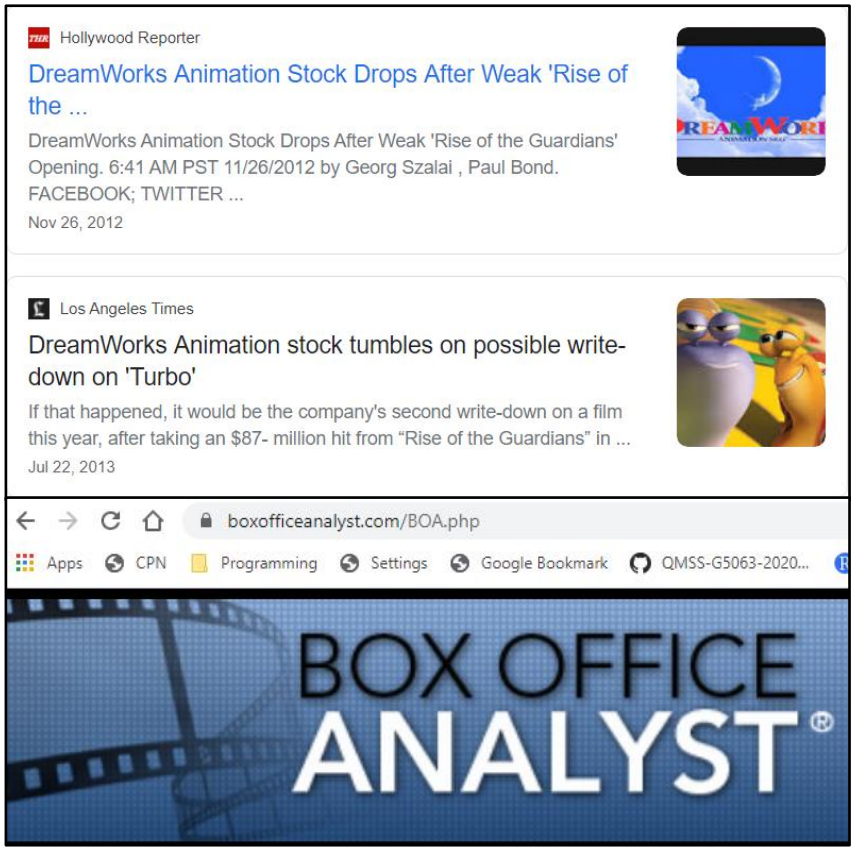
Predicting opening weekend box office for top movies

Tolu Bukola

June 2020

Why predict opening weekend box office?

- Openings are disproportionate to total gross:
  - Blockbusters make 30 – 40% of total gross on the opening weekend and 60 - 70% the first month.
  - Top 20 movies are typically 80% + of quarterly box.
- Stock prices of exhibitors (AMC, CNK) and small studios (LGF, DWA formerly) are highly sensitive to opening box office. Big studios (DIS, FOX) are less so.
- Value of predictions decreases after opening weekend. Expectations are priced in and new data is received by everyone in real time. Stocks move on “surprises”.
- Someone will totally pay you for it (Doug Stone – Boxofficeanalyst.com).



Opening weekend predictions have significant economic value

Data gathering and features

- ~960 of the biggest box office opening weekends gathered from Box Office Mojo (https://www.boxofficemojo.com/). Additional data from IMDB (https://pro.imdb.com/), and the-numbers (https://www.the-numbers.com/) via beautiful soup and Selenium. Missing data (i.e. budgets, mpaa rating, and prequels) filled in via Google search.
- Would prefer more data, limited by manual entry of prequel data.

Feature	Description
Opening gross (dependent)	**Opening weekend box office gross**
Prequel (categorical)	1/0 for whether movie has prequel
Prequel_gross	Opening weekend box for prequel
Budget	Movie budget
Theater	Number of theaters in opening
Starmeter_1 / Starmeter_2	Ranking of lead actor's current popularity
DirAveBox	Average box office gross of director's movies
Distpermove	Average box office gross of distributor's movies
Year	Year of movie opening
Genre (categorical)	One-hot encoded
MPAA rating (categorical)	One-hot encoded
Release quarter (categorical)	One-hot encoded
Distributor (categorical)	One-hot encoded

Seven numerical and four categorical variables were evaluated

Model refinement process

Model	Validation R^2	Delta to baseline	Training R2	Test R^2
Baseline model - Exclude distributors as categorical	**54.2%**		**58.8%**	**64.2%**
Model using only continuous features	51.9%	(2.3%)	54.0%	
Models using all features +	56.5%	2.3%	60.3%	65.9%
A theater squared feature	**58.6%**	**4.5%**	**64.4%**	**67.4%**
A theater squared feature and a budgets squared feature	**58.7%**	**4.5%**	**60.3%**	**69.2%**
Square features and interaction term for starmeter 1 and 2	**58.8%**	**4.6%**	**60.3%**	**69.6%**
term for prequel and prequel_gross	58.9%	4.7%	60.3%	69.4%
Polynomial features model	50.7%	(3.5%)		
Regularization				
Ridge regression	56.5%			65.9%
Lasso regression	56.5%			65.9%

- Pair plots appeared to show opening gross might have a polynomial relationship with number of theaters and budgets. **Final model included squared feature for theaters.**
- Slight improvements in the model could be made by adding extra polynomial and interaction terms, but tests suggested these models would be overfit.
- The choice model is highly interpretable and does not benefit from regularization.
- **Model RMSE is $22.4M, relative to an average opening gross of $45M and a range of $20 – 360M.**

The choice model is simple without sacrificing significant performance

Choice Model and Coefficients

Continuous features

prequel_gross : 0.23
theaters : -66595.36
budget : 0.04
starmeter_1 : 10.35
starmeter_2 : 0.77
prequel : -4579656.09
DirAveBox : 0.11
year : -808335.85
distpermov : -0.17
theatersq : 14.81

Mpaa rating

mpaa_PG : 4603328.69
mpaa_PG-13 : 10849903.84
mpaa_R : 13832593.68

Release quarter

quarter_2 : -2765017.37
quarter_3 : -4313900.34
quarter_4 : -2871123.30

Genre

Fantasy : -3774945.59
Musical : 1449510.72
Drama : 4615438.66
Mystery : 522779.74
Adventure : -2391272.12
Horror : 1467538.72
Romance : -1819614.86
Thriller : -4465848.25
Biography : -3150211.83
History : -6729892.15
Family : -6650115.12
Action : 1485912.23
Animation : -483880.49
War : -6533020.42
Crime : 244567.62
Documentary : 262744.59
Western : -13275795.47
Sci-Fi : 1261418.19
Sport : -3135899.19
Comedy : -955192.71

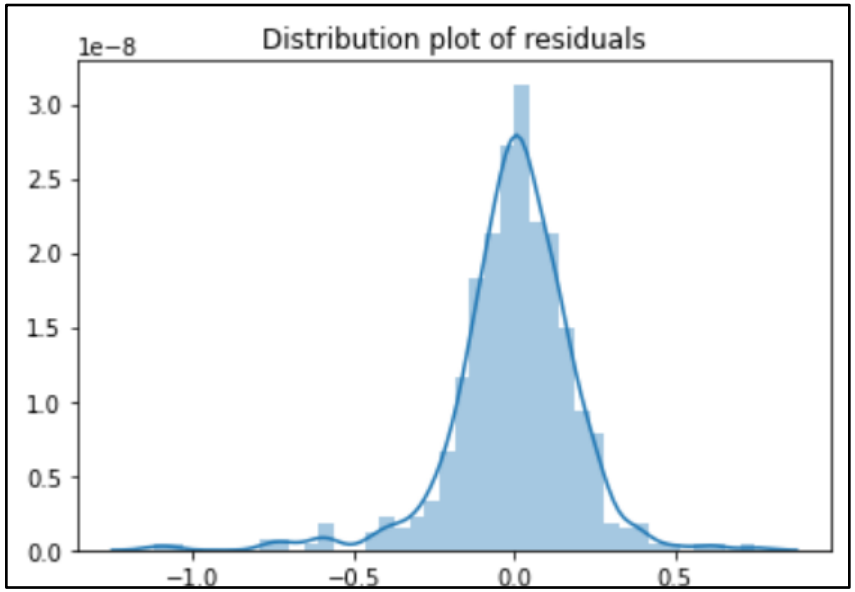
Distributor

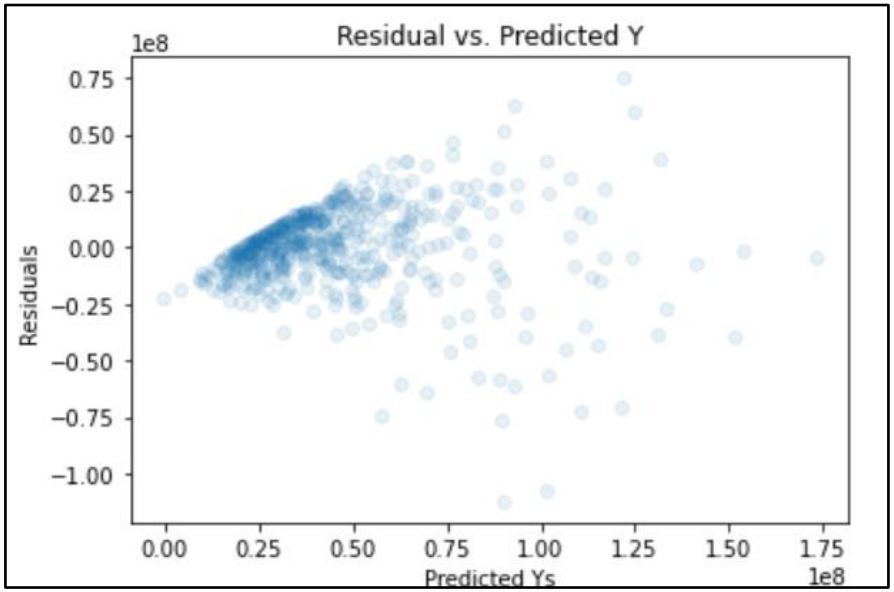
distributor_Lionsgate : 5228730.44
distributor_Otherstudios : 2665002.90
distributor_Paramount Pictures : 3987844.00
distributor_Screen Gems : 11159103.00
distributor_Sony Pictures Releasing : 2091336.02
distributor_Twentieth Century Fox : 3034012.05
distributor_Universal Pictures : 10819256.81
distributor_Walt Disney Studios : 16096868.89
distributor_Warner Bros. : 3970754.02

- Coefficients are largely Higher prequel gross, budget, DirAveBox, and theatre square, the key continuous models all drive higher opening weekend gross.
- Walt Disney Studios has the highest coefficient among distributors.
- Starmeter has opposite of expected signs. This could be due to feature construction since starmeter is not representative of at-the-time actor popularity.

Highly interpretable model mostly gives expected relationships

Error checking and residuals





- Residuals are symmetric around 0.
- Residual plot suggests that variance increases for higher values of the fitted. This heteroskedasticity can be partially addressed by a log-fitted model.
- The log model (see appendix) fits better on training data but worse on test.

Residuals are symmetric around 0 but heteroscedasticity is an issue

3-Fold Cross – Validation on Choice Model

Cross validation scores	Mean	Std.dev	95% interval	
Linear	57.5%	6.7%	44.1%	70.9%
Ridge	58.0%	6.1%	45.8%	70.2%
Scaled Ridge	57.1%	2.9%	51.3%	62.9%
Lasso	0.575	0.067	44.1%	70.9%

- 3 – fold cross validation is employed given the limited size of the data set.
- Cross validation scores are fair, but standard deviation is high. There is a wide range for the true score of the model.
- More data is needed to improve cross-validation.

Cross validation suggests model validity but with low confidence

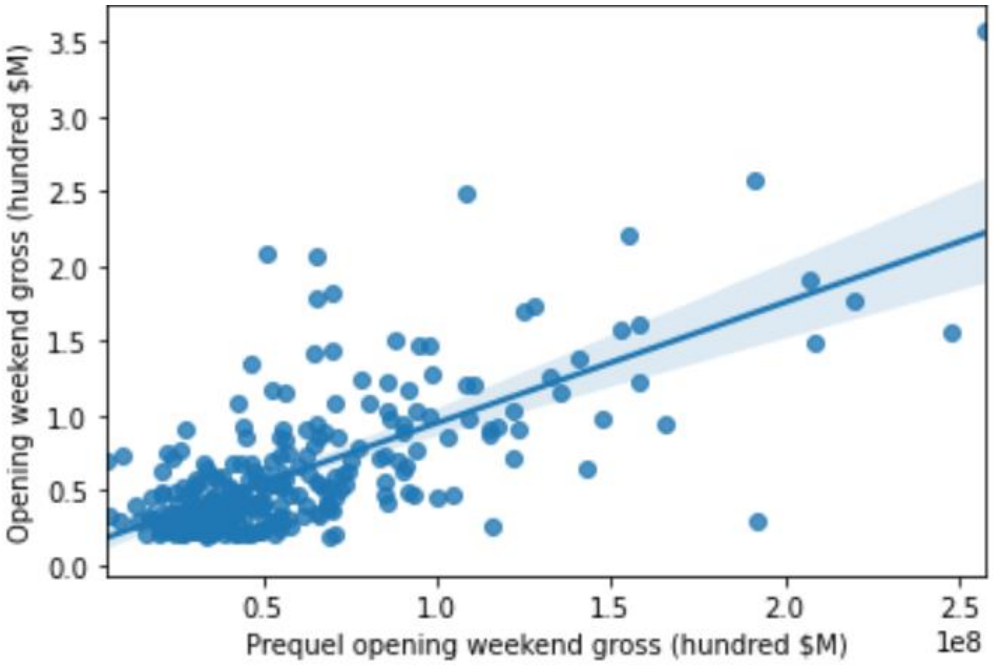
Unresolved issues and additional work

- More datapoints would increase confidence in model and validation measures.
- Starmeter is current ranking of actors. Should be replaced by time-of-release rankings.
- Director gross is average over director life, not at time-of-release.
- Several additional factors would likely have been helpful:
  - Total gross of prequel included along with opening weekend gross.
  - Ratings (rotten tomatoes?) of prequel.
  - Are there multiple prequels?
  - Social media sentiment is a high-efficacy real time gauge.
  - Whether movie was released on holiday/ long weekend.

There is significant scope for model improvement

APPENDIX

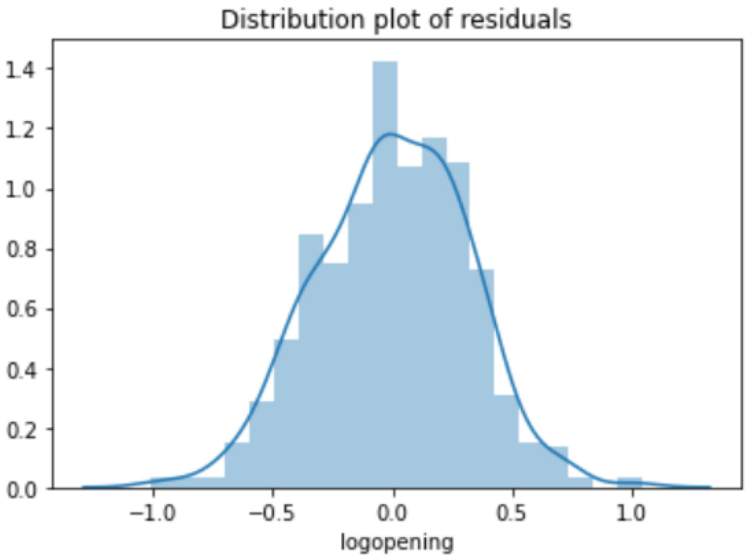
Opening weekend gross vs. prequel opening weekend gross for movies with prequels

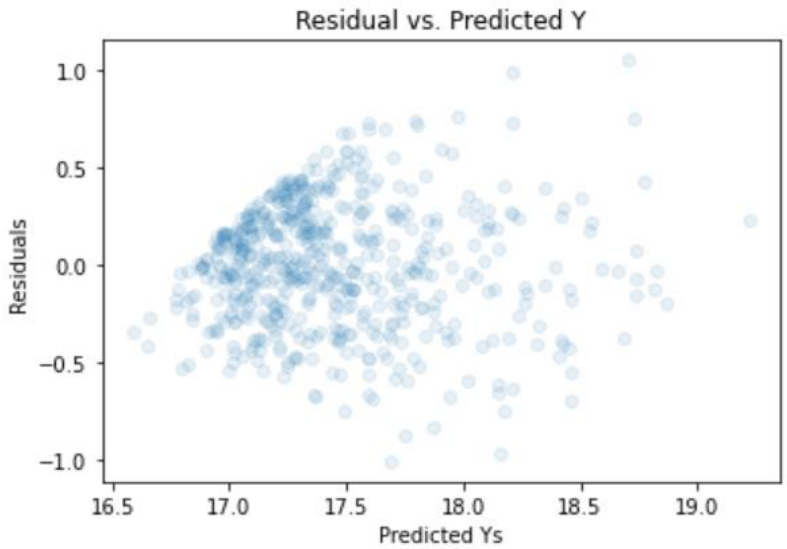


Prequel performance is a strong predictor for movies with prequels

Log model

	Validation R^2	Delta to baseline	Training R2	Test R^2
Log transformed model	59.2%	5.1%	64.9%	63.1%





Residuals are symmetric around 0 but heteroscedasticity is an issue

Continuous feature correlations

	opening_gross	prequel_gross	theaters	budget	starmeter_1	starmeter_2	prequel	DirAveBox	year	distpermov
opening_gross	1.000000	0.582843	0.568775	0.594811	-0.032785	-0.082590	0.323732	0.520300	0.220561	0.210931
prequel_gross	0.582843	1.000000	0.422253	0.449627	-0.021552	-0.056942	0.770942	0.299546	0.226516	0.066192
theaters	0.568775	0.422253	1.000000	0.637480	-0.039332	-0.101104	0.310906	0.401155	0.488012	0.199178
budget	0.594811	0.449627	0.637480	1.000000	-0.077899	-0.097390	0.253351	0.454966	0.164940	0.290459
starmeter_1	-0.032785	-0.021552	-0.039332	-0.077899	1.000000	0.380084	-0.018012	-0.040986	-0.012929	-0.061429
starmeter_2	-0.082590	-0.056942	-0.101104	-0.097390	0.380084	1.000000	-0.045357	-0.084113	-0.008775	-0.095999
prequel	0.323732	0.770942	0.310906	0.253351	-0.018012	-0.045357	1.000000	0.138649	0.186725	-0.063108
DirAveBox	0.520300	0.299546	0.401155	0.454966	-0.040986	-0.084113	0.138649	1.000000	0.122809	0.278169
year	0.220561	0.226516	0.488012	0.164940	-0.012929	-0.008775	0.186725	0.122809	1.000000	-0.013288
distpermov	0.210931	0.066192	0.199178	0.290459	-0.061429	-0.095999	-0.063108	0.278169	-0.013288	1.000000