

Kai Yan. Data Mining for Analyzing and Predicting the Success of Movies. A Master's paper for the M.S. in I.S. degree. April, 2021. 51 pages. Advisor: Arcot Rajasekar

As the movie industry has been growing rapidly as well as facing more and more challenges in recent years, a need to predict the success of movies arises. This study explores the IMDb movie data to reveal the underlying patterns related to genres, regions, directors, investment, ratings, etc. Using the movie metadata and constructing features of directors and cast members, the study builds machine learning models that present high performance in predicting *success* of movies, which is the range of ratings they received on the IMDb website. Based on the findings, this study provides suggestions on strategy-making for movie production, and it also proposes predictive models that could contribute to research works as well as real-world applications.

Headings:

IMDb movie data

Data visualization

Movie features

Random forest classifiers

DATA MINING FOR ANALYZING AND PREDICTING THE SUCCESS OF MOVIES

by
Kai Yan

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina
April 2021

Approved by:

Arcot Rajasekar

Table of Contents

1. Introduction	2
2. Literature Review	5
2.1. Movie Success	5
2.2. Movie Features	6
2.3. Machine Learning	9
3. Methodology	11
3.1. Data Collection	11
3.2. Software Configuration	11
3.3. Data Preprocessing	12
3.3.1. Basic Movie Information	12
3.3.2. Investment Information from TMDb	14
3.3.3. Directors and Cast	14
3.4. Feature Construction	16
3.5. Machine Learning Models	23
4. Results and Discussion	25
4.1. Visualization Analysis	25
4.1.1. Number of movies released	25
4.1.2. Average Ratings	28
4.1.3. Regions	30
4.1.4. Profitability	33
4.1.5. Directors	36
4.2. Predictive Models	39
4.2.1. Decision Trees	39
4.2.2. Random Forest	40
5. Conclusion	43
References	45
Appendix I	49
Appendix II	51

1. Introduction

In recent decades, the movie industry worldwide has been producing a large number of movies every year, and it has become a large market with great influence and revenue. In the United States, the major motion picture companies release approximately 500 movies in a year with on average \$60 million of investment capital per film (Simonoff & Sparrow, 2000). However, among the huge volume of movies, only a few can be successful and highly ranked. It is always a problem for producers to evaluate a proposed movie before setting up the project, in order to meet the expectation of box office and critical acclaim.

Besides, the movie industry faced serious challenges in the past year 2020. The impact of the COVID-19 crisis has suspended Hollywood and closed cinemas around the world. Due to the pandemic, most film-related events including festivals and awards have been canceled, halting film production as well as pushing back worldwide release dates of movies filming in 2020 (Stedman & McNary, 2020). It is reported that the global film industry incurred a \$5 billion loss during the coronavirus outbreak (Ozili & Arun, 2020). After the great losses on the investment and the sense of oppression, there is a great need to revitalize the movie industry with a strain on the budget and a despairing desire for success.

However, despite the large capital investment that must be made, the success of a movie is extremely uncertain. It is highly demanded by the movie producers that a reliable approach to the prediction of a movie's success during the time in which investments are being garnished. The past studies have two limitations: First, the studies focus almost solely on total box office revenue or theater admissions; Second, many of these studies employed features that are only available just prior to, or even after, the official release of the movie (Lash, Fu, Wang, & Zhao, 2015). Given this situation, the analysis of movie features before the start of shooting becomes necessary, and models to reliably predict the success of a movie can help improve the business significantly.

Benefited from the increasing computing techniques and exploding volumes of data, machine learning has emerged as a promising approach to accomplishing the prediction of movie success. With optimistic prospects for research advisability as well as real-world application, this paper hypothesizes that there is significant potential in mining movie data and visualizing important information, and it is also of vital value to develop machine learning models to predict the success of movies, which could help various stakeholders such as producers, financiers, and directors make more informed decisions on movie production. Having this motivation, this paper adopts movie data from IMDb, the world's most popular and authoritative source for movies, to investigate the relationship between movie features and ratings, and develop predictive models which leverage the approach of machine learning.

The significance of this paper is twofold. First, from the perspective of data analysis, different options and possibilities for movie features are explored, including genres, regions, directors, cast, budget, etc. Based on research works, it is revealed that all of these features have underlying relationships with the ratings of movies. For various categories of data, description from different angles is provided and potential explanations are proposed to help look insight what movies are more likely to succeed. Second, in terms of real-world application, machine learning models for movie success prediction prior to the release may yield significant value. In this project, the experiments are mainly conducted on decision tree and random forest models, and the performance of the models is compared with parameters tuned. This project provides the potential method that can be used to extract information to help analyze the movies, predict the success, and help make the decisions.

Our research questions are stated here:

RQ1: What information can we retrieve from the IMDb movie data?

RQ2: How can we construct machine learning models to predict the success of movies?

The reminder of this report proceeds as follows. First, this paper offers a literature review about the related topics and technical concepts. Then, this report elaborates on the methodology guiding the development of this project. Then, available data visualization and experiment results are presented and discussed. Finally, this paper ends with the conclusion.

2. Literature Review

2.1. Movie Success

Because the purpose of this paper is to predict *success* of movies, it is of paramount importance to define the success. In this paper, two measures of success will be discussed. The first one is profitability. As it is mentioned in the introduction section, many past studies regarding the movie industry focused on the box-office performances of movies. However, the cost to produce a movie is also very important when we consider the investment factors and thus it cannot be ignored when we check the success of a movie. Analysis of historical data also found that revenues are not directly related to profits (Lash & Zhao, 2016). Therefore, a more meaningful measure of success should be the numeric value of profits or the return on investment.

Another very important metric in movie success research is audience rating which is about the potential reception of a movie. It reflects the degree of preference by the audience and may lead to a great market power, which produces a word-of-mouth effect. Such information can be retrieved from different types of media, such as IMDb, Yahoo! Movies, Twitter, trailer ratings from YouTube, and so on. Although the demographic profile of the visitors who participate in the evaluations may not be the same as that of an actual audience, the evaluations can function as a good proxy if a sufficient number of participants can be secured (Chang & Ki, 2005). In this paper, we use the IMDb ratings as the measure of

success because IMDb is the most popular movie database, and it has the most structured rating system.

2.2. Movie Features

The abundance of movie data on the internet has encouraged many researchers to formulate techniques to analyze the underlying patterns. It is of vital importance to extract features from the data, and the accuracy of predictive models depends on the engineering of features. When it comes to studying movie success, three types of features have been explored by the previous works. The first one is audience-based features. These features are usually composed of text data and are mainly retrieved from social media, blogs, news articles, and movie reviews. The volume of discussions and the sentiment of reviews or comments are used as a measure for assessing the audience's excitement towards a movie (Lash & Zhao, 2016). The second one could be called content-based or visual-based features. It avoids human manual tagging but adopts deep learning technology to retrieve information from movie posters and clips (Deldjoo, Quadrana, & Cremonesi, 2018). It is a more advanced approach involving image or video analysis and it has great potential to be applied to movie recommendation systems (Deldjoo, Elahi, & Cremonesi, 2016). The third one is movie-based features which are the metadata directly related to a movie itself. The concentration of these

features includes discovering the role of budget, the significance of movies of a particular era, influence of particular actors or actresses, etc. in the success of a movie.

In this paper, we will mainly focus on movie-based features. Though the other two features are very informed and powerful in many situations, they usually reveal themselves only after a movie premiere. In this study, for real-world application of movie success prediction, we hope our input close to the practice that features mainly include those which the producers can influence during planning and production (Ericson & Grodman, 2013). Admittedly, limiting the set of features chosen may lead to an insufficient description of the attractiveness of a movie from the point of view of target users, but the rationale for selecting movie-based features is that we wanted to explore whether we can build working models with as few features as possible (Alspector, Kolcz, & Karunanithi, 1998).

Among the features that can be selected, the cast members are one of the most popular and intuitive features. It was agreed that the higher movie-star power is helpful for a movie's success. Research shows that the stronger a cast already is, the greater is the impact of a new movie of box office successes or with a strong artistic reputation (Elberse, 2007). From the perspective of producers, a stochastic consumer choice process based on the assumption that audiences choose to see films of their familiar actor/actress generates a particular distribution of financially successful films among film types (Albert, 1998).

This paper will also investigate directors as a movie feature. In the previous studies, renowned directors are assumed to have similar attracting power, like the above case of movie stars who contribute to the box office by attracting a persona-based audience (Chang & Ki, 2005). However, there is a lack of comparison between the significance of directors and movie stars. Some old research (Litman & Ahn, 1998) reported that the effect of directors was not significant, while new research shows that certain directors are more often on the top movie list than others (Meenakshi et al., 2018).

Regions, including the production region and the target market region, are another factor affecting the performance of a movie, especially in the cases of dubbing films in emerging markets. Studies reported that cinema attendance is already decreasing steadily in the U.S even before the pandemic outbreak, while on the contrary, in emerging markets such as China, cinema revenues have been strong, and the American film industry gains a significant profit by tapping into the Chinese market (Chiu, Chiu, Ho, & Zu, 2020). Research also found that consumers' rating behaviors are affected by cultural influences. For example, under-reporting is more prevalent among US online networks, while online reviews from China and Singapore are less likely to be extreme (Koh, Hu, & Clemons, 2010).

Genres are also very common features when it comes to movie classification. The association between movie preferences and ratings with the genre has been found in previous studies, such as children being fans of animation movies. It is reported that viewers would

give their favorite or frequently viewed genres high ratings because they are internally predisposed to like that category of movies (Moon, Bergey, & Iacobucci, 2010). In this paper, we want to know if genres have any relationship with the overall success of movies.

2.3. Machine Learning

As mentioned in the previous section, machine learning has emerged as a promising approach to discovering trends and patterns with a large amount of data, providing an effective method of analyzing and predicting movie success (Bramer, 2013). In this paper, we will adopt the supervised learning method, which is based on the notion of being able to predict an unknown attribute of an object based on the attributes we actually know.

Supervised learning could be used for predicting either discrete attributes with a finite number of distinct values or categories through classification, or continuous numerical attributes through what is known as regression (Bramer, 2013). Movie ratings, the variable we want to use as the target of prediction, are originally continuous numerical values.

However, it is meaningless to predict an exact number as the reference of a movie's success.

Instead, in most studies, researchers convert the number to a categorical label so that the problems are converted from regression to classification. Classification through machine learning has been very successful in predicting attributes in the form of a binary true/false value, in another word, whether the movie is good or bad (Ahmad, Duraisamy, Yousef, &

Buckles, 2017). Studies also suggest classifying the ratings into multiple reasonable classes, for example, Excellent, Average, Poor, and Terrible, for better usability of the predictive models (Asad, Ahmed, & Rahman, 2012). In this paper, we will mainly look at two machine learning methods: Decision Tree and Random Forest.

The Decision Tree is one of the simple and possible approaches to multistage decision-making. The idea behind this algorithm is to break up a complex decision into a union of several simpler decisions, trying to get the final solution that would resemble the intended desired solution (Safavian & Landgrebe, 1991). The Random Forest is an algorithm based on the decision tree. It was developed by Breiman (2001) with the goal of improving prediction performance over previous models, built upon the idea of combining an ensemble of multiple decision trees through a voting system. It generates a predictive model based on a number of combined individual models, developed with the goal of improving prediction performance. A forest of decision trees generated by RF will usually contain a number of hundreds or thousands of unique decision trees depending on the application (Persson, 2015). As the Random Forest algorithm is considered to yield good performance, ease of use, and transparency considering the importance of attributes in many studies and applications, we hope it would benefit our prediction of movie success.

3. Methodology

3.1. Data Collection

This paper adopts the dataset files downloaded from IMDb Datasets (<https://datasets.imdbws.com/>) and they cover all the movie titles available on IMDb. The database is refreshed daily, so the data is up-to-date when the experiment is conducted. The datasets contain the information of movies including the unique identifier of the title, the movie title, the region for this version of the title, the language of the title, the release year of the title, primary runtime of the title, up to three genres of the movie, the identifiers of principle cast/crew for movies, the average ratings and total votes received by the movie. The datasets also contain the information of directors and actors/actresses, including the identifiers, names, primary professions and so on. A detailed description of the datasets can be found in Appendix I.

Another dataset was fetched from TMDb (<https://www.themoviedb.org/>) as it provides the API to fetch the box office information of movies. In this dataset, mainly four fields are used in our study: the IMDb identifier, production companies, budget, and revenue.

3.2. Software Configuration

In this study, the data processing and model construction tasks will mainly be completed using Python 3 with Jupyter Notebook. Packages including “pandas” and “numpy” will be

adopted to facilitate data processing, and “matplotlib” and “seaborn” will be used for data visualization. For machine learning tasks, we will use “scikit-learn” to implement the algorithms and training the models. Weka will also be used for building machine learning models.

3.3. Data Preprocessing

3.3.1. Basic Movie Information

First things first, all the movies with release year before 2021 are selected from the original dataset `title.basics.tsv`, with 490,021 movies in total.

For genre analysis, we decided to split the genre field into duplicate rows as each movie contains up to 3 genres. After removing missing values, we counted the number of each genre assigned to the movies and got 28 genres showed in the following table. In our later analysis, we removed the genres with only few instances, including Film-Noir, Reality-TV, Talk-Show, Short, and Game-Show.

	Genre	Number		Genre	Number
1	Drama	178421	15	Music	9856
2	Documentary	96061	16	Musical	8836
3	Comedy	85909	17	Adult	8137
4	Action	39484	18	War	7833
5	Romance	39047	19	Sci-Fi	7708
6	Crime	28861	20	Western	6537
7	Thriller	26745	21	Animation	5990

8	Horror	22367	22	Sport	5309
9	Adventure	21642	23	News	1386
10	Family	13978	24	Film-Noir	786
11	Biography	12599	25	Reality-TV	365
12	Mystery	12452	26	Talk-Show	110
13	History	11120	27	Short	38
14	Fantasy	10396	28	Game-Show	14

Table 1. Genres with the number of movies.

The rating data `title.ratings.tsv` needs to be added to the movie data. After merging the two datasets, there are 259,350 movies left with valid ratings. When calculating the average ratings for each genre or each year, we use the number of votes `numVotes` as the weight.

The region information is contained in `title.akas.tsv`. Each movie might have more than one title which is identified by an ordering ID and differs in region, language, and a flag indicating if it's the original title. However, in this dataset, all titles labeled as the original ones are missing the region field. Instead, there is another entry sharing the same title and having the region, but the original field is labeled as false. To match the region information, firstly we need to separate the dataset based on `isOriginalTitle` is true or false. Then, we get the intersection of these two datasets on `titleId` and `title` to form the production region dataset. Finally, we merge the region information to the movie dataset we previously obtained. The final output shows there are 131,530 with non-missing region variable, with a total of 190 regions counted. For movies with both region and rating information, there are 97,696 valid

entries. When calculating the average ratings for each region, we use the number of votes `numVotes` as the weight.

3.3.2. Investment Information from TMDb

It is very common that a film is produced under the cooperation of several companies or studios. In the dataset fetched from TMDb, the field `production_companies` might contain more than one name separated by commas. For our company analysis, we split the `production_companies` string by comma and count this entry for each company.

The dataset contains the budget and returns on investment, then it would be easy to derive the profit and return on investment by the following definition::

- Profit = Revenue – Budget
- Return on Investment = Profit / Budget

As mentioned before, instead of the box office, these two variables will be meaningful as the measure of the movie's profitability in our later analysis.

3.3.3. Directors and Cast

The mappings between movie id and director id can be found in `title.crew.tsv`. In this dataset, the field “director” may contain two directors separated by a comma. We split this field and assign each director to a duplicated row. Then, we merge this dataset to the movie

data we got before on the movie identifier and we got the data of movies with corresponding directors and ratings. Table 2 shows a sample of data, where “tconst” is the identifier of the movie, “nconst” is the identifier of the director, “averageRating” is the IMDb rating of the movie, and “numVotes” indicates the number of people who rated the movie. The variables “averageRating” and “numVotes” are from the file title.ratings.tsv.

tconst	primaryTitle	year	genres	nconst	averageRating	numVotes
tt9916190	Safeguard	2020	Action,Adventure,Thriller	nm7308376	3.2	140
tt9916270	Il talento del calabrone	2020	Thriller	nm1480867	5.7	883
tt9916362	Coven	2020	Action,Adventure,Drama	nm1893148	6.1	280
tt9916428	The Secret of China	2019	Adventure,History,War	nm0910951	3.6	13
tt9916538	Kuambil Lagi Hatiku	2019	Drama	nm4457074	8.3	6

Table 2. Sample of director with movie rating data

To rank the directors by the rating and votes they received, previous studies usually set a threshold on the number of votes to filter out the movies that are watched by few people, and then calculate the average rating of the rest for the directors (Ahmad al et., 2017). However, we noticed that the decision of the threshold is very subjective, and the results differ a lot even though the change of the threshold value is slight. In this paper, we rank the directors in the following way. First, we define the total score of a movie as:

$$totalScore = averageRating^2 * numVotes$$

The reason we square the average rating is that we want to give this variable more weight.

For example, consider there are two movies where the first one has a rating of 10 with 100 votes while the second one has a rating of 1 with 1,000 votes. Apparently, we hope a total score will reflect that the first one is more popular than the second one.

Then we calculate the average total score of the movies by each director to get the value as the standard for ranking directors. To get the names of directors, we merge the table with the dataset `name.basics.tsv` by the identifier `nconst`.

3.4. Feature Construction

We want to build machine learning models for the prediction of movie success; thus, we create the target variable derived from the movie's rating:

- Excellent: $\text{rating} \geq 8$
- Good: $6 \leq \text{rating} < 8$
- OK: $4 \leq \text{rating} < 6$
- Bad: $\text{rating} < 4$

Given the labels, our model will perform a 4-class classification.

When constructing movie features, we include information about the directors and cast by the following rules:

For directors, we construct two variables: `director_rating` and `director_movies`. The variable `director_rating` is the average rating received by the director from the movies before the current one, with the number of votes as the weight. The variable `director_movies` is how many movies the director has made before the current movie. If there are two or more directors for one movie, we take their average as the features. We assume these two features will reflect the profession and popularity of the directors.

For actors/actresses, we construct two variables similar to the director variables: `act_rating` and `act_movies`. Different from directors, IMDb includes a lot of subordinate roles in the cast data, which makes it difficult to balance the influence of different actors/actresses. Therefore, we decide to pick the first cast member on the list for each movie as the principal and apply the same method to construct the features. For example, in the file `title.principals.tsv`, a movie might have three lines of actors with identifier “ordering” 1, 2, and 3. We pick the first one to represent the cast members of the movie. Finally, the variable `act_rating` is the average rating received by the actor from the movies before the current one, with the number of votes as the weight. The variable `act_movies` is how many movies the actor has acted before the current movie. We assume these two features will reflect the profession and popularity of the actor.

We keep the number of votes as a feature in our dataset based on two assumptions. First, it reflects the extent to which a movie is recognized by the people, which could provide some

information about the movie's success. The second assumption is that for predicting movie success, this variable can be treated as a measure of the movie-watcher group size.

A summary of all features is shown in Table 3.

Feature Name	Description
isAdult	The movie is for adult or not.
runtimeMinutes	The length of the movie in minute.
year	The releasing year of the movie.
region	The production region of the movie.
numVotes	The number of people rating the movie.
director_rating	The weighted average rating received by the director from the movies before this one.
director_movies	The number of movies the director made before this one.
act_rating	The weighted average rating received by the actor from the movies before this one.
act_movies	The number of movies the actor attended before this one.
genre_1	The first genre of the movie.
genre_2	The second genre, or NA.
genre_3	The third genre, or NA.

Table 3. Movie Feature Description.

There are 89,757 valid entries in total. To feed our data, we need to code the regions and genres from text to numbers. As these variables can be treated as categorical data, we use 1~25 to replace the genres and 1~190 to replace the regions. We save this data called “movies.csv”.

The correlations of the features are showed in Table 4. We noticed that director_rating and act_rating have a relatively high correlation because they are derived by the similar methods and cooperation between directors and actors/actresses is quite common. We keep

both variables for the later machine learning tasks as we assume they provide different information based on their definition.

	isAdult	runtimeMinutes	year	region	numVotes	director_rating
isAdult	1.000000	-0.025402	-0.034252	0.080961	-0.025094	-0.062353
runtimeMinu	-0.025402	1.000000	0.039470	-0.039457	0.034677	0.035597
year	-0.034252	0.039470	1.000000	-0.097812	0.066016	-0.121943
region	0.080961	-0.039457	-0.097812	1.000000	-0.016051	-0.120707
numVotes	-0.025094	0.034677	0.066016	-0.016051	1.000000	0.164914
director_rati	-0.062353	0.035597	-0.121943	-0.120707	0.164914	1.000000
director_mov	0.031713	-0.019974	-0.493430	0.141320	0.058883	0.082446
act_rating	-0.073222	0.033141	-0.132920	-0.114933	0.149845	0.682237
act_movies	0.016287	0.009437	-0.349832	0.059228	0.185563	0.126730
genre_1	0.374422	-0.016345	-0.052523	0.076757	0.031219	0.015973
genre_2	0.004644	-0.013982	-0.032903	-0.066672	-0.087689	0.010588
genre_3	0.049145	-0.014452	-0.035851	-0.057715	-0.147988	-0.010150

director_movies	act_rating	act_movies	genre_1	genre_2	genre_3
0.031713	-0.073222	0.016287	0.374422	0.004644	0.049145
-0.019974	0.033141	0.009437	-0.016345	-0.013982	-0.014452
-0.493430	-0.132920	-0.349832	-0.052523	-0.032903	-0.035851
0.141320	-0.114933	0.059228	0.076757	-0.066672	-0.057715
0.058883	0.149845	0.185563	0.031219	-0.087689	-0.147988
0.082446	0.682237	0.126730	0.015973	0.010588	-0.010150
1.000000	0.072152	0.367992	0.101645	-0.043995	-0.071313
0.072152	1.000000	0.274523	0.004308	-0.016873	-0.035190
0.367992	0.274523	1.000000	0.049208	-0.111209	-0.141267
0.101645	0.004308	0.049208	1.000000	-0.009578	-0.050925
-0.043995	-0.016873	-0.111209	-0.009578	1.000000	0.428239
-0.071313	-0.035190	-0.141267	-0.050925	0.428239	1.000000

Table 4. Movie Feature Correlations.

The distribution of classes is shown in Table 5.

Class	Number
Excellent	2788
Good	48197
OK	32703
Bad	6069

Table 5. Class distribution of movies.csv

Additionally, we want to include budget into our features but construct another dataset because the number of movies with the available budget is quite limited. After merging the budget to the above dataset, we got 5,030 instances, and the variable isAdult has only a single value 0, thus we have to remove it from the new dataset. We save this dataset called “movie_budget.csv”. Table 6 shows the correlations of the new features. In this table, the variable budget has a relatively high correlation with year and numVotes. The explanation would be that the recent movies usually got higher budget, and movies with higher budget usually attracted more viewers. Moreover, compared with Table 4, the correlation between director_rating and director_movies becomes significantly higher. By observing the dataset, we noticed that it is due to the focus on the popular movies and the exclusion of movies with missing budget data. This is an limitation of this dataset.

	runtimeMinutes
runtimeMinutes	1.000000
year	0.099251
region	-0.064905
numVotes	0.205411
director_rating	0.279958
director_movies	0.270765
actor_rating	0.194140
actor_movies	0.195407
genre_1	-0.126953
genre_2	-0.074080
genre_3	-0.059346
budget	0.264592

year	region	numVotes	director_rating	director_movies	act_rating	act_movies	genre_1	genre_2	genre_3
0.099251	-0.064905	0.205411	0.279958	0.270765	0.194140	0.195407	-0.126953	-0.074080	-0.059346
1.000000	0.085515	0.110495	-0.222764	-0.369638	-0.193861	-0.020810	-0.361781	-0.017369	-0.163470
0.085515	1.000000	0.028397	-0.113845	-0.138140	-0.095399	-0.103668	-0.023707	0.057735	0.034734
0.110495	0.028397	1.000000	0.396012	0.179931	0.350717	0.145437	0.146646	0.040940	-0.053171
-0.222764	-0.113845	0.396012	1.000000	0.399376	0.614786	0.120505	0.163127	-0.020709	-0.039042
-0.369638	-0.138140	0.179931	0.399376	1.000000	0.262221	0.269258	0.104964	-0.089351	-0.010292
-0.193861	-0.095399	0.350717	0.614786	0.262221	1.000000	0.309572	0.151304	-0.015445	-0.005963
-0.020810	-0.103668	0.145437	0.120505	0.269258	0.309572	1.000000	-0.014002	-0.128174	-0.063069
-0.361781	-0.023707	0.146646	0.163127	0.104964	0.151304	-0.014002	1.000000	-0.040470	-0.151972
-0.017369	0.057735	0.040940	-0.020709	-0.089351	-0.015445	-0.128174	-0.040470	1.000000	-0.117001
-0.163470	0.034734	-0.053171	-0.039042	-0.010292	-0.063069	-0.151972	-0.117001	0.123684	1.000000
0.349354	0.063429	0.397698	0.029292	-0.005963	0.052823	0.162592	0.106354	0.054925	-0.183683

budget	0.264592	0.349354	0.063429	0.397698	0.029292	-0.005963	0.052823	0.162592	0.106354	0.054925	-0.183683	1.000000
--------	----------	----------	----------	----------	----------	-----------	----------	----------	----------	----------	-----------	----------

Table 6. Movie Feature Correlations including Budget.

The distribution of classes is shown in Table 7.

Class	Number
Excellent	357
Good	3552
OK	1050
Bad	71

Table 7. Class distribution of movies_budget.csv

3.5. Machine Learning Models

Now we have two datasets: “movies.csv” with 89,757 entries and “movies_budget.csv” with 5,030 entries. We use them separately for machine learning tasks.

Weka is adopted for training Decision Tree Classifiers for its ease to use and powerful functions. We perform 10-fold cross-validation with the default parameter the confidence factor as 0.25. The decision tree models will be used as the baseline models and we want to conduct experiments on Random Forest to see if we can achieve better performance.

For the random forest models, scikit-learn which is a powerful python package for machine learning is used to implement the algorithm. Random forest makes use of internal

estimators to calculate the feature importance (Breiman, 2001), of which the interface is provided by scikit-learn so that take the advantages to investigate the features. To train the random forest models, firstly we randomly pick 80% data from the dataset as the training data and the rest as the validation data. For the hyper-parameters of random forest, the most important one is the number of estimators `n_estimators`, which represents the decision trees in the random forest. To determines the best `n_estimators`, a series of experiments are conducted to compare the prediction accuracy of the models. The result suggests that 100 is good enough for both two datasets.

4. Results and Discussion

This section will present the visualization analysis of the movie data and the experiment results of the machine learning models.

4.1. Visualization Analysis

4.1.1. Number of movies released

Figure 1 shows the total worldwide 490,021 movies released every year since the early 20th century. It can be observed during the past century, the development of the global movie industry is relatively slow in terms of the number of movies produced, and it suffered a decline during World War II. On the contrary, in the past 20 years, the number of movies has increased explosively from 5,000 to over 17,700 at the peak in 2017. However, it can be noticed that the number of movies has been decreasing to around 14,300 till 2020, which could be indirect evidence of the phenomenon mentioned before that the movie industry has been facing the challenge of the global economy (Chiu et al., 2020). It is also a signal for investors that the industry has entered the period of the bottleneck.

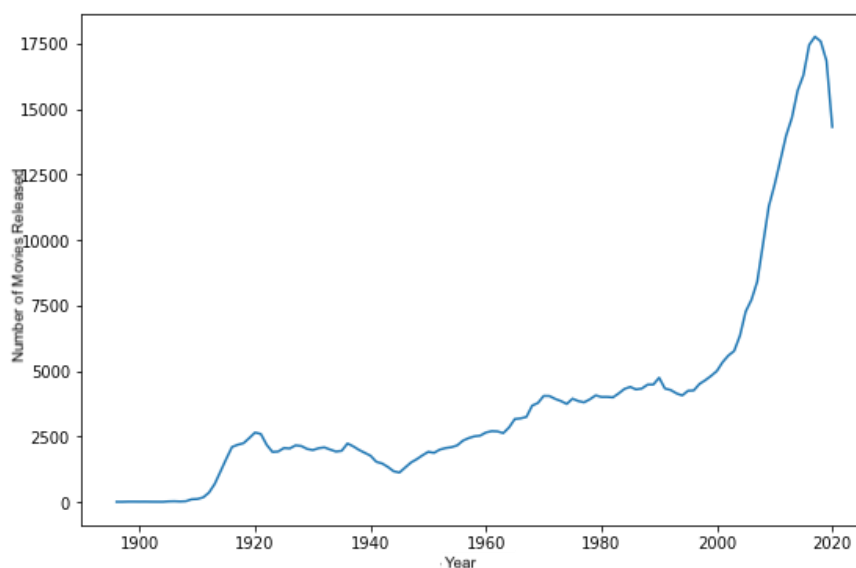


Figure 1. Number of movies released every year

Figure 2 is the pie chart showing the portion of different genres. Among all types of movies, Drama takes the most important part of movie production, followed by Documentary, Comedy, Action, Romance, Crime, and Adventure. Figure 3 shows the time tendency of the above major genres, from which we can see that Documentary has a huge increase since 2000 and passed most genres which were more popular before. It can be inferred that apart from the prosperity of other types of movies, the increasing number of movies in the 21st century is highly contributed by documentary movies. However, as a minor part of the movie industry and with low profitability, why documentary movies could make a steep rise in output? A potential explanation would be that the emerging digital technology has been leading to the revolution of films by enriching the tools of films including creation, distribution, and editing, and impelling new styles of movie-making (Fang

& Xiong, 2020). Compared with other commercial movies, documentary movies allow more people to enter this industry with the relatively low cost required instead of the traditional heavy and expensive equipment.

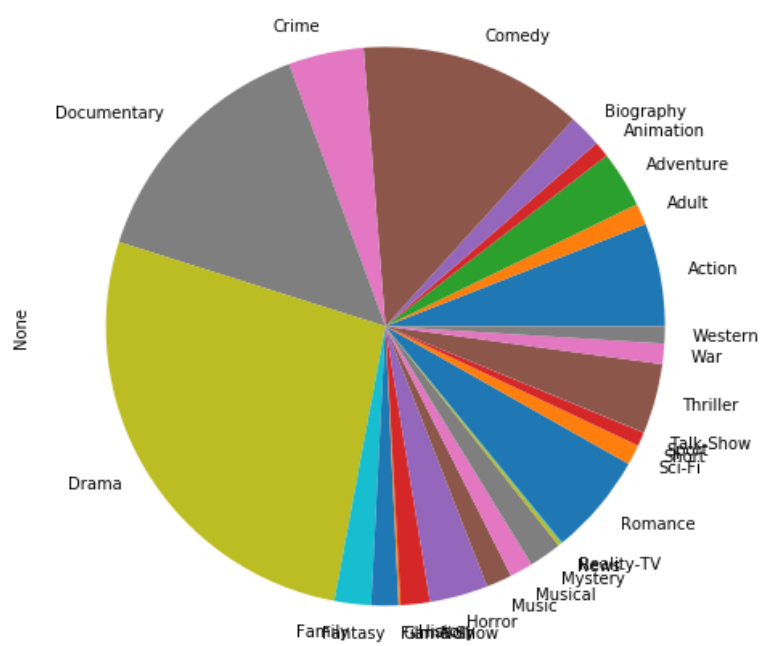


Figure 2. Pie chart of movie genres.

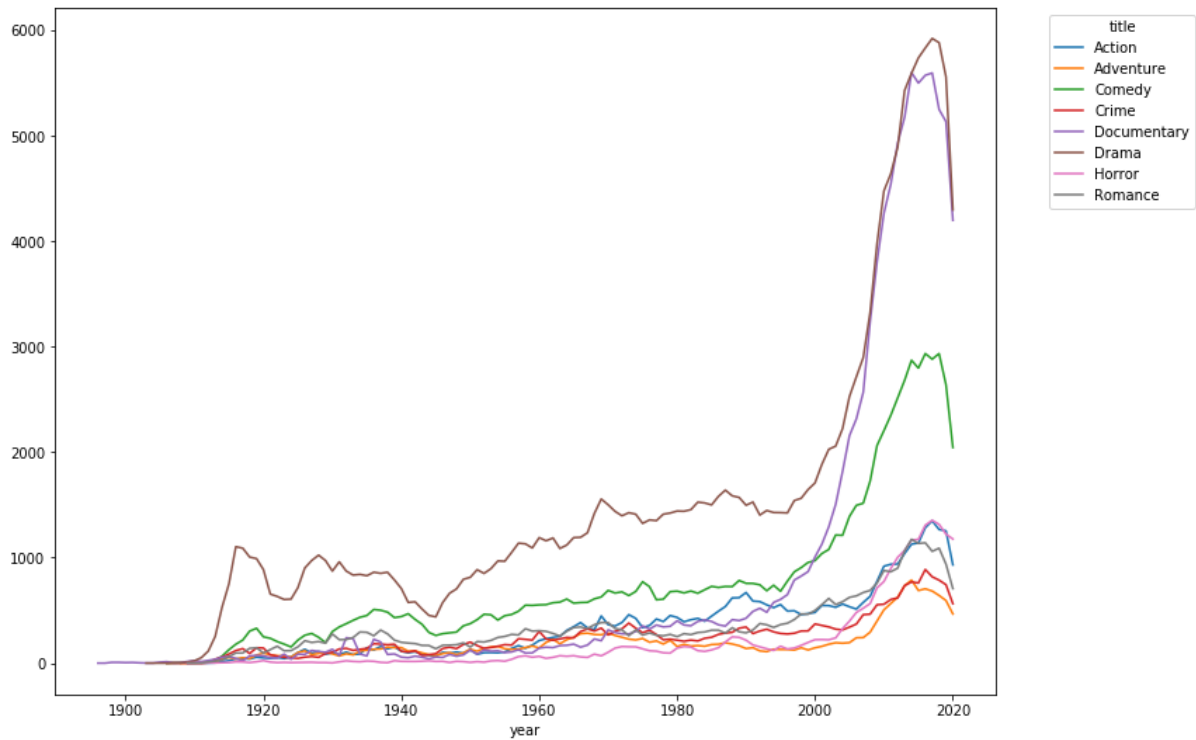


Figure 3. Number of movies of major genres according to the release year.

4.1.2. Average Ratings

To analyze the overall ratings of movies, we select the movies after 1920 because the previous movies and votes are very few which might not be representative enough. Figure 4 is the weighted average ratings of movies changing with the time in the past one hundred years, showing an obvious decrease over the past several decades. Combined with Figures 1 and 3, it can be inferred that despite the prosperity of the movie industry and the jumping in productivity, the average quality of movies has been sliding down over the years.

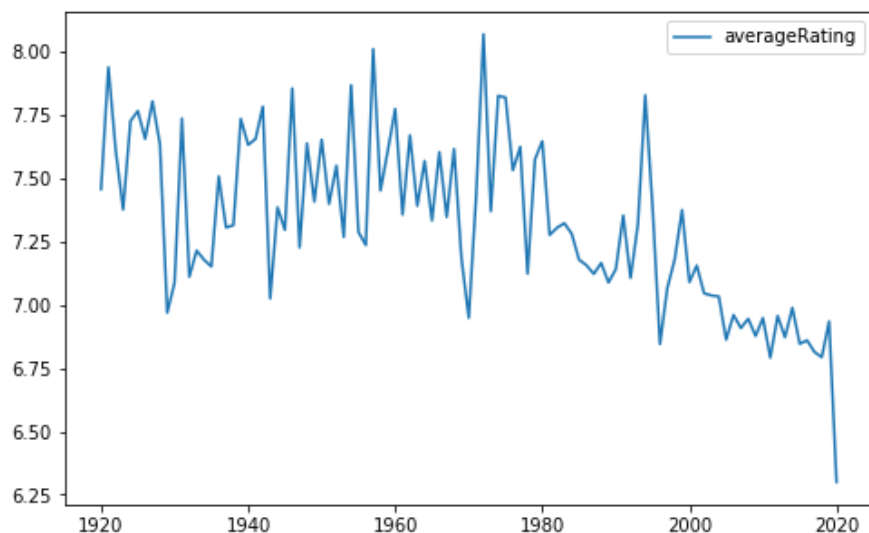


Figure 5. Weighted average ratings of movies.

To look insight the ratings of different genres, Figure 6 reveals that artistic movies like Documentary and Music are highly praised than the average and positively contribute to the ratings. Drama movies, the most popular genre, only perform moderately. The other popular commercial movie genres including Action, Thriller, and Horror usually receive relatively low ratings. These genres tend to be a hodgepodge of good and bad and more hardly satisfied the customers.

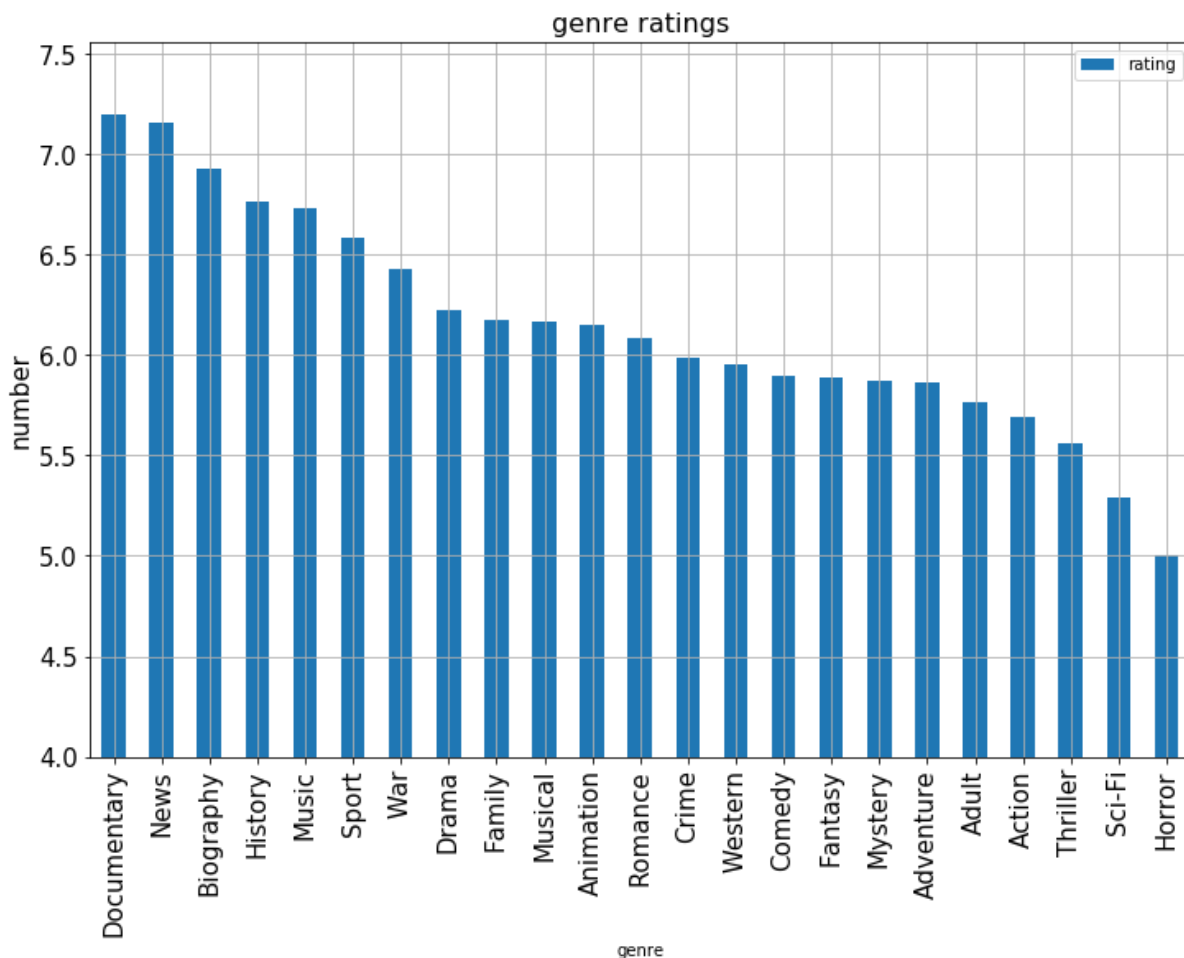


Figure 6. Weighted average ratings of different genres

4.1.3. Regions

Figure 7 is the pie chart of movies in terms of the production regions, and it shows the portion that the top 20 regions take. These 20 regions produce around 3 quarters of movies of the entire world, among which the U.S. produces over 25% of the global movies. As a comparison, Figure 8 shows how many movies were published in the top 20 regions, which more or less reflects the size of the market of different markets. Combined these two figures, the following interesting phenomena can be observed.

In terms of movie producers, France, Germany, and Italy are the top three following the United States. However, when it comes to the movie publication, the U.K. becomes the largest import country among the others. It reveals that language might be the barrier for the population of cultural products, given the condition that the British can easily accept movies from other English-speaking countries, while it cost a lot for the other European countries. On the contrary, it is very impressive that Japan is one of the largest markets for foreign movies as it takes a larger portion than most of the other countries in Figure 8. The region names and abbreviations can be found in Appendix II.

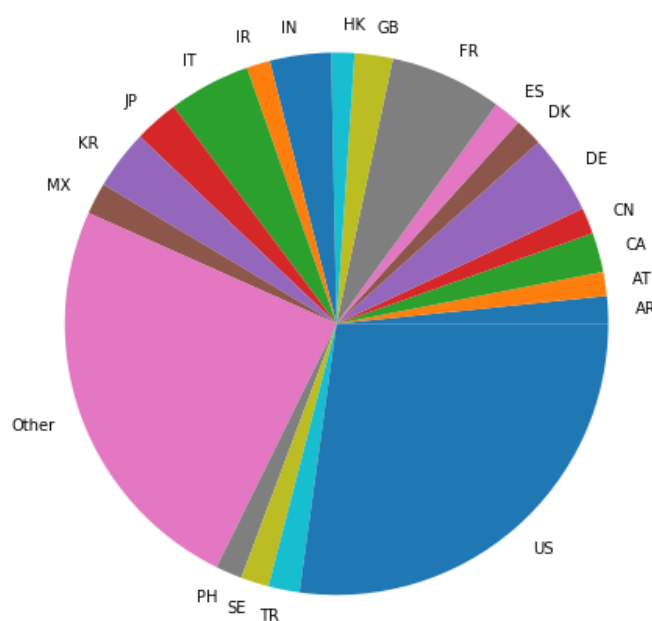


Figure 7. Pie chart of movie production countries in terms of movie numbers.

Figure 8 also has its limitation for measuring the size of the movie market. The literature reports that China has become the largest market for Hollywood companies (Chiu al et., 2020), it doesn't appear in this pie chart, which indicates that the size of the market depends

more on the total revenue instead of the amount number of movies. Nevertheless, it doesn't mean we do not need to care about the number of movies that could enter the country.

Nowadays, American companies, while receiving a lot of money from China, are also facing several challenges in this market, including the piracy concern and the local government censorship (Chiu al et., 2020). Given this situation, it becomes much necessary for companies that want to enter China and other emerging markets to carefully prepare the production of movies.

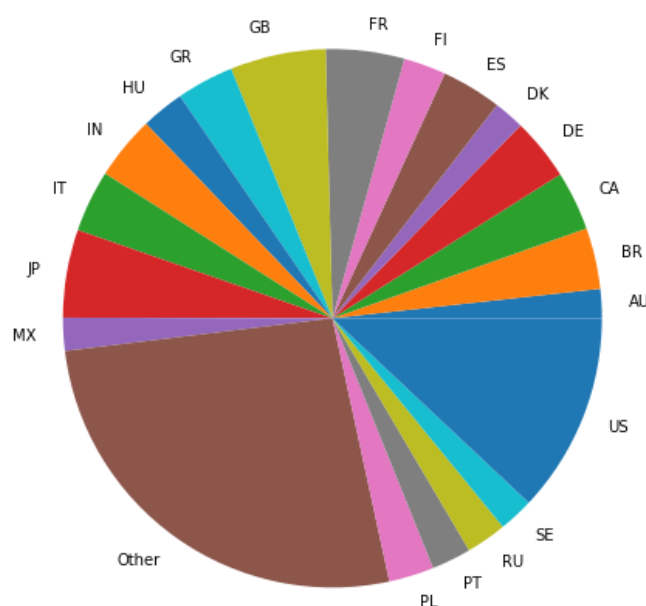


Figure 8. Pie chart of movie market regions in terms of movie numbers

Figure 9 shows the weighted average ratings received by the movies produced by the 20 countries in Figure 7. It shows that among the largest movie production countries, Japan and Germany receive relatively high praise from the customers, while the US and Italy perform below the average.

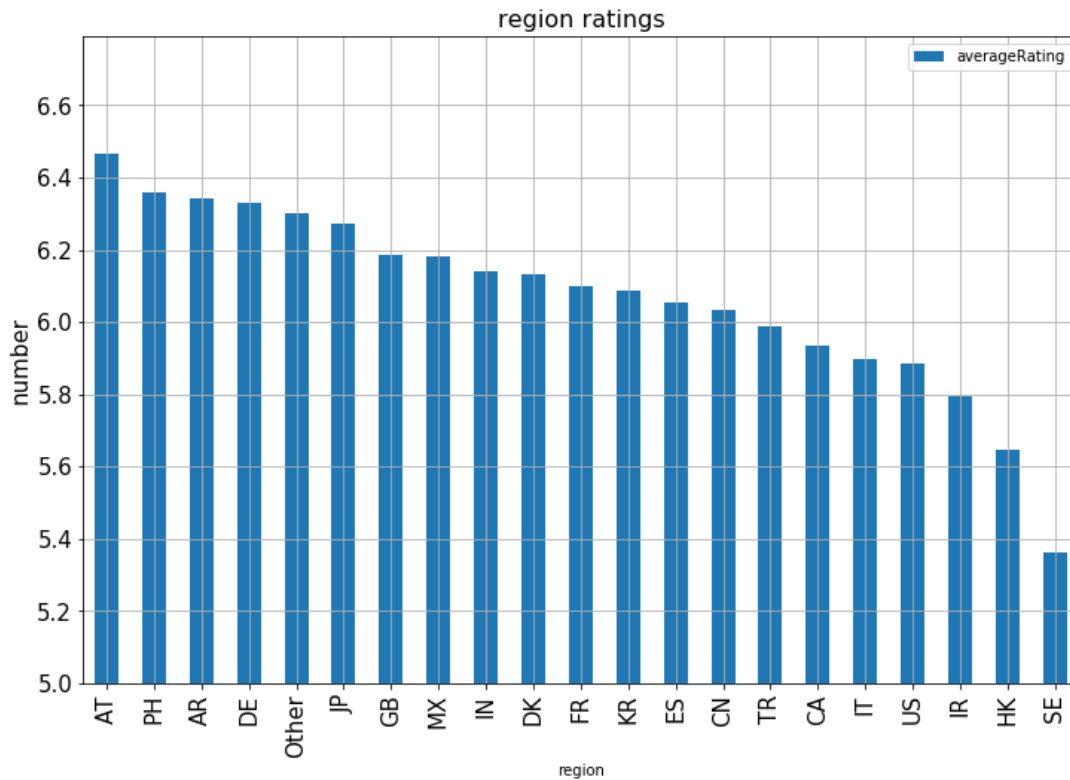


Figure 9. Weighted average ratings of different regions

4.1.4. Profitability

By calculating the revenue made by the motion picture companies in the past 50 years, we found that the world's most profitable ones almost come from Hollywood, and Table 8 shows the top 10 companies with the number of popular movies they made.

Company	No. of Movies
Warner Bros	454
Universal Pictures	446
Columbia Pictures	357
Paramount Pictures	345
21th Century Fox	290
New Line Cinema	192
Touchstone Pictures	155
Metro-Goldwyn-Maye	145

Walt Disney Pictures	144
TriStar Pictures	121

Table 8. Top companies with movie produced.

To illustrate the profit made by the companies changing along the time, we select the top 5 ones and plot the lines in Figure 10. It shows that the competition in the movie industry is extremely severe as every company has its own period of profitability, and for a certain, it is very hard to see all the companies share a common victory.

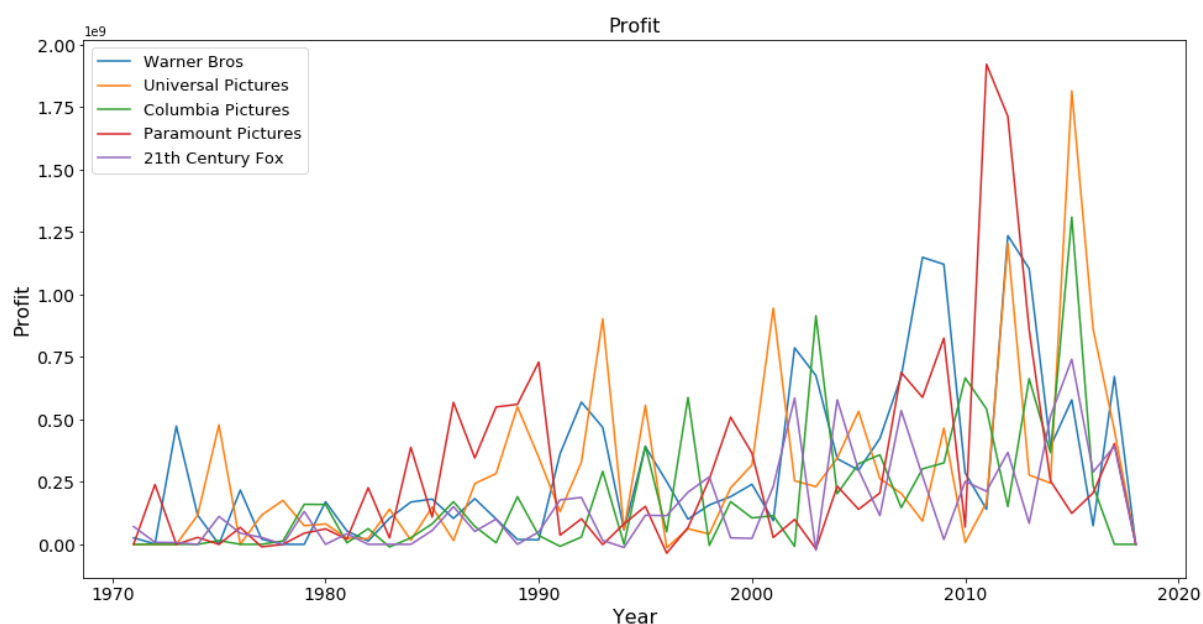


Figure 10. Profit made by the top Hollywood companies since 1970.

Figure 11 shows the top 20 movies based on their percentage return on investment. As it is mentioned before, since the box office earned by a movie does not give a clear picture of its monetary success, this analysis, over the value of the return on investment across the

movie budget, would provide better results. And pattern clearly revealed that the return on investment is high for those movies with a low budget and decreases as the budget of the movie increases.

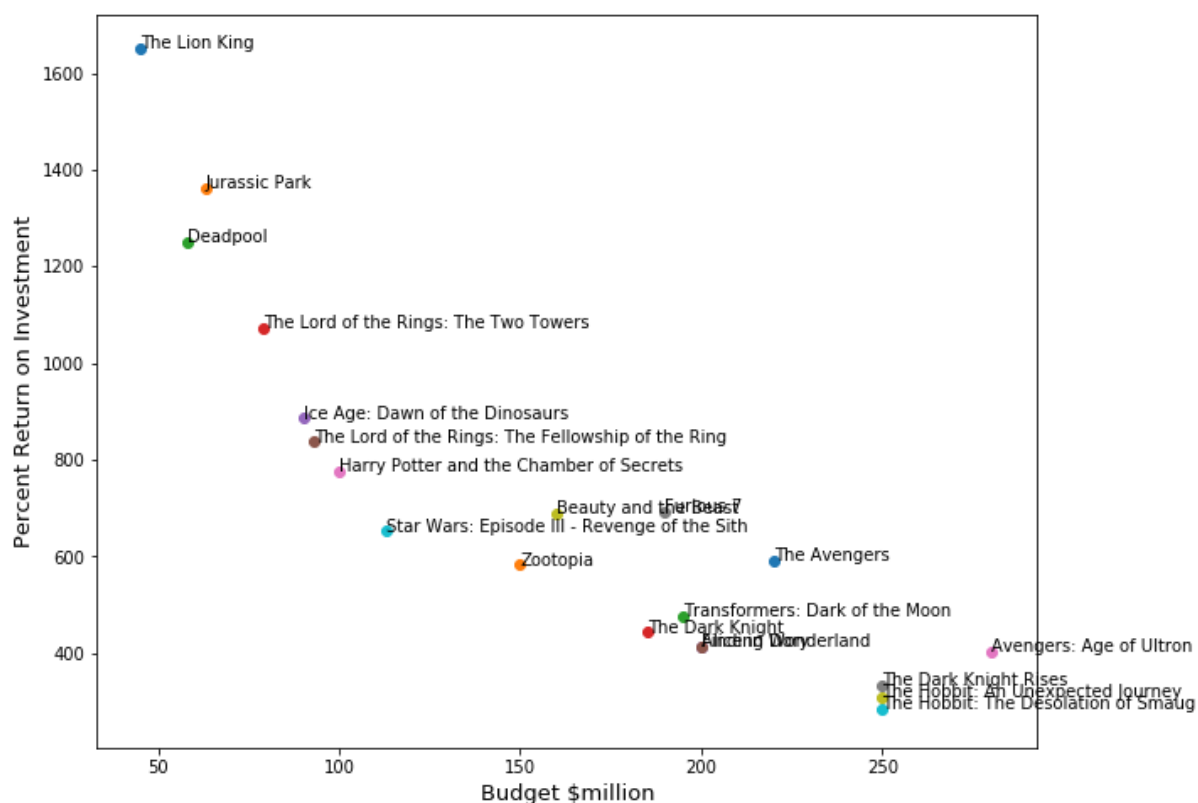


Figure 11. Top 20 profitable movies: Return on investment against Budget.

To further investigate this pattern, the top 100 most profitable movies are plotted in Figure 12. It can be found that the inverse correlation between the budget and the return on investment still holds to some extent, which suggests that it's usually very hard for a high budget to yield a high return on investment.

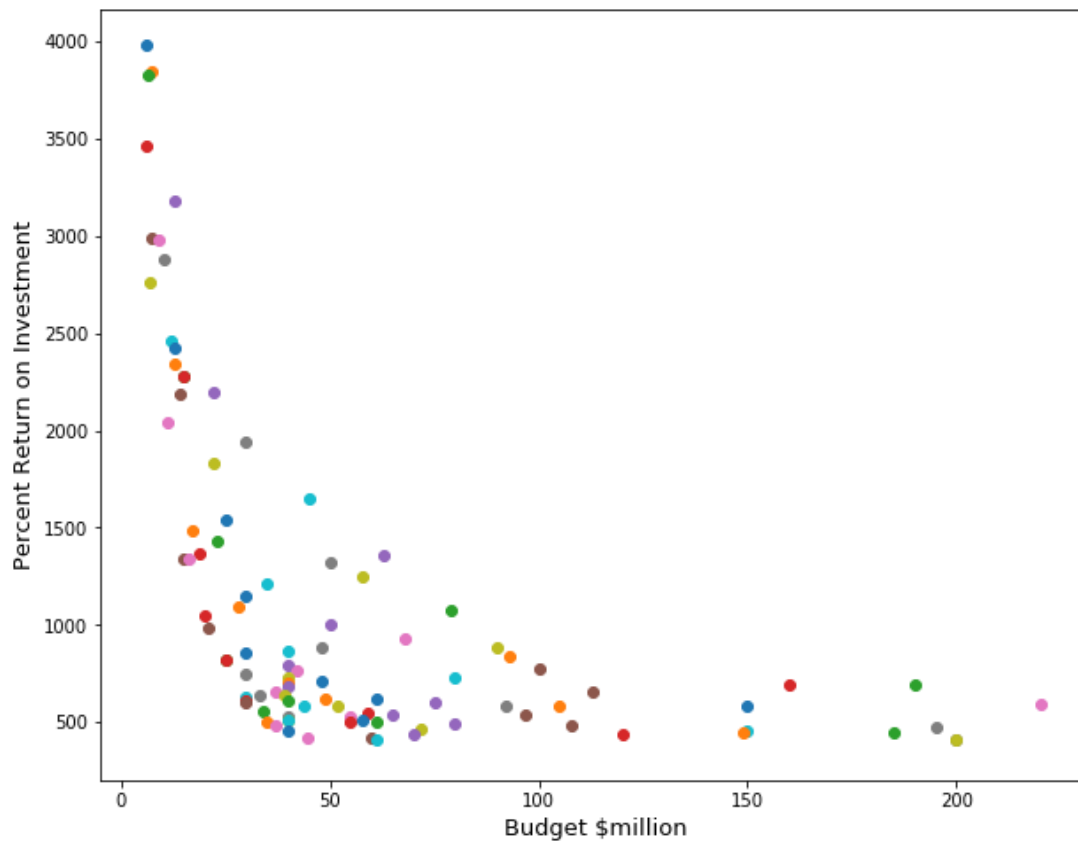


Figure 12. Top 100 profitable movies: Return on investment against Budget.

4.1.5. Directors

Table 9 shows the list of directors ranked by the ranking method using the weighted average total ratings received from IMDb. This list has two major differences compared with the traditional critical list of directors. First, there are more animation movie directors appearing in this list and even taking the front places. These names are all very successful in directing a bunch of animation movies while not as famous the others, which reveals these directors are highly recognized by the market but underestimated by previous research.

Secondly, the directors of the most popular movies in recent years like Avengers and DC movies show up on this list, which suggests that our measure of directors has

effectiveness for recent period of time. This measure gave reasonable value to the directors who are recently successful and popular, which allow us to expect their continuous success in the following years.

Rank	Name	Rank	Name
1	Christopher Nolan	16	Tim Miller
2	Frank Darabont	17	Joe Russo
3	Bob Peterson	18	Anthony Russo
4	Lee Unkrich	19	Lilly Wachowski
5	Quentin Tarantino	20	Lana Wachowski
6	David Fincher	21	Jared Bush
7	Jan Pinkava	22	J.J. Abrams
8	Ronnie Del Carmen	23	John Lasseter
9	Peter Jackson	24	Mel Gibson
10	Steven Spielberg	25	Nathan Greno
11	Joss Whedon	26	Brad Bird
12	Andrew Stanton	27	Damien Chazelle
13	Pete Docter	28	Matthew Vaughn
14	George Lucas	29	Neill Blomkamp
15	James Cameron	30	Stanley Kubrick

Table 9. Ranking of directors.

By picking the movies directed by the above directors and analyze the genres of those movies, we get the pie chart in Figure 13, which shows that these directors come from Drama, Comedy, Action, Adventure, etc. Compared with the distribution shown in Figure 2, it can be noticed that Documentary becomes negligible, while Adventure and Action play more important roles.

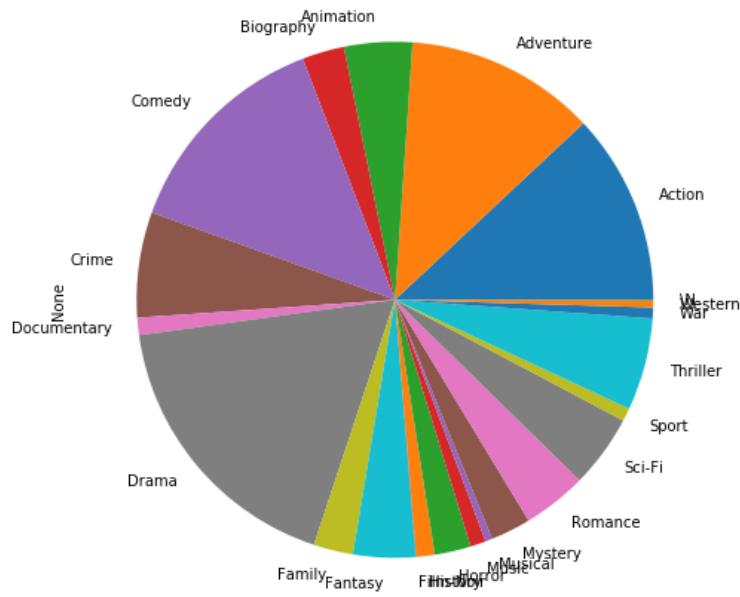


Figure 13. Genres of top 30 directors.

To further investigate our ranking method, we plot the box office of the top 20 profitable movies against their IMDb scores in Figure 14. By observing this figure, we found that even though our ranking method gives high advantages to the directors of popcorn movies, none of the directors of the movies appearing close to the left and bottom have a position on the director list. On the contrary, movies with high scores and revenues are usually directed by a top director. For example, The Dark Knight series, which is a masterpiece of Christopher Nolan, the director on the top of our list, receives both high revenues and ratings. Moreover, Christopher Nolan is also the most name appearing in IMDb's top 250 movie list.

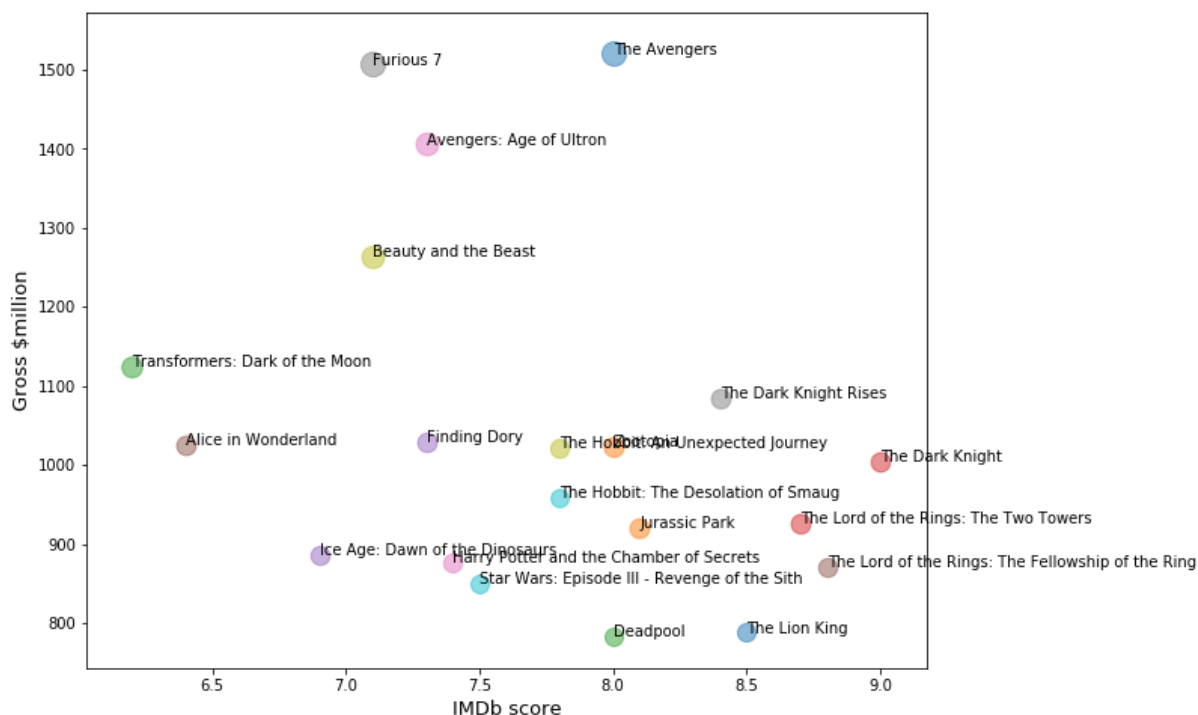


Figure 14. Gross revenue v.s. IMDb score for the top 20 profitable movies.

4.2. Predictive Models

4.2.1. Decision Trees

The decision tree method reaches 84.90% accuracy over 89,757 instances, and 93.61% accuracy over 5,030 instances with budget data. The confusion matrices are shown in Table 10 and 11.

a	b	c	d	← classified as	Precision
2320	401	64	3	a = excellent	0.905
195	43274	4535	193	b = good	0.866
44	6018	25919	722	c = ok	0.820
4	273	1100	4692	d = bad	0.836

Table 10. Confusion Matrix for Decision Tree on movies.csv

a	b	c	d	← classified as	Precision
325	32	0	0	a = excellent	0.948
18	3409	122	3	b = good	0.955
0	123	921	6	c = ok	0.873
0	5	12	54	d = bad	0.857

Table 11. Confusion Matrix for Decision Tree on movies_budget.csv

The accuracy suggests that our movie features have great potential in predicting the movie's success with decision trees. After introducing the variable of budget, the accuracy of the model significantly increased, indicating that budget is an important variable for predicting the success of movies. At the same time, we should notice the limitations of the comparison. Because the use of budget variables reduced the size of our data and enlarge the unbalance among different classes, the result may not precisely reflect the effectiveness of the features.

4.2.2. Random Forest

To further investigate our features and improve the models, the data has been trained and tested using random forest models, which achieves 87.71% accuracy on movies.csv and 95.53% on movies_budget.csv, both of which perform better than the decision tree models. It suggests that the random forest models are more promising for the real-world application of movie success prediction.

The feature importance given by the random forest models can be found in Tables 12 and 13. Both tables show that `director_rating` is the most important variable in movie success prediction, followed by `actor_ratings`. It is a piece of evidence that the popularity and recognition of directors and cast are the key factors determining the success of a movie, and to some extent, the attractiveness of the director contributes more than that of an actor/actress. On the contrary, the region and genres tend to be less important for predicting the success of movies. The table also indicates that budget is a useful variable for prediction and this is consistent with the result that the model accuracy got improved.

Nevertheless, there is also a problem that should be noticed. In Table 12, the percentage importance of `numVotes` is 6.85%, while for the other set of experiments, it rises to 14.71%. A potential explanation might be the smaller data size and unbalanced classes as mentioned in the previous results of decision trees. Therefore, in the future, we need to enlarge our dataset and re-examine this variable carefully.

feature	importance
<code>director_rating</code>	0.355197
<code>actor_rating</code>	0.250999
<code>actor_movies</code>	0.076365
<code>numVotes</code>	0.068451
<code>director_movies</code>	0.063738
<code>year</code>	0.049818
<code>runtimeMinutes</code>	0.048448
<code>region</code>	0.028171
<code>genre_1</code>	0.023117

genre_2	0.020811
genre_3	0.013846
isAdult	0.001040

Table 12. Feature importance given by random forest on movies.csv

feature	importance
director_rating	0.318954
actor_rating	0.182916
numVotes	0.147111
budget	0.063712
runtimeMinutes	0.059923
year	0.048244
director_movies	0.045451
actor_movies	0.043695
genre_2	0.029356
genre_3	0.024169
genre_1	0.023513
region	0.012958

Table 13. Feature importance given by random forest on movies_budget.csv

5. Conclusion

In the past few decades, the number of movies around the world has increased rapidly, but the average quality of movies keeps decreasing. The movie industry has entered an era of prosperity, but challenges are emerging at the same. As it becomes more necessary for producers and investors to address the ways to make successful movies, data mining is a promising method to discover underlying patterns from historical data and inspire strategy-making for movie production.

For our first research question, by adopting IMDb movie data, this paper has investigated and discussed several movie-based factors that can affect the reputation of the movies. It shows the distribution of movies in terms of time, regions, and genres, and their relationship with the average movie ratings has also been discussed. The paper proposed a ranking method for directors based on rating information from IMDb. Compared with the historical and critical ranking list of directors, this method attached more importance to animation movies and Hollywood blockbuster directors. In terms of research works, this paper has provided comprehensive and informative visualization about the movie-based data. From the perspective of real-world application, it is supposed to help indicate who and what will be more likely to make phenomenal movies in the coming years.

To answer our second research question, this paper has conducted experiments on two types of machine learning models which have given impressive performance on movie

success prediction using a few features. The results show that random forest models yield higher accuracy (87.71%) than that of decision trees on performing 4-class classification tasks using our basic dataset. With introducing the variable of movie budget, the prediction accuracy increased to 95.53%. The feature importance given by the random forest models revealed that the director and cast ratings are the most important determinants of predicting the success of movies. The ratings are derived from calculating the average rating received before the current movie. The experiment has also shown that genres and regions may not be necessary for predicting the movie's success. Given the performance of the models, we hope our methods of feature and model construction could contribute to research works as well as real-world applications.

Apart from the current achievements, we have also noticed the limitation of our study due to the relatively small data size of the movie budget and the unbalanced classes. In the future, more valid data is supposed to be involved to examine our research, and more advanced machine learning techniques are to be explored to further improve the prediction performance.

References

- Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B. (2017). Movie success prediction using data mining. *In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE.
- Albert, S. (1998). Movie stars and the distribution of financially successful films in the motion picture industry. *Journal of Cultural Economics*, 22(4), 249-270.
- Alspector, J., Kolcz, A., & Karunanithi, N. (1998, May). Comparing feature-based and clique-based user models for movie selection. *In Proceedings of the third ACM conference on Digital libraries* (pp. 11-18).
- Asad, K. I., Ahmed, T., & Rahman, M. S. (2012, May). Movie popularity classification based on inherent movie attributes using C4. 5, PART and correlation coefficient. *In 2012 International Conference on Informatics, Electronics & Vision (ICIEV)* (pp. 747-752). IEEE.
- Bramer, M.A. (2013). Principles of data mining. *Springer*.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

- Chang, B. H., & Ki, E. J. (2005). Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics*, 18(4), 247-269.
- Chiu, C. L., Chiu, J. L., Ho, H. C., & Zu, Z. (2020). American Film Industry Challenges in China: Before and During COVID-19 Outbreak. *Management Review: An International Journal*, 15(2), 118-149.
- Deldjoo, Y., Elahi, M., & Cremonesi, P. (2016, September). Using Visual Features and Latent Factors for Movie Recommendation. In *CBRecSys@ RecSys* (pp. 15-18).
- Deldjoo, Y., Elahi, M., Quadrana, M., & Cremonesi, P. (2018). Using visual features based on MPEG-7 and deep learning for movie recommendation. *International journal of multimedia information retrieval*, 7(4), 207-219.
- Elberse, A. (2007). The power of stars: Do star actors drive the success of movies?. *Journal of marketing*, 71(4), 102-120.
- Ericson, J., & Grodman, J. (2013). A predictor for movie success. CS229, Stanford University.
- Fang, J., & Xiong, W. (2020, March). Impact of digital technology and internet to film industry. In *IOP Conference Series: Materials Science and Engineering* (Vol. 768, No. 7, p. 072112). IOP Publishing.

- Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 9(5), 374-385.
- Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, 33(3), 874-903.
- Lash, M., Fu, S., Wang, S., & Zhao, K. (2015, March). Early prediction of movie success—what, who, and when. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 345-349). Springer, Cham.
- Lee, K., Park, J., Kim, I., & Choi, Y. (2018). Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*, 20(3), 577-588.
- Litman, B. R., & Ahn, H. (1998). Predicting financial success of motion pictures: The early '90s experience. In B. R. Litman (Ed.), *Motion picture mega-industry* (pp. 172–197).
- Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018, April). A Data mining Technique for Analyzing and Predicting the success of Movie. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012100). IOP Publishing.
- Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of marketing*, 74(1), 108-121.

Ozili, P. K., & Arun, T. (2020). Spillover of COVID-19: impact on the Global Economy.

Available at SSRN 3562570.

Persson, K. (2015). Predicting movie ratings: A comparative study on random forests and support vector machines.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*. 33, 1-2.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3), 660-674.

Simonoff, J. S., & Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3), 15-24.

Stedman, A., & McNary, D. (2020). Morbius, Ghostbusters Sequel, and more Sony Movies

Pushed Back to 2021. *Variety*. Accessed at:

<https://variety.com/2020/film/news/morbius-ghostbusters-afterlife-uncharted-pushed-back-2021-sony-1203549616/> [June 11, 2020]

Appendix I

Description of IMDb Datasets

dataset files can be accessed and downloaded from <https://datasets.imdbws.com/>. The data is refreshed daily. Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A ‘N’ is used to denote that a particular field is missing or null for that title/name. The available datasets are as follows:

title.akas.tsv.gz - Contains the following information for titles:

- titleId (string) - a tconst, an alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- title (string) – the localized title
- region (string) - the region for this version of the title
- language (string) - the language of the title
- types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
- attributes (array) - Additional terms to describe this alternative title, not enumerated
- isOriginalTitle (boolean) – 0: not original title; 1: original title

title.basics.tsv.gz - Contains the following information for titles:

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- originalTitle (string) - original title, in the original language
- isAdult (boolean) - 0: non-adult title; 1: adult title
- startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
- endYear (YYYY) – TV Series end year. ‘N’ for all other title types
- runtimeMinutes – primary runtime of the title, in minutes
- genres (string array) – includes up to three genres associated with the title

title.crew.tsv.gz – Contains the director and writer information for all the titles in IMDb.

Fields include:

- tconst (string) - alphanumeric unique identifier of the title
- directors (array of nconsts) - director(s) of the given title
- writers (array of nconsts) – writer(s) of the given title

title.episode.tsv.gz – Contains the tv episode information. Fields include:

- tconst (string) - alphanumeric identifier of episode
- parentTconst (string) - alphanumeric identifier of the parent TV Series
- seasonNumber (integer) – season number the episode belongs to
- episodeNumber (integer) – episode number of the tconst in the TV series

title.principals.tsv.gz – Contains the principal cast/crew for titles

- tconst (string) - alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- nconst (string) - alphanumeric unique identifier of the name/person
- category (string) - the category of job that person was in
- job (string) - the specific job title if applicable, else '\N'
- characters (string) - the name of the character played if applicable, else '\N'

title.ratings.tsv.gz – Contains the IMDb rating and votes information for titles

- tconst (string) - alphanumeric unique identifier of the title
- averageRating – weighted average of all the individual user ratings
- numVotes - number of votes the title has received

name.basics.tsv.gz – Contains the following information for names:

- nconst (string) - alphanumeric unique identifier of the name/person
- primaryName (string)– name by which the person is most often credited
- birthYear – in YYYY format
- deathYear – in YYYY format if applicable, else '\N'
- primaryProfession (array of strings)– the top-3 professions of the person
- knownForTitles (array of tconsts) – titles the person is known for

Appendix II

Region Abbreviations

Abbr.	Region
AR	Argentina
AT	Austria
AU	Australia
BR	Brazil
CA	Canada
CN	China
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GB	United Kingdom
GR	Greece
HK	Hong Kong
HU	Hungary
IN	India
IR	Iran
IT	Italy
JP	Japan
KR	South Korea
MX	Mexico