Predicting ROI for Hollywood Movies

Benison Pang

Question of interest

- Can we predict a movie's return on investment?
- In other words, what features of a movie are most important in predicting the success of a movie?

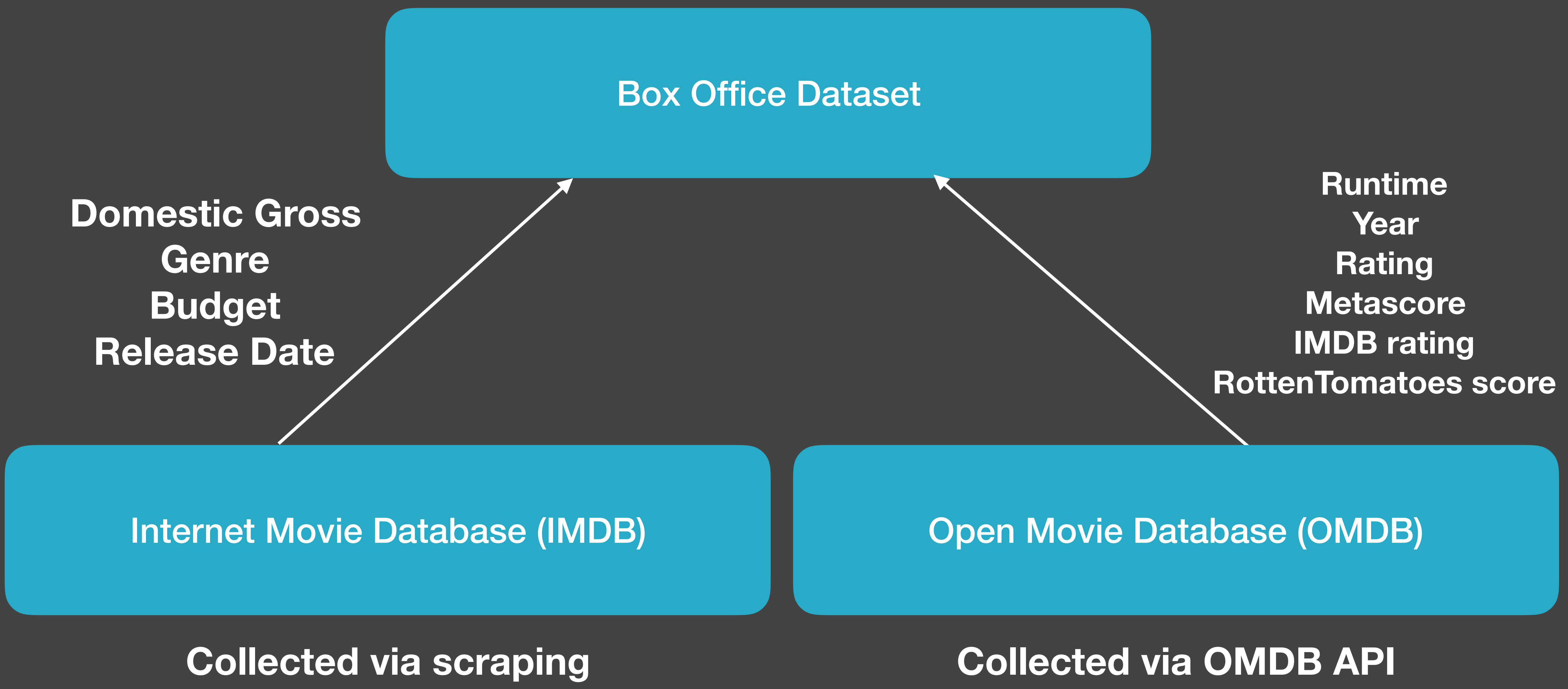
Some considerations for Movie Production

- Budget - The larger the budget, the higher the risk
- Genre - Some genres may be more profitable than others
- Runtime - The shorter the movie, the more screenings can be fit into a single day
- Critic and Audience Scores e.g. Rotten Tomatoes, IMDB, Metacritic

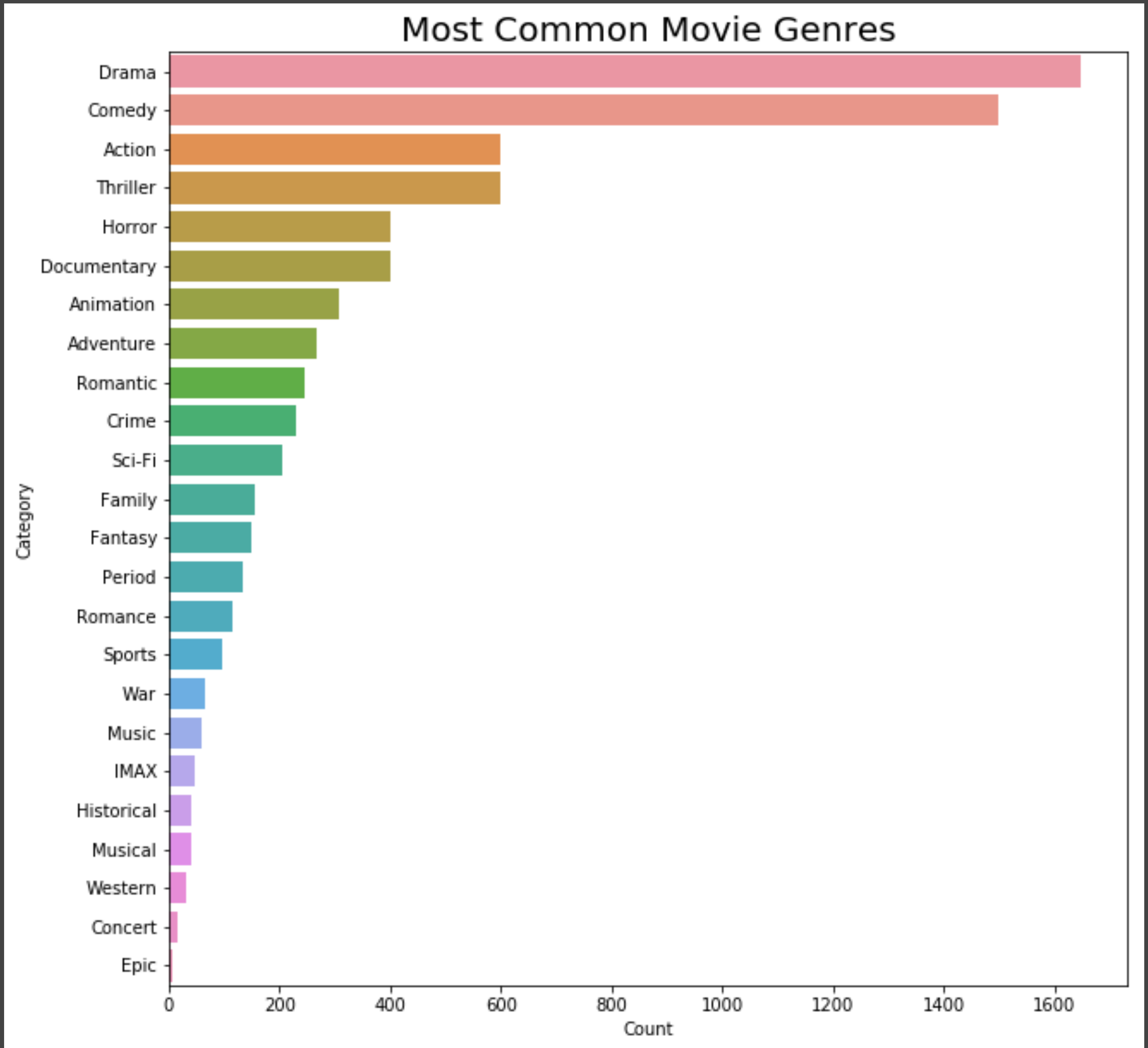
Approach

- Data Collection via web scraping and API
  - IMDB and OMDB (Open Movie Database)
- Exploratory Data Analysis
- Inferential Statistics (based on questions from EDA)
- Linear - based model
  - Extract feature coefficients to determine directionality
- Tree-Based Model (Random Forest vs XGBoost)
  - Extract feature importances

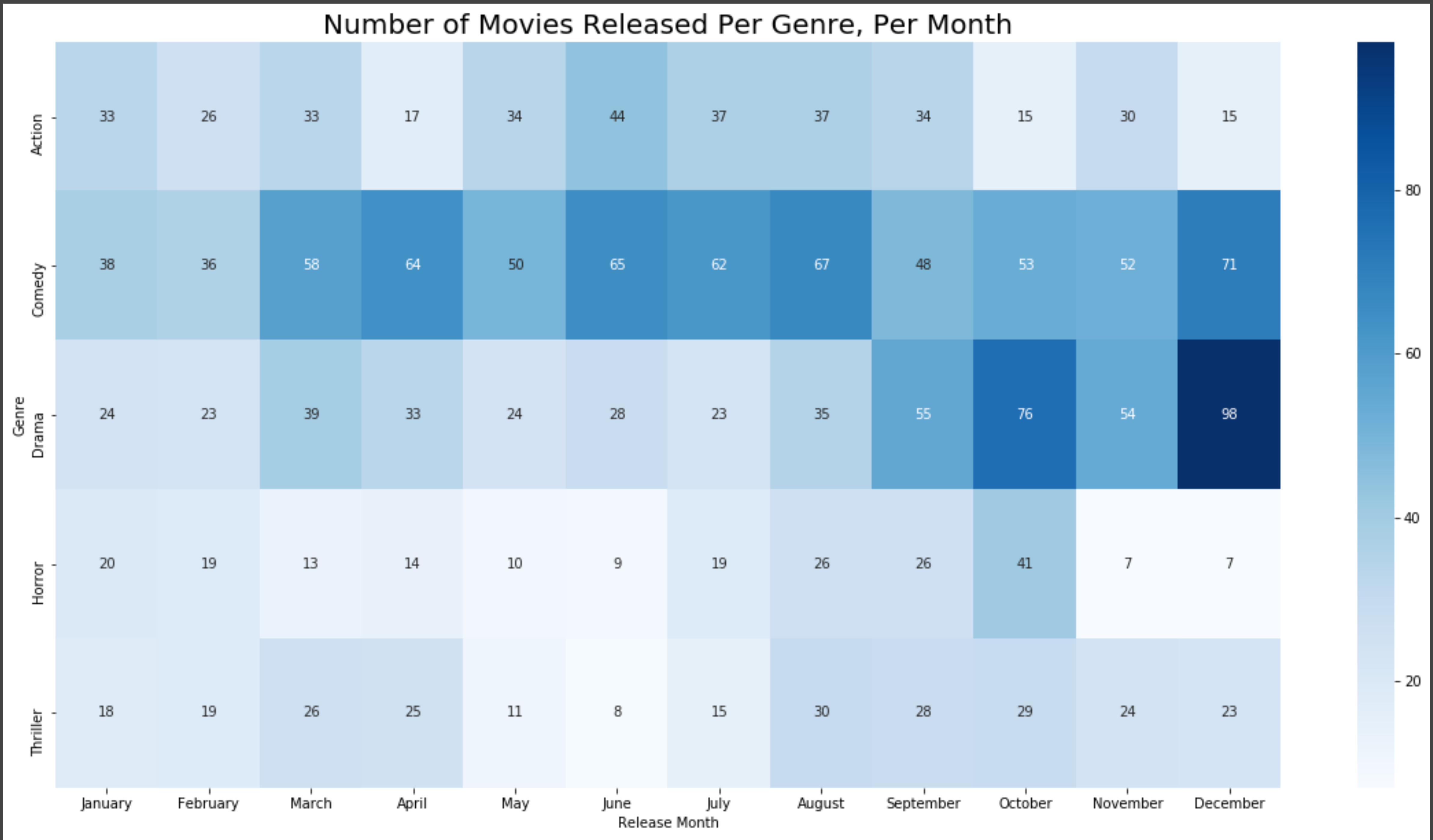
Data Collection



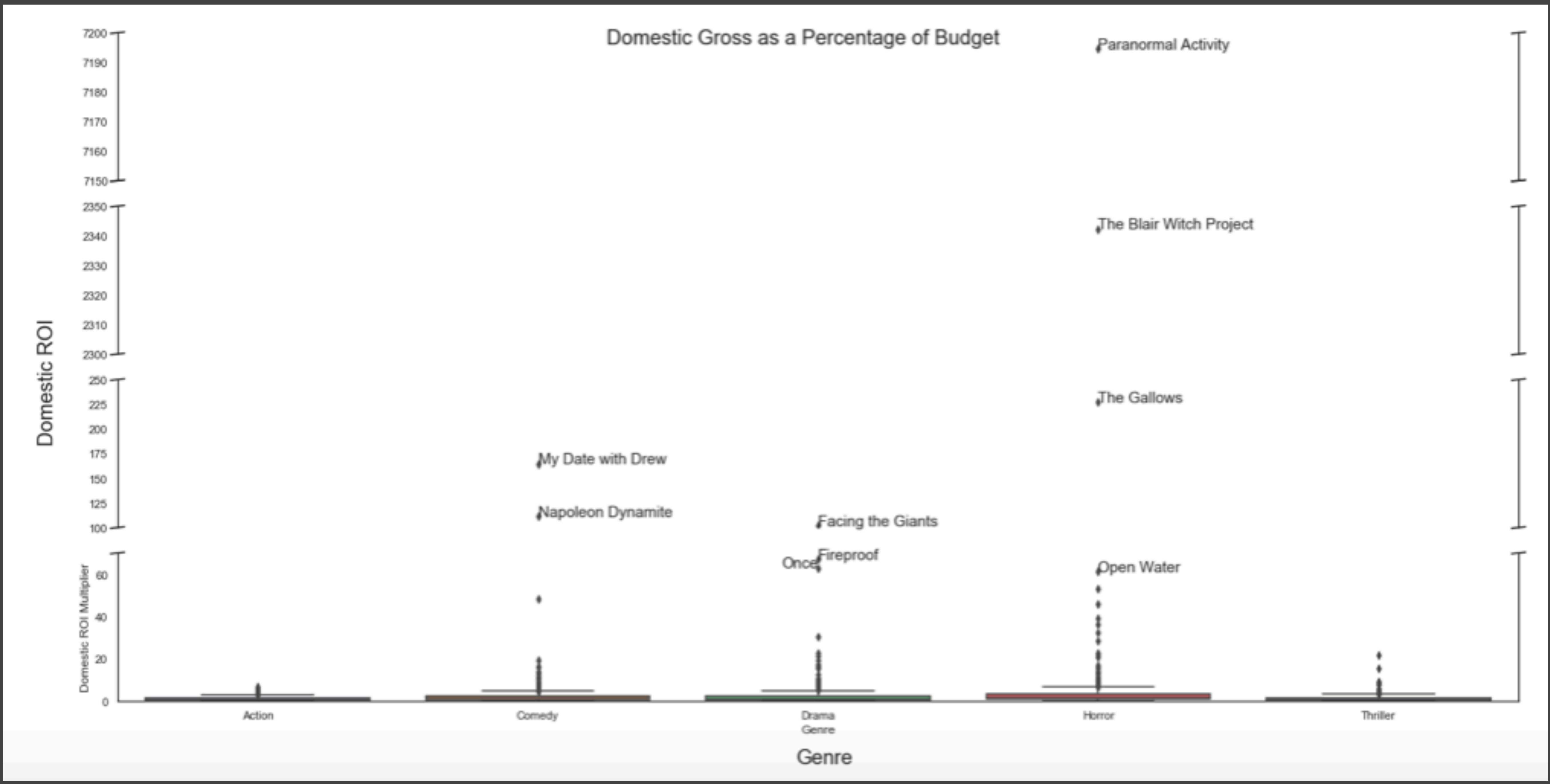
Which movies are most commonly made?



EDA - When are genre movies released?



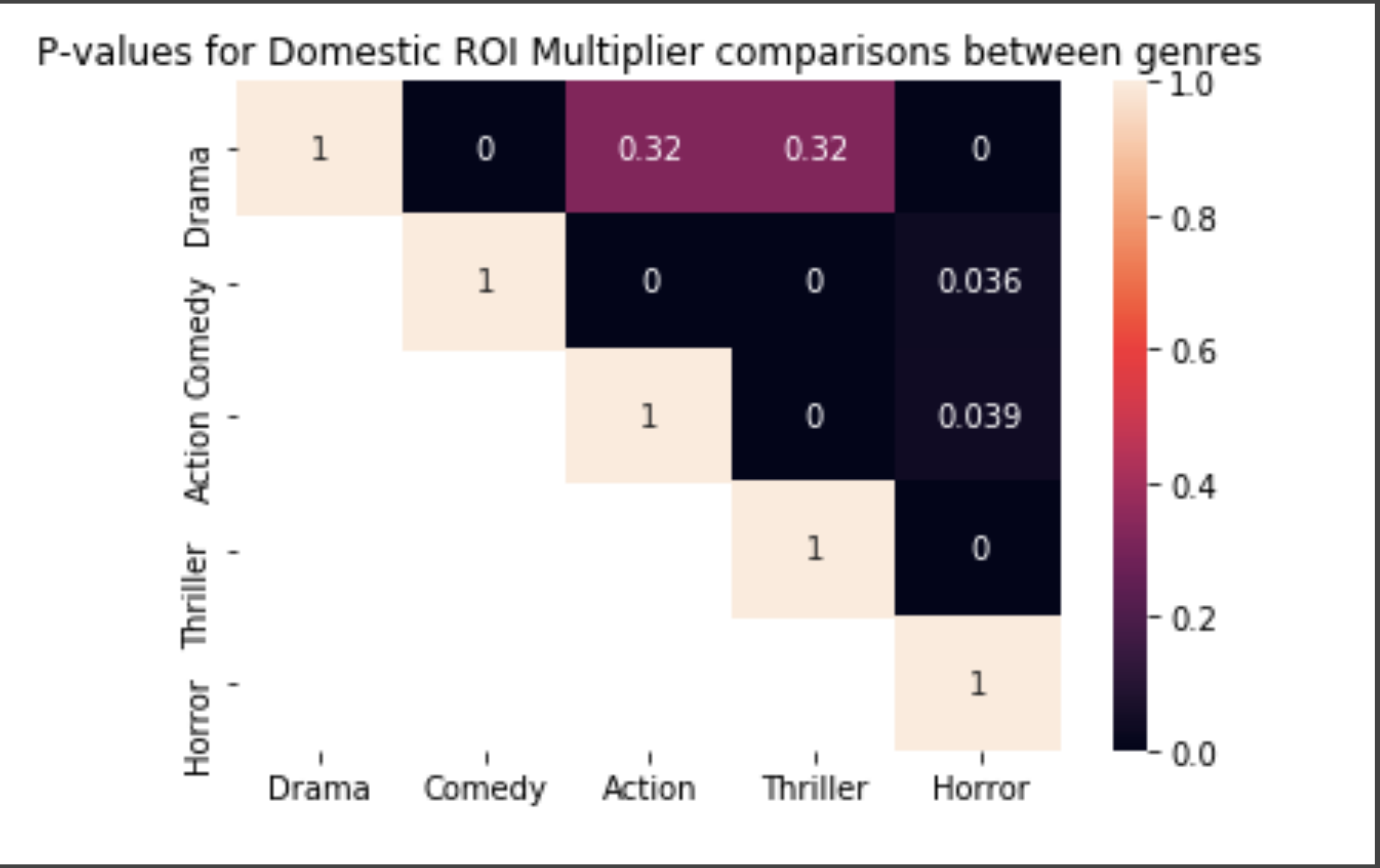
EDA - ROI by Genre



Statistics of Genre Finances

Genre	Budget (millions)	Profit (millions)	Domestic ROI Multiplier
Action	70.00	-0.989502	0.970905
Comedy	30.00	6.868437	1.296492
Drama	22.25	-0.701991	0.943328
Horror	20.00	10.490808	1.498722
Thriller	35.00	-1.841102	0.916314

Statistics of Genre Finances



Preprocessing - Standardization

- Standardization and scaling required to account for differences in ranges of numerical features
  - Standard Scaler to center on 0 with variance 1

Finding the best models - step by step

- Train-test split of 70/30, cross-validated grid search approach to tune hyperparameters.
- Compare model performance against each other using model evaluation metric
- Evaluation: Mean Absolute Error (MAE) vs Root Mean Squared Error
  - Used MAE in this project
  - The lower the error, the better

Models assessed

- Naive Model
- Linear Regression model
  - Elastic Net
- Tree-Based models
  - XGBoost (Gradient boosting)
  - Random Forest

Naive Model

- Simple model that always predicts the output to be the median of the ROI from the data
- Serves as a baseline we can compare to
- Worst performer, unsurprisingly

Elastic Net

- Chosen as the best current linear regression approach, compared to ridge and lasso regression.

- Benefit of a linear model: Feature coefficients allow us to assess directionality

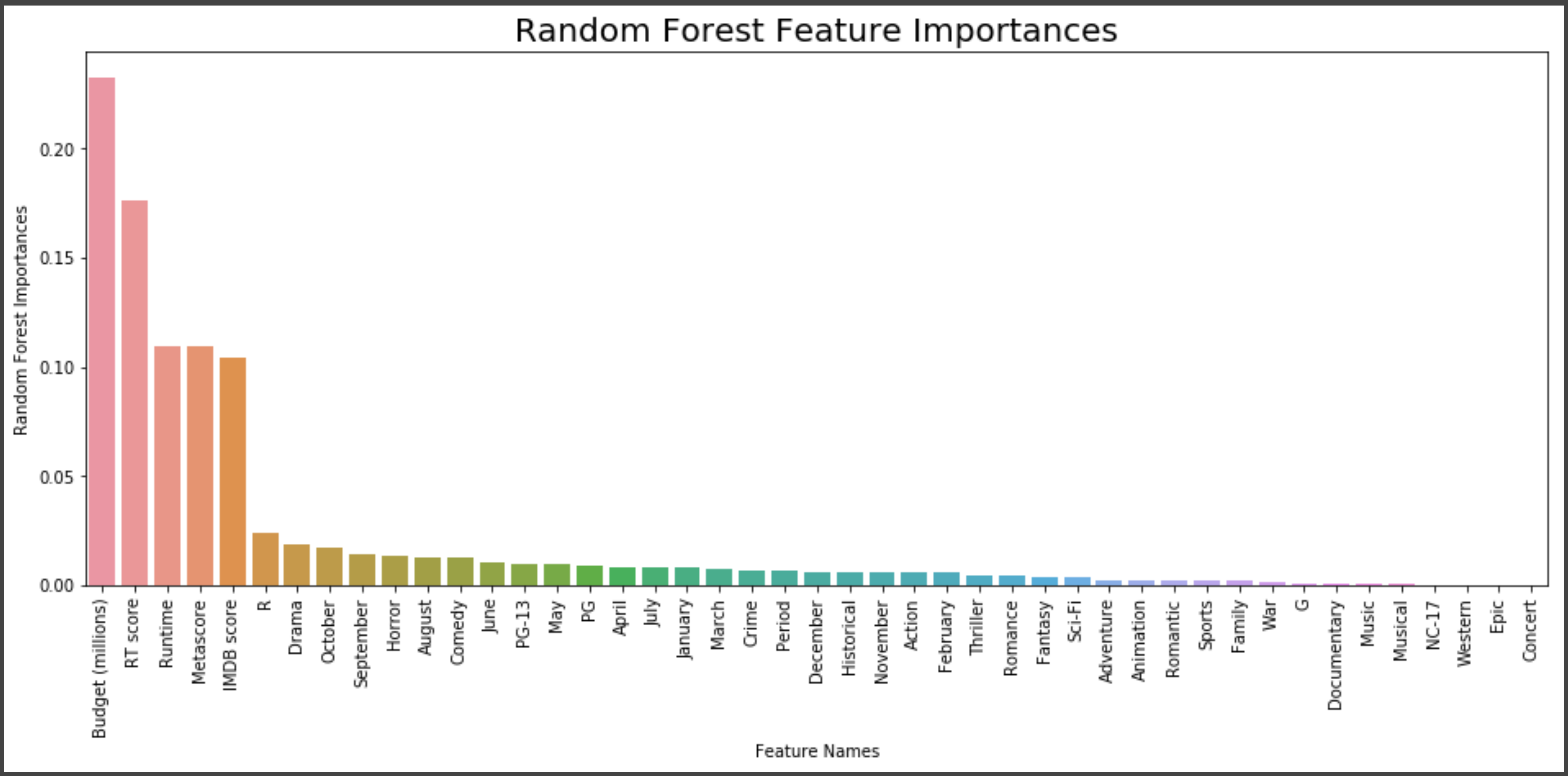
Random Forests

- Ensemble approach that applies the concept of ‘wisdom of the crowd’
- Problem: not very interpretable, prone to overfitting
- Benefit: Feature importances!

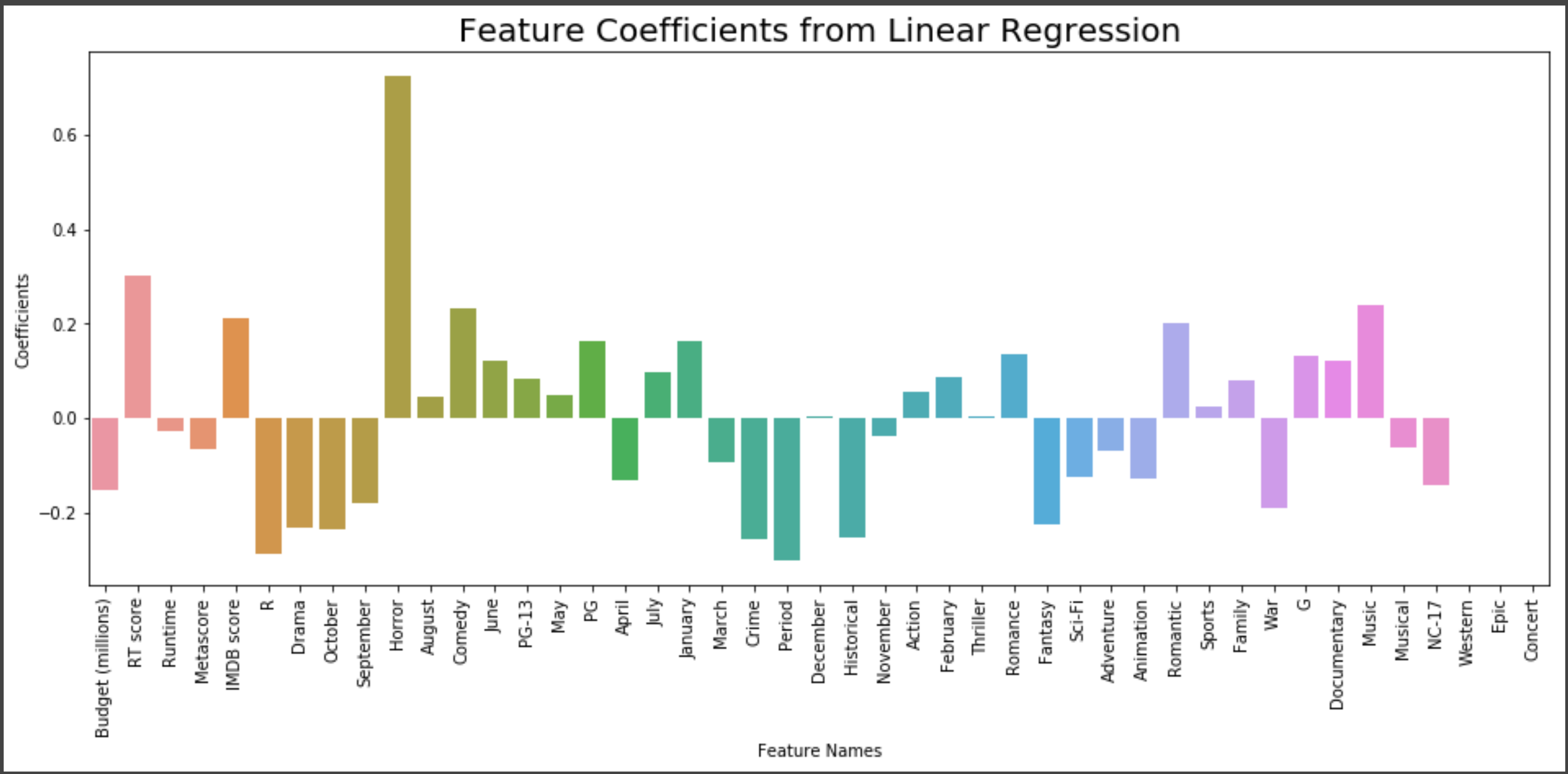
XGBoost

- Also a tree-based approach, fundamentally rooted in gradient boosting
- Problem: Highly prone to overfitting, difficult to train
- Benefit: Potentially give better results if tuned carefully

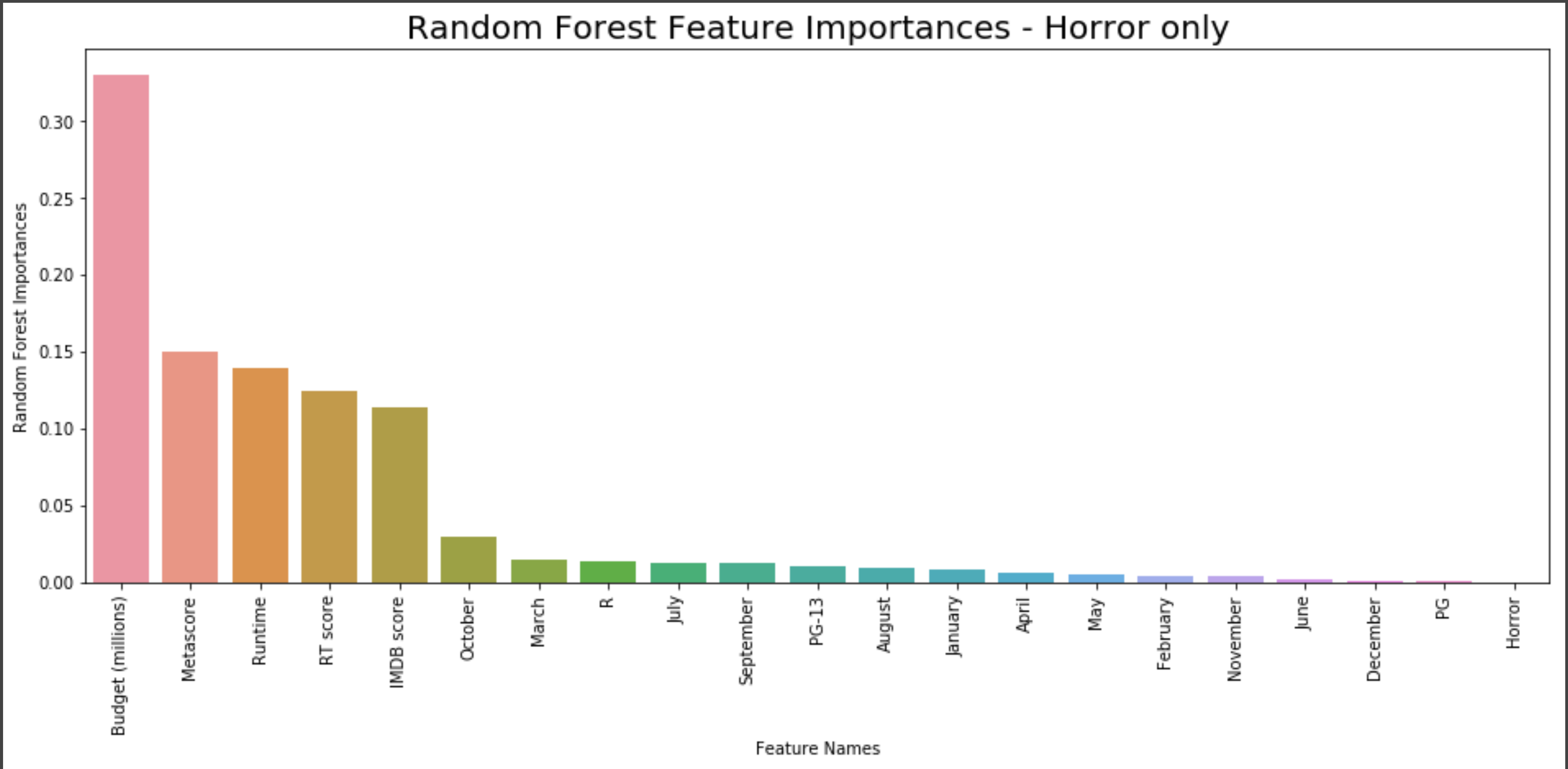
Feature Importances - All Movies



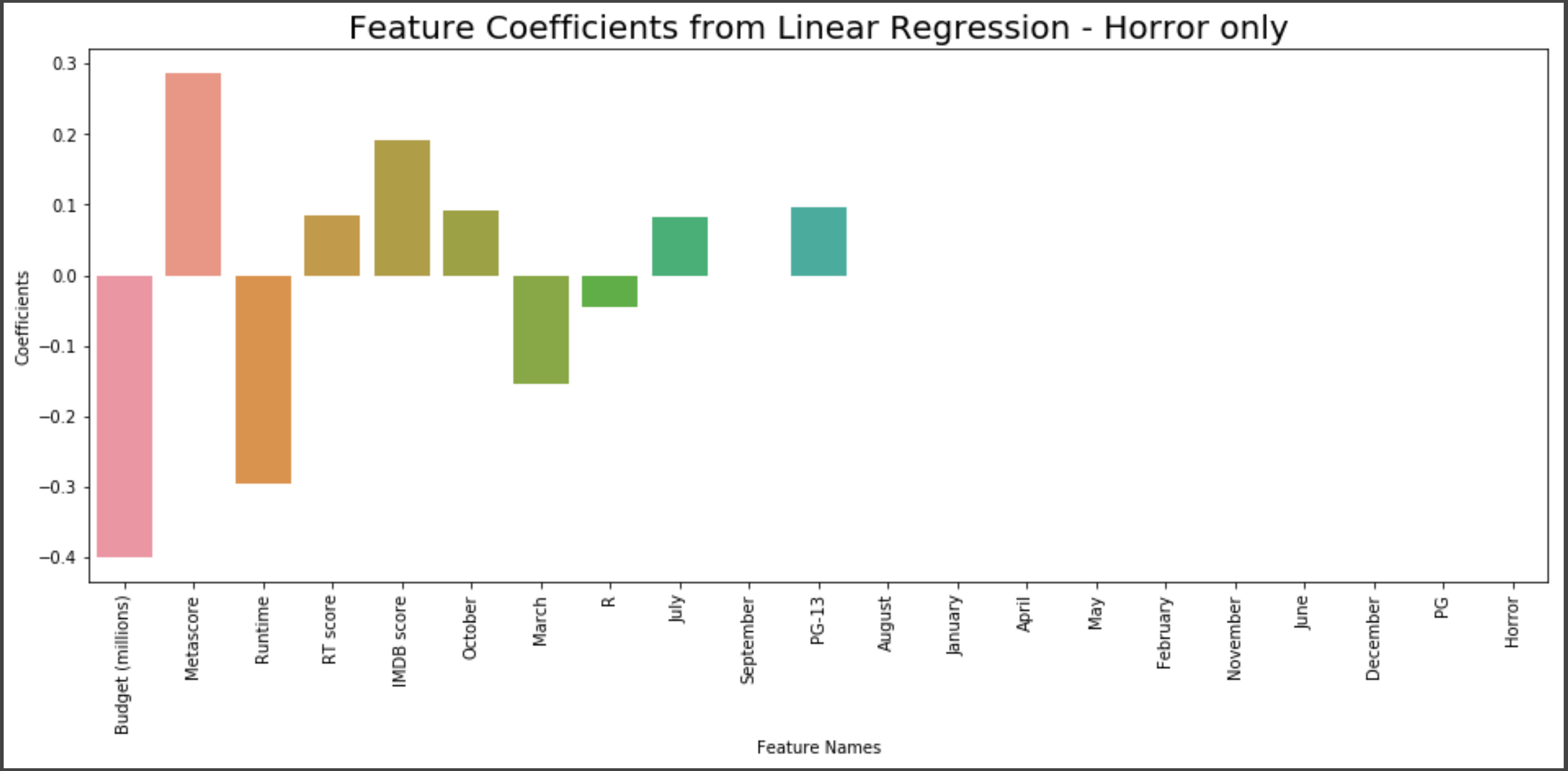
Feature Directionality - All Movies



Feature Importances - Horror



Feature Directionality - Horror



Actionable Insights

- Focus on making shorter, low-budget movies in the horror genre in order to maximize ROI
- Emphasize audience enjoyment above making a cinematically proficient movie
  - For horror, do both!
- Release horror movies in October as well as July