140 Final Team Project Proposal

140 Final Team Project Proposal

Figure 1: Team Picture: Cassia Ramelb, Alexis Adzich, Kevin Hamakawa, Jasmine Chu, Rohan Saklani, Annie Cen

Team Member Names and Associated Responsibility/Duties

- Data Analysts (analyze data using hypotheses and methods listed below & generate visualizations) - Alexis Adzich,
- Poster Editors - Jasmine Chu
- Presenters (present what we have on poster or make a slideshow) - Cassia Ramelb
- Editors (write final report) - each take a section or two of report? Cassia Ramelb, Annie Cen, Alexis Adzich, Rohan Saklani, Jasmine Chu, Kevin Hamakawa
- Final Editors (clean up report, create any additional data viz, and put it into rmd/qmd/latex) - Jasmine Chu, Rohan Saklani

Problem Statement

We want to identify the key factors driving a movie's financial success to help with budgeting, investment risks, and addressing representation gaps in the film industry

Motivation/Background:

- Help studios make data-driven decisions on budgets by understanding what factors contribute most to high revenue
  - Like should they invest in star power, specific genres, or runtime adjustments?
- For investors…investment risks: identify the "safe bets" by focusing on factors associated with profitable movies
- Marketing strategies
  - Optimize release time based on impact of timing (e.g., summer or holiday seasons) on revenue outcomes
  - Use results to tailor promotional campaigns based on popular factors like high IMDB scores or strong cast members
- See if diversity among top actors, directors, or genres correlates with financial success, which can help address representation gaps in the film industry

Dataset

Metadata for IMDb's top 1000 movies and TV shows Last updated: 2020

Original data: Obtained from Kaggle: IMDB Dataset of Top 1000 Movies and TV Shows Cleaned data: CURRENT_clean_imdbmovies_version5.csv https://github.com/aadzich/Stats140-Project

Lower Level Statements

- Determine how factors like genre, star power, and runtime influence revenue to allocate production budgets effectively
- Evaluate the predictability of financial success based on IMDB ratings, meta scores, votes, etc.
- Examine audience sentiment and engagement based on movie overviews and representation

Descriptive Statistics

- Mean domestic box office gross revenue: $60,513,599

- Max domestic box office gross revenue: $936,662,225 (Star Wars Episode VII The Force Awakens)
- Mean IMDB rating: 7.9493
- Max IMDB rating: 9.3 (The Shawshank Redemption)
- Mean runtime: 122.891 minutes → ~ 2 hours

Data Management

- Modifications made: Simplified column names, include sentiment analysis (based on overview), converted datetime, rm special characters in overview, handled nas in revenue by averages, rm duplicates, include genre count (movies per genre), include decade, include broad genre (reclassify genres into broader categories), rm duplicates, rm poster link.
- Further modifications: Possibly aggregate revenue for each star

Testable Hypotheses

- Movie Genre v Revenue: H0: The mean gross revenue is constant across all movie genres. HA: At least one movie genre has a mean gross revenue that is significantly different from the others.

- Director Influence on Revenue: (can explore similarly with influential/popular actors) H0: The mean gross revenue of movies directed by prolific directors (2+ films) is equal to the mean gross revenue of movies directed by other directors. HA: The mean gross revenue of movies directed by prolific directors (2+ films) is greater than the mean gross revenue of movies directed by other directors.

- Linear Regression: H0: The IMDB rating, number of votes, and runtime (or other variables) do not significantly predict a movie's total gross revenue. HA: The IMDB rating, number of votes, and runtime (or other variables) significantly predict a movie's total gross revenue.
- Random Forest Feature Importance: H0: The genre, director, (and other discrete/categorical variables we can't test w lin reg) do not significantly contribute to predicting gross revenue in a random forest model. HA: At least one of the genre, director, (and/or other variables) significantly contributes to predicting gross revenue in a random forest model.

Statistical Methods

- Hypothesis testing: T-tests, ANOVA, Chi-square (assess relationships, compare revenue across groups)
- Predictive modeling: linear regression, multiple regression, random forests, feature engineering/interaction (find key predictors, predict revenue for new movies, non linear models)
- Sentiment analysis: (star power, overview)
- Clustering: k-means (split groups into high vs low revenue)
- Explore which predictors have the most influence and find out more about the relationship with Revenue
- Consider the Confounding factors with each potentially influential predictor