What Makes a Hit?

Predicting Movie Success

Joseph Brazzale

Introduction

:

- The movie industry generated 42.5 Billion dollars in revenue in 2019
- WIth the rise in ticket costs, and more options for consuming media it has Become increasingly difficult to get people to go to the movies.
- There are many different factors that go into making a movie, which ones should Studios focus on?

Methodology

01

Obtain Data

Data was scraped from Boxofficemojo using Scrapy

02

EDA

Initial exploration of data performed with SQLite and Pandas

03

Initial discovery

Feature engineering was performed to find correlation

04

Modeling

Various linear models applied to find best R2

01

Target

Domestic revenue will be used as my target value

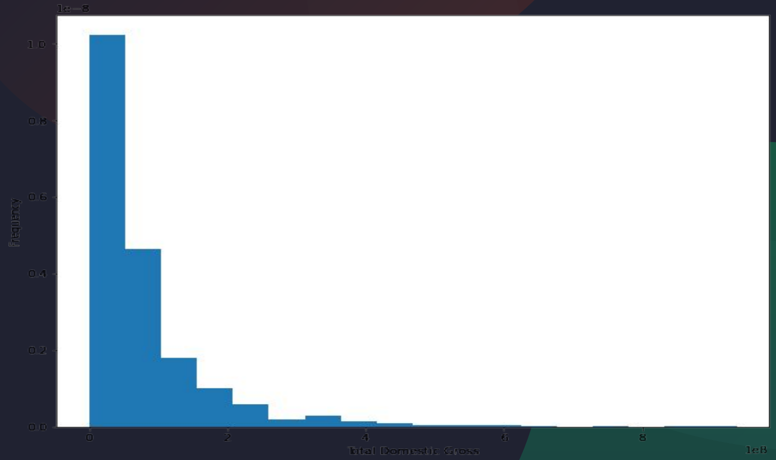
Features

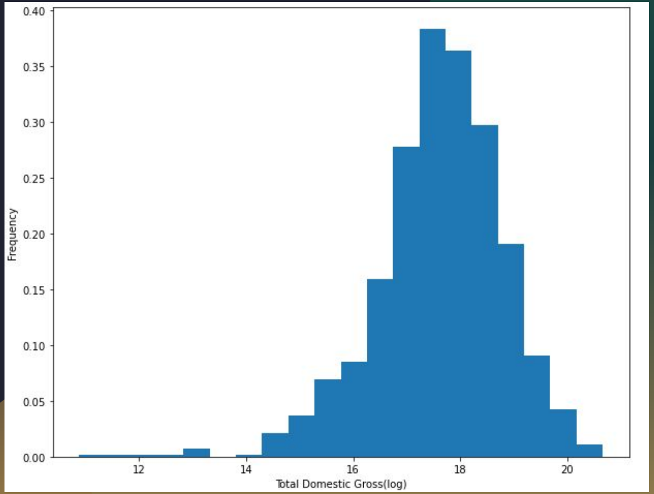
Budget
Opening Weekend
Release Date
MPAA-Rating

Exploration & Results

There are a few adjustments that needed to be made to get the best results

- Budget, revenue, and opening weekend data needs to be normalized
- All categorical data needs to have "dummy" variables assigned

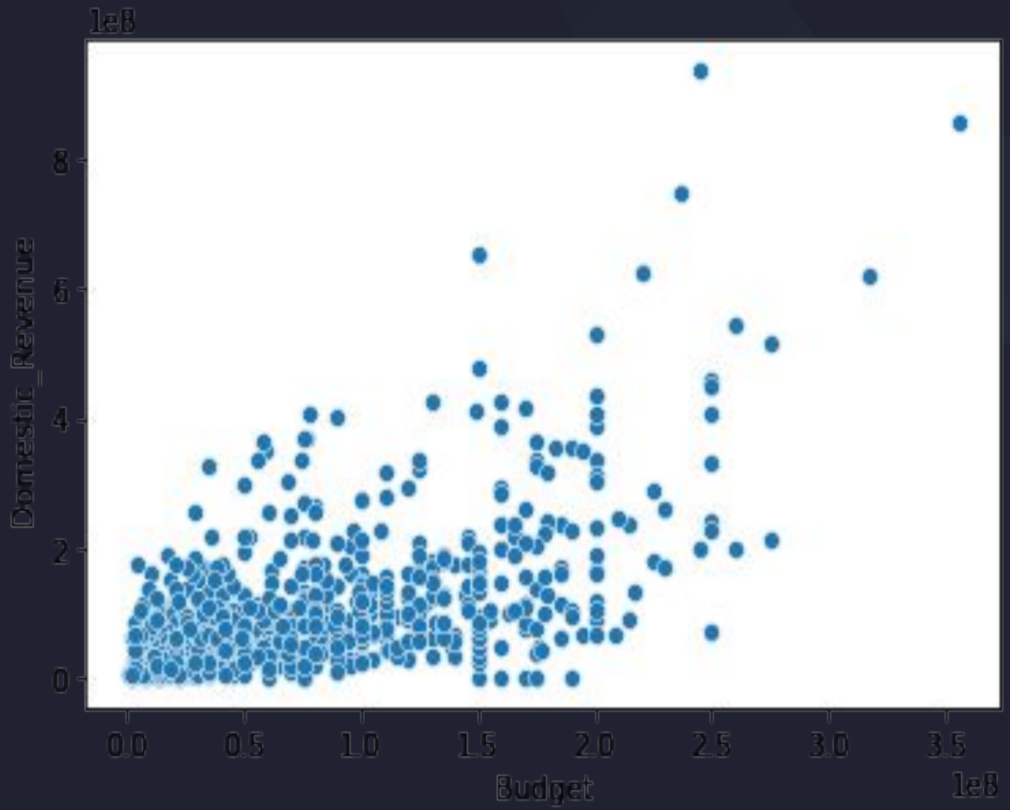




Initial Scores

Initial R-squared score of

0.52



Best features

Budget

Budget had a very strong correlation. You have to spend money to make money

Season

The Fall/Holiday season had the strongest correlation with revenue

Rating

G rating performed the best, R was the worst

Applying models to improve score

Which worked best?

Ridge regression showed the most improvement on r2

Other methods

Polynomial, and Lasso showed no significant improvements

Final Score

By applying Ridge, I was able to get a final score of .65

```
clf.score(X_test,y_test)
Out[54]: 0.653816745695938
```

Score Across Features

Dep. Variable:	dom_log_target	**R-squared:**	0.632
Model:	OLS	**Adj. R-squared:**	0.627
Method:	Least Squares	**F-statistic:**	125.1
Date:	Thu, 30 Sep 2021	**Prob (F-statistic):**	9.59e-151
Time:	23:14:56	**Log-Likelihood:**	-817.89
No. Observations:	739	**AIC:**	1658.
Df Residuals:	728	**BIC:**	1708.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
log_opening_w	0.6823	0.040	16.947	0.000	0.603	0.761
log_budget	0.2241	0.034	6.518	0.000	0.157	0.292
no_Theaters_open	-0.0003	5.96e-05	-5.495	0.000	-0.000	-0.000
Running_Time(min)	0.0074	0.002	4.072	0.000	0.004	0.011
G	1.7600	0.487	3.616	0.000	0.805	2.715
PG	1.3418	0.375	3.574	0.000	0.605	2.079
PG_13	1.3060	0.382	3.423	0.001	0.557	2.055
R	1.3012	0.369	3.527	0.000	0.577	2.025
winter	1.3751	0.380	3.619	0.000	0.629	2.121
spring	1.4056	0.380	3.696	0.000	0.659	2.152
summer	1.3553	0.379	3.575	0.000	0.611	2.100
fall_holiday	1.5729	0.382	4.118	0.000	0.823	2.323

Omnibus:	112.854	**Durbin-Watson:**	1.951
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	852.641
Skew:	-0.430	**Prob(JB):**	7.10e-186
Kurtosis:	8.192	**Cond. No.**	1.98e+19

Key Features have strong correlation

Future Work

How can we improve?

Thank you

The Genre Effect:
Due to the high number of genres attached to each film, I was unable to score based on genre

See if there is any correlation from specific directors or actors

Appendix: Checking Assumptions