Regression Analysis of Domestic Box Office Gross Profits

(2010-2016)

Chris Buie

Objective: Predict Box Office Success Measured Through Gross Profits.

Box office Movies are very expensive to make. Being able to accurately predict the drivers of a movies success would help make investment decisions.

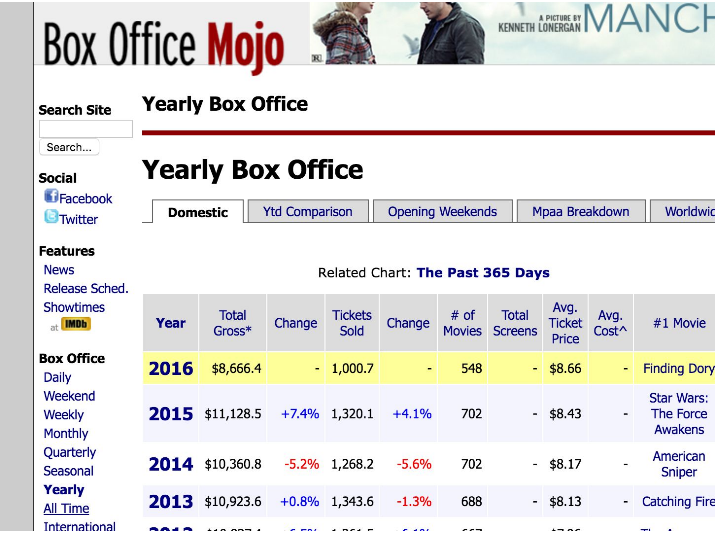
My goal is to create a regression model that will predict gross profits utilizing data obtained via web scraping and hopefully expose some of these drivers .

I will be using three regression models in this analysis, Linear Regression, Random Forest Regression, and Gradient Boosted Regression.

Data Assimilation

The data obtained for the analysis was scraped from two different sites, BoxofficeMojo.com and The-Numbers.com.

This data was scrubbed and merged for feature creation.

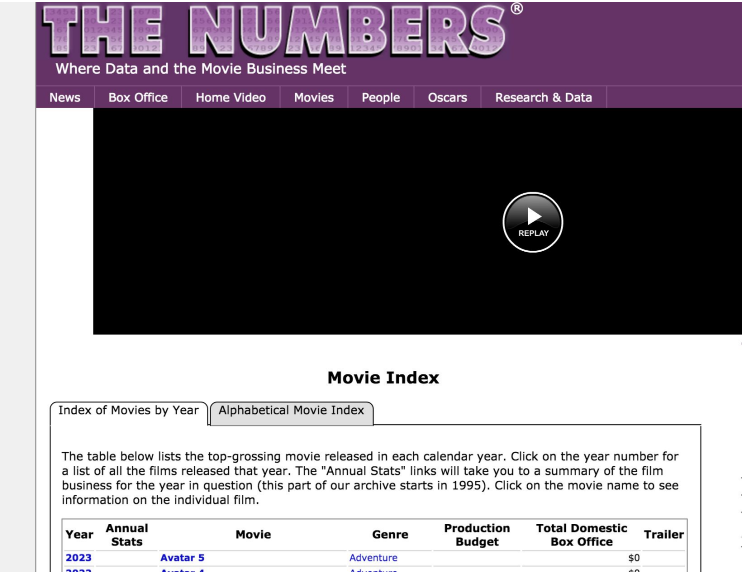


Yearly Box Office

Yearly Box Office

Related Chart: **The Past 365 Days**

Year	Total Gross*	Change	Tickets Sold	Change	# of Movies	Total Screens	Avg. Ticket Price	Avg. Cost^	#1 Movie
2016	$8,666.4	-	1,000.7	-	548	-	$8.66	-	Finding Dory
2015	$11,128.5	+7.4%	1,320.1	+4.1%	702	-	$8.43	-	Star Wars: The Force Awakens
2014	$10,360.8	-5.2%	1,268.2	-5.6%	702	-	$8.17	-	American Sniper
2013	$10,923.6	+0.8%	1,343.6	-1.3%	688	-	$8.13	-	Catching Fire



Movie Index

The table below lists the top-grossing movie released in each calendar year. Click on the year number for a list of all the films released that year. The "Annual Stats" links will take you to a summary of the film business for the year in question (this part of our archive starts in 1995). Click on the movie name to see information on the individual film.

Year	Annual Stats	Movie	Genre	Production Budget	Total Domestic Box Office	Trailer
2023		Avatar 5	Adventure		$0	

Initial Feature Selection:

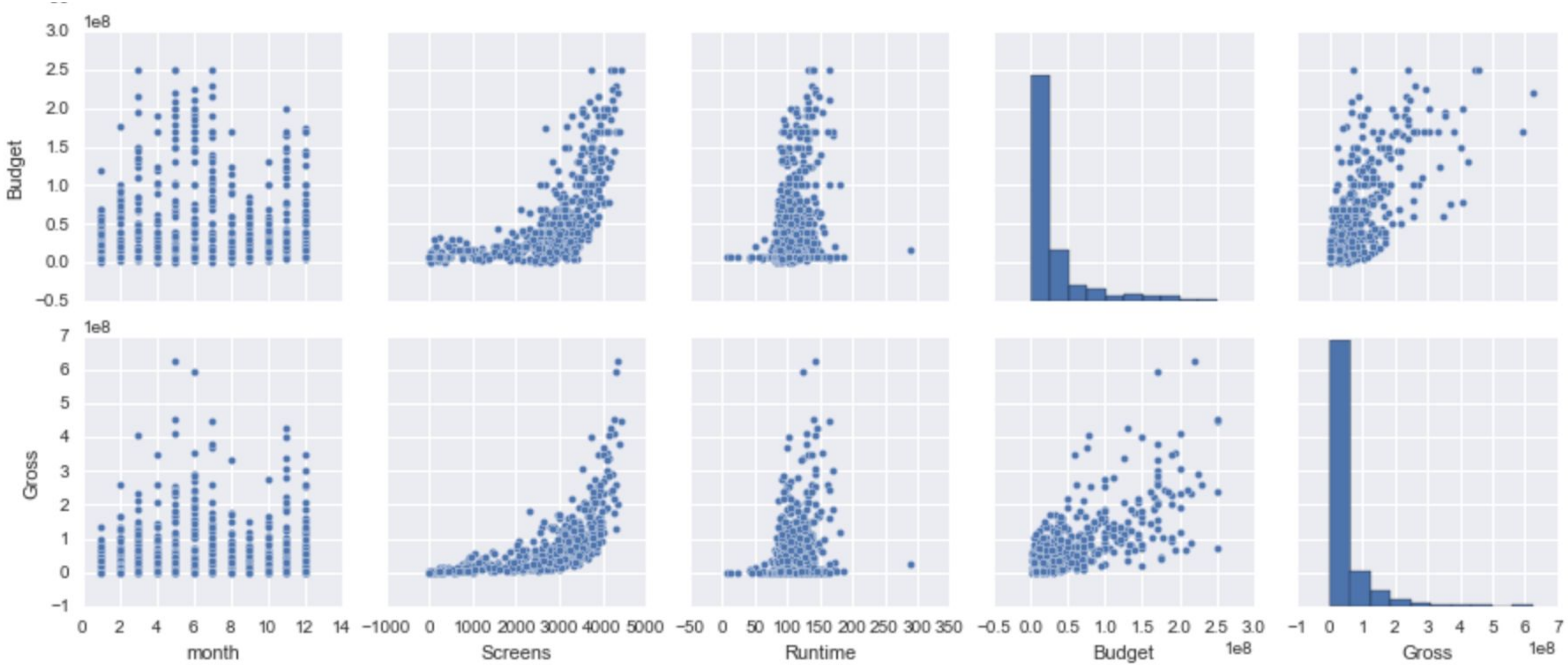
This analysis will look at the following 8 inherent features. Additionally, feature engineering will be attempted.

Budget	Screens
Director	Studio
Distributor	Genre
MPAA Rating	Release Date
Ticket Sales	

Data Exploration:

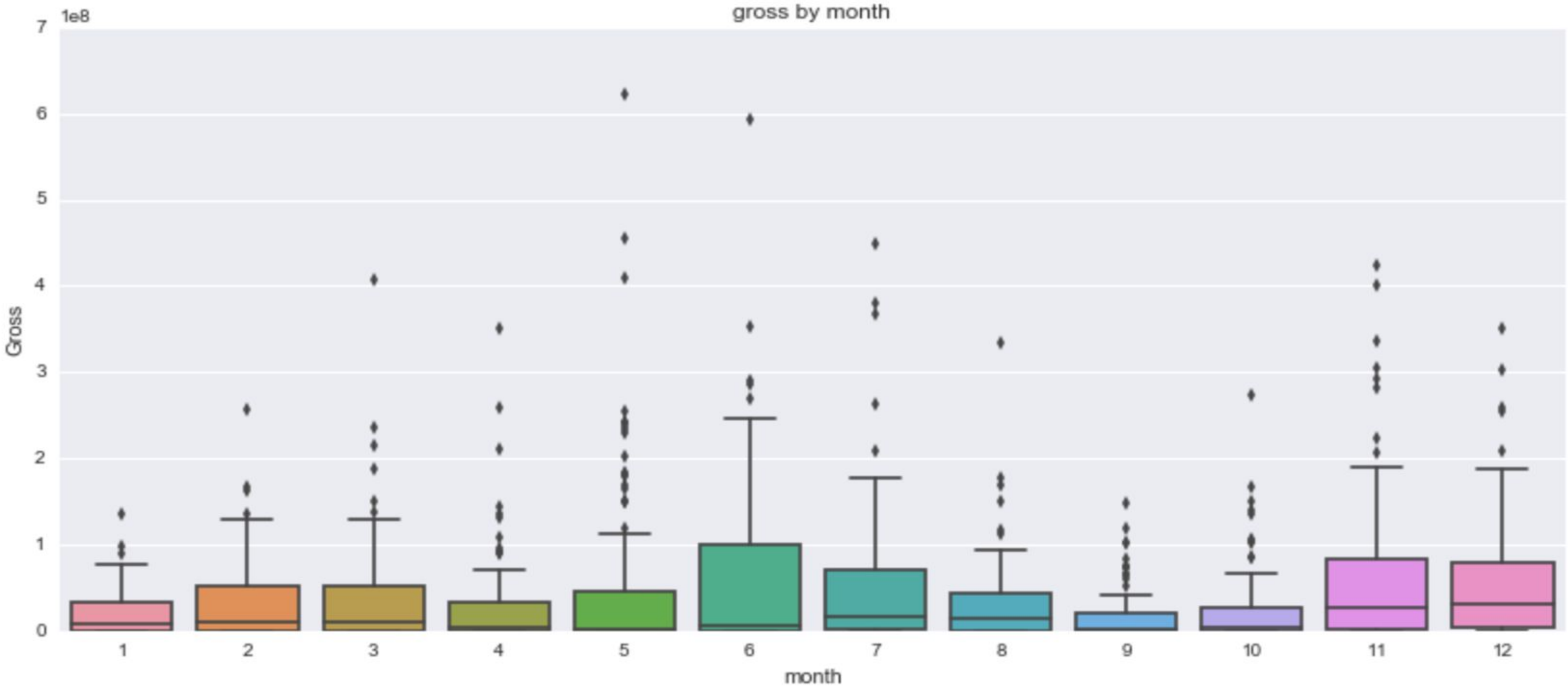
Based on pairs plot on Budget and Tickets Sold were dropped from the analysis due to collinearity with Screens.

Screens will need to be raised to higher powers to increase linear relationship



Data Exploration:

Seasonality features were created to account for holidays, summer, winter, and fall months.



Feature Selection:

This initial analysis will look at the following 11 inherent features.

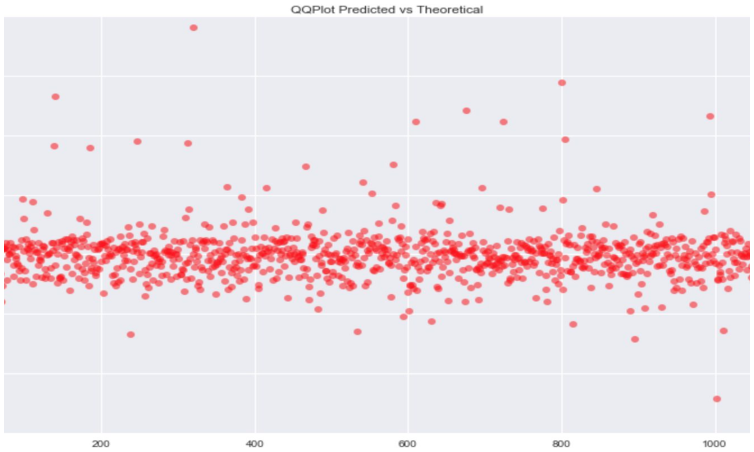
❌	Screens
Director	Studio
Distributor	Genre
MPAA Rating	Release Date
Holiday	Summer
Fall	Winter

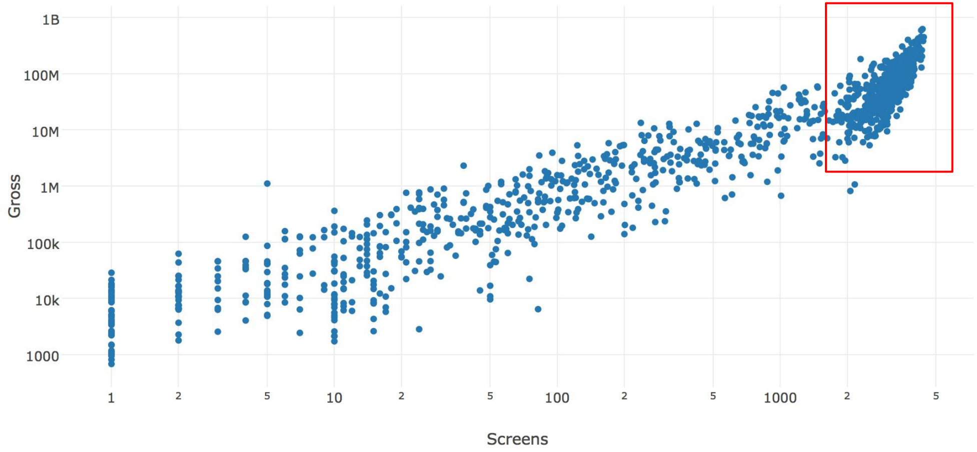
Linear Regression Results

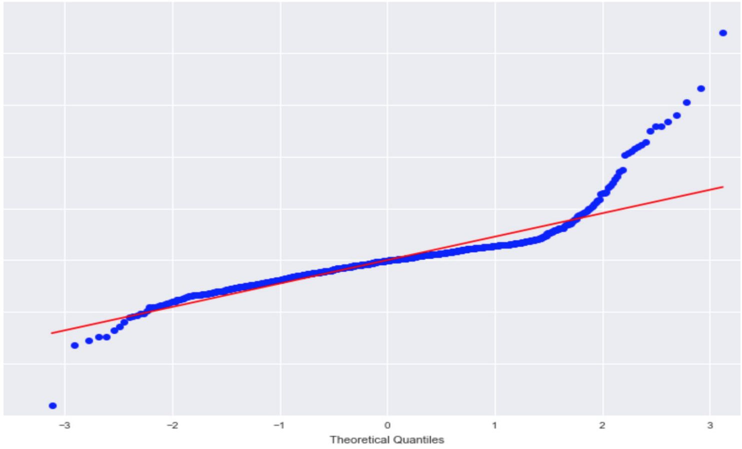
Results from three iterations.

LR with 10K CV	Model 1	Model 2	Model 3
# Observations	1103	1103	1103
Df Model	1013	197	21
Rsqd.	0.968	673	0.621
Ad. Rsqd.	0.498	0.589	0.611

Diagnostic Plots from Linear Regression Model







Normalized Score

Results from three Models

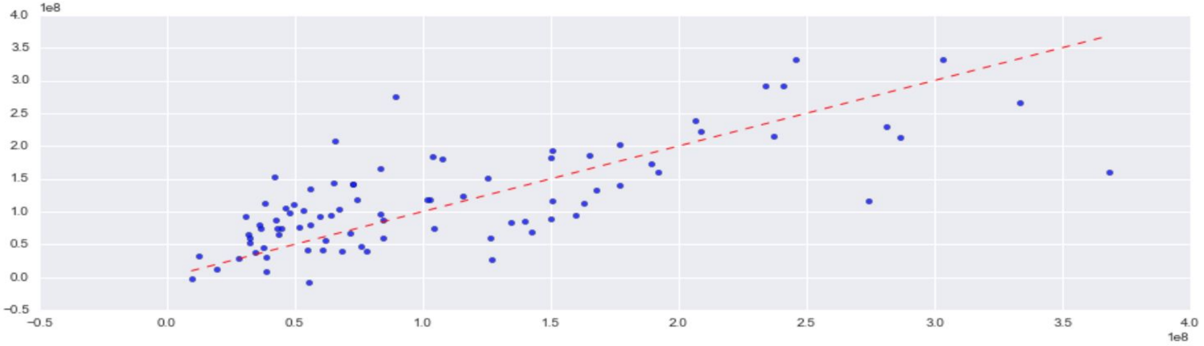
Model 1	Score	RMSE	
Linear Regression	0.69	109374829.4	
RandomForestReg.	0.81	30857763.7	
GradientBoostReg.	0.76	27603905.1	

Predicted vs Actuals

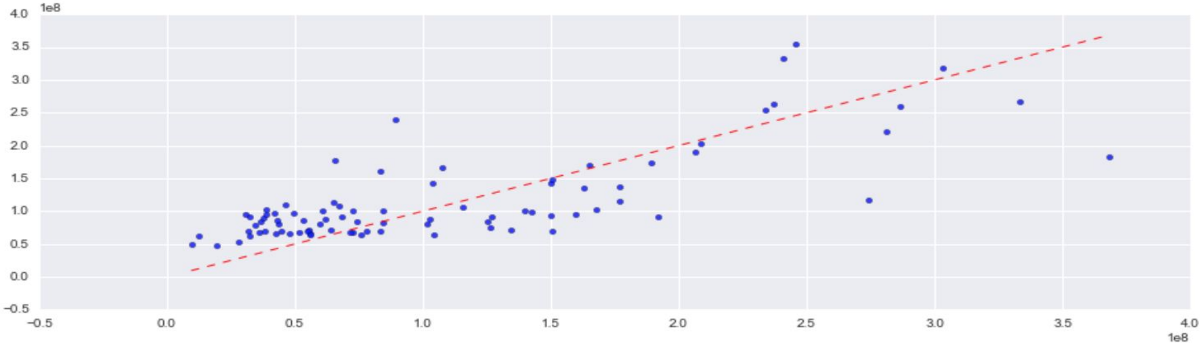
Results from three models.

Random Forest Regression Performed the best based on Score and RMSR

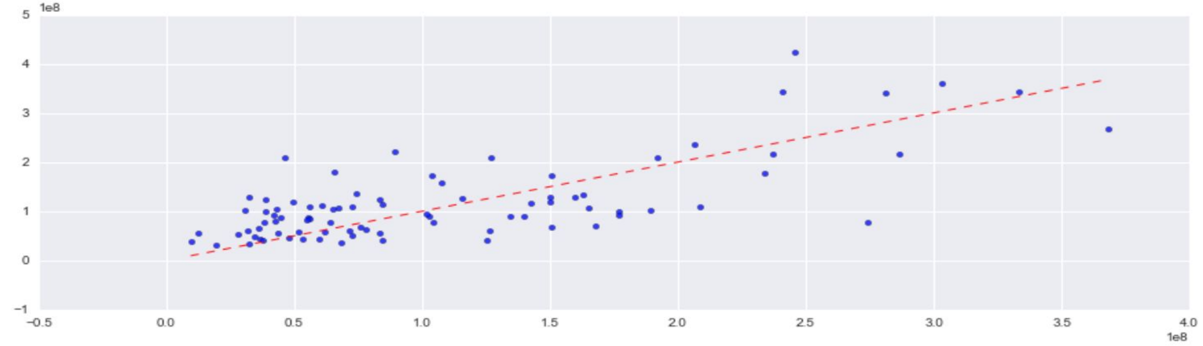
LR

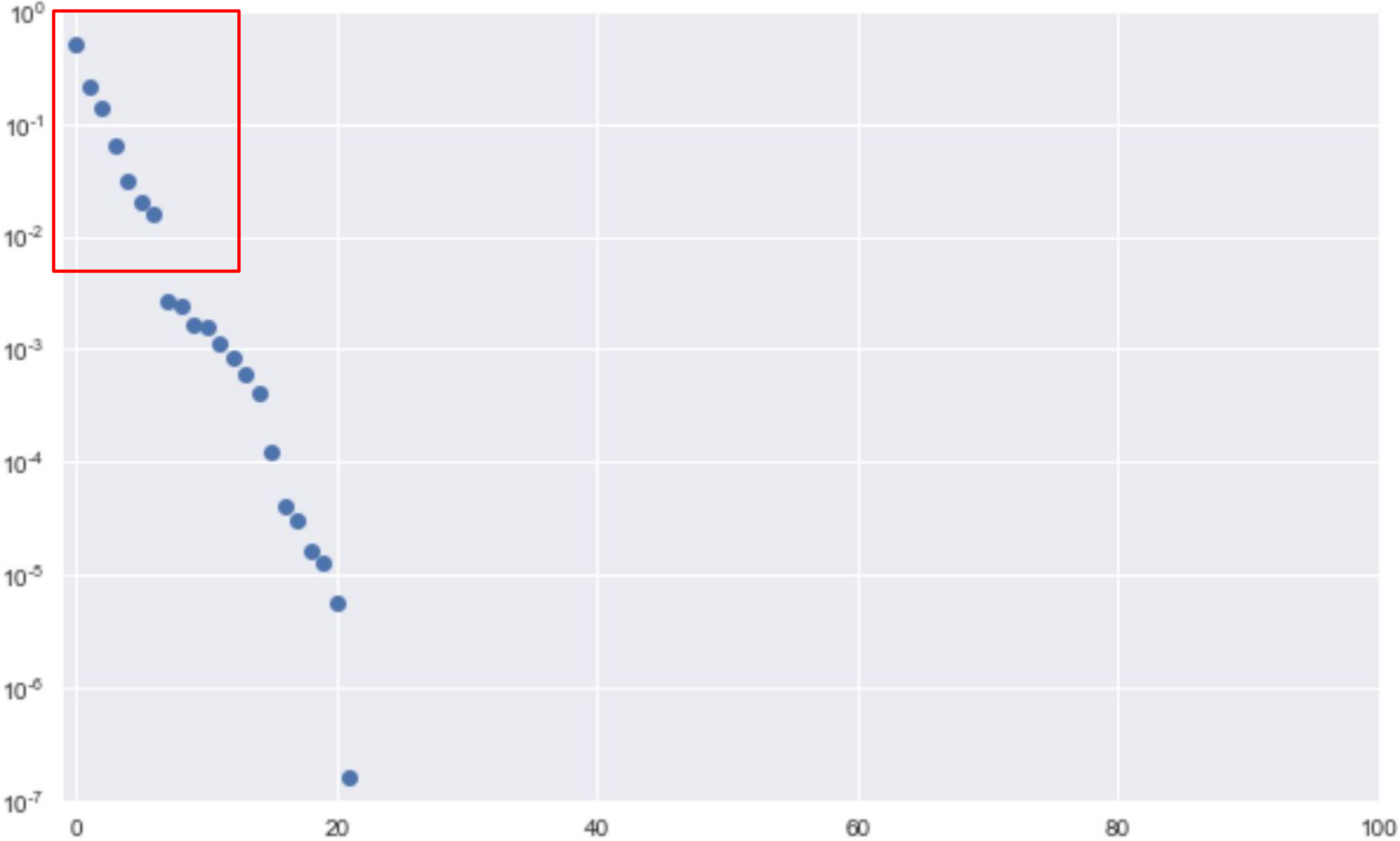


RFR



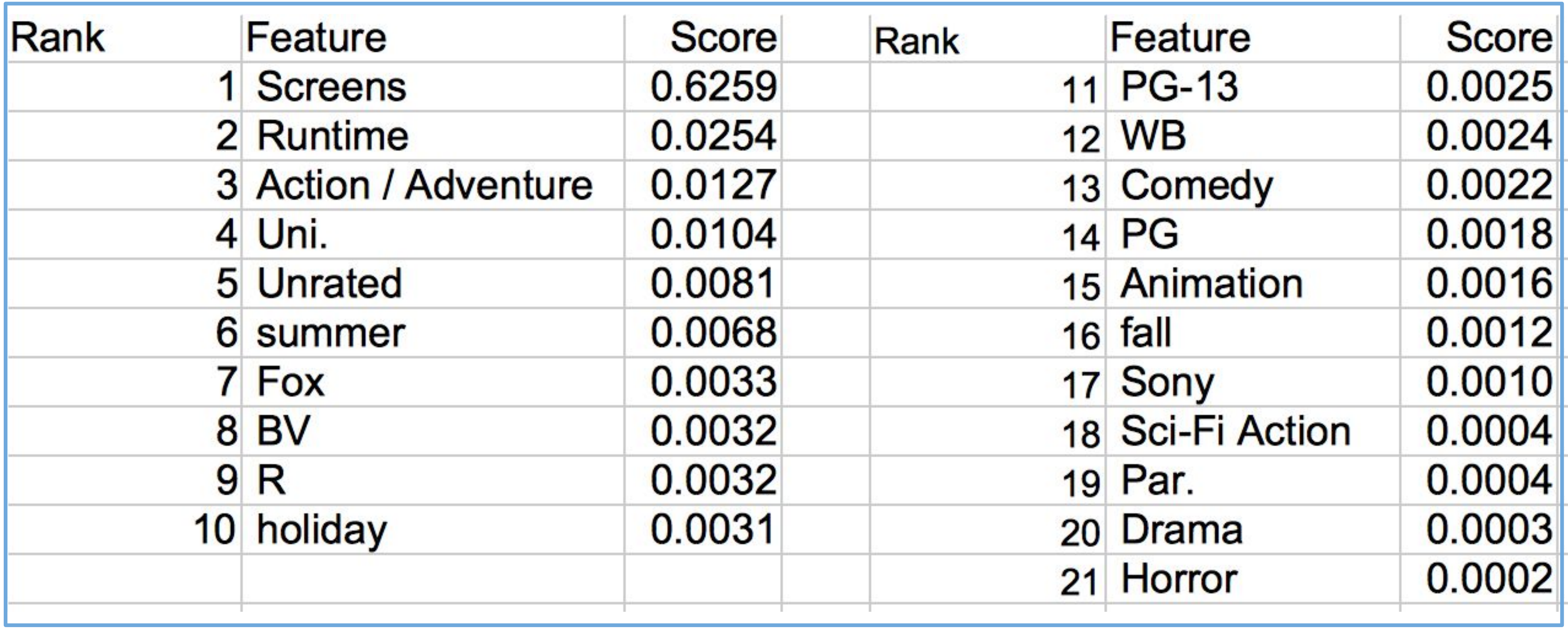
GBR




10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
0
20
40
60
80
100

Feature Importance Score

Based on Random Forest Regression



Rank	Feature	Score
1	Screens	0.6259
2	Runtime	0.0254
3	Action / Adventure	0.0127
4	Uni.	0.0104
5	Unrated	0.0081
6	summer	0.0068
7	Fox	0.0033
8	BV	0.0032
9	R	0.0032
10	holiday	0.0031
11	PG-13	0.0025
12	WB	0.0024
13	Comedy	0.0022
14	PG	0.0018
15	Animation	0.0016
16	fall	0.0012
17	Sony	0.0010
18	Sci-Fi Action	0.0004
19	Par.	0.0004
20	Drama	0.0003
21	Horror	0.0002

Final Thoughts

- Gross and screens seem to have a significant and non-linear relationship.
- Gross and budget seems to have a linear relationship.
- Gross and runtime seem to have negligible linear relationship.
- The relationship between Gross and release season are important.
- There seems to be a significant relationship between Gross and the certain Genre.

To Do:

- Need more data!
- Need more Budget data to calculate ROI.
- Screens data might be suspect.
- Explore separate populations.
- Explore Studio segmentation.
- Try Lasso.