Movies & Beyond

Metis Spring 2018

Jacob Levine

Problem

- Do movies associated with a more decorated cast/director typically perform better in terms of domestic gross total?
- How does the predictive value of ‘star power’ features compare to other features such as budget, rating, etc?

Sources and Methods

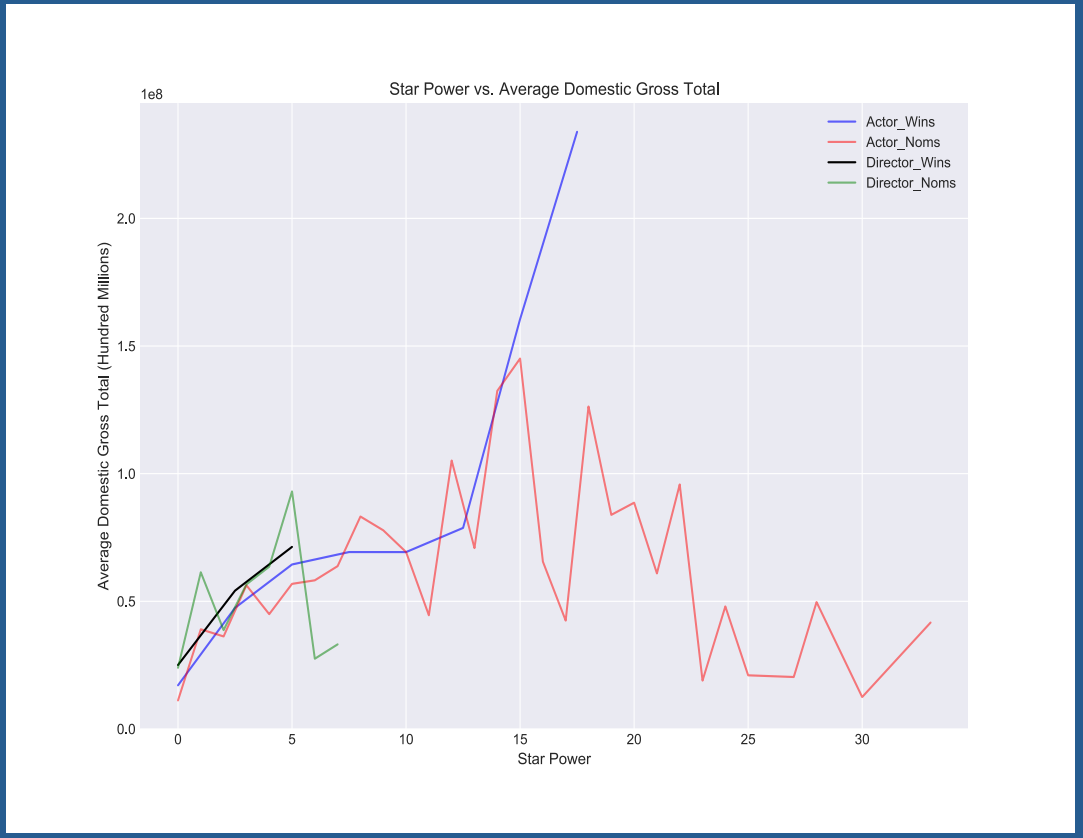
- Websites Scraped:
  - BoxOfficeMojo
  - Oscars.org
- Tools used:
  - BeautifulSoup
  - Scikit-learn
  - Statsmodels

Assumptions

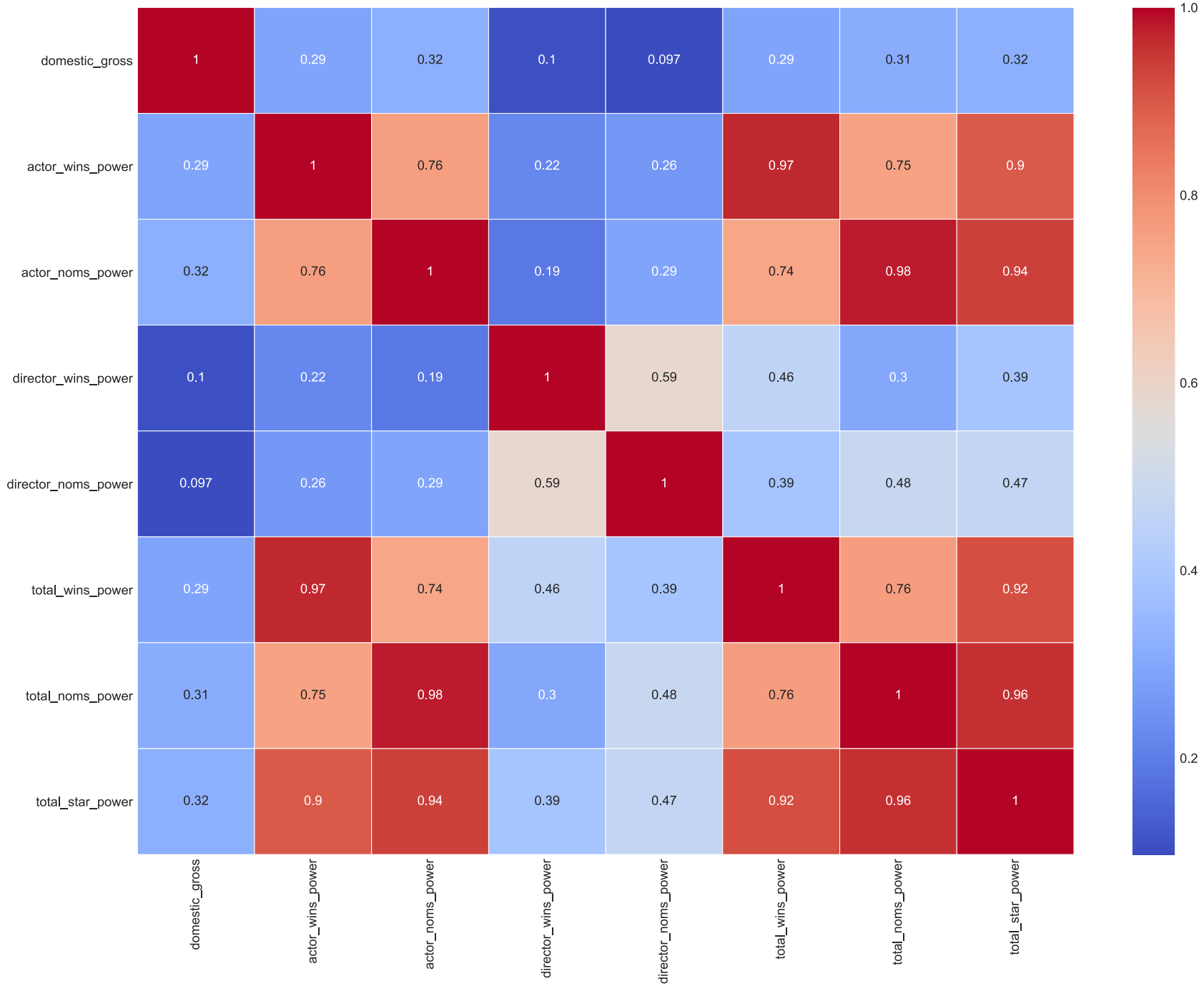
- Movies that perform better in the box office tend to feature more Oscar award winning/ Oscar nominated actors/director
- Oscar wins hold more weight than Oscar nominations
- Movies with higher domestic gross totals tend to have more information about them readily available on the internet

Analysis – Part 1

Type	Points
Actor Win	2.5
Actor Nomination	1
Director Win	2.5
Director Nomination	1

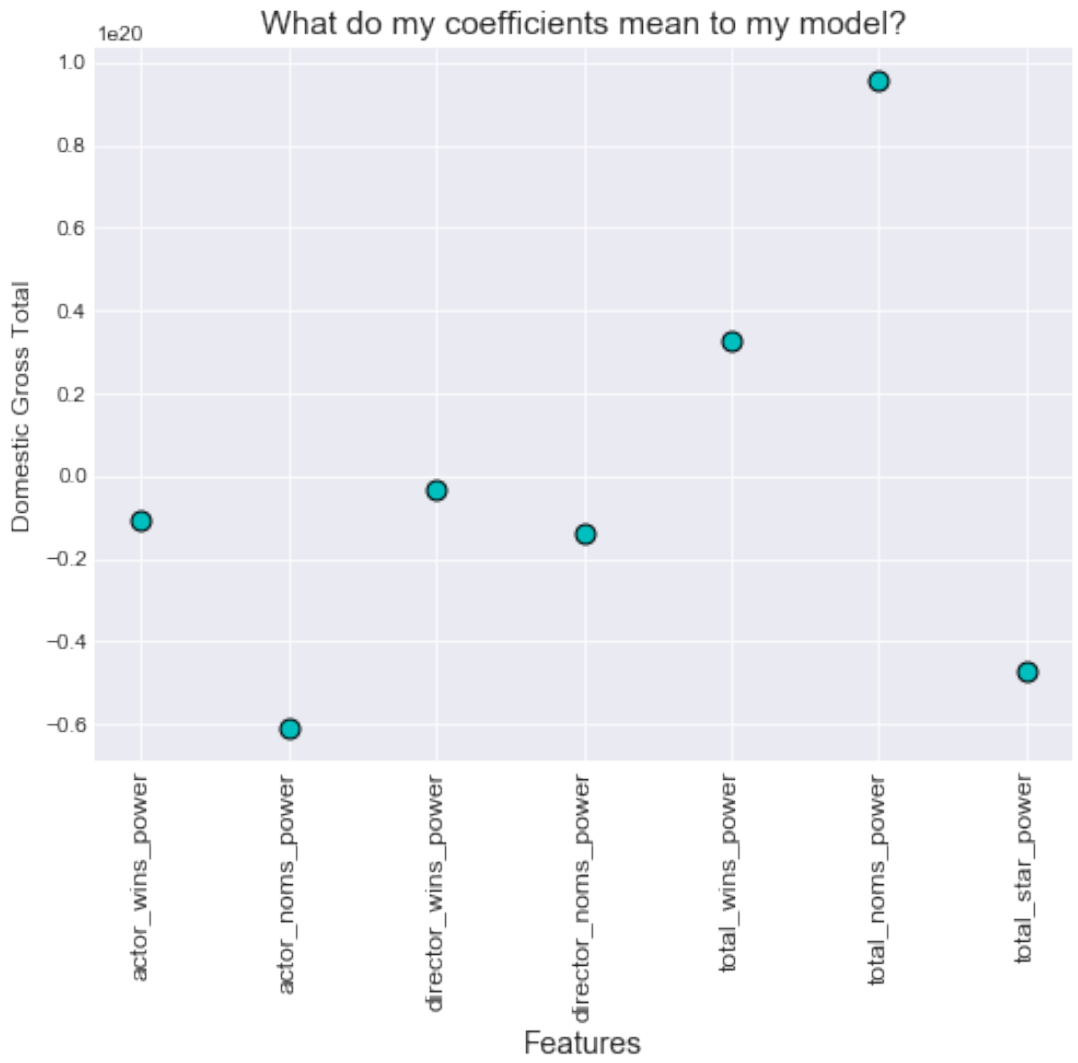


n = 6,103


domestic_gross
actor_wins_power
actor_noms_power
director_wins_power
director_noms_power
total_wins_power
total_noms_power
total_star_power
1 0.29 0.32 0.1 0.097 0.29 0.31 0.32
0.29 1 0.76 0.22 0.26 0.97 0.75 0.9
0.32 0.76 1 0.19 0.29 0.74 0.98 0.94
0.1 0.22 0.19 1 0.59 0.46 0.3 0.39
0.097 0.26 0.29 0.59 1 0.39 0.48 0.47
0.29 0.97 0.74 0.46 0.39 1 0.76 0.92
0.31 0.75 0.98 0.3 0.48 0.76 1 0.96
0.32 0.9 0.94 0.39 0.47 0.92 0.96 1
domestic_gross
actor_wins_power
actor_noms_power
director_wins_power
director_noms_power
total_wins_power
total_noms_power
total_star_power
1.0
0.8
0.6
0.4
0.2

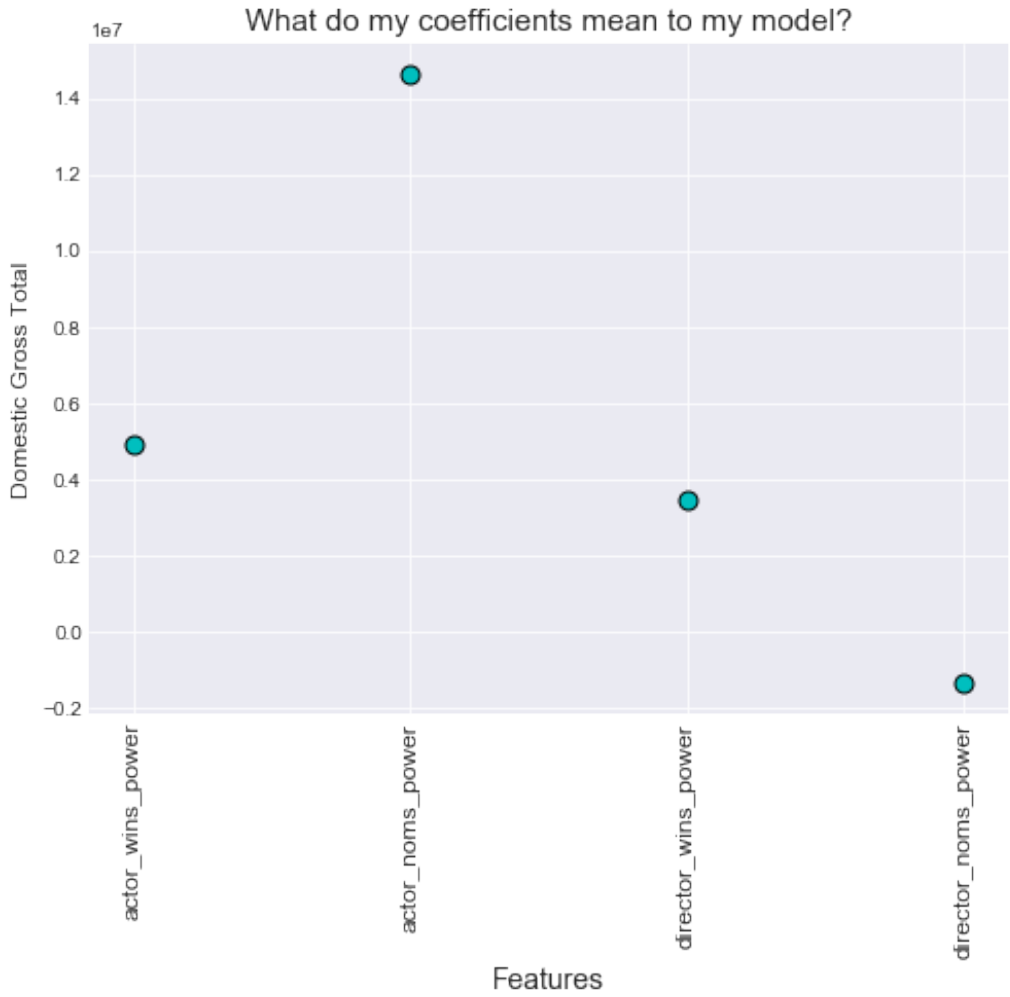
Model Interpretation

- Test R^2 – 0.127
- Total Nominations Power is the most telling feature



Model Interpretation

- Test R^2 – 0.127
- Total Nominations Power broken down
- Actor Nominations Power is up



StatsModels

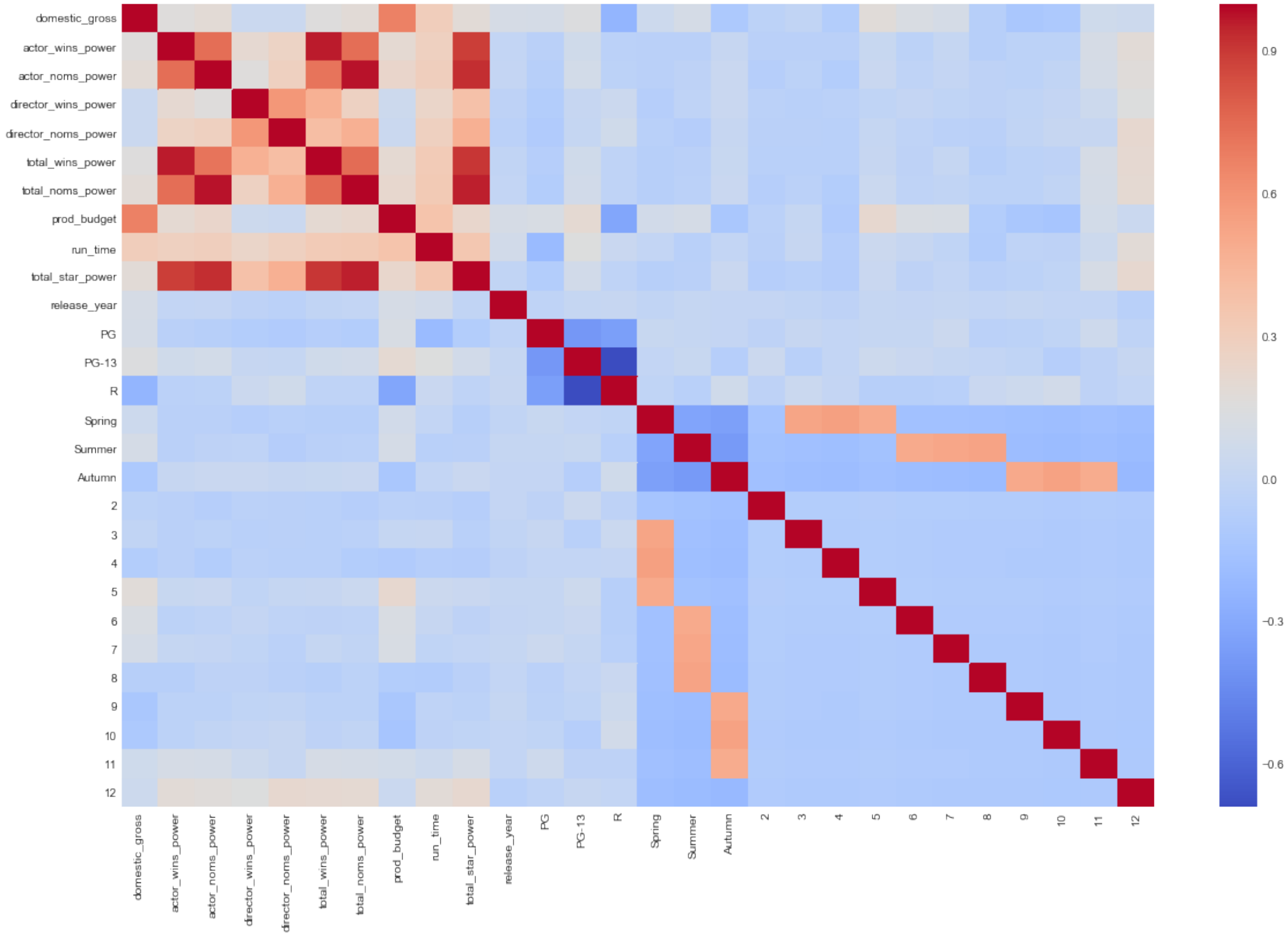
Dep. Variable:	y_train	**R-squared:**	0.103
Model:	OLS	**Adj. R-squared:**	0.103
Method:	Least Squares	**F-statistic:**	123.1
Date:	Thu, 26 Apr 2018	**Prob (F-statistic):**	1.53e-99
Time:	17:20:51	**Log-Likelihood:**	-82298.
No. Observations:	4272	**AIC:**	1.646e+05
Df Residuals:	4267	**BIC:**	1.646e+05
Df Model:	4		
Covariance Type:	nonrobust		

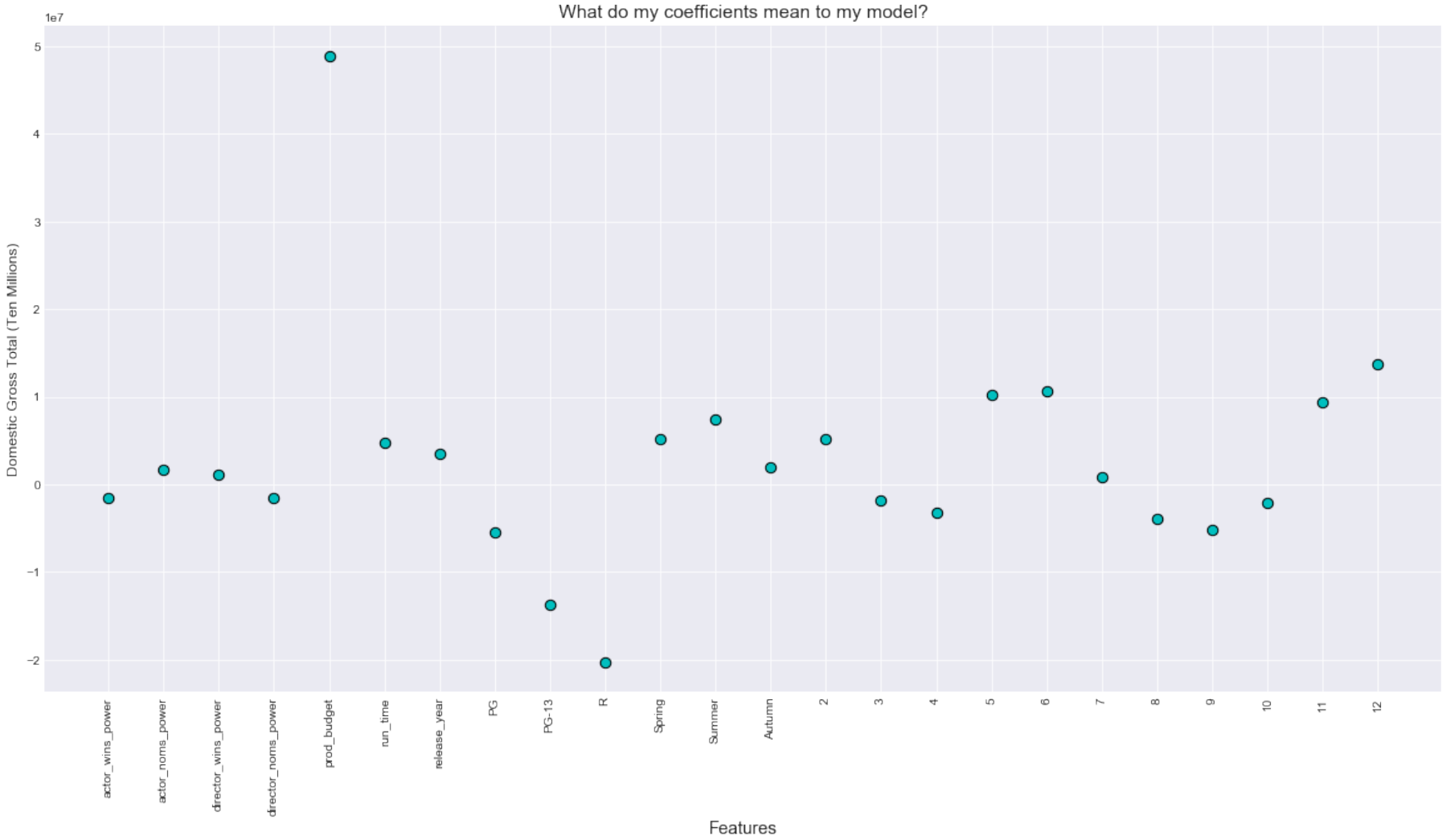
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.643e+07	8.61e+05	30.687	0.000	2.47e+07	2.81e+07
X_scaled[0]	4.898e+06	1.31e+06	3.730	0.000	2.32e+06	7.47e+06
X_scaled[1]	1.464e+07	1.31e+06	11.161	0.000	1.21e+07	1.72e+07
X_scaled[2]	3.467e+06	1.08e+06	3.209	0.001	1.35e+06	5.58e+06
X_scaled[3]	-1.365e+06	1.1e+06	-1.242	0.214	-3.52e+06	7.89e+05

Omnibus:	4221.834	**Durbin-Watson:**	1.959
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	288567.178
Skew:	4.727	**Prob(JB):**	0.00
Kurtosis:	42.138	**Cond. No.**	2.96

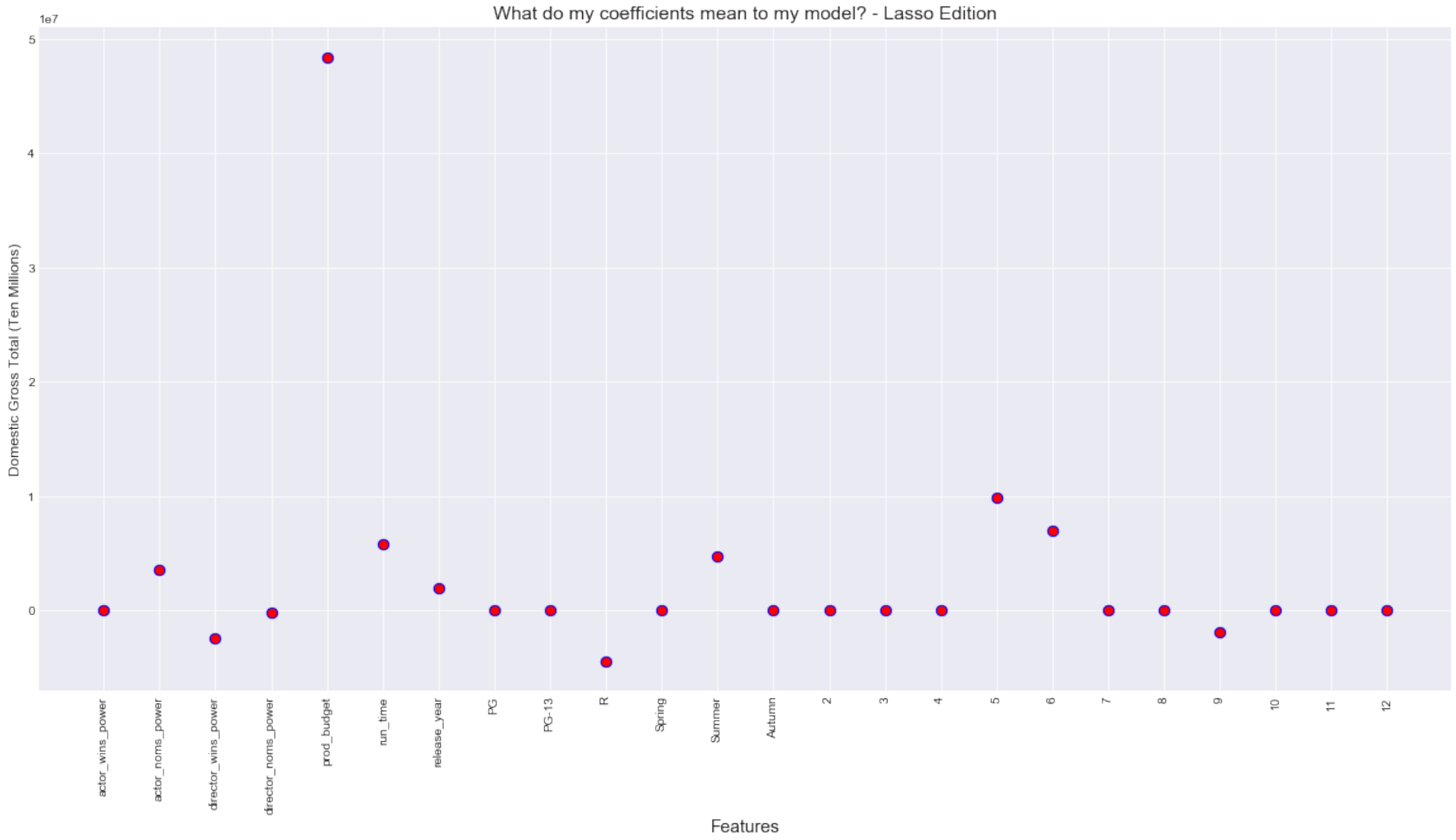
Analysis – Part 2

- Increased number of features
- Comparison between star power features and non star power related features
- Exclusion of nulls (budget, runtime, release date, titles with foreign genre) down to 2,001 samples


domestic_gross
actor_wins_power
actor_noms_power
director_wins_power
director_noms_power
total_wins_power
total_noms_power
prod_budget
run_time
total_star_power
release_year
PG
PG-13
R
Spring
Summer
Autumn
2
3
4
5
6
7
8
9
10
11
12
0.9
0.6
0.3
0.0
−0.3
−0.6



Test R^2 – 0.492 | Star Power Related features take backseat to Production Budget and certain seasonal features, there is still predictive power in Actor Nominations, Director Wins



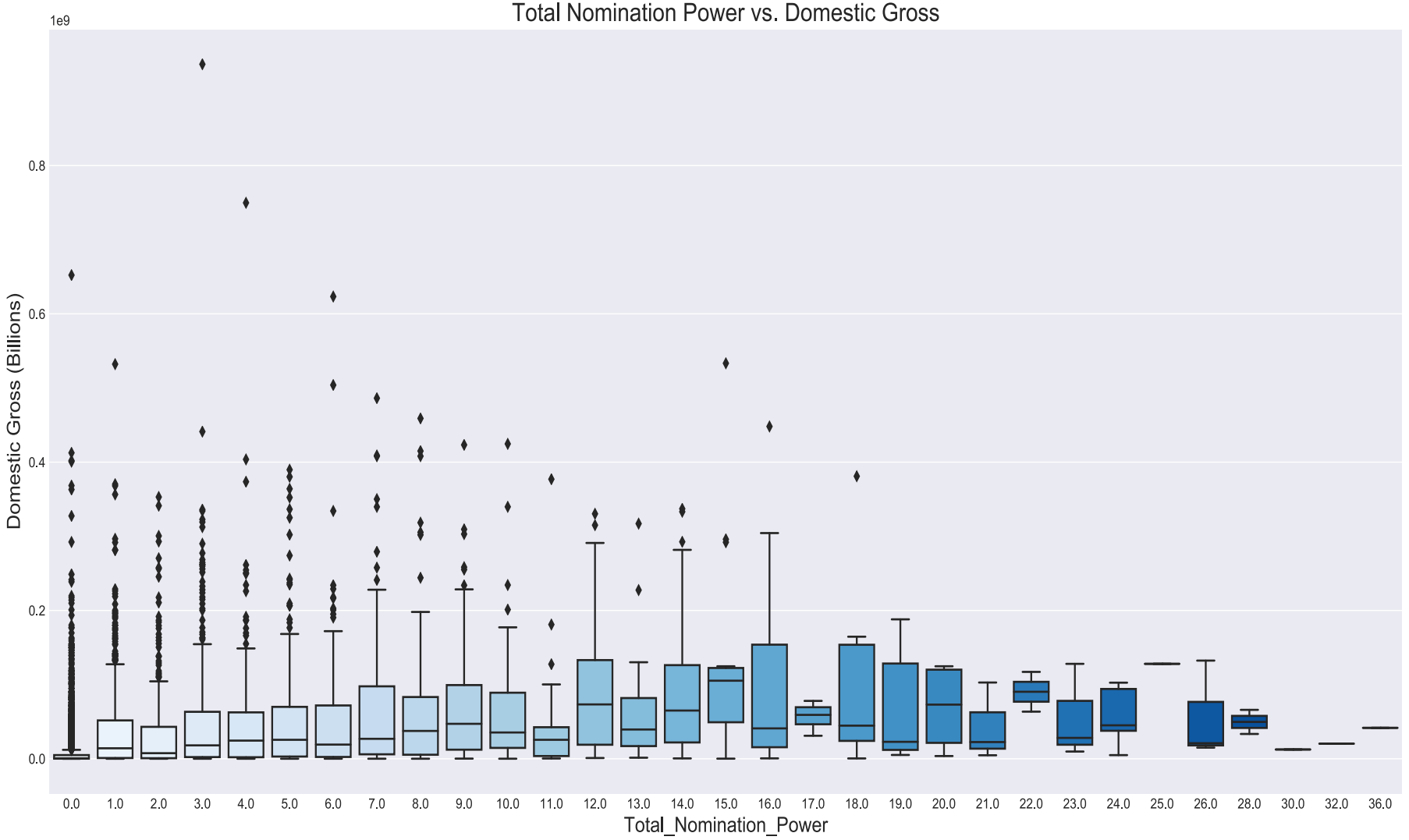
Test R^2 – 0.428 | Utilizing Lasso - Actor Nomination Power, runtime, certain seasonal features and of course budget hold weight.

Results

- Compared to Production Budget, star power features yield little predictive power. However, compared to other features such as runtime, year of release and season, certain star power features, particularly actor nominations power, hold their own
- Rating holds less power than I anticipated, late spring – early summer is best release period

Results

- Movies that feature Oscar winning and/or Oscar nominated cast/directors do tend to outperform movies that do not, however, more is not always better…


Total Nomination Power vs. Domestic Gross
1e9
0.8
0.6
0.4
0.2
0.0
Domestic Gross (Billions)
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
28.0
30.0
32.0
36.0
Total_Nomination_Power

Next Steps

- More Data!
- Many nulls in cast, production budget, genre, etc.
- Scrape titles from a wider range of years
- Experiment further with feature engineering and transformations

Thank You