Examining Bias in Public Opinion of Films

Wiley Battles, Grace Brasfield, Robert Kullberg

The University of Tulsa

02 May 2023

Abstract

 Our project takes a data mining approach to predict which factors will result in a film being highly rated by viewers. The predictors that we planned to assess are from IMDb datasets. The predictors are genres, actors, and directors. We created bins for the terrible, poor, average, and excellent rating levels. For analysis, four different models were utilized: logistic regression, naive Bayes, random forest, and bagging. The results found show that films have too many factors to accurately predict success based on the previous performance of other films. Each of the models used was very similar in accuracy. The coefficients for the predictors found in the models were largely unmeaningful. The four bins were our target variable, with a more specific focus on the excellent bin. The model performed proved inaccurate overall, as it demonstrated an insufficient ability to classify positive cases. This was due to an imbalance in the data. Most film ratings fell into the average bin, with very few landing in excellent.

Table of Contents

[**I. Introduction 4**](#_wn0totxx7ev9)

[**II. Analysis Approach 4**](#_i1pjndti0f3i)

[A. Data Processing 4](#_756232ffdsbm)

[B. Methods 5](#_qtr6g3nrq4lp)

[**III. Results 7**](#_5bl878xcildu)

[**IV. Summary and Conclusions 10**](#_4t34s7m0pxx9)

[**References 12**](#_ncha0kvkcu8u)

[**Appendices 13**](#_eh7fst2p374j)

[Tables 13](#_u6986j44ax10)

[Figures 13](#_fyr7ms5jqou1)

### Introduction

 IMDb (Internet Movie Database) is a popular website that gives information about movies, TV shows, and video games. It provides information like directors, actors, run time, and plot overviews. IMDb not only provides information about the films, but it also gives users the ability to rate them. These ratings are the basis for our analysis. With millions of user ratings, IMDb is one of the most widely used film rating sites.

 Public opinion is a major factor in determining a movie's overall success in the film industry. Because the public opinion of films is so subjective, we wanted to create a model that would help predict the factors that most contribute to the success of a film in terms of ratings. Such a model would be helpful for film companies to understand which factors that go into a movie are likely to have a large impact on the public perception of the film. There has been some research done using the same datasets we used from IMDb. These works included a Stanford article that examined movie factors on financial performance, a 2016 article that predicted box office success, and other studies with a narrower focus on the data (for example, data from only 2018 or data from only Russia). Our project differs from other research by examining movie traits, such as actors, genre, directors, and run time, to predict overall success. We also have a broader focus including movies from multiple years. Our approach to examining the impacts of these movie traits on overall public opinion is to create four models, logistic regression, naive Bayes, random forest, and bagging. We will use a supervised learning approach to predict the success of films in terms of ratings.

### Analysis Approach

#### A. Data Processing

 For our project, we used several publicly posted datasets that IMDb updates daily. Multiple datasets were consolidated to serve as the basis for analysis. We accessed the five IMDb datasets below on 3/3/2023, and the following variables were taken from each file:

| **Dataset:** | **Information Provided:** |
| --- | --- |
| title.akas | Language of title |
| title.basics | Year, Title type, Adult film, Runtime |
| title.crew | Director |
| title.principles | Lead actor |
| title.ratings | Ratings, Number of votes |

##### Table 1. Datasets Utilized

 The variables taken from the individual datasets were used to limit to only entries that fit the desired parameters. Non-english language titles were discarded to narrow our focus on the english speaking population. Adult and non-movie content was dropped to confine the data to only traditional film titles. The release year was used to remove any films released before 2000, allowing for an evaluation of current taste. All movies with under 20,000 votes were disregarded as insignificant measures of popular opinion. The final dataset included variables for the alphanumeric identifier for the film, the average rating 1-10, the runtime in minutes, three genres, the directors, and the lead actor.

 The dataset was preprocessed before classification. Genres, directors, and actors were converted into categorical dummy variables for ease of analysis. The movies were then divided into four bins based on their rating. These bins make up the target variables, with a specific focus on the Excellent category.

| **Class** | **Rating** |
| --- | --- |
| Excellent | 7.5-10 |
| Average | 5.9-7.4 |
| Poor | 3.6-5.8 |
| Terrible | 0-3.5 |

##### Table 2. Class Consideration

#### B. Methods

 All models were created using python in a juyptr notebook. Pandas was used for any data organization and dataframe creation. All model creation functions are included in the sklearn module. Graphical outputs were scikitplot or matplotlib.

 Our study utilized four unique prediction techniques in evaluating the data collected from the IMDb datasets. These included logistic regression, naive bayes, bagging, and random forest. Due to the smaller sample size, we used 60% of the data for our training set and 40% for the testing set to avoid overfitting. We employed the models to predict the probability of a title's inclusion within the Excellent bin, represented by a binary categorical variable. This allowed us to create models that predicted whether or not a movie would have an optimal high viewer opinion.

*1) Logistic Regression*: Logistic regression was chosen due to its usefulness in determining binary factors. Logistic regression has its basis in the comparison between the odds of an event and the odds of an event not happening [[1]](https://www.ahajournals.org/doi/epub/10.1161/CIRCULATIONAHA.106.682658). Each factor is evaluated through its effect on the likelihood. The algorithm uses the coefficients to calculate the odds of each outcome :

ln(odds)=β0+β1X…βnX

The binary classifications of the model made it a perfect fit for our data.

We chose to fit our data with a max iteration limit of 1,000,000 due to the technological requirements of the subject dataset’s size. A SAGA solver equation was also used due to the large amount of data and computational limits of the machines available.

*2) Naive Bayes*: Naive Bayes is a classification model based on the Bayes theorem. It differs from the other implementations of the formula in that it assumes all attributes are independent of each other [[2]](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8944604). The “naive Bayes assumption” simplifies the task, allowing for datasets with larger numbers of attributes [[3]](http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf). This model was chosen for this ease of implementation due to the substantial amount of variables created in our data.

We chose to use a multinomial naive Bayes model, due to its increased performance and reliability with data [[3]](http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf). We then evaluated our model using a grid search technique to determine the optimal hyperparameters. This returned an alpha of 0.5 and stated that the model should consider class prior probabilities.

*3) Bagging*: Bagging (bootstrap aggregating) creates subsets of the data and trains various models based on those new samples. Paired with data prediction models, it can be a highly useful tool in dealing with complex problems with high-dimensional data [[5]](https://www.jstor.org/stable/1558692)*.* Bagging is particularly useful in reducing the bias found in decision trees [[5]](https://www.jstor.org/stable/1558692)*.*

We used a decision tree model and specified the bagging hyperparameters. The maximum samples were set to 0.5, and the number of generated samples was limited to 1,008. Based on these, we constructed the final prediction model.

*4) Random Forest*: Random forest consists of grouping numerous generated tree-structured classifiers into a single prediction model. The model is centered around analyzing independent vectors [[2]](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8944604). Various trees created from subsets of the data vote on which outcome will make up the final prediction. Random forest has an added layer of variability, as each split trys a random subset of the features. This additional functionality helps to further prevent overfitting [[4]](https://link.springer.com/article/10.1023/a%3A1010933404324). Using a random forest allowed us to reflect the large degree of variability in the data.

Our random tree had a no depth limitation with a minimum leaf sample of 1 and a minimum split sample of 5. These resulted in highly extensive and close-fitting trees. A total of 100 trees was generated for each forest.

### Results



###### Figure 1. Classification Accuracy by Model

 All models performed with a similar level of accuracy in predicting class membership to the Excellent bin. Though the methods behind each model vary significantly, the predictive capabilities generate similar responses. Therefore no meaningful differentiations can be made on the basis of accuracy alone, and there must be some factor influential enough to affect each of the different models. As each technique handles the data in a unique way, an outside factor must be forcing similarity.

 Measuring the models by the root mean squared error also returns the same level of similarity. Each model performed within a narrow window of error, with numbers all rounding to 0.4. This similarity also suggests an outside factor influencing the models. Although the summaries indicate that all models perform well, the parallel measurements suggest a problem with all conducted analyses.

| **Models** | **RMSE** |
| --- | --- |
| Logistic Regression | 0.3968 |
| Naive Bayes | 0.4021 |
| Bagging | 0.3956 |
| Random Forest | 0.3909 |

##### Table 3. Root Mean Squared Error by Model

 Examining a more detailed classification summary demonstrates the issue with predicting membership to the Excellent bin. The F1 score is determined by taking the mean of precision and recall for each class. Precision is the percentage of correct predictions, and recall is the percent of the class identified by the model. These figures show that the models are only being rated as accurate due to their ability to predict non-cases. Each predictive technique employed shows a high F1 score for negative identifications but a noticeably low score for positive identifications. The imbalance suggests that it is easier for any model to predict that a film will not fall into the Excellent category. This pattern continues to demonstrate that there is a deficiency with the data.

| **Model** | **Excellent** | **F1 Score** |
| --- | --- | --- |
| Logistic Regression | No | 0.91 |
|  | Yes | 0.13 |
| Naive Bayes | No | 0.91 |
|  | Yes | 0.30 |
| Bagging | No | 0.91 |
|  | Yes | 0.34 |
| Random Forest | No | 0.91 |
|  | Yes | 0.25 |

##### Table 4. F1 Score by Model

 Basic exploratory analysis reveals that the problem is created through a deficiency in the dataset being analyzed. The predictions are biased due to the limited number of Excellent movies in the dataset, as the majority of entries fall into the defined Average bin. This explains the success in predicting negative cases, as a movie has a substantial random chance of being non-Excellent. Positive cases are harder to predict due to their diminished representation in the dataset.

 This imbalance in the data is the reason for the higher accuracy metrics, despite the models being unable to predict positive cases. In predicting a hit movie, the models are not considerably more reliable than a guess. The distribution of data shows the infectivity of the algorithms in determining movie success.



###### Figure 2. Movies per Ratings Bin

 Another issue with the analytical models is the innefectivity of the variables. With the immense number of crew members represented in the dataset, accurate representation of all their effects on film quality is impossible. Each actor and director will only have a small number of films represented in the entire data, as their work cannot include an amount of films substantial enough to establish an analytical basis for their contribution. This represented number decreases with sampling, as an individual actor or director would have even fewer films within the subset. The coefficient for an actor could potentially be created on the basis of 1 or 2 films. Those not represented in the sample will be left with even more inaccurate and meaningless measurements.

 This variable deficiency can be seen in the logistic regression summary. The coefficient for actor Leonardo DiCaprio is the third highest predictor (β: 0.2068) for an Excellent film within the model created. Upon examination, only 17 of his films are featured within the dataset of 4,777 movies. He is placed above the Drama genre (β: 0.1792) despite 2,575 drama films appearing on the list. The same imbalance of appearances in the data is demonstrated by the genre categories. The next most frequent genre is Comedy, with only 1,684 votes. These baseless results leave a statistically unsound model, as evidenced by the poor performance.



###### Figure 3. Movies per Genre

###  Summary and Conclusions

 The objective of the project was to predict the rating of a movie based on key factors such as genre, run time, director, and top billed actor in each film. These unique factors differentiate our research from other projects utilizing IMDb movie ratings. Other reports include outside factors such as studio decisions or condition of a film’s release, but our models deal only with the content inherent within the film. After preprocessing the data, four models were used to predict a movie’s rating: logistic regression, naive Bayes, random forest, and bagging. After separating the films based on rating bins created using ranges of values 1-10, we focused specifically on the Excellent bin. This allowed us to focus on predicting film success. Upon analysis, each unique modeling technique proved to have a similar accuracy. The accuracy measure in the percentage of correct predictions as well as by the root mean squared error indicate sufficient performance. On another note,

 Our large number of variables worked against our model. Each individual predictor, whether genre, actor, or director was effectively lost in the numerous other predictor variables. Measuring the coefficients in each model proved challenging, in part because of this large number of predictors. The presence of an actor and director in each model, with many films introducing new variables, made for a significant amount of coefficients. As a result, the models suffered from inaccuracy. This was especially true of positive classification since there were so few that the models would anticipate negative classifications. Some individual predictors were attached to numerous films, but the vast majority of predictors lacked a sufficient basis. The variables were unbalanced, with the number of films contributing to the measure of each factor differing substantially. Essentially, our predictor variables were spread too thin to create a highly effective model.

 Our results show that films do not have a basis in previous performance. Public opinion is too subjective, and there are too many diverse factors to accurately predict movie ratings based on a chosen few. Additionally, the processed data was unable to reflect the true measure of popular opinion, as it did not account for the number of votes while predicting film success. The cutoff for the number of votes was set at 20,000 total ratings. Due to this metric, the models treated films with 20,000 votes equally as those with +100,000 user ratings. While this was designed to ensure that a few major blockbusters did not skew the data, it does decrease the effectiveness of the models when implemented in real world circumstances. For future research, a weighted rating system would provide a more usable and accurate predictive analysis.

 The model’s utility is limited to public opinion of movies. However, the concept of the model can be applied to any instance where votes measure public opinion to predict ratings. For example, this could have applications in the music, arts, and sports industries, as long as data from each individual area of application is used. Moving forward, public opinion biases in each of these categories are strong candidates for future research as long as the number of predictors is limited in relation to the number selected in this project. Even within the movie industry, using different and fewer variables from the original datasets in this project could create unique and accurate models to predict public opinion of movies.

###

### References

[1] LaValley, Michael P. "Logistic regression." *Circulation* 117.18 (2008): 2395-2399.

[2] Bristi, Warda Ruheen, Zakia Zaman, and Nishat Sultana. "Predicting imdb rating of movies by machine learning techniques." *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019.

 [3] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization*. Vol. 752. No. 1. 1998.

[4] Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32.

[5] Büchlmann, Peter, and Bin Yu. “Analyzing Bagging.” *The Annals of Statistics*, vol. 30, no. 4, 2002, pp. 927–61. *JSTOR*, http://www.jstor.org/stable/1558692. Accessed 29 Apr. 2023.

###

### Appendices

#### Tables

[Table 1. Datasets Utilized 4](#_27s0st3ppmr1)

[Table 2. Class Consideration 5](#_f1tfzoleeqj1)

[Table 3. Root Mean Squared Error by Model 7](#_j5v78mdlmy6q)

[Table 4. F1 Score by Model 8](#_xp0b915mkwv8)

#### Figures

[Figure 1. Classification Accuracy by Model 7](#_tzxvfb317w5h)

[Figure 2. Movies per Ratings Bin 9](#_4k9vmuk74wem)

[Figure 3. Movies per Genre 10](#_l4g7rxapfuwp)