IMDb Analysis

•••

Ajay Patel, Andrew Mashhadi, Daniel Kwon, Sofia Alcazar, Dylan Jorling

STATS 405

June 7, 2022

Introduction

- Our goal is to help guide an executive producer through what it takes to create an iconic movie through the analysis of historical IMDb data.
- We answer the question of what traits and features are correlated with a successful movie by breaking down the goal of making a movie into three different perspectives:
  - Model and maximize box office profits
  - Model and maximize chances of an Academy Award nomination
  - Finding movies that are similar to a specified existing film

All Things Data

Data Collection

- Used IMDb website's HTML and Xpath to collect unique movie identifiers
- Used these identifiers with IMDb-API to collect JSON files with information for each unique movie

Data Cleaning

- Minimal NAs in data, JSON API had clean data
- Extracted additional variables from string information.
- Used relative frequencies in dataset or records for other variables

Data Storage

- Extracted data fields from JSON files and stored as records in AWS database tables
- Collaborated with read-only privileges on AWS database
- Used GIT and GITHUB for code collaboration

Variable Overview and Exploratory Data Analysis

Variables Overview and EDA

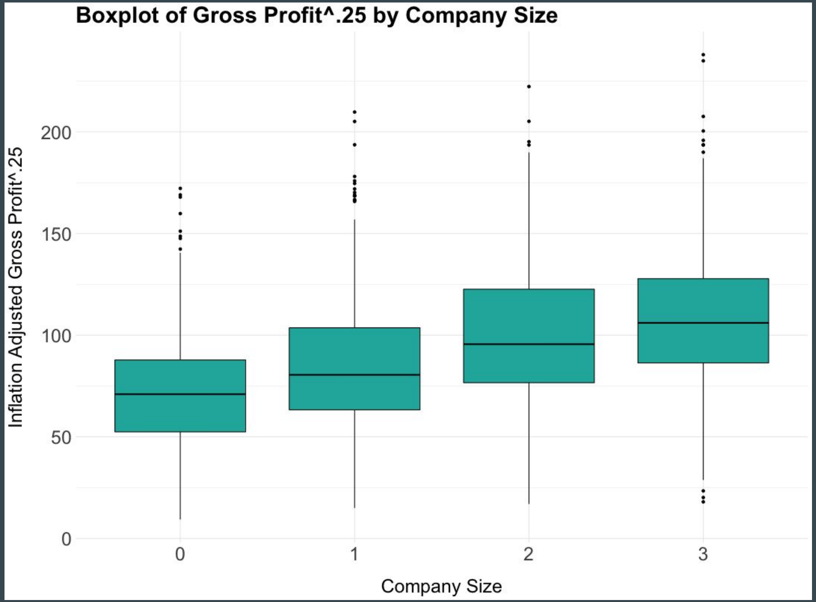
Response Variables:

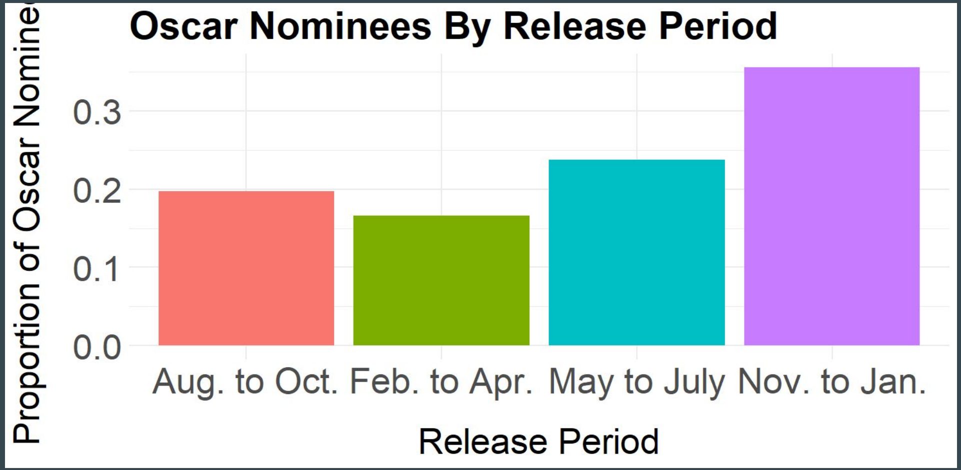
- Oscar Nomination
- Gross Profit

Predictor Variables:

- Runtime
- Genre
- Rating
- Language
- Star Power
- Writer Popularity
- Director Popularity
- Company Size
- Release Period
- Inflation Adjusted Budget

Variables Overview and EDA





Methodology

Methodology: Modeling Box Office Profits

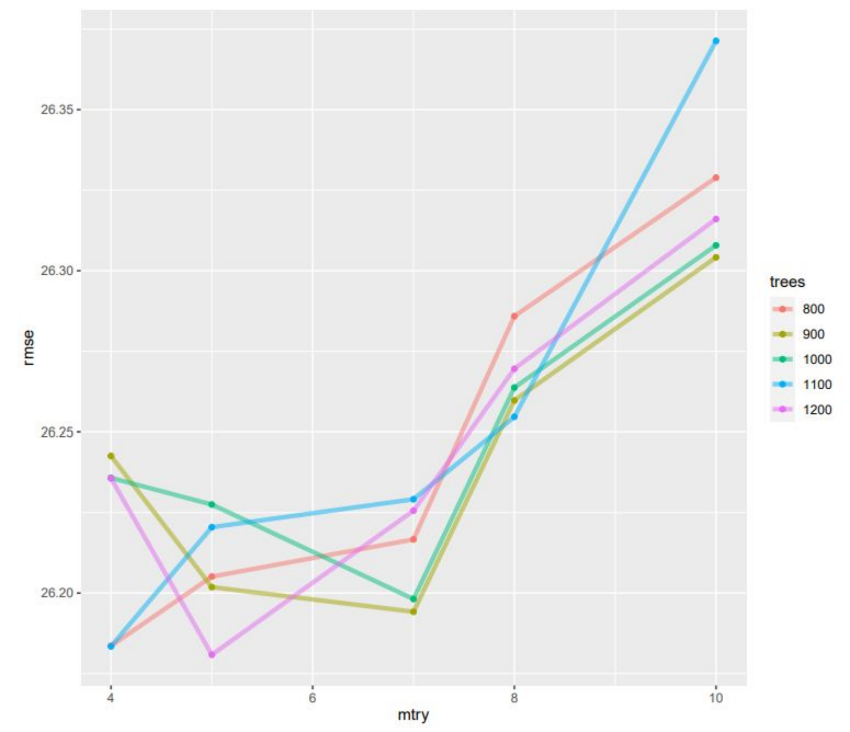
Random Forest (Regression)

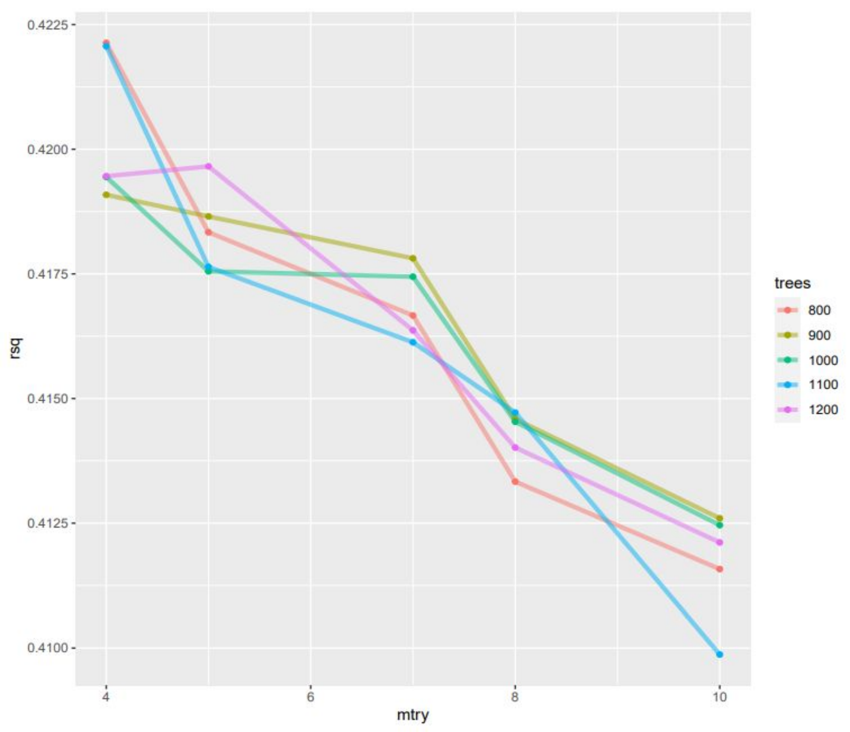
- Leverages resampling techniques to fit model using numeric *and* categorical variables
- Tuned parameters to minimize RMSE and maximize R-Squared without overfitting
- Variable Importance scores using *Permutation Importance Method* (using R-Squared)

Linear Regression Model

- Fit standard model with mostly numeric variables
- Interpretability of coefficients can be used to support our random forest's key insights
- Minimizes RSS with Ordinary Least Squares

Box Office Profit Model Tuning - Random Forest





Methodology: Modeling Box Office Profits

Random Forest (Regression)

- Leverages resampling techniques to fit model using numeric *and* categorical variables
- Tuned parameters to minimize RMSE and maximize R-Squared without overfitting
- Variable Importance scores using *Permutation Importance Method* (using R-Squared)

Linear Regression Model

- Fit standard model with mostly numeric variables
- Interpretability of coefficients can be used to support our random forest's key insights
- Minimizes RSS with Ordinary Least Squares

Methodology: Modeling Oscar Nomination

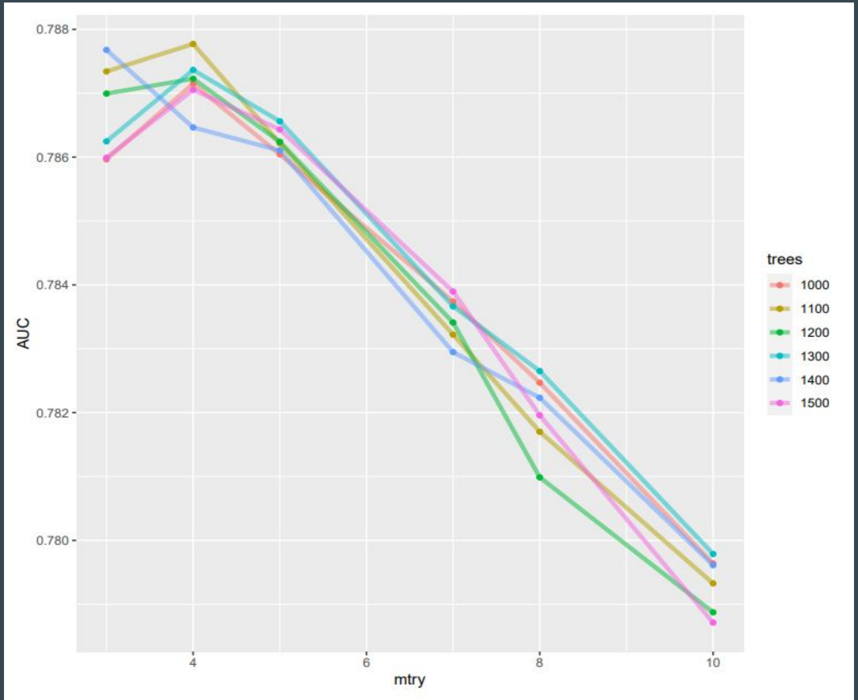
Random Forest (Classifier)

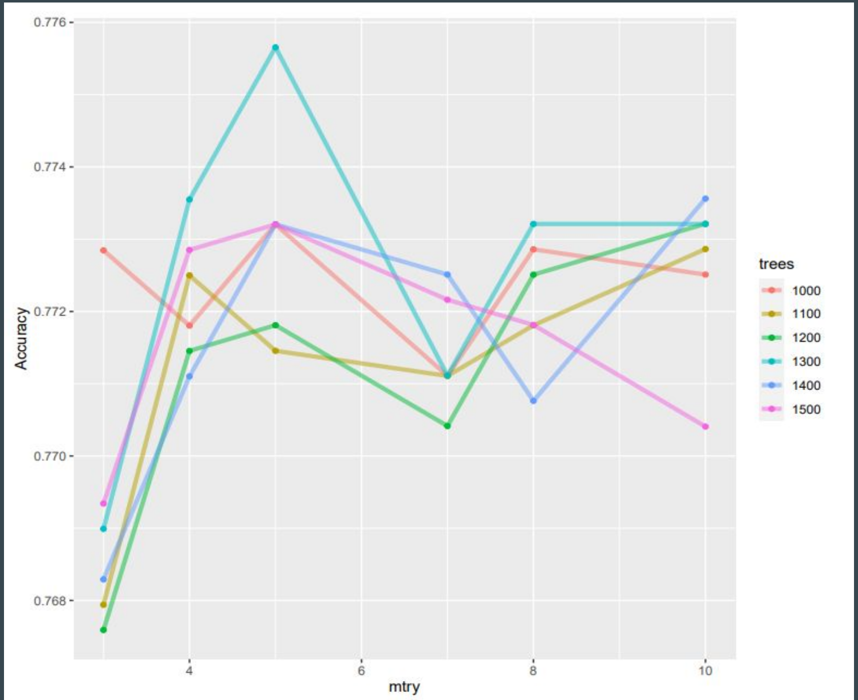
- Leverages resampling techniques to fit model using numeric *and* categorical variables
- Tuned parameters to maximize accuracy and area under the ROC curve, without overfitting
- Variable Importance scores using *Permutation Importance Method* (using accuracy)

Logistic Regression Model

- Supplements the RF model for a variety of reasons, such as interpretability of the coefficients
- Used regularization to shrink non-important variables' coefficients close to 0
- Tuned the penalty term to maximize area under ROC curve

Oscar Nomination Model Tuning - Random Forest





Methodology: Modeling Oscar Nomination

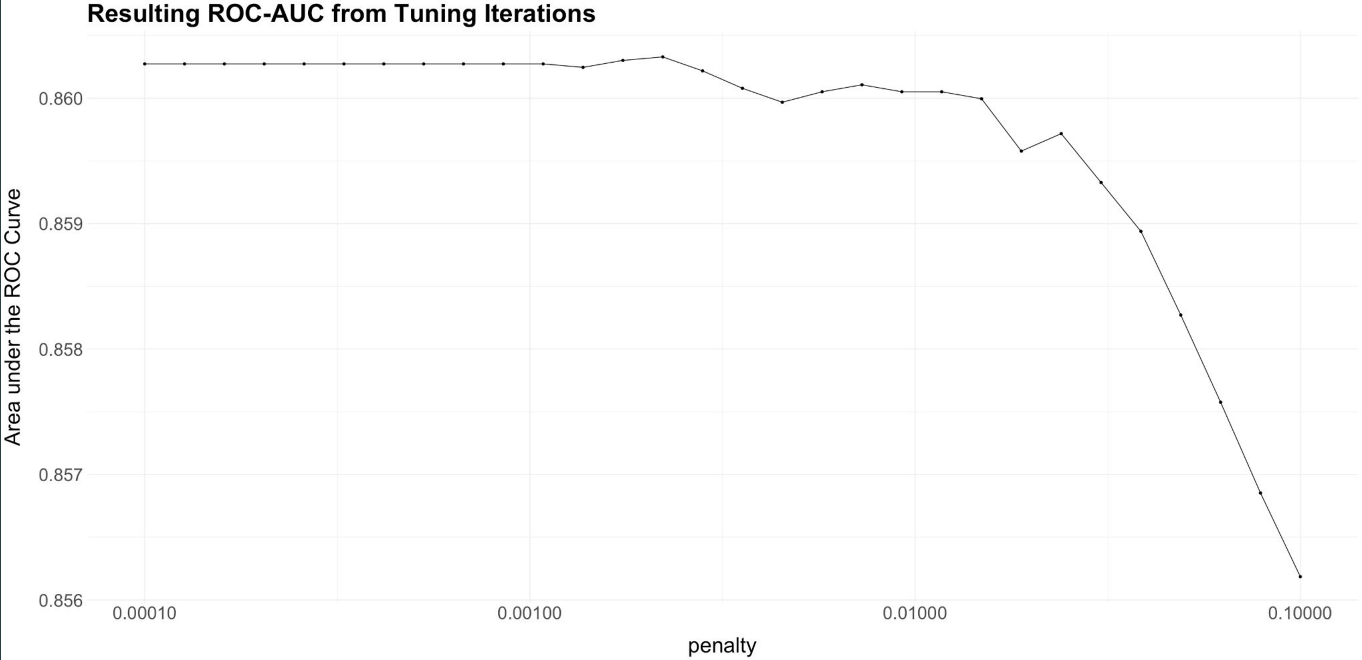
Random Forest (Classifier)

- Leverages resampling techniques to fit model using numeric *and* categorical variables
- Tuned parameters to maximize accuracy and area under the ROC curve, without overfitting
- Variable Importance scores using *Permutation Importance Method* (using accuracy)

Logistic Regression Model

- Supplements the RF model for a variety of reasons, such as interpretability of the coefficients
- Used regularization to shrink non-important variables' coefficients close to 0
- Tuned the penalty term to maximize area under ROC curve

Oscar Nomination Model Tuning - Logistic Regression



Methodology: Modeling Oscar Nomination

The performance metrics we used to compare these two binary classifiers were:

- Overall classification accuracy
- Area under the ROC and Precision/Recall curve
- Precision

$$\text{precision} = \frac{\text{true Oscar nominees}}{\text{predicted Oscar nominees}}$$

- Recall

$$\text{recall} = \frac{\text{true Oscar nominees}}{\text{actual Oscar nominees}}$$

Methodology: Finding Similar Movies

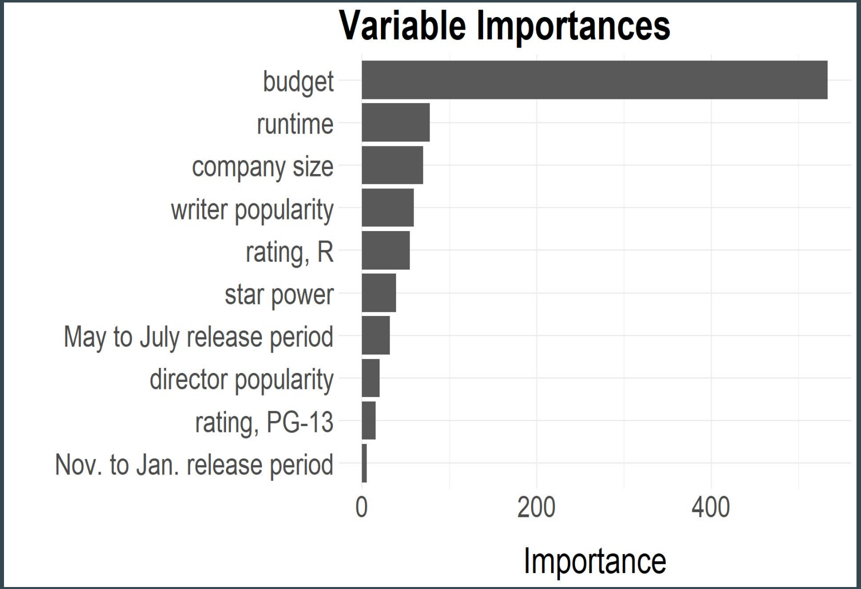
- Used a K-Nearest Neighbors algorithm to find the K-most similar movies to a given movie
- Used both numeric and categorical variables
    - Dummy coded the categorical variables
- Standardized the data and used Euclidean distance to find the K-nearest neighbors

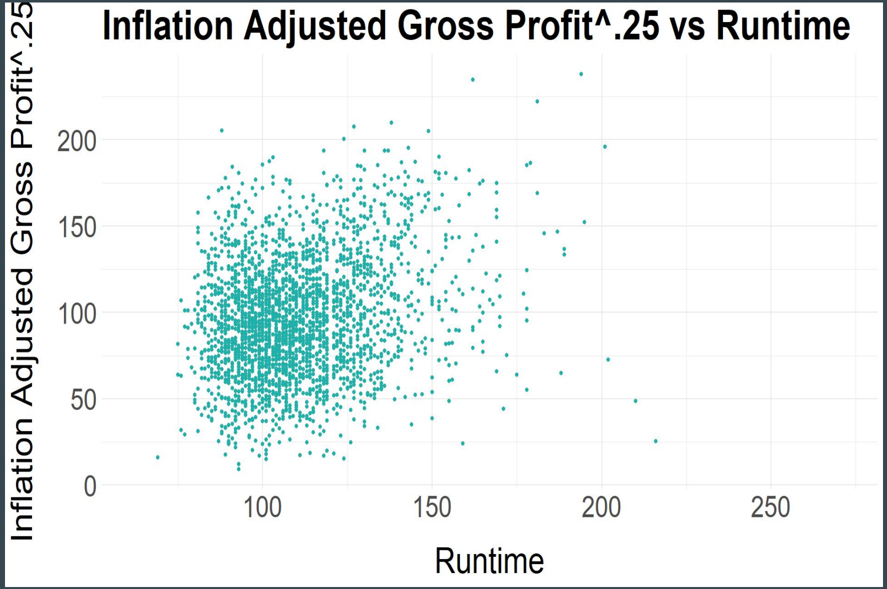
Results and Discussion

Results and Discussion: Box Office Profits

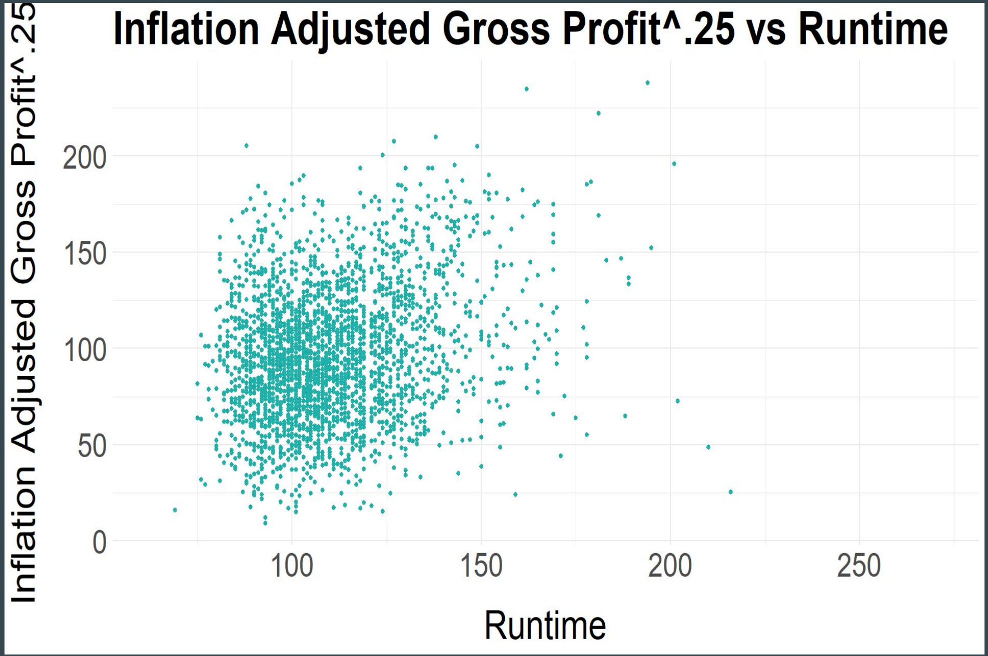
- Tuned our random forest hyperparameters using 10-Fold Cross Validation
- Found the following to minimize RMSE while maintaining a high R-Squared:
  - 1100 trees (bootstrap resamples)
  - 7 variables randomly sampled as candidates for each split
- With optimal parameters, the random forest achieved an approximate RMSE 429,496.70 U.S. dollars and an R-Squared of about 0.457 on test data
- Note that the reported RMSE is less than 0.2% of the standard deviation in our box office profit data

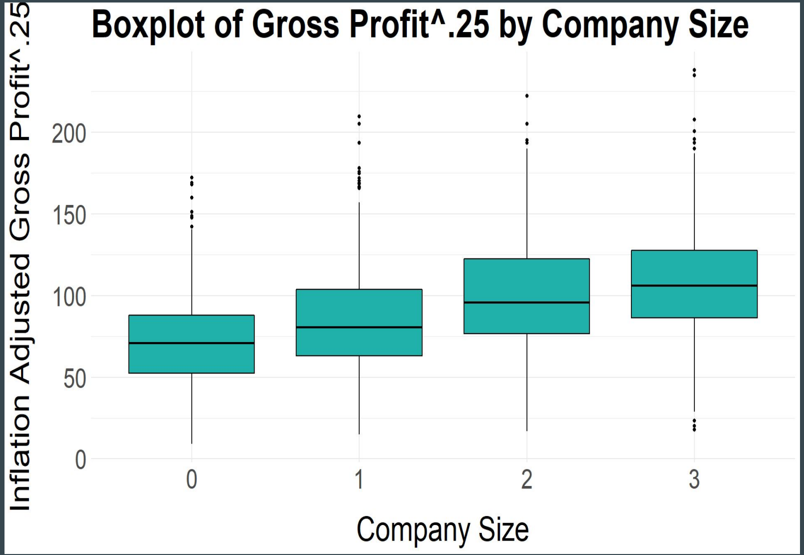
Results and Discussion: Box Office Profits – RF



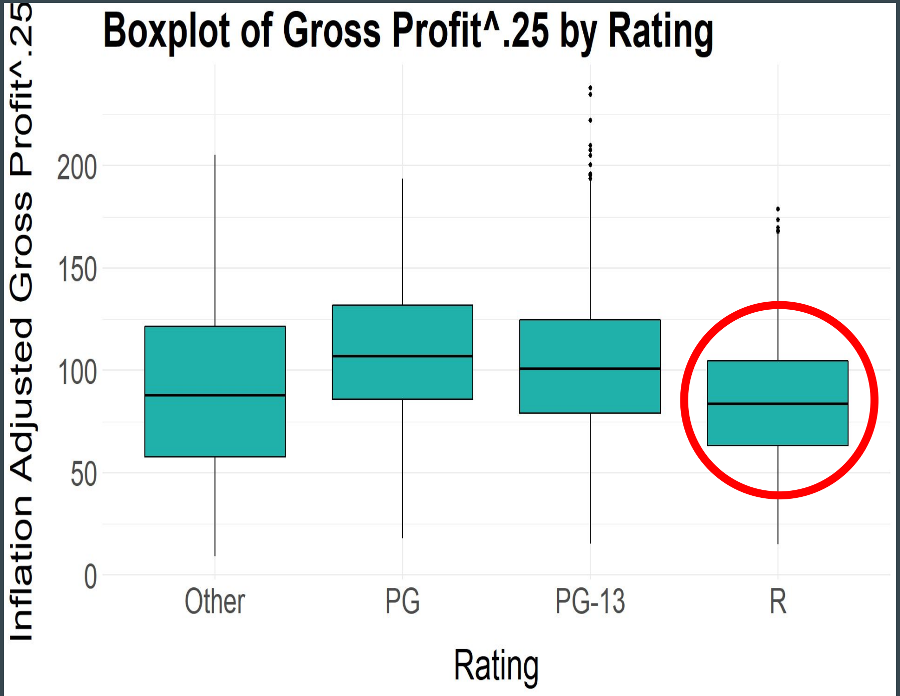


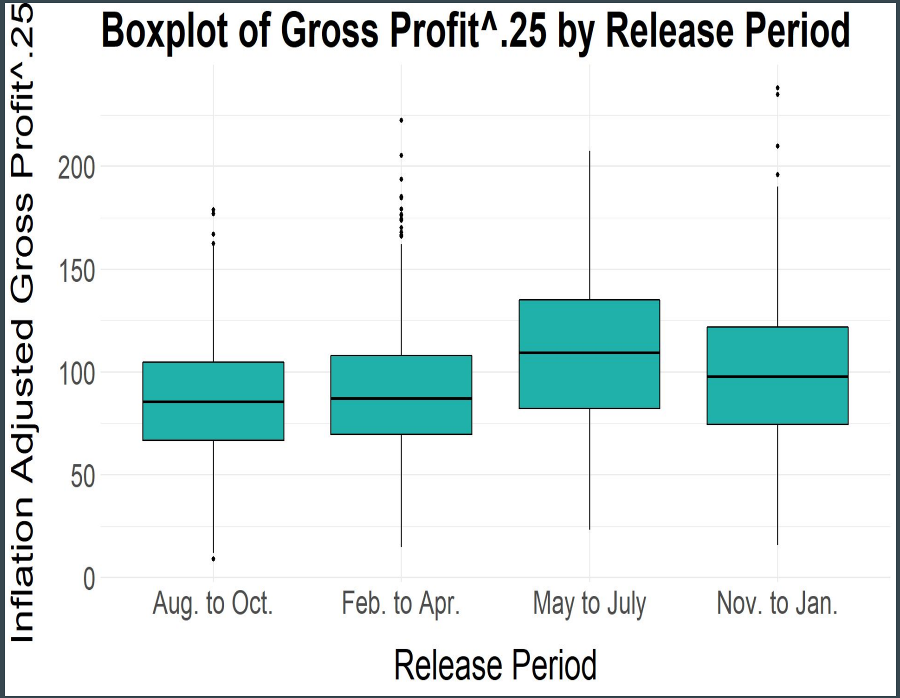
Results and Discussion: Box Office Profits – RF





Results and Discussion: Box Office Profits – RF





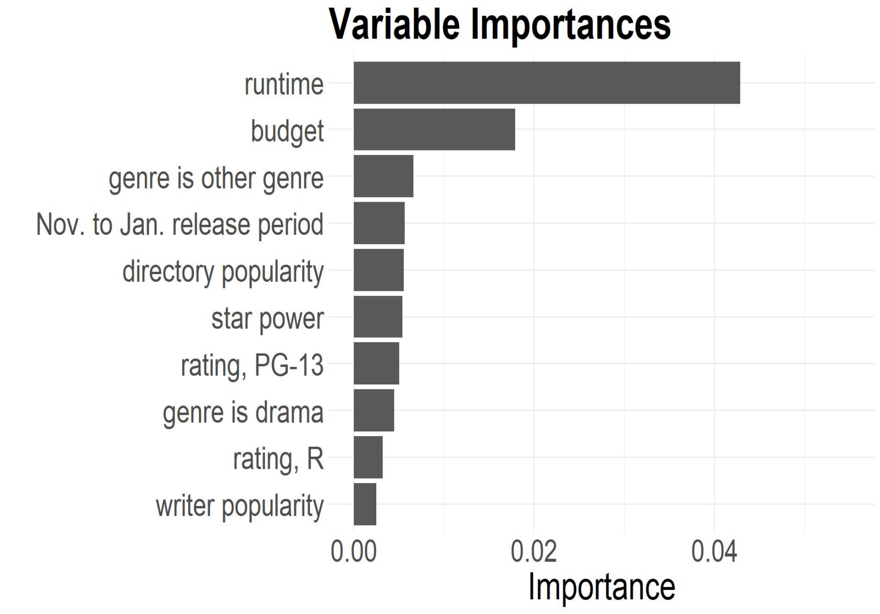
Results and Discussion: Box Office Profits – LM

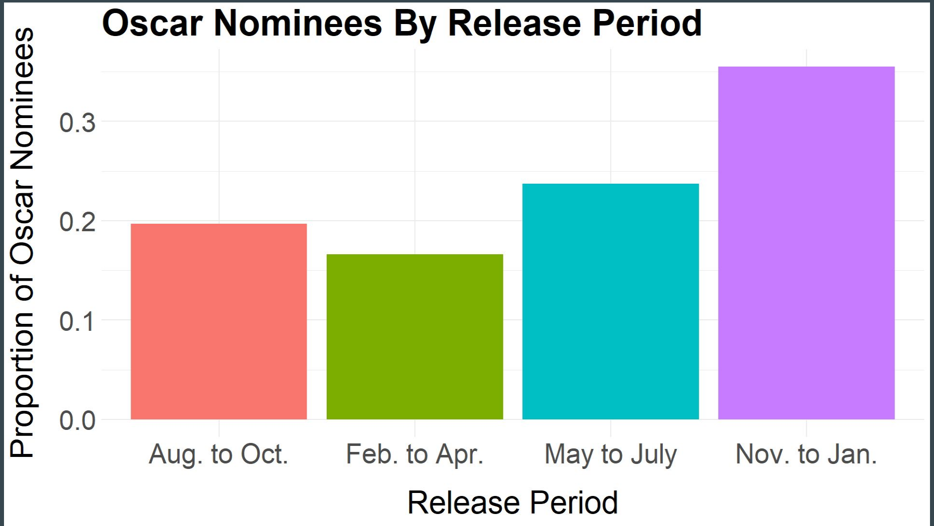
Coefficient	Estimate	Std. Error	P–Value
(Intercept)	93.9317	1.2583	2e-16
budget adj	16.7476	0.7439	**2e-16**
runtime	2.3632	0.6835	**0.000557**
dir pop fac	-0.1901	0.6980	0.785423
feb to apr release	-0.5301	1.7848	0.766505
may to jul release	8.4182	1.7899	**2.74e-06**
nov to jan release	4.4887	1.7222	0.009221
co size	4.1257	0.6689	**8.42e-10**
star power	0.1404	0.6969	0.840306

Results and Discussion: Oscar Nomination

- Random Forest
  - Found the following to achieve highest accuracy while maintaining a high area under the curve:
    - 1250 trees (bootstrap resamples)
    - 7 variables randomly sampled as candidates for each split
  - With optimal parameters, the random forest is approximately 80.4% accurate
  - Precision = 68.5%, Recall = 42.0%, Area under the ROC curve = 0.80
- Logistic Regression
  - Precision = 70.2%, Recall = 43.1%, Area under the ROC curve = 0.88
  - Found that 60 minutes of additional runtime correlates with a 4% increase in the odds of an Oscar nomination
  - Budget’s relationship with the odds of an Oscar nomination was non-linear and therefore better modeled by the random forest

Results and Discussion: Oscar Nomination – RF

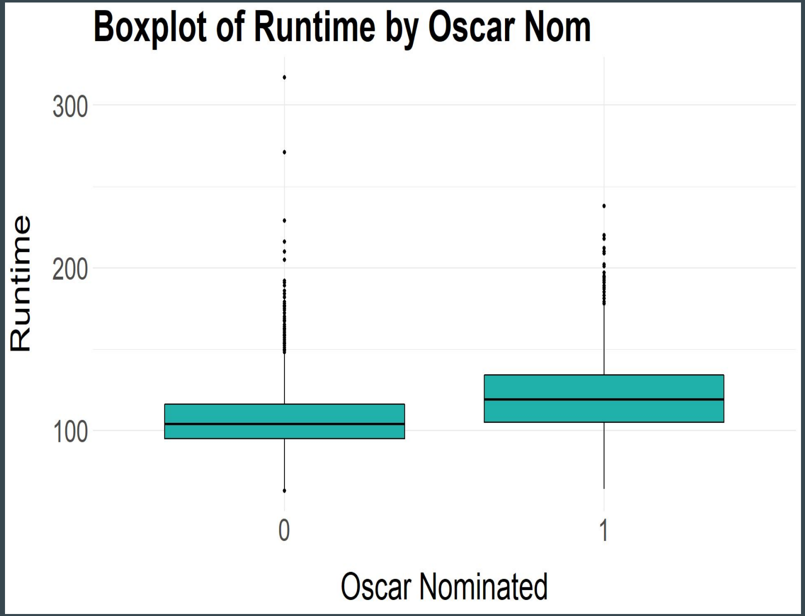


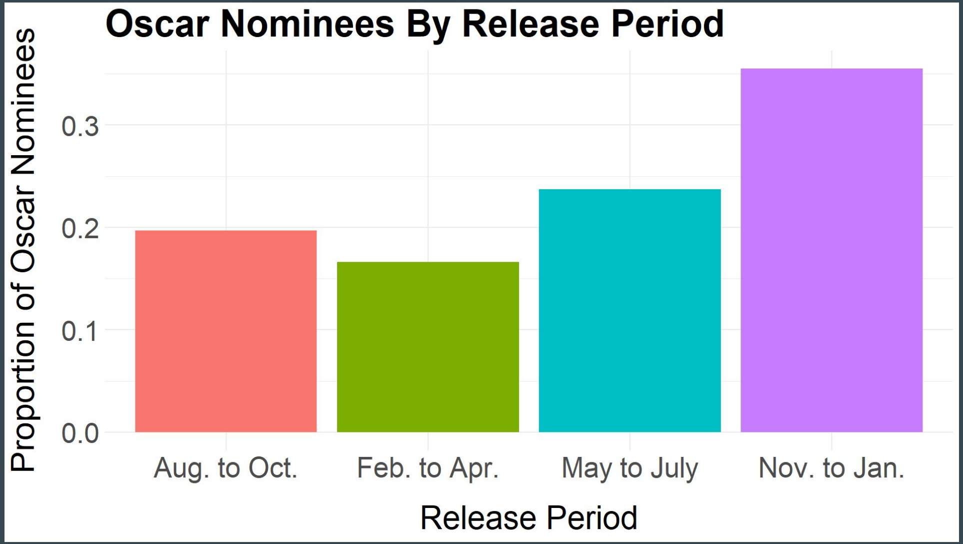


Results and Discussion: Oscar Nomination

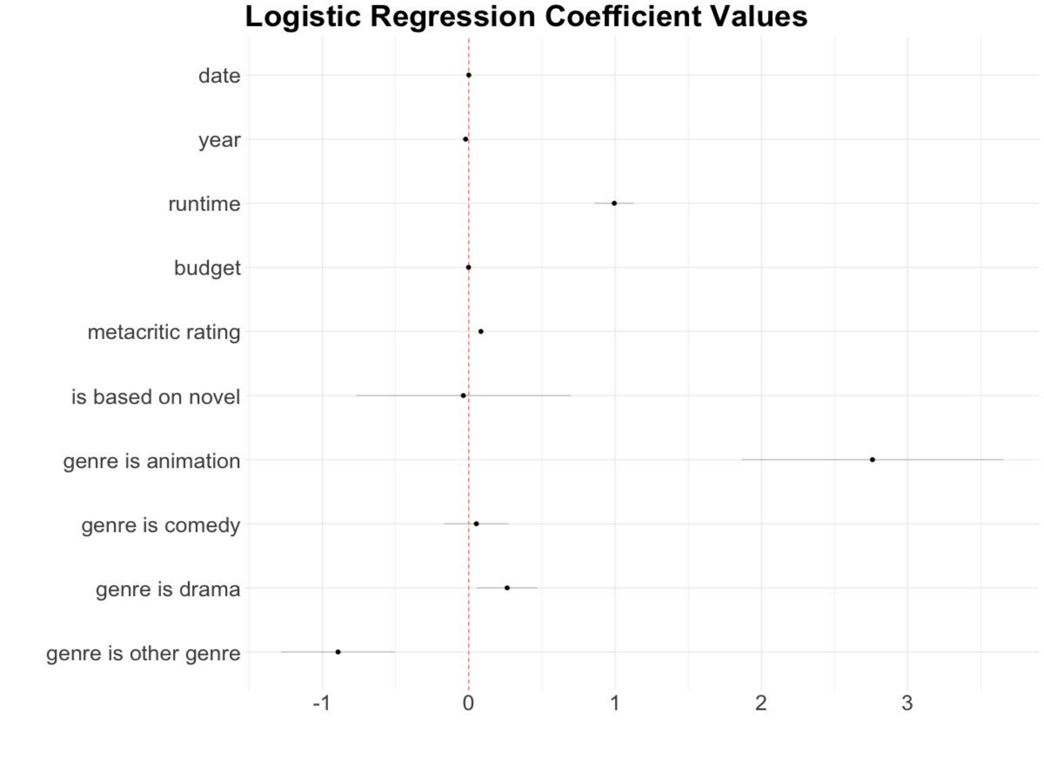
- Random Forest
  - Found the following to achieve highest accuracy while maintaining a high area under the curve:
    - 1250 trees (bootstrap resamples)
    - 7 variables randomly sampled as candidates for each split
  - With optimal parameters, the random forest is approximately 80.4% accurate
  - Precision = 68.5%, Recall = 42.0%, Area under the ROC curve = 0.80
- Logistic Regression
  - Precision = 70.2%, Recall = 43.1%, Area under the ROC curve = 0.88
  - Found that 60 minutes of additional runtime correlates with a 4% increase in the odds of an Oscar nomination
  - Budget's relationship with the odds of an Oscar nomination was non-linear and therefore better modeled by the random forest

Results and Discussion: Box Office Profits – RF





Results and Discussion: Box Office Profits – RF



Results and Discussion: Finding Similar Movies

Input

Movie Title	Gross USA ($)
Star Wars: Episode VII - The Force Awakens (2015)	936,662,225

K-Nearest Neighbors

Movie Title	Gross USA ($)
Spider-Man: No Way Home (2021)	804,617,772
Avengers: Infinity War (2018)	678,815,482
Avengers: Endgame (2019)	858,373,000

Most Common Words: world, superhero, marvel cinematic universe

Results and Discussion: Finding Similar Movies

Input

Movie Title	Metacritic Rating	IMDb Rating
The Godfather (1972)	100.00	9.20

K-Nearest Neighbors

Movie Title	Metacritic Rating	IMDb Rating
Goodfellas (1990)	90.00	8.70
The Godfather: Part II (1974)	90.00	9.00
The Girl with the Dragon Tattoo (2011)	71.00	7.80

Most Common Genres: drama and crime

Conclusions, Limitations, Shortcomings

- Satisfied with the gross-profit model and the two Oscar-nomination models
- We found the most important variables to each of the models that a producer would find valuable
- Believe a producer could use any combination of models and recommendations to make a new film
- Only 25% of movies in the data were Oscar nominated
    - Likely the cause for low recall
- The created variables might not be entirely accurate, but we feel they worked as a decent proxy