MGSC661 Midterm Project Report

Ge Gao, Hongyi Zhan, Yu Lu, Yanhuan Huang

November 1, 2023

1 Introduction

In today's data-driven world, the ability to make predictions based on past and present data is becoming increasingly essential. This is especially true in areas like the entertainment industry, where a film's success can have significant financial and reputational outcomes. Our group decided to take on "The 2023 IMDB Prediction Challenge" to try and predict the IMDB ratings of twelve upcoming blockbuster movies.

Using what we've learned in our courses, we're hoping to build models that can handle the many challenges that come with data analytics. This includes dealing with issues like biases, heteroskedasticity, collinearity, and outliers. One of our main concerns is avoiding overfitting, so we're aiming to create a model that's versatile and can work well with different kinds of data.

We're working with a dataset that has information on around 2,000 movies from IMDB. It's a pretty detailed dataset, with everything from basic movie details to more specific production information. We believe that by analyzing this data, we can get a good idea of what factors might influence a movie's ratings.

The objective of this project transcends mere academic inquiry. We genuinely believe that if we can accurately predict movie ratings, it could have real-world implications. This could influence how movies are marketed, how budgets are allocated, and even decisions about what movies get made. As we dive into this project, we're committed to combining the academic knowledge we've gained with real-world application, hoping to provide insights that are both academically sound and practically useful.

2 Data Description

2.1 Data Exploration

The dataset comprises a diverse array of variables from 1,930 predominately English-language movies, offering a comprehensive perspective on various film attributes and their potential influence on IMDB ratings. The dependent variable, IMDB score, is accompanied by 16 categorical variables and 25 numerical variables serving as independent variables. Among these, 13 binary variables represent different movie genres.

Our analysis began by examining the distribution of the dependent variable, IMDB_score, using both the `summary()` function and histograms (Figure 1). The results revealed that

the majority of movies in the dataset received scores between 6 and 7, with a minimum score of 1.9 and a maximum of 9.3. The distribution of IMDB scores is left-skewed, evident as the mean (6.512) is less than the median (6.6).

Similar analysis were conducted for independent variables, excluding dummy variables representing movie genres. In the case of *movie budgets*, several peaks were observed at 15 million, 20 million, 25 million, 35 million, 40 million, and 50 million. Despite these peaks, there was no apparent skewness within this variable, and no clear correlation with IMDB score was evident.

Regarding *movie duration*, the data exhibited a significant right skew with a magnitude of 2.68. Most movies had duration between 1.5 to 2 hours. Although there was one movie with a duration exceeding 300 minutes, the residual plot of the simple linear regression between duration and IMDB score did not indicate a severe outlier issue.

Analysis of *release year* and *release day* indicated that movies in the dataset were predominantly released within the past 40 years and were relatively evenly distributed across each day of the month, with minor peaks observed on the 1^{st} and the 25^{th}. Bar chart for *Release month* lacks a distinct peak; records are scattered randomly throughout this variable.

Additionally, *aspect ratios* tended to concentrate around values of 1.85 and 2.35, which are common formats for widescreen cinemas. A significant majority of movies (75%) featured only 2 *faces on their posters*. An outlier was noted in the movie "The Master," which displayed 31 faces on the poster; however, this outlier did not seem to pose a significant issue when fitting a simple linear regression of IMDB score. Moreover, it was observed that the outlier represented only three distinct faces recurring multiple times. In the case of *number of news articles*, *"Star Wars: Episode IV-A New Hope"* distinctly marked as an outlier with a value greater than 60,000 in the linear regression analysis with movie score, three times larger than the second-largest observation.

2.2 Data Preprocessing

In this project, a primary focus during the preprocessing phase was on the extraction and categorization of movie genres and plot keywords. While the original dataset provided separate columns for genres represented as binary values, it was noted that some genres, such as comedy, were not covered. To address this, the genres variable was split using the separator "|", and an enumeration process was applied to identify unique genres across the dataset. Similarly, plot keywords were disaggregated and indexed alongside their respective movies. This granular representation of plot keywords was then saved as a separate CSV file for further analysis. This meticulous preprocessing lays the groundwork for potential subsequent analyses, such as genre-based or keyword-based movie rating predictions.

Subsequent to the preprocessing of the IMDB dataset, the movie plot keywords underwent topic modeling using Latent Dirichlet Allocation (LDA). The preprocessed data, `pre_plot.csv`, was imported and transformed into a document-term matrix (DTM) suitable for LDA. An LDA model was trained on this matrix, targeting eight distinct topics. After modeling, the topic-word distributions were extracted, and for enhanced clarity, the top 30 keywords for each topic were isolated based on their significance. This approach provided a clear overview of the predominant thematic elements across the movie plots.

Then, the IMDB dataset underwent further preprocessing to refine and categorize its features. Initially, irrelevant columns such as movie titles, movie IDs, and IMDB links were excluded to streamline the dataset. The movie genres were then reclassified, and genres not included in the original dataset, such as 'Biography', 'Comedy', and 'Music', were isolated and encoded as binary variables. Each movie was mapped to its plot keywords, and these keywords were associated with one of the eight identified topics. This resulted in the creation of binary columns for each topic, indicating the presence or absence of a particular thematic element in a movie's plot. The refined dataset, encapsulating both genre and plot topic information, was saved for subsequent analyses.

We then proceed to address the four main model issues on training data. First we delved into the correlation matrix among individual numerical variables. The highest correlation (0.299) was observed between *actor2_star_meter* and *actor3_star_meter*, yet it did not reach a level that could be considered problematic. Hence, it was ascertained that apprehensions regarding collinearity among the selected variables were unsubstantiated. This conclusion was supported by the examination of variance inflation factors (Figure 2). We extended similar analysis to the binary columns representing plots and genres and investigation yielded no signs of multicollinearity within these categories.

Further categorizing variables into film, cast, and production characteristics, we conducted three distinct regression analyses to scrutinize heteroskedasticity, non-linearity and outlier issues within each category.

Regression involving cast-related variables revealed a ncvTest value exceeding 0.05, indicating the presence of heteroskedasticity. Tukey tests conducted on the three regressions identified non-linear patterns within film-related variables, specifically in attributes like duration, the number of news articles, and the 2023 movie ranking.

Performing a Bonferroni test on the regression between IMDB score and film characteristics, we identified six outliers. Additionally, Normal Q-Q plot revealed data point 492, deviating by 6 standard deviations below the mean (Figure 3). This particular data point corresponded to the movie *"Star Wars: Episode IV - A New Hope"*, a finding consistent with our earlier data exploration. We excluded this outlier and preserved the remaining records as an additional CSV file for the purpose of comparing model performance. Regressions on cast and production characteristics did not exhibit p-values exceeding 0.5 in the Bonferroni test, indicating the absence of outlier issues.

3 Model Selection

3.1 Initial Dataset Refinement

Our initial step in the model selection process involved refining the IMDB dataset to focus on variables with the most significant impact on film ratings. We removed attributes like *genres* and *plot_keywords*, which preliminary analysis suggested had minimal influence on IMDB scores. This decision was crucial for reducing model complexity and avoiding noise in the data.

3.2 Binary Transformation of Variables

We transformed *country* and *language* into binary variables, categorizing films as either American or non-American, and English-speaking or non-English-speaking. This binary coding was also applied to *release_month* to capture seasonal release patterns. These transformations aimed to simplify the dataset and emphasize the potential influence of geographical, linguistic, and seasonal factors.

3.3 Exclusion of Unbalanced Variables

The *maturity_rating* variable, initially with three categories, was excluded due to its unbalanced distribution. This decision was made to prevent potential bias in the model, as an unbalanced distribution of a variable could skew the results.

3.4 Reclassification of Director and Actor Variables

Director and *actor1* were reclassified based on their frequency in the dataset. This reclassification aimed to more effectively assess the impact of these personnel on film characteristics.

3.5 Development of Linear Regression Models

We developed two primary linear regression models: the BCS Model and the LASSO Model. Each model incorporated a different set of predictors, chosen based on their theoretical relevance and potential impact on IMDB scores. Variables like movie budget, release details, and film characteristics were included due to their perceived influence on audience ratings and film success.

3.6 Incorporation of Spline Transformations

To capture potential non-linear relationships, spline transformations were applied to variables *duration*, *movie_budget*, and *release_year*. The knots for these splines were strategically placed at the 25^{th}, 50^{th}, and 75^{th} percentiles, based on the distribution of these variables. This approach allowed us to model more complex relationships than would be possible with simple linear terms.

3.7 Balancing Model Complexity

The complexity of the model applied backward stepwise selection, focusing on improving the Akaike Information Criterion. This process involved removing the least significant variables one at a time, and refining the model to include only those predictors that contributed most to explaining the variability in IMDB scores.

3.8 Model Comparison Using ANOVA

ANOVA was used to compare different models, we compared models with and without the *maturity_rating* variable to understand its contribution to the model's explanatory power. This comparison was essential to ensure that our final model was both efficient and effective.

4 Results

4.1 Predictive Performance

R-squared: The model achieved an R-squared value of 0.4191, indicating that approximately 41.91% of the variability in IMDB scores is explained by the model. This level of explanatory power is moderate, suggesting that while the model captures a significant portion of the variance in IMDB scores, there are other unaccounted factors influencing these scores.

Adjusted R-squared: The adjusted R-squared value is slightly lower at 0.4105, accounting for the number of predictors in the model. This adjustment provides a more accurate measure of the model's explanatory power when considering the complexity of the model.

4.2 Predictive Accuracy

The final linear regression model, `model_final_lm`, demonstrates moderate predictive accuracy in estimating IMDB scores, as indicated by a Mean Absolute Error (MAE) of approximately 0.61 and a Mean Squared Error (MSE) of around 0.70. These values suggest that, on average, the model's predictions deviate from the actual scores by about 0.61 points, with some instances of larger errors highlighted by the MSE. This level of performance indicates reasonable effectiveness for practical applications, though it also points to potential areas for refinement in the model.

4.3 Significance of Predictors

- **Movie Budget:** The negative coefficient for movie budget suggests a counterintuitive relationship where higher budgets might not always correlate with higher IMDB scores.
- **Release Year:** The negative coefficient for release year indicates a potential decline in scores over time.
- **Duration:** The spline transformation for duration, with knots at specific percentiles, captures the non-linear relationship between movie duration and IMDB scores. Some spline components were significant, indicating varying effects of duration on scores at different levels.
- **Country:** The binary variable for country showed a significant effect, highlighting the influence of geographical factors.

- **Maturity Rating:** The maturity rating had mixed effects, with some categories showing significance.
- **Number of News Articles:** A positive coefficient for this variable underscores the role of media coverage in influencing IMDB scores.
- **Number of Faces in Poster:** The negative coefficient suggests that a higher number of faces on movie posters might correlate with lower scores.
- **Genres:** Different genres showed varying impacts, with drama, animation, and documentary genres positively influencing scores, while action and horror had negative effects.
- **IMDBPro Movie Meter:** The negative coefficient indicates a complex relationship between industry metrics and audience ratings.
- **Director and Lead Actor Categories:** Certain categories of directors and actors showed significant impacts on the scores, reflecting the influence of industry reputation and star power.

4.4 Predictions for 12 Movies

The model was used to predict IMDB scores for 12 movies from the test set. These predictions are a practical demonstration of the model's applicability and provide a snapshot of its predictive capabilities. The accuracy of these predictions, as reflected in the MAE and MSE, is crucial for evaluating the model's utility in real-world scenarios.

In conclusion, the `model_final_lm` presents a comprehensive analysis of factors influencing IMDB scores. While it demonstrates a moderate level of explanatory power, the model's predictive accuracy and the significance of each predictor offer valuable insights into the dynamics of film ratings. The model's performance on the test set, particularly the predictions for the 12 movies, further underscores its practical applicability in predicting IMDB scores.

5 Appendices

5.1 Table

Table 1: Regression Results for IMDb Scores

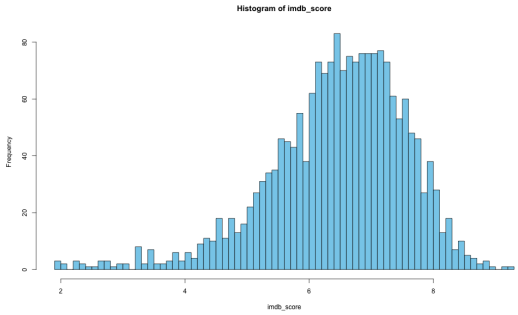
Variable	*Coefficient*
Release Year	-0.019^{***}
	(0.002)
Movie Budget	-0.000^{***}
	(0.000)
duration_bs_term1	-2.485^{**}
	(1.138)
duration_bs_term2	-2.603^{***}
	(0.804)
duration_bs_term3	-1.722^{**}
	(0.818)
duration_bs_term4	-0.583
	(0.847)
duration_bs_term5	-1.345
	(1.097)
duration_bs_term6	-0.774
	(1.117)
Country-USA	-0.230^{***}
	(0.050)
Maturity Rating-PG	0.160
	(0.112)
Maturity Rating-R	0.427^{***}
	(0.112)
Number of Articles in News	0.0002^{***}
	(0.00002)
Number of Faces on Poster	-0.031^{***}
	(0.010)
Genre-Action	-0.278^{***}
	(0.056)
Genre-Thriller	-0.082
	(0.052)
Genre-Horror	-0.348^{***}
	(0.071)
Genre-Drama	0.311^{***}
	(0.050)
Genre-Animation	1.047^{***}
	(0.204)
2023 Movie Rating	-0.00000^{***}
	(0.00000)
Genre-Biography	0.179^{**}
	(0.078)

Continued on next page

Table 1 continued from previous page

Variable	*Coefficient*
Genre-Comedy	-0.116^{**}
	(0.052)
Genre-Documentary	1.309^{***}
	(0.262)
Director-Medium	-0.070
	(0.109)
Director-MediumPlus	-0.099
	(0.127)
Director-Other	-0.295^{***}
	(0.107)
Actor1-Medium	-0.126
	(0.102)
Actor1-MediumPlus	-0.056
	(0.113)
Actor1-Other	-0.333^{***}
	(0.100)
Constant	46.127^{***}
	(4.032)
Observations	1,929
R^2	0.419
Adjusted R^2	0.411
Residual Std. Error	0.844 (df = 1900)
F Statistic	48.955^{***} (df = 28; 1900)
	Note: *p<0.1; **p<0.05; ***p<0.01

5.2 Figure



(a) Histogram of IMDb score

```
> summary(imdb$imdb_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.900   5.900   6.600   6.512   7.300   9.300
```

(b) Summary output

Figure 1: IMDb_score

```
> vif(reg4) # result shows that no collinearity issue with the selected features
       movie_budget         release_day        release_year            duration        aspect_ratio    nb_news_articles
           1.135798            1.003811            1.187913            1.145433            1.127742            1.027422
 actor1_star_meter   actor2_star_meter   actor3_star_meter           nb_faces movie_meter_IMDBpro
           1.037518            1.138971            1.100815            1.008444            1.023941
```

Figure 2: VIF Test for Numeric Variables

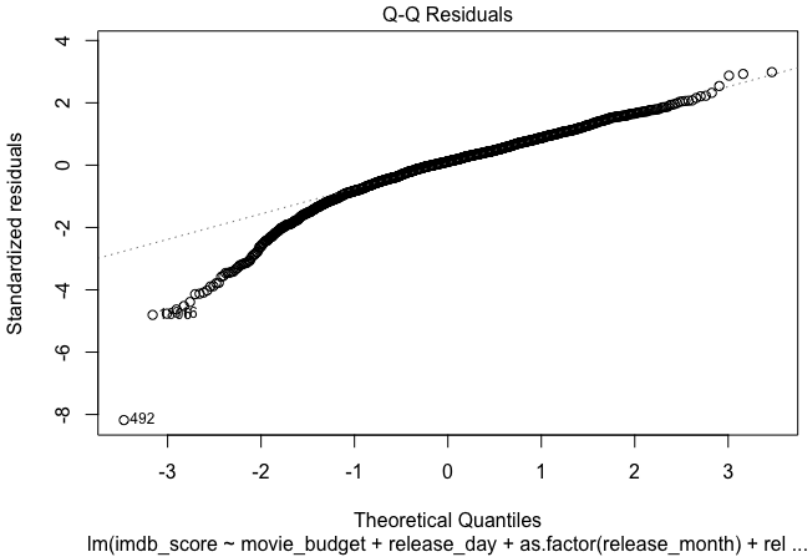


Figure 3: Normal Q-Q Plot for SLR on film characteristics