Cliffhanger - cpbryant, lymeraga, felia
Socio-Historical Context and Impact Report

Socio-Historical Context

Understanding the socio-historical context when analyzing movie data is important. A portion of our project involved analyzing the popularity of movies which is affected by societal factors. Economic conditions like recessions affect movie budgets which then affect popularity. Our data suggests that movies with higher budgets perform better from the popularity perspective. Furthermore, globalization also influences film themes due to interconnected markets. We saw that the language of a movie is not particularly a significant factor of popularity. Social movements can influence movies that start broader debates. When a movie discusses a “hot topic” people will be more likely to view it which brings into play our linear regression graph which visualizes vote_count versus vote_average rating of movies. Overall, considering all these factors helps us see how society shapes movies’ performance.

Our project affects various stakeholders. Firstly, it impacts movie industry professionals by providing them with insights into audience preferences and budget allocation strategies. Film critics and reviewers may be affected as our findings might challenge common perceptions about average audience ratings and language influences on movie ratings. On the other hand, general audiences could benefit from more informed viewing decisions based on our analysis. Those who might be harmed by our findings include filmmakers or studios relying on traditional assumptions who may need adaptation based on the revealed statistical relationships. Overall, our research findings have implications for industry practices and audience reception.

Sahu et al. (2022) suggested that although there is a lot of money put into movies, few of them achieve success. The study focuses on addressing the challenges of predicting the success and audience appeal of movies at the early stages of production. Their objective is to develop an expert system that can accurately predict the popularity and target audience of movies using historical movie data from TMBD and IMDb. This “expert system” is meant to be used as a tool to optimize production strategies and maximize success and it uses key attributes to create a prediction.

Ethical Considerations:

Underlying historical or societal biases may inadvertently be present in our data as the dataset may reflect existing biases in the film industry, such as gender or racial bias in casting, directing, or budget allocation. Movies featuring or created by underrepresented groups might receive lower budgets or audience ratings not due to quality but due to societal biases.To counteract these biases, we could include a normalization or

weighting mechanism in your analysis that accounts for known industry biases, or perform subgroup analyses to explore how movies from different demographics perform relative to each other. The systems and processes used to collect our datasets from sources might have inherent biases. For example, user ratings on websites are typically from more tech-savvy and possibly younger demographics, which may not accurately represent broader audience opinions.

Some biases that might be present in the interpretation of our data may come from how we gathered and cleaned our data. We made the choice to join three datasets with different attributes by completely removing any movies (rows) that had any missing values. This may present bias in that the movies we included in our dataset were likely inherently more popular, which would explain our finding that the average vote rating was above what we expected.

The data is being used in a manner agreed to by the individuals who provided the data. We got our datasets off of Kaggle, which is a collection of publicly available datasets that can be used in data science projects such as this. The original data also comes from IMdb data that is free to the public.

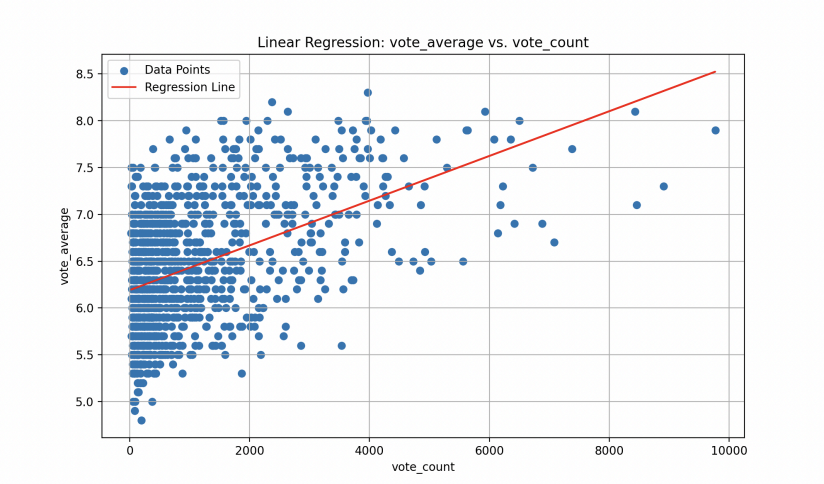
Our project findings may be misused to reinforce stereotypes about what kinds of movies are "successful." For example, implying that a movie needs to have a high budget to be successful. It's important to emphasize that correlation does not imply causation and that multiple factors influence movie success. Providing context and limitations in our report can help prevent misinterpretation.

References

1. Sahu, Sandipan, et al. "Movie popularity and target audience prediction using the content-based recommender system." IEEE Access 10 (2022): 42044-42060.
2. D, Lucas and K, Hurbert and Adeoye, Emmanuel, Ethical Considerations in Movie Recommendations (January 22, 2024).
3. “Recommendation Systems: Ethical Challenges and the Regulatory Landscape.” *Holistic AI*, www.holisticai.com/blog/recomedmendation-systems.

Visualization Component

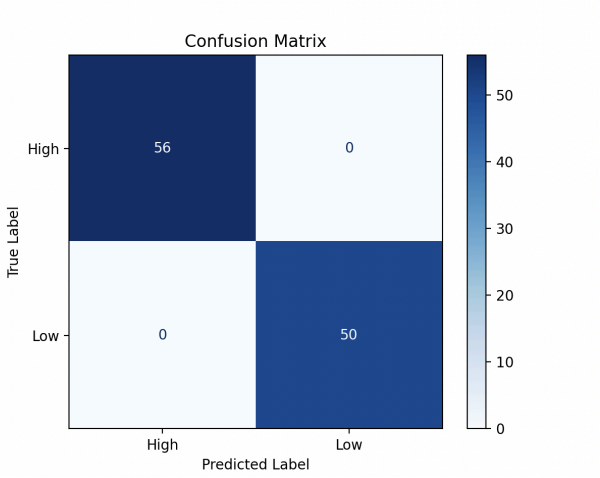
1. Linear Regression



- We chose to use a linear regression graph to explore the relationship between the variables vote_count and vote_average. By plotting vote_count on the x-axis and vote_average on the y-axis we can see whether there is any correlation between a movie's popularity and its average rating. The linear regression line on this graph helps us notice that overall as vote_count increases the vote_average increases as well, so it helps us indicate that there is a trend.
- Instead of a linear regression graph, we could have also used a bubble chart to plot the same information. Similarly, like the visualization we chose, we would again have vote_count on the x-axis and vote_average on the y-axis. The size of the bubbles plotted at any coordinate would give us additional information about the count of movies with those specific values. If we see a large bubble at a particular point (x, y), it means that many movies in our dataset have both vote_count around x and vote_average around y. This way we can identify regions of the plot where movies are more concentrated or sparse based on the two mentioned variables.
- Given that our dataset is large, it is a bit of a challenge to identify individual points because they sometimes overlap. That being said, we made sure that our graph is still easy to interpret and that viewers will be able to recognize the trend.
- The visualization itself is designed to be as standalone as possible, conveying the main trend and relationship visually. However, if the audience is not familiar with the dataset or the variables being plotted,

some additional text would be helpful so they could understand the aim behind this visualization.

2. Confusion Matrix



- We chose to use a confusion matrix to explore the number of false/true positives and false/true negatives of movies classified as either low or high popularity. The confusion matrix provides us with a clear and structured way to visalize the performance of our binary classification model. It shows the true positives (correctly predicted high popularity movies), false positives (incorrectly predicted high popularity movies), true negatives (correctly predicted low popularity movies), and false negatives (incorrectly predicted low popularity movies). This breakdown helps in understanding how well the model is performing in terms of predicting movie popularity.
- Instead of a confusion matrix, we could have also used ROC curves which are a different way to check how well a binary classification model works. They show us how often the model correctly identifies positive cases (like high popularity movies) versus how often it incorrectly identifies negative cases (like low popularity movies) across different cutoff points. While a confusion matrix gives us a single snapshot of this information, ROC curves offer a more dynamic view of the model's performance.
- A challenge with visualizing results using a confusion matrix is that the audience needs context in order to be able to interpret the visualization.

The matrix is a concise summary of our model's prediction performance and so understanding what we were measuring is necessary.

- Our visualization includes some text to explain what the predicted labels and true labels are, however, the viewers would need to have a basic understanding of a confusion matrix since we do not provide them with definitions of true positives, false positives, true negatives, and false negatives.