Domestic Box Office Gross:

Predicting the best of the best

Nicole McBride
Ishikawa Games, LLC

Why we're here

Background: ***Ishikawa Games, LLC*** wants to predict top domestic grosser movies so they can collaborate with the distributors to produce video games related to the movie.

Goal: Produce a regression model that can best predict domestic gross revenue based on varying movie related features.

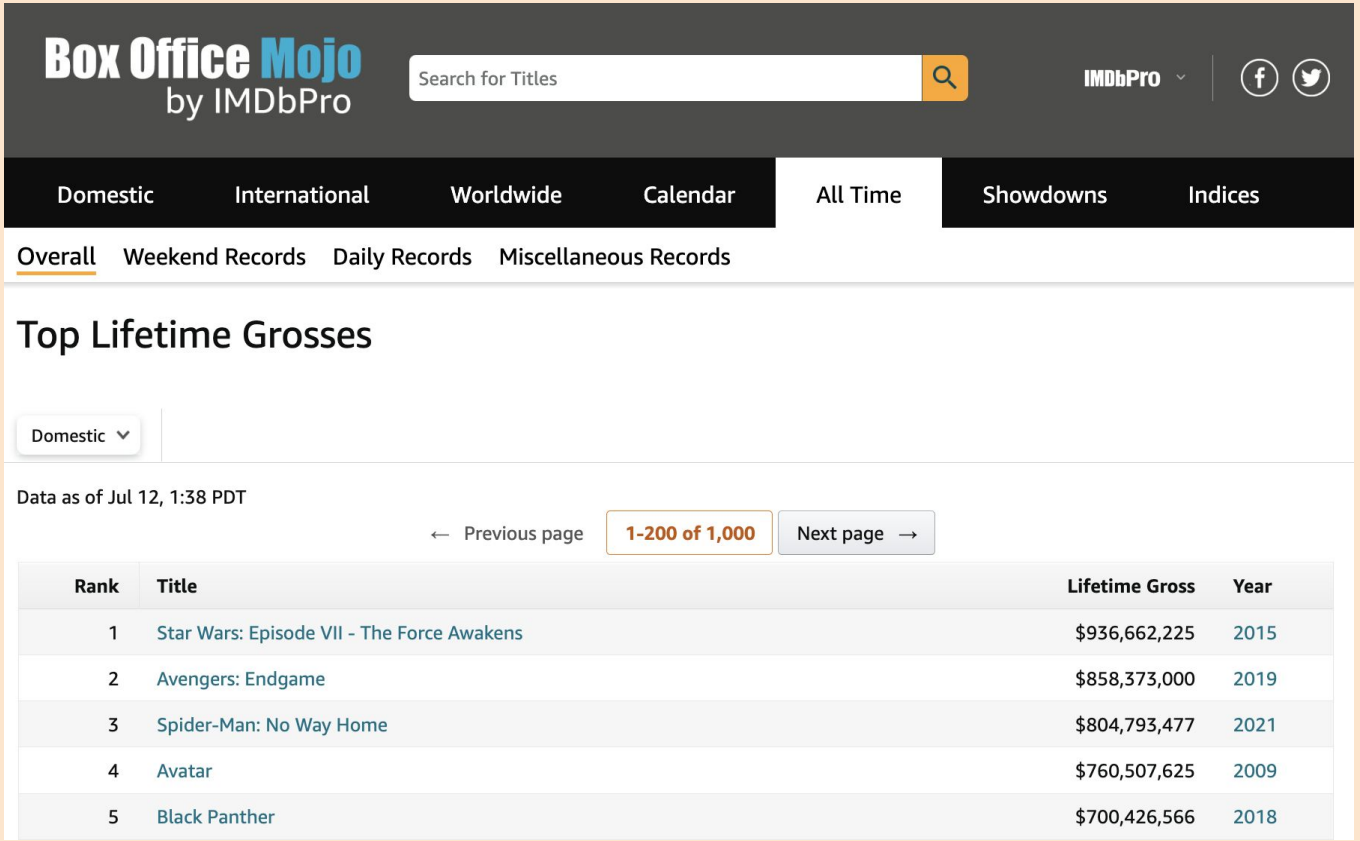
Design

The top 1000 highest domestic grossed movies was scraped on 5 June 2022 from Box Office Mojo

Each row represents a movie and its scraped features

Tools Used:

- BeautifulSoup
- Numpy
- Pandas
- Scikit-learn
- Statsmodels
- Matplotlib
- Seaborn



Rank	Title	Lifetime Gross	Year
1	Star Wars: Episode VII - The Force Awakens	$936,662,225	2015
2	Avengers: Endgame	$858,373,000	2019
3	Spider-Man: No Way Home	$804,793,477	2021
4	Avatar	$760,507,625	2009
5	Black Panther	$700,426,566	2018

Star Wars: Episode VII - The Force Awakens (2015)

As a new threat to the galaxy rises, Rey, a desert scavenger, and Finn, an ex-stormtrooper, must join Han Solo and Chewbacca to search for the one hope of restoring peace.

Independent Variables

Title Summary | All Releases

All Releases

DOMESTIC (45.3%)
$936,662,225
INTERNATIONAL (54.7%)
$1,132,859,475
WORLDWIDE
$2,069,521,700

Domestic Distributor	Walt Disney Studios Motion Pictures See full company information
[redacted]	
Budget	$245,000,000
Earliest Release Date	December 16, 2015 (EMEA, APAC)
MPAA	PG-13
Running Time	2 hr 18 min
Genres	Action Adventure Sci-Fi
IMDbPro	See more details at IMDbPro

Dependent Variable

Performance | Cast and Crew | All-Time Rankings | Related Stories | Similar Movies

Data Cleaning & Feature Engineering

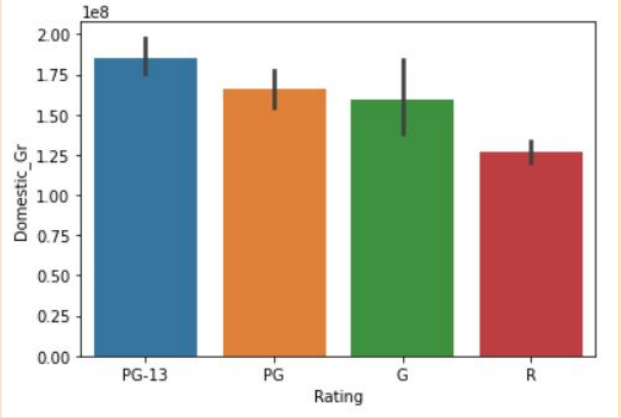
Missing values

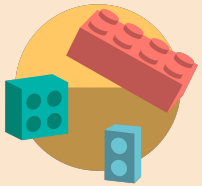
Budget → explore median and mean imputation

Distributor → manually fill n=1

Rating → manually fill n=124

Main Cast & Director → manually fill n=5



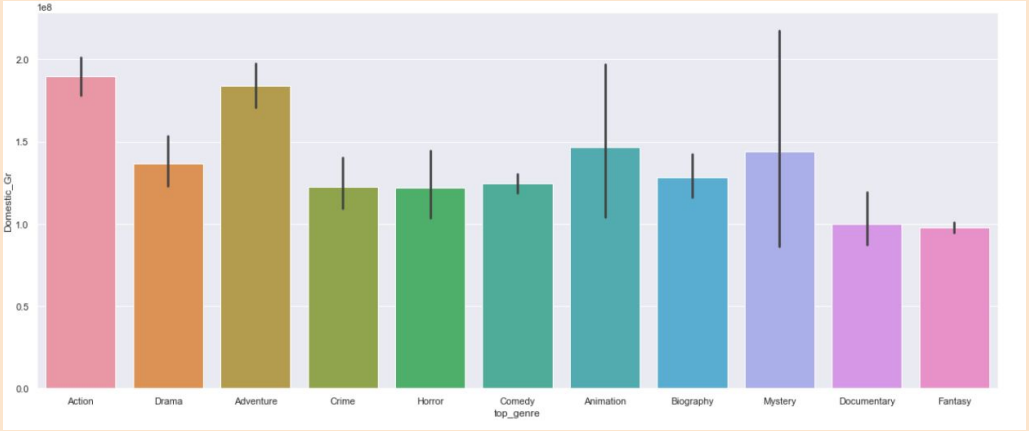


Dummy Code

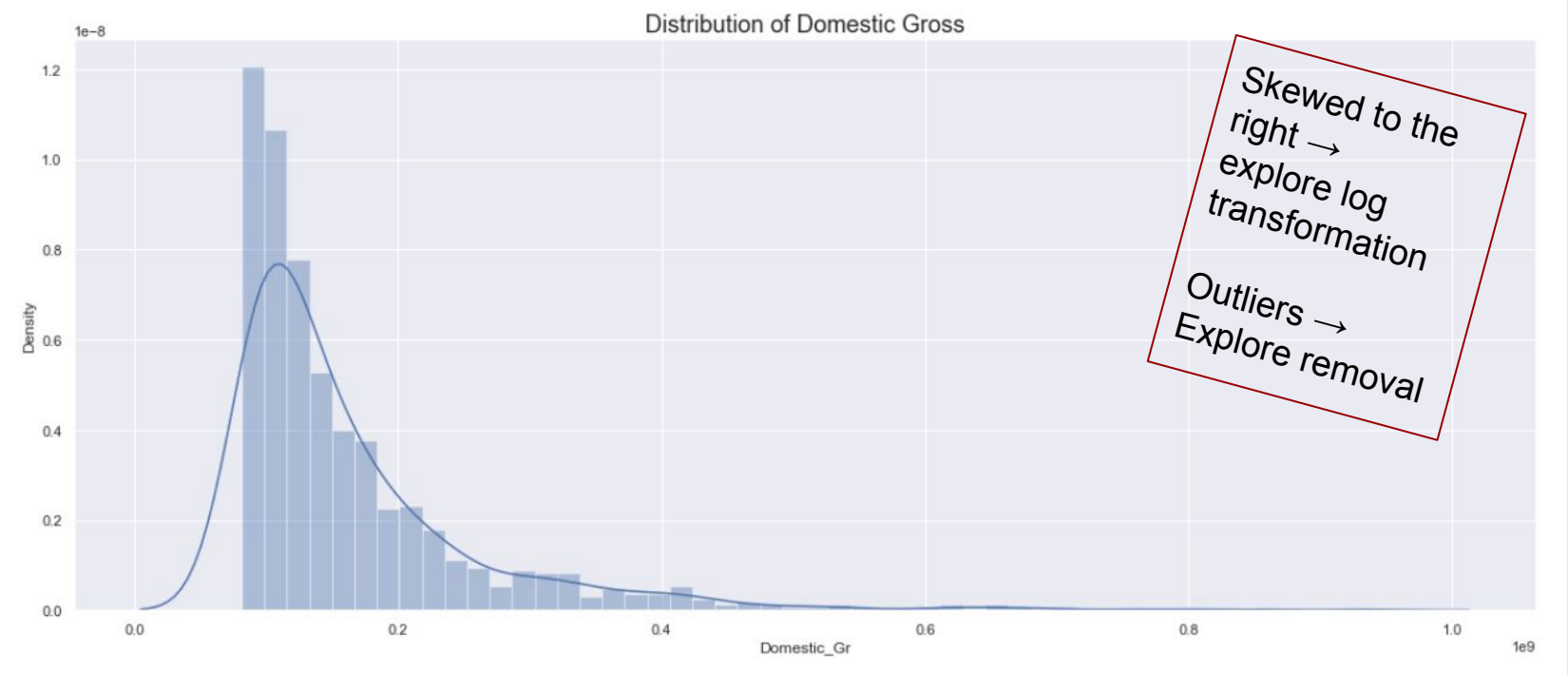
Distributor, Rating, Genre

Created

Age of movie since release
Calendar year quarter
Total runtime in minutes

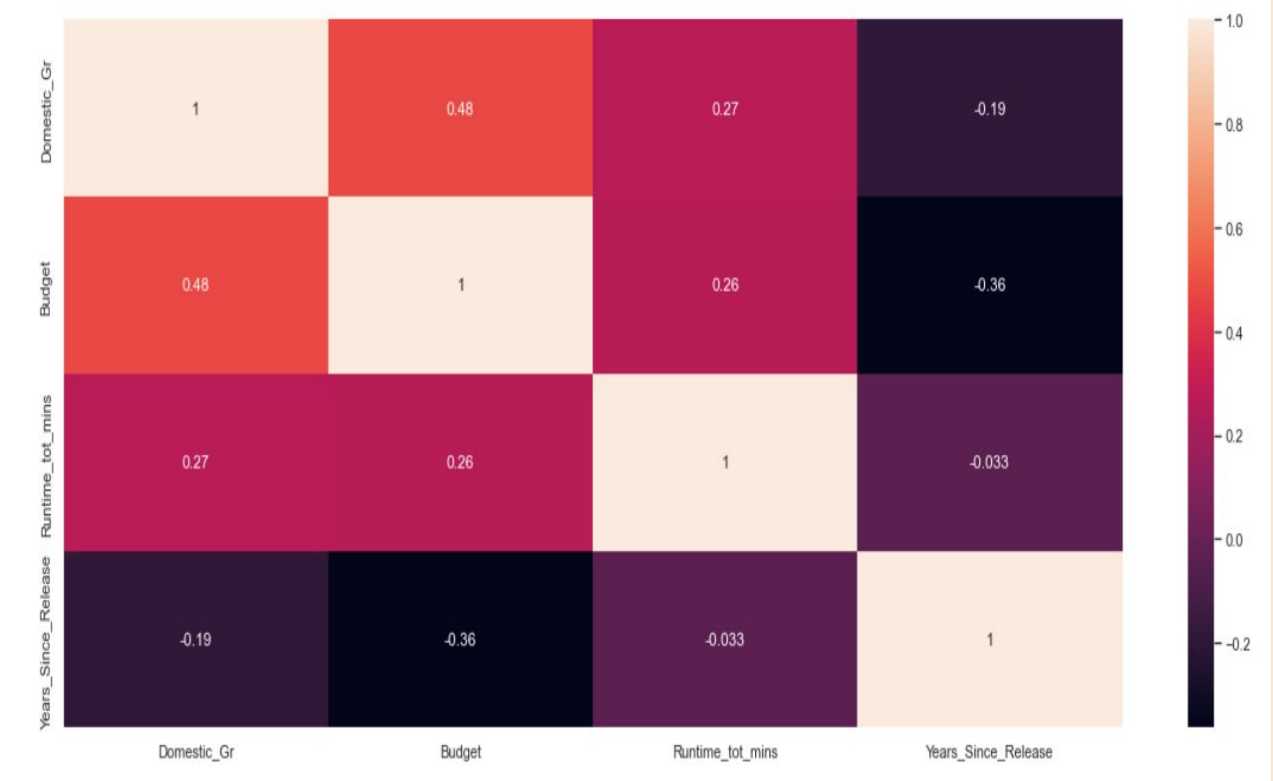


Exploratory Data Analysis



Exploratory Data Analysis

Budget seems most promising!



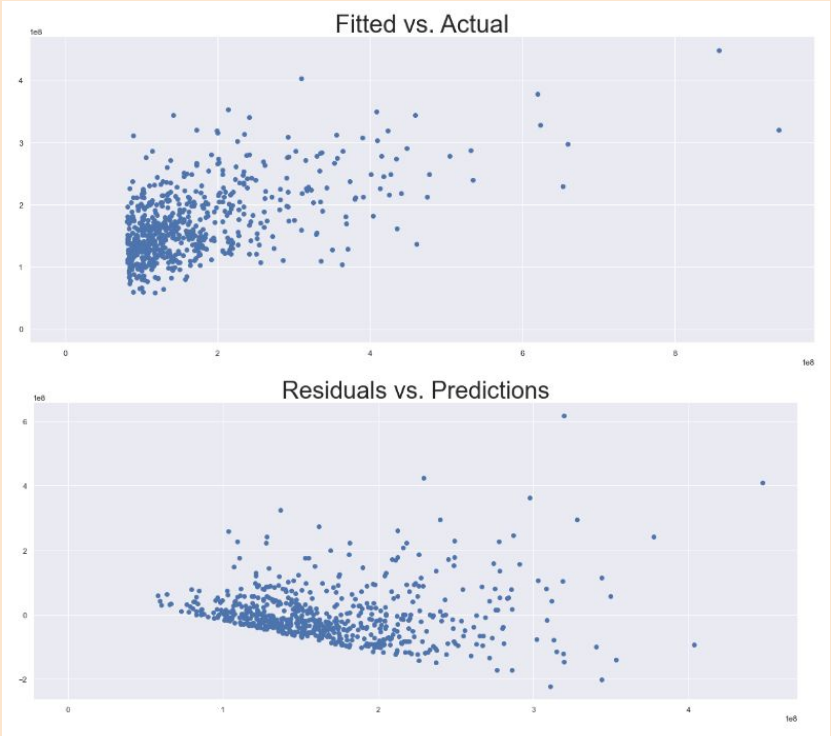
Baseline Model

Independent Variables (IV)	Dependent Variable (DV)
Budget	Domestic Gross
Total Runtime	
Age of movie since release	
Calendar year quarter*	
Total runtime in minutes*	
Genre*	
Distributor*	

*categorical

Kfold = 5 Cross Validation

Mean R^2 = .21 +/- 0.09

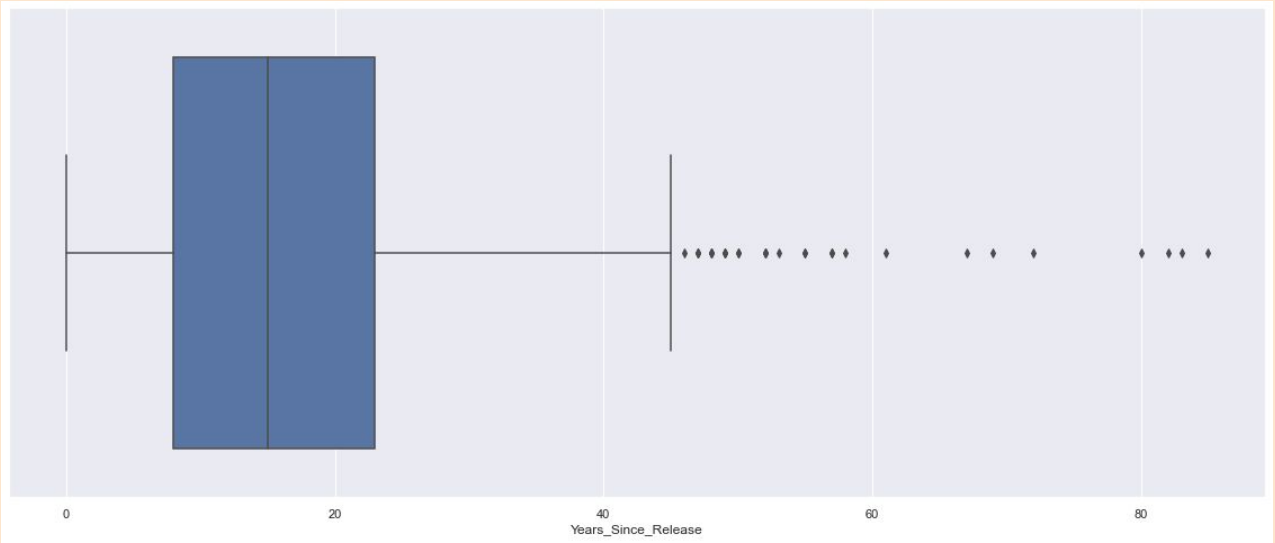


Outliers removed, new model

Independent Variables
Budget
Total Runtime
Age of movie since release
Calendar year quarter*
Total runtime in minutes*
Genre*
Distributor*

Dropping 13 outliers in **Age of movie since release** **improved** model, but dropping others didn't

Mean R^2 = .27 +/- 0.06



Merge Oscar data

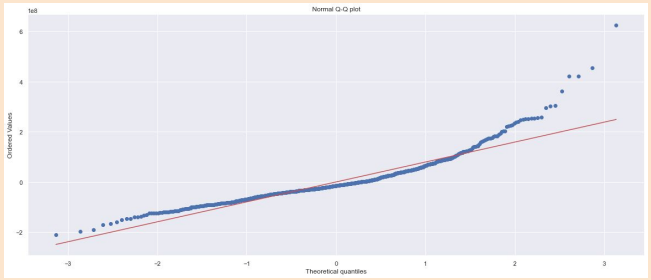
Independent Variables	Dependent Variable
Budget	Domestic Gross
Total Runtime	
Age of movie since release	
Calendar year quarter*	
Total runtime in minutes*	
Genre*	
Distributor*	
Actor/Actress Wins+Noms	
Director Wins+Noms	

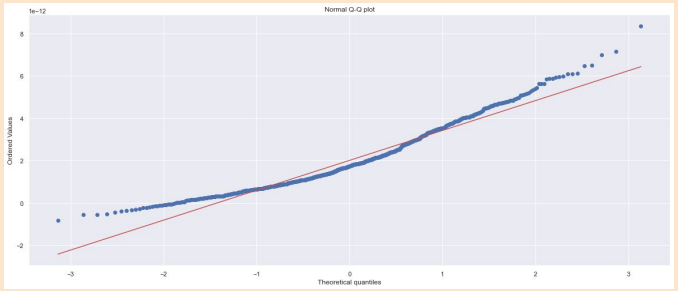
Oscar wins or nominations up until the year of movie release for Film Director (Lead) & Actor/Actress (Lead)

No change in model performance

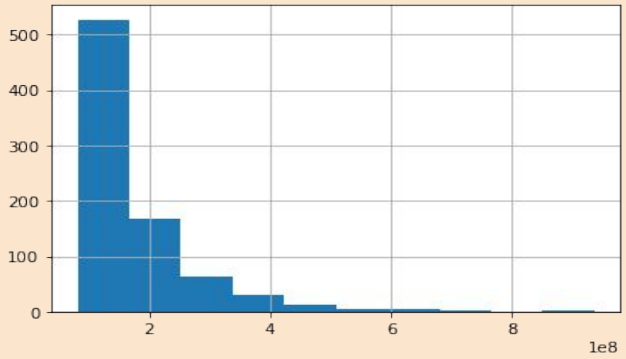
Mean R^2 = .26 +/- 0.06

Applying log transformation on DV

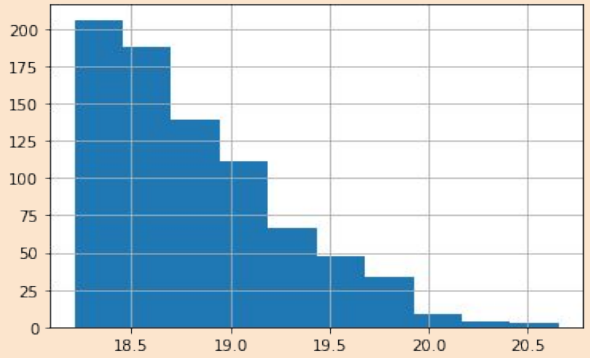




BEFORE



AFTER



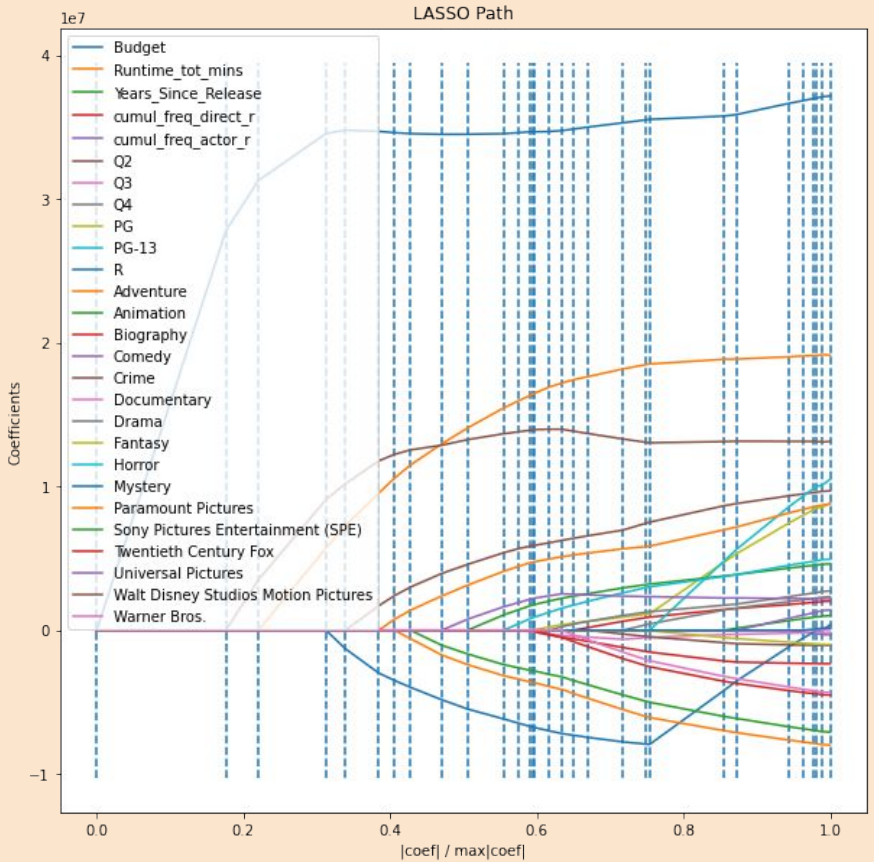
No change in model performance

Mean R^2 = .26 +/- 0.05

Feature Selection

Variance Inflation Factor

	variables	vif
0	const	35.979463
1	Budget	1.302339
2	Runtime_tot_mins	1.243191
3	Years_Since_Release	1.174818
4	cumul_freq_direct_r	1.169888
5	cumul_freq_actor_r	1.090906

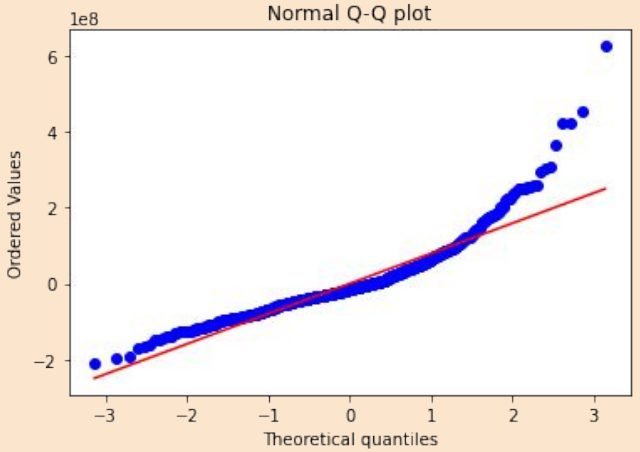


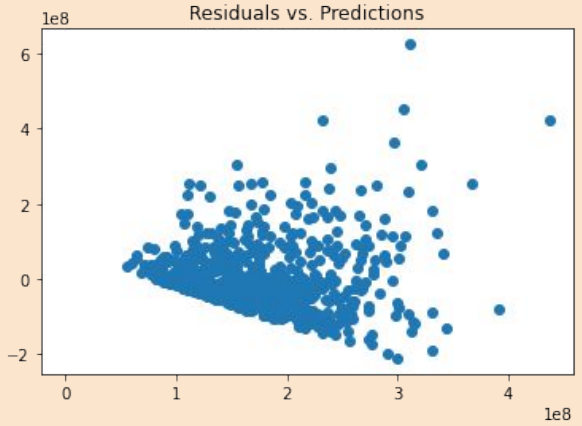
	coef	std err	t	P>|t|
const	-1.62e+07	2.77e+07	-0.584	0.559
Budget	0.5772	0.075	7.689	0.000
Runtime_tot_mins	1.006e+06	1.87e+05	5.386	0.000
Years_Since_Release	8.728e+04	3.83e+05	0.228	0.820
cumul_freq_direct_r	5.938e+06	4.36e+06	1.361	0.174
cumul_freq_actor_r	-5.661e+05	2.17e+06	-0.261	0.794
Q2	2.164e+07	9.9e+06	2.186	0.029
Q3	-2.777e+06	1.03e+07	-0.271	0.787
Q4	3.409e+06	9.98e+06	0.341	0.733
PG	1.406e+07	1.71e+07	0.824	0.410
PG-13	1.334e+07	1.91e+07	0.699	0.485
R	-8.538e+06	2.01e+07	-0.426	0.671
Adventure	1.562e+07	1.08e+07	1.444	0.149
Animation	2.869e+07	3.99e+07	0.719	0.473
Biography	-2.492e+07	1.79e+07	-1.392	0.164
Comedy	6.277e+05	9.8e+06	0.064	0.949
Crime	-1.315e+07	1.87e+07	-0.705	0.481
Documentary	-5.515e+06	8.69e+07	-0.063	0.949
Drama	6.152e+05	1.35e+07	0.045	0.964
Fantasy	-3.944e+07	5.06e+07	-0.780	0.436
Horror	3.34e+07	1.97e+07	1.691	0.091
Mystery	-8.734e+06	6.15e+07	-0.142	0.887
Paramount Pictures	-2.276e+07	1.18e+07	-1.926	0.054
Sony Pictures Entertainment (SPE)	-1.785e+07	1.16e+07	-1.544	0.123
Twentieth Century Fox	-7.439e+06	1.1e+07	-0.673	0.501
Universal Pictures	7.098e+05	1.08e+07	0.066	0.948
Walt Disney Studios Motion Pictures	3.01e+07	1.2e+07	2.507	0.012
Warner Bros.	-1.549e+07	1.05e+07	-1.473	0.141

Final Model

Independent Variables
Budget
Total Runtime
~~Age of movie since release~~
Calendar year quarter*
Total runtime in minutes*
Genre*
Distributor*
Actor/Actress Wins+Noms
Director Wins+Noms

	coef	P>|t|
const	-1.471e+07	0.585
Budget	0.5703	0.000
Runtime_tot_mins	1.016e+06	0.000
cumul_freq_direct_r	6.002e+06	0.168
cumul_freq_actor_r	-5.776e+05	0.790
Q2	2.2e+07	0.025
Q3	-2.608e+06	0.799
Q4	3.588e+06	0.718
PG	1.369e+07	0.420
PG-13	1.258e+07	0.503
R	-9.426e+06	0.632
Adventure	1.533e+07	0.154
Animation	2.823e+07	0.479
Biography	-2.564e+07	0.146
Comedy	5.722e+05	0.953
Crime	-1.321e+07	0.479
Documentary	-6.066e+06	0.944
Drama	3.619e+05	0.979
Fantasy	-3.891e+07	0.441
Horror	3.307e+07	0.093
Mystery	-8.961e+06	0.884
Paramount Pictures	-2.275e+07	0.055
Sony Pictures Entertainment (SPE)	-1.807e+07	0.117
Twentieth Century Fox	-7.559e+06	0.493
Universal Pictures	4.989e+05	0.963
Walt Disney Studios Motion Pictures	3.005e+07	0.012
Warner Bros.	-1.567e+07	0.135





$R^2 = 0.28$

Future Work

Scrape different data. E.g., past revenue of all domestic movies in last 10 years

Re-engineering on Distributor and Genre.

Import additional features such as: expansion of OSCAR/star power, movie sequel/triology.

Explore additional metrics to measure model performance.

Better performance model based transformations.