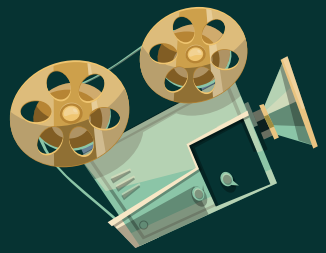
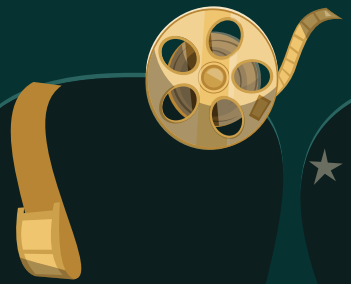


IMDb Movie Data Analysis



IMDb

Katherine Mulligan, Kanyarat Suwannama,
Nam Nguyen, Venkat Deenamsetty, Aaron Johnson

Background Information

The Internet Movie Database (IMDb) is one of the largest online databases and authoritative sources for movie, TV, and celebrity content.

- ➔ The IMDb Metascore is a weighted average of the published critic reviews, ranging from 0–100.
- ➔ The IMDb Rating is based on the votes of the website's users, ranging from 1–10 system.
- ➔ The better the movie, the higher the Metascore and rating are given.

Movie rating is a great way to share users' opinions, keep track of movie records, receive movie recommendations and provide practical implications for producers, distributors and consumers.

Research Problems

However, the IMDb rating methodology remains unclear, including which factors contribute to the ratings and how these ratings are measured...

Research Question 1
Which factors contribute most to the IMDb score?

Research Question 2
What is the impact of votes and revenue on IMDb score?

Research Question 3
How do the categories affect IMDb score?

Research Question 4
Are there any description keywords that are correlated with a higher IMDb score?

Data Description

Preparation Process:

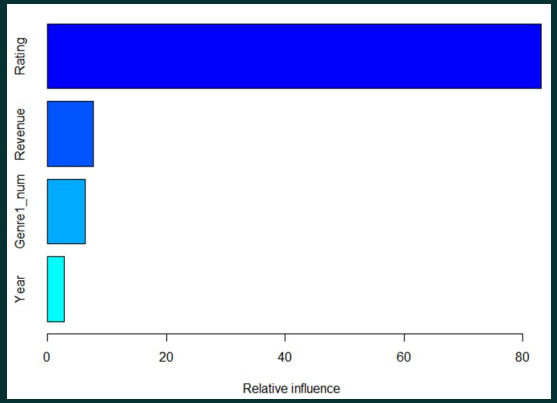
1. **Data Collection** – Collect IMDB Data from data source [IMDB Data from 2006 to 2016 – dataset by promptcloud | data.world]

2. **Data Exploration** – 12 variables for 1000 movies [rank, title, genre, description, director, actors, year, runtimes, rating, votes, revenue and metascores]

3. **Data Cleaning** – remove missing values and replace with mean value for movies with the same rating

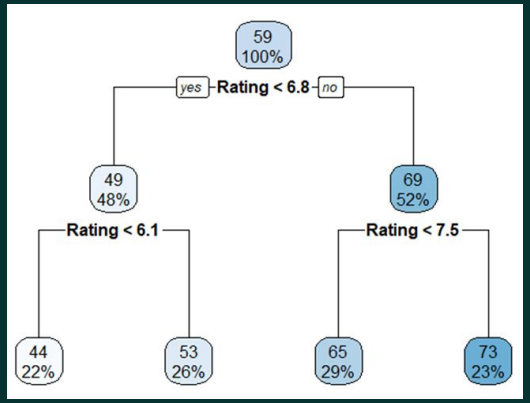
(1) Which factor(s) contribute most to the IMDb score?

Feature Selection



- **4 features** selected for predictive model: Rating, Genre1_num, Revenue and Year.
- In which, Rating has the biggest influence on IMDb score (Metascore), far more than other variables

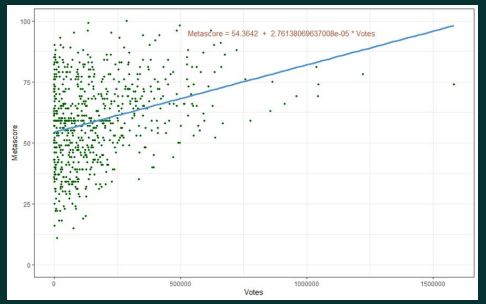
Metascore prediction tree

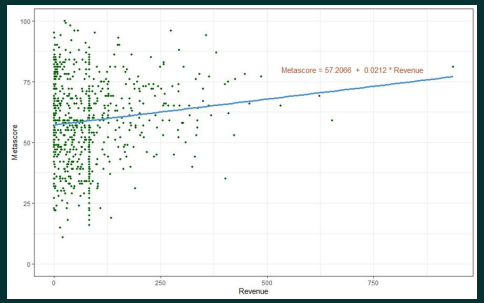


- **Rating** is key evaluation value in each node
- Based on above result, Rating could be predictive of IMDb score.

(2) What is the impact of votes and revenue on score?

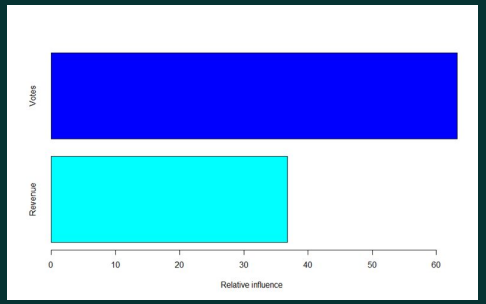
Linear Regression Model

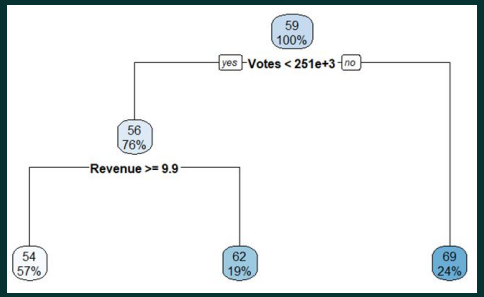




- **Two models are statistically significant,** so we can see both votes and revenue have positive impact on IMDb score

Metascore prediction

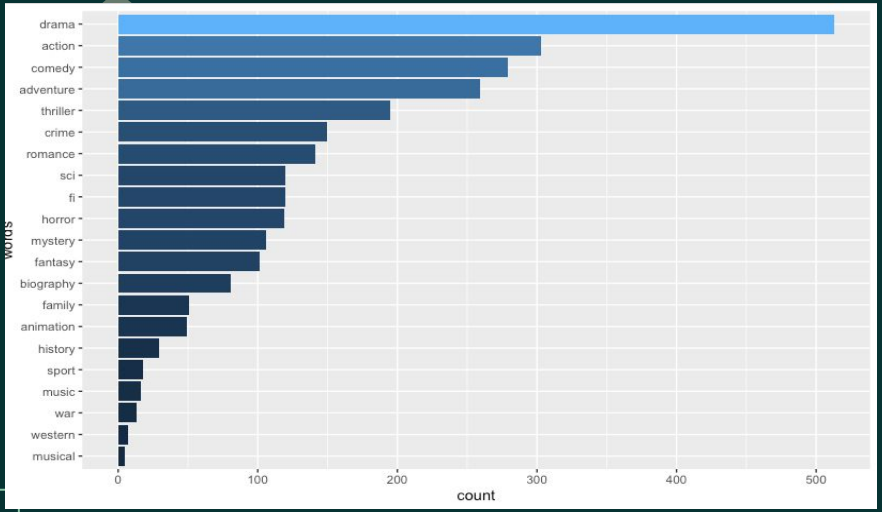


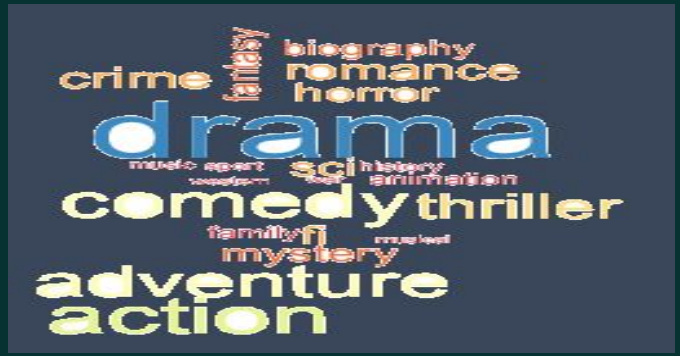


- **Votes** has the bigger influence than Revenue on IMDb score
- We can also see this influence from decision tree model, both are predictive of IMDb score.

Research 3 – How do the categories affect rating?

- Dataset had Genre column indicating possible categories for the movie.
- Objective is to determine how the movie categories influence the rating of the movie .
- Figure shows the count of categories in all the movies





Predictive Analysis on the Categories

Perform predictive analysis of the categories in Genre column

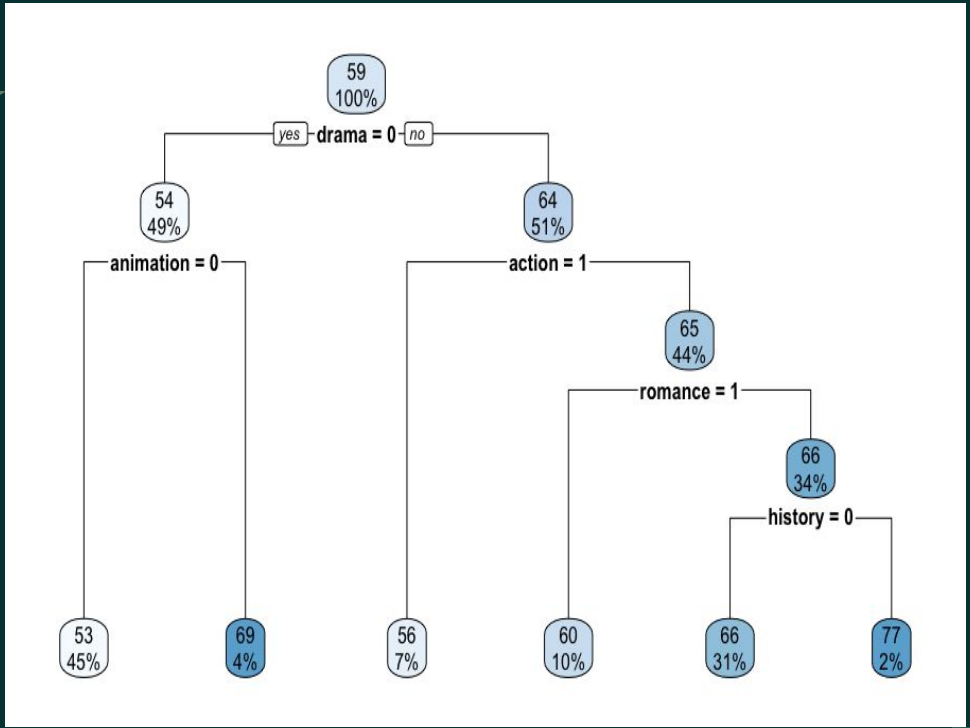
- Creating a corpus from the variable ‘Genre’
- Using tm_map to transform text to lower case
- Creating a dictionary
- Creating a DocumentTermMatrix (dtm)

(3) How do the categories affect rating?

run a CART model (decision tree)

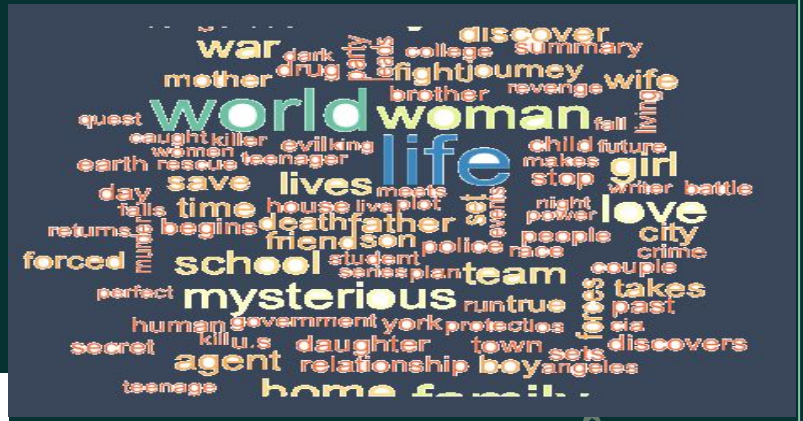
From this decision tree, we have some interesting predictions:

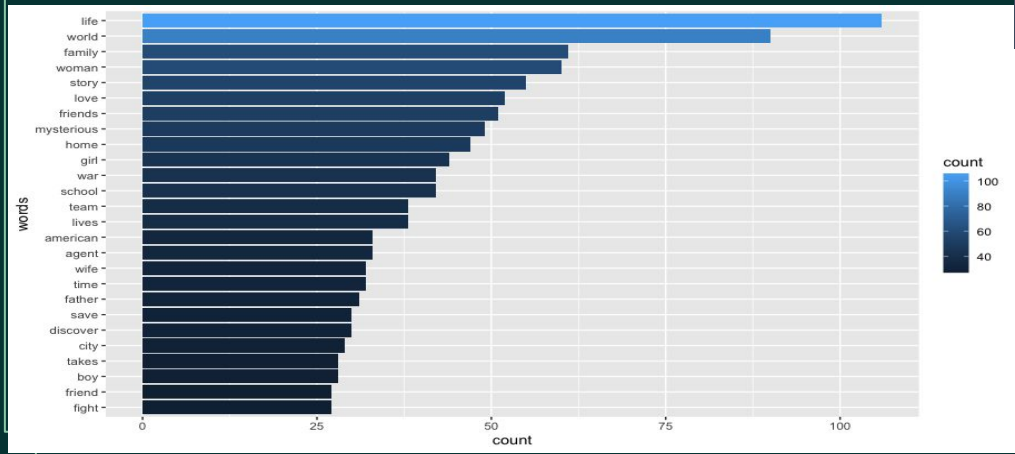
- Highest rated movies: a combination of drama and history without any action would have an average Metalcore of 77, and this is the recipe for success.
- Lowest rated movies: no drama, no animation movie would have an average Metascore of 53, and this is recipe for failure.
- However, animation movie without drama could perform well, with average Metascore of 69. Therefore, animation could be a big boost / guarantee for the success of a movie.

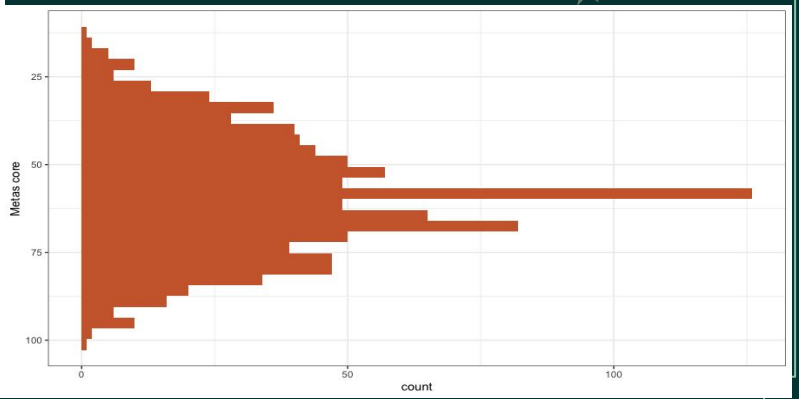


Research 4 – Are there any description keywords that are correlated with a higher IMDb score?

- Dataset had Description column indicating possible keywords about the movie.
- Objective is to determine any keywords which will influence the Metascore of the movie .
- Figure shows the count of key words in the title for all the movies

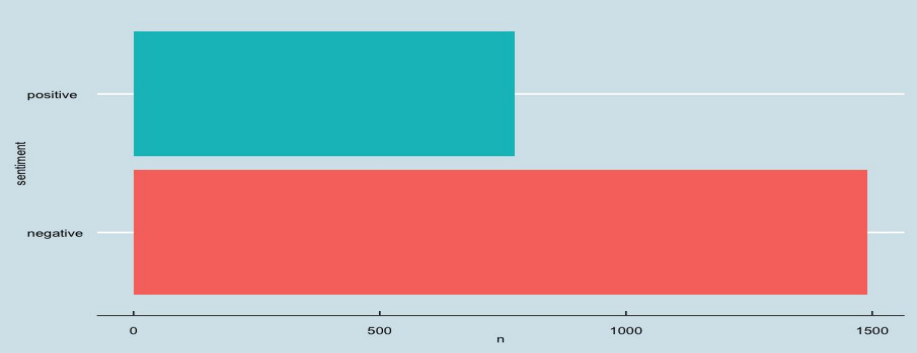






(4) Are there any description keywords that are correlated with a higher IMDb score?

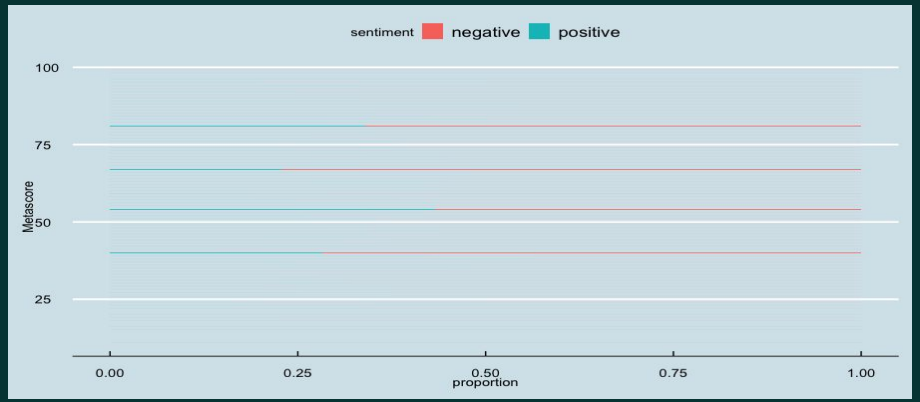
- Word cloud map shows positive and negative word for the top 200 words
- Figure shows the total number positive and negative words when processed through a Binanry sentiment Lexicons



(4) Are there any description keywords that are correlated with a higher IMDb score?

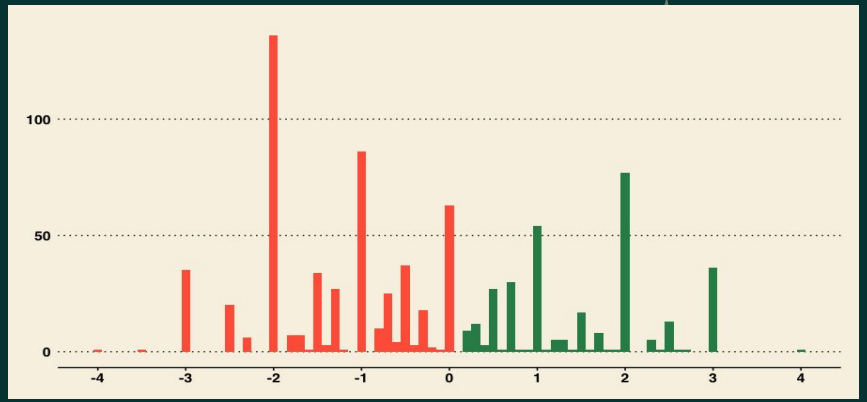
Bing Sentiment Analysis

- **Proportion of positive and negative words for each rating.**
- Correlation of positive words with the Metascore had a ratio 0.019



afinn Sentiment Analysis

- Figure below shows the distribution of sentiment
- Figure has dominance of negitve sentiment indicating no relation to the Metascore



Research Analysis

Question 1

- Metascore from Metacritic is weighted based on fame of critic
- Metascore is a standardized process of rating vs. IMDb score from casual viewers
  - Movie buffs may be more interested in critics choice
  - Casual viewers may be more interested in other casual viewer opinions

Question 2

- Votes are a measure of continued interest/success in movie
- Revenue is measure of initial (box office) success
- Revenue generated before watching vs. Voting placed after watching

Research Analysis

Question 3

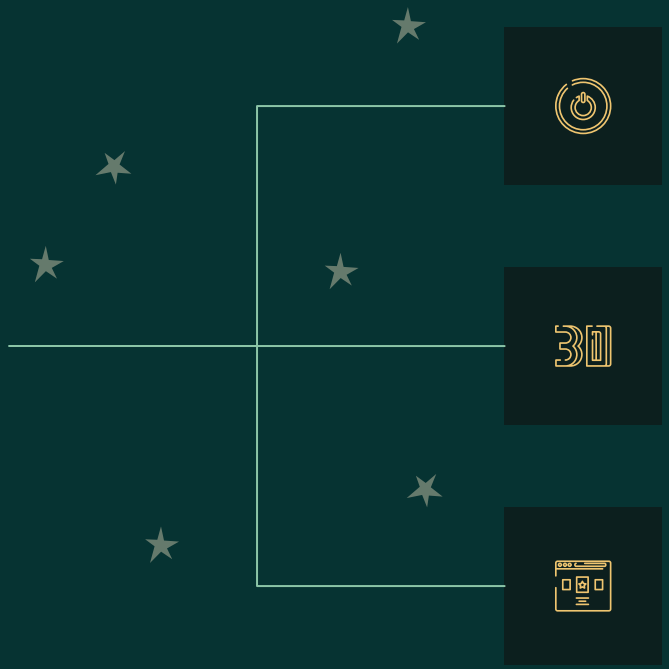
- Highest rated movies are drama and history with no action (only 2% of IMDb movies)
- 45% of movies not dramatic or animated:
  - Comedy, Horror, Thriller, Fantasy, Biography, Documentary
  - Combined Predicted Score: 53

Question 4

- Proportion of negative to positive sentiments is indicative to common movie plots
  - Negative sentiment approx. double the amount of positive sentiment → even "feel-good" movies describe negative problems

RECOMMENDATIONS

- Further research into movie genres based on awards (Academy Awards, Golden Globes, BAFTAs, Screen Actors Guild, etc.)
  - Create genre decision tree of Best Picture Winners
  - Chronological visualization of winners based on genre
- Goal:
  - Determine most successful movie genres for awards
  - Identify genre preference changes over time
- Outcome:
  - Movie makers can decide between making a movie for viewer rating or for award prestige



Thank You !