# Prediction Factors Analysis of Movie Rating, Popularity and Revenue

Olha Prots
*Computer Science*
*at Applied Sciences Faculty*
*Ukrainian Catholic University*
*Lviv, Ukraine*
*Email: prots@ucu.edu.ua*

*Abstract*—**The movie industry is very competitive and profitable. Every year, hundreds of movies get released and tens of billions of dollars get earned by the industries best minds. It would be beneficial for the studios, the screenwriters, the directors and movie theaters owners to know what patterns are present in movie success. Knowing beforehand what makes people like the movie, or what will increase the chance of a high revenue would not only save money, but tell us more about the moviegoers‘ habits and preferences. The most efficient predictors of success are late predictors such as revenue or budget, but we can analyse factors such as genres, number of tropes, run time, cast. Here we try to see the dependencies between features and find whether it is possible to achieve predicting power from early movie data. We try to predict popularity, revenue and the average rating of the films using simple Logistic Regression. Code can be found on my Github, links to data sets in the end.**

## 1. Introduction

A data set was created from an online database of movie tropes [4], or, as The Art Direction Handbook for Film defines it, ”a universally identified image imbued with several layers of contextual meaning creating a new visual metaphor”. We decided to analyse whether any separate tropes or the number of them present in the movie would have any correlation with the success or the rating. Along with data about tropes, we used common predictors taken from The Movies Dataset on Kaggle [5] that has general information like run time, release date, budget and more. Using these features together in a simple model, we can see if there are any valuable and strong predictors that do not require script analysis or late information such as reviews, rating and revenue.

## 2. Problem Definition

We analysed which of the features were of great importance in the model, which ones correlated with the target values and compared two models: a model with many separate popular tropes present as features and a simpler, more understandable and compact model with mostly nominal features. We tried to predict all three features of success: rating, revenue and popularity.

### 2.1. Related Work

Research had been done in this industry, predicting success based on early, hype and late features. Early features mostly include plot summary [3], user generated keywords and script, as well a combination of genres, writer and actor names [1]. These demanded an NLP approach and half of the times did not bring great results. Deep learning methods in [3] used with this information do not outperform simpler methods, though have noticeable results. [3] used sentiment analysis in 1D CNN and LSTM models to predict the level of success, predicting 68% of successful movies and 70% of unsuccessful. Although the use of a Decision Tree that takes as input Box Office Gross, Budget, Year, Genres, cast and crew appeared to have potential in predicting the rating, giving around 14% error rate [1]. Another study [2] achieved around 90% accuracy by analysing actor power and using ROI as a profit predictor instead of gross revenue.

## 3. Proposed methods

### 3.1. Data Collection and Preparation

We merged the PicTropes Dataset [4] with the Movies Dataset using movie titles, stripped of most punctuation symbols and lower cased. In the tropes data set, new variables were created from the list of tropes for each film - number of tropes, number of popular tropes (those present in more than 30 movies) and number of rare tropes (the rest). Also we took as features the most used tropes as a base for one of the models. Changing the cutoff for popular tropes to higher appeared to decrease the accuracy. From the other data set, we took the genres and made dummy variables for all of them. We also took these features: -run time -date of release -revenue -popularity -vote count (number of votes that were input by the users) -vote average -belonging to a collection of movies -production countries -production companies There were missing values present in columns budget and revenue, and both leaving them as null them and imputing them with the mean of the column gave the same result. It makes sense that imputation doesn't help here as the mean would be a bad indicator of the budget when the films can be independent or made by a huge company. Usually, the missing values were present in the movies that had little popularity, but it was not always the case. We decided to go with the mean imputation to prevent strong bias. With mean imputation, the test score for predicting also increased by 1%. We tried using the profit indicator created from budget and revenue called ROI (Return on Investment), as used in paper [2], but it decreased the accuracy of our models.

### 3.2. Feature Importance and Correlation

When we applied the model that consisted of popular tropes such as Shout Out or Karma Houdini as well as nominal variables, it showed great performance on guessing revenue and popularity, but only showed 0.4 F-Score on rating prediction. When we looked into the Feature Importance, it showed that no trope feature had importance bigger than the nominal features. Although, the numbers of tropes and popular tropes had quite a big significance. As we see in figure 2, the correlation between the number of tropes is high so we excluded the overall number of tropes from the next model. The correlation between number of tropes and revenue was 0.42, which is quite significant.

High importance of vote count and budget in popularity tells us that it might be hard to get the information on popularity in early stages of the development. We also see that vote average does not seem to have a lot of value to popularity, since it is more about people going to see the movie and talking about it, not necessarily liking it. We see tropes being just as important or even more than some nominal variables in popularity and revenue, but in rating we see some prevalence of non-trope variables, especially genres.

It is interesting that the rating prediction features have one trope important - Bittersweet Ending. From the name we can see already that such a trope is hard to calculate with the use of nominal variables, but we have it! We also see here the variable countries points sum, which was created by adding the movie production countries‘ times of use overall. However, this feature does not show high correlation when the number of features is smaller, thus is not included in the correlation plot. In this kind of model, the importance fluctuated a lot, since they all are extremely small. That mostly changed in the smaller model.

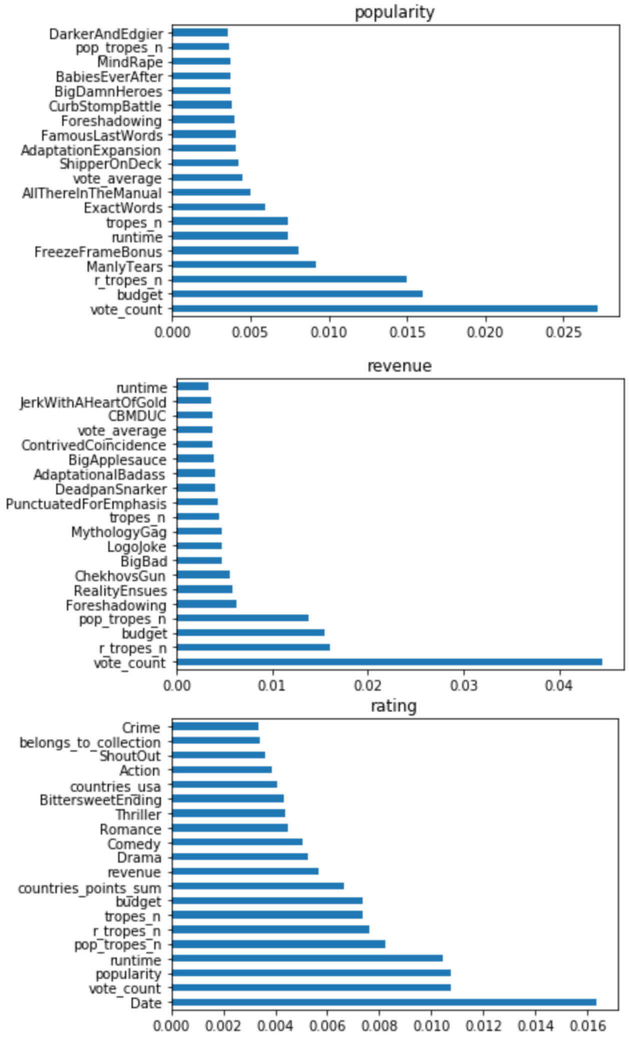

Figure 1. These are the most important features of a model with 1770 features

The next model we tried did not have the whole set of separate tropes included, but had the ones that proved to be significant such as Bittersweet Ending for rating or Manly Tears for popularity. We saw that the values of most genres did not have a lot of meaning to the model, just like the production companies dummy variables or vote average (for other targets than rating). We decided not to include vote count for popularity and revenue as it is quite understandable that a popular movie would get rated by many viewers, and that a movie that received revenue also garnered enough attention to be rated frequently. As for the rating, it is interesting that the relationship is positive between these two. The fact that few people voted on a movie does not mean that it will be biased to the positive side easily, and many votes do not decrease the average vote that much. Run time appeared to be a good indicator for rating (0.34 correlation) , which suggests that people might like longer than average movies more, since those usually tell the story in more detail, or in some other way add to the overall product such that it gives the user more satisfaction. Date was another great predictor, which is quite surprising, also considering that there is negative correlation to the target. Perhaps, newer movies have harder time satisfying the audiences, or they simply lack in quality more often than not. It was interesting as well that production companies usually did not have a high correlation with rating, although Marvel Studios had a 0.22 correlation with both budget and

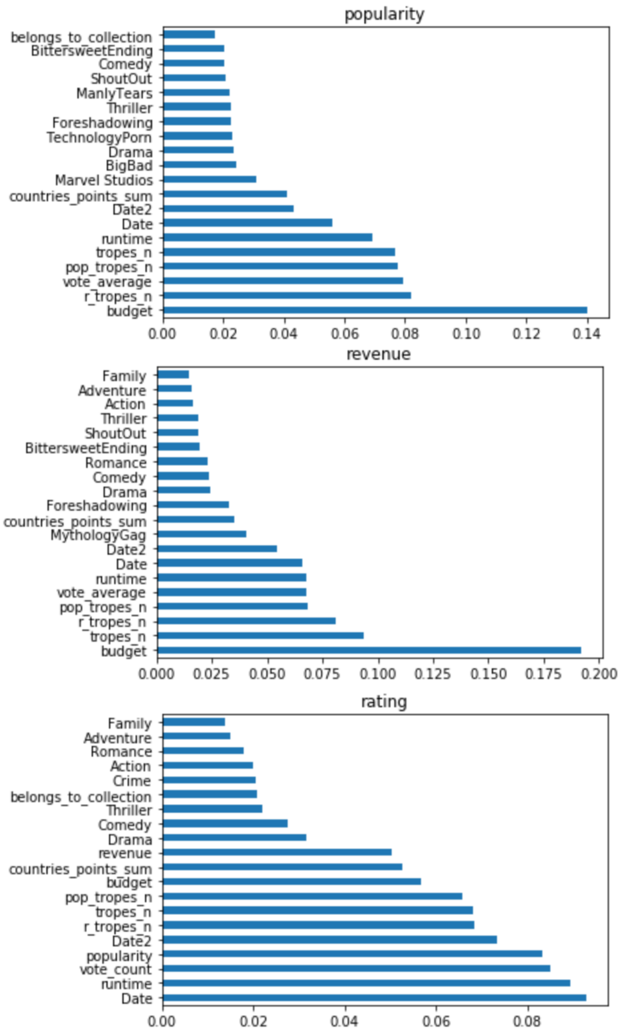

Figure 2. Most important features on reduced sized model

revenue, which did not come as a surprise.

### 3.3. Model

For predictions, we used Logistic Regression. Although the data is continuous, it seemed more efficient to try to group the movies and predict whether they get a rating of around seven or their popularity is, let's say, 14 out of 15. This approach proved to be great for ratings, but for popularity or revenue most of the movies were in the one or two highest clusters. The more classes we separated them into, the better the diversity of them seemed and the worse was the accuracy. We decided to separate films into 25 classes by popularity, 15 by revenue and 10 by rating, and the data set appeared to be imbalanced in such case. Most of the movies appeared to be successful and have quite a revenue, but rating was quite balanced compared to other two. On a test size of 1613 movies out of 4887 we achieved 0.95 accuracy with popularity and budget, and 0.48 on rating prediction. The three models had some distinctions, depending on their top important features, but overall had the same structure. Looking at the figure 3, we see that the highest correlation is between revenue and budget, which means that rating did not have even one strong predictor. The closest to it was run time. Also, considering the relationship of budget and revenue, it would be beneficial in the future to create a more representative value for the monetary success than just the revenue. Also, we should try to account for inflation.

## 4. Future Work

Overall, there are a lot of ways we can improve this analysis. Firstly, we can try to analyse the structure of groups of tropes present in the films, however we would need to create a new data set, since the one used here appeared to be quite outdated as well as have wrong movie titles. Also, it is useful to consider that the movies that have more tropes are just more closely inspected and, maybe, more liked by the viewers, so this measure might not be as good for an early prediction. Also, it is created by manual work, which is not always accessible to a machine learning model. Also, one feature that we were not able to find in public data sets was MPAA rating, which can also be a valuable predictor, both for revenue and popularity.

As for other contextual features, we could try to analyse keywords, plot summaries and cast and crew members. These might prove to be stronger than tropes or genres. The model proved that simple regression might not achieve high results on such a task. Also, we saw that there exists correlation between such features as Drama and belonging to collection (negative), Comedy and Thriller (also negative), movie budget and a trope Comic Book Movies Don't Use Codenames (also correlated with Marvel Studios), that suggest links between features we could further explore.

Another approach we could take in further experience would be to take into account the clustering of the movies. They are greatly divided by genre, and genres often have particular cast and crew (such as directors) that work there, as well as whole companies (like Marvel). Although tropes are shared among genres oftentimes, there is still a chance ti find valuable information for a decision making process of a model by examining them .

## 5. Conclusion

We have shown that the data obtained from public data sets on movies has some prediction potential for budget and revenue, as well as a little potential for rating prediction. We achieved high accuracy predicting successful movies, however we need to better classify them in order to insure our results‘ accuracy better. As for rating, we managed to find certain patterns in data we used, particularly in run time and tropes, however, we are still far from a good model. We can clearly see that many features present here only work for a certain part of the movies, because of the nature of this industry. Thus, it might mean that certain clustering analysis can be of use. This research will see more progress since the movie industry does not stop and actually produces more and more each year, presenting us with more data and getting more people interested in a good predictor of success.

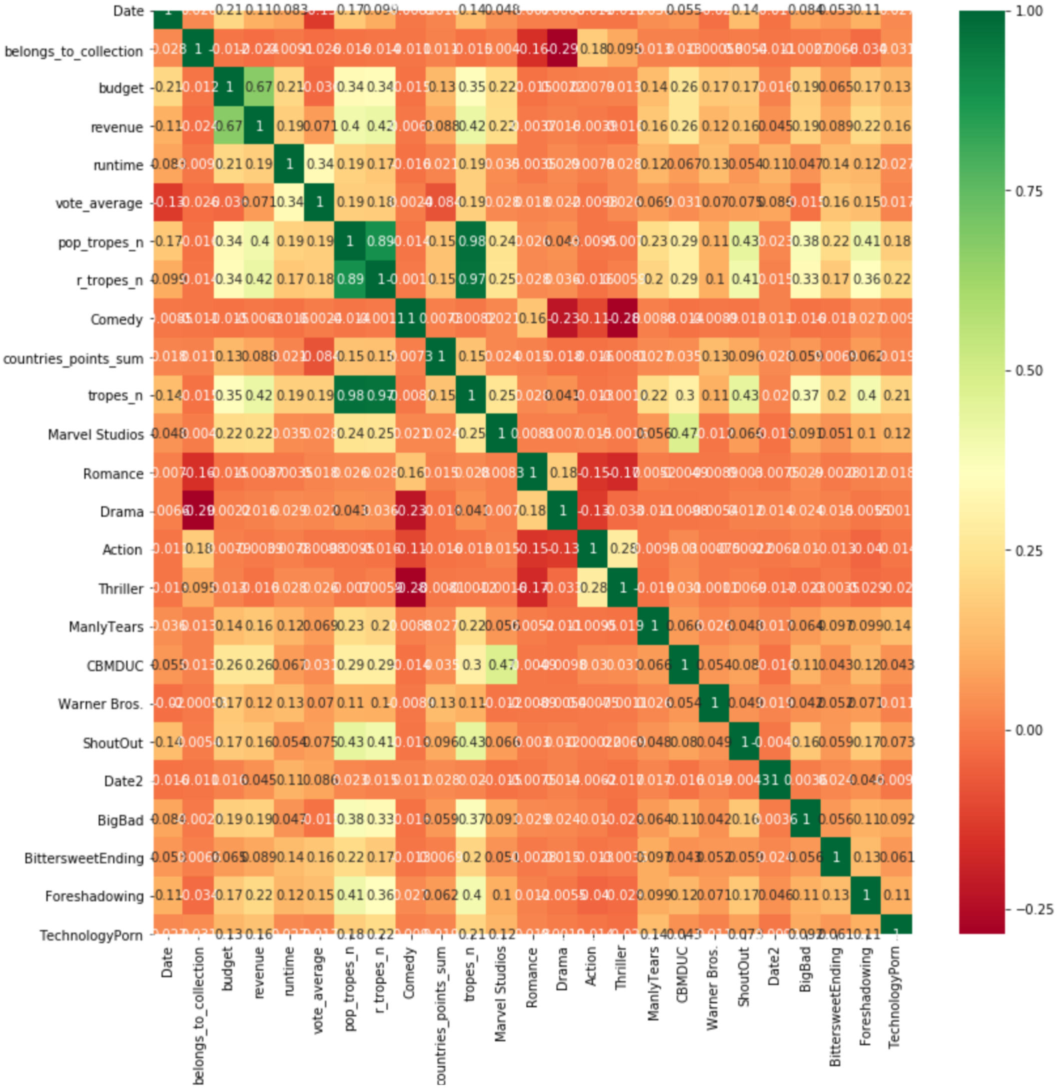

Figure 3. Correlation of features present in the smaller model

## Acknowledgments

The authors would like to thank the creators of PicTropes Dataset and The Movies Dataset.

## References

[1] N. Armstrong and K. Yoon, *Movie Rating Prediction*, Carnegie Mellon University

[2] M. Lash and K. Zhao, *Early Predictions of Movie Success: the Who,What, and When of Profitability*,24. 2016

[3] You-Jin Kim, Jung Hoon-Lee, Yun-Gyung Cheong, *Prediction of a Movie's Success From Plot Summaries Using Deep Learning Models* 2016

[4] R. Garcia-Ortega, J. Merelo Gustavus, P.Sanchez, G. Pitaru, Overview of PicTropes, a film trope dataset. 2018

[5] Rounak Banik. The Movies Dataset