Predicting Movie Domestic Total Gross

By: Lubna Alhenaki

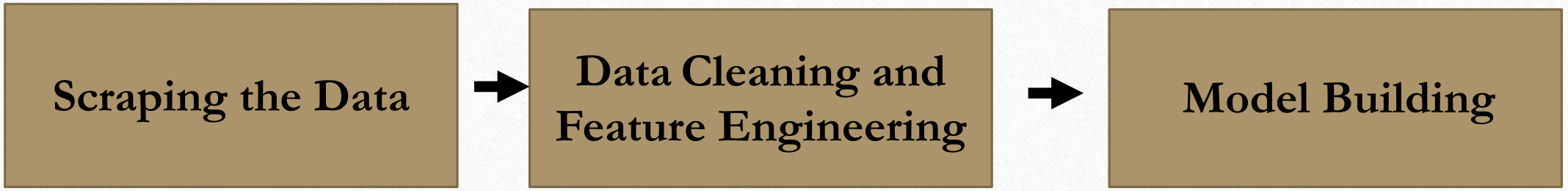
Outline

- Introduction
- Methodology Design
- Conclusion and Future work

Introduction

- The film industry is a significant player in the global economy.
- The business question is whether we could use a regression model to identify domestic total gross movies .
- The major goal is to predict a domestic total gross value by using a linear regression model.

Predicting Movie Domestic Total Gross Workflow



Scraping the Data

- Box Office Mojo
- BeautifulSoup

Features
News
Release Sched.
Showtimes
at IMDb

Box Office
Daily
Weekend
Weekly
Monthly
Quarterly
Seasonal
Yearly
All Time
International

Indices
Studios
People
Genres
Franchises
Showdowns
Theater Counts

2018 DOMESTIC GROSSES

Total Grosses of all Movies Released in 2018

#1–100 - #101–200 - #201–300 - #301–400 - #401–500 - #501–600 - #601–700 - #701–800 - #801–879

< Previous Year Data as of: Today or (Month / Day / 2018) Go Next Year >

Rank	Movie Title (click to view)	Studio Filter	Total Gross / Theaters	Opening / Theaters	Open	Close
1	**Black Panther**	V	**$700,059,566** 4,084	$202,003,951 4,020	2/16	8/9
2	**Avengers**					
3	**Incredibl**					
4	**Jurassic V Kingdom**					
5	**Aquaman**					
6	**Deadpool**					
7	**Dr. Seuss (2018)**					
8	**Mission: Fallout**					
9	**Ant-Man**					
10	**Bohemia**					
11	**A Star is**					
12	**Solo: A S**					
13	**Venom (2**					

Search Site
Search...

Features
News
Release Sched.
Showtimes
at IMDb

Box Office
Daily
Weekend
Weekly
Monthly
Quarterly
Seasonal
Yearly

Dr. Seuss' The Grinch (2018)

Domestic Total Gross: **$270,620,950**

Distributor: **Universal**	Release Date: **November 9, 2018**
Genre: **Animation**	Runtime: **1 hrs. 30 min.**
MPAA Rating: **PG**	Production Budget: **$75 million**

Summary | Daily | Weekend | Weekly | Foreign | Similar Movies

Total Lifetime Grosses
Domestic: $270,620,950 52.9%

The Players
Directors: Yarrow Cheney

Related Stories
1/30 A Look Back at 2018's Record Year at the Domestic

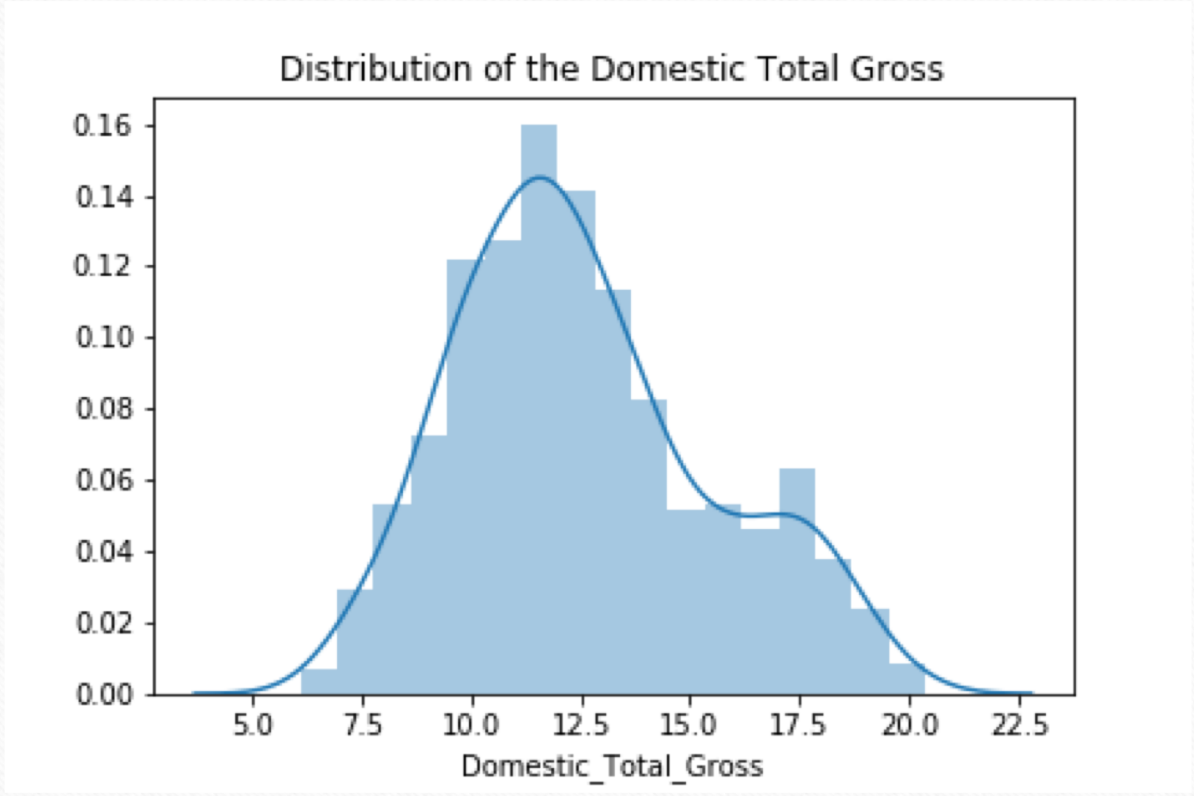
Movie Dataset

- All top Movies Released in 2018
- 879 movies and 8 features
- Title, Domestic_Total_Gross, Distributor, Release_Date,Runtime Production_Budget, MPAA_Rating, and Genre

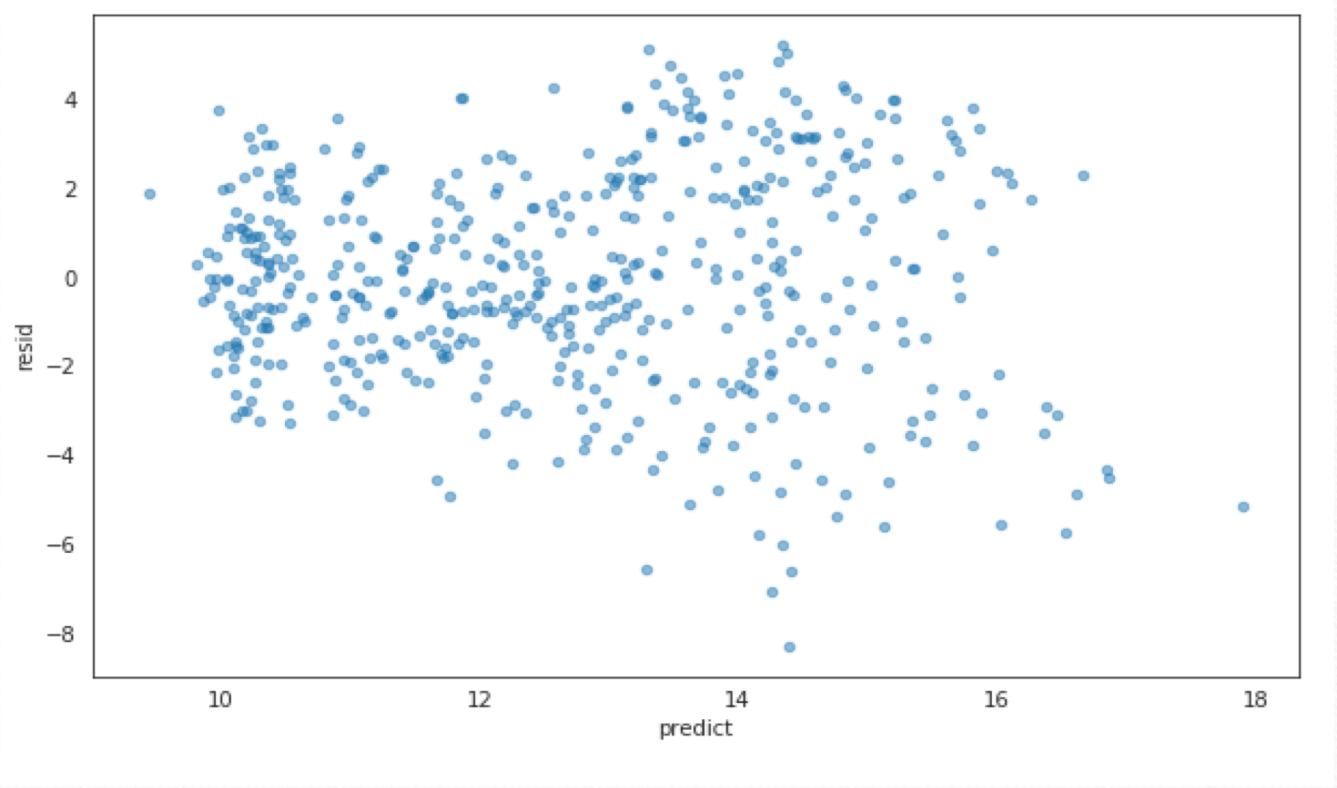
Data Cleaning and Feature Engineering

- Turning the Release Date strings into datetime objects
- Turning the runtime string from "1 hr. 52 min." to minutes float format
- Getting the Month , Day of Year and Day form Release Date
- Dealing with categorical data such as Genre and MAPP Rating
- Drop the Null values
- Shuffling Data

Model Building



Residual plot



Model Selection Process

- Stage 1 :

BaseLine model: Validation R^2= 0.119

- Stage 2:

ElasticNet model: Validation R^2=0.281

- Final Model : Test r^2= 0.298

Conclusion and Future Work

- Using a linear regression model, a month feature influence prediction of the domestic total gross.
- The results of the model shown that there was no significant relationship between a genre and domestic total gross.
- For future work, other features will be selected and improve model r^2.

Thank You

Appendix

- P-Values for Variables in a Regression Model
- Regression Coefficients
- Correlation Matrix

P-Values for Variables in a Regression Model

Covariance Type:		nonrobust				
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	16.1525	5.006	3.227	0.001	6.318	25.987
Runtime	0.0125	0.004	3.049	0.002	0.004	0.021
MPAA_Rating	-0.9863	0.073	-13.591	0.000	-1.129	-0.844
Day	0.0372	0.159	0.23	0.815	-0.275	0.349
day_of_year	-0.0236	0.159	-0.14	0.882	-0.335	0.288
Month	0.7491	4.826	0.15	0.877	-8.733	10.231
Komedy_True	-0.9823	0.433	-2.267	0.024	-1.834	-0.131
Othors_True	-0.5187	0.503	-1.032	0.303	-1.506	0.469
Anoimation_True	-2.1299	0.627	-3.395	0.001	-3.362	-0.897
DoKumentary_True	-2.2678	0.566	-4.007	0.000	-3.380	-1.156
Dorama_True	-1.6523	0.443	-3.728	0.000	-2.523	-0.782

Omnibus:	6.328	**Durbin-Watson:**	2.077
Prob(Omnibus):	0.042	**Jarque-Bera (JB):**	6.439
Skew:	-0.257	**Prob(JB):**	0.0400
Kurtosis:	2.820	**Cond. No.**	1.55e+04

Df Residuals:	513	**BIC:**	2430.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	16.7923	0.589	28.525	0.000	15.636	17.949
Runtime	0.0132	0.004	3.247	0.001	0.005	0.021
MPAA_Rating	-0.9915	0.072	-13.709	0.000	-1.134	-0.849
Komedy_True	-0.6542	0.291	-2.251	0.025	-1.225	-0.083
Anoimation_True	-1.6305	0.436	-3.740	0.000	-2.487	-0.774
DoKumentary_True	-1.7847	0.340	-5.255	0.000	-2.452	-1.118
Dorama_True	-1.2415	0.248	-5.007	0.000	-1.729	-0.754

Omnibus:	6.313	**Durbin-Watson:**	2.055
Prob(Omnibus):	0.043	**Jarque-Bera (JB):**	6.408
Skew:	-0.255	**Prob(JB):**	0.0406
Kurtosis:	2.810	**Cond. No.**	695.

Regression Coefficients

```
Runtime : 0.01
MPAA_Rating : -0.99
Day : 0.04
day_of_year : -0.02
Month : 0.75
Komedy_True : -0.98
Othors_True : -0.52
Anoimation_True : -2.13
DoKumentary_True : -2.27
Dorama_True : -1.65
```

Correlation Matrix