Predicting A Movie's Success

DNSC 6211: Programming for Analytics

Zheng Li
Rithvik Mundra
Nikitha Reghu
Zhian Wang
Qinhui Xu

Abstract

The most relatable form of entertainment for people all over the world is movies, and we wanted to identify and test the factors that can determine a movie's success. We chose the likelihood of a movie being nominated for Best Picture at the Academy Awards as our indicator of success. To determine the likelihood of a nomination, we used critic ratings (from Rotten Tomatoes), public ratings (from IMDB), popular sentiment (from Twitter and Facebook), box office revenue, movie budget and movie length as factors that can predict a nomination. We scraped Rotten Tomatoes and IMDB for highly rated movies from 2012-present and used these movies to scrape Twitter for sentiment scores. We used an IMDB movie dataset on Kaggle (1) to acquire budget, revenue, movie length and Facebook likes for most of these movies. Lastly, we ran logistic regression models on this data. Movies selected from 2012-2014 were used for training, 2015 movies were used for validation, and 2016 movies were used to test our model. After running our model, average ratings, Facebook likes, and budget were the most important factors in predicting a movie's likelihood of being nominated. We also predicted 8 movies that could be nominated for Best Picture in 2017.

Contents

1 Introduction

Since movies are a common form of entertainment across all regions and cultures in the world, we wanted to identify which factors can best predict a movie's success. According to PriceWaterhouseCoopers (who handles the Oscars voting process), the global box office revenue will grow from $36.7 billion in 2014 to $48.45 billion in 2019 (2). The astronomical figures of both the current and projected revenue indicate this industry will always be lucrative. This is why we were motivated to identify the factors that contribute to a movie's popularity and profitability. We chose to use the likelihood of a movie being nominated for Best Picture at the Academy Awards as the most recognizable form of success. Factors such as a movie's length (minimum of 40 minutes) and theatrical run time (playing in theatres for at least 2 weeks) are qualifiers to be eligible for a nomination. Therefore, we chose to incorporate movie duration, along with popular sentiment, critic ratings, public ratings, box-office revenue and budget data to determine a movie's chances of being one of up to ten movies nominated for Best Picture.

2 Background

Initially, we wanted to answer which movie will win the Best Picture award at the Oscars but we broadened this to which movies will be nominated for Best Picture. We wanted to predict these nominations by looking at Rotten Tomatoes (RT) or IMDB ratings, and found a Kaggle dataset that contained 5000 movies and their ratings from IMDB. Since we planned on using logistic regression for predicting the likelihood of a nomination, we needed more independent variables to build an effective model. Therefore, we expanded our list of prediction factors to include ratings from both IMDB and RT, sentiment from Twitter and Facebook, revenue, budget, and movie-duration. Using actors, directors, and genre were other variables we considered using but this information was hard to acquire and use in our model. We chose RT ratings as our main predictor since critic ratings might be most influential in predicting a nomination. To build our final datasets, we scraped RT and IMDB for highly rated movies each year from 2012-2016. There were missing fields in the Kaggle dataset, but we searched for and completed this information for the selected movies. For our final datasets, we averaged the RT and IMDB ratings, calculated a Twitter sentiment score, and inputted the number of Facebook likes, the gross revenue, the budget, and movie-duration for each movie.

3 Method

The overall question that we answered was if we can predict a movie's likelihood of being nominated for Best Picture at the Oscars. We conducted this analysis by building our own datasets and using logistic regression. For our logistic regression model, we used 2012-2014 movies as the training dataset, 2015 movies as validation data and 2016 movies as test data. To build our datasets, we first scraped Rotten Tomatoes to select the top 100 highly rated movies for each year. Next, we scraped IMDB for the top 1000 highly rated movies per year. We joined these datasets together to select movies that were highly-rated by both critics and the public. This join resulted in roughly 40 movies selected per year. Next, we averaged the Rotten Tomatoes and IMDB ratings so ratings represented both the public and critic opinion. A Twitter sentiment score analysis was conducted on the selected movies. Then, values for budget, revenue, duration and the number of Facebook likes were acquired from the Kaggle dataset for the selected movies. There were missing fields in this dataset but we searched the Internet to input these values for our selected movies. Values for budget were unavailable for six movies in our test dataset, but we used 60% of the gross revenue as the budget value for these

movies (to keep budget as an independent variable in our models). Lastly, we ran the logistic regression model on the training, validation and test datasets and used Shiny to visualize our results.

4 Organization

All group members were actively involved in discussing and formulating the main questions and methods needed to acquire the data and to analyze, visualize and present it. Qinhui and Anita took ownership of running the Twitter sentiment analysis and visualizing our outputs through Shiny. Scraping the various websites was divided amongst all group members. Rithvik lead the process for using SQL to join the Rotten Tomatoes, IMDB, Twitter and Kaggle datasets, and in building the logistic regression model. Leo and Nikitha assisted Rithvik with these steps. Nikitha wrote the reports for the initial and final presentations.

4.1 Workflow

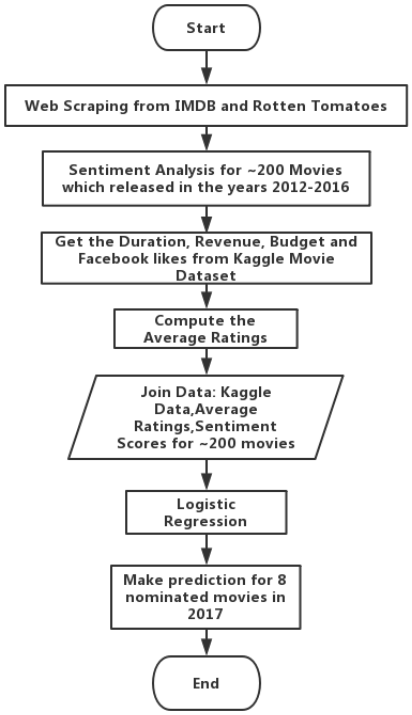


Figure 1: The project workflow

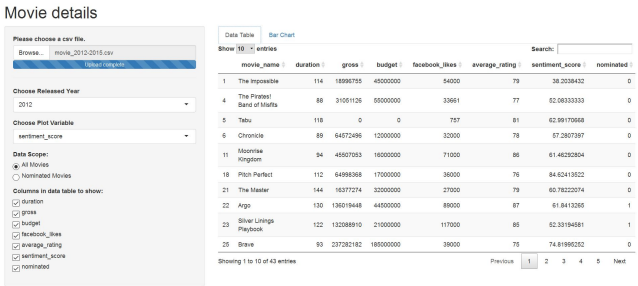
We first started by web scraping the Rotten Tomatoes websites for the top 100 movies for years 2012-2016. We then scraped the IMDB website for their top 1000 highly ranked movies for those years as well. Next, we selected movies from the IMDB dataset that were also present in the Rotten Tomatoes dataset. We then averaged the IMDB and Rotten Tomatoes ratings for each selected movie to have a combined rating score. We used the list of selected movies to acquire the revenue, budget, number of Facebook likes, and movie-duration data from the Kaggle IMDB dataset. We also ran a Twitter sentiment score analysis on these selected movies. Lastly, we ran our logistic regression models on the datasets from 2012-2016 (datasets include the six independent variables used to model our data: average rating, revenue, budget, Twitter sentiment score, movie-length and number of Facebook likes). We used the datasets for movies from 2012-2014 as our training data, the 2015 dataset as our validation data, and the 2016 dataset as our test data.

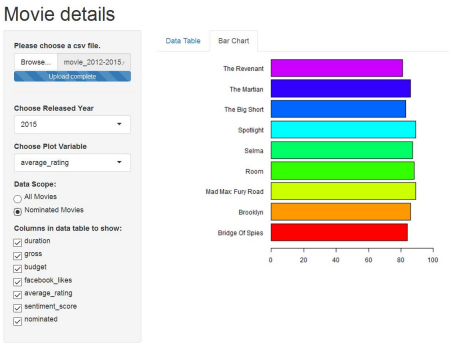
4.2 Project structure

Our first data source is the Kaggle IMDB dataset that contains a variety of information on 5000 highly-rated movies. The next primary data source are the 100 movies scraped from Rotten Tomatoes per year from 2012-2016. We also scraped IMDB for the top 1000 highly ranked movies per year from 2012-2016. Lastly, we ran a sentiment score analysis on Twitter for the roughly 200 selected movies. We also had to search the Internet for missing data in the Kaggle dataset for our selected movies. We scraped Rotten Tomatoes first because this dataset was the foundation of our research project. We assumed ratings to be a strong indicator of likelihood for being nominated for Best Picture at the Oscars, and decided to build the base of our datasets with movies that were highly ranked by critics. We also scraped IMDB for highly rated movies so our final dataset reflected movies that were highly-ranked by both the critics and the public. Since we chose to answer our research question with logistic regression analysis, we wanted to include more independent variables to improve the predictability of our model. Therefore, we used Twitter sentiment score analysis to have another variable in our analysis. The Kaggle dataset provided us with additional independent variables for our logistic regression model - such as revenue, budget, movie-length, and the number of Facebook likes.

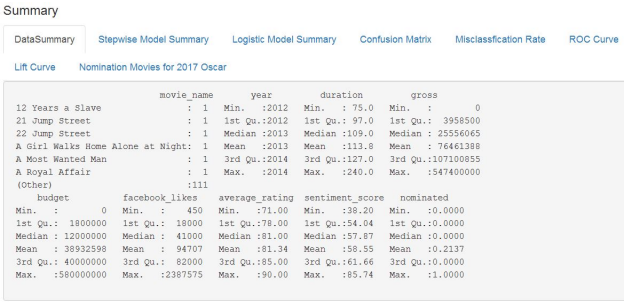
4.3 Figures and Tables

- Exploratory Data Analysis (First Shiny App)
  - This application shows all the movies selected for analysis for each year in the first tab. The second tab shows the movies in each year from 2012-2015. The second drop-down option in this tab enables us to visualize trends in duration, revenue, budget, Facebook likes, average ratings, and sentiment score.





- Data Summary Tab
  - Overview of each of the independent variables used in the model.

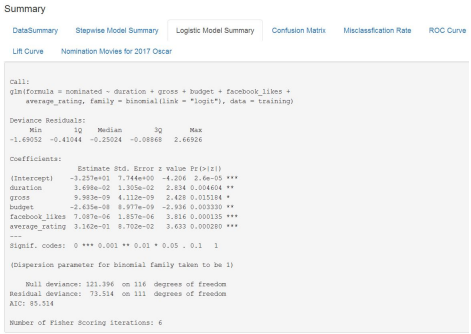



```
                          movie_name       year         duration         gross
12 Years a Slave                :  1   Min.   :2012   Min.   : 75.0   Min.   :        0
21 Jump Street                  :  1   1st Qu.:2012   1st Qu.: 97.0   1st Qu.:  3958500
22 Jump Street                  :  1   Median :2013   Median :109.0   Median : 25556065
A Girl Walks Home Alone at Night:  1   Mean   :2013   Mean   :113.8   Mean   : 76461388
A Most Wanted Man               :  1   3rd Qu.:2014   3rd Qu.:127.0   3rd Qu.:107100855
A Royal Affair                  :  1   Max.   :2014   Max.   :240.0   Max.   :547400000
(Other)                         :111
     budget          facebook_likes    average_rating  sentiment_score    nominated
Min.   :        0   Min.   :    450   Min.   :71.00   Min.   :38.20   Min.   :0.0000
1st Qu.:  1800000   1st Qu.:  18000   1st Qu.:78.00   1st Qu.:54.04   1st Qu.:0.0000
Median : 12000000   Median :  41000   Median :81.00   Median :57.87   Median :0.0000
Mean   : 38932598   Mean   :  94707   Mean   :81.34   Mean   :58.55   Mean   :0.2137
3rd Qu.: 40000000   3rd Qu.:  82000   3rd Qu.:85.00   3rd Qu.:61.66   3rd Qu.:0.0000
Max.   :580000000   Max.   :2387575   Max.   :90.00   Max.   :85.74   Max.   :1.0000
```

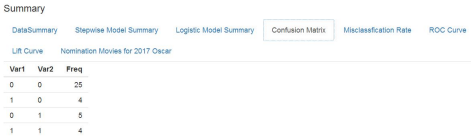
- Stepwise Model Summary Tab
  - Outputs the best variables to use in the final logistic regression model.

```
Start:  AIC=86.58
nominated ~ duration + gross + budget + facebook_likes + average_rating +
    sentiment_score

                  Df Deviance     AIC
- sentiment_score  1   12.578  85.092
<none>                 12.523  86.580
- gross            1   13.139  90.199
- duration         1   13.524  93.583
- budget           1   13.909  96.859
- average_rating   1   14.388 100.824
- facebook_likes   1   15.604 110.320

Step:  AIC=85.09
nominated ~ duration + gross + budget + facebook_likes + average_rating

                  Df Deviance     AIC
<none>                 12.578  85.092
+ sentiment_score  1   12.523  86.580
- gross            1   13.200  88.738
- duration         1   13.653  92.687
- budget           1   13.947  95.179
- average_rating   1   14.396  98.894
- facebook_likes   1   15.606 108.331
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
nominated ~ duration + gross + budget + facebook_likes + average_rating +
    sentiment_score

Final Model:
nominated ~ duration + gross + budget + facebook_likes + average_rating


             Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                    110   12.52274 86.58014
2 - sentiment_score  1 0.05486588       111   12.57760 85.09163
```

- Logistic Model Summary Tab
  - Output helps visualize the significance level (number of asterisks) of each of the independent variables used in the model.

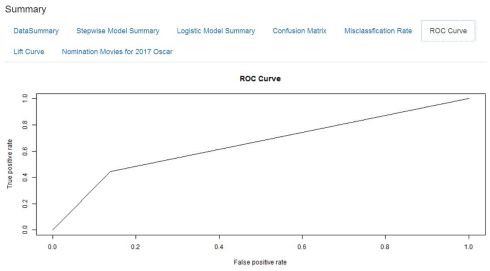


- Confusion Matrix Tab
  - Depicts how well the model correctly predicts the actual nominations for the validation dataset.

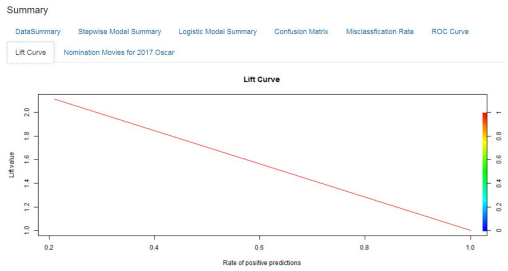


- Misclassification Rate Tab
  - Quantifies the accuracy of the model.

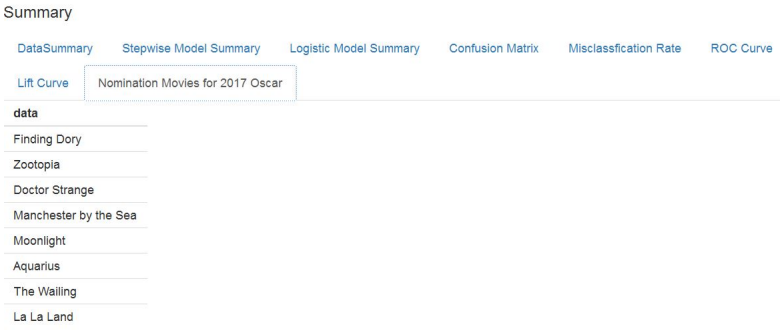
- ROC Curve Tab
  - Visualizes all possible classification thresholds for the model.



- Lift Curve Tab
  - Measure of the performance of the logistic regression model: our model is 2x better at predicting the likelihood of a nomination for a movie in the top 20% of the probabilities than selecting a movie at random.



- Nomination Movies for 2016 Tab
  - Outputs model's test-data prediction of 8 movies that could be nominated for Best Picture at the Oscars.



5 Discussion

Being able to predict which movies are going to be nominated for Best Picture at the Oscars was the primary selling point of our project. If a movie can be successful so visibly, everyone involved in the making of the movie achieves success by improving their prospects for future movies. Determining which factors contribute to a movie being nominated was the other selling point. Through our analysis, we learned that average ratings, number of Facebook likes, budget, duration and gross revenue (in that order) were the most important in predicting a movie's likelihood of being nominated. Sentiment score was not significant in predicting this outcome. As predicted, ratings play an important role in determining a movie's success. Therefore, movies should continue to please and entertain both the public and the movie critics.

5.1 Learnings

One of the best moments in this project was seeing our logistic regression model work. Since we have not yet learned how to work with advanced prediction models in our program, we were

skeptical that we could predict movies that are nominated for Best Picture at the upcoming Academy Awards. When we saw our validation dataset correctly predict four out of the eight movies nominated in 2015, we found cause for celebration and had more confidence when running our model on the test data. By working on this project, we learned how to conduct logistic regression analysis in R, how to use LaTeX to write a formal report, and how to use new applications in Shiny to visualize our data. We also solidified our skills in scraping data from various websites and in using sentiment score analysis on Twitter.

5.2 Challenges

We encountered difficulty when we realized the Kaggle dataset has missing information for the movies we selected for our models. We had to put in considerable time and effort to input values manually, to clean up and to build an accurate dataset. We also had to come up with an algorithm for the empty values for budget in our training dataset. Another difficult moment was when we tried to use additional categorical variables in our logistic regression models. We wanted to use the actors, directors and genres in our analysis, but this was too difficult. There were too many options and combinations of each variable to easily incorporate in our model. Running a sentiment score analysis on 200 movies also proved to be time consuming. Lastly, developing new Shiny applications required lots of online searches to have them functioning properly.

6 Conclusion

We wanted to answer if we can predict a movie's likelihood of being nominated for Best Picture at the Oscars. We conducted a logistic regression analysis to predict the likelihood of a movie being nominated. The factors we used in our model were critic and public ratings (from Rotten Tomatoes and IMDB), public sentiment (from Twitter and Facebook), gross revenue, movie-budget, and movie-length. After running our model through the training, validation and test datasets, we saw that average ratings, the number of Facebook likes and budget were the most significant factors in determining the likelihood of a nomination. Our validation model, after calculating the misclassification rate (23%), is about 77% accurate. Therefore, we proceeded with our test model and have predicted 8 movies that could be nominated for Best Picture in 2017:

- Finding Dory
- Zootopia
- Doctor Strange
- Manchester by the Sea
- Moonlight
- Aquarius
- The Wailing
- La La Land

7 References

1. (2016, August). Retrieved from https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset
2. (2015). Retrieved from https://www.pwc.com/gx/en/global-entertainment-media-outlook/assets/2015/filmed-entertainment-key-insights-2-global-box-office.pdf
3. Link to our Project on Github: http://github.com/rithvikmundra/DNSC-6211-Final-Project
4. Link to our Youtube video: https://www.youtube.com/watch?v=PLiws94MQ_U