Predicting International Box Office Revenue

Matt Ranalletta

Goal

Create a model to predict the international box office revenue of a movie, based on known attributes.

- Budget
- Runtime
- Rating

- Release month
- Release year
- Genres

Use cases

1 Production companies

3 Film distributors

2 Studios

4 Theaters

Web Scraping

BoxOfficeMojo.com

Avengers: Endgame (2019)

After the devastating events of Avengers: Infinity War, the universe is in ruins. With the help of remaining allies, the Avengers assemble once more in order to reverse Thanos' actions and restore balance to the universe.

IMDbPro

Cast information
Crew information
Company information
News
Box office
Brand rankings
Franchise rankings
Genre keyword rankings

Title Summary   All Releases
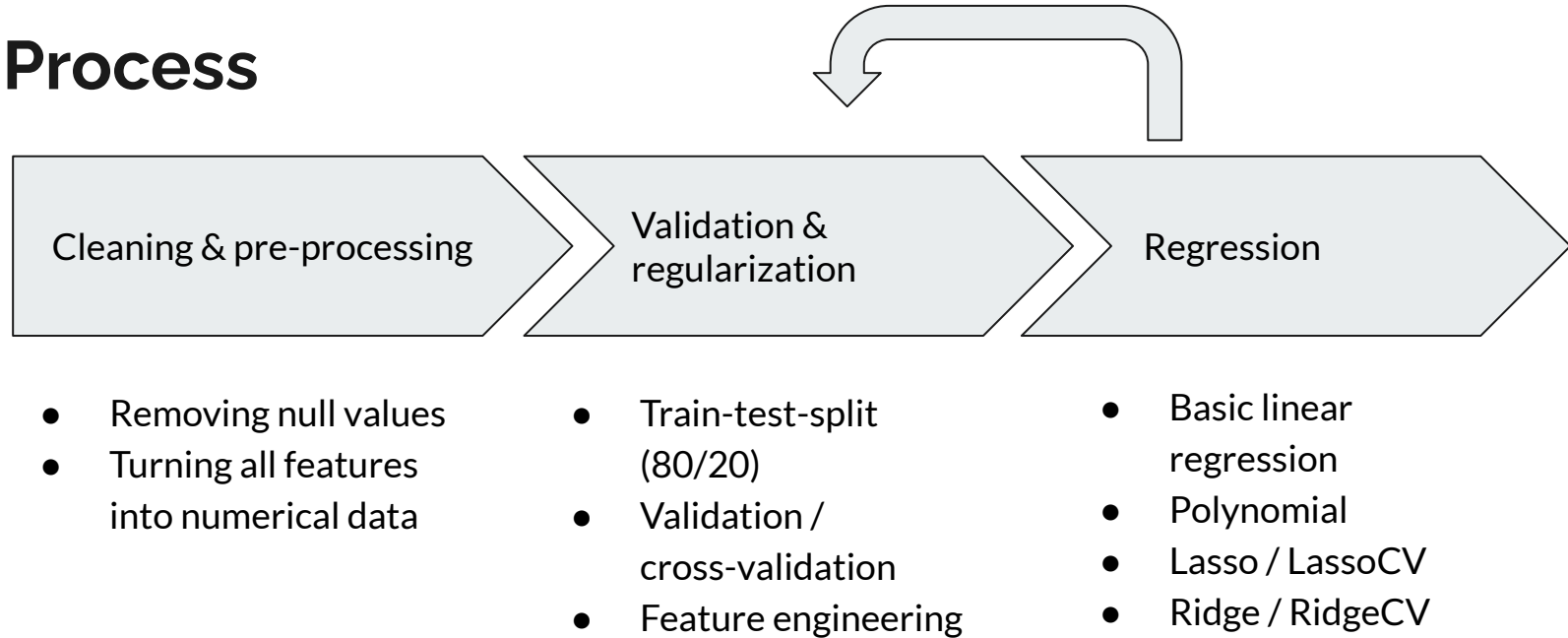
All Releases
DOMESTIC (30.7%)
$858,373,000
INTERNATIONAL (69.3%)
$1,939,427,564
WORLDWIDE
$2,797,800,564

Domestic Distributor	Walt Disney Studios Motion Pictures
	See full company information ↗
Domestic Opening	$357,115,007
Budget	$356,000,000
Earliest Release Date	April 24, 2019 (21 markets)
MPAA	PG-13
Running Time	3 hr 1 min
Genres	Action Adventure Drama Sci-Fi
IMDbPro	See more details at IMDbPro ↗

Top Lifetime Grosses

Worldwide ∨

Data as of Oct 9, 2:35 PDT

← Previous page   1-200 of 1,000   Next page →

Rank	Title	Worldwide Lifetime Gross	Domestic Lifetime Gross	Domestic %	Foreign Lifetime Gross	Foreign %	Year
1	Avengers: Endgame	$2,797,800,564	$858,373,000	30.7%	$1,939,427,564	69.3%	2019
2	Avatar	$2,790,439,092	$760,507,625	27.2%	$2,029,931,467	72.8%	2009
3	Titanic	$2,195,169,869	$659,363,944	30%	$1,535,805,925	70%	1997
4	Star Wars: Episode VII - The Force Awakens	$2,068,224,036	$936,662,225	45.3%	$1,131,561,811	54.7%	2015
5	Avengers: Infinity War	$2,048,359,754	$678,815,482	33.1%	$1,369,544,272	66.9%	2018
6	Jurassic World	$1,670,401,444	$652,270,625	39%	$1,018,130,819	61%	2015
7	The Lion King	$1,656,963,790	$543,638,043	32.8%	$1,113,325,747	67.2%	2019
8	The Avengers	$1,518,815,515	$623,357,910	41%	$895,457,605	59%	2012
9	Furious 7	$1,515,048,151	$353,007,020	23.3%	$1,162,041,131	76.7%	2015
10	Frozen II	$1,450,026,933	$477,373,578	32.9%	$972,653,355	67.1%	2019

Process

Cleaning & pre-processing	Validation & regularization	Regression

- Removing null values
- Turning all features into numerical data

- Train-test-split (80/20)
- Validation / cross-validation
- Feature engineering

- Basic linear regression
- Polynomial
- Lasso / LassoCV
- Ridge / RidgeCV

Features

- Year
- ~~Month~~
- Budget
- Rating (G, PG, ~~PG-13, R~~)
- Runtime

Genres:

- ~~Action~~
- Adventure
- Animation
- ~~Biography~~
- ~~Comedy~~
- Crime
- ~~Documentary~~
- ~~Drama~~
- ~~Family~~
- Fantasy
- History
- ~~Horror~~
- Music
- Musical
- ~~Mystery~~
- ~~Romance~~
- ~~Sci-Fi~~
- Sport
- Thriller
- War
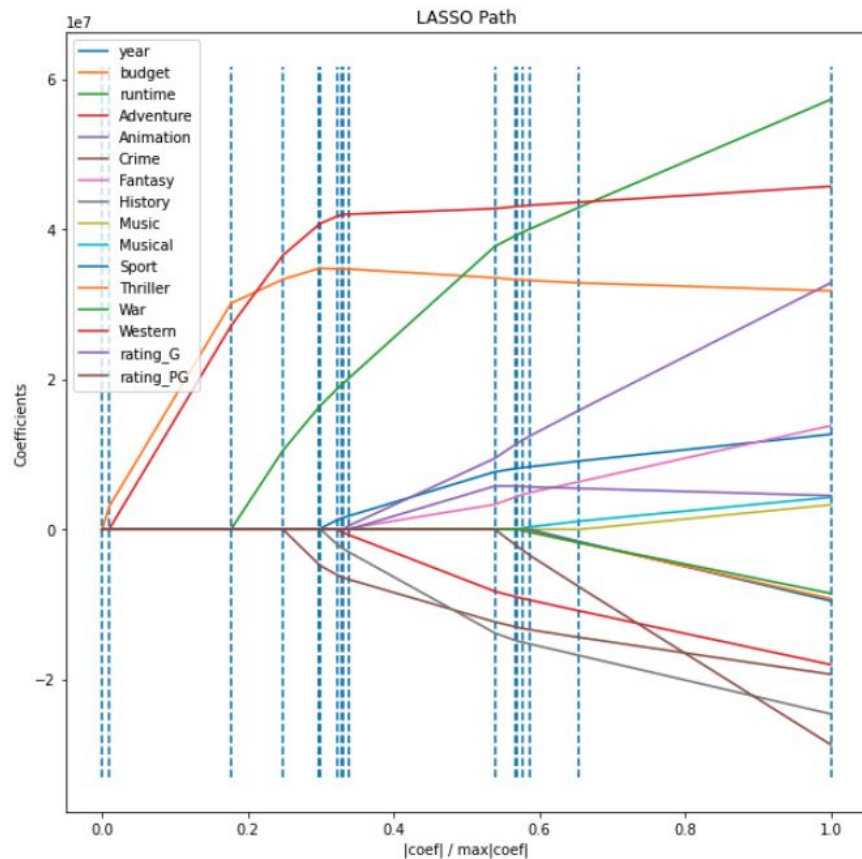- Western

Results

R SQUARED

.32

Represents the proportion of variance for international box office gross explained by the list of features in the model

	index	std	lasso_coeff	original_scale_coeff
0	year	7.59840	11527130.70974	1517047.04149
1	budget	54061604.57866	32136822.42849	0.59445
2	runtime	20.43380	52666297.10130	2577410.97430
3	Adventure	0.49662	45034433.48011	90681807.41809
4	Animation	0.34216	27382269.67810	80028020.34230
5	Crime	0.33260	-17727012.41789	-53298044.39756
6	Fantasy	0.43825	11409810.36302	26034900.64932
7	History	0.18726	-22079735.04599	-117907023.70267
8	Music	0.15535	2225483.41714	14325547.51359
9	Musical	0.18726	3255940.33505	17386904.03034
10	Sport	0.13101	-6974372.98379	-53237025.96386
11	Thriller	0.45960	-6819128.49383	-14837105.18238
12	War	0.19432	-6349894.77160	-32678077.93408
13	Western	0.11371	-15709756.51346	-138161137.35361
14	rating_G	0.11977	4813753.94973	40192186.10633
15	rating_PG	0.41712	-21901198.37373	-52506065.69534

Lasso Path

Most influential variables:

- Budget
- Genre: Adventure
- Runtime

Further Exploration

- Much larger dataset

- More features from other sources

- U.S. domestic box office

Thank you!