Predicting Box Office Gross with Linear Regression

...

By Tim Martin

The Data

Scraped from film site, The-Numbers:

- Budget, gross
- Release date
- Cast, production companies
- Runtime, rating, genre
- Movies from 1915-2016
- 1600 released in the US (and possibly also abroad)

The Goal

- Predict international box office gross *before* a film opens

Top 5 Highest Grossing Films

Wikipedia:

1	Avatar
2	Titanic
3	Star Wars: The Force Awakens
4	Jurassic World
5	The Avengers

Timmypedia:

1	Avatar
2	Titanic
3	Star Wars VII: The Force Awakens
4	Jurassic World
5	The Avengers

Top 10 Highest Grossing Films - Inflation Adjusted

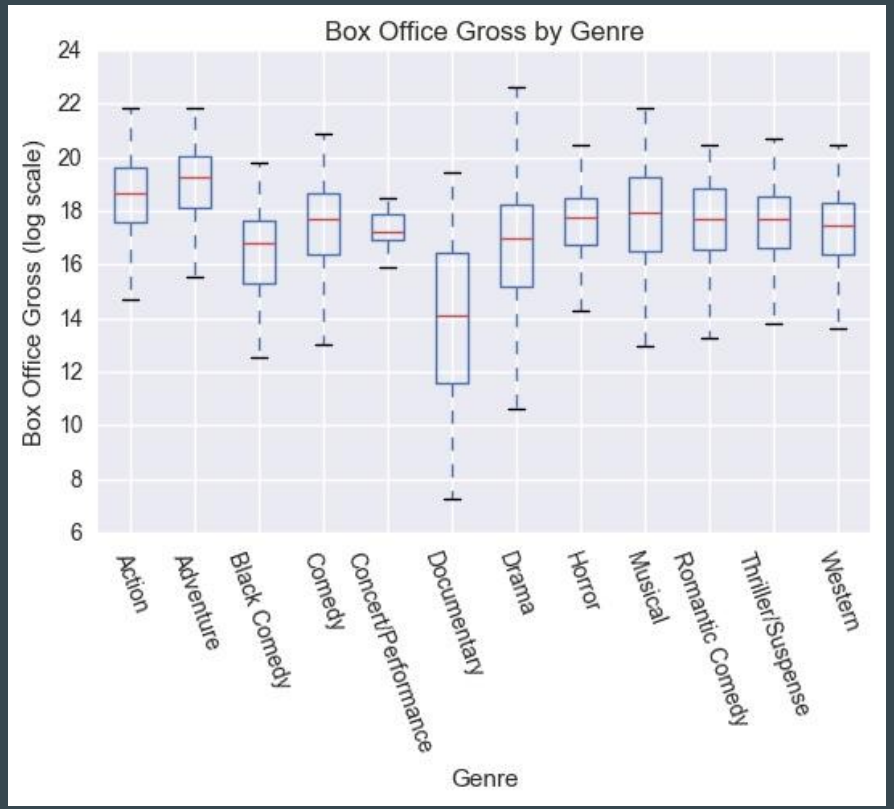
Wikipedia:

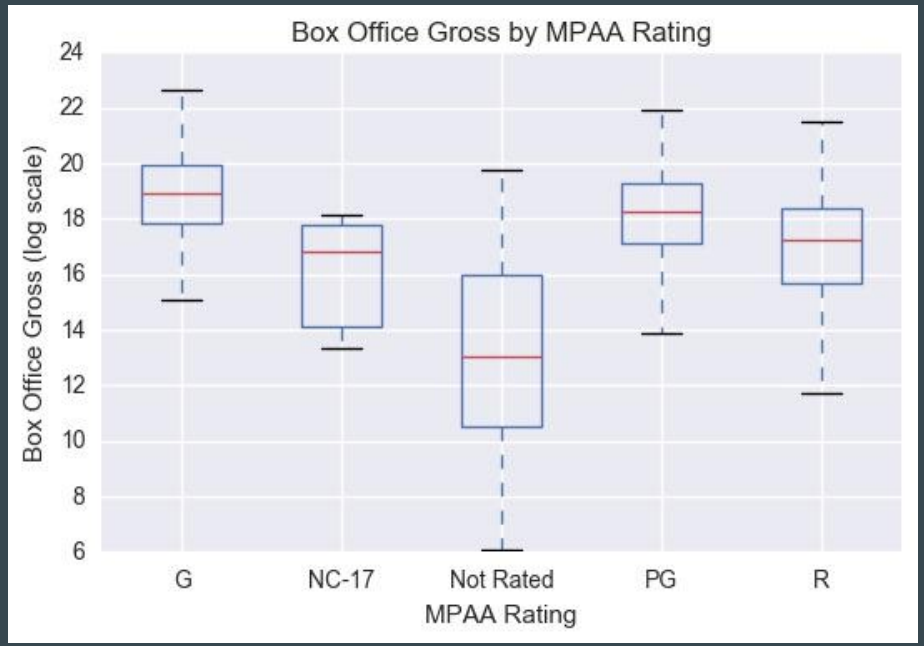
1	Gone with the Wind
2	Avatar
3	Star Wars
4	Titanic
5	The Sound of Music
6	E.T. the Extra-Terrestrial
7	The Ten Commandments
8	Doctor Zhivago
9	Jaws
10	Snow White and the Seven Dwarfs

Timmypedia:

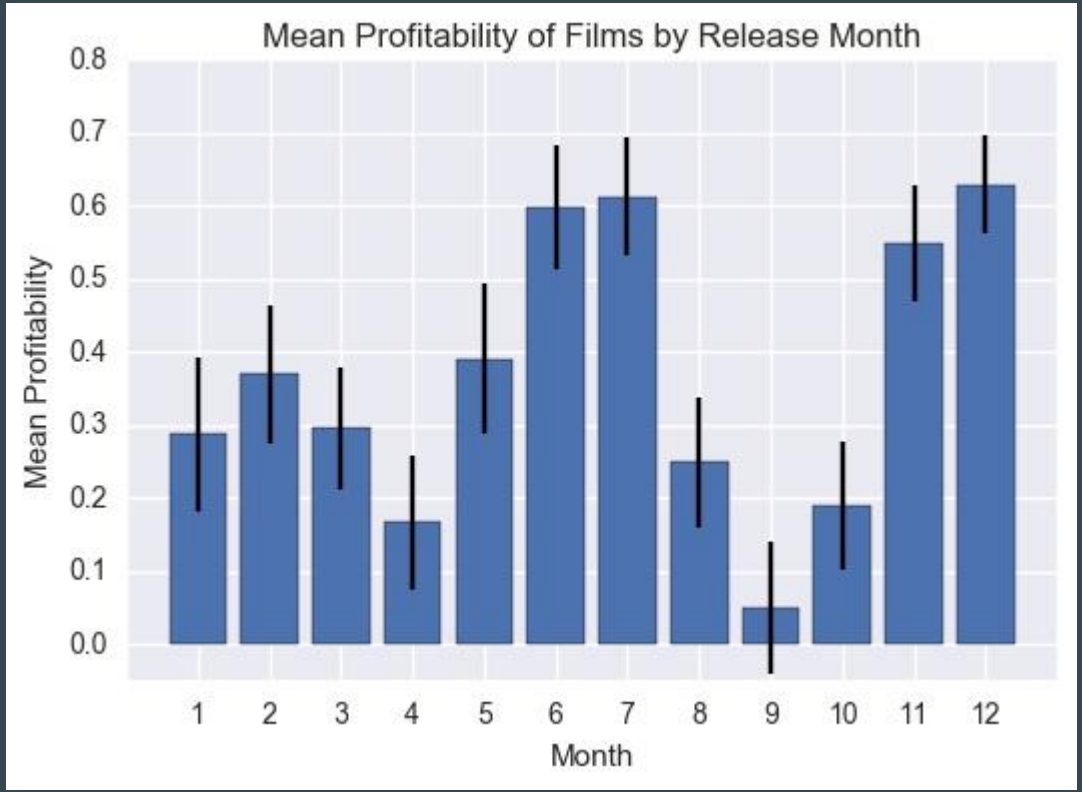
1	Gone with the Wind
2	Bambi
3	Titanic
4	Star Wars Ep. IV: A New Hope
5	Avatar
6	Snow White and the Seven Dwarfs
7	Cinderella
8	The Sound of Music
9	The Exorcist
10	Jaws

Lots of choices...





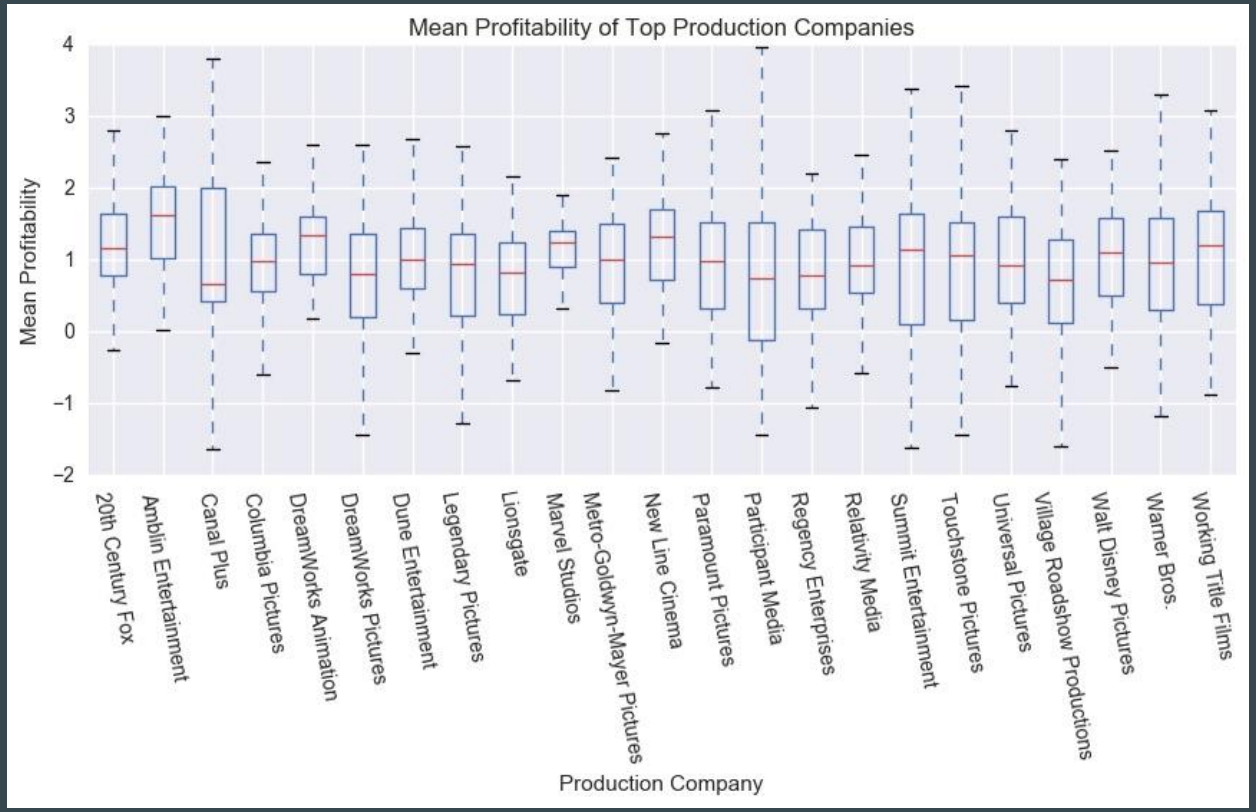
From Summer Blockbusters to the Holiday Rush...



Turning Production Companies into Features

Movie	Production Companies
Avatar	Dune Entertainment, 20th Century Fox, Ingenious Film Partners
Star Wars VII: The Force Awakens	Lucasfilm, Bad Robot
Pirates of the Caribbean: At World's End	Walt Disney Pictures, Jerry Bruckheimer
Spectre	Eon Productions
The Dark Knight Rises	Warner Bros., Legendary Pictures, Syncopy

Turning Production Companies into Features



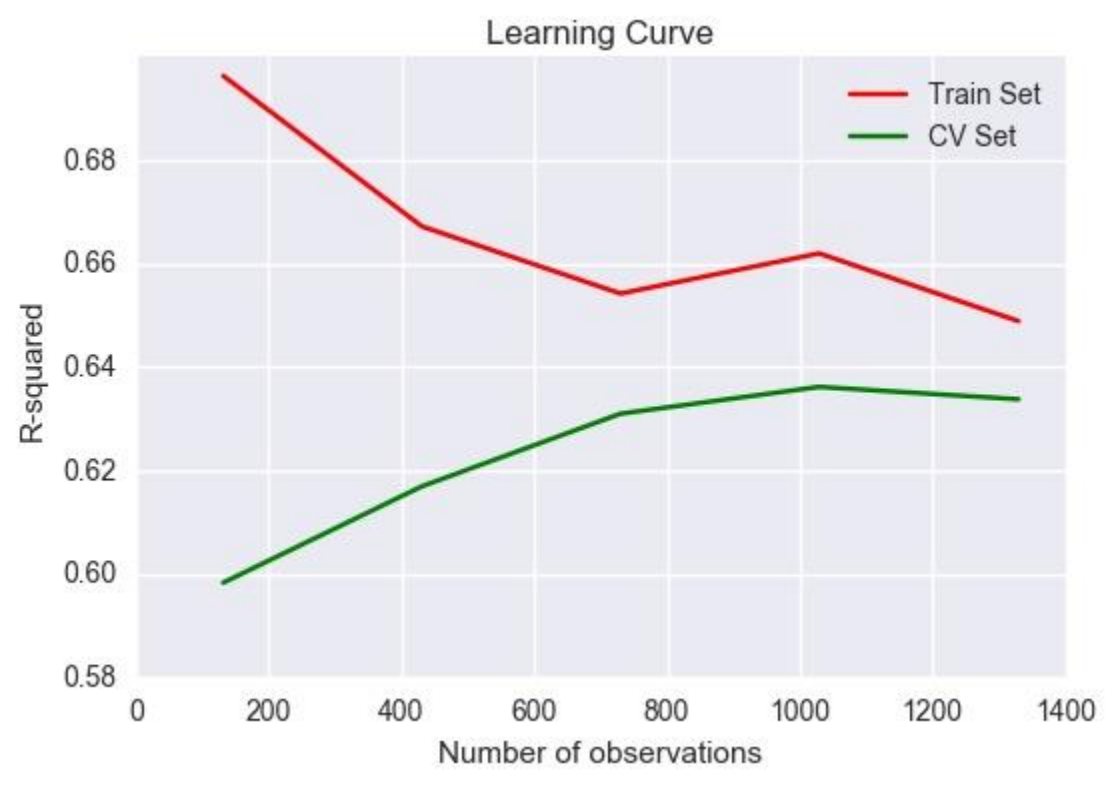
Turning Production Companies into Features

Movie	Production Company	Rank	Historical Profitability
0	20th Century Fox	61	1.19
0	Dune Entertainment	54	1.48
0	Ingenious Film Partners	7	0.42
1	Bad Robot	10	1.21
1	Lucasfilm	12	1.43
2	Jerry Bruckheimer	19	1.2
2	Walt Disney Pictures	69	0.85
3	Eon Productions	2	0.8
4	Legendary Pictures	33	1.1
4	Syncopy	4	.87
4	Warner Bros.	110	1.57

Lasso Model with 3-fold CV

Coefficient	Feature
16.74	Budget
-2.32	Rating - Not Rated
1.95	Production Cast Score
0.8	Cast Size
0.67	Production Company Profitability
0.61	Genre - Horror
0.41	Release Month Profitability
-0.36	Rating - R
0.34	Production Method - Digital Animation
-0.3	Production Method - Live Action

How did my Machine Learn?



Training R-squared:
.647

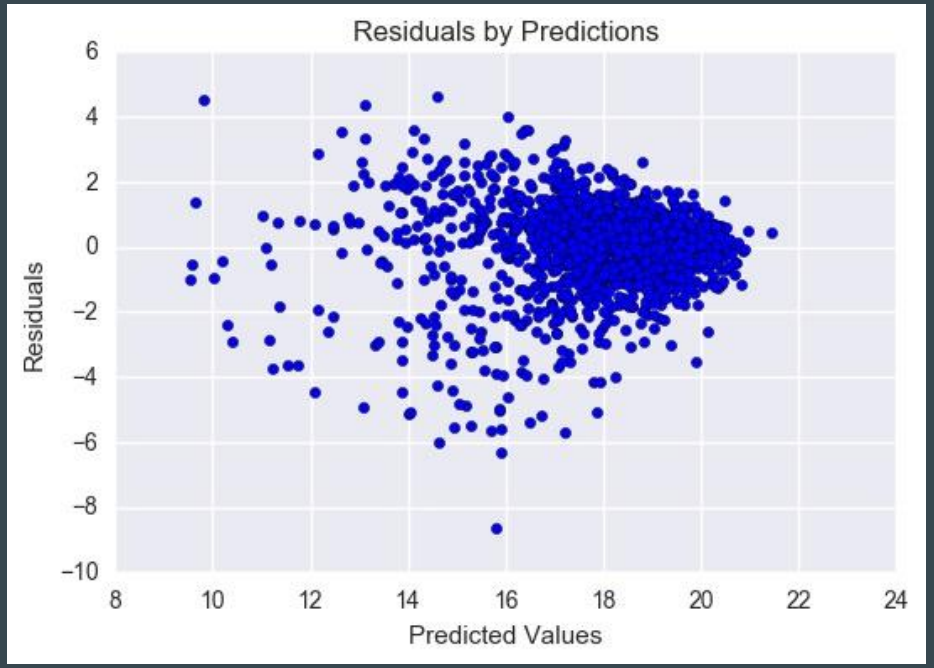
Test R-squared:
.652

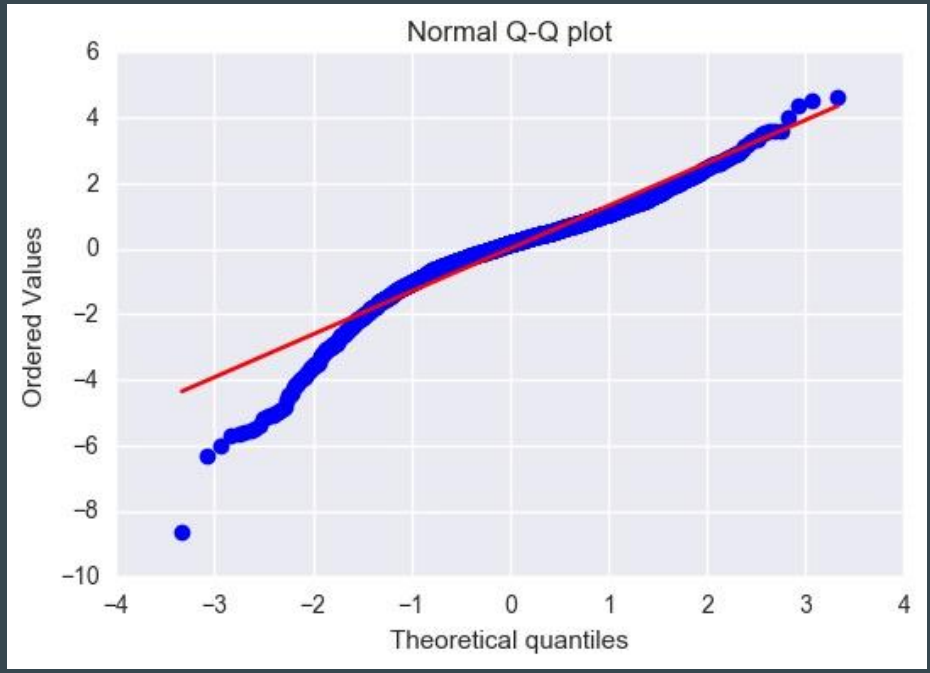
Conclusions

- Budget matters (obviously)
- G, PG, or PG-13 rating
- Top production cast
- Top production company
- Large cast of actors
- Release month
- Horror films!

Thank you!

Appendix


Residuals by Predictions
Residuals
Predicted Values
6
4
2
0
−2
−4
−6
−8
−10
8
10
12
14
16
18
20
22
24


Normal Q-Q plot
Ordered Values
Theoretical quantiles
6
4
2
0
−2
−4
−6
−8
−10
−4
−3
−2
−1
0
1
2
3
4