# *Avatar* or *Mortal Engines*? Predicting Movie Profitability From Scripts

Varun Tandon, German Shabanets Enik, Katie Mishra

## Motivation

Hollywood movie studios consider thousands of scripts annually in order to find the few they produce and distribute on the film market. Even so, many films still flop, either failing to return a profit, or being poorly received by critics. Thus, an artificially intelligent tool which processes film propositions could provide insight to movie studios.

## Objective

We sought to engineer an intelligent tool that predicts both the profitability and the potential critical acclaim of a film proposition given raw text of a script, as well as basic information such as casting, rating, and a brief description.

## Approaches

**Data Cleaning/Preprocessing**

- SMOTE Oversampling
- Majority Undersampling
- Feature scaling

**Manually Generated Features (Script Profitability and Critical Reception)**

- Number of words*^
- Number of pages*
- Script Author^
- Title words^
- Word usage*^
- Director^
- Cast^
- Runtime^
- Rating (G, PG, etc.)^
- Genre^

**Models (Manually Generated Features)**

- Logistic Regression
- K-Nearest Neighbors
- LinearSVC
- SVM
- Bagging Classifier
- Random Forest
- Voting Classifier
- MLP Classifier
- Naive Bayes

**Models (Unsupervised Features)**

- BERT
- Doc2vec

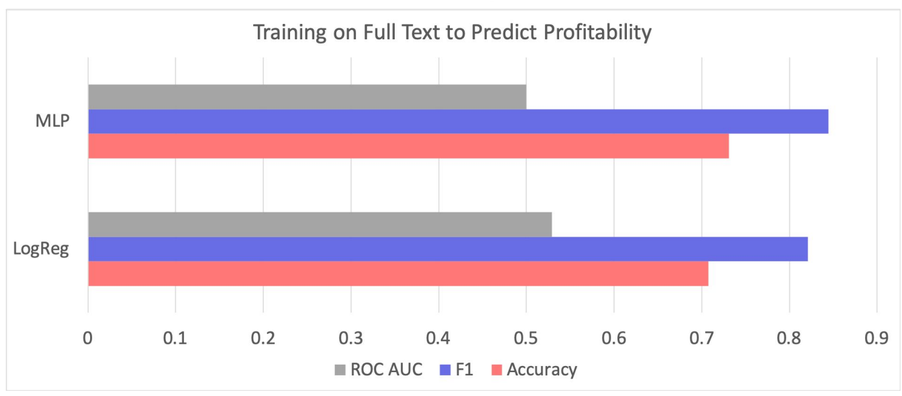
## Challenges

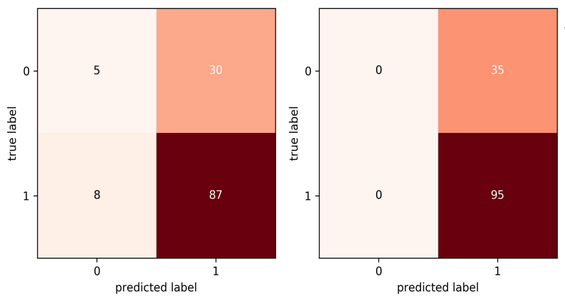
- We began this project with the goal of predicting the profitability of a movie given the script data. Through this process, we discovered that there is little to no correlation between the financial success of a film and the merit of a script. (see Figure 3).
- Furthermore, since the average script length is 120 pages, we faced issues with computational intensity, particularly while parsing and processing our dataset of over 1000 movie scripts.
- With these discoveries in mind, we made the decision to present our current findings, as well as to explore the potential for an artificially intelligent model to predict the critical success of a film given its synopsis. This would still serve as a useful tool for movie studios in evaluating scripts.

## Results

*Trial 1: Trained on full-length scripts**

We first extracted features marked with with * from 1238 full scripts.




Above - Figure 1: performance data when trained on full scripts.

Left - Figure 2: Confusion matrices of the LogReg classifier (left) and MLP classifier (right).
Data indicate that the models achieve high accuracy due to bias in datasets.

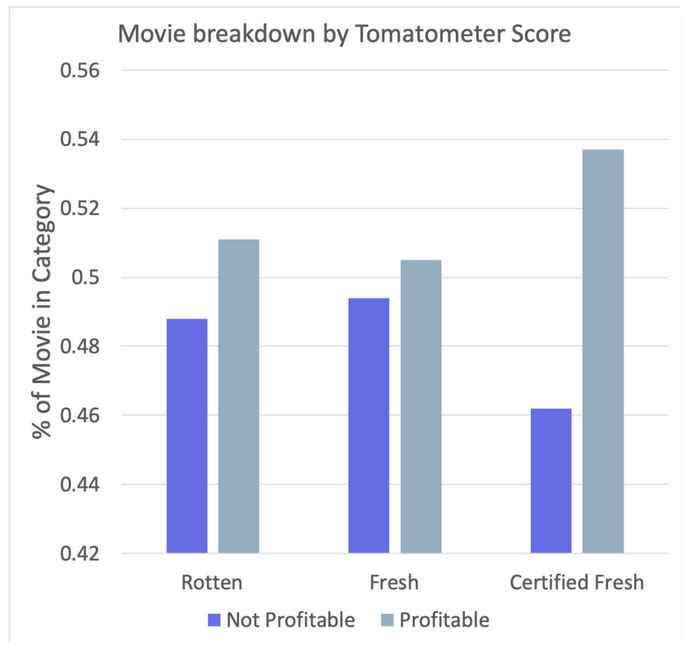



Figure 3: out of the movies that were ranked Rotten, Fresh, and Certified Fresh on Rotten Tomatoes, percent of Profitable and Not Profitable is approximately equal.

*Trial 2: Trained on synopsis + movie info^ to Predict Critical Reception*

After examining the results above, we decided to train on less text data (using synopsis instead of full-length script). Furthermore, we included more predictors beyond text, as can be seen on the left hand side (marked with ^).
We had around 5000 data points.
After extracting features, we trained various models and obtained results summarized in Figure 4 below.

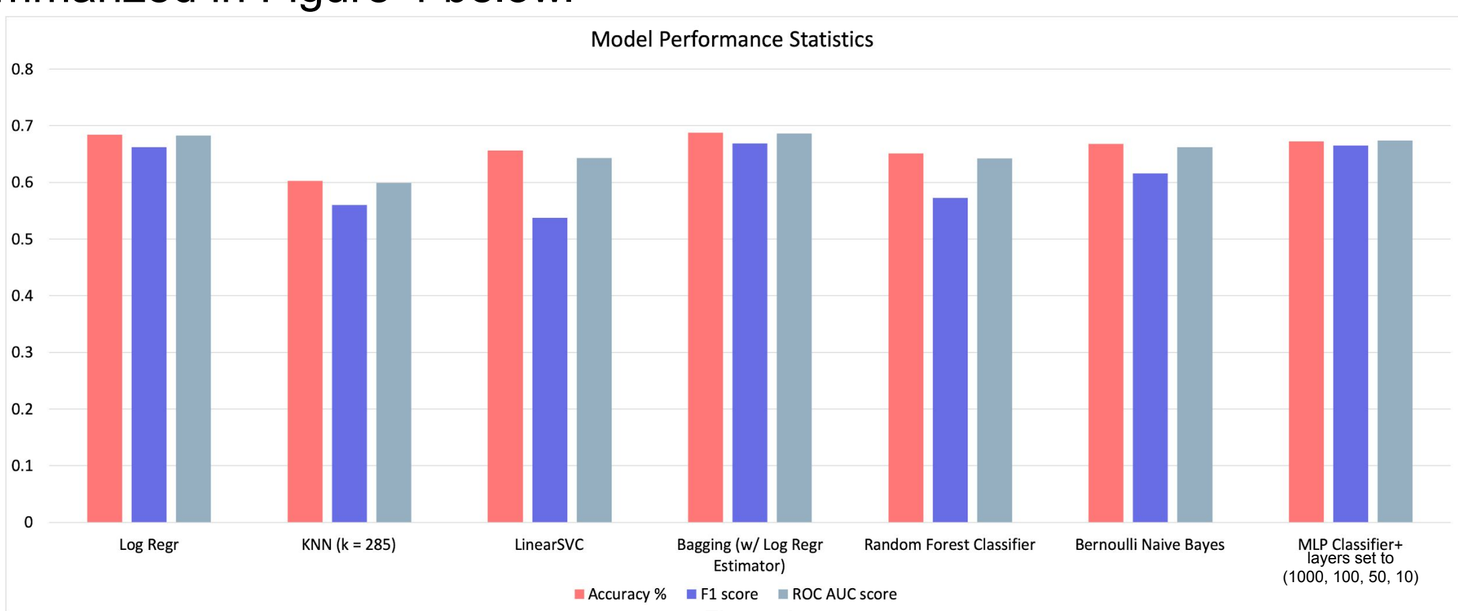



Figure 4

## Error Analysis and Conclusions

- Our initial models had high accuracy as a result of bias in our data (as shown by ROC AUC and confusion matrices).
- Minority oversampling and majority undersampling failed to yield accurate models and the following analysis of our data revealed the content of a script has no correlation with profitability of a movie.
- We found that our selected features and models could still provide insights to studios by informing the potential critical acclaim of a film. We generated generalizable models that achieve a nearly 70% accuracy in prediction of a movie's *Rotten Tomatoes* "Freshness".