PROJECT BLOCKBUSTERS

IEOR 4523 PROJECT REPORT

Team Members:
eq2150
hh2821
ww2547
wx2226

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

CONTENTS

1. Introduction

Our Project aims to predict US Box Office Revenue. The innovation of our project is that we took a different perspective when building our prediction model by breaking down the problem into two parts in terms of factors that may contribute to the revenue of movie-making: internals and external factors. Internal factors being the attributes of the movie itself, such as the budget, the cast, and the director. External factors being the overall economy and release time. In this project, we built a classification model with an accuracy of 65% predicting which category of revenue will the movie be in.

2. Project Design

2.1 Problem Breakdown

When you think about movie revenue prediction, what are some attributes that you think might contribute to the success? Is if famous actors? Fame might bring more exposure and publicity. Is if the production company? Prestigious companies may have more resources to produce higher-quality movies. Is if the genre of the movie? Horror movies may have a lower budget thus increasing the return rate. Is it the economy? A good economy might have a positive impact on the total sales of movie tickets. Is it the release time? Low budget movies might have a better chance to do well in non-holiday seasons. To answer these, we first needed to collect the necessary datasets.

2.2 Data Collection

Our data collection mainly consists of two parts. The first part being data collection of movie internal factors (factors of the movie itself). We scrapped 19 different factors from the IMDB and TMDB website. Then we used two factors, budget, and USA Gross GDP, to calculate return rate. Finally, we got 21 internal factors. The second part is collecting data for factors that indicate the wellness of the economy and the release time. To do that, we gathered data for about 2000 different economy indicators from the Bloomberg Terminal and we narrowed down our collection to 14. The final internal and external factors are shown in the table below:

	Internal Factors	*External Factors*
Factors, Annotations & Data Type	* Title: title in English (string) * Runtime (integer) * Genre (list) * Homepage: url (string) * IMBD ID (string) * Rate: movie rate (integer) * Rating Count: amount of people rating movie on IMDB (integer) * Country: production country (list) * Language (list) * Release date (datetime) * Original Title: movie title in original language (string) * Director (list) * Writer (list) * Cast: if number of casts exceeds 30, we only keep top 30 casts (list) * Production Company (list) * Budget (integer) * Gross USA (integer) * Cumulative Worldwide Gross (integer) * Review (string) * Belongs to Collection: name of collection, i.e. Star War (string) * Return Rate: $\frac{Gross\ USA}{Budget}$ (float)	* GDP Deflator * US Real GDP 2 * US Real GDP: Gross Domestic Product * US Real GDP: Personal Consumption Expenditures * US Nominal GDP: Gross Domestic Product * US Nominal GDP: Personal Consumption Expenditures * House_Price_Index * China Personal Expense: All Households * China Personal Expense: Rural Households * China Personal Expense: Urban Households * Gini: Gini Coefficient * Gini: White Alone * World CPI 2 * World CPI 2

Table 2-1

Altogether there are 306534 movie instances, and economical data (external factors) ranging from 1930 to 2019. As shown in the above table, we have a mix of quantitative and qualitative factors, which we will do further cleaning below in the modeling section.

2.3 Data Cleaning

Glancing over the data for movie internal factors, the first thing we noticed is the missing data. We got rid of the missing data and narrowed down the data set from 300k rows to 7k. Further cleaning, we realized a few abnormal return rates (revenue/budget). We found out that for some of the foreign country movies, the currency was not correctly converted into USD. We discussed ways to handle this situation and decided to drop these as well since they consist of only about 10% of our dataset. For our economy data, we gathered more than 2k attributes. However, we realized that many of our attributes are subcategories of other attributes. For Consumer Goods CPI is a subcategory of US CPI etc. These attributes are strongly correlated and if we include them all, it might mess our model. We chose the main attribute that summarizes the other attributes. This way, we narrowed it down to 14 factors.

3. Business Analysis

Along with our data cleaning and model building process, we found a few interesting insights about our data.

3.3.1 Collection

We assumed that for series/collection movies, the box office of the previous movie may influence the box office of the latter movie. Thus we did and analysis for series/collection movies independently. We first grouped movies by series/collections (i.e. Star War collection and Harry Potter collection). To ensure the integrity of our data, if there exists a null value, we'll drop the whole series. Table 3-1 reflects the relationship between ith movie in a series/collection and its return rate. We found that with the increase of the length of the series/collections, the average return rate decrease.

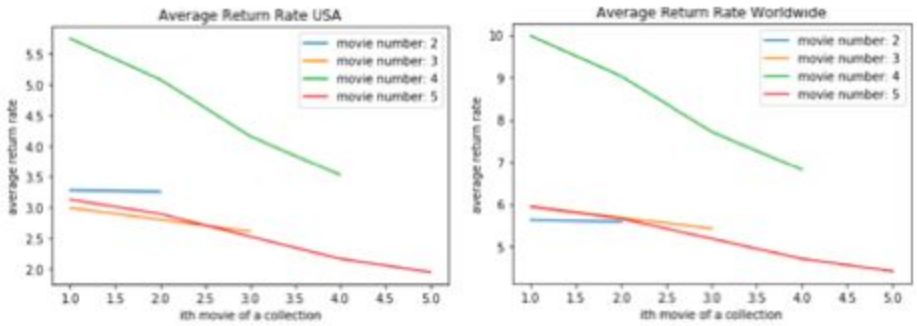


Table 3-1 Collection Movie Return Rate Change Trend

To explain this trend, we plotted the relationship between the ith movie and its gross as well as the relationship between the ith movie and its budget. From Table 3-2 and Table 3-3, we found that the increase in series/collections length is accompanied by a decrease in gross and an increase in budget. This means that the production company increases investment in production, but cannot match the expectation of the production company.

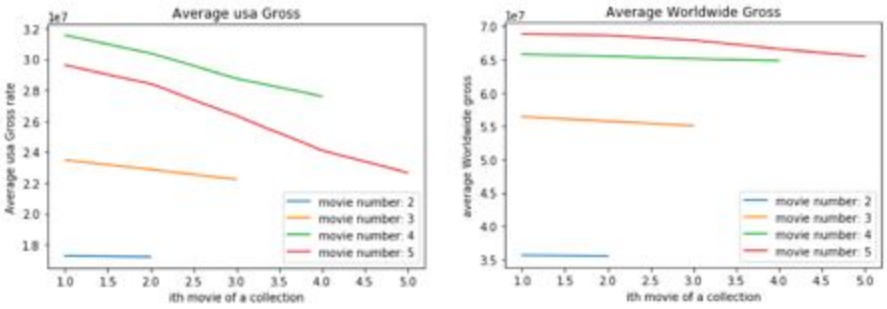


Table 3-2 Collection Movie Gross Change Trend

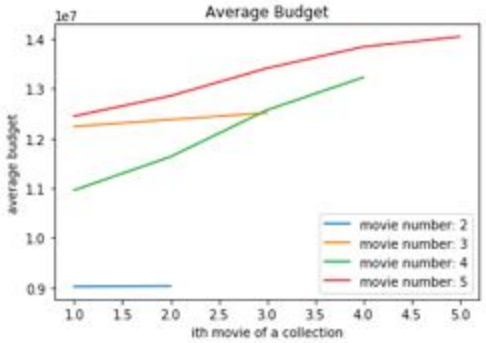


Table 3-3 Collection Movie Budget Change Trend

In summary, the box office of the previous movie may influence the box office of the latter movie. So our assumption was proved.

3.3.2 High Low Budget Return Rate

When we divided the movies into two groups: high (more than 50M USD) and low budget (less than 50M USD), we discovered a few trends. For high budget, about 30% of the movies yield negative revenue. Seasonality also seems to be affecting high budget movies, with higher return rates during the summer and winter break. On the contrary, all of the low budget movies yield a positive return rate and does not depend on release time. From Table 3-2, we see that there is a higher risk of investing in high budget movies. However, it is also possible that there is a proportion of low budget movies that cannot successfully release. That information we cannot require.

Table 3-2

4. Modeling

We use two machine learning models to predict the revenue range of a movie using both internal and external data.

4.1 Quantify Data

Before building models, we need to quantify the non-numeric attributes including genre, release date, country, director, writer, cast, and production company.

Genre: One-hot encoding transforms genre attribute to 21 binary attributes. All values are 0 except for 1.

Release date: Since external economic data changes over time, release date is used to map a movie to a specific value of the external attributes;

Country: One-hot encoding transformation is used. We split this attribute into three attributes: USA, UK and other_countries.

Director, Writer, Cast, Production Company: To judge a person, his past achievement is the most important factor that should be considered. We judge this from two sides:

Business Value: We first score movies (highest-revenue movie gets 100 points, lowest-revenue movie gets 0 points, and other movies are linearly distributed between 0 and 100). Then, for each person/company of each movie, we calculate the average score of all the movies he participated in before this movie as his score at that moment. In which $x<i$. $movie_x$ represents all the movies released before $movie_i$. Finally, we calculate the maximum and sum of the scores of all people/companies in a movie.

Number of Works: The number of previous works of a person/companies are also calculated to measure how experienced he is in making movies.

In addition, to reduce the difficulty of prediction, we don't predict the actual revenue. Instead, we calculate the 0%, 33.33%, 66.67%, 100% quantile, and divided all movies into three groups: low (0 - 33.33%), medium (33.33 – 66.67%) and high (66.67 – 100%), which are represented by 1, 2, 3 respectively and stored as variable gross_usa_adjusted_level .

4.2 Building Models

We finally got 5960 movies and randomly choose 20% as our test set. Decision Tree, Random Forest, and KNN revenue prediction models are built and compared.

4.2.1 Decision Tree

Since we divided our movies into only three groups, Decision Tree Regressor perform poorly in our case. Instead, Decision Tree Classifier is used to achieve classification, and the result is as below.

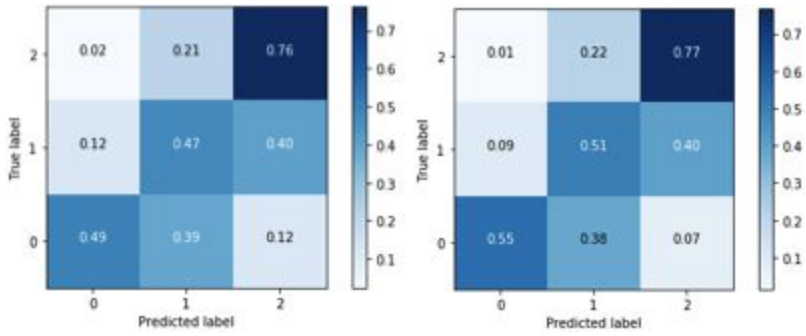


Fig X.1 confusion matrix of train set (left) and test set (right)

The test accuracy of our decision tree classifier model is 64.6%.

```
Test_Accuracy: 0.6459627329192547
Train_Accuracy: 0.6207279183311141
```

Fig X.2 confusion matrix of train set (left) and test set (right)

4.2.2 Random Forest

Random Forest model is used for comparison. We have defined 10 trees in our random forest and used entropy as the criterion. The test accuracy is 62.47%, and the complex matrixes of prediction results are shown below.

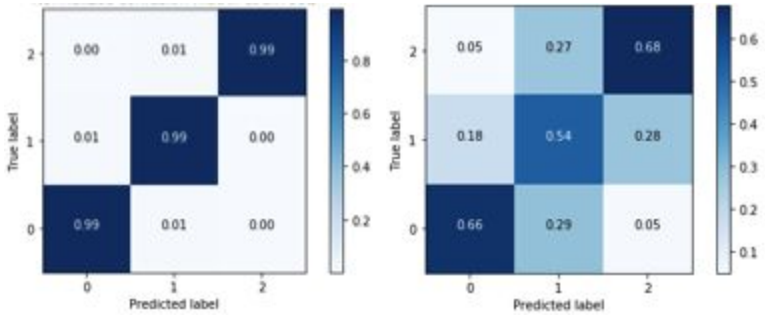


Fig X.2 confusion matrix of train set (left) and test set (right)

Random Forest is used to show the importance of features, which are shown below. This graph indicates that being produced by an experienced company, writer, director, and cast is the most important thing for a successful movie.

Fig X.3 Feature importance distribution, where 9 represents works_of_ProductionCompany, 7 represents works_of_writer, 6 represents works of director, 8 represents works of casts.

4.2.3 KNN

KNN model is used for comparison. The number of neighbors in our model is 7. When applied to the test set, it gives us an accuracy of 52.09%, which is the worst among all three models.

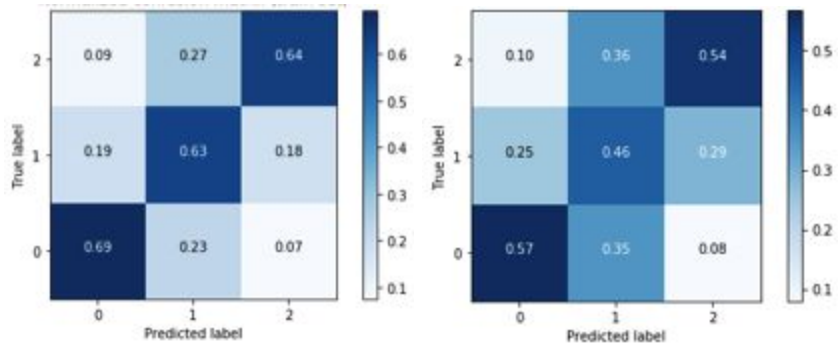
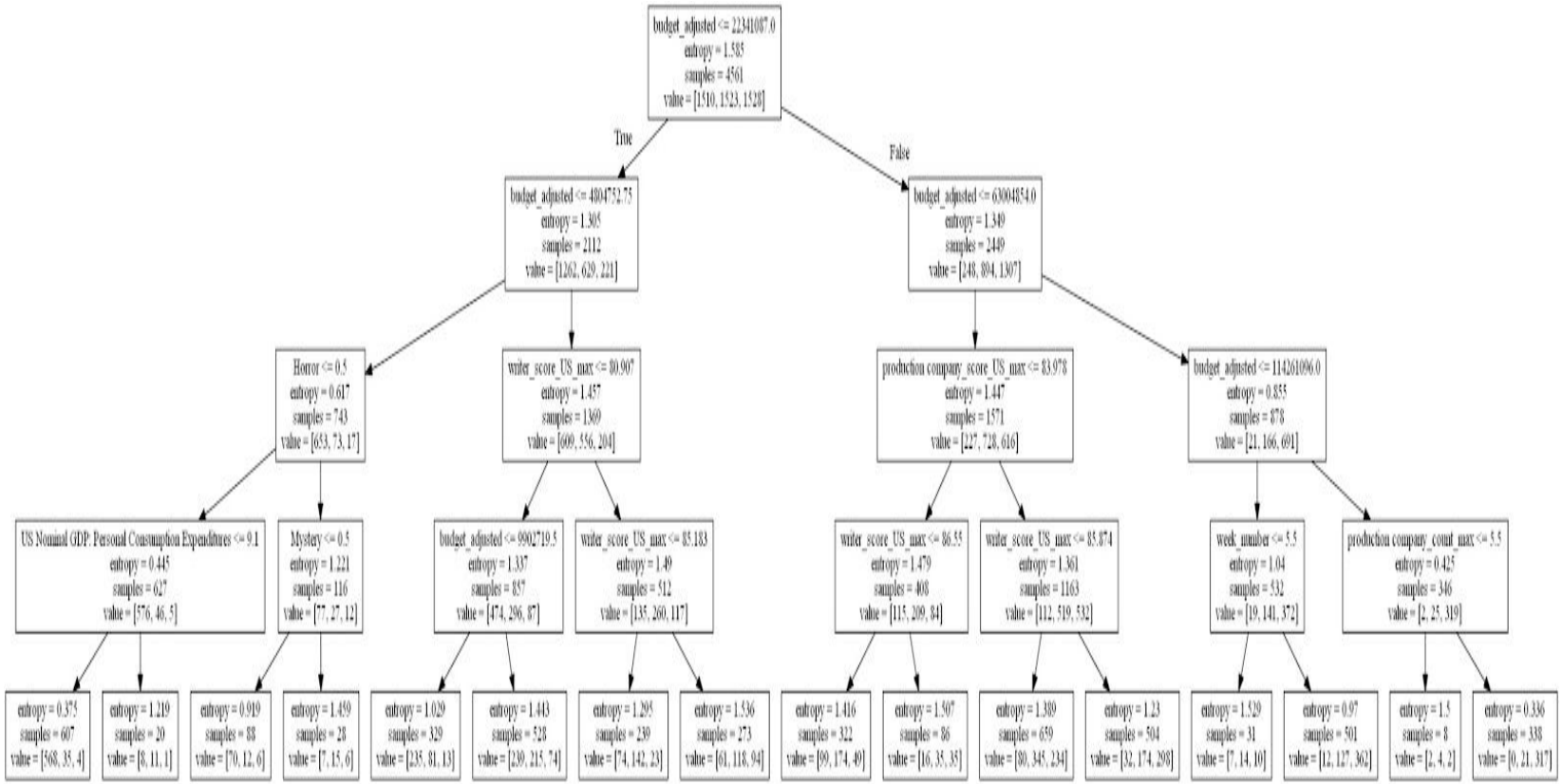


Fig X.4 confusion matrix of train set (left) and test set (right)

5. Prediction & Conclusion

After the model was built, we predicted a soon to be released movie Mulan (by Disney) and our model tells us that this movie will generate high revenue. Taking a final look at our decision tree, there are a few recommendations that we thought we could give the investors. First, if an investor do not have much money to invest (less than 4M USD), and decided NOT to invest in horror movies, we are quite confident that this investor will lose money. On the contrary, if you have a

moderate amount of money (60M USD-100M USD), we recommend that you do not release the movie in January. Finally, if you have more than 100M to invest, in this case, which production company you work with matters. We recommend collaborating with one of the more prestigious companies.


budget_adjusted <= 22341087.0
entropy = 1.585
samples = 4561
value = [1510, 1523, 1528]
True
False
budget_adjusted <= 4804752.75
entropy = 1.305
samples = 2112
value = [1262, 629, 221]
budget_adjusted <= 6300485.0
entropy = 1.349
samples = 2449
value = [248, 894, 1307]
Horror <= 0.5
entropy = 0.617
samples = 743
value = [653, 73, 17]
writer_score_US_max <= 80.907
entropy = 1.457
samples = 1369
value = [609, 556, 204]
production company_score_US_max <= 83.978
entropy = 1.447
samples = 1571
value = [227, 728, 616]
budget_adjusted <= 114261096.0
entropy = 0.855
samples = 878
value = [21, 166, 691]
US Nominal GDP: Personal Consumption Expenditures <= 9.1
entropy = 0.445
samples = 627
value = [576, 46, 5]
Mystery <= 0.5
entropy = 1.221
samples = 116
value = [77, 27, 12]
budget_adjusted <= 9902719.5
entropy = 1.337
samples = 857
value = [474, 296, 87]
writer_score_US_max <= 85.183
entropy = 1.49
samples = 512
value = [135, 260, 117]
writer_score_US_max <= 86.55
entropy = 1.479
samples = 408
value = [115, 209, 84]
writer_score_US_max <= 85.874
entropy = 1.361
samples = 1163
value = [112, 519, 532]
week_number <= 5.5
entropy = 1.04
samples = 532
value = [19, 141, 372]
production company_count_max <= 5.5
entropy = 0.425
samples = 346
value = [2, 25, 319]
entropy = 0.375
samples = 607
value = [568, 35, 4]
entropy = 1.219
samples = 20
value = [8, 11, 1]
entropy = 0.919
samples = 88
value = [70, 12, 6]
entropy = 1.459
samples = 28
value = [7, 15, 6]
entropy = 1.029
samples = 329
value = [235, 81, 13]
entropy = 1.443
samples = 528
value = [239, 215, 74]
entropy = 1.295
samples = 239
value = [74, 142, 23]
entropy = 1.536
samples = 273
value = [61, 118, 94]
entropy = 1.416
samples = 322
value = [99, 174, 49]
entropy = 1.507
samples = 86
value = [16, 35, 35]
entropy = 1.389
samples = 659
value = [80, 345, 234]
entropy = 1.23
samples = 504
value = [32, 174, 298]
entropy = 1.529
samples = 31
value = [7, 14, 10]
entropy = 0.97
samples = 501
value = [12, 127, 362]
entropy = 1.5
samples = 8
value = [2, 4, 2]
entropy = 0.336
samples = 338
value = [0, 21, 317]