Bryan Rogers
7/8/2024
Bellevue University
DSC680 – Applied Data Science

The Second Project: "Unveiling Movie Magic with IMDB Data" – Milestone 3

Milestone 3 – White Paper:

Unveiling Movie Magic with IMDB Data: A Data Science Exploration

Business Problem/Research Questions

There are many different factors that that contribute to a movie's success in the film industry and people are constantly trying to understand audience preferences. This white paper explores how data science can be applied to analyze movie attributes and audience reception using data from the Internet Movie Database (IMDb). The research focuses on two key questions:

- **Question 1:** What are the key factors in the film industry that influence a movie's critical reception (IMDB score)?
- **Question 2:** How does movie popularity and critical reception (IMDB score) interact across different genres?

By answering these questions, we can gain valuable insights that can inform film production strategies, content recommendations for streaming services, and ultimately, enhance the moviegoing experience.

Background/History

The film industry has traditionally relied on intuition, experience, and market research to understand audience preferences. However, the explosion of data in the digital age offers new opportunities for data-driven decision making. IMDb, a prominent online movie database, provides a rich source of information on movies, including cast, crew, ratings, reviews, and release details. This project leverages this data to explore the factors influencing movie success from a data science perspective.

Data Explanation

The primary data source that is being used will be the IMDb dataset from the IMDb website. I will attempt to work with IMDb's public API to collect data on movies, including many different factors such as title, genre, director, cast, release year, budget, IMDB score, user ratings, and number of votes. The data will undergo Data Preprocessing to clean and preprocess the data to ensure accurate and complete data. Some of the cleaning steps include handling missing values, checking for inconsistencies and biases, and creating new features and variables if necessary.

Methods

This project will be mostly Exploratory Data Analysis (EDA) based with a little focus on modeling and machine learning, I will be using visualization techniques to help understand data distribution, identify patterns, and explore potential relationships between variables. These visuals will include box plots, scatter plots, and heatmaps to visualize correlations between movie attributes and critical reception score metrics. I will also be looking into statistical analysis by looking at correlation matrices and analysis to assess the strength and direction of relationships between movie features and IMDB score/user ratings. Hypothesis testing can also be done to determine the statistical significance of these correlations. After all this EDA, I will take a brief look into some machine learning, by creating a Random Forest Regression model to predict IMDB score based on various movie attributes. The model performance will be evaluated using metrics like R-squared and mean squared error.

Analysis

The EDA revealed interesting patterns. For instance, comedies and drama movies led the way with the most movies released in those categories. The correlation analysis identified statistically significant positive correlations between IMDB score and Metacritic reviews from professional reviewers, while user ratings showed a weaker correlation with IMDb votes. Box plots and bar graphs were used to display some of the preliminary EDA features.

Assumptions

There are a few assumptions that need to be taken into consideration when working with the IMDb dataset. There is a hope that the data that is collected from IMDb is accurate and reflects user preferences. The features that were chosen to present an accurate reflection that capture the factors influencing movie critical reception score. The Random Forest model is able to run and reflects the underlying relationships within the data.

Limitations

Along with the assumptions, there also a few limitations when dealing with this IMDb dataset. The data relies on user contributions and is varied in the responses, which can skew the data and make it subjective and biased. Since the dataset is focused on just IMDb data the analysis may not generalize to the entire movie population. The accuracy of the Random Forest model can be limited by the complexity of factors influencing movie success.

Challenges

There are some potential challenges that I could face when working with this dataset. There could be a challenge with data quality issues like missing values and inconsistencies required careful cleaning methods. Identifying and addressing potential biases in user-generated content could be a challenge due to the nature of the dataset. Interpreting the complex relationships between numerous movie attributes and reception metrics could be a challenge with a limited dataset.

Future Uses/Additional Applications

Using a popular dataset such as the IMDb dataset, this project lends itself to many potential future uses. There is potential use for the findings to inform movie recommendations on streaming services by considering user preferences and genre-specific trends. Filmmakers can leverage the insights to have production strategies that align with higher score movies to target audiences within specific genres. Further analysis can be implemented to explore the impact of marketing campaigns and advertising, as well as social media buzz on movie reception.

Recommendations

After looking at the dataset, and beginning to play around with the data, I have looked into and taken down some notes for things to add on to this project in the future. Use the given dataset with additional data sources (e.g., Box Office Mojo) to add financial performance metrics and data to potentially identify financial driven factors. Explore more advanced machine learning models like Gradient Boosting Machines for potentially improved prediction accuracy. Since the dataset can be seen as a text format, you can conduct sentiment analysis on user reviews to gain deeper insights into audience preferences beyond ratings.

Implementation Plan

After implementing the first round of EDA and model training, the next steps of the implementation plan, includes a five-step process. Step one, would be data enrichment by exploring and adding data from additional data to create a full picture of movie data analysis. After that, step two would be to enhance and refine the model with the new data and look at other models for improved performance. Step three would be to incorporate sentiment analysis by using Natural Language Processing (NLP) methods to analyze and review sentiment. Step four is to develop visualizations and reports that effectively present the data in a comprehensive and effective manner. The fifth and final step in the process would be deployment of the model in a future state that allows the data to be continuously used and reported on.

Ethical Assessment

This project makes sure to uphold the ethical principles throughout the data science process by focusing on a handful of different topics. There is consideration for data anonymization by ensuring that privacy protection methods are put in place. There is transparency about data bias by recognizing user bias and they are addressed in the analysis and reporting. By agreeing to any terms or service put in place by using publicly accessed data, the data is collected responsibly. The Random Forest Model would hopefully be easy to interpret and allow the factors to show accurate predictions. Measures will be considered to ensure that the data being used is secure during the storage and analysis processes.

Conclusion

By applying data science techniques to IMDb data, this project unveils valuable insights into the factors influencing movie reception. These findings can empower filmmakers, streaming services, and ultimately, enhance the moviegoing experience for audiences. Future work will look more in-depth deeper into data enrichment, model refinement, and explore the potential for real-world applications. This project is committed to conducting this research ethically and responsibly.

10 Questions from the Audience

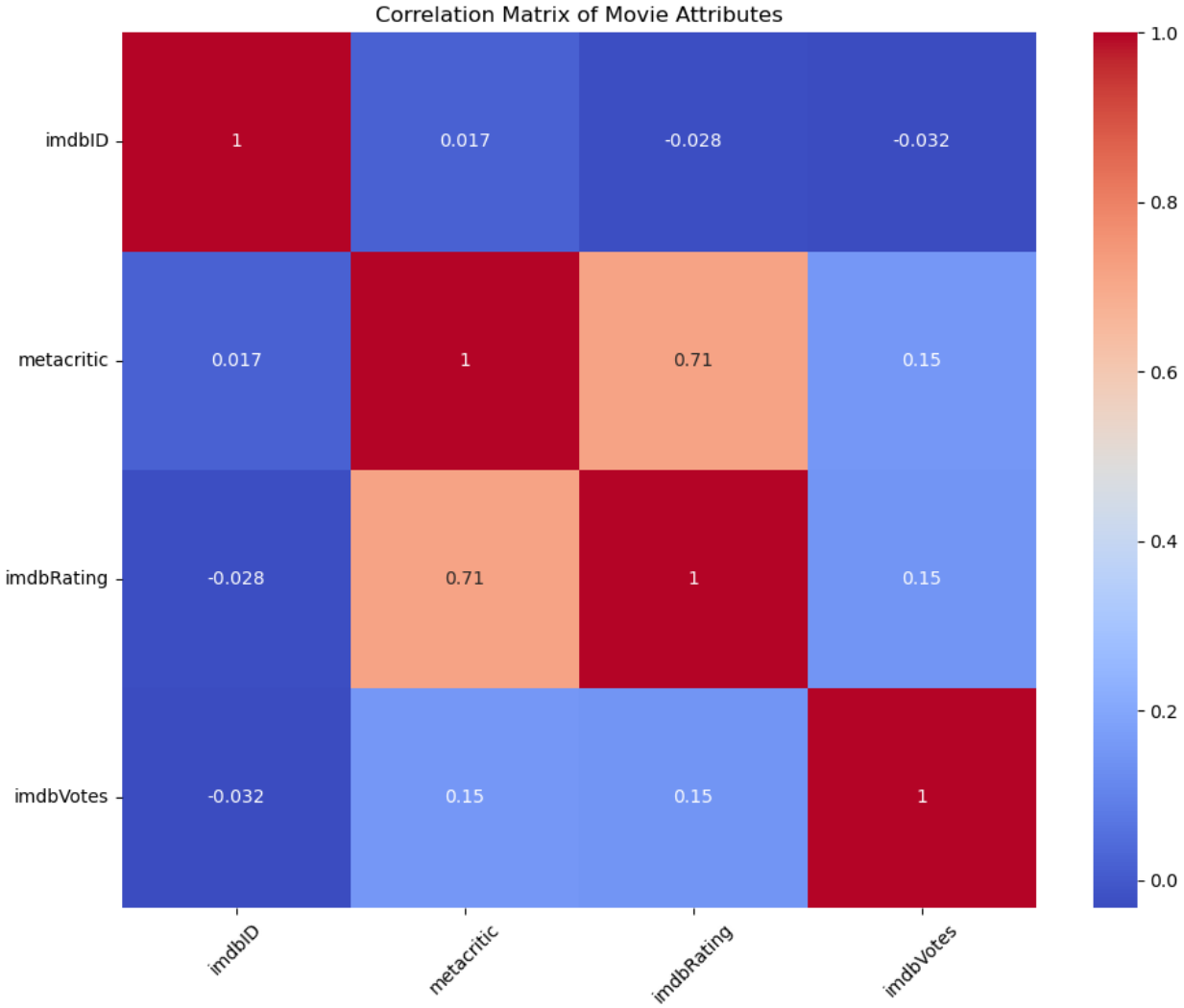
- Can your model actually tell me what movie to watch next?
    - *The model is not setup to recommend a specific movie as there are too many outside factors and variables. Hopefully, when the model is running, it can look for and find patterns that can recommend movies similar to ones you've enjoyed in the past, based on your preferences and what you are looking for.*
- Can your analysis tell me if people are just being harsh or if a movie is truly bad?
    - *Based on the EDA, this analysis cannot do this at its current stage. In the future, if/when sentiment analysis comes into play, reading the text to find those reviews can be possible. In the future, the model will be able to pick up positive, negative, or neutral sentiment, but it cannot make judgements or access the credibility of the sentiment.*
- What about cult classics? Those movies might not have high ratings initially but become favorites later.
    - *The EDA and current state of the data does not take "cult classics" into consideration, as there is no true way or determination of what defines or what a "cult classic" is.*
- Does your analysis consider big-budget marketing campaigns that might inflate a movie's popularity?
    - *At the current time, this analysis does not take any marketing campaigns into effect. This has potential to be added on in the future, when financial and other data is added from additional sources.*
- Isn't it all about big stars? Won't a movie with A-listers automatically do well?
    - *The influence of star power is a good question. The analysis will show what cast is featured in the movies, and there is a possibility that there is a strong correlation between A-list stars and movies that have high scores and rating. This is one of the many factors being looked at and a big name isn't always a guarantee of success.*
- What about foreign language films? Can your model handle those?
    - *This all depends on the data source and chosen features used to represent the dataset. The model and analysis will only show movies that are in the dataset, and if foreign language films are on IMDb and show up in the dataset, they will be represented and able to be part of the model.*
- Is there anything surprising that your analysis revealed?
    - *No, I think that the analysis that was performed with the dataset, seemed to be around what I was expecting. To me, it was no surprise that comedies and dramas led the way, although, I am a bit surprised to see not more action movies listed near the top of the ratings and votes.*
- Can horror movies be critically acclaimed?

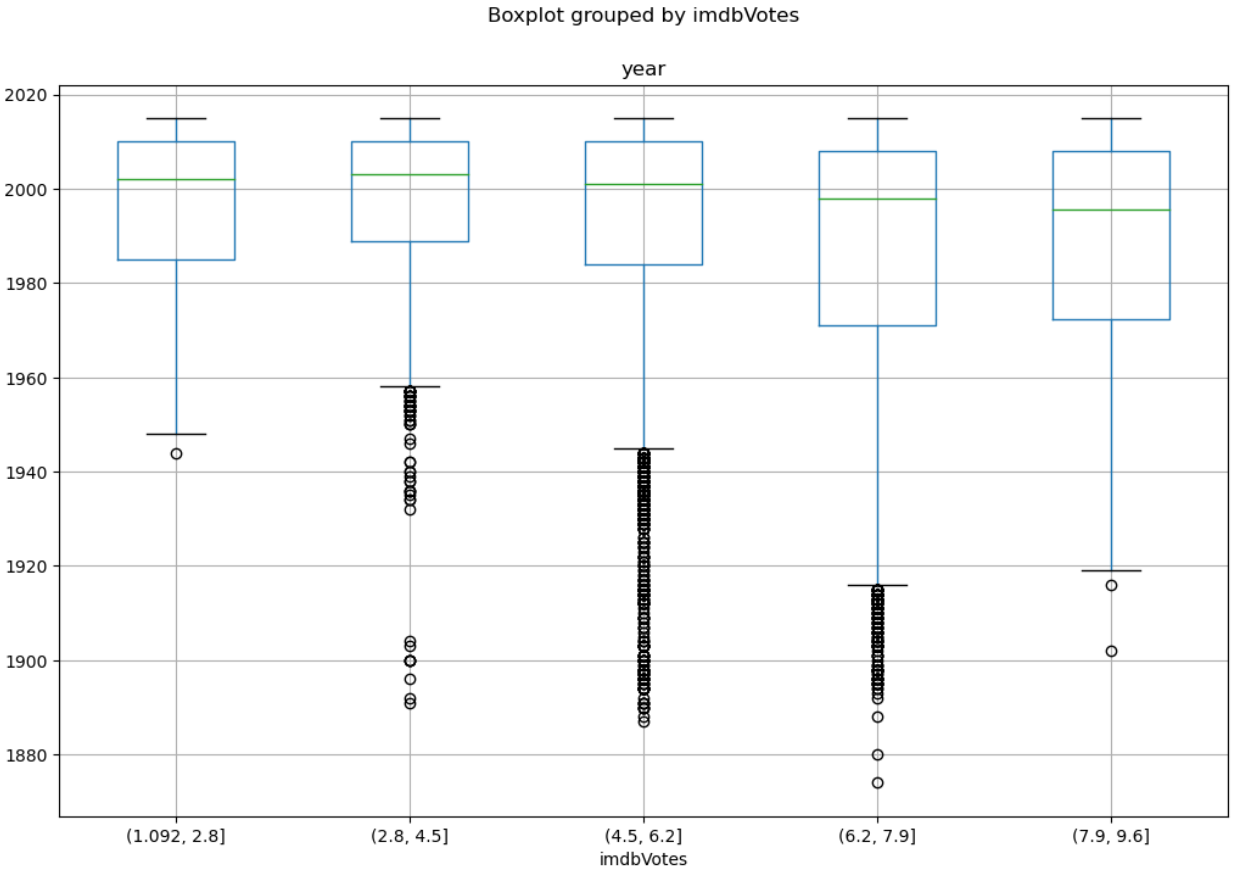
    - *It all depends on what the analysis shows in the ratings and votes for movies in the horror genre.*
- Are documentaries included in your analysis? How do they compare to other genres?
    - *This goes back to foreign language films, documentaries get the same treatment as other categories and if they show up in the dataset, they are represented and able to be put through the model.*
- This is all about reception, but what about the impact of movies? Can your analysis tell us anything about that?
    - *The current analysis only focuses on reception metrics and the data from IMDb, but in the future, analysis can be added to explore the societal or cultural impact of movies using different data sources and metrics.*

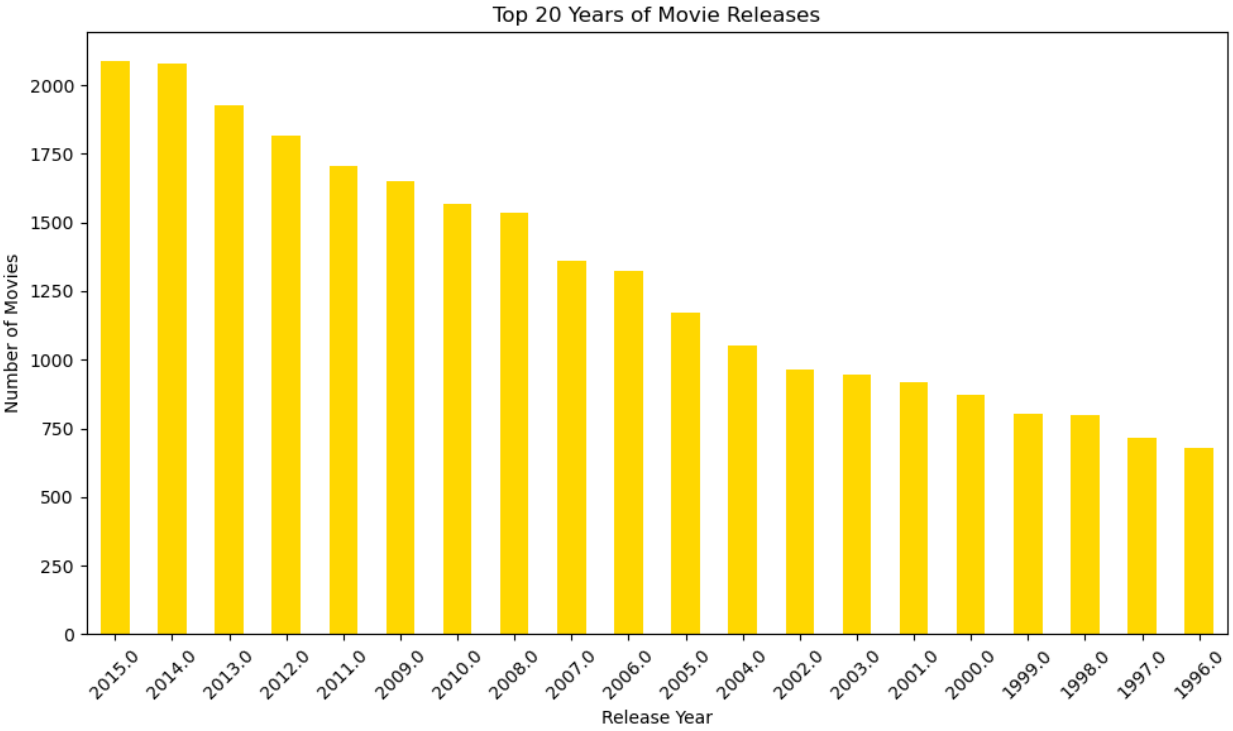
References:

- IMDb: https://www.imdb.com/
- Box Office Mojo: https://www.boxofficemojo.com/
- Kaggle: https://www.kaggle.com/code/robinjrjr/imdb-movie-data-analysis

Note: This proposal is a starting point and may be adjusted as the project progresses.


Correlation Matrix of Movie Attributes
imdbID
metacritic
imdbRating
imdbVotes
1
0.017
-0.028
-0.032
0.017
1
0.71
0.15
-0.028
0.71
1
0.15
-0.032
0.15
0.15
1
imdbID
metacritic
imdbRating
imdbVotes
1.0
0.8
0.6
0.4
0.2
0.0


Boxplot grouped by imdbVotes
year
2020
2000
1980
1960
1940
1920
1900
1880
(1.092, 2.8]
(2.8, 4.5]
(4.5, 6.2]
(6.2, 7.9]
(7.9, 9.6]
imdbVotes


Top 20 Years of Movie Releases
Number of Movies
2000
1750
1500
1250
1000
750
500
250
0
2015.0
2014.0
2013.0
2012.0
2011.0
2009.0
2010.0
2008.0
2007.0
2006.0
2005.0
2004.0
2002.0
2003.0
2001.0
2000.0
1999.0
1998.0
1997.0
1996.0
Release Year