



# May the Odds be with you

## Re-Mapping the Formula for Box Office Success

Ed Chin

# The Bottom Line

- Can one maximize the probability of making a high grossing movie **without** starting with a mega budget, highly paid actors and entitled directors?

***"You must unlearn what you have learned" - Yoda***

# Rationale of Initial Feature Selections

- Focus on features that can be controlled or at least influenced by a studio BEFORE the movie is released
- No Critic Ratings. Do not want be at the whim of some opinionated critic
- No User Ratings because they can not be reliably obtained prior to a movie release
- No award information
- No DVD sales

## Classic Features

- MPAA Rating
- Runtime
- # of Theaters
- Seasonality
- Genre
- Studio

## Social Media "Buzz" Features

- Aggregated Wikipedia Pageviews 30 days before release
- # of Wikipedia Edits/Users 365 days prior to release

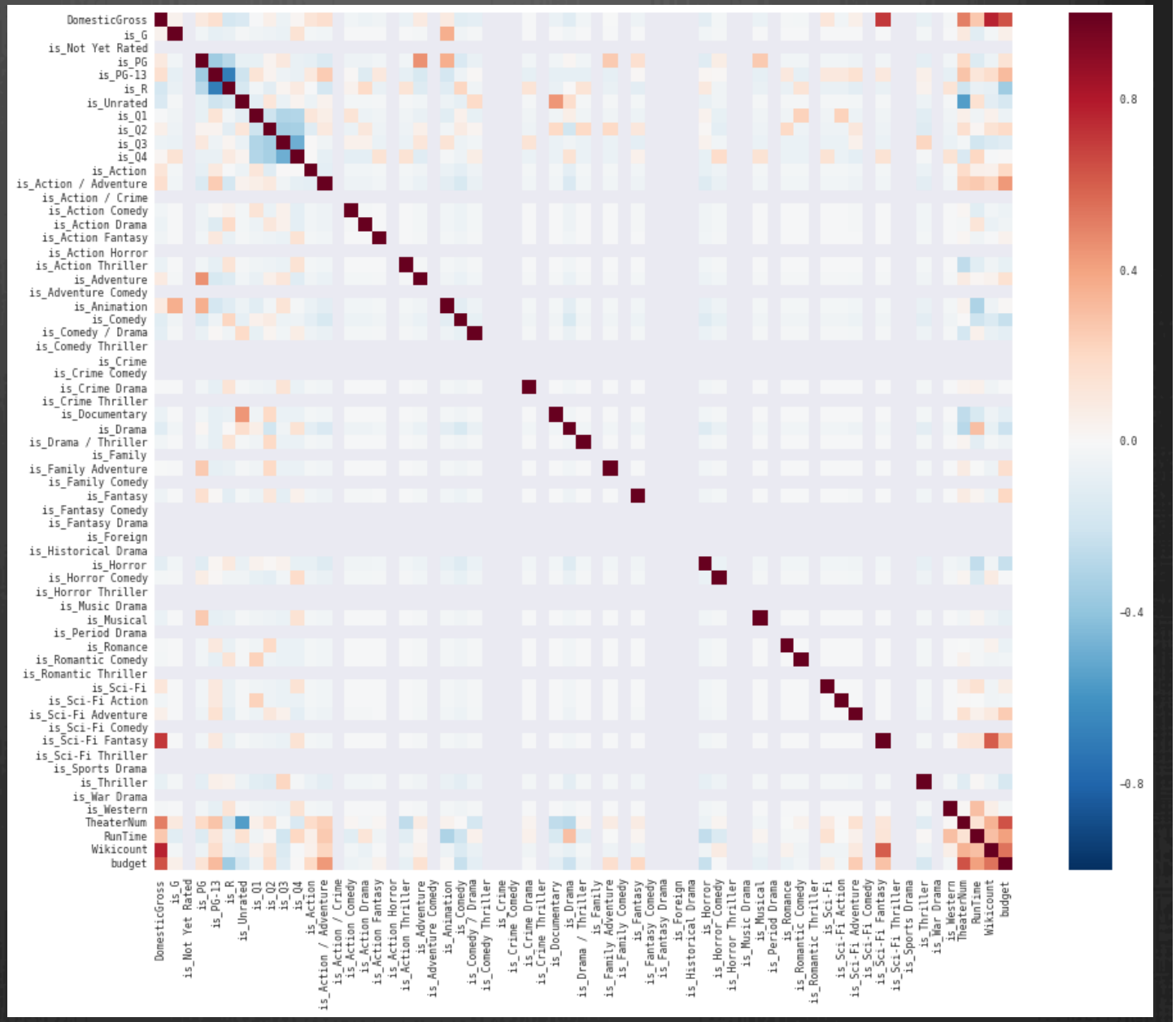
# Features at a glance

## Positive Correlation

- Budget
- Wiki Traffic
- # of Theaters
- Sci-Fi Fantasy

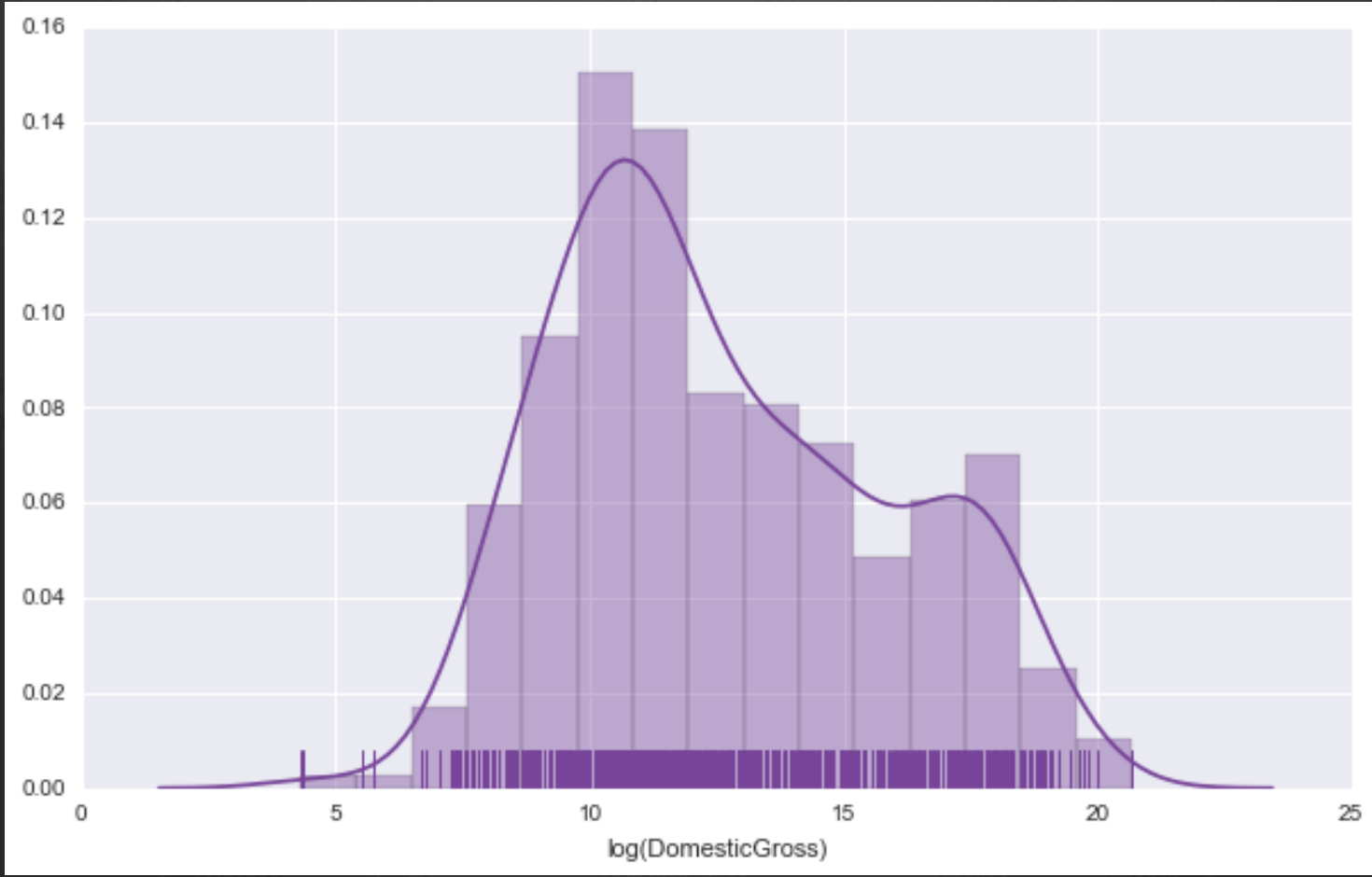
## Negative Correlation

- R rating
- Unrated
- Documentary



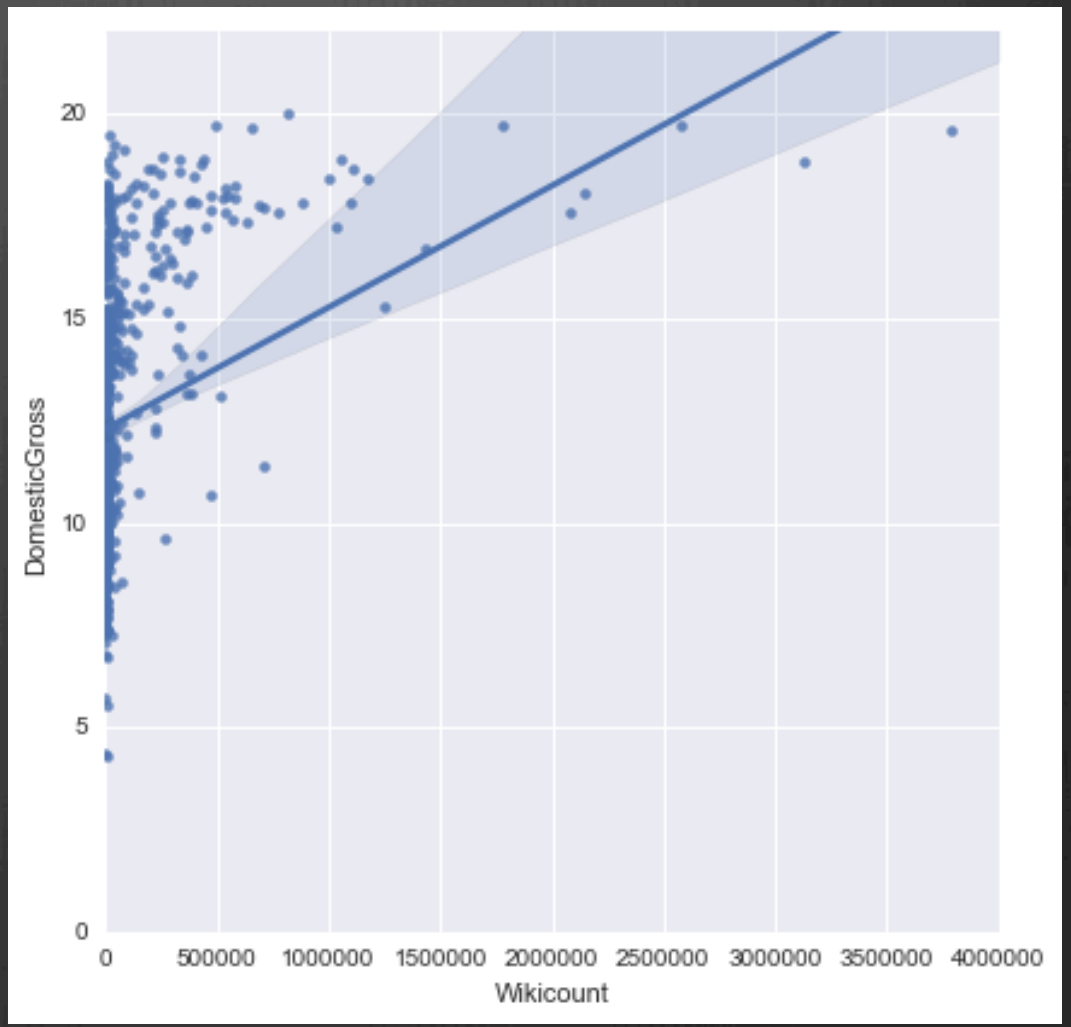
# Observation #1

- Box office exhibits bimodal natures with extreme asymmetric outliers so data has been log-transformed before fitting



# Observation #2

- Usage of Wikipedia traffic as a source of information to detect and predict events in the real world has yet to be properly explored
- While box office success can possibly be achieved without much social media buzz, the chances of a flop decreases dramatically once Wikipedia traffic reaches a certain critical threshold prior to the release



# Observation #3

- A wide theater release and high Wikipedia traffic appear to be common denominators for top grossing movies

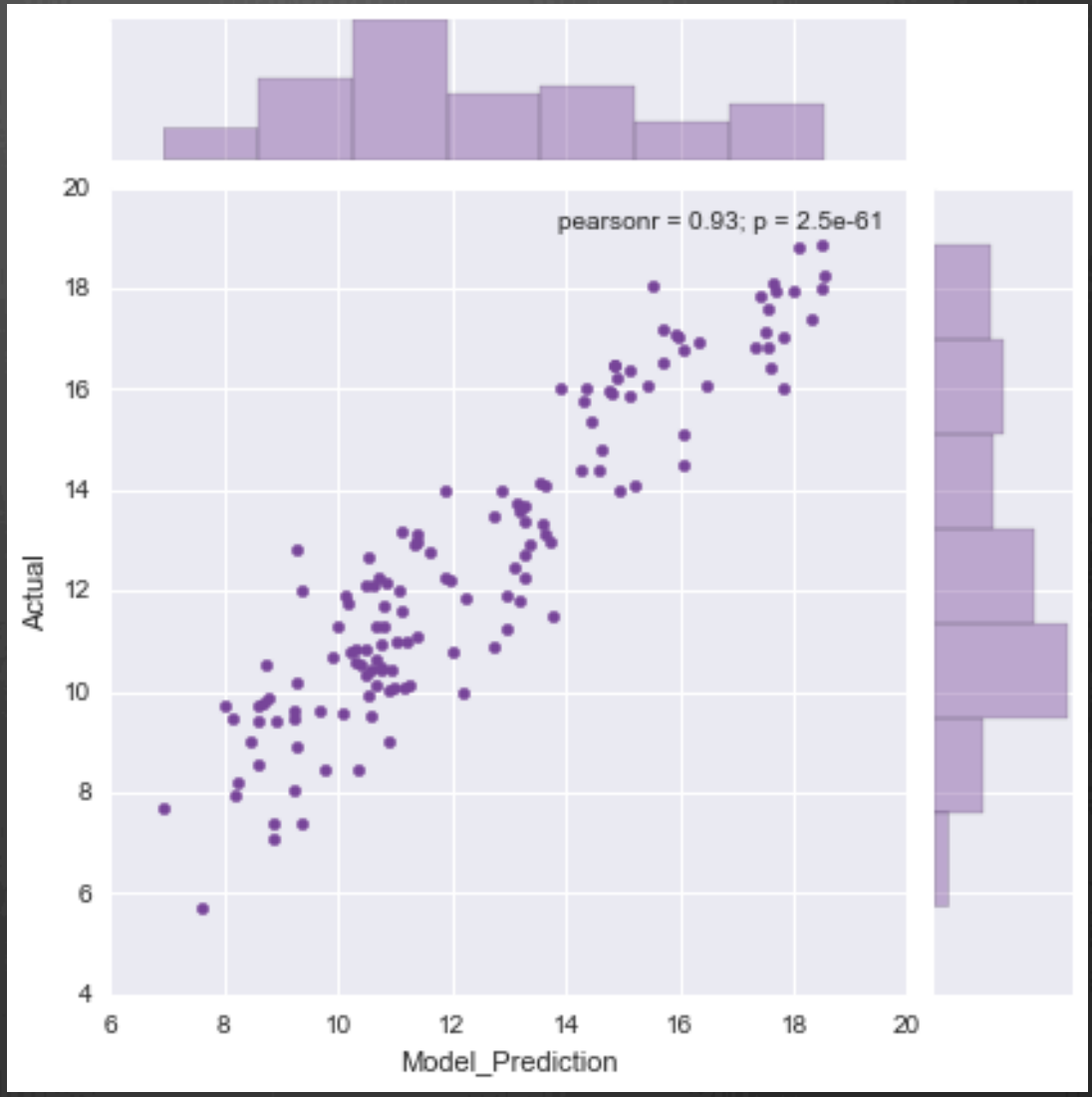
| | All Movies | | Top 50 Gross | |
|---|---|---|---|---|
| | **Mean** | **Median** | **Mean** | **Median** |
| Domestic Gross | $16.2mm | $144k | $166mm | $111mm |
| Wikipedia Views | 101,000 | 1,928 | 754,974 | 291,210 |
| No. of Theater | 675 | 27 | 3,629 | 3,745 |
| Runtime | 106 | 102 | 115 | 111 |

# Procedures

- Data were obtained on 695 domestic movies from Aug 2015 through Aug 2016
- Train the dataset (80% of data) across 9 models using cross validation techniques beginning with all features
- PCA and Lazzo were used to select the primary feature components that drive box office performance
- The model with the best mean score was then selected and further optimized to minimize MSE
- Validate final model against the test set (20% of data)

# Model Results

- Model explains ~90% of the test sample
- Traditional linear regression models have fared poorly
- Ensemble learning methods that can approximate non-linear relationships have scored in the 86-90% range

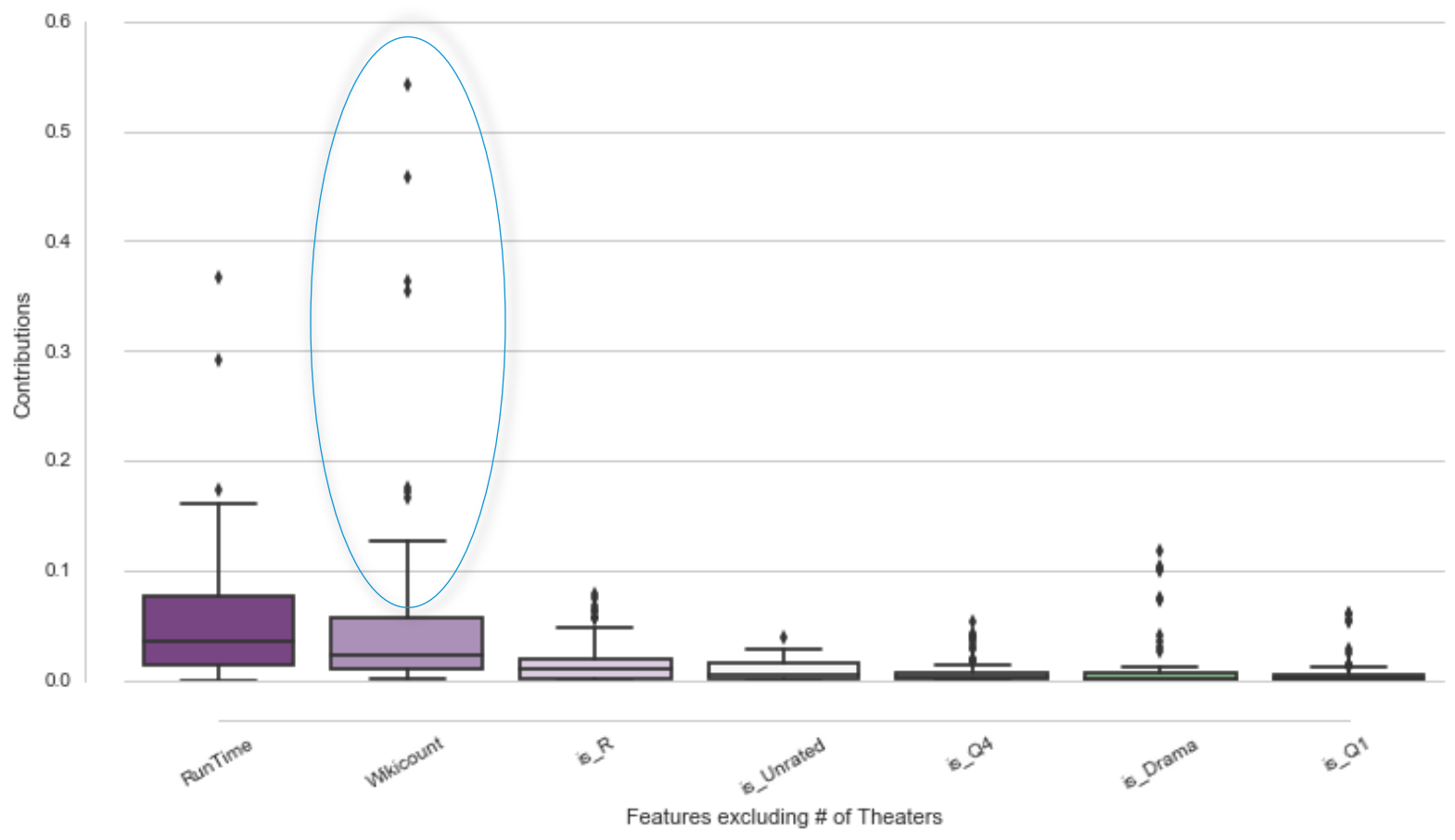


# Why does the model fit so well?

- Theater count remains the most important feature in predicting the bulk of the box office performance
- However, theater count does a poor job in estimating the tail
- In other words, theater count is a necessary condition for box office success, but it does not explain the bulk of the variance observed in the higher end of the box office range
- For example, 3500+ theaters are needed to have an expected $50+mm box office or higher. After that, what makes a movie gross $50mm vs $150mm is not something that can be fully explained by theater count
- Social media traffic, and in this case Wikipedia traffic, is the missing link that does a fantastic job in approximating the tail

# Diving into Feature Importances

- Traditional interpretation of feature importance in models like random forest often rank its components by their average absolute contributions
- That is not a bad idea, but it is a poor construct when it comes to understanding box office performance. One should focus on the tail contributions instead



# Recommendations

- Build a movie franchise on a trendy subject matter (popular book, video game, comic book, franchise etc) that has pre-existing social media following.
- Alternatively, allocate resources to build social media "buzz". For a moderate budget movie, aim to have 400k Wikipedia pageviews in the month leading up to the release. For a blockbuster release, aim to have at least 1+mm traffic count
- Focus on comedy, animation and/or drama
- Release movies in the summer, avoid Q1 release
- Avoid documentary, R-rated and unrated films

# Appendix

## Python Packages

- Sklearn
- Pickle
- Tree Intrepreter
- Seaborn

## Tools

- Scrapy
- Selenium

## Website Visited

- Box Office Mojo
- Wiki PageView API
- IMDB
- RottenTomatoes
- Metacritic

## Multi-Linear Regression

- Classic OLS
- Ridge
- Lasso
- ElasticNet

## Supervised Machine Learning

- Decision Trees
- Extra Trees
- Adaptive Boosting
- Random Forest
- Gradient Boosting

Ed Chin
echin6@gmail.com

*"Always pass on what you have learned"* - Yoda