“Let’s All Go to the Movies!”

Erika Ibarra, Bryce Lunceford, Daniel Smith, Paul Smith, Luke Wright

December 2021

1 Research Question and Data Overview

Unless you’re from the deepest crevice of the jungle or the most desolate corner of the desert, you have probably experienced the excitement that comes from going to the movies. Sitting in a dark hall, crammed together with others, you soak up the energy of anticipation until the lights dim, the music starts, and the opening credits herald in the start of the story you have been so eager to see. Whether an audience member or a critic walks away from this experience satisfied might seem entirely subjective; however, we hypothesize that some features of a film will make an audience quantifiably more satisfied. Our goal is to determine which features a filmmaker should focus on in order to maximize critic reception and audience reception.

The model and analysis we set out to produce would potentially maximize profits for the entertainment industry and ensure positive reception from both critics and audiences alike. It could change the way the film industry is run and alter the choices that surround how features and aspects of films are combined.

We used the dataset `IMDB movies` taken from IMDB.com which contains information for 85,856 films [3]. There were many features in the raw dataset, so we limited our view to the following independent variables: `movie genre`, `duration` (the length of the film,) `director`, `writer`, `production company`, `actors`, `average vote`, `number of votes`, and `budget`. We used `metascore` and `worldwide gross income` as our dependent variables to quantify critical reception and popular reception respectively. The next step was to clean and engineer the data for analysis.

2 Data Cleaning and Feature Engineering

To clean our data we adjusted the `budgets` and `worldwide gross income` to US dollars accounting for inflation. We also one-hot encoded the categorical features.

2.1 Adjusting Budget to U.S. Dollars

Our dataset contained a `budget` feature that described how much money was spent on each film's production. However, the amount specified was in the currency of the country of origin and was not adjusted for inflation. Therefore, for consistency, it was necessary to convert the amount used from the original currency to U.S. dollars.

Our research brought us upon the `OECD` data set which contains the historical U.S. dollar exchange rates for a large number of currencies dating back to 1950. Within the data set each currency's country of origin is identified [1]. We built a Python dictionary that matched each currency type to a country code in the `OECD` dataset and then joined the `IMDB` dataset to the `OECD` dataset by the currency's country code and year. We assigned a `NaN` value to all currencies which were not documented.

After joining the two datasets together, the historical U.S. dollar equivalent was computed by dividing the budget amount by the respective exchange rate. All films with a `NaN` budget or exchange rate value were dropped as well as those that were made before 1950.

After converting the historical budget amounts to the equivalent amount in U.S. dollars, we applied an adjustment for inflation to regularize the budgets to their current-day equivalents. We describe the methods we used in the following section.

2.2 Inflation

A common measure of inflation over time is the Consumer Price Index. The Consumer Price Index is defined to be

$$CPI_t = \frac{C_t}{C_0} \cdot 100 \tag{1}$$

where C_t is the cost of a "basket" of goods in year t and C_0 is the cost of a "basket" of the same goods in some base year. The ratio of the cost of an item in one year to the cost of the item in another year can then be calculated using the ratio of the consumer price indices.

$$\frac{CPI_1}{CPI_2} = \frac{C_1 \cdot 100}{C_0} \cdot \frac{C_0}{C_2 \cdot 100} = \frac{C_1}{C_2}$$

Thus, a formula for the value of an amount after inflation is given by

$$C_2 = C_1 \frac{CPI_2}{CPI_1} \tag{2}$$

To perform this computation we used the `Bureau of Labor Statistics' Consumer Price Index` dataset[2], loading it and then inner-joining it with the IMDB dataset by `year`. We then selected the `Annual` column from the CPI dataset that gives the average CPI for the year that the film was released. We chose to regularize all currency amounts to their value in 2020 to give the amounts a sense of relatability. We took the CPI for 2020 from the dataset and used equation (2) to calculate the gross income and budget for the films adjusted to the equivalent value. With this done, we moved on to deal with our categorical variables.

2.3 One-hot Encoding Categorical Variables

We began by dropping every film with data missing for any independent variable. We then took the non-numerical features `genre, director, writer, production company`, and `actors` and one-hot encoded them. However, this resulted in an unruly number of uniquely valued columns. We ended up choosing to remove those that we determined would not affect our results.

To retain only statistically significant, categorical variables, we only considered `directors` and `writers` who had worked on at least three films, production companies that had produced at least ten films, and actors who had been in at least five films. This dramatically reduced the column size of our dataset from 113,562 to 15,585.

3 Data Visualization and Basic Analysis

Preliminary analysis showed that after cleaning the data, some movies still had missing metascores. As an amendment to cleaning the data, we calculated metascores for films that didn't have one.

3.1 Data Augmentation: Predicting MetaScore

To predict the missing metascore values, we used a Linear Regression model (OLS) with vote and average vote as the independent variables. We trained the function with the data we had and tested the function back on that same data to compute a set of residuals, the residuals being the difference between the predicted metascore values and their metascore counterparts. In order to have confidence in the OLS results, the residuals needed to have mean zero, constant variance, and normality in their distribution.

3.1.1 Mean Zero, Constant Variance, and Normality

The calculated residual mean comes out to be $\hat{\mu} = 3.34 \cdot 10^{-13} \approx 0$. To determine if the residuals have constant variance, we standardized and plotted them against the index as follows: $\hat{\epsilon}_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$. There were no outstanding patterns, so we concluded that the residuals for metascore have constant variance.

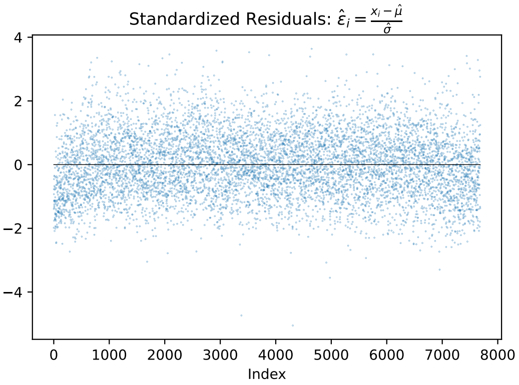
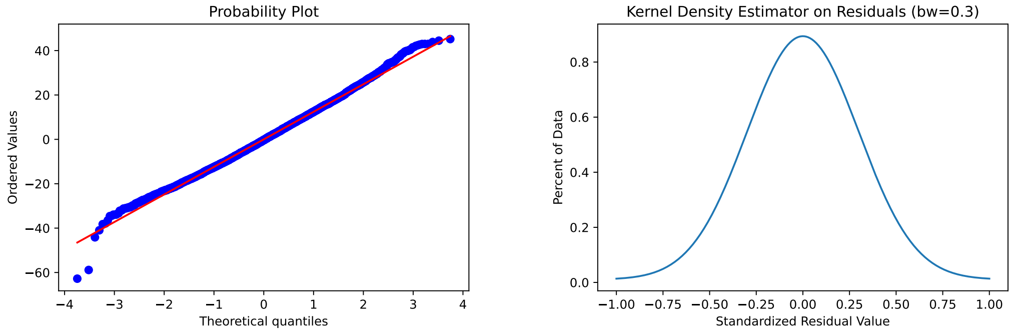


Figure 1: Standardized Residuals

To verify normality of the residuals, we used a Quantile-Quantile Plot (QQ Plot) and a Kernel Density Estimator (KDE). The linearity of the QQ Plot confirms normality as does the KDE. The calculated residuals yield the following charts:

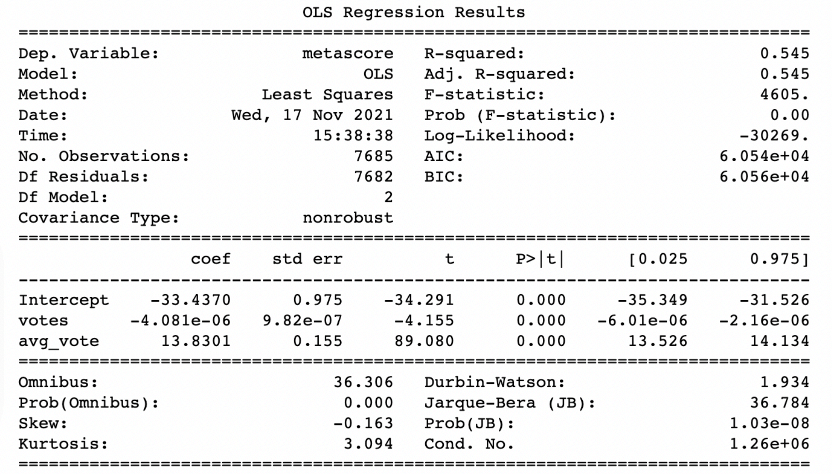


(a) QQ Plot

(b) KDE

3.1.2 The Metascore OLS Model

Since the residuals satisfy the necessary assumptions, we can analyze the OLS model output. This is the summary of the OLS model for predicting `metascore`.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              metascore   R-squared:                       0.545
Model:                            OLS   Adj. R-squared:                  0.545
Method:                 Least Squares   F-statistic:                     4605.
Date:                Wed, 17 Nov 2021   Prob (F-statistic):               0.00
Time:                        15:38:38   Log-Likelihood:                -30269.
No. Observations:                7685   AIC:                         6.054e+04
Df Residuals:                    7682   BIC:                         6.056e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -33.4370      0.975    -34.291      0.000     -35.349     -31.526
votes      -4.081e-06   9.82e-07     -4.155      0.000   -6.01e-06   -2.16e-06
avg_vote      13.8301      0.155     89.080      0.000      13.526      14.134
==============================================================================
Omnibus:                       36.306   Durbin-Watson:                   1.934
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.784
Skew:                          -0.163   Prob(JB):                     1.03e-08
Kurtosis:                       3.094   Cond. No.                     1.26e+06
==============================================================================
```

Figure 2: OLS Regression Results

The coefficients have excellent p-values, so `votes` and `average vote` are significant features. We used this model to predict the missing `metascore` values. Our next objective was to build complete datasets for analysis.

3.2 Constructing Datasets for Analysis

We created three different datasets for our in-depth analysis. Our first dataset only included films with a recorded `worldwide gross income`. Our second dataset only included films with a recorded `original metascore`. For the third dataset, we used the `engineered metascores`, which retained all 22k films instead of dropping down to the 7k films that only had their `original metascores`.

4 Learning Algorithms and In-depth Analysis

With our clean data, we attempted to find what factors would make the best possible movie as determined by the two metrics `worldwide gross income` and `metascore`.

4.1 Linear Regression

We ran a scikit-learn Linear Regression against the target values (`worldwide gross income`, `original metascore`, and `engineered metascore`), and with a 70-30 train test split, we received an R^2 score of 0.475 for `worldwide gross income`, 0.079 for the `original metascore`, and 0.00006 for the `engineered metascore`. While Linear Regression seemed like a good first attempt, these results made it clear that we needed to try other methods to more accurately fit our model.

4.2 Random Forests

The second model that we tried was a Random Forest Regressor. Using scikit-learn's `GridSearchCV` it was determined that the hyperparameters that gave the best results were `n_estimators=200` and `min_sample_split=200`. Using these hyperparameters, we were able to predict `worldwide gross income` somewhat better, increasing to an OOB score of 0.499. We were able to predict the `metascores` significantly better than before, but the accuracy was still only marginal. We predicted the `original metascores` with a score of 0.226, and could predict the `engineered metascores` with a score of 0.323.

4.3 Boosted Trees

In an attempt to increase our regression score, we tried a scikit-learn Boosted Tree. We found that 50 `n_estimators` with a `squared error loss`, `learning_rate` of 0.3, `min_samples_split` of 20, `max_depth` of 7, and `min_samples_leaf` of 3 gave the best results. This Boosted Tree gave us a slightly improved OOB score of 0.545 for `worldwide gross income`, 0.249 for the `original metascores`, and 0.330 for the `engineered metascores`. Surprisingly, applying the Boosted Tree method consistently outperformed applying the Random forest method. Therefore, the Boosted Tree method was used to fit our model for all project results.

5 Results of the Model

After running the regressions, we identified the highest ranked `director`, `writer`, and `production company`, the top two `genres`, and the top seven `actors` by finding those variables with the highest feature importance scores. To find an appropriate value for the numerical values of `budget` and `duration`, we calculated the mean values of the `budgets` and `durations` of the films that the top ranked `directors`, `writers`, `production companies`, and `actors` each worked, which were also of the top ranked genres.

	Gross Income	Original Metascore	Engineered Metascore
Top Genres	Adventure, Drama	Drama, Animation	Drama, Horror
Top Director	Steven Spielberg	Alfred Hitchcock	Sergey A.
Top Writers	George Lucas, Andrew Stanton	Chuck Konzelman, Woody Allen	John Waters, Jean-Claude La Marre
Top Studio	Lucasfilm	Pixar Animation Studios	The Asylum
Top Actors	Ben Wright, Anthony Daniels, Josh Gad, Julie Andrews, Jeff Goldblum, Mark Hammill, Murray Hamilton	Cuba Gooding Jr., George Clooney, Adam Sandler, Joaquim de Almeida, Jennifer O'Neill, Gerard Butler, Jessica Biel	Michael Madsen, Eric Roberts, Ajay Devgn, Michael Paré, Simon Phillips, Christopher Guest, Kane Hodder
Duration (mins)	117.9	109.2	94.4
Budget	$ 90,254,770	$ 60,906,835	$ 10,112,244
Predicted Income	$2,545,820,423	-	-
Predicted Metascore	-	45.8	13.7

Table 1: Calculated Film Characteristics and Predictions

Once we identified these specific feature values, we created a “film” by taking a row from our dataset, zeroing out the features, and then filling in the appropriate features with the data we found. We then ran this film back into our Boosted Tree to make the predictions. The results can be seen in Table 1.

6 Results and Conclusions

Running our model again and again, we noted an interesting phenomenon: Our code seemed to be very good at generating movies with features that would be met with positive popular reception, however, it was not able to do the same with critical reception. Instead, the model features that were “most important” for predicting `metascore` were mostly those that were likely to affect the score *negatively.* For instance when we reference the second and third columns of Table 1, we observe `metascore` values 45.8/100 and 13.7/100 respectively. This indicated that these films would be regarded as unpopular among critics.

This could indicate a couple different possibilities. The first possibility is that predicting `metascore` in general is not a feasible task, at least not with the data we have. The second possibility is that it may not be feasible to predict a high `metascore`, but it *may* be feasible to predict a low `metascore`. In plain English, we can not predict good movies, but we can predict bad ones. In either case, it seems that the quantifiable features of a film are less important to the critical reception of that film than potentially unquantifiable aspects like classic, artful storytelling. For instance, even the best of actors can show up in bad films; even the underdog studios can create a masterpiece; Critics don’t seem to be groupies to any one director, writer, studio, or combination of actors.

Unfortunately, it appears that the same cannot be said for general audiences. Our results indicate that it *is* possible to take a formulaic approach to creating a movie in order to achieve high box office revenue. This leads us neatly into our ethical concerns.

7 Ethics

The first ethical concern is that filmmakers could abuse the results we found and only create films with the limited set of specific features that lead to high revenue. Studios might limit who they hired to the set of most popular actors which would create negative feedback loops. And movies in general would lose their variety and (most likely) quality.

On top of that, we also noticed that according to our program the `actors`, `writers`, and `directors` who tended to lead to high box office sales were predominantly Caucasian males from the United States. Women and people of other races often failed to appear in the results. Modern films seem to push for greater inclusivity in their casts and crews and greater diversity in their plots and storytelling methods. If they took our approach, theoretically they would make a decent amount of money, but they'd be missing out on an enormous aggregated section of the creative market.

The results that our model returned may be a reflection of the fact that the film industry was created and primarily developed in America throughout the late 1900s. With that being said, this may an old historical trend of discrimination. If we had created our model over films produced in the last 20 years alone, we likely would have gotten different results. One way or another, filmmakers should probably not consider the monetary benefit of race at all as they produce movies. They should focus on creating compelling stories.

Too bad Hollywood isn't doing that either.

References

[1] *Exchange rates (indicator).* Organisation for Economic Co-operation and Development (OECD), 2021.

[2] United States Bureau of Labor Statistics. *All items in U.S. city average, all urban consumers, not seasonally adjusted. Series ID CUUR0000SA0.* United States Bureau of Labor Statistics, 2020.

[3] Ashirwad Sangwan. IMDB, 2019.