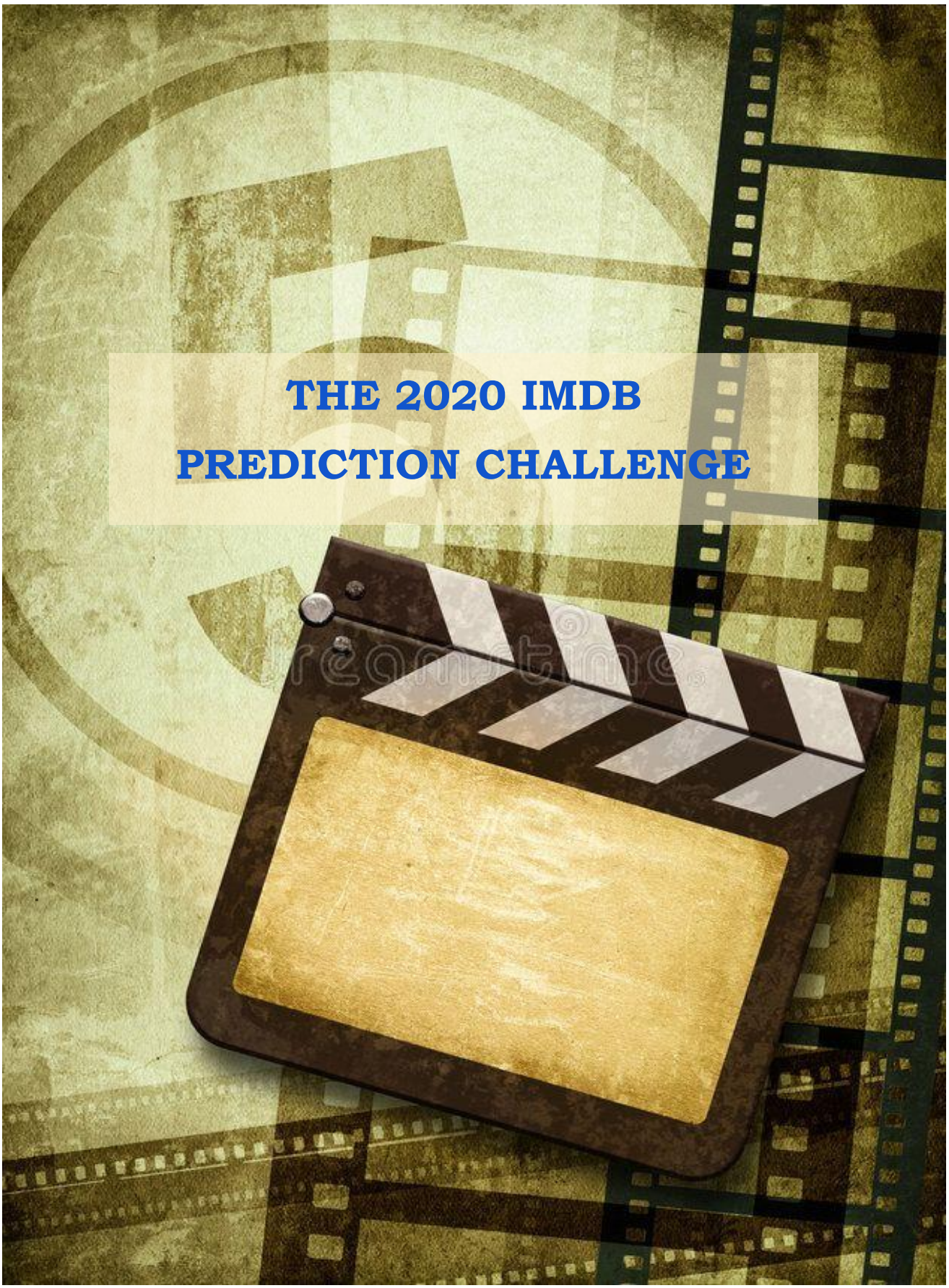
THE 2020 IMDB
PREDICTION CHALLENGE

TABLE OF CONTENTS

Introduction..... 2

Data Description..... 3

Model Selection..... 7

Managerial Implications..... 11

Appendix..... 14

INTRODUCTION

IMDb, or "Internet Movie Database", is a website which provides information on movies, TV programs, and digital entertainment content. Founded in 1990 as a platform of film-related content in list-format (i.e. casts, crew, and ratings), IMDb was purchased by Amazon in 1996 and quickly evolved into a crowdsourcing database which allows registered users to search, rate, and review films[1]. Today, IMDb is one of the most reputable sources for movies and film related information, receiving over 250 million unique visitors a month[2]. While IMDb's revenue model is largely advertisement-driven, it is undeniable that landing a spot on the coveted "250 Top Rated Movies" list can significantly represent a movie's popularity and may further drive its future reputation. Given that each rating is determined by up to thousands of verified, crowdsourced user reviews, there is a significant opportunity to explore what factors may influence a movie's average score. Knowing such factors could shed light on whether there are certain critical ingredients which influence IMDb user preferences and provide valuable information for entertainment executives to use when attempting to craft the next popular hit.

This project aims to build a regression-based model for predicting IMDb film scores. The primary goal of this project is to use the data gathered on nearly 3,000 films from IMDb to make an accurate and robust predictive model. The model will take various film-related attributes such as number of actors, genres, and budget as input, and generate a predicted score rated on a scale of 0-10 that a film with said attributes would earn on IMDb. To assess the predictive accuracy of the model once it is built, the model will be validated using an additional 40 films as input to assess the model's out-of-sample performance. To build our model, we analyzed 47 film features and their relationship to IMDb score, and used these findings to compile the features we thought had the greatest predictive power. We subsequently used cross-validation methods to test our model performance, and further adjusted the parameters of the model as necessary to arrive at the model with the most optimal predictive power, measured by the lowest mean-squared error (MSE) of the regression prediction. Unless otherwise specified, all data analysis was performed in R Studio.

[1] Sawers, P. (2015, October 30). 25 years of IMDb, the world's biggest online movie database. Retrieved November 10, 2020, from https://venturebeat.com/2015/10/30/25-years-of-imdb-the-worlds-biggest-online-movie-database/

[2] IMDb. (n.d.). Press Room. Retrieved November 10, 2020, from https://www.imdb.com/pressroom/press-pull-quotes/

DATA DESCRIPTION

We began our analysis by examining the distribution of each of the numerical variables in the dataset – IMDb score, budget in millions, year and month of release, duration in hours, number of actors, number of producers, production companies and production countries, number of languages, and number of directors.

First, we examined the target variable, **IMDb score**. The histogram in Figure 1 exhibits an approximately normal distribution with a mean of around 6.8, and a standard deviation of roughly 0.7. Further analyzing this distribution via the boxplot in Figure 10 reveals that the mean is about 6.8, but the normality of scores likely does not hold due to the large number of low outliers, and no outliers above the mean. We conclude that the distribution of IMDb scores takes the approximate shape of a normal distribution, but is left-skewed.

The first predictor examined is the **budget in millions**. Budget is heavily right-skewed, as seen in both Figure 2 and Figure 11. The mean of this distribution is about $20m as per Figure 11, and there is a large concentration of points in the $0-50m range. There is a long positive tail in this distribution. When analyzing the distribution of budget in millions with respect to the IMDb score of that film, we first looked at the residuals of a simple linear relationship. The residuals follow a clear trend, implying that this variable is non-linear. We then created a scatterplot and used local regression with a span of 0.2 to assess the relationship – the results of which can be seen in Figure 19. The trend does not appear to follow an easily defined form – which was confirmed through performing an analysis of variance (ANOVA) on regression models between IMDb score and budget. The ANOVA test compares the difference between tested models of varying polynomial degrees and suggests that a quintic (degree = 5) polynomial would fit the data most optimally.

The next variable of interest is the **year of release**. Year of release is highly negatively skewed – there are a very large number of films in the dataset that were made in the most recent years, and far fewer made before 1980. The histogram and boxplot of this variable can be viewed in Figures 3 and 12, respectively. The p-value of the residual test confirms that the residuals are not linearly distributed, such that the relationship between year of release and IMDb score is non-linear. Using local regression with a span of 0.2, as seen in Figure 20, we found that the relationship between year of release and score is, like budget, not easily defined. ANOVA suggests that a polynomial relationship of degree 3 most accurately captures this trend.

The histogram and boxplot distributions for **month of release** were not particularly telling, and so we decided to fit this variable in a simple linear relationship with IMDb score to explore possible trends. While the residuals confirm that the linearity assumption holds, there

seems to be very little trend in the data – the slope of the linear relationship which fits this regression is very small, indicating that the relationship between month of release and IMDb score is not very telling. A plot of the local regression is shown in Figure 26.

The next variable that we analyzed was **duration in hours**. This variable has a large concentration around 1.25 to 2.5 hours, and a long positive tail – seen in Figure 4. The boxplot in Figure 13 confirms this by showing many positive outliers. The whiskers of the boxplot are rather long, and roughly symmetrical. Thus, it seems very unlikely that a film is under 1.25 hours, while it is uncommon that a movie is about 2.75+ hours, but not exceedingly rare. Looking at a simple linear relationship between duration and IMDb scores, we found that the relationship is non-linear, evidenced by the very low p-value when conducting a residual test. Using local regression with a span of 0.2, we found that the relationship was certainly non-linear, but it did not follow a clear trend – evidenced in Figure 21. ANOVA indicates that the best relationship between duration and IMDb scores is a degree 5 polynomial.

Total number of actors is the next variable that we analyzed. Similar to budget, this variable has a large concentration around 10-20 actors, and a very long positive tail (Figure 5). The boxplot in Figure 14 confirms this – exhibiting a tight concentration around a low mean, and a large number of positive outliers. Looking at the simple linear relationship between total number of actors and IMDb score, the relationship exhibits strong evidence of non-linearity. Using LOESS and a span of 0.2 (Figure 22), we found that the relationship could possibly be of degree 3, but we still leveraged ANOVA to find the optimal polynomial relationship between the variables. ANOVA suggested that a degree 3 polynomial was indeed the optimal degree.

The next variable of interest was the **total number of producers**. This variable has a large concentration around 2 producers, and a positive tail in Figure 6, but the number of unique values for this predictor in the dataset is not very large. The boxplot in Figure 15 exhibits 6 positive outliers, and none below the mean. By examining the relationship between number of producers and IMDb scores, we find that the relationship is significantly non-linear, but the line of best fit in the linear relationship had slope of nearly 0. Using local regression and a span of 0.6 (the number of unique values is not large enough to use much lower values) indicates a similar sentiment (Figure 25). While the relationship seems somewhat non-linear, the trend is practically non-existent.

Next, we looked at the total **number of production companies**. The distribution of this variable was similar to the number of producers (Figure 7). There is a concentration at low values (around 1-5), and a long positive tail. The boxplot distribution also confirms this trend (Figure 16). Looking at the residuals between production company count and IMDb scores, we found that the relationship was on the cusp of being linear and nonlinear, as the p-value of the residual

test is 0.055. After plotting the relationship using LOESS with a span of 0.7 (Figure 27), we found the relationship did seem to be non-linear due to a small perturbation near 0. Experimenting a bit, we found that a degree 2 spline with 4 knots mirrored this relationship most optimally.

We then viewed the **total number of production countries**. This predictor has a strong concentration around 1 country, then it very quickly drops off (Figure 8). The mean is 1, and the boxplot has no whisker below (Figure 17). There are a few positive outliers in this plot. The relationship between the number of production countries and IMDb score is well modeled as linear according to the residual test's p-value, but the slope of this relationship is practically 0 – as shown in Figure 23 with local regression.

The final variable of interest is the **number of languages** for the film. Most films in the dataset contained only one language, and the number of languages spoken seemed to decay exponentially – according to the histogram in Figure 9. The boxplot in Figure 18 provided a similar sentiment, although it interestingly indicates that there is at least one film in the dataset which has no language – perhaps a silent film or a set of silent films. The relationship between the number of languages and IMDb scores is modeled best by a linear relationship. While the trend is non-zero, the slope of this line of best fit is rather small, indicating that the relationship may not have much predictive power – at least in isolation (Figure 28). Furthermore, there are very few unique values that the number of languages takes on in this dataset, so a linear relationship between this count and score may not be appropriate.

The **total number of directors** in a film was examined, but the histogram and boxplot indicated that nearly every film in the dataset had only one director, and thus this variable does not seem to be very helpful in modeling.

In summary, to examine all the variables and their relationship to IMDb score, we created a simple linear model between each variable and score, and examined the model coefficients, the R^2 value, and the p-value of the residual test of each relationship. These attributes assisted us in identifying the predictors which have the strongest predictive power for IMDb score. We found that duration in hours, drama genre, and year of release had the strongest power as per their R^2 value. The p-values for each of the residual tests help to confirm our findings from above regarding which variables are linear vs. non-linear. Additionally, we found that the predictors that exhibit nonlinear relationships are budget, year of release, duration in hours, total actors, and total producers, while the number of languages, number of production companies, and the number of directors are best modeled by linear relationships. Finally, the month of release and the total number of production companies are predictors that are on the cusp of linear and nonlinear.

We constructed a correlation matrix for each numerical variable to ensure that the predictors are uniquely determined and not influenced by each other (Figure 29). No two variables had a correlation coefficient higher than about 0.44 – well below the ~0.8-0.85 threshold which would suggest a collinear relationship between variables. Since we are interested in prediction rather than causality, collinearity is not a strong concern; however, our results nonetheless indicate that collinearity is not present in the dataset.

Although outliers were detected in many of the predictors, we chose not to remove outliers from any of the variables. Estimates of predictive power are more accurate by retaining outliers, as removing them would merely be ignoring data that the model does not fit well and would reduce the ability of the model to predict accurately in a test-setting.

MODEL SELECTION

Using our understanding of each predictor's relationship on IMDb score, we performed a number of tests to select the best features to use in our final predictive model. We opted to exclude all text-based categorical variables such as main producer name or main actor name from the subsequent analysis. This decision was made as many text-based variables included a high number of unique instances; for example, there were 1190 unique instances of the lead main actor, and the most frequently appearing actor in that category represented less than 1% of films in the dataset.

First, we transformed categorical variables (i.e. genre) into factors using the *as.factor()* function in R Studio. As non-numerical categories cannot be input into regression equations, this process re-codes categorical variables into a series of binary dummy variables which can then be input into regression type equations. The reality TV and short film genres were eliminated from the analysis at this stage as each variable only included one instance.

To perform feature selection and detect which variables have the strongest relationship with IMDb score, we first used the scikit-learn machine learning package in Python 3.7 to perform a LASSO regression with all variables (continuous and dummy categorical variables) with varying levels of λ to find the most important features. λ is a tuning parameter which imposes a shrinkage penalty on the predictor coefficient with respect to its relative influence on IMDb score, and penalizes regression coefficients of variables with minor contributions to IMDb score to zero. We varied λ between 0.01 and 0.1, with increasing λ corresponding to a higher shrinkage penalty. At a λ of 0.1, there were seven predictors remaining: year of release, duration in hours, total number of actors, action genre, comedy genre, drama genre, and horror genre.

These results were consistent with the ones found by comparing the R^2 value for simple linear relationships between each predictor and the IMDb score, as duration in hours, the drama genre, and year of release were both deemed important by the LASSO and in the simple linear regression. We decided to use this set of seven variables as one potential set of predictors to input into the subsequently regression models.

Following the LASSO regression, we tested a second feature selection technique for comparison by building a multiple regression model and using a backward-elimination process to identify the most important features. Starting with the entire set of 10 continuous variables and 28 categorical variables, we repeatedly ran the multiple regression model to test their joint effect on the IMDb score. In each subsequent step, we eliminated the least significant predictors according to the p-value of the coefficient until all the remaining predictors were significant at the 95% level. It is important to note that while performing multiple comparisons may lead to a higher occurrence of false predictor significance, this method can nonetheless prove powerful for feature selection. This process resulted in 16 final predictors: **5 continuous and 11 categorical.**

By comparing the two feature selection processes, we found that the entire set of variables deemed as important predictors by the LASSO regression were significant in the multiple regression. As such, we tested all subsequent models with both the entire set of 16 predictors from the multiple regression, as well as the condensed set of seven predictors from the LASSO regression.

From the initial data preprocessing stage, we had verified and eliminated the possibility of collinearity between quantitative variables introducing bias to the model. We moved forward to perform residual tests for both regression models, and again found that all variables in both models were non-linear except for the month of release (Figure 32). The p-value of the Tukey test for linearity was <0.05, suggesting that linearity assumptions were not satisfied for either set of predictors. These findings are however consistent with what was observed when examining linearity between the individual predictors and IMDb score.

Given the clear evidence of non-linearity of the models, we opted to utilize non-linear regression techniques by building a multiple polynomial spline regression model. As opposed to applying a combination of polynomial functions to fit the predictors, spline regression can improve the tendency of a simple polynomial transformation to overfit the data by dividing the dataset into several sections connected by "knots" and fitting each section of the data with a separate model. For each categorical variable, we varied the polynomial degree transformations from one to four. Since there was no clear evidence of changes in pattern from the individual regression plots of the variables, the spline knots were placed uniformly at either one knot at the **50th percentile** or two knots at the **33rd and 66th percentiles**. We opted to set a limit of two

knots and a quartic degree transformation, as to avoid overfitting the in-sample performance of the data.

Subsequently, the polynomial spline function was formed by using a basis function to connect varying polynomial segments of degree d, at k number of knot locations, where d = *{1,2,3,4}* and k = *{1,2}*. We tested the polynomial spline regression model using all combinations of knots and degree transformations applied to the continuous variables. Note that the polynomial spline was not applied to the variable month of release, as this variable was previously determined to demonstrate a linear trend with regards to IMDb score. To test the out-of-sample predictive performance of the model, each iteration was performed using K-folds cross-validation. We chose **15 folds** to divide the data into 15 equal sets. The K-folds cross validation performs 15 iterations of the model, where each iteration retains $1/15^{th}$ of the dataset as test data to validate the out-of-sample performance of the model. This is repeated until all 15 sets have been utilized as test data. A mean-squared error (MSE) of the test set is then generated, which is an average across all cross-validations. While this test was repeated for both the set of either 16 or seven predictors, the best predictive accuracy was obtained using the 16 variables (5 continuous and 11 categorical) obtained from the multiple-regression feature selection process. As such, the smaller model containing seven predictors was subsequently dropped from the analysis, and its results are not reported. The variables used for the final polynomial spline model are as follows:

- Budget in Millions
- Month of Release
- Year of Release
- Duration in Hours
- Total Number of Actors
- Genre Action
- Genre Animation
- Genre Comedy
- Genre Documentary
- Genre Drama
- Genre Family
- Genre History
- Genre Horror
- Genre Romance
- Main Actor 1 is Female
- Main Actor 2 is Female

The final number of knots and polynomial degree transformations chosen for the respective categorical variables are as follows:

$a_1 = 2,\ b_1 = 3$ $\quad$ $i = 1$: Budget in Millions

$a_2 = 2,\ b_2 = 2$ $\quad$ $i = 2$: Total Number of Actors

$a_3 = 1,\ b_3 = 1$ $\quad$ $i = 3$: Year of Release

$a_4 = 1,\ b_4 = 3$ $\quad$ $i = 4$: Duration in Hours

where a_i = knots, b_i = degree

The final model equation is as follows:

$$IMDb\ Score = f_{spline(Knots=\mathbf{2},Degree=\mathbf{3})}(Budget\ in\ Millions) +$$
$$f_{spline(Knots=\mathbf{2},Degree=\mathbf{2})}(Total\ Number\ of\ Actors) +$$
$$f_{spline(Knots=\mathbf{1},Degree=\mathbf{1})}(Year\ of\ Release) +$$
$$f_{spline(Knots=\mathbf{1},Degree=\mathbf{3})}(Duration\ in\ Hours) +$$
$$Genre\ Action + Genre\ Animation + Genre\ Comedy + Genre\ Documentary + Genre\ Drama + Genre\ Family +$$
$$Genre\ Horror + Genre\ History + Genre\ Romance + Main\ Actor\ \mathbf{1}\ Female + Main\ Actor\ \mathbf{2}\ Female + Month\ of\ Release$$

A visual representation of the polynomial spline transformations is presented in Figure 30. Using the final set of 16 predictors, the lowest MSE, which measures the model's out-of-sample performance and ability to predict new observations, was observed to be **0.5821065**. Note that due to the random splits used to create training and testing data when employing the cross-validation approach, the MSE may fluctuate around the observed point. To ensure replicability, the MSE reported was observed at the following random state: *set.seed(0*).

To further tune our model, we plotted the residuals of the model predictors to test for model heteroskedasticity. We found that the residuals were uncorrelated and uniformly distributed, ruling out the need to adjust for heteroskedasticity (Figure 32).

Next, we utilized step-by-step elimination of each of the 11 categorical variables from the model individually to reassess model performance. We found that removing any categorical variable would increase the MSE, leading us to keep the original model detailed above. While we opted to ultimately exclude text-based variables in our model, we also experimented with creating dummy variables for such variables, such as by grouping the main language of film to English or non-English films, but found that its effect on the final model seemed marginal, at best. Many other variables had a high number of categorical types, which would increase the number of categorical variables substantially. However, it remains possible that factors such as the main production company could have an influence on IMDb rating, as certain production companies such as Disney may bring a level of "star power" and reputation to a film. To explore this alternative, we attempted to take the top seven main production companies from the main production company variable as dummy variables, with all other companies as "Other", but again found that the decrease in MSE when added to the above model was marginal-to-none.

MANAGERIAL IMPLICATIONS

According to the Motion Picture Association of America, between 2010 and 2015, the total number of films made worldwide increased by 15.6%[3]. With the increase of streaming services over the years producing their own content, this growing trend persists. As the business market within the film industry continues to grow, so does competition. Therefore, predicting a movie's rating is also increasingly important to filmmakers in order to ensure that they capture high viewership, which will increase the likelihood of obtaining a high ROI in the form of box office revenues and/or distribution deals. Measuring the opinions of people who have seen a film is far from exact, and being able to predict what the audience will think of a film prior to its release is even more complex. A film score usually plays an important role in convincing people to spend their time and money to see a film.

Currently, most films are given a score after their release based on opinions by film critics and/or the general audience. Two of the most cited measures of film rating are: 1) formal polls of people who have seen a film (e.g. CinemaScore), and 2) collection of scores from individuals who willingly submit reviews online (e.g. IMDb and Rotten Tomatoes)[3]. Using IMDb's movie database, we have endeavored to predict a film's score prior to its release by relying on objective characteristics of the film such as genre and budget. This way, filmmakers will have insights into what features of a film are more likely to lead to a higher rating, and can customize their movies using these insights to boost their potential film score. Most viewers choose the films they watch based on the rating, therefore, a higher rating for a film tends to increase the likelihood of the film succeeding in terms of viewership and revenue.

Based on intuition, we might assume that if a film is translated into multiple languages, this is an indication that the film is successful as a larger budget has been allocated towards translation in order to reach a wider audience. Our analysis showed that this is not the case. The number of languages in which a movie is produced is not a significant predictor of its rating, and as such, this variable was not included in our model. Throughout our analysis, it was interesting to observe such insights that demonstrated that predictors of a film's rating differ from what is commonly assumed or expected and is more intricate than anticipated. Likewise, when creating our predictive model, we identified and eliminated several of such non-significant

[3]Richeri, G. (2016). Global film market, regional problems. *Global Media and China, 1*(4), 312-330. Retrieved November 8, 2020, from https://journals.sagepub.com/doi/pdf/10.1177/2059436416681576

[4]Wilkinson, A. (2018, August 13). CinemaScore, Rotten Tomatoes, and movie audience scores, explained. Retrieved November 8, 2020, from https://www.vox.com/culture/2018/8/13/17657264/cinemascore-rotten-tomatoes-audience-score-metacritic-imdb

variables. Holding all other factors constant, we ultimately observed that when quantifying the impact of each variable in isolation on the IMDb score, the best predictors for a movie's rating are as follows:

- **Month of Release:** with each increase in month (i.e. from January to December), the film rating increases by 0.0357 points on average
- **Genre (Animation):** animated movies have a score that is 0.115 points higher on average than non-animated movies
- **Genre (Documentary)**: documentaries have a score that is 0.851 points higher on average than non-documentaries
- **Genre (Drama)**: films that fall within the drama category have a score that is 0.578 points higher on average than other genres
- **Genre (Action)**: action movies are 0.325 points lower on average than other genres.
- **Genre (Comedy):** comedy movies are rated 0.389 points lower on average than other genres.
- **Genre (Family)**: family movies are rated 0.325 points lower on average than other genres.
- **Genre (History)**: history movies are rated 0.598 points higher on average than other genres.
- **Genre (Horror)**: horror movies are rated 0.693 points lower on average than other genres.
- **Genre (Romance)**: romance movies are rated 0.0114 points higher on average than other genres.
- **Primary Main Actor is Female**: having a female lead actress decreases the film's score by 0.224 on average.
- **Secondary Main Actor is Female**: having a second female lead actress decreases the film's score by 0.0856 on average.
- **Total Number of Actors***
- **Duration (in hours)***
- **Budget (in millions)***
- **Year of Release***

** These predictors are non-linear, as such, their impact on the film score cannot be quantified by observing the coefficient of the linear regression model.*

Overall, we can observe that films tend to have a better rating if they fall within the genre of animation, documentary and/or drama, are longer in duration than the average film, and are released later in the year. This information is particularly important for filmmakers when creating their movies. For instance, using our model, a director could infer that making a comedy with a female lead which will be released in May will garner lower ratings than a drama with a male lead set to be released in September. Therefore, the director may choose to produce the latter film if his primary goal is securing higher ratings. In terms of casting, the director would lean more towards a larger number of actors in the cast, budget-permitting, of course.

Unlike traditional film ratings which are obtained after a film is released, movie directors and producers can use our predicted rating to more confidently plan the release date, distribution channels, marketing strategies, and other operational aspects of the film prior to its release. For instance, a high film score tends to correlate with a longer run in theaters, which is understandable: if the film has positive reviews in its opening weekend audience, this creates a "buzz" and word-of-mouth will encourage more audiences to go see the film. The film score can therefore help distributors decide whether, and how rapidly, the film should expand to more theaters. Our model would help the directors and distributors tailor their strategies beforehand in anticipation of real audience feedback once the film is released. Additionally, streaming networks such as Netflix tend to pay a higher premium to buy the rights to more popular films, therefore, these streaming services could use our model to adjust their acquisition budgets prior to a film's release.

An important caveat to note is that when building our predictive model, the IMDb scores we used to train our model are based on past audience scores. Therefore, the predicted score we generate will be a good measure of one type of audience reaction: the audience most primed to like the film in the first place. It is not, however, a measure of the film's ability to persuade those who are skeptical about the film or its genre to love it or view it.

Our model produced a Mean Squared Error (MSE) of 0.582. This means that our model's average error when predicting a film's score is about 0.761 points (which is the square root of the MSE). In other words, on average, our model's predicted score differs from a film's actual score by ±0.761 points.

APPENDIX

FIGURES 1-9: HISTOGRAMS OF PREDICTORS

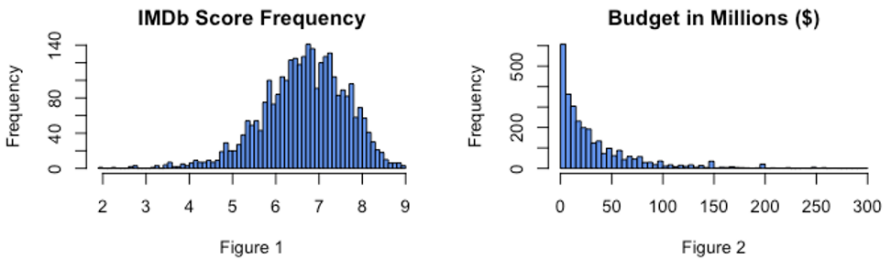

Figure 1

Figure 2

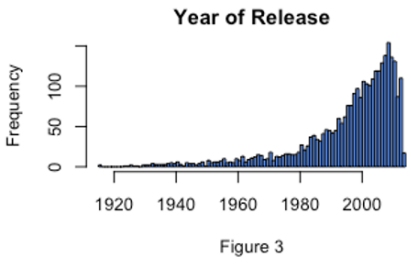

Figure 3

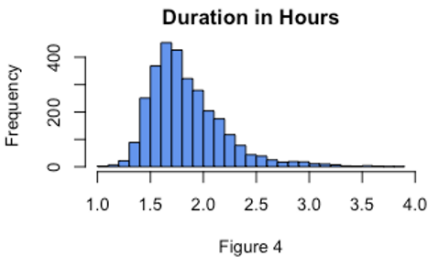

Figure 4

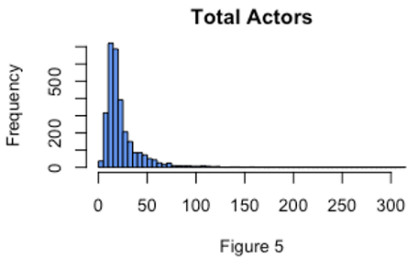

Figure 5

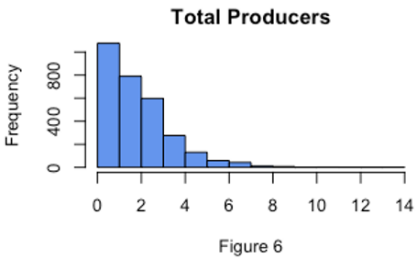

Figure 6

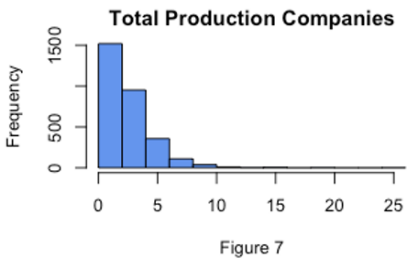

Figure 7

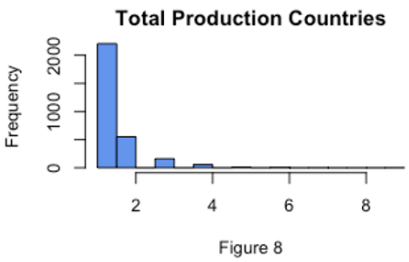

Figure 8

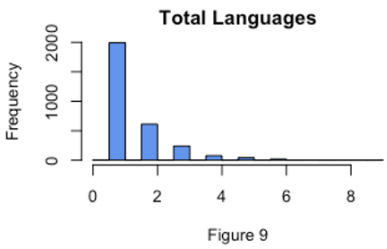

Figure 9

FIGURES 10-18: BOX PLOTS OF PREDICTORS

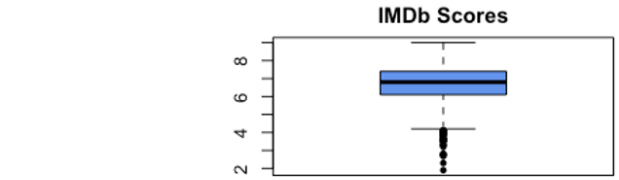

Figure 10

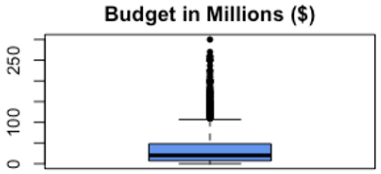

Figure 11

Figure 12

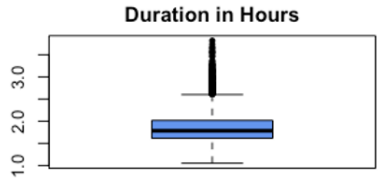

Figure 13

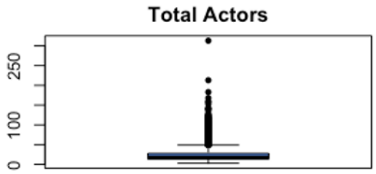

Figure 14

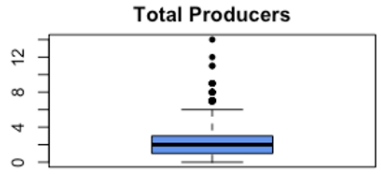

Figure 15

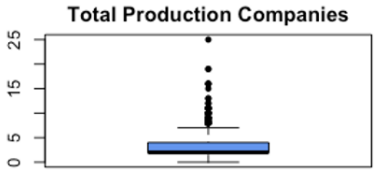

Figure 16

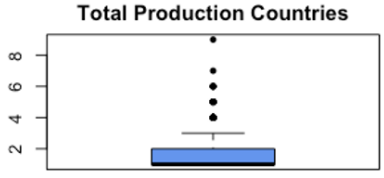

Figure 17

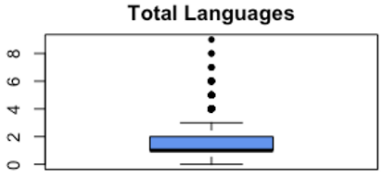

Figure 18

FIGURES 19-28: POLYNOMIAL REGRESSIONS OF NON-LINEAR CANDIDATE PREDICTORS

Note that in the below graphs, the y-axis is IMDb Score for each plot. The label was omitted to increase visibility of each graph. The predictor is labeled below each graph:

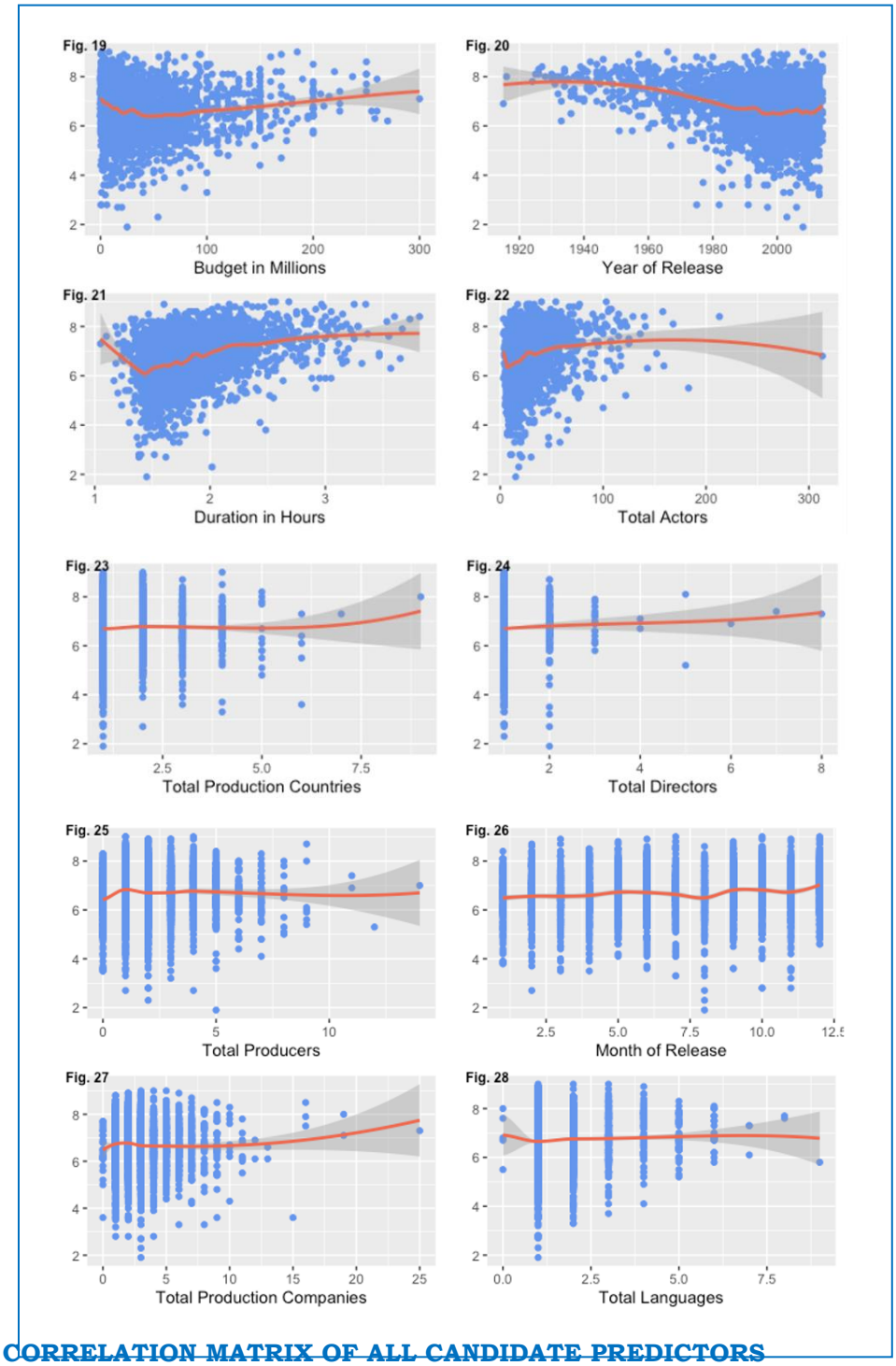

Figure 29: CORRELATION MATRIX OF ALL CANDIDATE PREDICTORS

Note that in the correlation matrix, the variables have been omitted for readability.

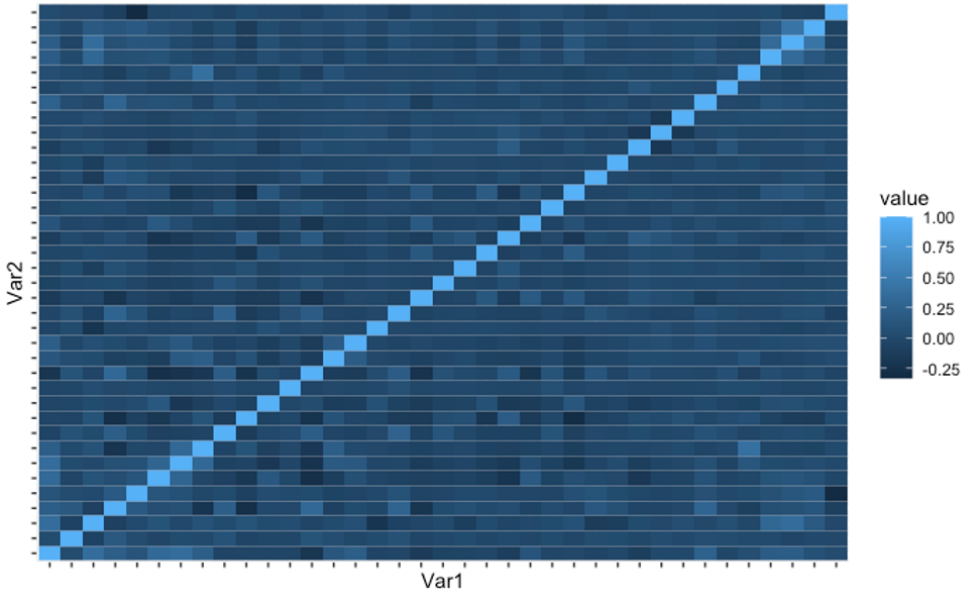

FIGURE 30: POLYNOMIAL SPLINES FOR TRANSFORMED MODEL PREDICTORS

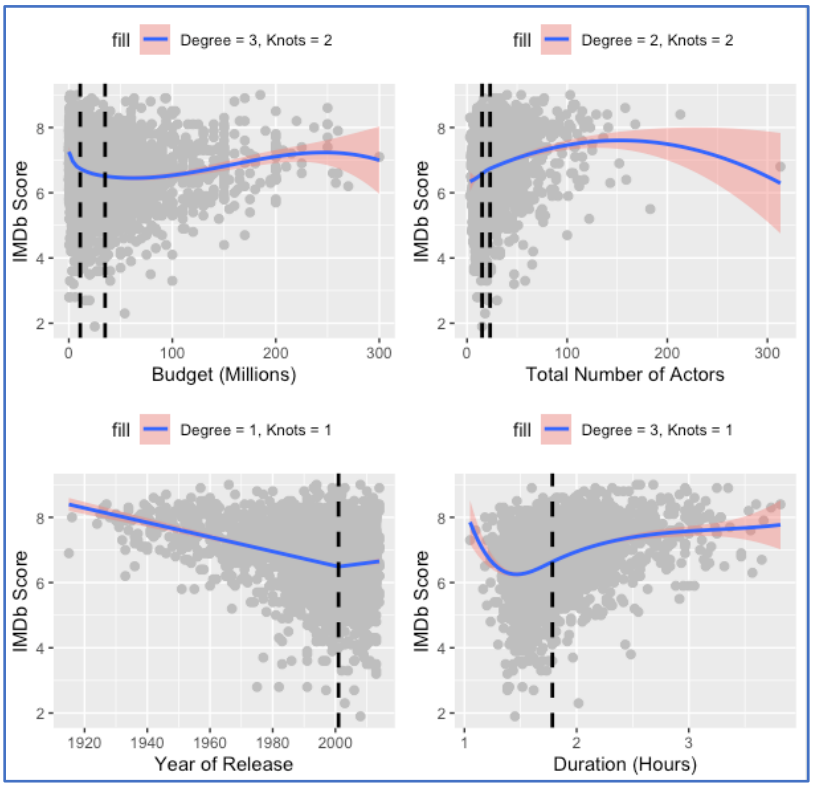

FIGURE 31: HETEROSKEDASTICITY PLOTS FOR FINAL MODEL PREDICTORS

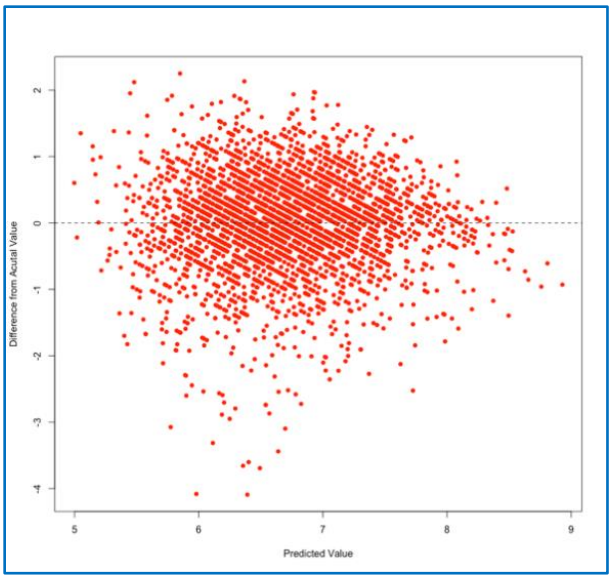

FIGURE 32: RESIDUAL PLOTS FOR FINAL CATEGORICAL PREDICTORS