Predicting Movie Success from Screenplays to Make Assessments about the Film Industry

Ambika Acharya **Nicole Crawford** **Nadav Lidor**
{aacharya, nicolecr, lidor}@stanford.edu
Stanford University, Department of Computer Science

Abstract

Predicting movies' box office success and viewers' rating based on screenplays alone is both an interesting real-world problem, and one that is directly related to a wide array of natural language classification tasks. In this study, we explore textual and pre-production features to predict a movie's genre, IMDb rating and box office success. We obtain moderate results for all three tasks, with a peak accuracy of 88% for rating prediction. From our analysis, we find that genre plays a significant role in predicting IMDb rating and box office success. Passing the Bechdel test correlated positively with box office success but negatively with ratings, having male leads correlated positively with box office success while having a female lead correlated negatively, showing that gender appears to be significant factor for both tasks. This work deepens our understanding of the movie industry and viewers' expectations.

1 Introduction

The most important question movie producers ask when assessing a production is whether the film will succeed financially or critically. Indeed, there is no formula for addressing this question and the types of movies that tend to succeed in the box office often differ from those that are critically acclaimed. Yet today there is ample data available of successes and flops to attempt to tackle these questions computationally.

In this study, we predict a movie's success based on its screenplay. To address this task, we used a machine learning approach with a focus on feature engineering and natural language understanding. We divided the task into three sub-tasks: (1) building a model to predict a movie's genre and building models to predict two forms of success, (2) IMDb rating and (3) box office success. Lastly, by analyzing the model's weight assignment, we discuss social implication and insight we gain about the movie industry.

1.1 Genre Prediction

The task of genre prediction is not only an interesting task related to sentiment analysis and essay scoring, but is also an interesting gateway screenplay textual analysis for future work. This was a multi-class classification task where we classify a movie as one of six distinct classes: drama, comedy, thriller, documentary, action, sci-fi. This constitutes a challenging classification problem as movie genres are generally fluid and not well-distinct, and can often overlap to form specific genres like dramedy (drama-comedy) or rom-com (romantic comedy). We mitigate that by limiting the number of classes and grouping highly similar genres. We then analyzed the resulting scores for this task and looked at the weights of our final model to discern which features were better at indicating differentiation than others. We further use genre as a feature for movie success, which proved highly informative.

1.2 IMDb Rating Prediction

The user rating of a movie is generated from user submission on the IMDb website. The IMDb movie rating is one which is widely used as a metric of audience reception, and is based of thousands of raters, thus providing a sound metric of success for our study. We note that our concern is popular viewer sentiment and not critical acclaim. Indeed, IMDb ratings reflect viewer reception, and often are not directly correlated with critics' views or official award scores. For our task, we defined two classes: movies with an IMDb rating of 0 - 5 are considered 'bad' films, and those with a score

between 7.5 - 10 are good. We decided to split the two classes with a significant score margin to handle the ambiguity of closely related mid-scores.

1.3 Box Office Success Prediction

Anticipated box office success is often a critical factor for a production, with revenues driving the film industry. We obtained box office metrics for a subset of our corpus films, and used that to predict financial success, defined as a movie making at least double its budget (domestic box office reported). Any film that did not at least double its budget was considered a flop. Other than predicting movie financial success, a sub-goal was to compare and contrast the features of box office success with those of the IMDb rating task to learn of the relations between a good IMDb rating and film's grossing. One reasonable hypothesis is that films with mediocre or bad screenplays but lots of commercial hype would receive higher box office success and lower IMDb ratings, and vice-versa for good screenplays for, as an example, independent films.

2 Prior Literature

The tasks of predicting genre and movie success with respect to rating and box office have been studied in various forms throughout the literature, directly as well as indirectly as part of other common NLP problems such as sentiment analysis and summarization. We now introduce a few major papers in the field.

With a clear economic incentive, the most studied movie-related problem is predicting box office success. While most works make use of both pre-production textual features and production non-textual features, such as movie budget or "star" selection, (Eliashberg et al., 2014) is the only work we are aware of that used only pre-production script features to predict movie success. They defined a similarity metric to compare new scripts to previously produced movies in order to determine whether the new script should be green-lit. Using textual features based on movie scripts, they removed stop words, used a stemming algorithm, and an importance index, described as similar to tf-idf weighting. They then used two different kernel algorithms: Kernel-I had all features of equal weight, and Kernel-II used unequal weights. They compared these results to choosing films based solely on similarity with respect to genre and budget, as well as other algorithms including multiple regression. Both kernel algorithms performed better, indicating they picked the highest return on investment sets of films.

Another related study by (Lindner et al., 2015) explored the effects of female characters on box office success. Using a few simple measures, including the Bechdel Test, to capture whether a film has 'an independent female presence', they examined over 1000 films to find that 64.4% of the top-grossing films between 2000-2009 did not pass the test. We do not deduce a causal relationship, but rather an overall picture of Hollywood's current relationship with gender equality.

An additional set of interesting features can be extracted through common Network Analysis tools. (Elvevåg et al., 2007) showed that a word graph based on a movie script can produce insight with respect to the film's box office success. In particular, looking at 170 films, they found a significant correlation between the size of the main connected component of a film and its box office success, and proposed other graph related metrics, such as degree centrality and betweenness, to continue this line of research.

As noted, many studies incorporated production and often post-production features. For example, (Eliashberg et al., 2007) examined whether movie spoilers can be used to predict return on investment (ROI). Primarily employing bag-of-words features, they achieved 61.7% accuracy. However, as spoilers are most commonly written by people who have already seen the film, i.e. post-production, this work is primarily less concerned with indicating a movie's future success.

On the task of genre prediction, (Blackstock and Spitz, 2008) use a variety of textual and voice features, including average length of speech by a character, ratio of personal pronouns, mentions of locations (cities, countries) or organizations (FBI, KGB, etc.) and more. They found that the most significant features for genre prediction were the normalized number of personal pronouns and the average length of character's speech. Similarly, (Ashok et al., 2013), showed that structural features such as parts of speech, presence of punctuation, concentration of proper nouns, words per sentence, and word frequency are all significant in this prediction task. Another study by (Hunter et al., 2016) examined writing styles to predict a novel's success. They explored POS tagging,

number of phrases and clauses, grammar rules and bigrams and unigrams. They found that novels with more noun phrases tend to be successful while those with verb phrases are less so. In addition, they identified connectedness (words like with, together, etc.) and negation to be associated highly with successful pieces of work.

3 Data

3.1 Corpus

The Cornell Movie-Dialogs Corpus contains 220,579 conversations between a total of 9,035 characters from 617 movies. This corpus also includes details such as a movie's associated genres, as well as each character's gender and position in the credits. Approximately a third of the characters in the corpus had a gender listed. As for position in the credits, we assumed, as the corpus authors did, this indicates the level of importance of a character in their movie, meaning that a character listed first in the credits has a larger role in the movie than a character position at the tenth spot.

We performed a 80-20 split on the data for our train and test sets, and then another 80-20 split on our train set to create our train and dev sets.

3.2 Scraped Sources

To get data on which movies pass the Bechdel Test, we scraped `http://bechdeltest.com/`. Approximately 456 movies were found on the Bechdel Test website. In order for a movie to pass the Bechdel Test, it must 1) have two named women, 2) the women must talk to each other, and 3) they must talk about something other than a man.

We also scraped `http://boxofficemojo.com` to get information on movie budget and domestic box office results. 247 movies were found on the Box Office Mojo website. When using these attributes as features for our model, we limited our analysis solely to the movies that had a result from the respective website.

3.3 Parser Classes

After parsing the corpus data, we created different classes to easily analyze the various attributes of our data. This includes Movie, Character, and Line classes. Our setup made running our models and analyzing our data much easier, and also made it simple to add additional features or attributes to movies or characters.

4 Model

4.1 Classifier

4.1.1 Genre Prediction

For this task we used a multi-class Support Vector Machine (SVM) with a RBF kernel. This implements a one-against-one approach to the multi-class problem.

4.1.2 IMDb Rating and Box Office Success Prediction

For both these tasks, we implemented two classifiers, logistic regression and an SVM with an RBF kernel. We compared the performances of these two models to determine the best classifier for this task with our feature set.

4.2 Features

A major component of this study was the semantic and feature-level analysis of the tasks, and thus we spent much time building and tuning our feature set. The following is a description of the feature engineering we worked on. The number following the name of a feature is its feature id, which is referenced in subsequent sections and figures of this paper.

4.2.1 High Level Features

We began by looking at the metadata of a movie and built features surrounding the attributes of a film such as the characters, gender, length, and genre. We summarize the intuition behind the features we constructed for our final model, though we experimented with many more.

- **number of characters (1)**: The number of characters in a given film. We use the raw counts and believed that this might help differentiate between genres, since we often see comedies with large casts and dramas with few antagonists.

- **main character gender (2):** The leading character's gender, if indicated. We were interested in finding correlations between gender and the success of films, and added this feature as a way to investigate this question.

- **male characters (3):** The proportion of male characters in a film out of the total number of characters in the film. There were data gaps

with this feature as some characters appeared in our data without a gender label. In this case, we added them to the total number of characters but didn't add them as male or female characters.

- **female characters (4):** The proportion of female characters in a film of the total number of characters in that film. See "male characters" for explanation and reasoning.

- **two female leads (5):** A binary feature indicating if the gender of the first two leading characters are women. We realized that a lot of films do not have women in leading roles and were interested to examine how films lead by female characters performed at the box office. We also speculated this could help in genre prediction because movies with two females leads might fall under similar genres.

- **average line length (6):** The average number of words in a line for a given film. Dialogues tend to have shorter lines while monologues have one character speaking for a long time in one line. We thought this feature would help differentiate between "fast-talk" genres, e.g. comedies, and slower, more wordy genres such as dramas or documentaries.

- **number of lines (7):** The number of lines in a film. This might help differentiate genres with longer movies from those with shorter films.

- **genre:** For the rating and box office tasks, we speculated genre might be a good indicator. We added a binary feature for each genre (is-comedy (8), is-action(9), is-thriller(10), is-sci-fi(11), is-doc(12), is-drama(13).

4.2.2 Semantic Features

After looking at the high-level attributes of a film, we focused on a semantic analysis with token-based features.

- **bechdel test (14):** A binary feature indicating whether or not the film passes the bechdel test.

- **vocabulary size (15):** The number of unique words used in a film. We thought that more complex films, such as dramas and documentaries would have a larger vocabulary size than comedies.

- **unigrams (16), bigrams (17):** These are large feature vectors with all of the words across the movies in our data set and the count of each words in the given movie. For bigrams, we used all the bigrams across the movies in our data set instead. We normalized over the total number of words in that movie. We thought this was a good way to represent a film in the vector space.

- **pronouns (18):** We used a part of speech tagger to label all the words in a film and then calculated the proportion of pronouns. Movies with more use of the first person might be distinguishable from those without.

- **exclamation points (19), question marks (20):** The number of exclamation points and the number of question marks per movie, normalized by movie length. We thought this would be particularly helpful in differentiating genre.

- **sentiment (21):** We used a sentiment corpus to determine the proportion of positive words in a movie to the negative words.

5 Results

5.1 Model Metrics

Here we present the results of the machine learning models and the feature weights we generated. We discuss the implications of these results in the analysis section.

5.1.1 Genre Prediction

Our accuracies for this task are compared against a baseline of 16%, chance-level for a 6-class classifier. Because we used a 6-class classification, it was challenging to predict classes even after equalizing the number of examples we had for each class. Without unigram and sentiment, we had a prediction accuracy of 19%, which was slightly above baseline. However, we got peak results when including unigrams and sentiment. With these semantic-level features our accuracy increased to 30%. Because this is a 6-class problem, we see this as a significant gain.

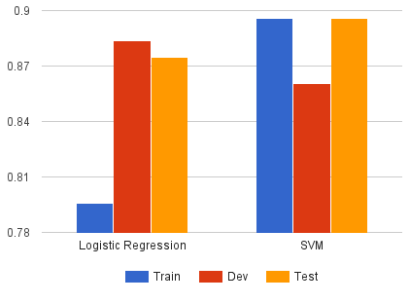

Figure 1: IMDb Rating Accuracy

5.1.2 IMDb Scores Prediction

Figure 1 presents a summary of the models' accuracies on our training, development and test sets. They are compared against a baseline of 50%, chance for a binary classifier. We found that the SVM performed the best on both the dev and test sets, at 0.86 and 0.89 accuracy values respectively.

5.1.3 Box Office Success Prediction

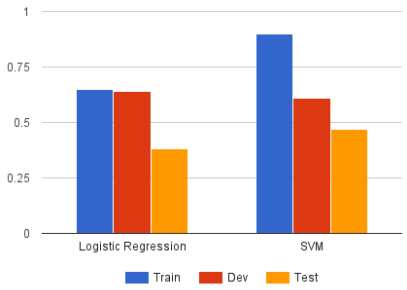

Figure 2: Box Office Accuracy

Figure 2 shows a summary of the models' accuracies on our training, development and test sets. They are compared against a baseline of 50%, chance for a binary classifier.

6 Analysis

6.1 Limitations

The dataset available was fairly limited in a number of ways. Our initial corpus was of only 617 films, with a maximum number of training examples of at most 394 for genre prediction. However, this number was significantly reduced for IMDb ratings and box office success, as many films did not have this data readily available and so were neglected. Here, we found that our dev and especially our test accuracies decreased from the training.

Secondly, our corpus contained only partial scripts, a subset of the movies' actual dialogue. We do not know how these partial scripts were sampled, but this fact can very well effect performance. This is particularly important for conversation long features, were we track pronouns for example. Additionally, some data for feature extraction was only partially available, such as character gender, actor's position in titles (signaling character's importance), and Bechtel test results.

Most movies in our dataset had multiple genres listed. However, to predict genre, we had to assign each movie to a single genre class, although many movies could not easily be classified in this manner (e.g. drama / comedies). We therefore defined six super-genres (e.g. "super-comedy" includes the genres: "family", "comedy", "musical", and "animation") and assigned movies' genre based on max overlap with our super-genres. We note that this is still a potential pain-point, as many film span across multiple genres, styles and cinematic features. In the future we would want to use a new dataset or manually generate the main genre for a film, though we acknowledge that this may be challenging as a film's genre tends to be fluid.

6.2 Genre Prediction

The Sixth Sense

Actual: thriller
Predicted: drama
This film has a lot of conversations about high-level concepts, suggesting a large vocabulary size. Since we believe dramas to be highly correlated with a large vocab, our model predicted this film to be a drama, even though it is really a thriller.

This task was much more challenging than it appeared a first glance due to our having to predict 6 distinct classes. Therefore, the highest accuracy we achieved for this task was 30%, using sparse features such as semantic analysis, bigrams, and unigrams. These features likely performed better than our dense features, like main character's gender, because a movie's genre typically depends on the actual words used versus metadata like the number of lines. For example, a romance is more like to contain words like "love" than a comedy,

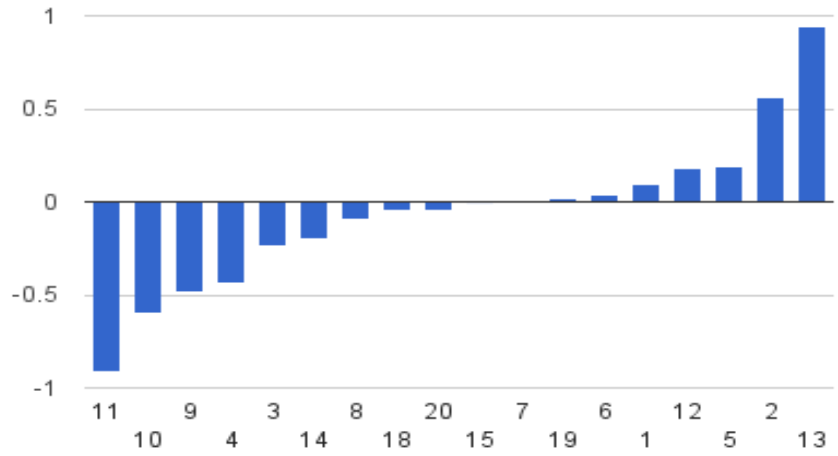

Figure 3: IMDb Rating Feature Weights

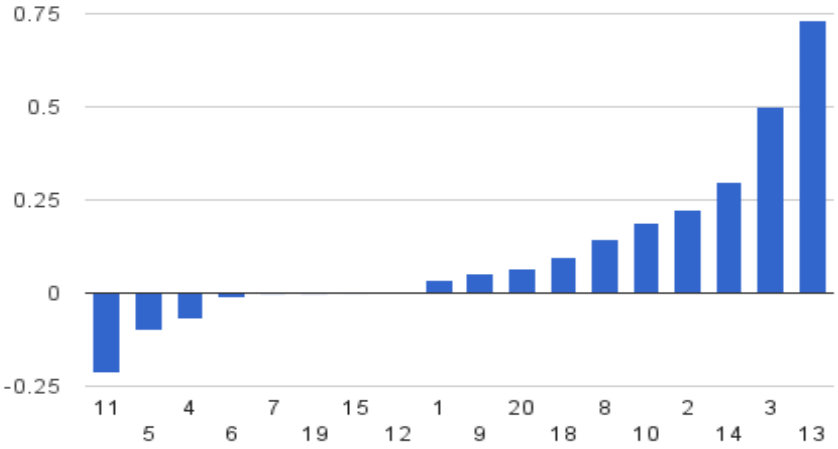

Figure 4: Box Office Feature Weights

but the main characters in both films can have the same gender.

We at first assumed doing this task would lead us to valuable features we could use in our other tasks, but it turned out to be much more time intensive than both IMDb rating and box office prediction. However, we still used features generated in this task for the others.

6.3 Prediction Success

As shown in figures 1 and 2, we reached above chance-level success on all three tasks for our dev set, with IMDb classification reaching 88% accuracy and box office accuracy of 64%. However, we did not reach the same level of accuracy with our test set. We assume this fluctuation is due to a very small dev / test sets, producing uneven results.

Moreover, predicting a movie's financial success based on pre-production features is a particularly challenging task, given that production companies go to great efforts to increase revenues through production and post-production decisions, such as casting, movie and marketing budget, commercial theater distribution and release

date. Finally, as noted previously, genre prediction has proved especially challenging due to the very nature of film genres—a multi-class problem with highly fluid and vague classes.

6.4 Feature and Weight Analysis

Following is the main insight we extracted from analyzing the feature weights for our logistic regression models. We extracted weights for both box office and IMDb rating, but not genre prediction, as we used a non-linear multi-class SVM classifier.

6.4.1 IMDb Rating Prediction

As shown in figure 3, the highest positive weighted features, making a movie more likely to be classified as 'good', are (in order of importance, parentheses indicate feature index in chart) genre-drama (13) and main-character-gender (2). The highest negative weighted features, making a movie more likely to be classified as 'bad', are (in order) genre-thriller (11), genre-scifi (10), genre-action (9). We note that some features that had negligible weights, such as the length of the script (7), number of exclamations (19) and movie vocab size (15). It is clear that the genre of a film has a significant impact on IMDb rating, with dramas ranking highest, while thrillers are commonly low-rated. We note that passing the Bechdel test (14) is negatively associated with a high IMDb score, despite our intuition. A related recurring feature is gender in the film, particularly that of the main characters.

Thelma and Louise

Actual: rating higher than 7.5
Predicted: rating lower than 5
It's likely that we predicted this movie to have a low rating because it has two female leads. Our model downweights this feature, since movies with two female leads have been shown to have low IMDb ratings. Even though this is a critically acclaimed film, our model predicted it to have a low rating.

6.4.2 Box Office Prediction

From Figure 4, the highest positive weighted features, making a movie more likely to be classified as 'good', are (in order of importance) genre-drama (13), male-characters (3) and Bechdel test (14). The highest negative weighted features, making a movie more likely to be classified as 'bad', are (in order) genre-thriller (11), two-female leads (5), female-characters (4). The following features did not have significant impact on our model: genre-documentary (12), movie-vocab-size (15) and number of exclamation points (19).

As in IMDb rating, it is evident that the genre of a film has significant impact on box office success rating, with dramas rank once again highest, while thrillers have higher probability of being failure. This could be due to the significant budget costs—dramas are often much cheaper production as compared to thrillers. Unlike IMDb ratings, passing the Bechdel test (14) is now positively related associated with a high IMDb score, despite our initial intuition. This shows that although we often hear of women playing stereotypical male-obsessed characters in Hollywood movies, movies that do show women in a better light do perform well at the box office, but audiences do not rate them highly.

Saving Private Ryan

Actual: success
Predicted: fail
It's likely that we predicted this movie as a flop because it doesn't pass the Bechdel test, which our classifier weights as a very important feature for box office success.

10 Things I Hate About You

Actual: success
Predicted: fail
This film was likely predicted to be a box office failure because it's a comedy, which our model down-weights heavily.

7 Conclusion and Future Work

We built machine learning models to predict movie genre, IMDb ratings and box office success based solely on pre-production textual features. Our study achieved moderate results, despite multiple challenges, most importantly due to partial data and a small movie corpus. Surprisingly, genre prediction proved to be particularly challenging due to the imprecise nature of movie genre as a human defined classification, and movies often classified under multiple genres. Both IMDb ratings

and box office success are hugely impacted by production and post-production decision, from casting movie stars, to budgeting and cinematographic choices.

In addition, it was interesting to see the differences between a successful movie on IMDb and at the box-office. Though we saw gender and genre playing a role in both, the fact that movies featuring women in lead roles or as a dominant part of the cast trend towards having positive IMDb ratings and are highly correlated with poor box office success says something about the film industry. Gaining such insights into Hollywood was a unique experience.

Going forward, a set of features we were not able to explore are dialogue based. Given that movie scripts are composed of conversations, it should be beneficial to examine features such as conversation dominance, how conversations are impacted by gender, and how pronouns are used through multi-turn dialogue. For this sake, we need full movie scripts and scene labels, data not currently available to us.

Lastly, we want to acknowledge that while our project might seem to aim at creating the formula for a perfect movie, we don't think this is possible nor something the movie industry should aim for. The fluidity of genre and unpredictability of success based on a screenplay speak to the artistry of the field, something which keeps both viewers and critics going back for more.

Acknowledgments

We thank Christopher Potts for his enthusiasm and great feedback throughout this project. We also thank Bill McCartney and the entire CS224U course staff.

References

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. *Poetry*, 580(9):70.

Alex Blackstock and Matt Spitz. 2008. Classifying movie scripts by genre with a memm using nlp-based features.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Jehoshua Eliashberg, Sam K Hui, and Z John Zhang. 2007. From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, 53(6):881–893.

Jehoshua Eliashberg, Sam K Hui, and Zhongwei Jake Zhang. 2014. Assessing box office performance using movie scripts: A kernel-based approach. *Knowledge and Data Engineering, IEEE Transactions on*, 26(11):2639–2648.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1):304–316.

III Hunter, Starling David, Susan Smith, and Saba Singh. 2016. Predicting box office from the screenplay: A text analytical approach. *Journal of Screenwriting*, 7(2):135–154.

Andrew M Lindner, Melissa Lindquist, and Julie Arnold. 2015. Million dollar maybe? the effect of female presence in movies on box office returns. *Sociological Inquiry*, 85(3):407–428.

Julia M Taylor. 2010. Ontology-based view of natural language meaning: the case of humor detection. *Journal of Ambient Intelligence and Humanized Computing*, 1(3):221–234.