ABC Movie Production: Investment Optimization

Team member: Zhongjie Xu, Youai Qin, Abbas Khan, Yi Wang, Akanksha Mishra

Business Understanding

Business Problem: ABC is a production company currently looking to invest in new movies that have the potential for market success. Our metrics for movies' performance is based on their IMDB Popularity and Revenue. As a business, we will be looking to achieve maximum predicted revenue, but at the same time we also need to balance that decision based on the popularity prediction of that movie(IMDB score). For this purpose, we're going to analyze historical data of a list of 5000 movies understanding how the best revenue and popularity can be obtained in the movie business, bearing in mind that success is striking a good balance between these two metrics.

Proposed Data Mining Solution: Through analysis, we want to optimize ABC's investment decision. Our data mining analysis can help us determine which various variables potentially impacting the revenue of a company and how. These variables could be production company/companies we should collaborate with, which director we should collaborate with and which genres have the highest potential for the highest profitability. We will also predict IMDB scores, using a model, to determine the popularity of the movie. It will be important to see how IMDB scores and ROI are correlated in order to guide our business decisions while maintaining a good balance between these two metrics.

Data Understanding

Identify and describe the data (and data sources) that will support data mining to address the business problem. Include those aspects of the data that we routinely talk about in class and/or in the assignments. (If appropriate highlight potential bias, full disclosure is always better, and identify potential bias directions.)

● Target Variable: Revenue & IMDB Score

● Features: Genre/ Languages/ Production companies/ Production countries/ Release Month and Year / Runtime/ Vote average/ Vote count/ # of production companies/ Keyword

Data Preparation

1. Data Mining Method Selection: We narrowed down two data mining methods to meet our agenda.

- **Revenue Prediction -** We identified what factors influence revenue generated by a movie to be able to optimize these factors while deciding our investment. We thought linear regression will help us answer this question. Within linear regression, we choose influencing factors using techniques such as stepwise regression, lasso and post lasso. We later tested model accuracy by comparing OOS R-square.
- **Movie IMDB rating Classification:** In order to answer what makes a movie popular amongst people we built a classification model. For this we used various techniques ranging from - Logistic regression, Classification tree, Multinomial model. We compared both Lasso and Post Lasso version of these models with the Null Model to choose the most accurate one in terms of prediction.

2. Merging: Two data sets available to us i.e. Movie IMDB ratings and Movie cast details.

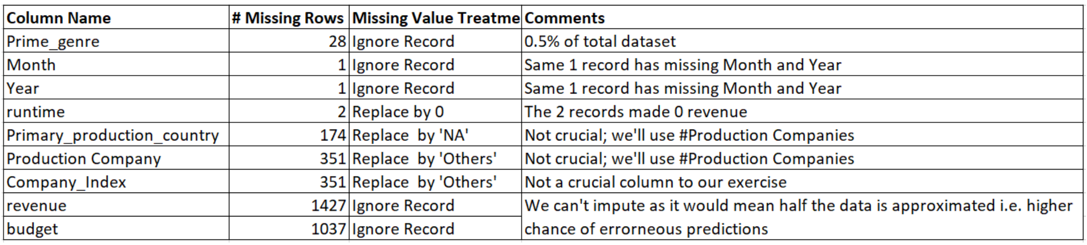
3. Data Cleaning: Transformed raw data, making it more suitable for the analysis:

- Descriptive *'Keywords'* column → created 'Number of keywords'
- Multiple *'Production_country'* columns → kept only the first & created another variable for the 'Number of countries'
- Multiple *'Production_company'* column → kept only the first & created another variable for the 'Number of companies'
- Segregated '*Date'* column → Kept only month and year value and dropped the date
- *Revenue* and *Budget* → Transformed to million dollars

4. Data Tailoring

- After careful consideration and testing, we dropped redundant columns from data, i.e. *Keyword_1, Keyword_2, Keyword_3, Secondary_genre, Third_genre, tagline, overview.*
- We also removed all movies before Year 2000 for a fair comparison of the value of money.

5. Missing Value Treatment- Identified missing values and decided missing value treatment on a case by case basis. Below is a summary of what we identified will be doing:



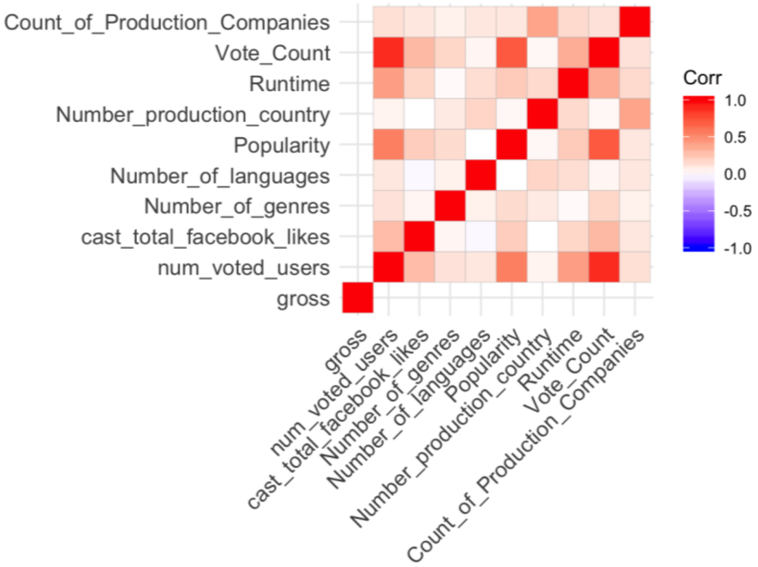
Column Name	# Missing Rows	Missing Value Treatme	Comments
Prime_genre	28	Ignore Record	0.5% of total dataset
Month	1	Ignore Record	Same 1 record has missing Month and Year
Year	1	Ignore Record	Same 1 record has missing Month and Year
runtime	2	Replace by 0	The 2 records made 0 revenue
Primary_production_country	174	Replace by 'NA'	Not crucial; we'll use #Production Companies
Production Company	351	Replace by 'Others'	Not crucial; we'll use #Production Companies
Company_Index	351	Replace by 'Others'	Not a crucial column to our exercise
revenue	1427	Ignore Record	We can't impute as it would mean half the data is approximated i.e. higher chance of errorneous predictions
budget	1037	Ignore Record	

We ended up with a descriptive dataset for 1841 movies and various analyses of relevant variables.

Exploratory Data Analysis

1. Correlation

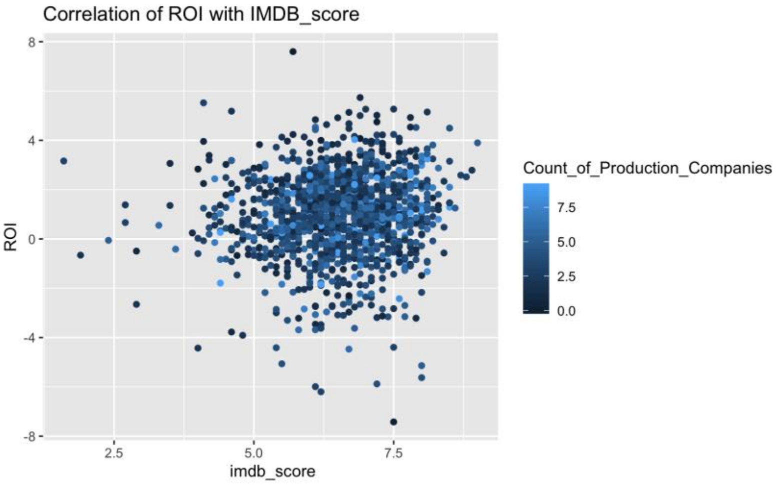
Popularity and Vote Count has a high correlation as expected. This is because Popularity is defined as a metric that uses Vote Count in its calculation. Thus, we need to use them exclusively in our analysis. We also see that month does not have a relevant influence on the Revenue and Popularity unlike expected.



2. ROI and IMDB score

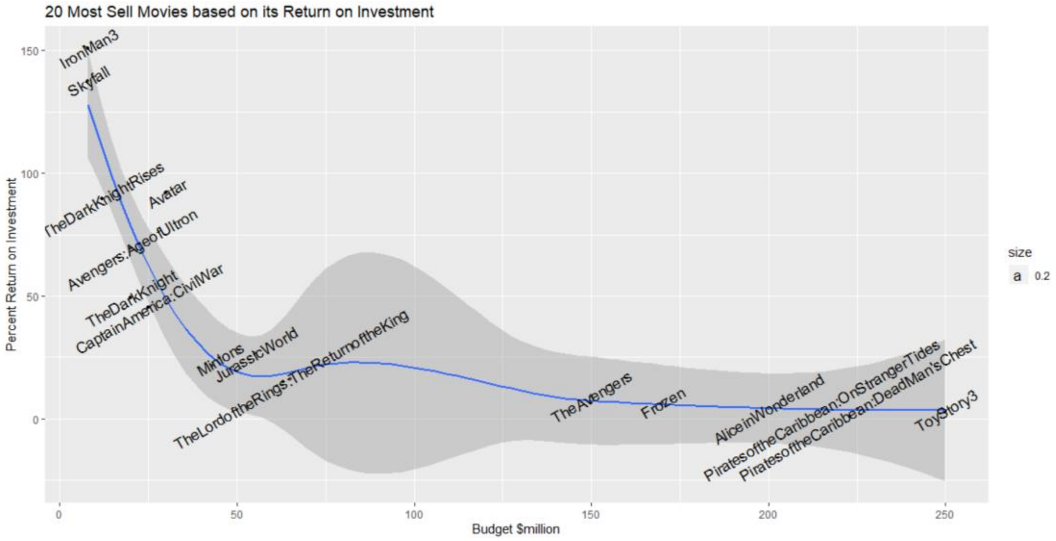
From the scatter plot, we can conclude that there is no direct linear correlation between the two ROI and IMDB score. However, we can find that many movies have imdb_score between 5 to 7.5

also have relatively high ROI.



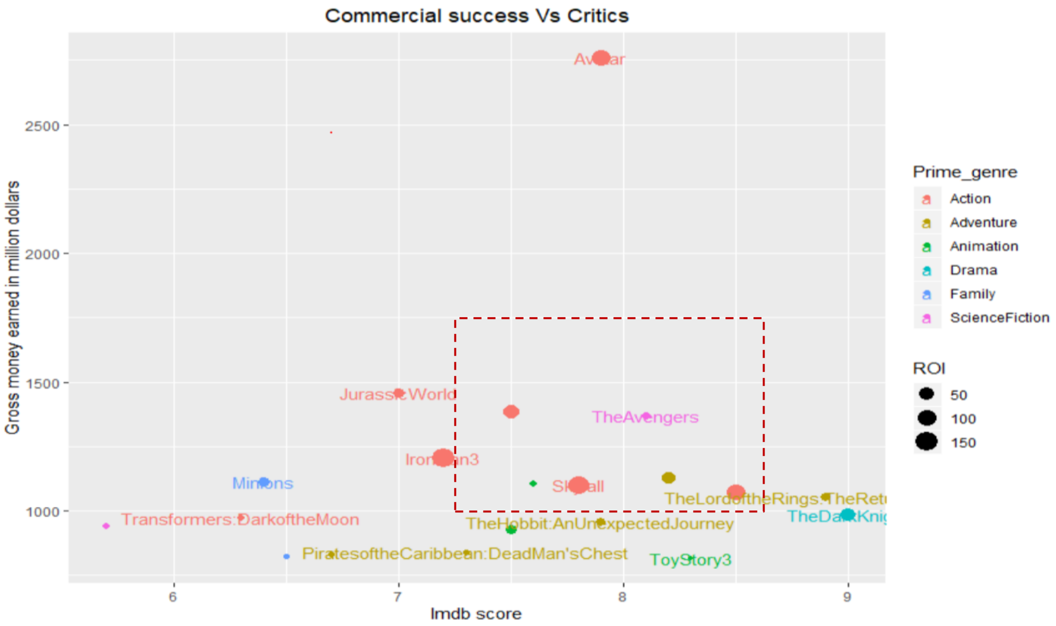
3. Fare between Budget and Revenue

We assumed that a high budget does not guarantee the high revenue i.e. it's not linear. In order to test the same we compared this in the top 20 movies and it confirmed our hypothesis to be true. It can be seen that Iron Man3 and Skyfall have the highest ROIs.

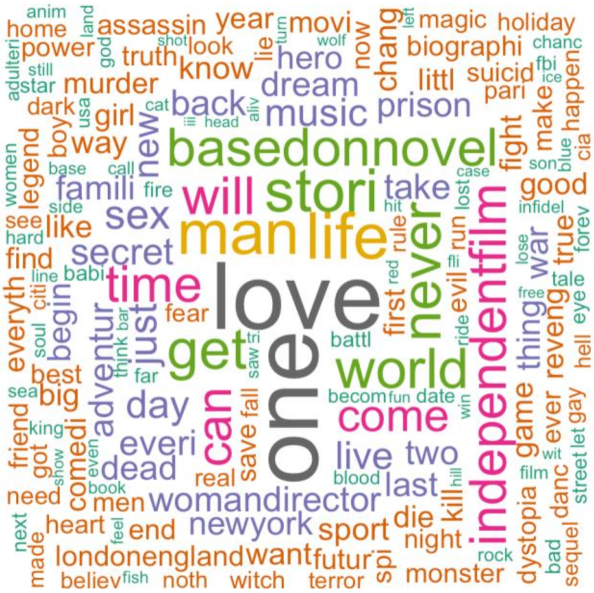


4. Critical v.s. Commercial

From the graph, we can tell that a critically acclaimed or popular movie can be bad commercially as well. We visualized this amongst the top 20 again. From this graph we identify that the investors should focus on the movies in the rectangle (seen in graph) because it represents a good balance between ROI and IMDB score.



6. Sentiment Analysis



Firstly, we extrapolate 3 keywords for each movie in the dataset. The size implies the frequency of a certain word. According to our word cloud, "love" tops the list among all the keywords, indicating the popularity. of romantic movies. Other common words are "man", "life", "time" which are closely related to everyday life.

Moreover, we find also "New York" and "London" are also common words in movies which suggests that many movies set their stages in metropolitan cities. However, there are other words like battle and fear indicating the various genres of movies.

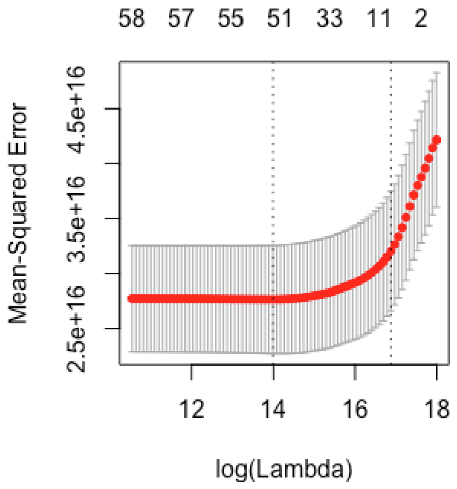
Secondly, IMDB_score is an internally generated popularity benchmark in the database. ROI is defined by (revenue- budget)/budget. From the graph, it can be concluded that most films with a high IMDB_score also has a relatively high ROI, which is usually the case as the more popular a film is, the larger revenue does it generate.

Modeling

1. Movie Revenue Prediction

What will the revenue be like? This is one of the things investors care most about. That being said, we want to predict the revenue for films we've been working on to estimate our profit. Hence, we developed three models i.e. Lasso with lambda. min, lasso with lambda.1se and Stepwise model with backward selection. By comparing the out of sample R squares and mean squared error (MSE), we will be able to pick model with the best performance.

Three Models, Two Metrics



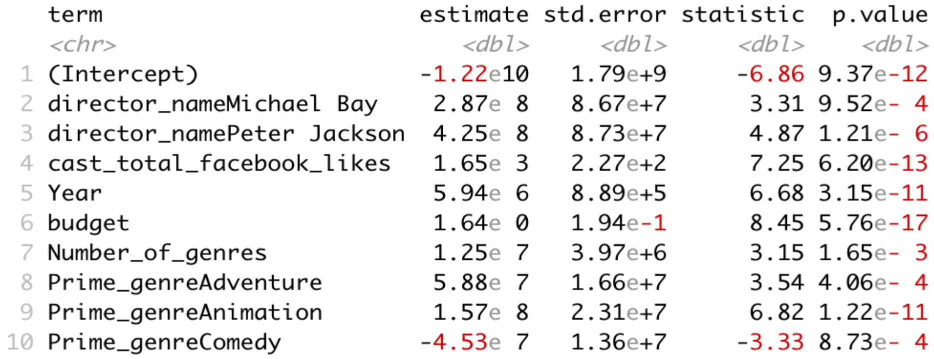
We split the dataset by using a 5-fold cross-validation dataset based on cleaned data (without any NA and dropping any possible variables that might cause high collinearity to revenue). Then, we run lasso on our train data and come up with lambda 1se (21,271,649) and lambda min (1,189,258) expressed as log in the plot indicated by the dotted line.

Final Model

We choose the stepwise model as the final model as it had the highest OOS R square and lowest MSE amongst the three. The significance level is 0.05.

From the model, we know that Michael Bay, Peter Jackson, total facebook likes for the cast, Year, budget, number of genres, adventure and animation are significantly increasing the revenue.

However, if a firm produces a comedy genre film, it's going to generate less revenue compared to other genres like adventure which tends to make more money. The cast total facebook likes variable is just shows how if you have cast that is more famous, the movie is likely to more money. Increasing the total genres of the movie also makes it more lucrative, which makes sense intuitively because it can appeal to different segments of the population.



```
   term                        estimate std.error statistic  p.value
   <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                 -1.22e10   1.79e+9     -6.86 9.37e-12
 2 director_nameMichael Bay     2.87e 8   8.67e+7      3.31 9.52e- 4
 3 director_namePeter Jackson   4.25e 8   8.73e+7      4.87 1.21e- 6
 4 cast_total_facebook_likes    1.65e 3   2.27e+2      7.25 6.20e-13
 5 Year                         5.94e 6   8.89e+5      6.68 3.15e-11
 6 budget                       1.64e 0   1.94e-1      8.45 5.76e-17
 7 Number_of_genres             1.25e 7   3.97e+6      3.15 1.65e- 3
 8 Prime_genreAdventure         5.88e 7   1.66e+7      3.54 4.06e- 4
 9 Prime_genreAnimation         1.57e 8   2.31e+7      6.82 1.22e-11
10 Prime_genreComedy           -4.53e 7   1.36e+7     -3.33 8.73e- 4
```

Pros, Cons and Alternatives

What's good about our model is that we compared across different models to arrive at the best model to go ahead with to drive our business decisions. We also trained our data through a k-5 fold validation so that we could make as much use of our dataset as we could.

However, a major issue with our data was a lot of missing and NA values that we ended up having to drop. We felt it was a better decision to drop those observations because intuitively it did not make sense to fill them with averages or other imputing techniques. For further exploration, we can definitely try the non-linear model to predict the revenue.

2. Movie IMDB Rating Classification

We want a model to determine IMDB Rating for a movie, given Investment details. For this we attempted several modeling techniques broadly under the Multinomial and Binomial Classification types. Below is a summary of the same:

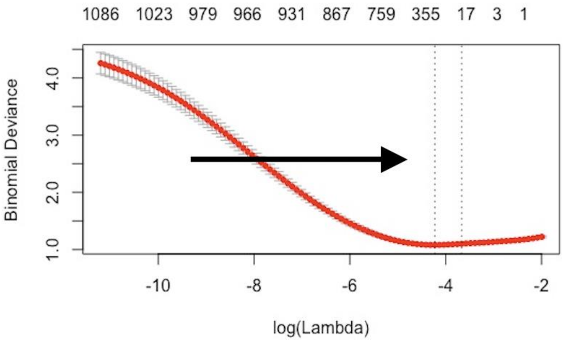
Multinomial Classification:

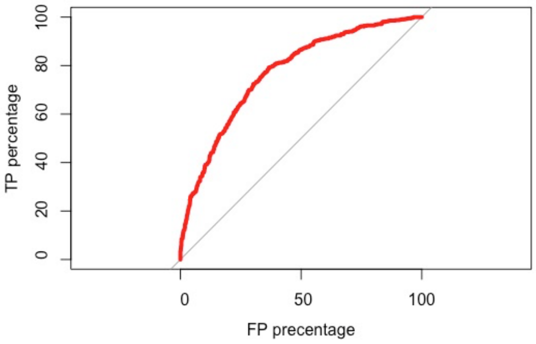
Translated IMDB score (out of 10) into 4 categories i.e. Bad (0 to 5), Acceptable(5 to 6), Good(6 to 7), Excellent(7 to10) and made multinomial classification model using Neural Networks and Tree.

Model Accuracy: 54% for Neural Networks and 53% in Tree. we moved to simpler options because we couldn't find a good baseline and evaluation tool for our model accuracy based on FPR_TPR.

Binomial Classification:

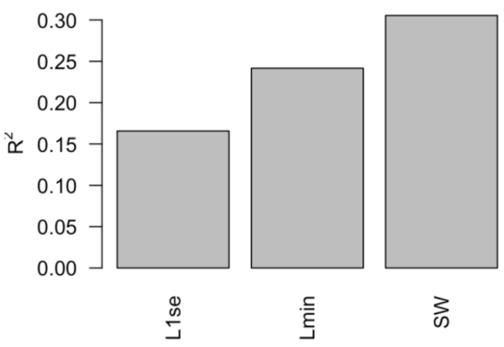
Moved on to Binomial classification, thus reducing the complexity of classification i.e. Bad (less than 6) and Good (more than 6). Within Binomial we compared 2 models both before and after Lasso. We compare five models - Null model, Classification tree, Logistic regression,Post-Lasso Logistic Regression , Post-Lasso Classification tree.

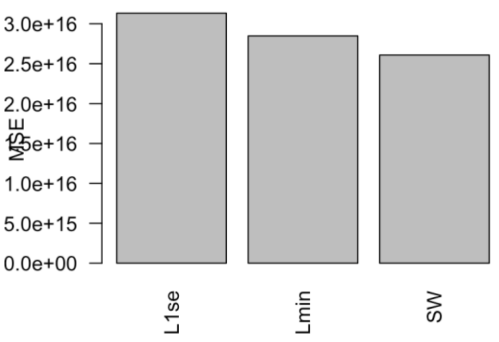




Evaluation

1. Movie Revenue Prediction - Evaluated the three models (Step-wise regression, Lasso with min lambda and Lasso with 1SE lambda) using, OOS R-Square and MSE. As can be seen from the plots below, step-wise again had the best performance of the three models with the smallest MSE value and the best R-square. The Lasso with 1 SE lambda performed the worst across both of the metrics.



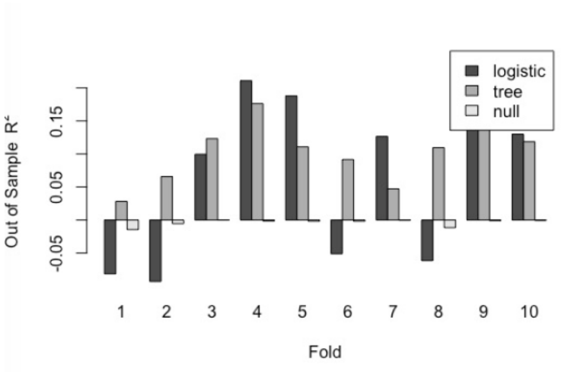


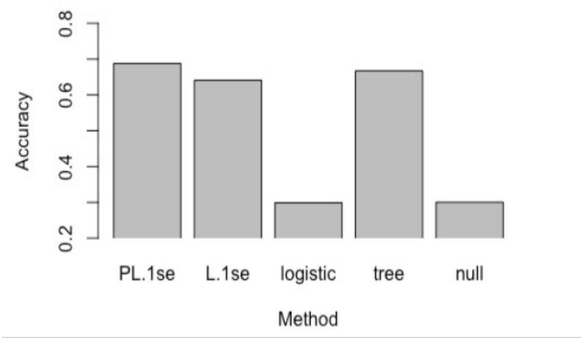
2. **Movie IMDB Rating Classification -** Evaluated using 10-fold Cross-Validation:

After using 10-fold CV, we found out using Post-Lasso to choose variables would help us on improving our model accuracies by a large amount as you can see from the graphs above.

We choose 0.3 as our threshold because we want a conservative strategy for now so that we'd rather suffer from False Negative than from False Positive.

Our final model used Post-Lasso to predict test outcomes (with 68.7% accuracy with 0.3 threshold)





Cons and Alternative: We couldn't get Post-Lasso with Lambda.min because the model took too long to run inside the CV loop. Therefore, we choose to use Lambda.1se for lasso and post-lasso.

Deployment

1. Movie Revenue Prediction

Based on our lasso modeling and out of sample tests, we as investors can predict the ROI of the movies so that better decisions can be made with the help of statistical modeling. Even though the models ' accuracy needs to be refined, it still gives us a hint of whether to invest in a movie or not.

2. Movie IMDB Rating Classification

According to our final model, we will use it to identify the "bad" movies-IMDB rated under 6.

As producers and investors, we are not only concerned on the commercial value it could bring to us but also, we should put certain weight on IMDB's rating-improve our reputation in the industry. Looking forward, we will adjust the weight of commercial value and reputation (IMDB ratings) accordingly to our investing strategy. Furthermore, we might even change our threshold from 0.3 to a lower one when we could afford to take risks.

Reference

Kaggle: https://www.kaggle.com/tmdb/tmdb-movie-metadata

IMDB: https://data.world/data-society/imdb-5000-movie-dataset