The Art of Movie Success

Predicting a film’s box office gross based on artistic metrics

Kevin Zecchini

Contents

- Tools Used
- Motivation/Assumptions
- Data Collection
- Data Analysis
- Pitfalls and Challenges
- Areas for Improvement - Next Steps

Tools

Requests

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

SM StatsModels

Statistics in Python

matplotlib

scikit learn

NumPy

What assumptions are we making?

- **Success**
  - domestic total gross
- **Artistic metrics**
  - average shot length
  - runtime of film

Data Collection

“There’s no crying in data science!”

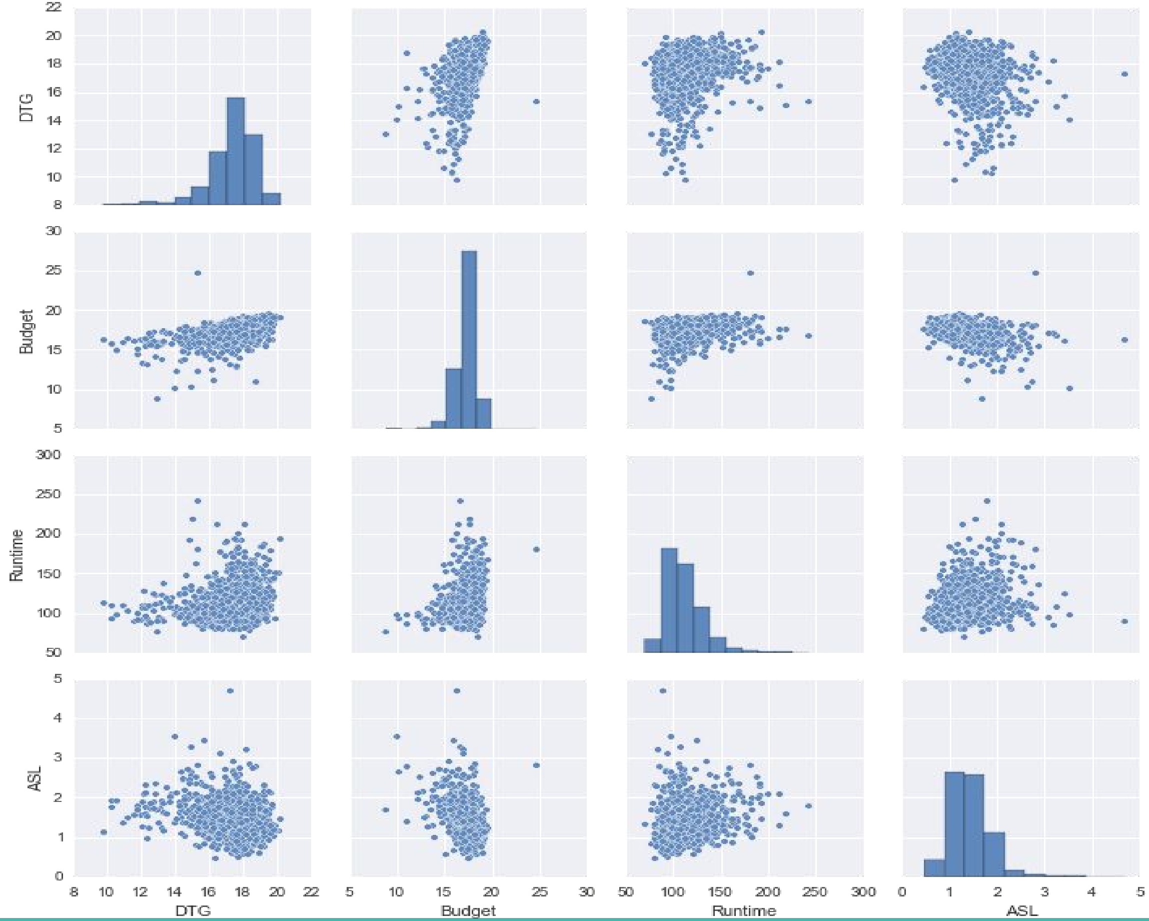
- Scrape BoxOfficeMojo
  - Domestic total gross, runtime, and other features that may be important
- Scrape Barry Salt’s Database hosted on Cinemetrics
  - Average Shot Length
- Merge two databases
  - 1409 films total

Box Office Mojo

CINEMETRICS

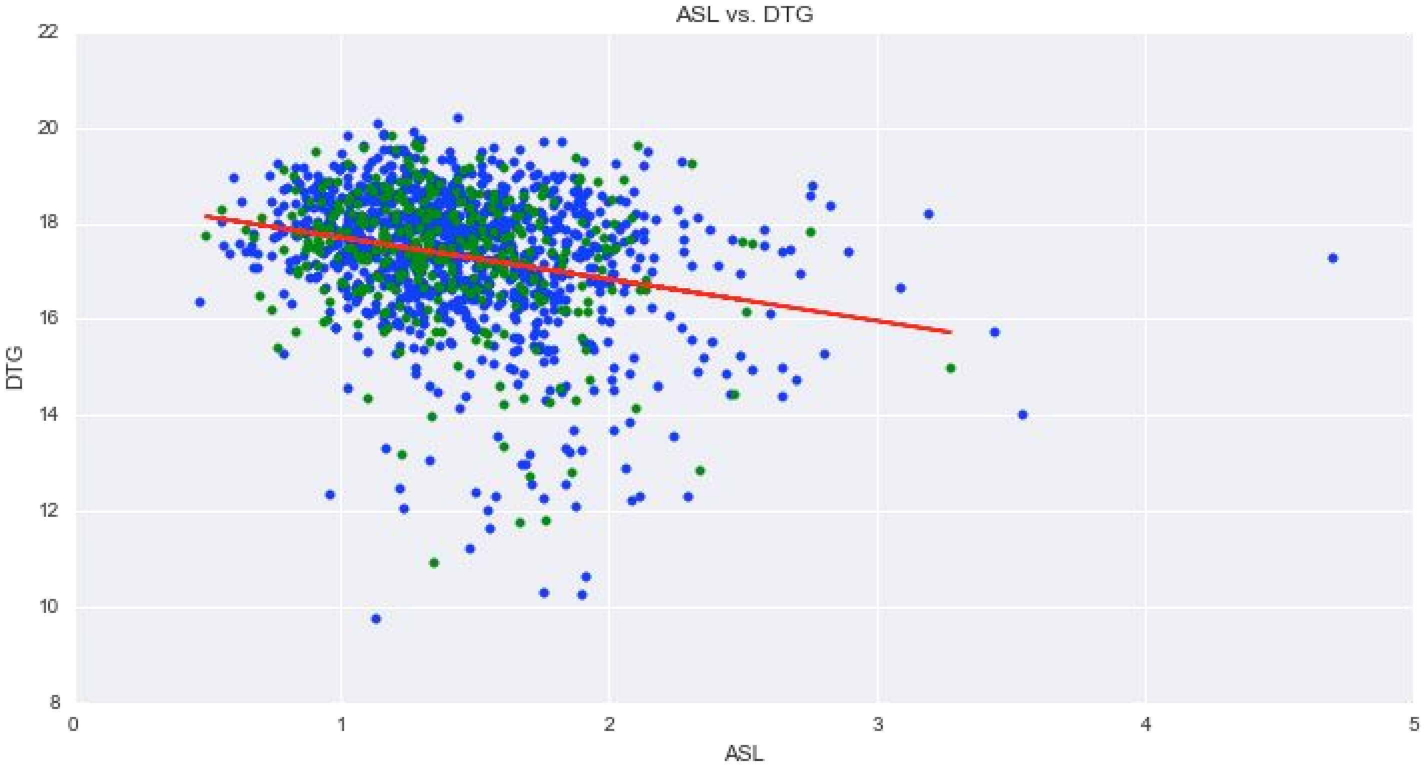
Data Analysis - Feature Interaction

- Features
  - Average Shot Length
  - Runtime
  - Budget
- Dependent variable
  - Domestic Total Gross
- Data was transformed to be approximately normal and scaled

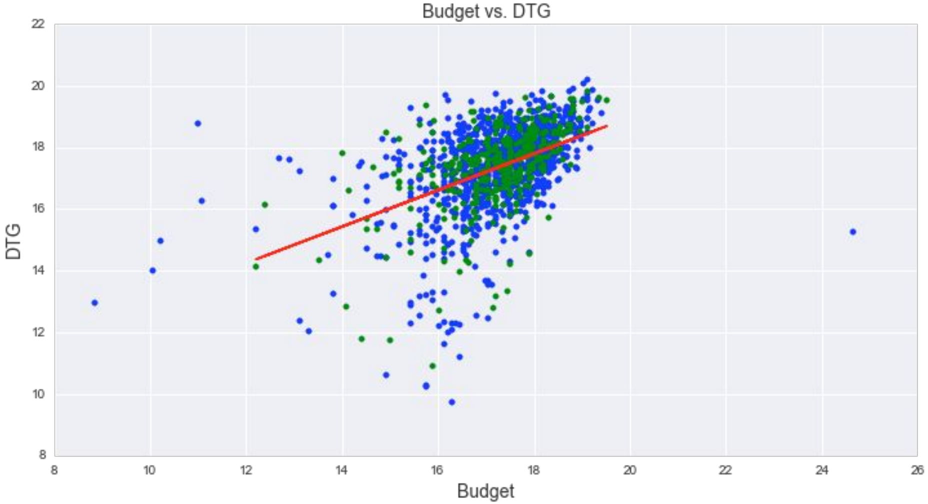


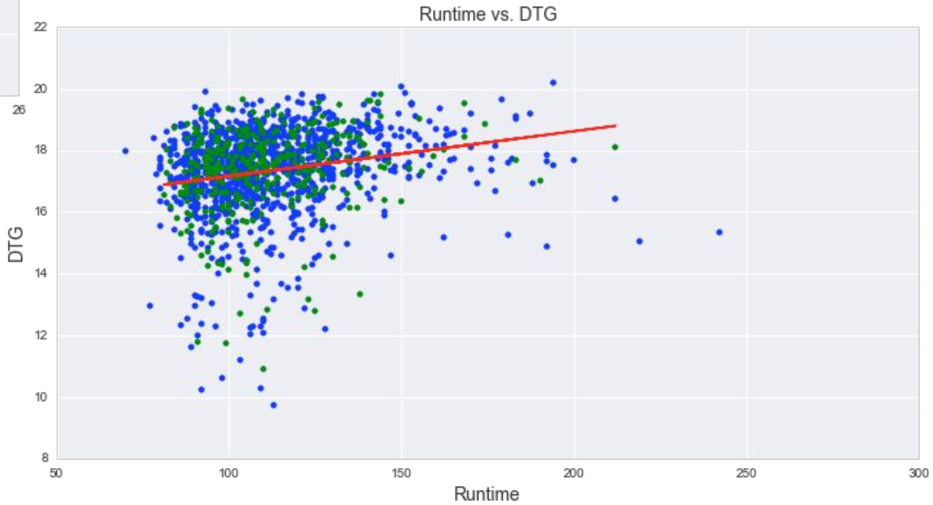
Data Analysis - ASL Regression

“Show me the money!”



Data Analysis - Budget and Runtime Regression

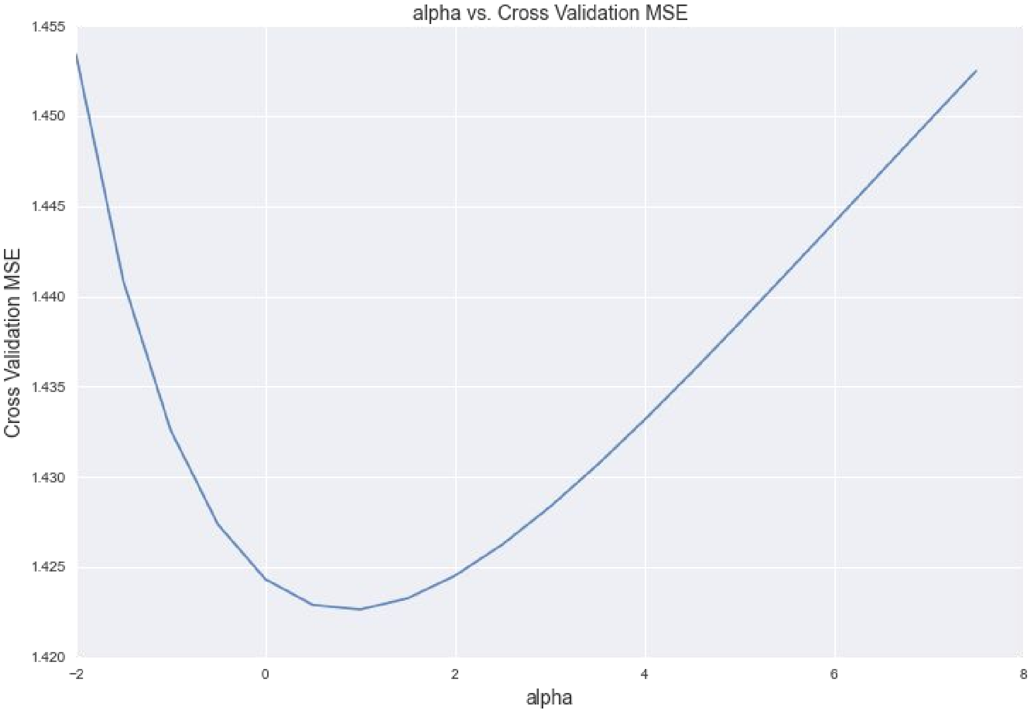




Data Analysis - Combining Features

“You’re gonna need a bigger model…”

- R^2 value increased slightly
  - $R^2 = 0.300$
- Main feature is budget
- Ridge regression shows runtime and asl do not contribute very much



Challenges

- Scraping websites and cleaning data
    - It’s a pain
    - Finding artistic data is difficult!
- Domestic Total Gross
    - Not the best descriptor of success
    - We see budget has a much greater influence
    - Don’t have adjusted values - years might come into play
- Movies are from all years, genres, and ratings
    - Preserved to keep data set large
    - Could subset or add as features to see if more conclusions could be drawn

Next Iterations - Look at more data

“I’ll be back.”

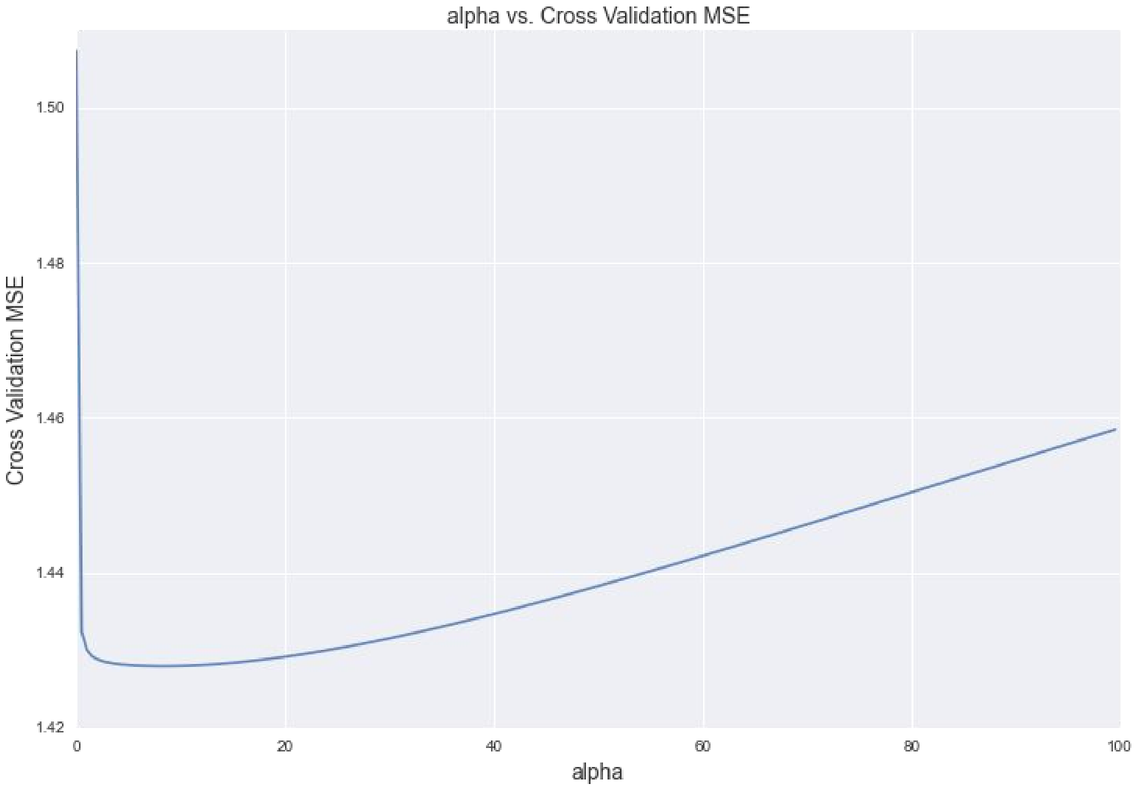
- Redefine assumptions
  - Success could be...
    - Number of film related nominations or wins
    - Website ratings
  - Artistic value could be...
    - Percentage of film with dialogue
    - Percentage of film scored with music
    - Color palette
    - Number of types of shots (high angle, low angle, zoom)
      - Barry Salt has a small database
- Subset based on year, genre, rating

But for now…

Average shot length does not have a strong impact on predicting domestic gross

Questions????

Backup Slides


alpha vs. Cross Validation MSE
Cross Validation MSE
alpha