Final Report

Wei Kong

Predicting Commercial Success in the Movie Industry

Note: The code to produce the results below and later in the report are hidden. For the original code please refer to the code on github repository (https://github.com/kongwei981126/Movie-Industry/blob/Wei/Final%20Report.ipynb)

Executive Summary:

We all know that movie industry is risky, but it can also bring enormous profit at the same time. That's why countless companies and people have been attracted to this industry. However, how risky is it? We all saw super hits like Titanic, and we all have seen big productions fail. There hasn't been an universal formula to make a popular movie, but it is always to our interest to choose the right movie to invest in. Personaly, I would also like to find out how to prevent the emergence of bad movies: that is how to identify them before they even started shooting.

In this analysis, I aim to produce a study to figure out specific reasons that lead to the success of a movie economically, and give advices for companies or individuals who want to enter this industry. In other words, I decided to explore how to select a movie that will be successful economically to invest in for companies and individuls who want to make a profit in this field.

The whole project was carried out in the following procedue:

- Collecting data
- Data Cleaning
- Outlier Removal
- Feature Engineering
- Hyperparameter Tuning / Model Selection
- Detailed Tuning of model
- Analysis of results

Data description/preprocessing

For the first half semester, I have been using "The Movies Dataset" on Kaggle (https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv). These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages,

production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies.

However, this dataset consists of large amount of missing entires for budget and revenue. There were only around 6000 entries with both of these features, and a large amount of data with problem, new data was chosen for the study. For detailed information on this old dataset, please refer to https://github.com/kongwei981126/Movie-Industry/blob/Wei/midReport.ipynb in the github repository.

New Dataset

The new dataset was gathered from The Numbers (https://www.the-numbers.com/), a website operated by Nash Information Services, LLC., consisting of detailed information (especially on the business aspect) about various movies. Code to get and assemble the dataset were recorded in file "scraping info" (https://github.com/kongwei981126/Movie-Industry/blob/Wei/scraping%20info.ipynb) and "Scraping more" (https://github.com/kongwei981126/Movie-Industry/blob/Wei/Scraping%20more.ipynb) in the github repository.

Overview on the dataset

```
Rows: 6382
Colums: 35
---------Info of dataset--------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6382 entries, 0 to 6381
Data columns (total 35 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   moviename                        6382 non-null   object 
 1   Budget                           6382 non-null   int64  
 2   link                             6382 non-null   object 
 3   Domestic Box Office              6382 non-null   float64
 4   International Box Office         6382 non-null   float64
 5   Worldwide Box Office             6382 non-null   float64
 6   Domestic Release                 5967 non-null   object 
 7   Dom Year                         5967 non-null   float64
 8   International Release            2942 non-null   object 
 9   Int Year                         2942 non-null   float64
 10  Rating                           5983 non-null   object 
 11  Runtime                          5399 non-null   float64
 12  Franchise                        1333 non-null   object 
 13  Keywords                         5242 non-null   object 
 14  Genre                            6224 non-null   object 
 15  Production Method                6212 non-null   object 
 16  Creative Type                    6096 non-null   object 
 17  Production/Financing Companies   4050 non-null   object 
 18  Production Country               5887 non-null   object 
 19  Languages                        5388 non-null   object 
 20  Leading Cast                     4644 non-null   object 
 21  Supporting Cast                  4381 non-null   object 
 22  Production and Technical Credits 5769 non-null   object 
 23  Director                         5769 non-null   object 
 24  Est. Domestic DVD Sales          2418 non-null   float64
 25  Est. Domestic Blu-ray Sales      2061 non-null   float64
 26  Total Est. Domestic Video Sales  2672 non-null   float64
 27  Lead Ensemble Members            704 non-null    object 
 28  Cameos                           583 non-null    object 
 29  Uncategorized Crew               1757 non-null   object 
 30  Narrator(s)                      81 non-null     object 
 31  Extras                           7 non-null      object 
 32  Documentary Subject(s)           26 non-null     object 
 33  Interviewee(s)                   26 non-null     object 
 34  Revenue                          6382 non-null   float64
dtypes: float64(10), int64(1), object(24)
memory usage: 1.7+ MB
None
```

Out[7]:

	moviename	Budget	link	Domestic Box Office	International Box Office	Worldwide Box Office	Domesti Releas
0	The Super Mario Bros. Movie	100000000	Super-Mario-Bros-Movie-The-(2022)#tab=summary	260,823,700.00	248,398,143.00	509,221,843.00	April 5t 202

In this dataset, we have 6382 observations across 34 features. The output above also mentioned the amount of null-values in each feature. However, some of them need futher processing before any other action can be taken. The following part will go through each of the feature in the dataset.

Features removed

These features were removed due to inability to incorporate NLP analysis in this analysis.

- moviename: the name of the movie.
- Keywords: keywords of the content of the movie.

These features were removed due to not sufficient information included in them.

- link: the link to the information page of the movie
- Production Country: Most of the movies were made in the United States, causing this feature to be not useful. (Refer to table below)
- Languages: Most of the moview were in english, causing this feature to be not useful. (Refer to table below)
- Production/Financing Companies: The large number of companies in this category made it hard to use.
- Supporting Cast
- Production and Technical Credits

Out[20]:

```
Top of Frequency table of Production Country
['United States']                      4325
['United Kingdom']                      203
['United Kingdom', 'United States']     179
['France']                               97
['Canada']                               59
Name: Production Country, dtype: int64

Top of Frequency table of Language of the movies
```

Out[19]:
```
['English']               4620
['French']                  72
['English', 'Spanish']      64
Name: Languages, dtype: int64
```

These features were removed due to large amount of null values (refer to the data overview outputs above) and inability to impute them.

- Est. Domestic DVD Sales: Estimated domestic DVD sales
- Est. Domestic Blu-ray Sales: Estimated domestic Blu-ray sales
- Total Est. Domestic Video Sales: Total estimated video sales
- Cameos
- Uncategorized Crew
- Narrator(s)
- Extras
- Documentary Subject(s)
- Interviewee(s)

Features merged

The following features were merged into two new features: Realease year and month. Here, whichever earlier between domestic and international release (if differs) is taken as the final result:

- Domestic Release: Domestic release date
- Dom Year: Domestic release year
- International Release: International release year
- Int Year: International release year.

The following features were merged into one: leading cast. These two names were used interchangably for different movies:

- Leading Cast
- Lead Ensemble Members

The following features were merged into one: Revenue (The total boxoffice of a movie, which is what we aim to investigate in this report)

- Domestic Box Office
- International Box Office
- Worldwide Box Office

Features changed

- Franchise: Due to the large amount of franchises out there, this feature is turned into a binary True or False indicating if a the movie is in a Franchise.

Features unchanged

- Rating: Motion Picture Association film rating system rating of the movie
- Runtime: Length of movie (in minutes)
- Budget
- Production Method: 'Live Action', 'Digital Animation', 'Animation/Live Action', 'Hand Animation', 'Stop-Motion Animation', 'Rotoscoping', 'Multiple Production Methods'
- Creative Type: 'Contemporary Fiction', 'Historical Fiction', 'Dramatization', 'Science Fiction', 'Fantasy', 'Kids Fiction', 'Factual', 'Super Hero', 'Multiple Creative Types'
- Franchise
- Director

Outlier detection

In this part, a new variable was introduced to measure the economic return of movie. ROI was conputed by (Revenue - Budget)/Budget * 100%. Following are some information on distribution of ROI

Out[46]:
```
Descriptive Statistics of ROI
count     6,053.00
mean        391.31
std       2,895.68
min        -100.00
25%         -36.71
50%          82.06
75%         292.78
max     179,900.00
Name: ROI, dtype: float64

Lowest ROI movies
```

Out[48]:

	moviename	Budget	link	Domestic Box Office	International Box Office	Worldwide Box Office	Do
5892	Kirikou Et Les Hommes Et Les Femmes	7500000	Kirikou-Et-Les-Hommes-Et-Les-Femmes-(France)(2...	17.00	0.00	17.00	
16	Capricorn One	5000000	Capricorn-One#tab=summary	401.00	0.00	401.00	
4990	Pandaemonium	15000000	Pandaemonium#tab=summary	1,438.00	0.00	1,438.00	
6012	Perrier's Bounty	6600000	Perriers-Bounty#tab=summary	828.00	0.00	0.00	M
15	Born of War	5000000	Born-of-War#tab=summary	671.00	0.00	671.00	

Highest ROI movies

Out[51]:

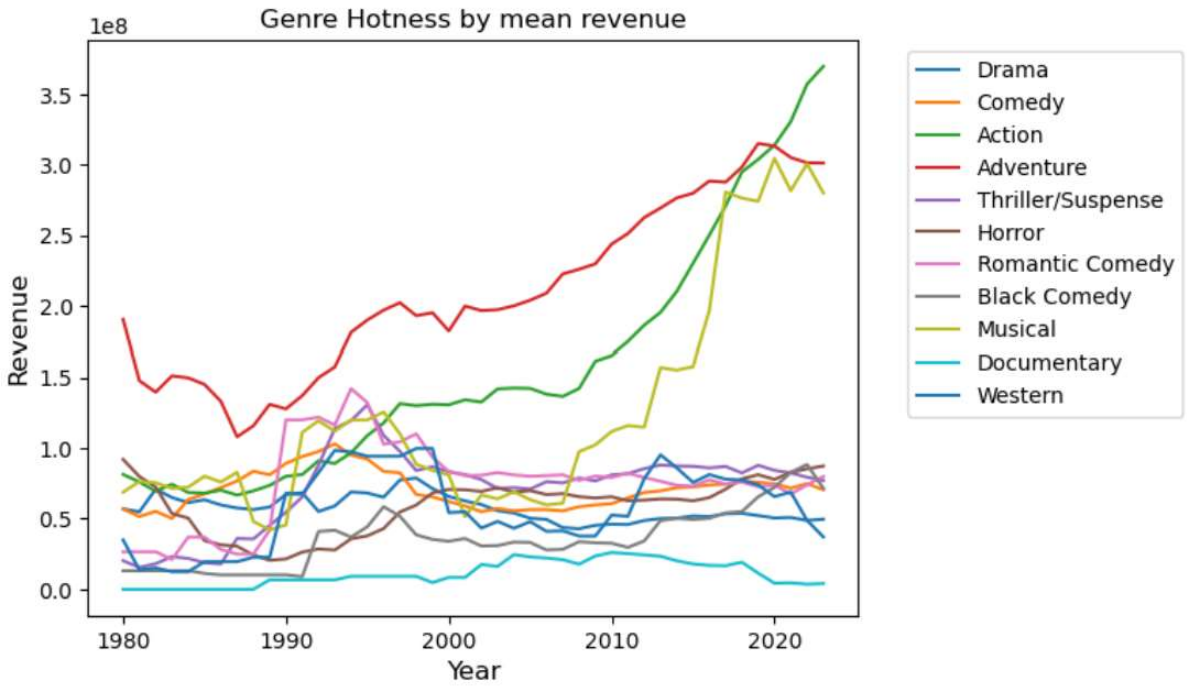
	moviename	Budget	link	Domestic Box Office	International Box Office	Worldwide Box Office	Dome Rele
1117	The Blair Witch Project	600000	Blair-Witch-Project-The#tab=summary	140,539,099.00	107,760,901.00	248,300,000.00	July 1 1
1409	The Gallows	100000	Gallows-The#tab=summary	22,764,410.00	18,892,064.00	41,656,474.00	July 1 2
1207	Paranormal Activity	450000	Paranormal-Activity#tab=summary	107,918,810.00	86,264,224.00	194,183,034.00	Septem 25th, 2
1342	Mad Max	200000	Mad-Max#tab=summary	8,750,000.00	91,000,000.00	99,750,000.00	M 21st, 1
1486	Deep Throat	25000	Deep-	45,000,000.00	0.00	0.00	June 3

Here, we do have some strange values, both with super large and small ROI. That being said, I decided to fact check the movies: For movies with low ROI, with evidence of ROI of -95% existing, -95% is used as a cutoff. For moview with high ROI, the one movie "Deep Throat" with incredibly high ROI is removed. No statistical outlier detection method were included to remove outliers since all movies are our subject for prediction.

More Feature Engineering

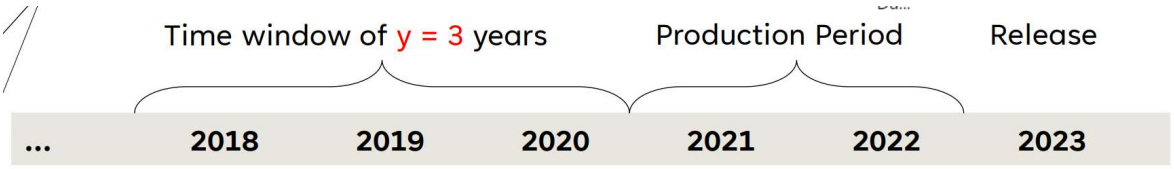
In analyzing Genre trend, we can see the following: The hotness of Genres change over time, so as the trend with actors and directors. Like shown in the following graph, the hotness of different genres rise and fall over time.

For creation of the graph below, refer to the the Descritive.ipynb in the repository (https://github.com/kongwei981126/Movie-Industry/blob/Wei/Descriptive.ipynb)



Thus, genre may not be a good feature to predict the performance of a movie. Instead, I propose the following method:

To make the explanation clear, use a movie released in 2023 as an example. As shown in the graph below, the production period of a movie usually takes 1 to 2 years, thus, we have to decide by 2020 if we want to make a movie that is gonna be released in 2023. Thus, only information till 2020 are considered available when considering the investment of a movie released in 2023.



Then, we consider a time window of y years till the time of decision. This time window of y years is used to measure the current trend in actors, directors, and genre.

In each time window, the following three things are measured:

- The total revenue of each actor in this y year period.
- The total revenue of each director in this y year period.
- The total revenue of each genre in this y year period.

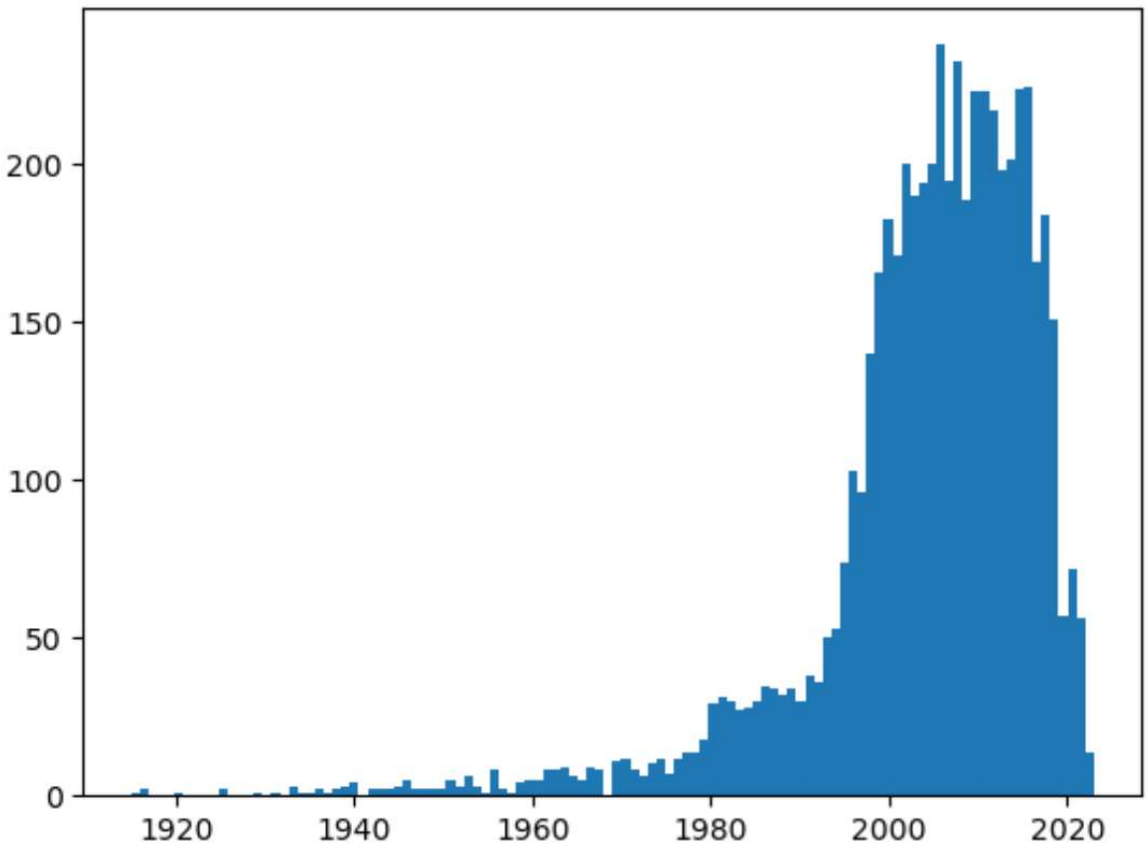
Then, from the information above, we define the top n actors and directors the "star actor" or "star director" at that time. After that, five new features were made:

- The total revenue of lead actors in the time window beforehand.
- The total revenue of directors in the time window beforehand.

- The total revenue of the genre of the movie in the time window beforehand.
- If there is a star actor in the leading cast by the time of production.
- If there is a star director at the time of production.

Finalizing

Due to the fact that the time window technique was used, it is mandatory for all data to have "year" feature. Thus, observations with NA in year were removed. In fact, due to the low number of observations before 1980, the dataset was further subsetted to only observations after 1980 to ensure better fitting. This left enough data for model and should not hinder validation of the model. Below is a histogram of numbers of movies versus years. It can be seen that few movie were produced before 1980. Refer to the the Descritive.ipynb in the repository for the production of the graph below. (https://github.com/kongwei981126/Movie-Industry/blob/Wei/Descriptive.ipynb)



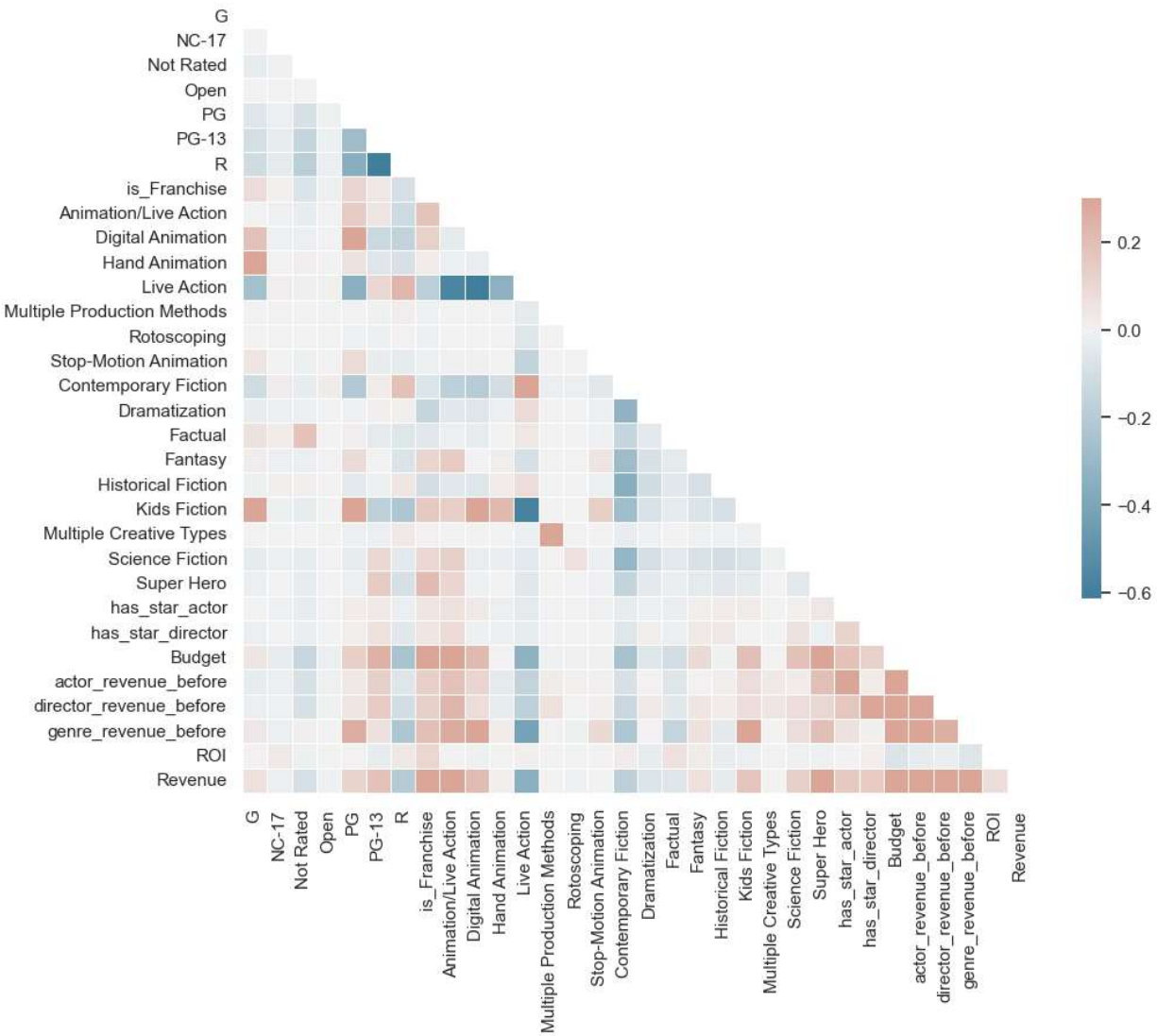
Modeling approach

Deciding on the objective

At first, the project aimed to predict ROI of a movie, due to the fact that ROI should counter the problem of inflation and give a better idea of relative return instead of absolute return. For instance, ROI for a big budget movie and a small budget one making the same revenue will be greatly different while the difference cannot be directly estimated by their revenue. (For further details on predicting ROI, please refer to the following report: https://github.com/kongwei981126/Movie-Industry/blob/Wei/passMidReport.ipynb)

One major reason is the fact that ROI has low correlation with all the features in the dataset, which can be seen from the correlation plot below. Therefore, instead of ROI, the project switched onto predicting Revenue of a movie yet to be made.

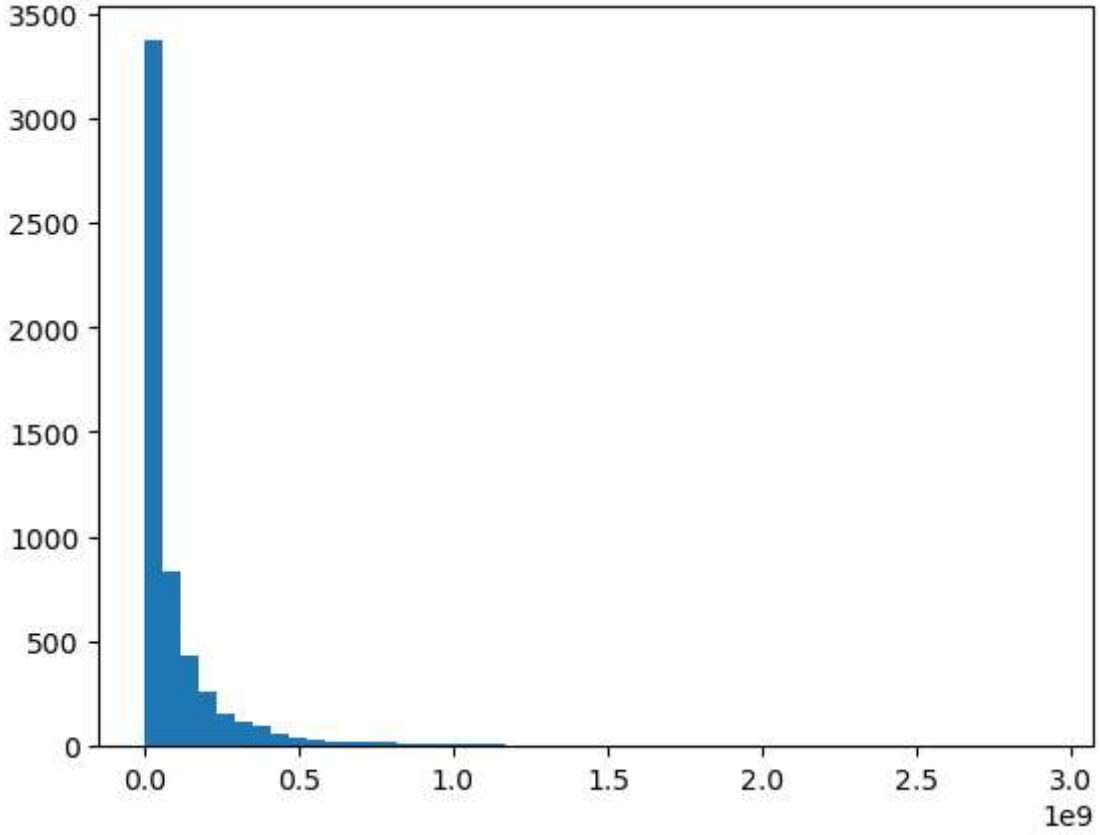
Out[450]: <AxesSubplot:>



Model Selection

From the histogram below, we can see that the response is hightly skewed. This ruled out most methods based on normality. Thus, tree based methods were selected for this project: mainly random forest and XGBoost.

```
Histogram of distribution of Revenue
```



Model Traning Process

The following parts describe the procedure of model training:

Tuning Hyperparameters:

In this part, there are in all 5 hyperparameters for tuning:

- Length of time window to consider for actors, directors and genre: Here these three time windows are not required to be equal length.
- Number of star actor and directors to be considered for each time window: these two are also not required to be the same.

A grid search was performed while 3, 5, 10 were allowed for each of the 5 hyperparameters. In each iteration of the search, a 5-fold cross validation search on hyperparameters (max depth and features) of the model (random forest/XGBoost) was performed to find a model to estimate the performance with that set of hyperparameters. In this process, RMSE was used for comparison.

The following shows the example result of hyperparameters of both grid searches:

```
Random Forest Grid Search Results
```

Out[85]:

	timewindow_actor	timewindow_director	timewindow_genre	star_actor_num	star_director_num	
53	3	5	10	10	10	'n
59	3	10	3	5	10	'n

XGBoost Grid Search Results

Out[86]:

	timewindow_actor	timewindow_director	timewindow_genre	star_actor_num	star_director_num	
54	3	10	3	3	3	{'r
43	3	5	5	10	5	{'r

Due to the fact that XGBoost out performed random Forest model for all set of parameters, It was chosen as the final model. Then, for the best of hyperparameters, another 5 fold cross validation was used to fine tune each of the model (XGBoost). Here, RMSE is again used for comparison.

Results and Insights

Firstly, below provides the descriptive results for the model, and this part will discuss further interpretations.

```
r2 score:  0.5738890076632135
MSE:  1.3877082380000316
```

Comparison with alternative models:

In order to compare and measure the progress in this model, alternative models were produced.

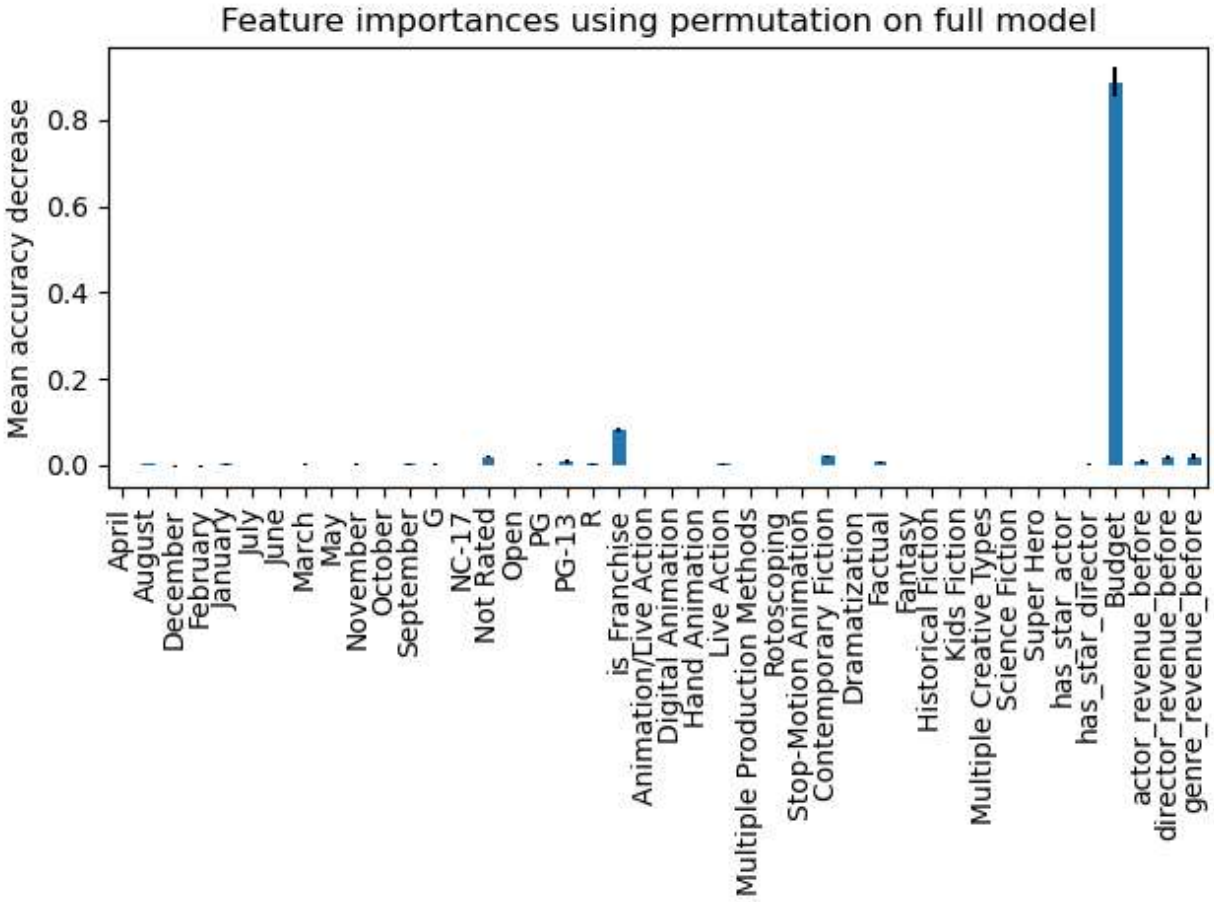
XGBoost model without feature engineering

This part below shows the result of an alternative model: This model is trained with XGBoost without all the features engineered in this project. Instead, the original ones were used in training. The great boost on r2 score and decrease MSE validates the importance of the features created in this analysis.

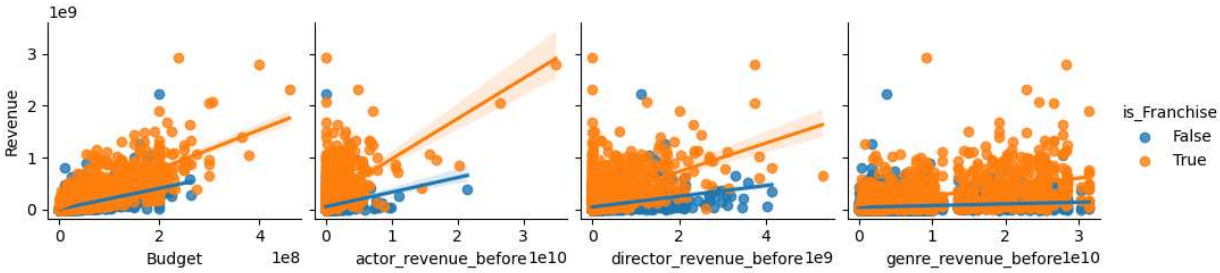
```
r2 score:  0.49796386164252404
MSE:  1.6210515813363064
```

Discussion on feature importance

As shown in the graph below, certain features are deemed important by the tree predictor. And the only one categorical feature being if the movie belongs to a franchise. In order to validate its importance, interaction analysis were conduct around this feature.



The graph below provides comparision for the two different groups. Revenue were plotted against the important numerical features in this project, and the two colors denote different groups. It is easy to see, in all four plots, movies that belongs to a franchise have a line above the ones that don't: this means that franchise movies produce a much higher revenue given the same circumstances compared to non-franchise ones.



To further support this argument, a linear regression was run on the whole dataset, Franchise movies, and non franchise movies to compare the results.

On the whole dataset:

```
r2 score:  0.1260607791334808
MSE:  2.023095142953555
```

On Franchise movies:

```
r2 score:  0.37567420925092965
MSE:  0.844808891064326
```

On non-franchise movies:

```
r2 score:  -0.28570693372472533
MSE:  2.367045898228619
```

The great difference in the latter two models validates the fact the two population performs really greatly. The boost in performance on Franchise movies means there is a linear correlation between features and confirms our findings. The bad result on non-franchise movies suggest they need more study: either they don't follow the linear trend line franchise movies, or there are other factors that plays the important part in their successes. Either way, this confirms our hypothesis that these two population behaves differetly.

Conclusions

In all, this project tried to find out important factors that lead to the success of a movie, and tried to predict its revenue. Some conclusions can be drawn from these results:

- The current trend in movie industry is that large budget movie and franchise movies tend to get larger success. Larger budget also usually means more money is used to hire more famous actors and/or directors, which are also important in our predictions.
- Franchise and non-franchise movies behave very differently in this industry. From the information this project has, it is hard to find better explanation or predictions on this matter. However, other information not included in this dataset (for instance, amount of money spent on advertisements, etc.) or other techniques (for instance, NLP analysis on keywords or titles, etc.) may introduce new insights into this project.
- The feature engineering in this project was successful and proved to be useful. It shows that it is possible to predict (to some extent) the success of a movie based on past performance of its cast and directors, and the average performance of movies of the same genre. Further analysis may dig deeper into how the trends shift through years, and may introduce deeper understanding of this matter.

Bibliography

Banik, Rounak. "The Movies Dataset." Kaggle, ROUNAK BANIK, 10 Nov. 2017, https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv.

De Vany, Arthur, and W. David Walls. "Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office?" Journal of Cultural Economics, vol. 23, no. 4, Nov. 1999, pp. 285–318., https://doi.org/10.1023/a:1007608125988.

De Vany, Arthur, and W. David Walls. "Bose-Einstein Dynamics and Adaptive Contracting in the Motion Picture Industry." The Economic Journal, vol. 106, no. 439, 1996, pp. 1493–514. JSTOR, https://doi.org/10.2307/2235197. Accessed 6 Mar. 2023.

Walls, W. David. "Modeling Movie Success When 'Nobody Knows Anything': Conditional Stable-Distribution Analysis of Film Returns - Journal of Cultural Economics." SpringerLink, Kluwer Academic Publishers, https://link.springer.com/article/10.1007/s10824-005-1156-5.

"Movie Budgets, Most Expensive Movies, Most Profitable Movies, Biggest Money-Losing Movies." The Numbers, www.the-numbers.com/movie/budgets. Accessed March 2023.

Appendix

All code used in this project are stored on github: https://github.com/kongwei981126/Movie-Industry/tree/Wei

For all code used in this report, please access through the original file: https://github.com/kongwei981126/Movie-Industry/blob/Wei/Final%20Report.ipynb

For references mentioned in this report, please access through the following files:

- For discussion on the original dataset: https://github.com/kongwei981126/Movie-Industry/blob/Wei/midReport.ipynb
- For creation of the dataset: https://github.com/kongwei981126/Movie-Industry/blob/Wei/scraping%20info.ipynb and https://github.com/kongwei981126/Movie-Industry/blob/Wei/Scraping%20more.ipynb
- For descriptive analysis: https://github.com/kongwei981126/Movie-Industry/blob/Wei/Descriptive.ipynb
- For discussion on ROI: https://github.com/kongwei981126/Movie-Industry/blob/Wei/passMidReport.ipynb
- Other code references for this anaysis: https://github.com/kongwei981126/Movie-Industry/blob/Wei/Goodbye_ROI.ipynb