Project Report : IMDb Movie Data Analysis Report

Group 10: Venkat Deenamsetty, Nam Nguyen, Kanyarat Suwannama, Katie Mulligan, Aaron Johnson

---

Statement of Research Problems

Social media and the rapid growth of the film industry worldwide have redefined how consumers get information by offering a variety of reviews such as user ratings, user reviews, and critic comments. Movie rating is a great way to share users' opinions, keep track of movie records, and receive movie recommendations. Also, the ratings are seen as good predictors of the quality of the movie and determine whether or not the movie is worth recommending. Thus, a good understanding of these ratings impact sales and the importance of such impacts would provide practical implications for movie producers, distributors, and consumers. Agencies and data scientists alike have been trying to extract insights from streaming and behavioral data to target the correct customers and recommend movies at the right time.

The Internet Movie Database (IMDb) is one of the largest online databases and authoritative sources for movie, TV, and celebrity content. The IMDb offers a rating scale that allows users to rate films with different appreciations. The rating is based on the votes of the website's users, ranging from 1-10 systems. In addition, the IMDb Metascore is a weighted average of the published critic reviews, ranging from 0-100. The better the movie, the higher the Metascore and rating are given. However, many factors such as description, experienced directors, and famous actors are considered for creating good movies but do not ensure a good rating on IMDb. Additionally, the methodology for generating the IMBd movie rating remains unclear, including which factors contribute to the ratings and how these ratings are measured.

Upon analyzing this data, we intend to answer the following research questions:
(1) Which factors contribute most to the IMDb score?
(2) What is the impact of votes and revenue on IMDb score?
(3) How do the categories affect rating?
(4) Are there any description keywords that are correlated with a higher IMDb score?

Machine learning insights can improve the success of streaming companies in many ways. We will try to understand what factors (user ratings, reviews, and comments) contribute most to the movie ratings, revenue and provide an advanced recommender system. In order to answer these questions, we will create models that will predict the correlation of the factors and how they contribute to the score.

Therefore, this project aims to see which factors contribute the most to the IMDB score and provide analytical insights into movies. To accomplish our goal, we performed the analytical process as follows:

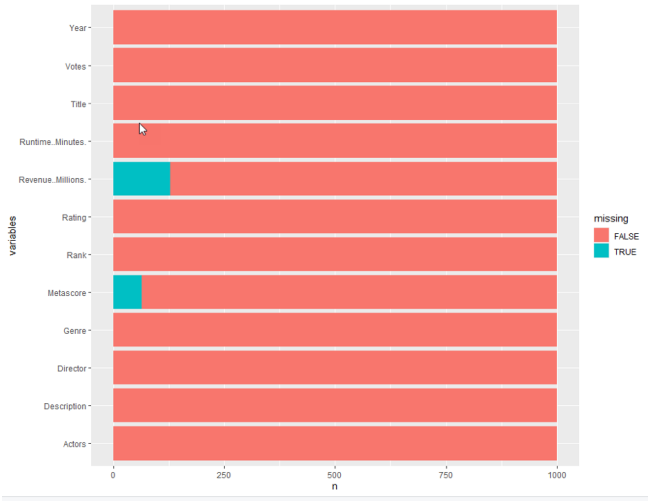
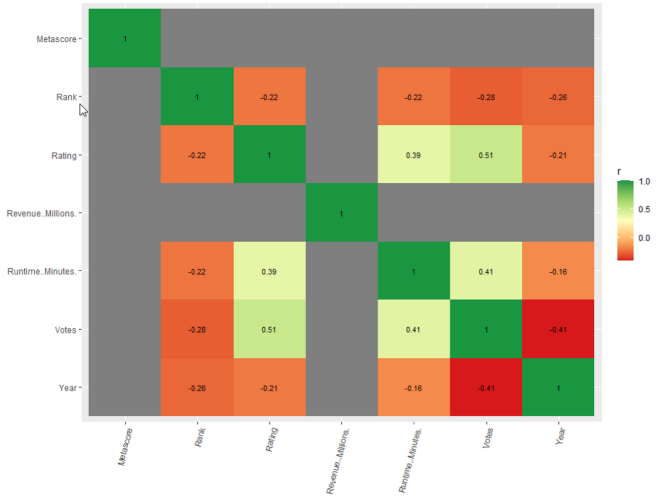
1. Initial Exploration - Load a dataset and understand the variables
2. Data Preparation - Clean data by removing duplicate and useless columns and inputting the missing value
3. Exploratory Data Analysis & Virtualization - plotting and calculating variables' frequency to understand the relationship better and identify important factors
4. Predict movie rating with Machine Learning
5. Final Conclusion and Recommendations

Data Description

To obtain a good data set several datasets were reviewed and the team felt collectively the data from data source [IMDB Data from 2006 to 2016 - dataset by promptcloud | data.world] will be a good source to pursue this analysis. The Raw data was downloaded and we followed the following to explore, analyze, clean and finalize the data.

The dataset contains 12 variables for 1000 movies. This is a summary dataset that contains all the data of IMDB from the year 2006 to 2016. The era where the greatest movies were made and what was the start of the Marvel Cinematic Universe and various other movies that went on to change the world. The dataset includes rank, title, genre, description, director, actors, year, runtimes(minutes), rating, votes, revenue(millions) and metascores for reviews that may impact a movie IMDB scores. “imdb_score” is the response variable while the other variables are possible predictors.

Initial exploring of the data found the 2 features of Revenue and Metascore have missing values. Also, in order to understand the data we extracted the numeric features in the dataset as shown in Figure 1 and generated Correlation heat maps as shown in Figure 2 below.

Figure 1: Numeric Features of Each Variable	**Figure 2**: Correlation Heat Maps of Each Variable
	

In the process of cleaning and preparing the data it was identified that the Metascore feature had 64 missing values . The missing data for Metascore is substituted with the main of Metascore from the movies with same rating mean of metascore , based on the similar rating. For the revenue, there were 129 rows of missing data. We substituted the revenue with mean revenue for movies with the same rating.

Choice of Analytical Technique(s) and Modeling Outcome

(1) Which factors contribute most to the IMDb score?

We perform feature selection to find the best fit model to predict IMDb score.
Firstly, we cleaned the data for analysis:

- We split the Genre variable into 3 categorical variables, namely Genre1, Genre2, Genre3. From domain knowledge, we select Genre1 as the main genre for the movie and will use this variable for analysis.
- We convert two categorical variables, Genre1 and Director, to numerical ones for further analysis.

Then we split the analysis data into train (70%) and test (30%) sets for modeling and validation. Then we selected 7 numerical features that are most potential in predicting Metascore: Rating, Votes, Revenue, Year, Runtime, Genre1_num (from Genre1), and Director_num (from Director variable)
.
Stepwise feature selection process shows that we could include 4 features in the model, namely (1) Rating, (2) Genre1_num, (3) Revenue and (4) Year.

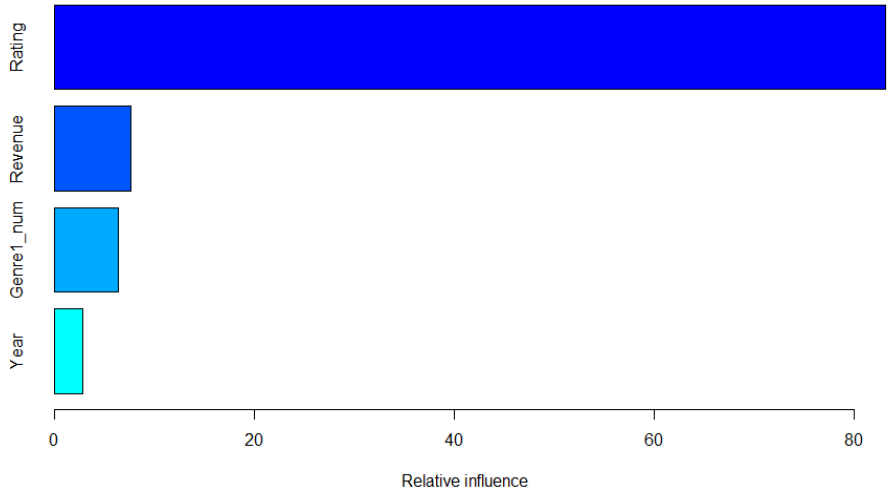
```
Step:  AIC=3625.65
Metascore ~ Rating + Genre1_num + Revenue + Year

               Df Sum of Sq    RSS    AIC
<none>                      117608 3625.7
+ Runtime       1       151 117457 3626.7
+ Votes         1       147 117461 3626.8
+ Director_num  1       144 117464 3626.8
- Year          1       602 118210 3627.3
- Revenue       1       675 118283 3627.7
- Genre1_num    1      3828 121436 3646.3
- Rating        1     71315 188923 3958.8
```

Next, we ran 3 different models, namely liner regression, decision tree, and boosting, on both train and test data sets to find the lowest RMSE.

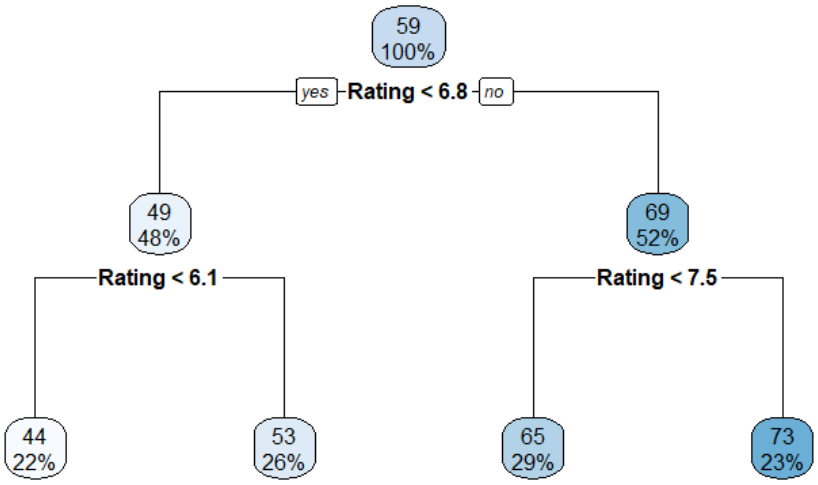
Model from previous section, with 4 features	**RMSE on training data**	**RMSE on test data**	**Other notes**

Model 1: Linear Regression	12.89758	13.13025	
Model 2: Decision tree	12.6402	13.25852	
Model 3: Decision tree – Boost	11.98625	12.5069	Best result

From the Influence matrix in Boosting model which has lowest RMSE, we could see that Rating has the biggest influence on IMDb score (Metascore), far more than other variables.

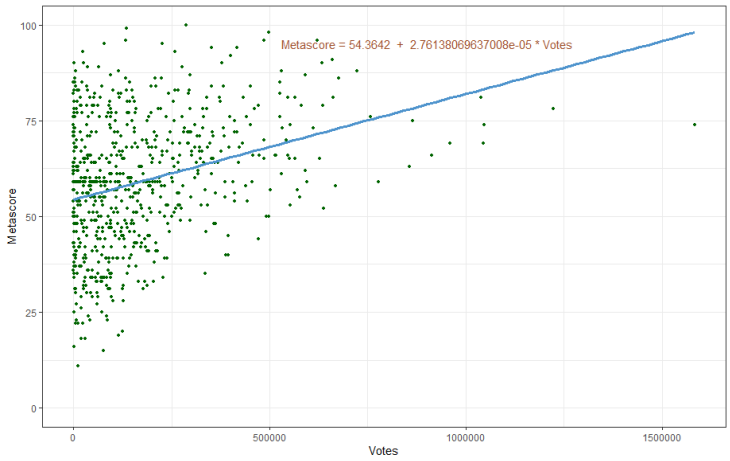


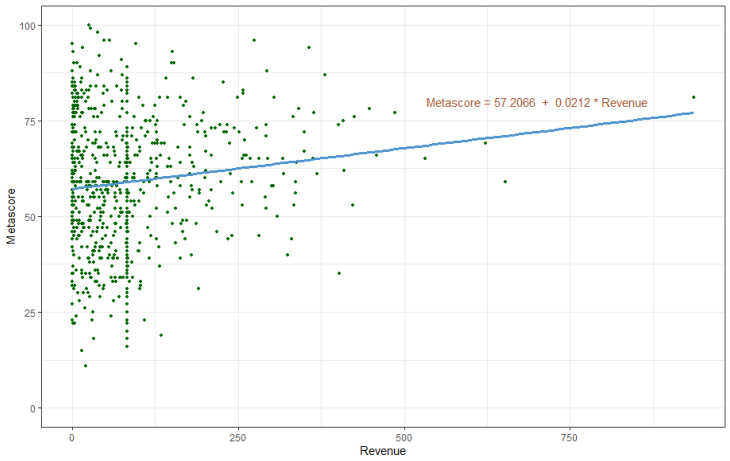
Metascore prediction from Decision Tree model:



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

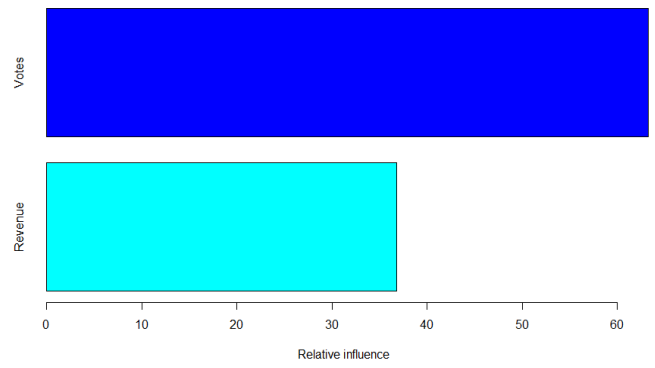
We run two linear regression models to evaluate the impact of each variable on IMDb score. These two models are statistically significant, so we can see both votes and revenue (independently) have a positive impact on IMDb score.



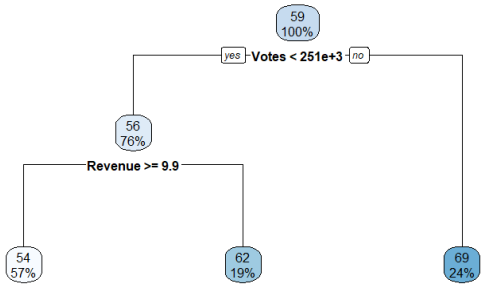


The correlation between Votes and Revenue is 0.61 which is significant. This can be explained that the more Votes a movie receives, the more people already saw it, indicating higher box office/ Revenue.

We then run a boosting model with two features, Votes and Revenue, to evaluate its impact on IMDb score. From the Influence matrix, we could see that Votes has the bigger influence than Revenue on IMDb score (Metascore).



We also have Metascore prediction from Decision Tree model as follows:

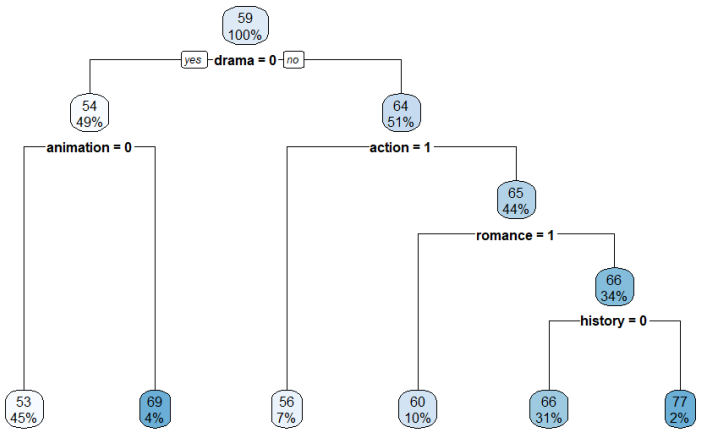


(3) How do the categories affect IMDb score?
We will apply text mining techniques to evaluate impact of movie genre on IMDb score (Metascore).

Firstly, we clean and tokenize the movie genres by:

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

Then we add Metascore back to the dataframe of genres. From that TF (term frequency) features, we can run a CART model (decision tree) to predict IMDb score (Metascore) from movie genre as follows:



From this decision tree, we have some interesting predictions:

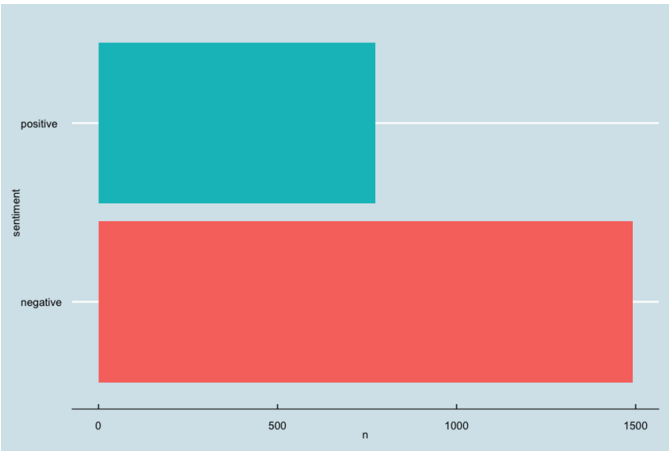
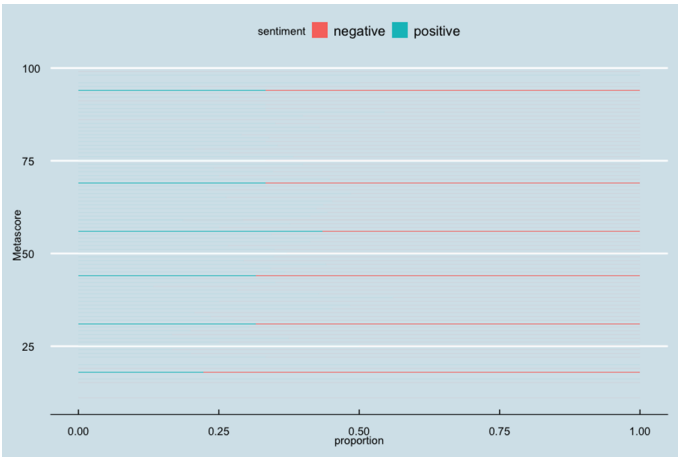
- Highest rated movies: a combination of drama and history without any action would have an average Metascore of 77, and this is the recipe for success.
- Lowest rated movies: no drama, no animation movie would have an average Metascore of 53, and this is a recipe for failure.
- However, animated movies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

(4) Are there any description keywords that are correlated with a higher IMDb score?
To determine that the words in the description as any impact on the IMDB score , we need to explore the text and conduct sentiment analysis

We shall try to explore 2 different ways to identify any correlation

- Binary Lexicon – Bing
- Sentiment Lexicon – afinn

Bing Lexicon

Sentiment count	Let us drill down a bit more to see whether the proportion of positive words has any impact on its helpfulness. We will look at the proportion of positive (and negative words) for each rating.
	

Positive Reviews

Let us compute the proportion of positive words for each review. The proportion of positive words is the ratio of positive words and the sum of positive and negative words. This differs from the analysis above as it is computed for each review.

	Rank	Title	Metascore	positive_words	negative_words	proportion_positive
	<int>	<chr>	<dbl>	<int>	<int>	<dbl>
1	1	Guardians of the Galaxy	76	1	1	0.5
2	3	Split	62	0	1	0
3	5	Suicide Squad	40	1	4	0.2
4	6	The Great Wall	42	1	3	0.25
5	7	La La Land	93	0	1	0
6	8	Mindhorn	71	2	1	0.667
7	9	The Lost City of Z	78	0	1	0
8	13	Rogue One	65	0	3	0
9	14	Moana	81	1	3	0.25
10	15	Colossal	70	1	0	1

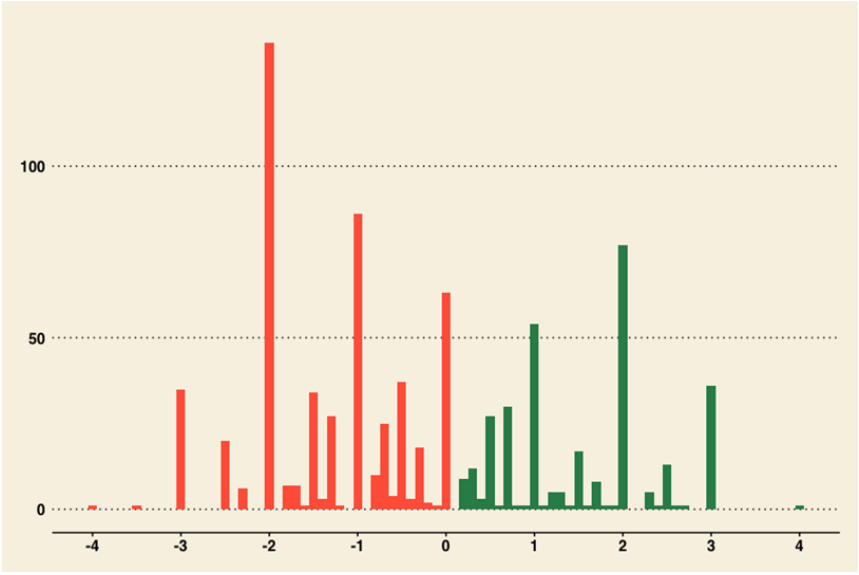
… with 861 more rows

Let us see if reviews with a lot of positive words are rated favorably.	Finally, here is a comparison cloud to contrast positive and negative words in the reviews.
	

Afinn sentiment

Distribution of the sentiment

afnn	min	max	median	mean
	-4	4	-0.5	-0.271



Looking at both the semantics, the words in description have no impact on IMDb score.

Analysis Results

Research Question 1:
We determined that the metascore is the strongest predictor of IMDb score. Metascore is a standardized process of measurement for the movies by trained professionals. The credibility of critics is a good determinant of whether or not casual viewers will enjoy the movie. The two scores can be used to determine what movie is the best option for viewing. Movie buffs are able to find the critics whose opinions they align with and can refine their movie search by the Metascore from said critics. In addition to passionate movie fans, these scores attract the casual viewers by displaying a reliable, numerical representation of the expectations for a given movie.

Research Question 2:
The results of research question 2 can be understood through the premise that votes are a measure of the continued interest and success of the movie, whereas revenue is a measure of initial success in the box office, especially as streaming services decrease the post box office purchases. Viewers submitting votes after watching a movie allows them to continue support for the production long after its initial release. However, revenue is mainly generated in box office which is dependent on high advertising budgets, popular actors or directors, and a seemingly interesting plot. It is less representative of quality of movie than the IMDb votes.

Research Question 3:
The highest rated movies are a combination of drama and history with no action. However, only 2% of movies on IMDb fall into that category. On the flipside, 45% of movies include no drama and no animation and the lowest metascore of 53. given that these movies are not dramatic or animated, we could conclude the 45% of movies may include genres such as comedy, horror, thriller, fantasy, and documentary. This means that comedies with 1.9 ratings are being included with adventure movies like Star Wars with an 8.7 rating. Based on this analysis, the information gained from answering research question 3 is most useful for someone interested in creating a movie based on the highest rated genres. However, further research can be done to dissect the remaining genres that make up the remaining 45% of movies.

Research Question 4:
While we were able to determine that the word choice in the description has no effect on IMDb score, the proportion of negative to positive sentiments is indicative to the plots of movies. This result leads us to conclude that even feel good movies rely on a storyline involving conflict and words of negative sentiment to describe said issues.

Recommendations and Conclusion

This project aims to see which factors contribute the most to the IMDB score and provide analytical insights into movies. To accomplish our goal, we performed the analytical process as follows:

1. Initial Exploration - Load a dataset and understand the variables
2. Data Preparation - Clean data by removing duplicate and useless columns and inputting the missing value
3. Exploratory Data Analysis & Virtualization - plotting and calculating variables' frequency to understand the relationship better and identify important factors
4. Predict movie rating with Machine Learning
5. Final Conclusion and Recommendations

Taking a step back we analyzed the factors that contributed to rating, categories that affected ratings, and the correlation between revenue and votes. These findings Provide insights that help directors, producers and studios outline the different objectives/outcomes when releasing a film. However, for future analysis we'd recommend analyzing correlation between (genres x awards). Understanding this helps producers define goals of revenue, consumer rating, or critical acclaim.

Secondly, popularity of actors / directors across revenue, awards, and consumer ratings to inform internal alignment on objectives. We could achieve deeper/more insightful results