Analyzing movies to predict their commercial viability for producers

Devendra Swami
University of Southern California
Los Angeles, California
dswami@usc.edu

Yash Phogat
University of Southern California
Los Angeles, California
phogat@usc.edu

Aadiraj Batlaw
University of California
Berkeley, California
batlaw33375@berkeley.edu

Ashwin Goyal
University of Southern California
Los Angeles, California
ashwingo@usc.edu

ABSTRACT

Upon film premiere, a major form of speculation concerns the relative success of the film. This relativity is in particular regards to the film's original budget – as many a time have "big-budget blockbusters" been met with exceptional success as met with abject failure. So how does one predict the success of an upcoming film? In this project, we explored a vast array of film data in an attempt to develop a model that could predict the expected return of an upcoming film. The approach to this development is as follows: First, we began with the MovieLens dataset [2] having common movie attributes along with genome tags per each film. Genome tags give insight into what particular characteristics of the film are most salient. We then included additional features regarding film content, cast/crew, audience perception, budget, and earnings from TMDB, IMDB, and Metacritic websites. Next, we performed EDA and engineered a wide range of new features capturing historical information for the available features. Thereafter, we used singular value decomposition (SVD) for dimensionality reduction of the high dimensional features (ex. genome tags). Finally, we built a Random Forest Classifier and performed hyper-parameter tuning to optimize for model accuracy. A future application of our model could be seen in the film industry, allowing production companies to better predict the expected return of their projects based on their envisioned outline for their production procedure, thereby allowing them to revise their plan in an attempt to achieve optimal returns.

1 INTRODUCTION

The film industry is one that stands out in every aspect. It is a world of it's own. For this project our motivation comes from the desire to provide a prediction model for producers to get an idea of the commercial viability of their movie. Joe Swanberg said in 2016, "The only way you're ever going to make any money is if you're investing in your own movies". Before a producer does this, they have to ensure that their investment is sound and understand how they see a return on that investment – this is where our model enters the movie. Our main contributions in solving this problem as part of this Datathon can be summarized as follows:

West Coast Regional, Citadel Securities, October 2020, CA, USA

- We investigated a wide range of features that are likely to be associated with commercial success of a movie. Different from other available studies, we have incorporated many novel features such as publicity, release date and movie cast & crew. We spent considerable time in feature engineering to better understand what factors make a movie financially lucrative.
- We extracted 11 different groups of features and built a random forest (RF) model to predict whether the return on investment (ROI) for the movie will be above or below median. After training RF, we have then identified the relative importance of each individual feature and groups of features.

We organize the remaining of the paper as follows. In Section 2, we describe our research methodology. In 3, we present the experiment results, followed by section 4, where we discuss the relationship between few important features and ROI. In Section 5, we identify threats to validity and finally in section 6, we conclude the paper and outline future directions.

2 RESEARCH METHODOLOGY

2.1 Problem Statement

The aim in choosing this problem statement is to investigate the potential of using a machine learning algorithm and to understand the features that potentially affect the commercial success of a movie. While doing so, we want to uncover the learned relationship between these features and ROI. The problem statement can be formally defined as follows

Task: Movie Success Prediction: Given a movie M, predict whether M will be a commercially successful movie.
While performing this task we will aim to answer the following research questions.

- **RQ1**: How effective is the random forest algorithm at predicting whether a movie will be a commercial success in terms of ROI?
- **RQ2**: Which individual features and groups of features are most important in predicting expected return of movies?

2.2 Features

In this section, we present the features we considered in this study. The choice of the features we selected are combination of what

features are generally used for performing analysis of movies industry, and other novel features that are publicly available and can be extracted easily using accessible tools.

Table 1 shows the summary of features considered in this study. We categorized our selected features into 11 groups based on the characteristics of a feature within the group. We believe that the short descriptions in Table 1 are explanatory enough for all features, but we will still provide a brief discussion for each feature group and include any features that may deserve a longer explanation.

2.2.1 ***Content.*** In this group, we consider *is_adult*, *is_english*, *languages_count*, *runtime*, *genome* and *genre*. The content of the film is important as film watchers typically have a distinct preference for these categories (ex. English speakers typically prefer movies in English), and therefore film content allows us to narrow down on the target audience.

2.2.2 ***Publicity.*** Several features were extracted to measure publicity efforts, like *is_collection*, *is_homepage*, *is_tagline* and *keywords_count*. Publicity is an important factor, as the marketing and accessibility of the film has a direct impact on the number of people that hear about the movie, and, by extension, the number of people that watch the movie (today in particular, films part of series tend to out perform other movies on the box office).

2.2.3 ***Audience Perception.*** In this group, we examined the features including *popularity*, *vote_average*, *vote_count*, *metacritic_score*, *imdb_rating*, *imdb_votes*. Naturally, audience perception has a huge effect on the success of movies. Films with high ratings tend to outperform movies with lower ratings. However, since this information is usually unavailable until after the film is released, we did not include this in our prediction model.

2.2.4 ***Release Date.*** In this group, we examined features like *release_month*, *movies_per_month*, *budget_fraction*, *movie_expense_score*. Several studies have indicated that the competition a film meets at the time of release has a major impact on the financial success of the film. This feature group strives to examine this effect.

2.2.5 ***Finance.*** This group comprises solely of the feature, *time_discounted_budget*, which is the discounted value of the budget used to produce the movie. Typically films with larger budgets tend to outperform lower budgeted movies in terms of revenue.

2.2.6 ***Production House.*** This group only includes the feature *production_house_embedding*. This is calculated from average performance of recent movies produced by the same production houses. Better, more established production houses are usually able to hire more renowned directors, writers, and stars, which, in addition to working with larger budgets, are factors that typically lead to box office success.

2.2.7 ***Writers.*** This group comprises solely of the feature *writers_embedding*, which was calculated from average performance of recent movies written by the same writers. Typically, writers who have written for successful films tend to write more successful films.

2.2.8 ***Directors.*** This group comprises solely of the feature *directors_embedding*, which was calculated from average performance of recent movies made by the same directors. Typically, directors who have directed successful films tend to direct more successful films.

2.2.9 ***Producers.*** This group comprises solely of the feature *producers_embedding*, which was calculated from average performance of recent movies made by the same producers. Typically, producers who have produced successful films tend to produce more successful films.

2.2.10 ***Main Cast.*** This group comprises solely of the feature *main_cast_embedding*, which is calculated from average performance of recent movies where members from the main cast featured. Several studies have shown that the star(s) of the film have a major effect on its success.

2.2.11 ***Support Staff.*** This group comprises of the features *female_count*, *male_count*, and *crew_length*. Some studies indicate the gender and size of the cast of a film can have an effect on its success.

2.3 Data Collection

We obtained most of the data from the Kaggle dataset "The Movies Dataset",[3] which provided us with data on the cast, crew, and general metadata on a large subset of films. In a dataset provided by the Data Open competition community, we obtained the genome tags, which we merged with our metadata dataset. We obtained further features by making API calls to the TMDB (The Movie Database), as well as scraping Metacritic scores from "IMDB: All U.S. Released Movies: 1972-2016". Initially, our merged dataset included slightly over 13K rows as we only had genome information for these many movies. We then decided to focus on movies after the color revolution (i.e. after 1965), and got rid of rows that had their budget values or revenue values set to 0. (There was forethought to interpolate these errors, however, as budget and revenue were such an important factor in determining rate of return, we concluded any method of interpolation would largely bias our model). This brought our total number of rows to 5,426.

2.4 Machine Learning Algorithm

We first thought of using regression as our machine learning algorithm due to the fact that the ROI values are continuous. However, our predictions from regression were rather inaccurate in many cases. Moreover, rather than being interested in knowing exact values, movie producers might be more inclined to know whether their movie will potentially perform well or otherwise. Thus, we have decided to predict above or below median ROI *(median calculated on training data)* rather than precise ROI values. This way, we have converted our original hard-to-learn regression problem into an easy binary classification task.

We have deployed random forest (RF) algorithm to perform our classification task since it is one of the most successful non-linear machine learning algorithm. This ensemble algorithm consists of numerous decision trees where each tree is trained on a random sample of the trained data drawn with replacement (known as bootstrapping). Moreover at each level in the tree, it randomly selects subsets of features when splitting nodes. Finally, the decision is carried out by taking average of predictions from each decision

Table 1: Movie features likely to affect the commercial success of a movie

#	Group	Feature	Description
1	Content	is_adult	Whether the movie is only suitable for adults or not.
		is_english	Whether the primary language used in the movie is English or not.
		languages_count	The total number of languages used in the movie.
		movie_runtime	The total length of the movie, in minutes.
		genome_tags [6]	Features extracted from movie content.
		movie_genres	The different genres under which the movie can be placed into.
2	Publicity	is_collection	Whether the movie is part of a collection or series.
		is_homepage	Whether the movie has a homepage or not.
		is_tagline	Whether the movie has an associated tagline or not.
		keywords_count	Number of frequently used keywords that can be assigned to the movie.
3	Audience Perception†	popularity	The movie popularity score provided by TMDB.
		vote_average	The average approval score for the movie, based on total votes at TMDB.
		vote_count	The total number of votes provided for the movie at TMDB.
		metacritic_score	The average approval movie score at metacritic.
		imdb_rating	The average approval movie score at IMDB.
		imdb_votes	The total number of votes provided for the movie at IMDB.
4	Release Date	release_month	The month in which the movie is released.
		movies_per_month	The total number of movies released in the same month as the given movie.
		budget_fraction	The budget of the movie in proportion to the cumulative budget of all movies of that month.
		movie_expense_score	The budget of the movie in proportion to the average budget of all movies of that month.
5	Finance	time_discounted_budget	The time discounted value for budget used to produce movie.
6	Production House	production_house_embedding	Calculated from performance of recent movies produced by the same production houses.
7	Writers	writers_embedding	Calculated from performance of recent movies written by the same writers.
8	Directors	directors_embedding	Calculated from performance of recent movies directed by the same directors.
9	Producers	producers_embedding	Calculated from performance of recent movies produced by the same producers.
10	Main Cast	main_cast_embedding	Calculated from performance of recent movies where members from the main cast featured.
11	Support Staff	female_count	The total number of females in the movie cast.
		male_count	The total number of males in the movie cast.
		crew_length	The total number of people in the movie crew.

†: Not used in prediction model as that information is not available before movie release.

tree. As a result, the random forest algorithm is less prone to overfitting. The algorithm is also flexible as it can take numerical or categorical variables as input and does not require feature scaling. For these reasons, we chose the random forest algorithm for our study. We have used Scikit-learn library of Python to implement the RF algorithm. The dataset is split into approx 80% training and 20% test data with movies earlier than 2011 are part of train data and from 2011 and on-wards comprising test set.

Dimensionality reduction: Due to presence of high dimensional sparse features like genome tags and genre information (after one hot encoding) in the dataset, we have used singular value decomposition (SVD) to reduce their number of dimensions while keeping most of the variance in data intact. In addition to that, we have also removed highly correlated features from our dataset. For removing correlated features, we have identified highly correlated pair of features, i.e. pairs having absolute correlation value of 0.75 or higher from Spearman's correlation. Among these identified pairs, the ones with a lower mutual information [4] are subsequently removed. It is done to remove features that do not provide any additional information and can help in reducing size of the dataset to speed up our training process.

Hyperparameters optimization: For obtaining optimal hyperparameters for our RF model, we have started by preparing a grid search space. The range of values used for different hyperparameters to create this grid search space are as follows:

- **n_estimators**: 100 to 1000 with step size of 100.
- **max_features**: integral multiple of $\sqrt{\#features}$ to #features.
- **max_depth**: 10 to 100 with step size of 10.
- **min_samples_split**: 0.01, 0.03, and 0.05.
- **min_samples_leaf**: 1, 3, and 5.
- **bootstrap**: always set to *True*.

As we can see that the above set of values leads to a grid search space of 12600 ($10 \cdot 14 \cdot 10 \cdot 3 \cdot 3$) and thus might be computationally expensive to search through all posssible permutations. Thus, we performed randomized search over uniformly drawn samples from it, using *RandomizedSearchCV* (from scikit-learn) to try out a wide spectrum of values from the above search space. We have used 100 iterations through 4–fold cross validation to identify the optimal hyperparameters. We report the hyperparameters obtained from this procedure in Table 2.

Table 2: Parameters of RF model tuned over validation set

Parameters	Optimal Value
n_estimators	500
max_features	14
max_depth	40
min_samples_split	0.05
min_samples_leaf	5

2.5 Model Evaluation

By virtue of the way we formulate our classification task, our dataset is perfectly balanced with two classes of equal importance. Therefore, *Accuracy* is initially thought to be a suitable evaluation metric. However, accuracy is sensitive to selection of threshold for converting predicted probability to the most likely output class. Different threshold can yield different accuracy score for the same ML predictions and thus makes it difficult to do performance comparison across multiple ML algorithms. In order to overcome this hurdle, we have selected the Area Under the Receiver Operating Characteristic (ROC) Curve to evaluate the performance on our task. We use the acronym *AUC* to denote this metric. The ROC curve shows how the true positive rate and false positive rate relationship changes as the threshold for identifying positives in the model changes. The area under this curve (AUC) indicates the performance of a model at separating classes. To evaluate the effectiveness, we compare the performance of the random forest algorithm described earlier with the random baseline approach. In **Random baseline**, we randomly assign a class to each movie in the test set.

Generally, if the performance of our model is too close to the baseline, it implies that the selected features have no association with commercial success of he movie or more features are needed. A perfect classification model would give an AUC value of 1.00, while a random model would give an AUC value of around 0.50. Generally, the higher the AUC value, the better the model.

2.6 Feature Importance Analysis

We have used permutation feature importance technique proposed by Breiman [1] to measure the importance of an individual feature and a group of features. The idea is that if a random permutation of a feature leads to a substantial decrease in the AUC value, then the model heavily relies on that particular feature for prediction. This decrease in value is referred as the **Importance Value (IV)** and a higher value indicates that the feature is more important. Similarly, the group importance is calculated by permuting all the features in the group together. It is important to not here that we can safely employ the permutation feature importance technique [5] since none of the features are highly correlated in our dataset as we have already removed highly correlated features.

3 RESULTS

3.1 RQ1: How effective is the random forest algorithm at predicting whether a movie will be a commercial success in terms of ROI?

Figure 1 shows the ROC curves of the baselines and the random forest algorithm . The AUC for the random forest algorithms is 0.78. Regarding the baseline, *Random* achieved expected AUC scores of around 0.500 due to its randomness.

Overall, the high average AUC score of 0.78 indicates that the models have a high explanatory power in predicting whether a movie will be commercially successful or not. This suggests that an application which applies feature engineering and machine learning would have the potential to help movie producers and analysts to estimate the expected ROI for upcoming movies.

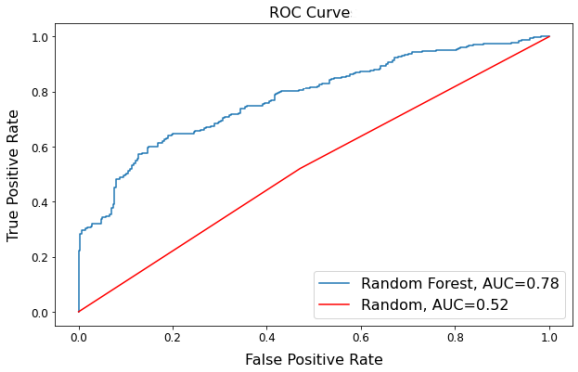

Figure 1: The ROC curve of the random forest algorithm that uses the selected features, and the baseline for our dataset.

3.2 RQ2: Which individual features and groups of features are most important in predicting expected return of movies?

Table 3 shows the top 15 most important features in predicting movies' ROI. We also observe that *is_collection* is among the most important feature. This makes intuitive sense, as sequels, spin-offs, and films extensions of a thematic universe tend to outperform movies without such characteristics. We also notice that the genome features (which as one might recall describes a film by its most salient characteristics) tend to offer valuable information. *key_word_count* offers similar information to the genome feature, and it also having a high information valuable implies that the characteristics of films are a key insight for determining the success of a film. The last noticeable observation in this table would be *movies_per_month*, which implies that competition upon a film's release date is indeed a major contributor in determining the success of a film.

Table 4 shows the top 5 most important feature groups. Here we observe that *Content* has the highest information value. This makes sense, as *genome_tags* is included as a feature in this category. With

Table 3: Importance of top 15 individual features

Feature Name	Information Value (IV)
is_collection	0.010
genome_0	0.005
keywords_count	0.004
genome_2	0.004
genome_13	0.004
movies_per_month	0.003
male_count	0.002
genome_3	0.002
genome_5	0.002
genome_12	0.002
is_homepage	0.001
time_discounted_budget	0.001
female_count	0.001
genome_1	0.001
genome_10	0.001

Table 4: Importance of top 5 feature groups

Feature Group	Information Value (IV)
Content	0.047
Main Cast	0.031
Publicity	0.021
Writers	0.017
Production House	0.016

Main Cast coming in second in IV, we reaffirm our hypothesis that the stars of a film have a major impact on its success.

4 DISCUSSION

Rather than solely rely on feature importance values discussed in previous section, we also want to investigate the causality and determine the direction of influence of important features on ROI values. Thus, in this section, we specifically explore the relationship between the key features and return. However, we note that this is a uni-variate analysis which has the shortcoming of not considering multivariate interactions.

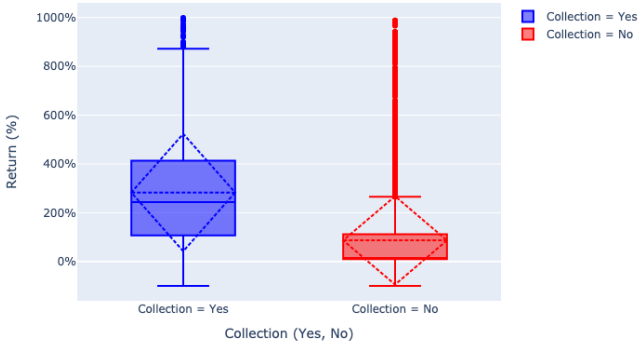

Figure 2: Plot between Is_collection and Return

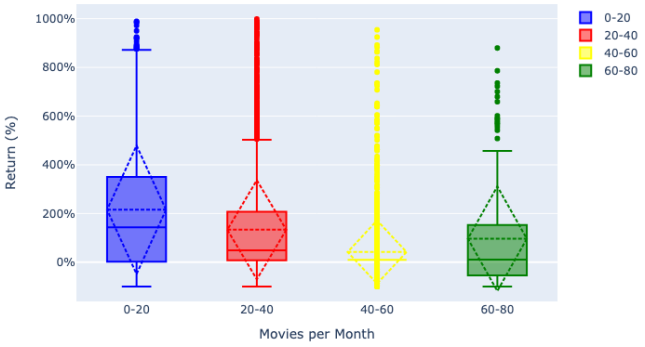

Figure 3: Plot between Movies per month and Return

Figure 2 shows that there appears to be a reasonable difference in the return on investment of films that are a part of a collection and those that are not. Movies that are part of a collection/series tend to offer higher ROI than those not part of any collection.

Figure 3 shows that there appears to be a fairly reasonable difference between the return on investment for movies released with different ranges of number of other movies released in that month. It appears that for the most part, the fewer movies released in the month of the released date of the film of interest, the higher the rate of return. It is important to note that there are large outliers for all these categories, signalling that some films across categories performed exceptionally well regardless of the number of films released in the same month of the film's release date.

Figure 4 explores if the existence of a homepage for a particular film is associated with a high return on investment. From this figure, it appears there is an association, but it does not seem to be particularly significant for the univariate case as we see there confidence intervals over lap.

Figure 5 compares the return on investment for different categories of budget. From here we see that there films produced with larger budgets tend to have a higher return on investment.

5 THREATS TO VALIDITY

In this section, we have identified several threats to the validity and limitations of our study.

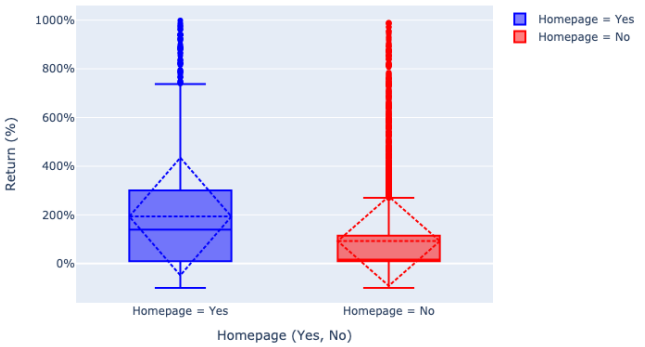

Figure 4: Plot between Is_homepage and Return

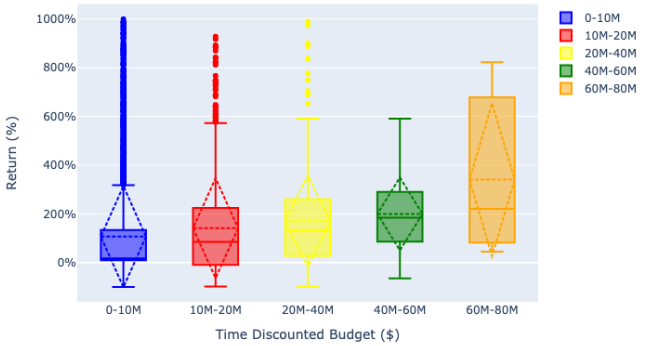

Figure 5: Plot between Time Discounted Budget and Return

Feature Selection Bias: Our results and conclusions were affected by the features we have selected and considered, which is limited by the human imagination. This means that others employing a different set of features and categorizing them differently may arrive at different research outcomes. Some features that we have considered for our prediction but could not incorporate yet include but are not limited to for example if the movie was nominated for an academy award, how well connected was the production team with the world? Yes, we have included popularity features but the broader question remains, how was the budget managed? How much was allocated to marketing and campaigning?. Addressing these and including more features is part of our plans for the future.

Tool and Method Reliability: While the tools we employed to are proved to be effective in literature, it is still possible that inaccuracies exist in them that we did not take into account. For example, we assumed that genome tags are perfectly computed and provide a better representation of the movie content. However, being derived from ML algorithm, these tags can prove to be inaccurate in some cases.

External validity: We acknowledge several limitations that may influence the generalizability of our findings. First, the movie sampling problem; to elaborate, we recognized that we don't have data to validate our results and hence will have to perform train/test split, this introduces yet another problem of bias, we realised that times have changed since early 1920's and hence we would need to consider only a limited time-frame before the targeted period, this led to reduction in data points and hence reduction in validity checks that we could perform. Only one machine learning algorithm used, namely Random Forest. We considered using neural networks but due a paucity of data points our model would have been flawed.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we sought to answer the question how might production companies predict future return of upcoming cinematic projects, attempting to help filmmakers better understand the commercial viability of their undertakings. To achieve this, we extracted features in the categories of Content, Publicity, Release Date, Finance, Production House, Writers, Directors, Producers, Main Cast, and Support Staff. Finally, we built a random forest classification model, which, after thorough hyperparameter tuning, was able to successfully predict the return of the films with an AUC score of 78%.

Future Work: The top priority is adding more features and expanding our data-set to include more data-points for available features. This includes web-scrapping social media websites to find hidden connections and deep networks that were previously unknown, for example, constructing a social media connectivity graph to better identify popularity and inner circles of the production team which in-turn provides better predictions for our core problems. (Social circles identify how powerful a single member's presence is - better the presence, more the return). Once we have done that, we would like to concentrate our efforts in employing other learning algorithms to obtain new models that would compare or ensemble with our current model. One such algorithm can be neural networks as we may discover additional real world relationships that are currently unknown. This, of course is possible once we have expanded our data-set sufficiently.

REFERENCES

[1] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.
[2] Harper, F. M., and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. 5*, 4 (Dec. 2015).
[3] Kaggle. The movies dataset. Online, 2017.
[4] Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical review E 69*, 6 (2004), 066138.
[5] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. Conditional variable importance for random forests. *BMC bioinformatics 9*, 1 (2008), 307.
[6] Vig, J., Sen, S., and Riedl, J. The tag genome: Encoding community knowledge to support novel interaction. *ACM Trans. Interact. Intell. Syst. 2*, 3 (Sept. 2012).