-0.21	0.33	0.14	-0.12	-0.2	-0.00092		-0.2	0.15	-0.16	-0.034	-0.45	-0.1	-0.048
Crime -	-0.077	-0.11	0.042	-0.044	-0.038	-0.022	-0.041		-0.21	-0.084	-0.042	0.0038	1	-0.039	0.049	-0.13	-0.19	-0.039	-0.14	-0.074	-0.078	0.076	-0.15	-0.16	-0.042	0.26	-0.077	-0.00018
Documentary -	-0.05	-0.075	-0.074	-0.079	0.023	0.038	-0.008	-0.068	-0.061	-0.028	-0.025	-0.047	-0.039	1	-0.077	-0.038	-0.043	0.038	-0.035		-0.018	-0.033	-0.035	-0.044	-0.014	-0.061	-0.015	-0.0092
Drama -	-0.18	-0.26	0.25	-0.3	-0.25	-0.043	-0.019	-0.26	-0.31	-0.17	0.32	-0.21	0.049	-0.077	1	-0.15	-0.17		-0.16	0.09	-0.033	-0.016		-0.23		-0.02		0.013
Family -	0.13		-0.29		0.097	0.083		-0.21	0.42	0.72	-0.055	0.33	-0.13	-0.038	-0.15	1	0.47	-0.061	-0.11	-0.0022		-0.097	-0.06	-0.028	0.057	-0.28	-0.06	-0.022
Fantasy -	0.1	0.28	-0.13		0.11	0.044	0.058	0.0013	0.39	0.36	-0.14	0.14	-0.19	-0.043	-0.17	0.47	1	-0.063	0.029	-0.045		-0.068	0.042	0.0036	-0.081	-0.24	-0.043	-0.03
History -	-0.052	-0.04	0.17	-0.098	-0.14	-0.079	-0.019	0.001	-0.04	-0.048	0.29	-0.12	-0.039	0.038	0.2	-0.061	-0.063	1	-0.081	-0.014	-0.041	-0.063	-0.055	-0.09	-0.003	-0.049	0.37	0.07
Horror -	-0.11	-0.23	-0.26	-0.012	0.11	0.013	-0.037	-0.15	-0.17	-0.064	-0.11	-0.2	-0.14	-0.035	-0.16	-0.11	0.029	-0.081	1	-0.078	-0.081	0.44	-0.12	0.019	-0.066	0.28	-0.037	-0.043
Music -	-0.03	-0.11	-0.008	-0.075	-0.027	0.028	0.091	-0.14	-0.095	-0.011		-0.00092	-0.074		0.09	-0.0022	-0.045	-0.014	-0.078	1	0.089	-0.075	0.077	-0.074	-0.032	-0.13	-0.034	-0.021
Musical -	0.099	0.067	-0.077	0.0029	0.017	0.083		-0.15	0.14	0.28	-0.021	0.14	-0.078	-0.018	-0.033	0.35	0.22	-0.041	-0.081	0.089	1	-0.049	0.096	-0.09	-0.033	-0.14	-0.035	-0.022
Mystery -	-0.12	-0.19	-0.096	-0.037	0.071	-0.012	-0.035	-0.13	-0.15	-0.094	-0.11	-0.2	0.076	-0.033	-0.016	-0.097	-0.068	-0.063	0.44	-0.075	-0.049	1	-0.085	-0.0063	-0.063	0.32	-0.05	-0.014
Romance -	-0.081	-0.14	0.00091	-0.06	-0.0063	-0.067	-0.038	-0.25	-0.14	-0.076	-0.065	0.15	-0.15	-0.035	0.16	-0.06	0.042	-0.055	-0.12	0.077	0.096	-0.085	1	-0.13	-0.05	-0.18	-0.055	-0.018
Sci-Fi -	0.28	0.38			0.11	-0.0047	-0.047	0.34	0.35	0.029	-0.14	-0.16	-0.16	-0.044	-0.23	-0.028	0.0036	-0.09	0.019	-0.074	-0.09	-0.0063	-0.13	1	-0.083	0.061	-0.046	-0.0098
Sport -	-0.043	-0.041	0.079	-0.028	-0.026	0.061	0.11	-0.031	-0.053	0.028	0.27	-0.034	-0.042	-0.014		0.057	-0.081	-0.003	-0.066	-0.032	-0.033	-0.063	-0.05	-0.083	1	-0.1	0.0066	-0.017
Thriller -	-0.11	-0.13	0.0085	-0.016	0.055	-0.013	-0.064		-0.17	-0.2	-0.13	-0.45	0.26	-0.061	-0.02	-0.28	-0.24	-0.049		-0.13	-0.14	0.32	-0.18	0.061	-0.1	1	-0.038	-0.018
War -	0.0029	-0.0041		-0.054	-0.086	-0.028	-0.016	0.095	-0.027	-0.055		-0.1	-0.077	-0.015		-0.06	-0.043	0.37	-0.037	-0.034	-0.035	-0.05	-0.055	-0.046	0.0066	-0.038	1	0.035
Western -	0.0045	0.059	0.12	0.042	-0.077	-0.032	-0.0098	0.063	0.019	-0.0037	0.0035	-0.048	-0.00018	-0.0092	0.013	-0.022	-0.03	0.07	-0.043	-0.021	-0.022	-0.014	-0.018	-0.0098	-0.017	-0.018	0.035	1
	domestic_gross -	movie_budget -	running_time -	widest_release -	rel_year -	rel_month -	.9	Action -	Adventure	Animation -	Biography -	Comedy	Crime -	Documentary	Drama -	Family	- Fantasy -	History -	Horror	Music	Musical -	Mystery -	- Bomance -	Sci-Fi	Sport	Thriller	War	Western

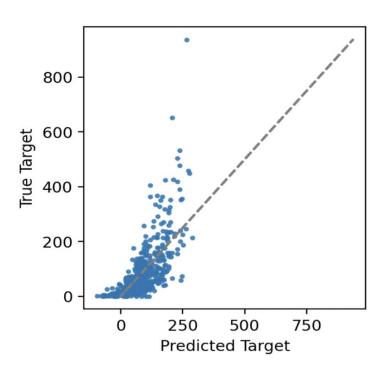


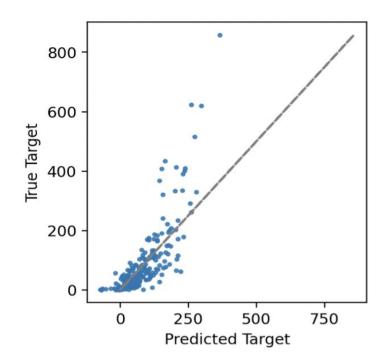
Feature to Feature relationship





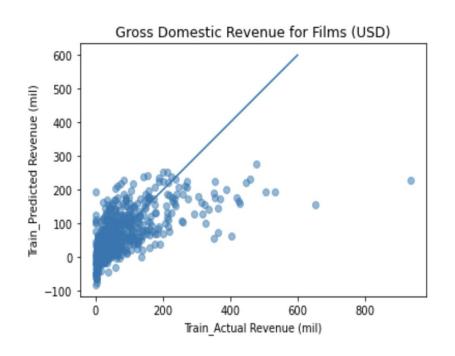
Linear Regression model on train/validation data

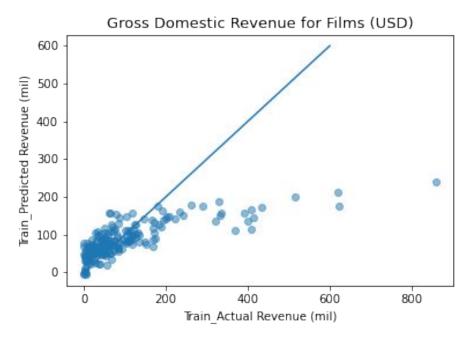




R2 for train = 0.49MSE = 66.87 R2 for val = 0.56MSE = 77

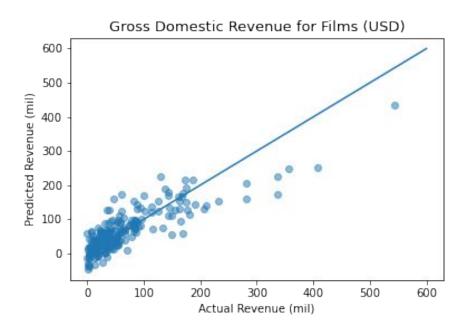
Lasso model on train /validation data set





R2 for train = 66MSE = 74.38 R2 for train = 77MSE = 91

Final prediction on test data set using Linear Regression model



R2	71
MSE	40

Conclusions

- Prediction with linear regression showing better result than lasso model
- R2 = 71
- Mean Squared Error (MSE) = 40

- Future Direction:
 - Focus on feature selection to avoid overfitting
 - minimize error
 - Including more feature