Project Luther

Will ‘Zubat’ Cosby

Problem

Should the studio pursue an international release for their movie?

What is the Data?

- Obtained from BoxOfficeMojo.com.
- Movies from 2010-2016.
- Primary fields of interest:
  - Genre
  - Budget
  - Time of Release
  - Staff involved

Cleaning the Data

Problems:

- Missing values.
- Blocked scraping.
- Scale.

Solutions:

- Used subset of scraped data with the information needed.
- Ran out of time.
- Scaled down Budget and Foreign earnings information.

Feature Engineering

Genre:

- One-hot vectors, broke apart “joint” categories.

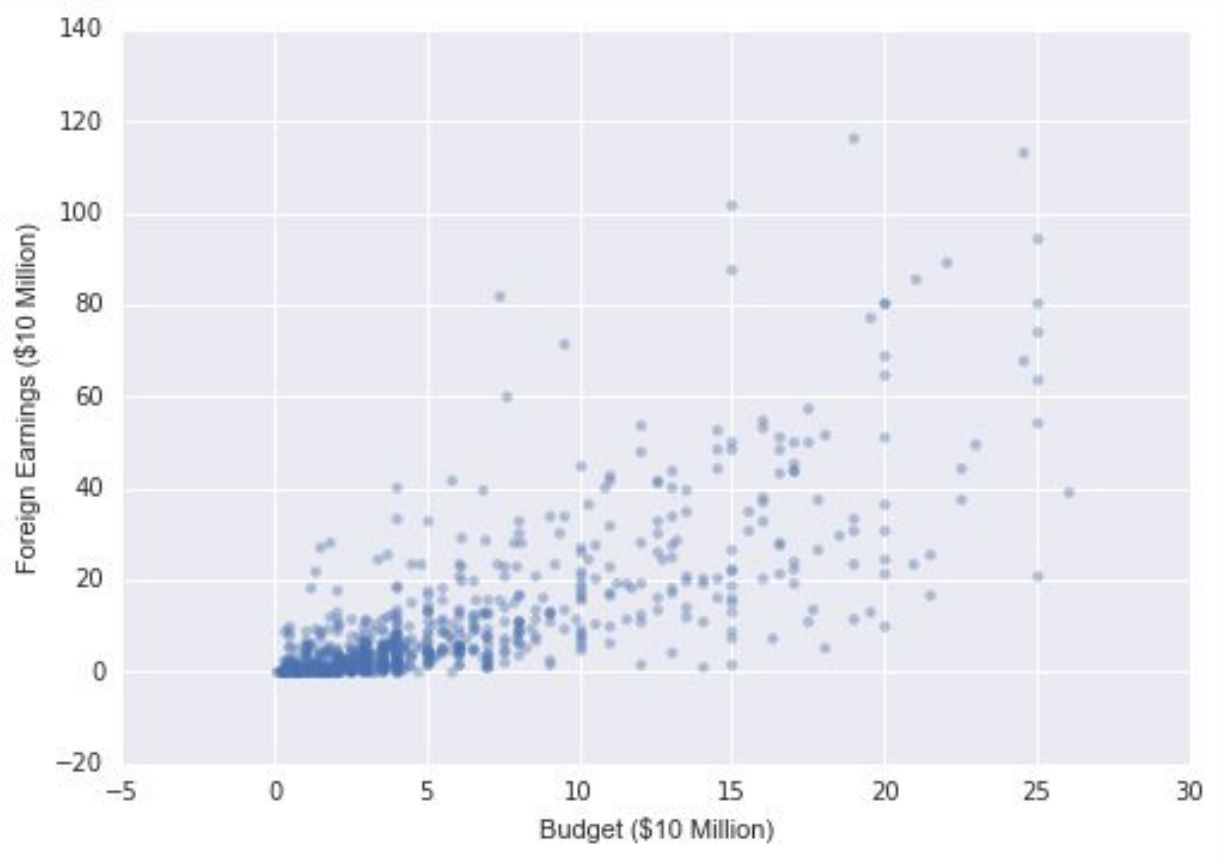
Team Experience:

- Summed the total actor, writer, director, and producer experience.

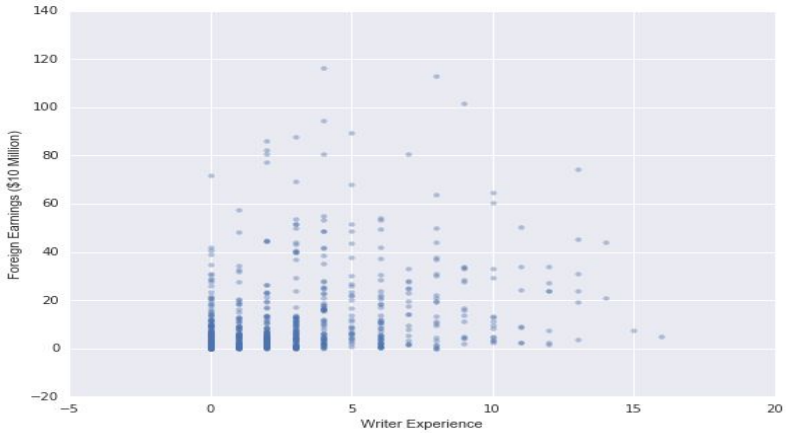
Release Date:

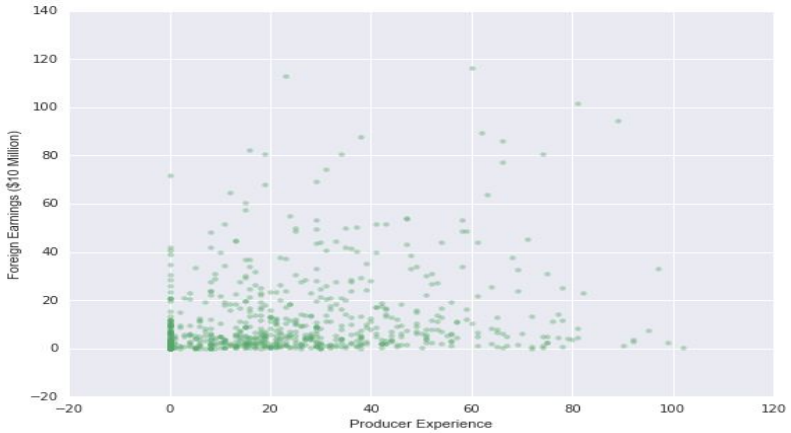
- Decomposed dates into financial quarter (Q1, Q2, Q3, Q4), represented as one-hot vectors

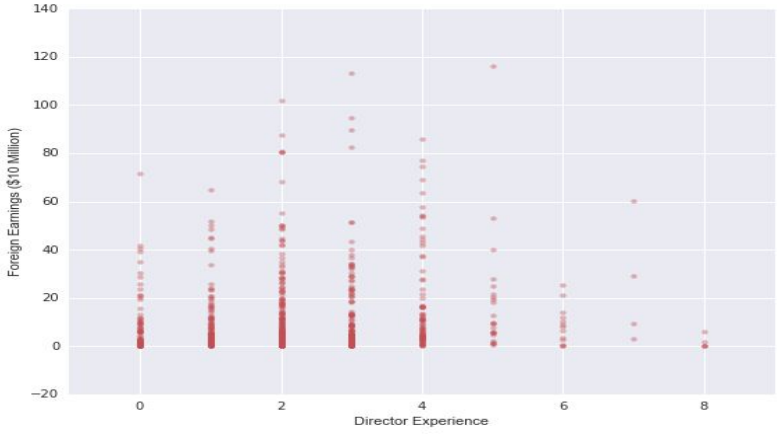
Budget Vs. Foreign Earnings



Team Experience







Modeling

Types of Models:

- Linear Regression
- Random Forests
- Gradient Boosted Trees

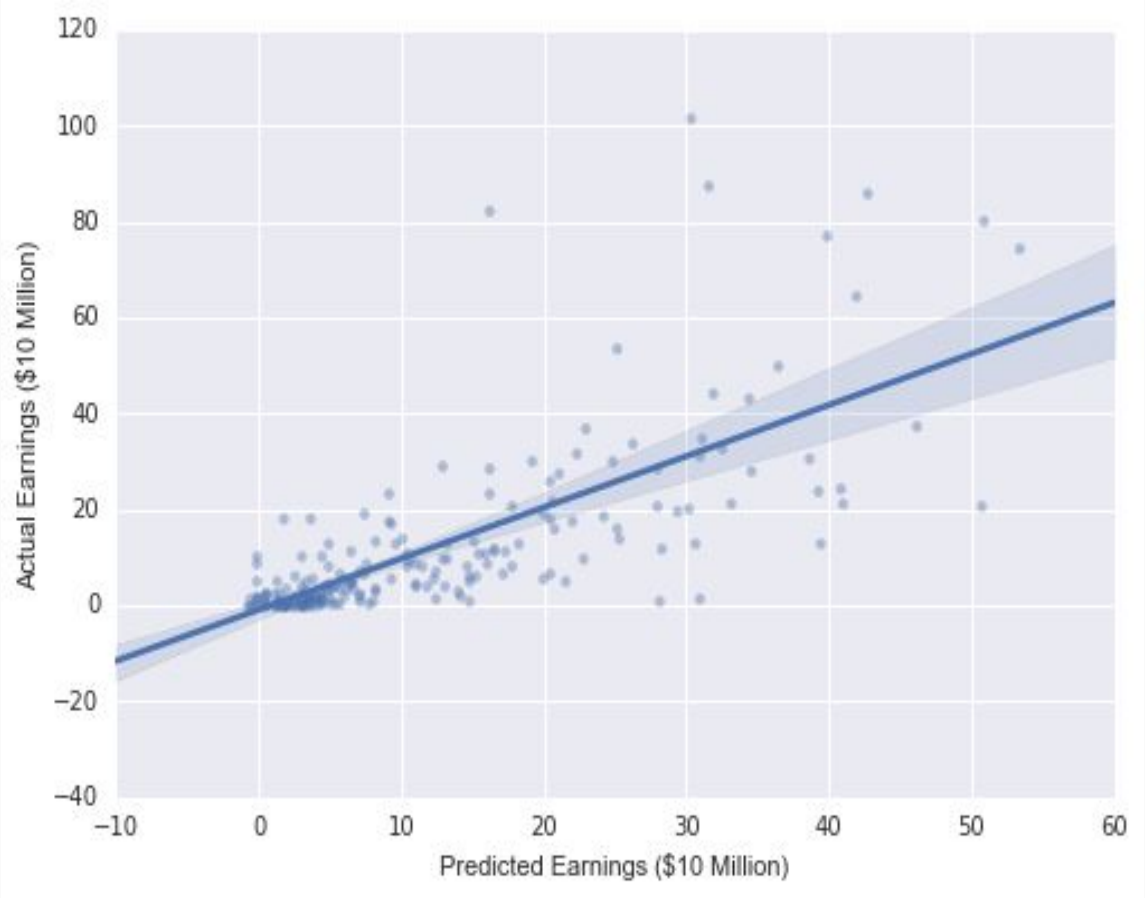
Technique:

- 10% of data held completely out of testing.
- 10-fold Cross Validation
- Grid Search over various lin-reg models
- Manual tuning for Random Forest and Gradient Boost

Best Model

Linear Regression:

- Lasso
- Test Set Performance: **56.1%**
- Average Cross Validation Score: **53%**
- Parameters:
  - Alpha: 0.0359
  - Normalized



Feature Coefficients

1. Budget: 2.138
2. Animation: 1.236
3. Comedy: -1.174
4. Release Q1: -0.565
5. Drama: -0.4699
6. Romance: 0.467
7. Sci-Fi: 0.426
8. Release Q4: 0.398
9. Horror: 0.067
10. Actor Experience 0.056
11. Director Experience: 0.042
12. Release Q2: 0.016
13. Producer Experience: 0.01
14. Writer Experience: 0.005

Conclusion

- Model is usable as an initial estimation of profit to drive further research.
- Model is not appropriate as a reliable projection tool.
- Need higher quality predictors and more data to improve the performance of the model.

Future Work

- Experiment with different methods of calculating “experience.”
- Obtain more data and information from other sources to fill in missing data.
- Include Social media and advertising efforts in prediction.
- Franchise information (Sequel? Results from previous movies in the franchise?).
- Experiment with more transforms of numerical data.
- Explore more parameter tuning with Gradient Boosted Trees.
- Look at performance in specific countries.

Thanks :)