Regression Analysis:

Predicting the International vs. Domestic Share of Box Office Revenue

ANALYSIS BY: Olivia Offutt, Metis Regression

INTRODUCTION

OVERVIEW

- Global markets are significant sources of revenue for US made movies
- Predicting what kind of movie will produce a high international response would be helpful for a movie distributor to have in its business decision making toolkit

PROJECT GOAL:

Using data from Box Office Mojo, build a predictive model for international revenue percentage (%) of revenue for domestic made movies

METHODOLOGY

DATA

Sample Frame: 1000 domestic movie web pages from Box Office Mojo Top Lifetime Grosses page

Variables:

- *Target Variable (1):* International Revenue %
- *Feature Variables (95):*
  - Numeric Vars: [Year, Run Time, Budget (adjusted for inflation)]
  - Categorical Vars: [Distributor, Rating, Genres, Directors, Actors, Release Month]

METHODOLOGY

TOOLS

- **Web Scraping:** BeautifulSoup, Pickle
- **Data Visualization:** Matplotlib and Seaborn
- **Regression Analysis:** SciKitLearn, statsmodels

METHODOLOGY

METRICS

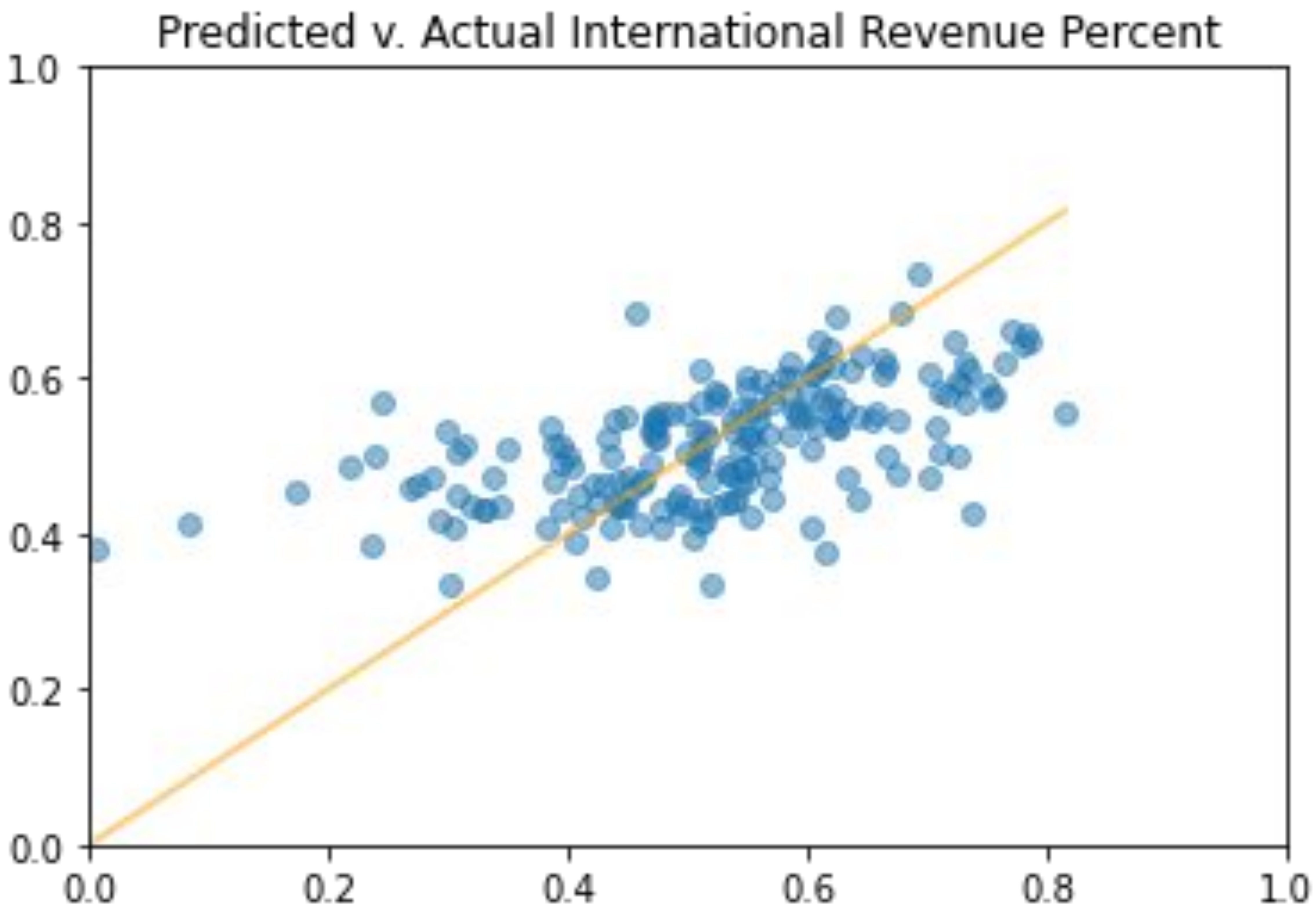
- **Model variable review (feature engineering)**
  - Fit linear regression to sample data
  - Pairplots, heatmaps, and VIF analysis to ID/address collinearity
  - Residuals scatter plot to ID/address heteroskedasticity
- **Trained and tested 4 regression models** scored for explanation power (R^2 value) and magnitude of error value (MAE):

METHODOLOGY

METRICS

- **Standard OLS, cross validated (KFold)*:**
  - R^2 = 0.33 +- 0.11
  - MAE = 0.1 +- 0.01
- **Polynomial OLS, cross validated (KFold))*:**
  - R^2 = - 0.21 +- 0.36
  - MAE = 0.13 +- 0.02
- **Ridge Regression, cross validated (RidgeCV))*:**
  - R^2 = .29
  - MAE = 0.09
- **Lasso Regression, cross validated (LassoCV))*:**
  - R^2 = .33
  - MAE = 0.09

RESULTS



Final Model:

Lasso Regression, cross validated (LassoCV)

R^2 = .33

MAE = 0.09

RESULTS

Variables with highest coefficient (standardized)

'budget_adj'	0.034
'year'	0.032
'Adventure'	0.022
'Animation'	0.012
'Thriller'	0.011
'Comedy'	0.015
'Will Ferrell'	0.016

Variables that were dropped to zero during Lasso Regression:

- Run Time
- All Directors
- All Distributors
- Al Months Released
- Most Actors
- Many Genres

CONCLUSIONS

Unfortunately, the model in its current form is not very predictive or useful for a business case.

Ways to improve the model

- Better sample Frame:
  - Use a much larger data set
  - Use a data set within a more recent time frame (last 10 years for example)

- Better target:
  - International % may be too broad and capturing too many variables
  - Model can be refined to predict more specific targets, e.g. Chinese box office revenue

- Better features:
  - Search for another movie data website that has data that promises to be more explanatory for predicting international market targets