Exploring the Influence of Key Media Features on Perceived Success: A Study of IMDb Ratings

Leo Kouskouris	**Hugh Buntine**	**Neel Singh**	**Eugene Yap**
1168671	1172794	1309156	1353623
COMP20008	COMP20008	COMP20008	COMP20008
lkouskouris@student.unimelb.edu.au	hbuntine@student.unimelb.edu.au	neels@student.unimelb.edu.au	eugeney@student.unimelb.edu.au

Executive Summary

This report explores the effect of genre, runtime, and choice of director on the perceived success of a movie, as measured by its IMDb score. We focused on the use of two supervised machine learning models in our analysis.

The first model used linear regression to predict the IMDb score of a movie based on the average IMDb score of a selection of the director's other movies. One iteration of this model calculated this average based on movies released before 2017, whereas another iteration randomly split the dataset and calculated the average performance of each director's movies in the first dataset to predict the performance of movies in the second dataset. The average coefficient of determination was 0.23 and 0.22 respectively, suggesting that the performance of a director's previous movies is a moderately good predictor of the performance of their future movies.

Our second model used a k-Nearest Neighbours (k-NN) algorithm to predict whether the media's performance would be 'Good' (IMDb score above the median of 6.6) or 'Poor' (IMDb score below the median of 6.6), using media runtime and genre as predictors. This model had some predictive power, classifying media with an accuracy of 67%, relative to a Zero-R model which had an accuracy of 52%.

Although we noted several limitations in our analysis, our findings do provide value for media producers. Our findings support the hypothesis that the choice of director is a key factor in determining the likelihood of media's perceived success. They also suggest that media producers should consider the runtime and genre as potentially significant factors which can influence the perceived success of their media.

Introduction

The report aims to investigate how three key attributes of media – runtime, genre, and the choice of director - influence the perceived success of a movie or TV show, as measured by its IMDb score.

In this report, we use the term 'media' to refer to both movies and TV shows.

Our analysis leverages two interconnected datasets: the Credits Dataset and the Titles Dataset. These datasets provide a variety of information regarding the characteristics of media content, as well as the actors and directors cast for each of these. After extensive exploratory data analysis, we chose to focus on the three key variables stated above. In total, we had 3744 rows of movies, 2106 rows of TV shows and 4550 director instances available for our analysis.

Methodology

General data preparation

We noted several issues with the datasets provided, including unclean strings, errors in production country codes and redundant age certifications. Although we rectified many of these issues to assist in the exploratory stages of our analysis, our discussion will focus on the methods used to process the runtime, genre, director, and IMDb score columns.

We noted that 482 media in our dataset were missing IMDb scores. These values seemed to be missing completely at random. Given the importance of IMDb score as our target label, we chose to use data imputation methods rather than eliminating these rows. We noted that both IMDb and TMDb scores were collated from user ratings and had a moderate positive correlation of 0.57. Given this, we chose to create a linear regression model which predicted IMDb scores from TMDb scores and then used this to fill in missing IMDb values. This linear regression had a mean squared error of 0.90. Relative to the scoring range of 0 to 10, this MSE suggested our predictions were accurate.

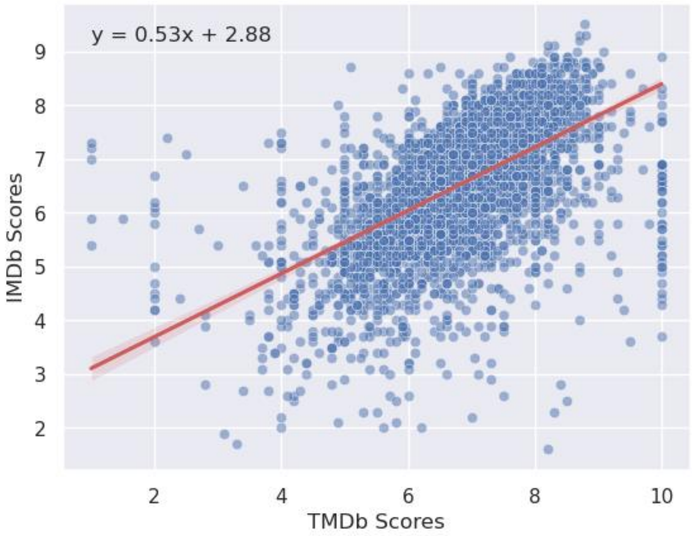

Figure 1: A scatterplot depicting our test data, with the regression line that was fitted on our training data.

IMDb Scores against Runtime and Genre

The first relationship which we explored in our analysis is whether the choice of genre and runtime for media has any meaningful relationship with its perceived success.

We tested this through the development of a k-nearest neighbours (k-NN) model. This model used the runtime and genre of a media to classify its expected performance into one of two categories: ‘Good’ (IMDb score in the range [6.6-10]) or ‘Poor’ (IMDb score in the range [0-6.6)). We discretized our numerical IMDb score data into one of these two bins, converting it into categorical data. Our choice of model here is justified as we are seeking to classify our data into one of these two categories.

We note here that IMDb score was chosen as a preferred target label over TMDb score, given the larger presence of the platform (83 million users compared to 3 million users).

The threshold value was chosen to be 6.6 as this is the median IMDb score for our original and imputed dataset. As such, ‘Good’ media is in the top 50% of performers, and ‘Poor’ media is in the bottom 50%.

After discretising our IMDb values, we performed one hot encoding on genre, converting our nominal genre variables into numerical variables. This resulted in 19 new columns, one for each genre variable. The variable runtime was also added as a model feature here, as runtime can impact the perceived success of a media, particularly when it is too long or too short.

We then implemented a cross-validation method with 5 folds. For each fold, we performed feature selection using mutual information on the genre columns in our training set. This would reduce the number of genre features from 19 to between 11-12. Our model then sets the number of neighbours to be used in our algorithm as 25. The high choice of n here is reasonable given the sparsity of our genre columns, as well as the substantial number of data points we had available to train our model.

IMDb Scores against Choice of Director

The second key relationship which we explored in our analysis is whether the choice of director for media has any meaningful relationship with its perceived success. Common knowledge suggests that media which are directed by high-profile directors will, in general, have more success. We aimed to test whether there was any truth to this claim.

We tested this through supervised machine learning methods, specifically by developing and evaluating linear regression models. This is justified as both our predictor and regressor are numerical variables. Our response variable was the IMDb score of media, and our predictor was the average IMDb scores of a selection of the director’s other media.

The first regression model we will focus on was built by ordering our data, and then splitting our data into media released before 2017, and media released after this period. Note that this split is not for training and testing purposes, but instead for constructing a new dataset.

Our chosen allocation method here provides us with the benefit of allowing us to explore the relationship between a director’s success over time. That is, we can explore whether a director who has produced successful media in the past is also more likely to produce successful media in the future.

For our earlier dataset, we calculated the average IMDb score across each director’s media, and then searched our second dataset to see if a director also appeared here. When a match appeared, we would insert the average IMDb score of the director’s previous media alongside the IMDb score of the media in the second dataset.

After completing this, we had 415 data points, each of which contained the IMDb score of our director’s media (our response variable), alongside the average IMDb score of the media that this same director had produced before 2017 (our predictor variable).

Note here that we cannot build this new dataset without initially splitting the data. If we were to do this, we would be calculating average IMDb scores from the same set of media whose IMDb score we are later attempting to predict. This would produce a model that is inherently biased.

Having formed our new dataset, we then split this into a training and testing set, with a 60-40 split, and our regression model was built.

We also explored this relationship by splitting our initial dataset into two datasets with random allocation, rather than making this allocation according to the release year.

In this second model, we assessed whether the success of a director's media could be predicted from the average success of a random selection of their other media. Note that this relationship no longer has a temporal factor.

Data Exploration and Analysis

In our analysis, our key variable of interest was IMDb Score. As such, we provide descriptive statistics (Table 1) and visualizations (Figure 2) highlighting this variable's distribution.

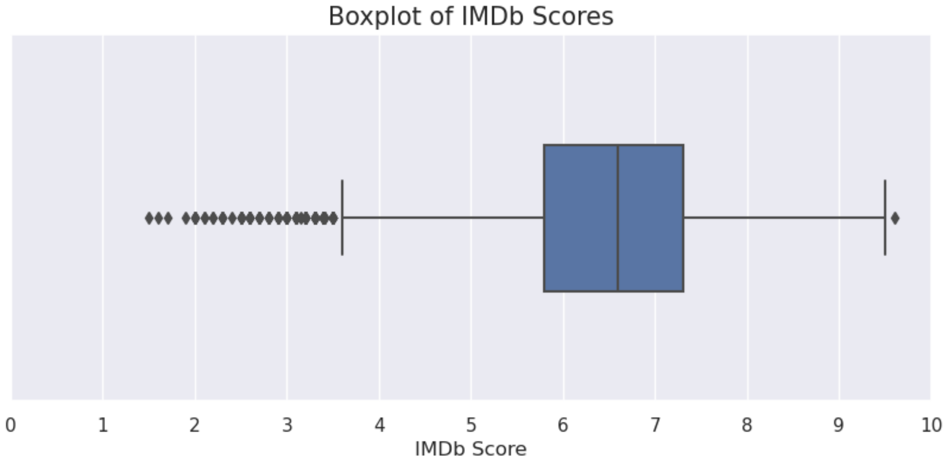

Figure 2: A boxplot depicting the distribution of IMDb scores in our cleaned dataset.

Descriptive Statistic	Corresponding value for IMDb Score
Mean	6.51
Median	6.6
Mode	6.5
Minimum	1.5
Maximum	9.6
Standard Deviation	1.14
Q1	5.8
Q3	7.3
IQR	1.5
Number of Outliers	93

Table 1: A table depicting descriptive statistics of IMDb score.

The central tendency of our IMDb scores, as noted by the mean and median, is given by 6.51 and 6.6, respectively. We also see that the spread of our data is low, as indicated by the values for our IQR and standard deviation, which are 1.5 and 1.14, respectively. Our distribution also exhibits some left skewness, primarily driven by lower-end outliers, representing movies which performed particularly poorly.

During our exploratory data analysis, we also analyzed correlations between numeric variables, which are summarized in Figure 3.

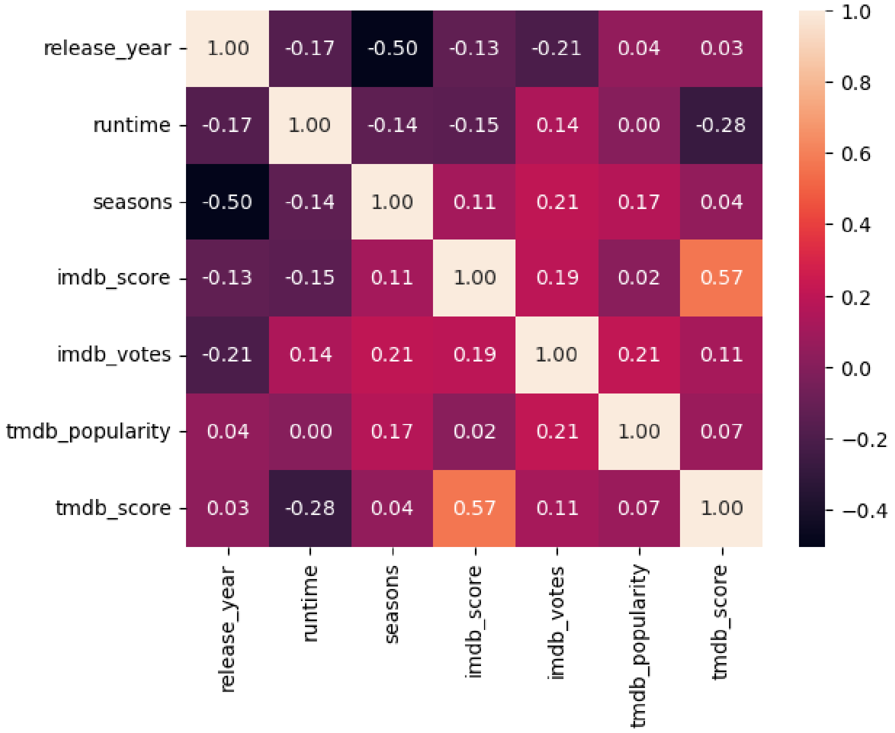

Figure 3: Correlation heatmap of the numeric features of all media.

Aside from TMDb score, which was discussed earlier, most numeric features exhibit a limited correlation with IMDb score. We note that the correlation between runtime and IMDb score appears to be significantly different from zero.

In Figure 4, we see that 'comedy' and 'drama' are by far the most common genres in our dataset, with 'western' and 'war' being the least represented.

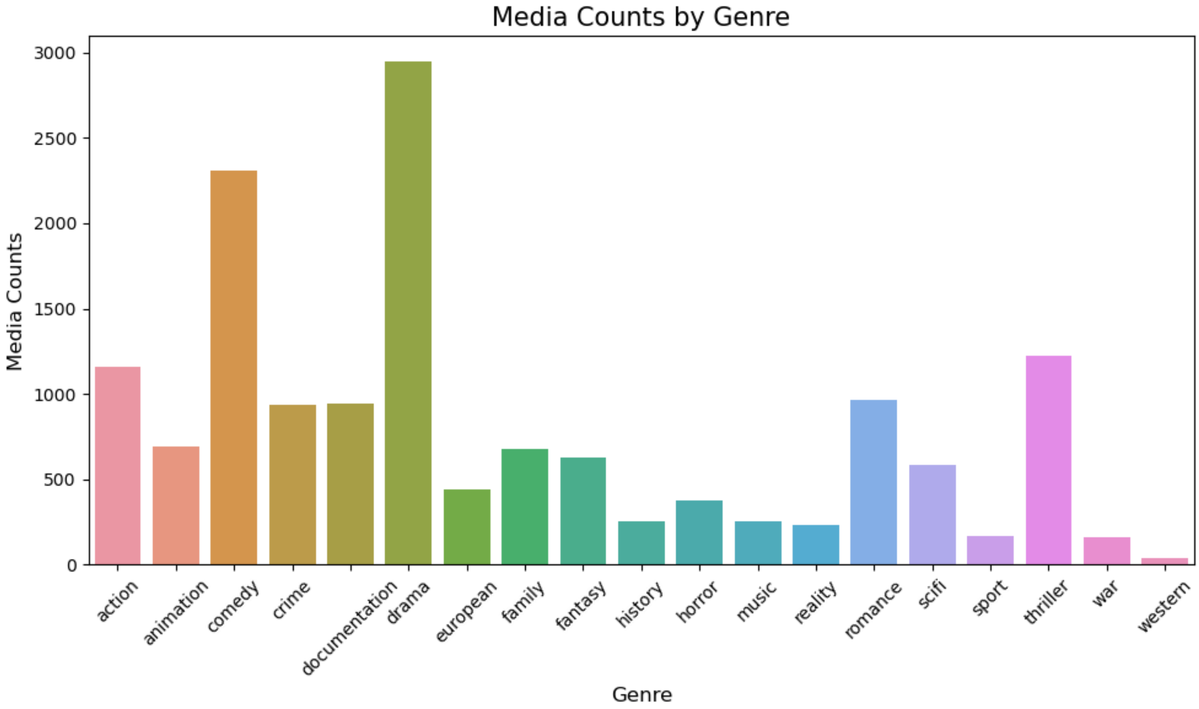

Figure 4: Bar chart showing the media count for each genre.

Results

IMDb Scores against Runtime and Genre

Our k-NN model which used runtime and genre to predict whether media would be perceived as ‘Good’ (IMDb score equal to or above 6.6) or ‘Poor’ (IMDb score below 6.6) performed moderately well. Its accuracy was on average ~67%. We found that this was stronger than similar models which were built using other categorical media features. We compare this to our Zero-R model which used the most frequent label in our training set to predict the label for all instances. The accuracy of this model was ~52%. Our confusion matrix (Figure 5) summarizes our average model performance over a 5-fold cross-validation process. Table 2 also includes a selection of corresponding performance metrics.

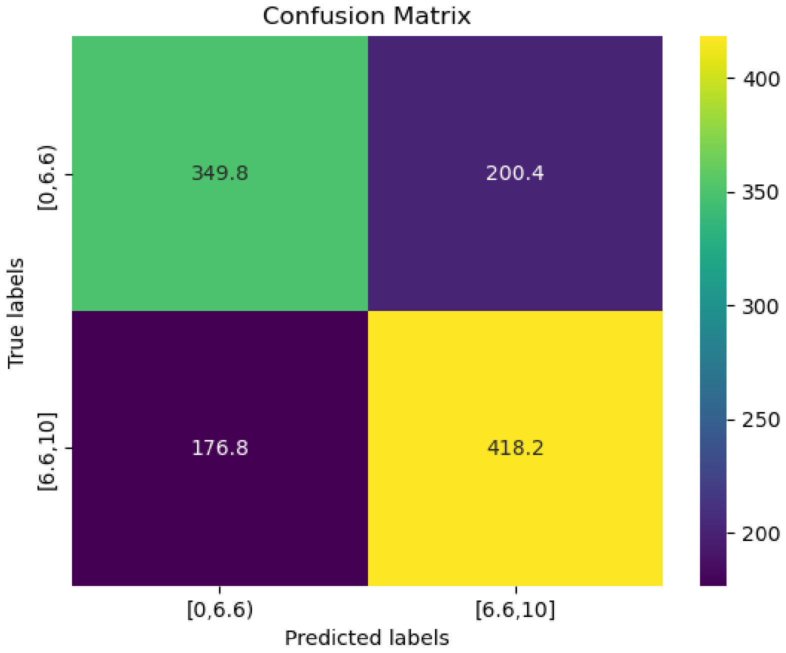

Figure 5: Confusion matrix with average values after 5-fold cross-validation.

Performance Metric	Average Score
Accuracy	0.6706
Precision	0.6646
Recall	0.6360
F1 Score	0.6496

Table 2: Performance metrics for our classification algorithm after 5-fold cross-validation.

Here, we have defined a positive’ selection to be media that receives a ‘Poor’ rating, as producers and investors are likely to be risk averse. They would be particularly interested in avoiding making poorly reviewed media as this would be more likely to result in financial losses.

Based on our precision score, we see that our model correctly predicts when media will perform poorly 66.5% of the time. From our recall score, we conclude that our model can find 63.6% of all poorly performing media in the sample.

We also compared the performance of our model when trained on subgroups of our dataset, specifically movies and TV shows, and obtained the results in Table 3.

Performance Metric	Average Score for Movies	Average Score for TV Shows
Accuracy	0.6589	0.7223
Precision	0.6860	0.5607
Recall	0.7720	0.2068
F1 Score	0.7263	0.2996

Table 3: Performance metrics for our classification algorithm after 5-fold cross-validation on subgroups.

We note mixed results here, likely due to the decreased amount of data available for training in each instance.

IMDb Scores against Choice of Director

From our regression model assessing whether directors with previously successful media tend to continue to produce successful media over time, we noted a moderative positive linear relationship. For a set of parameter estimates, we obtained a slope estimate of 0.5678 and an intercept estimate of 2.5329.

Conducting 5-fold cross-validation on this model, we received an average coefficient of determination of 0.23 and an average mean squared error of 0.86.

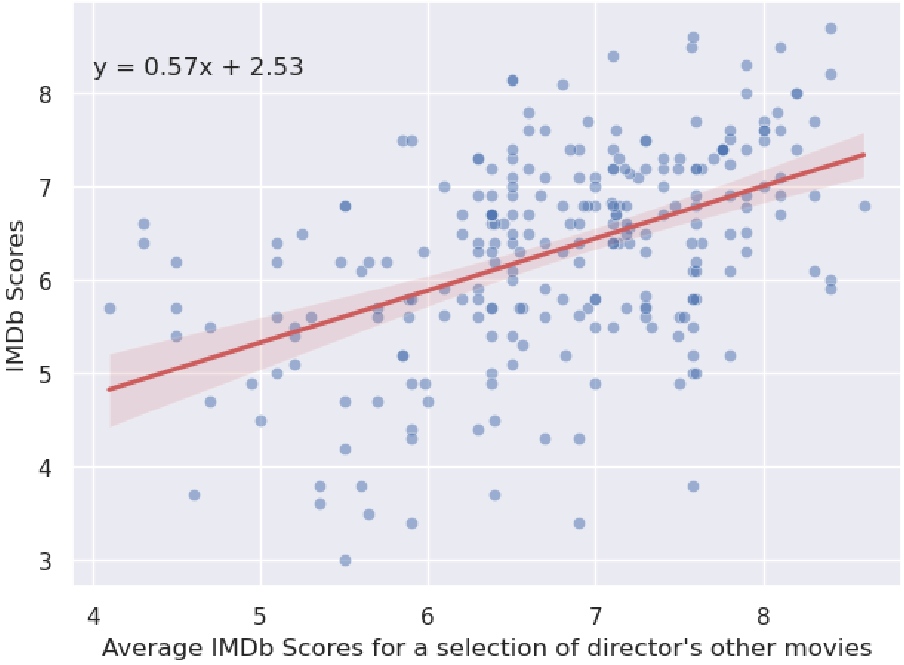

Figure 6: Scatterplot of test data with regression line obtained from first model.

From our regression model with a random allocation, we also noted a moderative positive linear relationship. We obtained a slope estimate of 0.5126 and an intercept estimate of 3.1478.

Conducting 5-fold cross-validation on this model, we received an average coefficient of determination of 0.22 and an average mean squared error of 0.91.

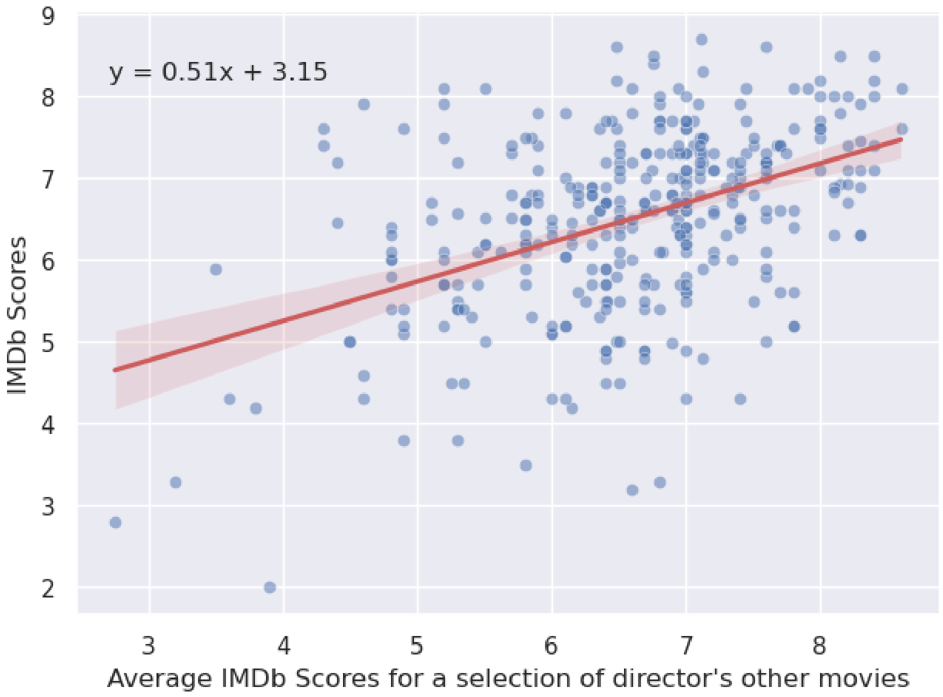

Figure 7: Scatterplot of test data with regression line obtained from second model.

Findings and Interpretation

IMDb Scores against Runtime and Genre

Due to our decision to use a k-NN algorithm, our model has limited direct interpretability. However, we do note that our model demonstrates that runtime and genre are both useful factors in predicting whether media will be perceived as ‘Good’ (above the median) or ‘Poor’ (below the median). Although we cannot conclude that this relationship is causal, it is still useful for media producers who wish to have some ability to predict the success of their media.

IMDb Scores against Choice of Director

From our previously stated results for our model which splits our data before 2017 and after 2017, we can say that a 1-point increase in the average IMDb score of a director’s medias previous medias is associated with a 0.4395-point increase in the expected IMDb score of a director’s media after 2017. Our model which randomly splits our dataset offers less direct interpretability here, however we can similarly see that directors who have been involved in other successful media are likely to continue to produce successful media.

From our coefficient of determination for our dataset split by time, we can say that 23% of the variation in IMDb scores of media released post-2017 can be explained by the average IMDb score of directors’ media released pre-2017. We note that this suggests that, at most, 23% of a media’s success is tied to the director leading a project.

These findings confirm the idea that a media's success does seem to be associated with the director leading the project. However, further analysis should be conducted to determine if this is truly a causal relationship.

Limitations and improvement opportunities

IMDb Scores against Runtime and Genre

During feature selection, the following genres were consistently omitted: 'fantasy', 'western', 'music', 'European', 'sci-fi', 'reality'. This likely means that our model has minimal predictive power for media of these genres. A dataset which contains a higher proportion of media in these genres would produce a model with greater utility for these media.

IMDb Scores against Choice of Director

When constructing the dataset for our regression analysis of IMDb scores against the average scores of a director's other medias, we note that some survivorship bias is likely to be present in the data. That is, directors who have made very poorly received media are less likely to be hired for a future media, and so our dataset is more likely to exclude poor performing directors, biasing the IMDb score distribution upwards. This effect would be particularly strong in our first model.

We note that because of this bias, our model is less likely to generalize well when predicting the IMDb rating of a show whose director has previously directed a small number of poor performing shows. However, it is still likely to hold predictive power in other cases.

Our choice of regression model here means that this bias cannot be eliminated completely, regardless of the size of the dataset. However, a future model may choose to explicitly limit the analysis to contain only directors who have directed more than a certain threshold number of media.

Another limitation we face in our model is a limited ability to test for relevant variables which may confound our relationship between the director and the IMDb score. An example of a confounding variable here could be the media budget. Media which have access to greater financial resources would be more likely to hire a notable director, however, more funding would also be available to hire skilled writers and producers, and to fund marketing campaigns, which are also likely to increase the success of a media. This means that our simple linear regression which analyses the relationship between a director and the success of their media may be overstating the positive effect which the choice of director has on media's IMDb rating. To overcome this limitation, we require data on these potentially confounding variables.

Other potential limitations

Although IMDb score is an excellent measure of the perceived success of a film or TV show, a high IMDb score does not necessarily imply that the media was a financial success. For media investors, an analysis focused on the media's return on investment would have greater utility for decision-making.

Conclusion

Our report highlights that genre, runtime and choice of director are all significant factors which can influence the perceived success of a movie. Although we noted several limitations in our analysis, we do believe that our results provide value for media producers and investors.

We have provided evidence to support the common idea that the success of media is often tied to the choice of director. As such, to increase the likelihood of media's perceived success, a director with a strong record of accomplishment should be chosen.

Similarly, genre and runtime are also potentially significant factors which can influence the perceived success of media. To increase the likelihood of media's perceived success, investors and producers should consider focusing on projects which have a similar runtime and genre choice to other previously well received media.

References

IMDb. (n.d.). IMDb Ratings FAQ. https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV

TMDb. (n.d.). General TMDb FAQ. https://www.themoviedb.org/faq/general?language=en-AU