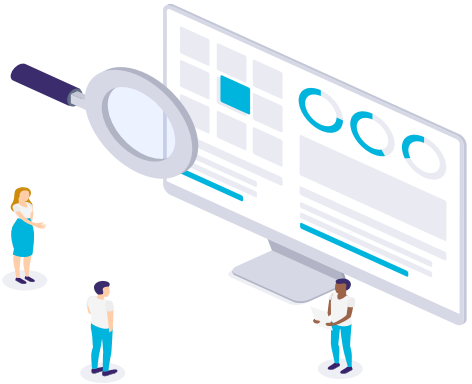
Building Regression Models for Movie: Predicting Domestic Box Office in North America

Presented by Jill Ke

Introduction

Objective: build regression models to predict domestic gross of the movie in North America based on the features and select the best model for accurate prediction.

Goal: provide insights to film investors who are involved in making strategic decisions during the process of a movie production.



Methodology

Data Collection

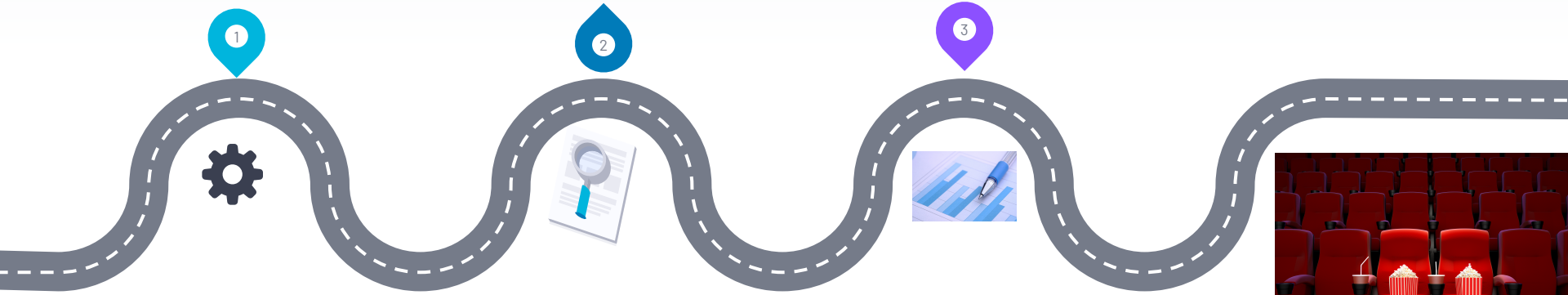
Data is scrapped from Boxofficemojo.com website.

Data Cleaning

Perform exploratory data analysis on missing data and irrelevant columns.

Models

Build regression models and use different evaluation metrics to evaluate the models.



Data

Mission: Impossible - Fallout

Ethan Hunt and his IMF team, along with some familiar allies, race against time after a mission gone wrong.

Title Summary | Original Release | Domestic

Grosses

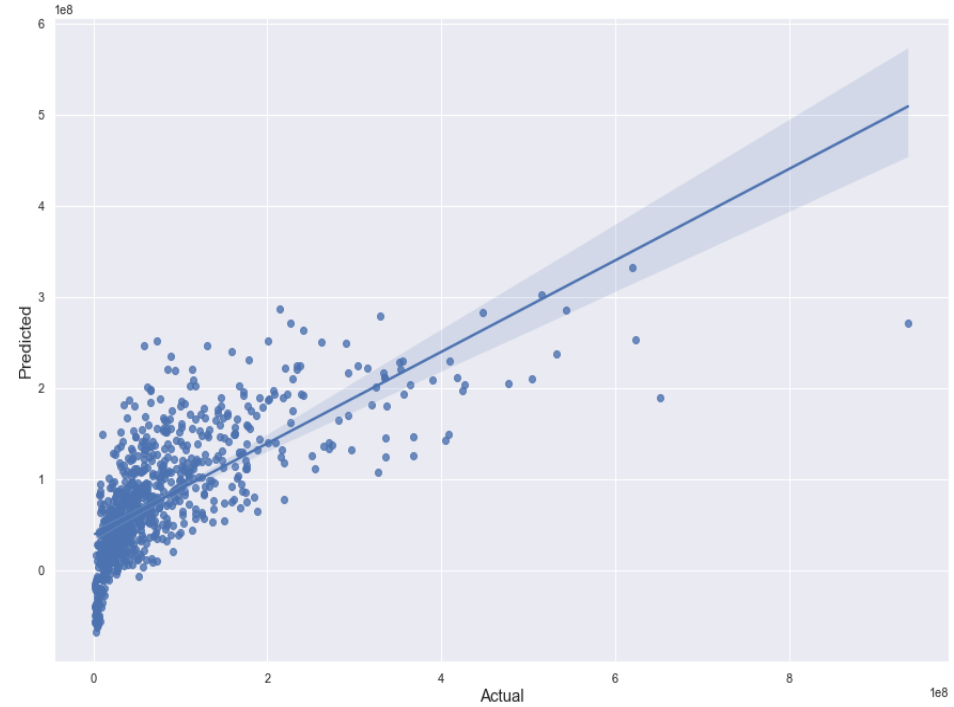
DOMESTIC (27.8%)
$220,159,104
INTERNATIONAL (72.2%)
$570,956,000
WORLDWIDE
$791,115,104

Distributor	Paramount Pictures See full company information
Opening	$61,236,534 4,386 theaters
Budget	$178,000,000
Release Date	Jul 27, 2018 - Oct 18, 2018
MPAA	PG-13
Running Time	2 hr 27 min
Genres	Action Adventure Thriller
In Release	158 days/22 weeks
Widest Release	4,395 theaters

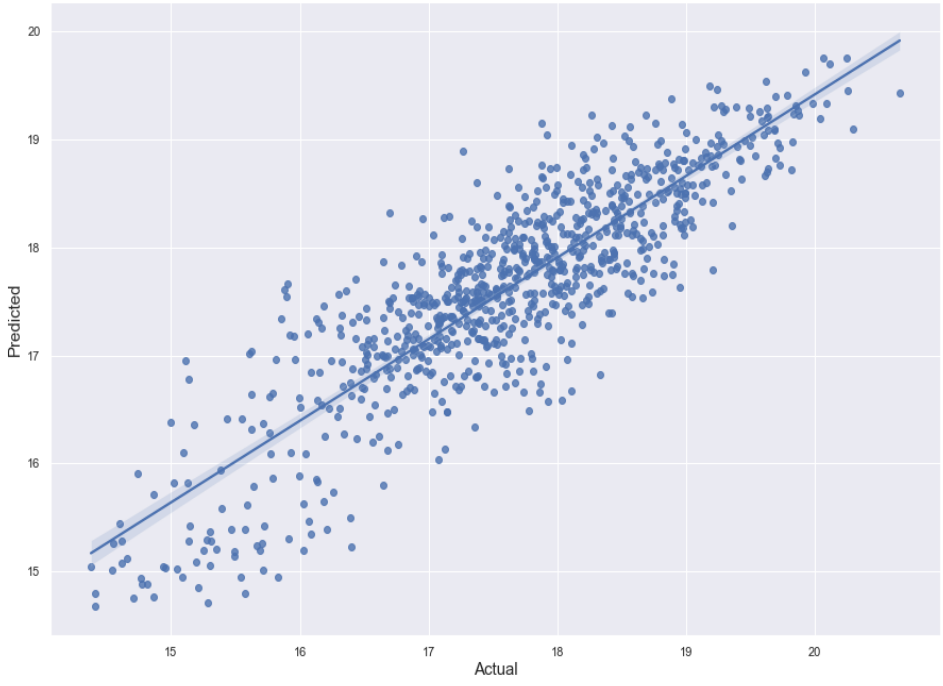
- from 2010~2019 domestic (North America)
- 1000+ movies
- 9 features

Linear Regression Model

Before log transformation of target variable



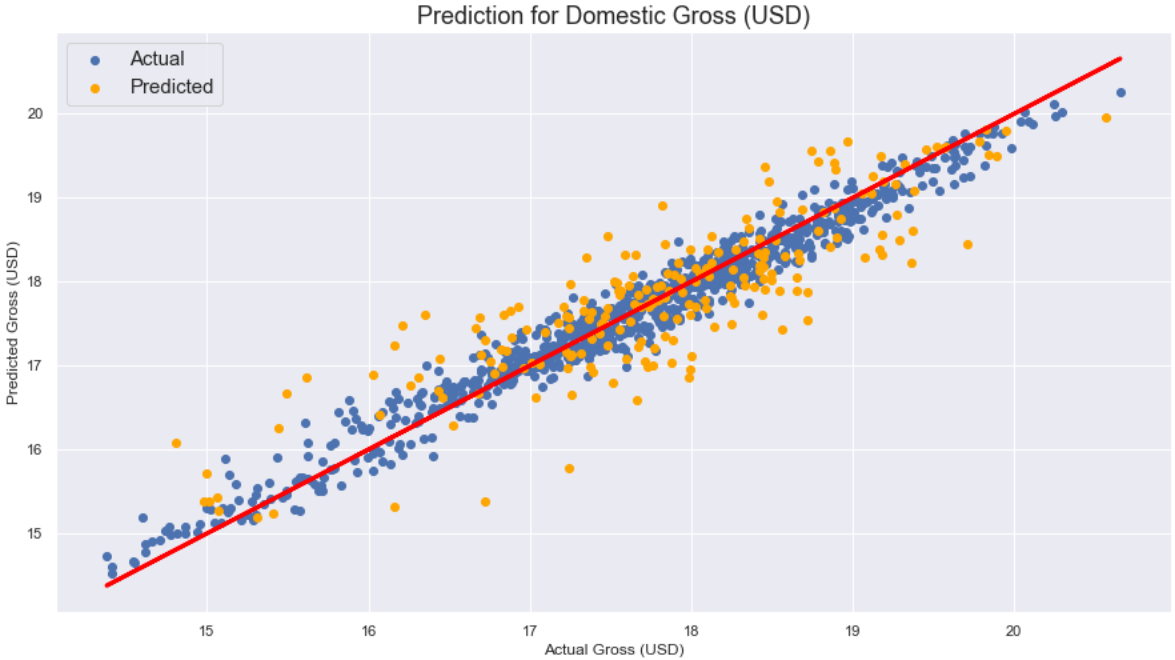
After log transformation of target variable



Regression Models

		After Log Transformation of Target Variable (Domestic Gross)			
	Linear Regression	**Linear Regression**	**Lasso**	**Ridge**	**Random Forest**
R^2	0.53	0.72	0.73	0.72	0.74
RMSE	69,031,376	0.56	0.55	0.56	0.53
MAE	45,718,340	0.44	0.44	0.44	0.42
K-fold Cross Validation of R^2	0.50	0.71	0.71	0.71	0.72

Conclusion



- Random forest is the best model to predict movie's domestic gross with $\mathbf{R^2}$ = 0.74 and MAE=0.42

Future Work:

- Collect more data from difference websites
- Do more data cleaning

Thank you!