Predicting Box Office Success through Linear Regression

Itelina Ma

July, 2015

Movie box office success can be predicted! Literature presents a variety of methods

Two major angles for predicting movie box office success:

Inherent Movie Characteristics

Movie Genre/ Franchise

Release Season

Cast/Director

Screens

Book to Movie Adaptation

Consumer Activity Prior to Release

Wikipedia Page Edits

Movie Trailer View Statistics

Facebook/ Twitter/Social Media Posts

Google Searches

Sources:

1. *Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. Mestyan et al.*
2. *The Drivers of Motion Picture Performance: The Need to Consider Dynamics, Endogeneity and Simultaneity. Elberse et al.*

This analysis focuses on using inherent movie characteristics as key predictive features

The following features were considered for the regression analysis

Brand

Marvel Comics, DreamWorks Animation, Walt Disney Animation Studios, etc

Studio

Fox, Buena Vista, Warner Brothers, etc

Series or Franchise

The Hunger Games, Avengers, Twilight, Spider-Man, Harry Potter, etc

Genre

Comedy, Action, Drama, Animation, etc

Screens

Describes the maximum number of screens that the movie was contracted for

Budget

Describes the approximate movie budget

Release Date

Date that the movie is released

Famous Stars

Describes how many members of the cast has been nominated for the Oscars

Famous Director

Describes whether or not the movie director has been nominated for the Oscars

Analysis Dataset: *All movies released from 2010 - 2015*
Data Sources: ***boxofficemojo.com, IMDB.com***

OLS Regression Results

Statistics on Model Performance

Comparison of Model Performance on Training and Testing Data

Various Models Considered	Original Model *Included all selected features*	Iteration 1 *Included significant features from original model(3)*	Iteration 2 *Included significant features from Iteration 1*
	Model Performance on Training Data		
Model Inputs	**258**	**39**	**23**
Average R^2 from Cross Validations	**73.4%**	**76.3%**	**76.0%**
SD of R^2 from Cross Validations	0.0533	0.0414	0.0486
	Model Performance on Testing Data		
R^2	**80.7%**	**81.2%**	**81.2%**

Cross Validation

Selected Model

Notes:

1. *Training data included 932 observations, while testing data contained 234 observations.*
2. *100 iterations of cross validation (hold out method) were performed.*
3. *Significant features are defined as those that have p value less than 5%.*

Can I predict the top 20 box office hits of a random list of movies?? Showing predictions vs. actuals from the test dataset

Actual Top 20 Movies	Predicted Top 20 Movies
Toy Story 3	Iron Man 3
Iron Man 3	The Hunger Games: Mockingjay - Part 1
Frozen	Transformers: Age of Extinction
The Hunger Games: Mockingjay - Part 1	The Hobbit: The Desolation of Smaug
Monsters University	X-Men: Days of Future Past
Captain America: The Winter Soldier	Cars 2
The Hobbit: The Desolation of Smaug	How to Train Your Dragon 2
Transformers: Age of Extinction	Toy Story 3
X-Men: Days of Future Past	Captain America: The Winter Soldier
Dr. Seuss' The Lorax	Monsters University
Dawn of the Planet of the Apes	The Wolverine
Cinderella (2015)	Dawn of the Planet of the Apes
Cars 2	The Lone Ranger
The Croods	Frozen
Pitch Perfect 2	Rango
How to Train Your Dragon 2	Cinderella (2015)
Rise of the Planet of the Apes	Dr. Seuss' The Lorax
The Help	The Croods
Gone Girl	Clash of the Titans (2010)
Clash of the Titans (2010)	Prometheus

Legend:

Correct

Incorrect

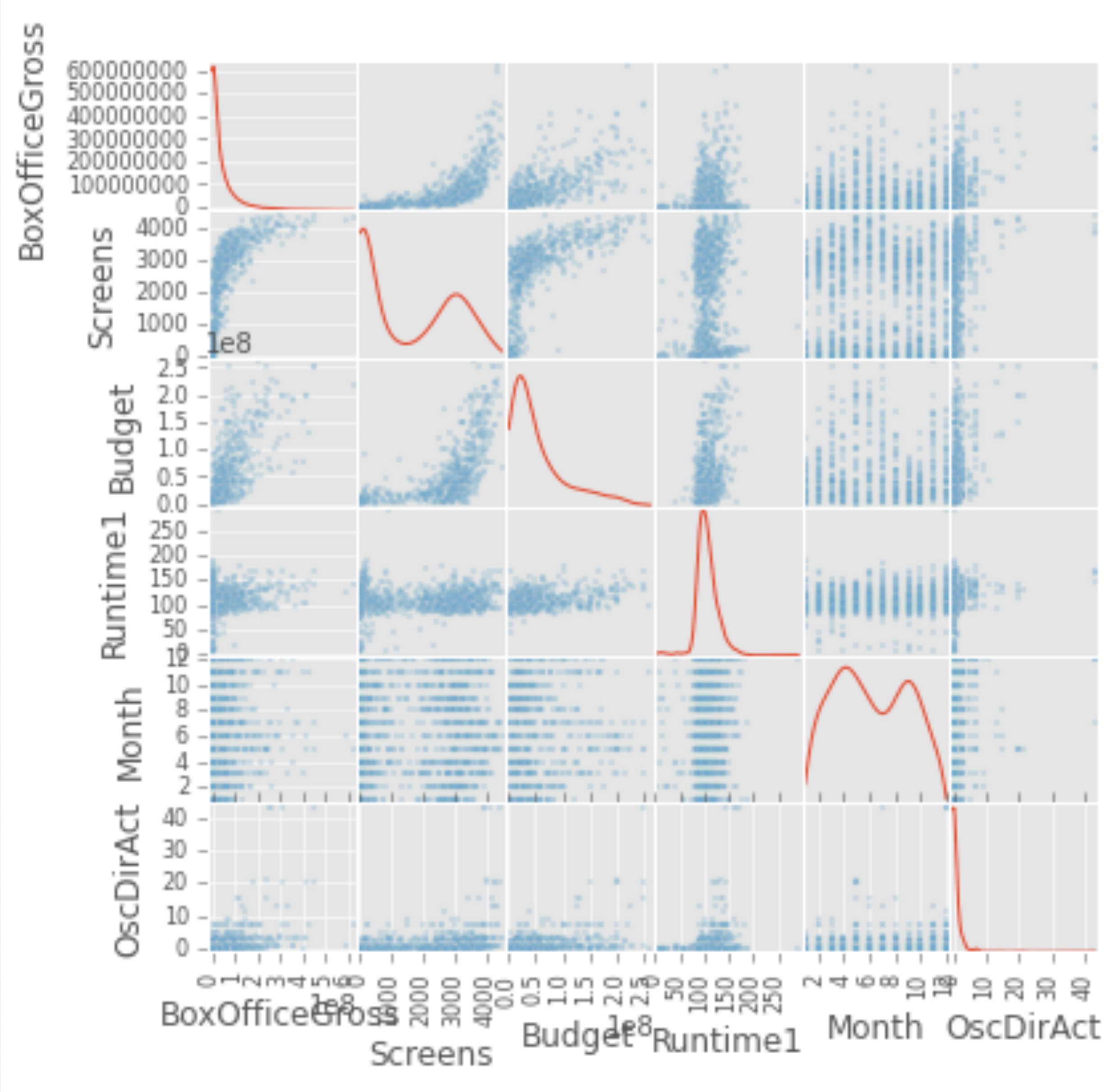
Percentage Predicted Correctly: 80%

Appendix

OLS Regression Results

Relationship Between Features (1 of 2)

Scatter Matrix Between All Variables



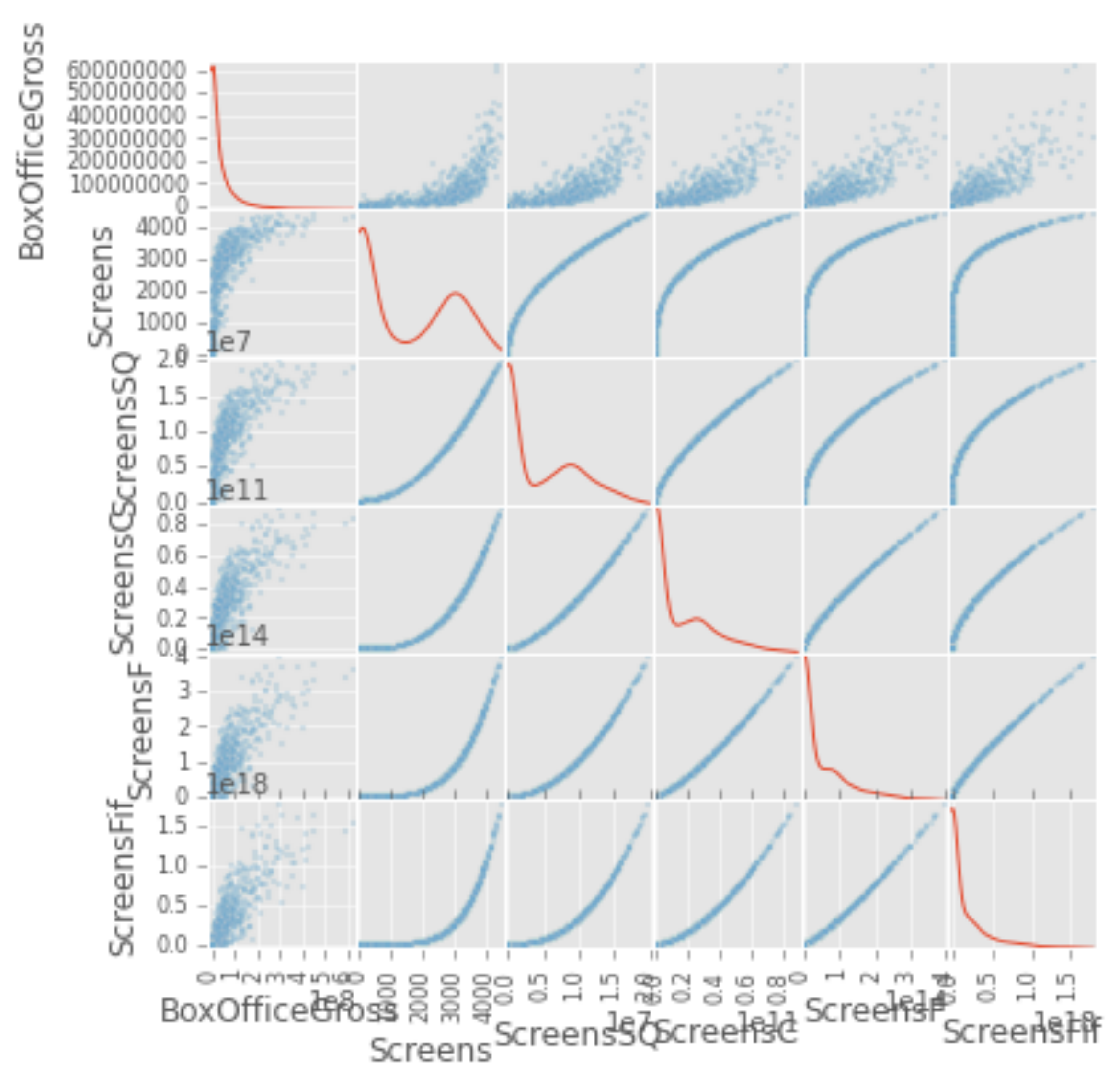
Key Takeaways

- Box office total and **screens** seem to have a **significant and non-linear relationship**
- Box office total and **budget** seems to have a **linear relationship**
- Box office total and **runtime** seem to have **no significant linear relationship**
- The relationship between box office gross and **months** seem to be less clear
- There seems to be a **significant relationship** between box office gross and the **"Star Factor"** (OscDirAct) variable

OLS Regression Results

Relationship Between Features (2 of 2)

Scatter Matrix Between Box Office and Screens Raised to Different Powers



Key Takeaways

- **As we raise screens to higher powers,** the relationship with box office gross seems to become **more linear.**
- In this model the various powers for screens are all included as features

OLS Regression Results

Summary of Final Model Features

Categories	Features	
Genre + Brand	Action-Marvel Comics Adventure-Tim Burton-Johnny Depp Animation-DreamWorks Animation Animation-Illumination Entertainment Animation-Marvel Comics	Animation-Pixar Animation-Walt Disney Animation Studios Biography-Legendary Pictures Drama-Stephen King
Studio	BV Fox KE	P/DW SGem Uni.
Screens	Screens^2 Screens^3 Screens^4	
Release Date	June November December	
Famous Stars + Famous Director + Series	Oscar Star(s) + Oscar Directors + Series + Oscar Star(s)* Series + Oscar Director * Series + Oscar Star(s) * Oscar Director	