Factors Contributing to Film Popularity

Yana Yuan	**Naoki Yuasa**	**Pei-hsiu Kao**	**Jenna Parkin**
1359969	1450462	1425485	1363232
COMP20008	COMP20008	COMP20008	COMP20008
yanayuan@student.unimelb.edu.au	yuasan@student.unimelb.edu.au	kaop@student.unimelb.edu.au	jparkin@student.unimelb.edu.au

Executive Summary

From applying random forest and linear regression to the features with popularity as determined by IMDB votes and TMDB popularity, we observe a high correlation between popularity with the following features:

- Horror (genre)
- Follows (description)
- Daughter (description)
- Thriller (genre)
- Crime (genre)
- Western (genre)
- TMDB and IMDB scores

The commonality between these features includes genre, description, and scores. Indicating that genres and descriptions are the better predictors of popularity other than TMDB and IMDB scores. Of the remaining categories, we also investigated age certification. Though age certification has an overall lower correlation with popularity, interestingly within this subset, R-rated films and other mature age restricted films tends to have higher popularity in comparison to the less restricted films. The report focuses on the categorical variables genre, description, and age restriction. After analysis, we conclude that the popularity of a film can be best predicted by IMDB score, TMDB scores, and the genre of the film.

Relation between Description, Age Restriction, Genres, Actors, and Popularity of the film

The relationships between each feature of description, age restriction, and genre against the popularity measure by IMDB votes and TMDB popularity are investigated. By applying various methods to the variables, we aim to identify which variable and method best predicts the popularity of the film. Methods and techniques were chosen based on the data types and features, and then we incorporated the use of two supervised machine learning processes: linear regression and random forest. Both were used separately for IMDB votes and TMDB popularity labels. The choice to use Random Forest was due to its ability to capture nonlinearity and robustness to outliers, and the use of linear regression is for its interpretability. Using both can validate results and provide more diverse insights. Other variables such as id, person id, IMDB id, character, and name were considered irrelevant as they are unique to each film. Whereas the variables roles, runtime, production country, and seasons have extreme levels of variability or are very sparsely associated with a film, therefore were discarded as well. IMDB scores and TMDB scores are closely related to popularity, however, the following focuses on the other variables as they provide more unique insight into popularity predictions.

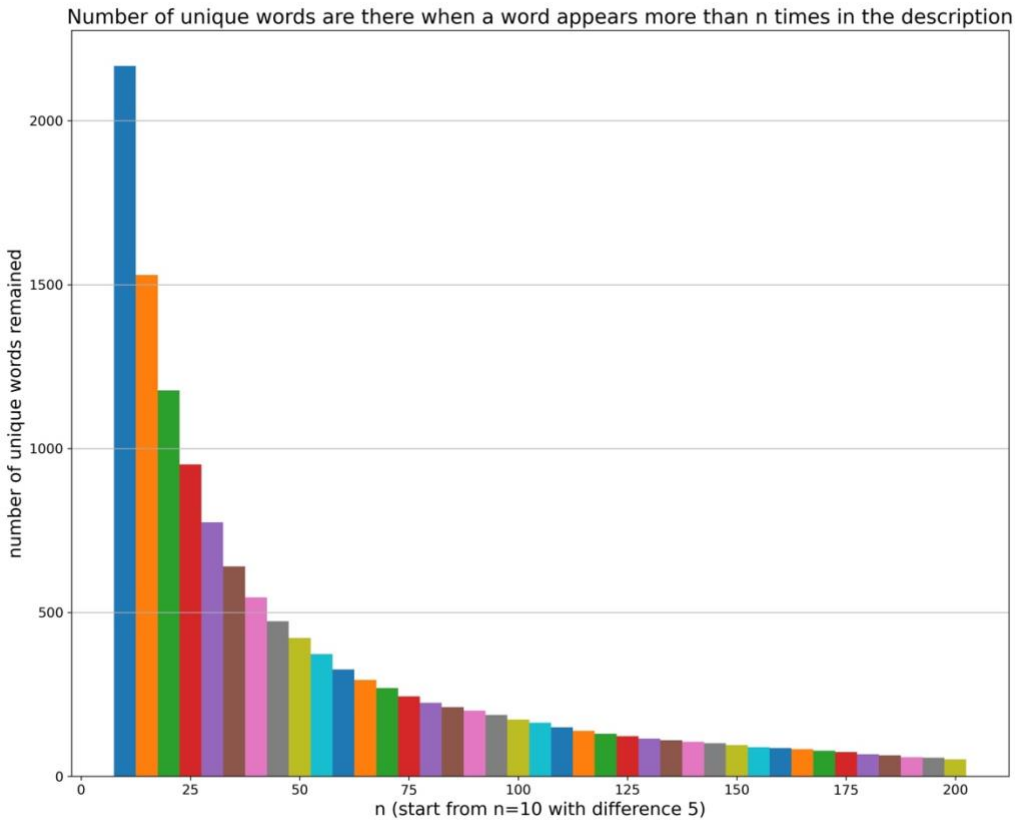

Figure 1: Number of unique words that remain when a word appears more than 'n' times in the description.

For description's relation with popularity, common descriptive words were isolated to compare the popularity of the film with that word in the description. This enables us to find commonalities between the descriptions, thus giving a basis for analysis. Through inspection, the descriptions all have many function words and of the remaining content words, many only appear in a low number of descriptions. Hence, before analysis, tokenization, stop word removal, and lemmatization was applied to remove the insignificant differences, allowing for more common words in description. Then, only word that appear in more than 150 descriptions were considered for further analysis. As shown in Figure 1, around 150 for appearances, the number of unique words decreases to approximately 95. After the dimensionality reduction we have a more manageable set of words, the higher frequency requirement also allows for

better popularity predictions as each word remaining is associated with a high number of popularity scores.

age_certification	tmdb_popularity_adj
G	2.285004
NC-17	3.837902
PG	3.452113
PG-13	3.679748
R	4.056666
TV-14	3.452012
TV-G	3.352476
TV-MA	3.390366
TV-PG	2.793306
TV-Y	2.351463
TV-Y7	3.606185

Figure 2: TMDB popularity of each age certification.

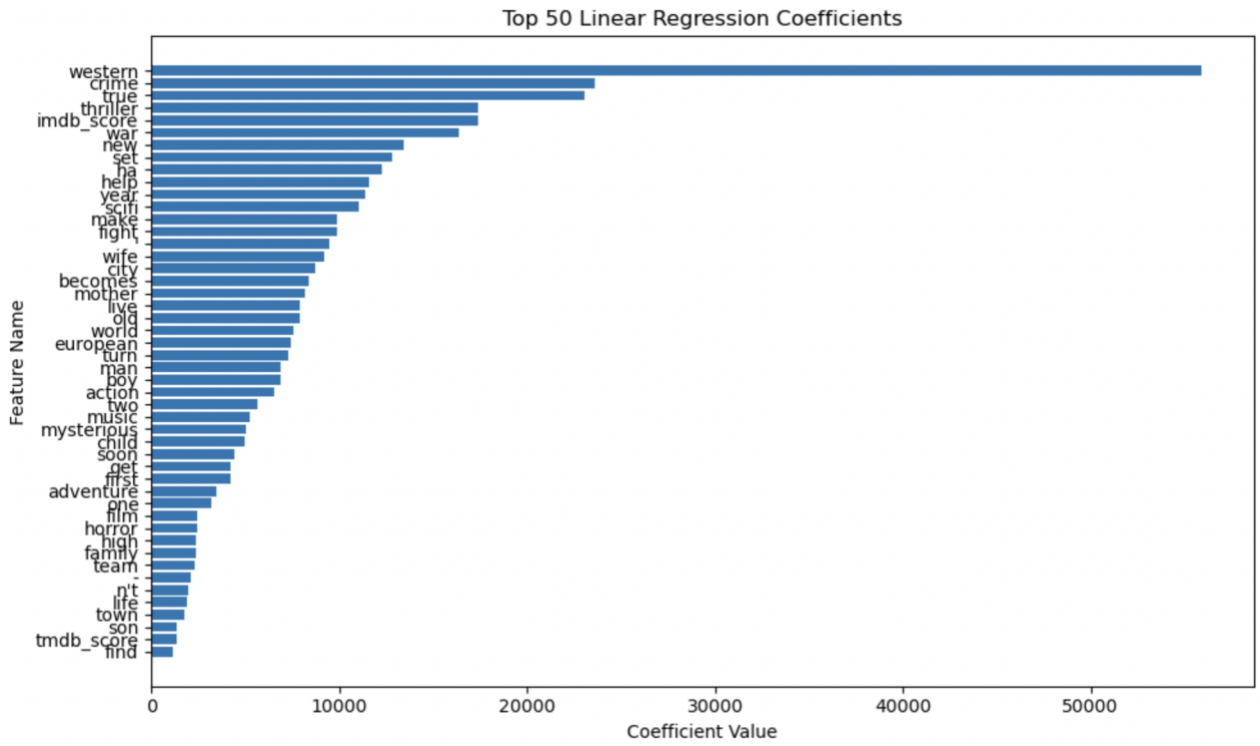

Figure 3: Top 50 linear regression coefficients for IMDB scores

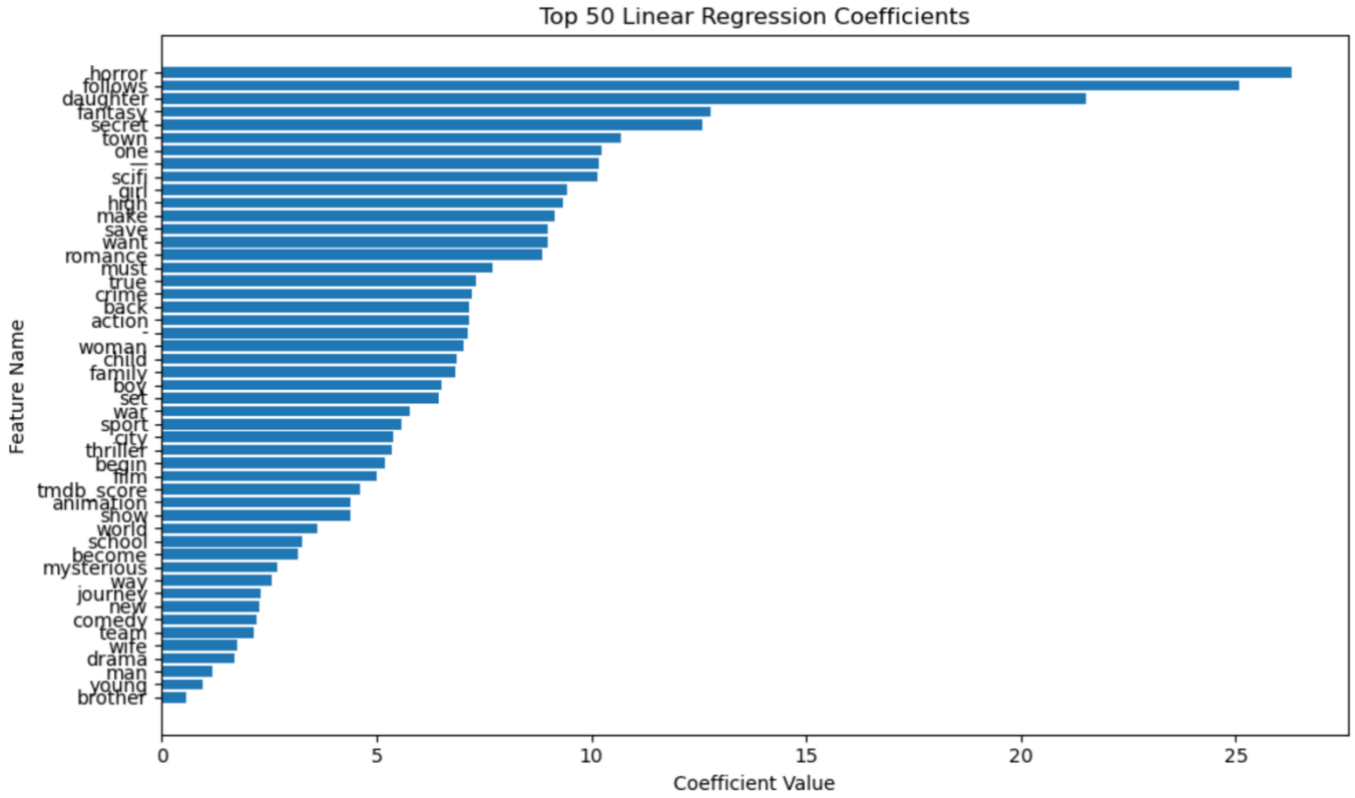

Figure 4: Top 50 linear regression coefficients for TMDB popularity

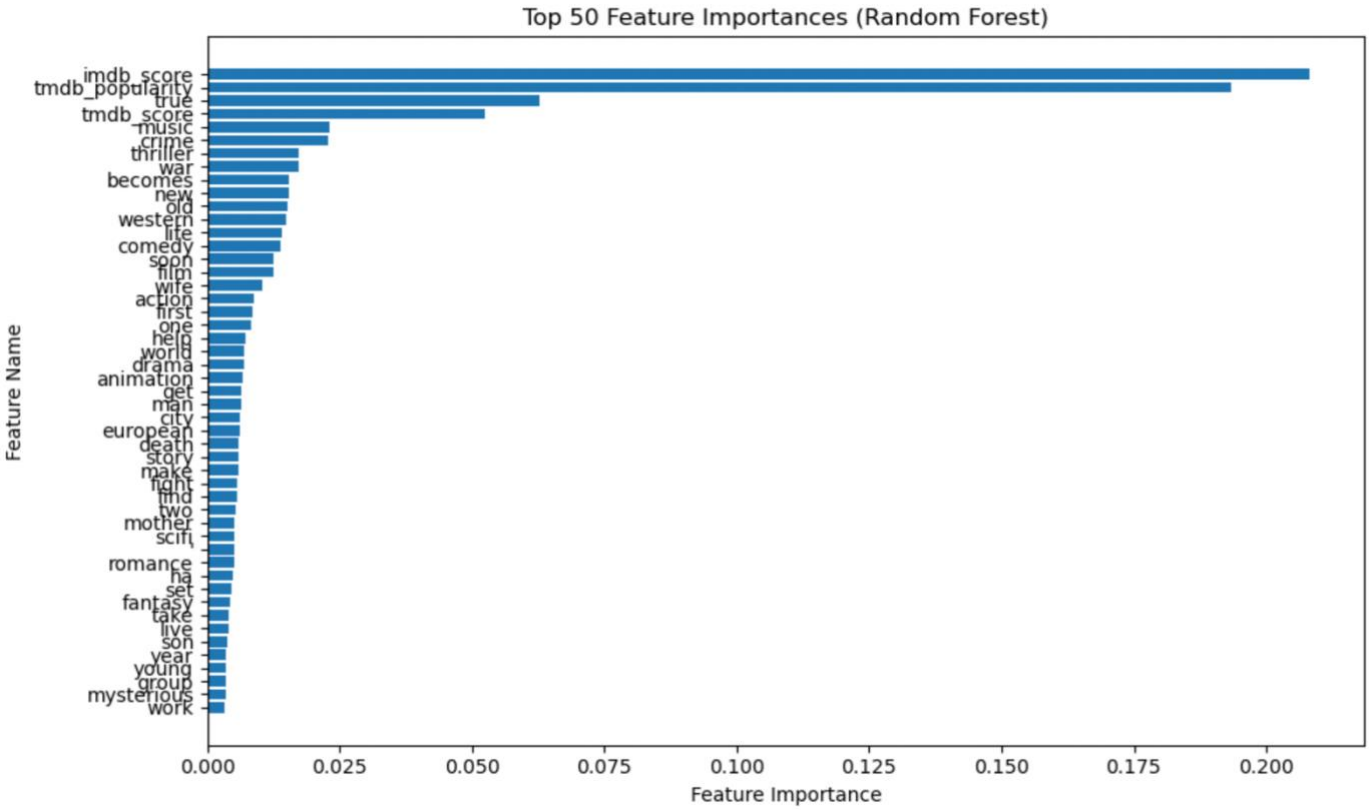

Figure 5: Top 50 important features by random forest for IMDB scores

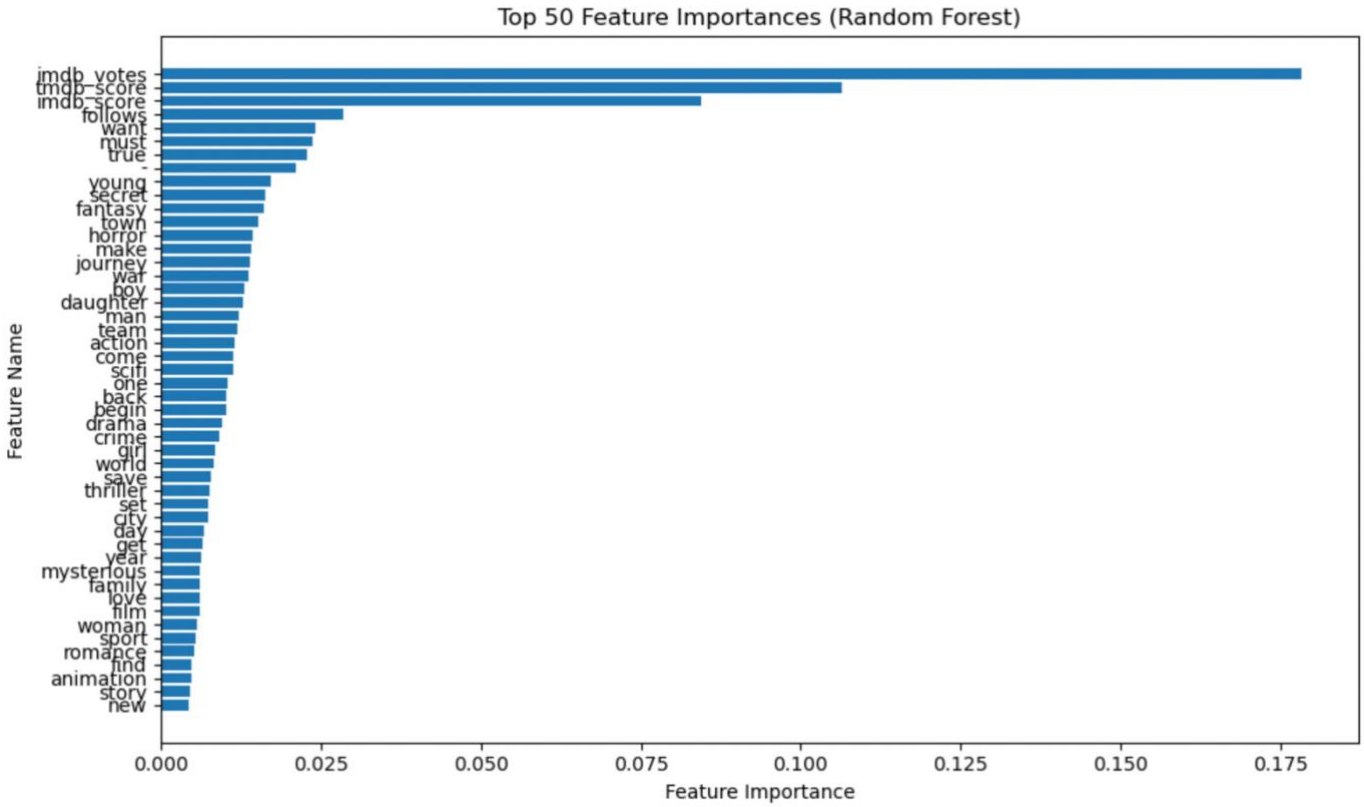

Figure 6: Top 50 important features by random forest for TMDB popularity

Then a linear regression and random forest is performed for all remaining words, along with other features later discussed. The coefficients are displayed in Figure 3 and Figure 4 and feature importance in Figure 5 and Figure 6. According to IMDB scores the most popular descriptions words are “crime”, “true”, “war”, “new”, “set”, and “help” in the coefficient rankings. Also, in coefficient rankings for TMDB popularity, the description words corresponding to the highest popularity are “follow”, “daughter”, “secret”, “town”, and “one”. From random forest, the top important description for IMDB scores are “follows”, “wants”, “must”, “true”, and “young”, for TMDB popularity they are “true”, “music”, “new”, “old”, “life”. There is little overlap between the sets of words in each type of comparison. Potential explanations for this phenomenon are that the popularity of the movies is not correlated or only slightly correlated with the words in the description, which would explain the variation, in this case the popularity is not dependent on the words in the description. Otherwise, it suggests that the IMDB score and TMDB popularity measure popularity differently and the correlation of description words may differ between linear and non-linear to explain the dissociation between linear regression and random forest results.

For the relationship between age restriction and population, initial analysis demonstrated that R-rated movies are overall more popular. From Figure 2, shows that R-rated films are the most popular followed by NC-17 and TV-Y7, all leaning towards a mature audience, similar trends are also present for the contrary as the least popular films, that is G and TV-Y rated have no age restrictions and are generally targeted towards children. To eliminate the potential effect of certain more popular genres being more likely to be rated R or other restrictions and vice versa more popular age certifications being more likely to be certain genres skewing our conclusions, analysis was conducted on each genre and the age restrictions within them. Pairwise comparison was done between each genre and age restriction using a Tukey test, the data of no significance as indicated by a high p-value (over 0.05 as a 95% confidence interval was used) were eliminated. From previous investigations, R-rated films are the most popular. After further analysis, R-rated drama, action, and thriller films are significantly higher than comedy R-rates films. As demonstrated in Figure 3 and Figure 4, action and thriller films of all age restrictions rank higher for correlation to higher popularity than comedy films, the difference between comedy and drama films are either not displayed or relatively small. Suggesting that the higher popularity of R-rated films and thriller and action films are correlated with each other, therefore cannot conclude if R-rated

films are popular because of their certification or for being more likely to be of a popular genre. Overall, as shown in Figure 3, Figure 4, Figure 5, and Figure 6, no age certification is amongst the top correlated features with popularity, so although trends exist within the category, in comparison to other features, age certifications are an inferior predictor.

Isolating genres and population, the observed ranking of the most popular genres is somewhat consistent in all the methods used. The top genres across all methods and popularity measurements are “Western”, “crime”, “thriller”, “war” “Sci-Fi”, “horror”, “fantasy”, and “romance”. All of which, other than “romance”, appear in two or more of the method’s top 5 most popular genres. The compatible results across different methods reveal that regardless of the method or popularity measurement, the listed genres have a strong correlation to high popularity. Linking back to the discussion on R-rated films being more popular now that we have established the strong correlation between the genres and popularity, also given that R-rated films occupy more of certain more popular genres. Perhaps we can infer that popularity is more dependent on genre than age certification.

[15]:

OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.162
Model:	OLS	**Adj. R-squared (uncentered):**	0.145
Method:	Least Squares	**F-statistic:**	9.452
Date:	Thu, 05 Oct 2023	**Prob (F-statistic):**	3.38e-144
Time:	19:15:57	**Log-Likelihood:**	-33715.
No. Observations:	5850	**AIC:**	6.766e+04
Df Residuals:	5733	**BIC:**	6.844e+04
Df Model:	117		
Covariance Type:	nonrobust		

Figure 7: OLS regression results for TMDB popularity

To reflect on the models used, an OLS regression was conducted. In Figure 7, the OLS regression is regarding TMDB popularity, an OLS regression was also conducted for IMDB votes, receiving similar results. The similarity in results is likely due to the usage of one-hot-encoding in preprocessing, which assigned variables with values between 0 to 1. The r^2 for TMDB popularity is 0.162, and 0.175 for IMDB votes. DB test both showing no autocorrelation between variables. JB test showing that the input variables are not normally distributed, but this might also be a side effect of the one-hot-encoding. The test also shows that there is strong multicollinearity for input variables. Generally, the models are statistically significant.

Conclusions and Limitations

Limitations may exist from the current models and analysis. Firstly, many of the variables have confounding variables and therefore are not independent of one another. For example, descriptions often overlap, and are somewhat determined by the genre. So, when comparing the two’s ability to predict popularity, we cannot credit the prediction to one variable only, also it may be possible for the two variables to simply have redundant information as they produce similar predictions. An improvement could be to perform regression between each variable that is not popularity, to see whether there is correlation between the current chosen explanatory variables or to calculate the mutual information for each pair of variables.

Secondly, the comparison of description against popularity may be over-simplified, as we only consider single words that are featured in the description. Rather, a description has many other factors that may contribute to popularity, such as how expressive it is, which relies on the comparison of word combinations. This can be improved upon by considering short clauses as well as single words, a machine learning model could be built to recognize and group description with similar known description's popularity.

Thirdly, as two models, random forest, and linear regression, were applied and returned differing results, it suggests that there is a significant difference between the linear and non-linear associations. Comparison between linear and non-linear associations currently cannot be done fairly. A potential next step would be to linearize all feature's regression, giving a foundation for comparisons, or first normalize the values then linearize before regression.

Furthermore, we have an imbalanced dataset. For example, the majority (approximately 80%) of all given age restrictions are classified as R, the model will then be biased towards R classifications as the training set will have more R-rated data causing overfitting for R-rated films. Having a lower number of test data that has a balance of each category would resolve the imbalance but considering the low fraction of non-R-rated films, this would dramatically decrease the training set, thereby increasing risk of underfitting. Further analysis should be done to determine an optimal complexity to avoid both underfitting and overfitting.

Finally, there are missing values for many films, this is especially prevalent for the category seasons and age certification. Our approach was to remove all films with a missing value in the category for analysis. Possible issues that may arise with this approach include the missing values being correlated with a specific popularity value and the lack of those data points will skew the result. Solutions could be to predict the missing value by using other information for the same film, however, this requires significant domain knowledge, human judgement, confidence in the other categories' correlation to the missing value being able to accurately predict and fill in the gap, and possibly require external data. A different simple alternative would be a zero-r or one-r approach to approximate the missing values.

Despite the limitations, valid conclusions can be made. We have established that IMDB and TMDB have slightly different measurements for popularity though they have high correlation with each other, the results from analyzing other variables against each popularity measure restored different results. Though individual words in descriptions appear to have high correlation with popularity, it is more likely a result of sheer chance rather than dependency as the highly correlated words do not mirror each other in the different methods. R-rated films, along with other more mature audience films tend to have higher popularity. However, after analysis with genre and popularity, we observe a high correlation between the two, so given the overlap between R-rates films and popular genres, the popularity of R-rated films is likely attributed to R-rated films being more likely to be of a popular genre. Also, worth noting that IMDB score and TMDB scores has extremely high correlation to popularity according to the random forest method as seen in Figure 5 and Figure 6. This aligns with intuition that highly rated films tend to be more popular. Overall, the best predictors of popularity are IMDB score, TMDB score, and genre.