Data science in practice
MGT-415

Project:
TMDB 5000 film rating forecasts

Nacer Ammor
Amor Chabchoub
Ridha Chahed
Nour Ghribi
Haitham Hammami

Lausanne, 10th May 2020

Contents

1 Introduction

The Movie Database (TMDb) is an open and community-based online database on film and television. The TMDB provides a large amount of information on films such as actors, directors, screenwriters and all the people and companies involved in the development of a film, TV movie. The project was founded in 2008 by Travis Bell to collect film posters. Access to public information is free of charge. It looks like a specialized version of Wikipedia where everything is editable but very specialized around the only data on films, TV and actors. When it was launched in 2009, the project started as a simple image sharing community, the initial database was a donation from the free Open Media Database (omdb) project, everything that has been added and edited since then is due to the large and very active TMDB community [1].
Today, TMDB, is one of the most active film databases on the Internet and is currently used by millions of people each month as well as many popular media centers around the world to enrich the user experience. The database contains 543,523 movies, 92,367 television shows and has a community of approximately 1,650,750 people [2].

1.1 Goal and expectations

The production of a film is an expensive and risky investment. Indeed, some great films with production costs exceeding $100 million can still fail, thus it can either be a great success and generate a lot of money, or it can flop and cost the production company a fortune. Therefore, predicting whether a film will be a success is a critical issue for the film industry. Many factors, such as experienced directors, famous actors, are in most cases crucial to produce a successful film, which may lead to the expected revenues, however, this does not always guarantee the success of the film with the general public and therefore a high Tmdb score. From the numerous data on films, the important factors that make a film more successful than others would be interesting to understand. Thus, the objective of this study is to understand what variables define the success of films, a successful movie being a movie with higher revenue and score.
Finally, the goal of this project is to apply the analytical data science tools to the TMDB 5000 film data. In this study, we will use the theory we learned during our studies at EPFL as well as the theory of the course: *Data Science in Practice*. The results can help film companies understand the secret behind the creation of a blockbuster film.

2 Data Analysis

The project begins with an analysis of the data. A clear understanding of the subject as well as of the data, is crucial for drawing relevant conclusions and developing an appropriate model. The data [3] is composed of 26 variables for 5043 films, covering 100 years in 66

countries. There are 52'234 crew members and 54'201 actors. A thorough analysis of the data is essential in order to be able to draw relevant conclusions and plan an effective model.

2.1 Description of the data

The dataset contains information about 5000 films grouped into 2 csv files:

- *tmdb_5000_movies.csv*: Provides film-specific information such as score, title, release date, genre, revenue etc.
- *tmdb_5000_credits.csv*: Includes information about the actors and crew of each movie.

The data set contains 24 variables that we can divide into two categories: qualitative variables (17) and quantitative variables (7).

- **Qualitative variables**: Includes 17 variables in the data sets and designates discrete units. It is used to label variables that have no quantitative value and can be found in both csv files.
  -In the **tmdb_5000_movies.csv** file the qualitative value include the title and the id of the film, the genre of the movie (Sci-Fi, Family, Horror, Comedy, Action,..), the links of the homepage of the film, the language of the film (English, Arabic, Chinese, French,..),the name of the production company (Walt disney, Columbia,..),the status of the film(Released or post production), the tagline of the movie, a small overview as well as keywords describing the film's plot.
  - The **tmdb_5000_credits.csv** file Contains 4 variables, two of which are also in the first file and which are the title and the unique ID of the film, the two other variables are in Json format: the **casting** which contains the names of the main actors with their gender (1 if male 2 if female, 0 if undefined) And the **crew**, which contains all the members of the film crew, including the name and gender of each crew member, as well as their function and the field in which they work.

The quantitative variables are also divided into 2 sub-parts and are only found in the *tmdb_5000_movies.csv* file:

- **Discrete variables:** Consists of 3 variables in the data set and are numerical data that can only take a certain set of values. The discrete variables are: the number of votes for each film in tmdb, the release date and the duration of the film.
- **Continuous variables :** Consists of 4 variables in the data set and represents measurements or quantities that can take any numerical value. The continuous variables are: film revenues in dollars, film budget in dollars, popularity, as well as average film votes.

2.2 Data cleaning

Unfortunately, the dataset is not perfect, and some problems prevent us from being able to work with it directly.

- The biggest issue with this dataset is the *json* format (JavaScript Object Notation), it is a syntax for storing and exchanging data between two computers that looks like a dictionary pair (key:value) embedded in a string. Therefore the first task was to browse all the data and replace the *json* character in *string* format using the *json.loads()* method. The columns that create this problem are the genre, keywords, countries of production, production companies, languages spoken in the film file and the crew and actors columns in the credits file.

- We also need to deal with the null values, five features cause this problem: home page, release date, overview, execution time and slogan. The homepage and the tagline contain the most null values and are not relevant for our study, therefore the entire columns were discarded. This leaves the release date, the run-time and the overview, which contain three, two and one null value respectively. For run-time, the two null values are replaced by the average value of the duration of the films. Then, for the overview, the null value can easily be fixed, we simply put unspecified instead of null in order to have a category for it and thus avoid any problem downstream. Finally, for the release date, we chose to drop this value.

- The last problem is that of zero values. There are five features that cause this problem: budget, revenue, run time, average vote and number of votes. Since our primary objective is to forecast revenue, we are going to drop the zero value for revenue and budgets. We have decided to drop everything under $1,000, there are cases where the budget is $5 but in reality it is worth $5 million; however, we opted to drop them instead of replacing them one by one as they only represent a small part of our dataset. Finally, for the other features , we decided to drop them.

- In addition, there are some movies that have the same title, therefore by adding the year of release, we make the title unique and remedy this problem.

- Also for convenience, in the release date, it is preferable to change the year, month, day format initially in the same column to that of each in a different column.

The data are now ready to be properly observed in order to extract as much information as possible, allowing them to be subjected to statistical tests.

2.3 Exploratory Data Analysis (EDA)

Before getting to the core of the subject, it is a good idea to first draw the correlation matrix in order to assess the dependency between the different variables (Figure 1). Each cell in the table shows the correlation between two variables.

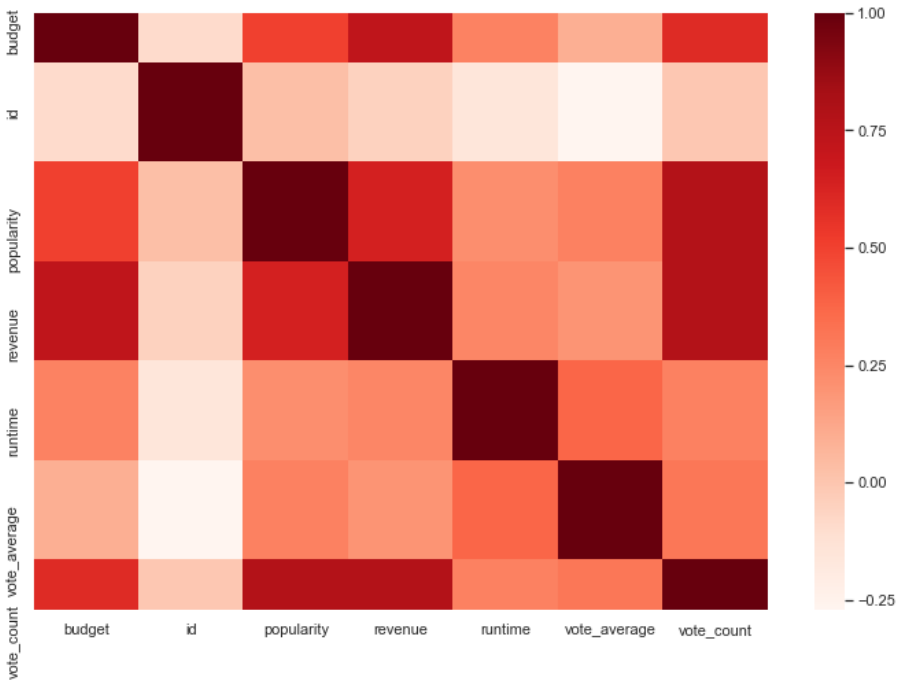

Figure 1 – Correlation matrix

Statistics summary of movies quantitative features is shown in figure 2.

	budget	id	popularity	revenue	runtime	vote_average	vote_count
count	4.803000e+03	4803.000000	4803.000000	4.803000e+03	4801.000000	4803.000000	4803.000000
mean	2.904504e+07	57165.484281	21.492301	8.226064e+07	106.875859	6.092172	690.217989
std	4.072239e+07	88694.614033	31.816650	1.628571e+08	22.611935	1.194612	1234.585891
min	0.000000e+00	5.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	7.900000e+05	9014.500000	4.668070	0.000000e+00	94.000000	5.600000	54.000000
50%	1.500000e+07	14629.000000	12.921594	1.917000e+07	103.000000	6.200000	235.000000
75%	4.000000e+07	58610.500000	28.313505	9.291719e+07	118.000000	6.800000	737.000000
max	3.800000e+08	459488.000000	875.581305	2.787965e+09	338.000000	10.000000	13752.000000

Figure 2 – Features statistics

From the correlation matrix, we observe that the number of votes variable is highly correlated with popularity as well as revenue with a moderate correlation with budget, while the popularity variable is moderately correlated with revenue and slightly less correlated with budget. In addition, the budget and revenue variables are highly correlated. While the run time and average number of votes variables do not seem to have a dependency relationship with the other characteristics.

Then, we begin our data exploration by comparing the genre of each film and its recurrence, we find that the dramatic genre is the most common; it represents 25.18% of all genres, followed by the comedy and thriller and action genres which represent 18.89%, 12.65% and 9.8% respectively. The least recurrent genres are foreign films, westerns and documentaries with a percentage of 2%.
On the other hand, action and adventure films, followed by dramas and thrillers are the most revenue-generating categories, while foreign films, westerns and documentaries are the least revenue-generating categories. Figure 3 shows the revenues generated for each film category between 1990 and 2015, where we see a net increase in revenues for each category.

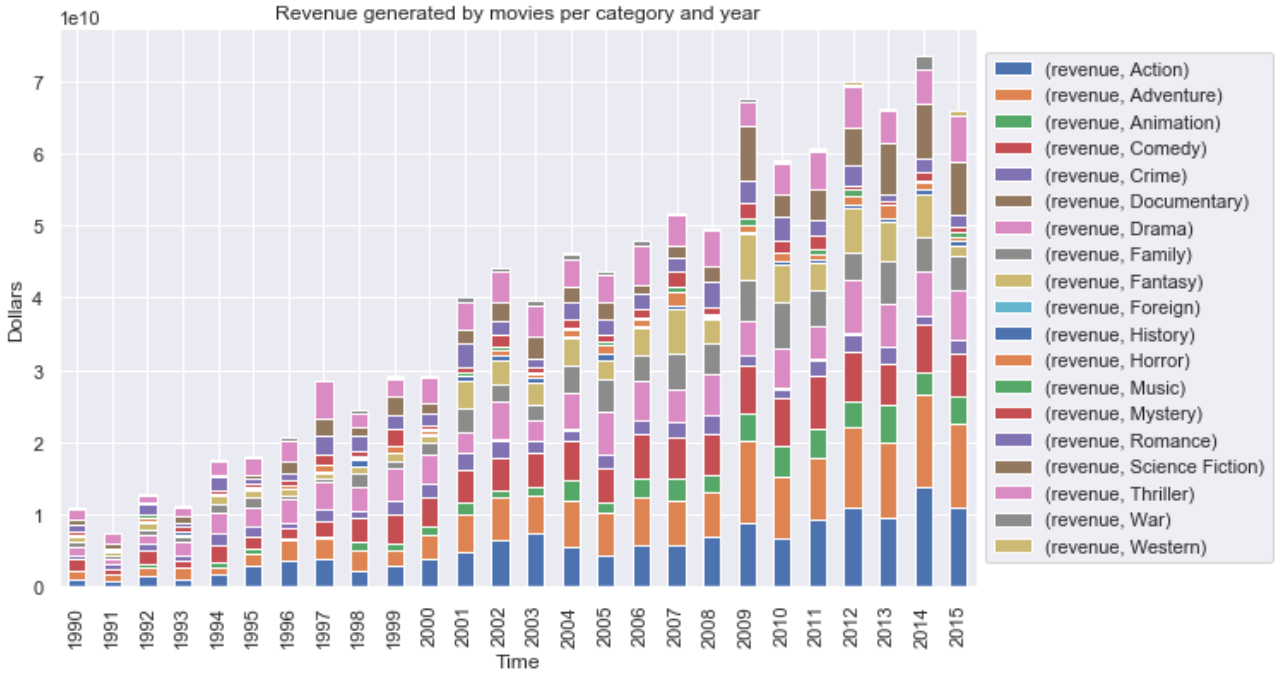

Figure 3 – the total revenues generated for each film category during 1990-2015

However, this evolution must be nuanced, as the economical phenomenon of inflation must be took into account. The effect of inflation refers to the general and sustainable increase of goods and services prices in the economy. To only take the movie industry as example in 1939, the Americans paid 0.25 $ to watch the blockbuster Gone With the Wind while they paid 5 $ to see Titanic in 1997 and 9 $ for Avatar in 2009. Therefore, in order to really see if the film industry is generating more revenue, it is wise to contextualize the revenues in such a way that the money generated can be more fairly compared, this is done using the inflation increase formula shown below:

$$Rise\ in\ inflation = \frac{CPI_{ref,year}}{CPI_{fin,year}} \tag{1}$$

with:

- $CPI_{ref,year}$: The Consumer Price Index for the reference year
- $CPI_{fin,year}$: The Consumer Price Index for the final year(here is 2015)

This can be seen clearly in figure 4, the orange curve is the curve of the inflation-adjusted revenue while the blue curve is the curve without inflation adjustment, it can be seen that the closer we get to the final time (here 2015) the more the two curves overlap while the further we get from 2020 the more the curves differ. moreover we notice that outliers emerge, as Gone with the wild in 1939 and Star wars V in 1980 or Avatar in 2009.

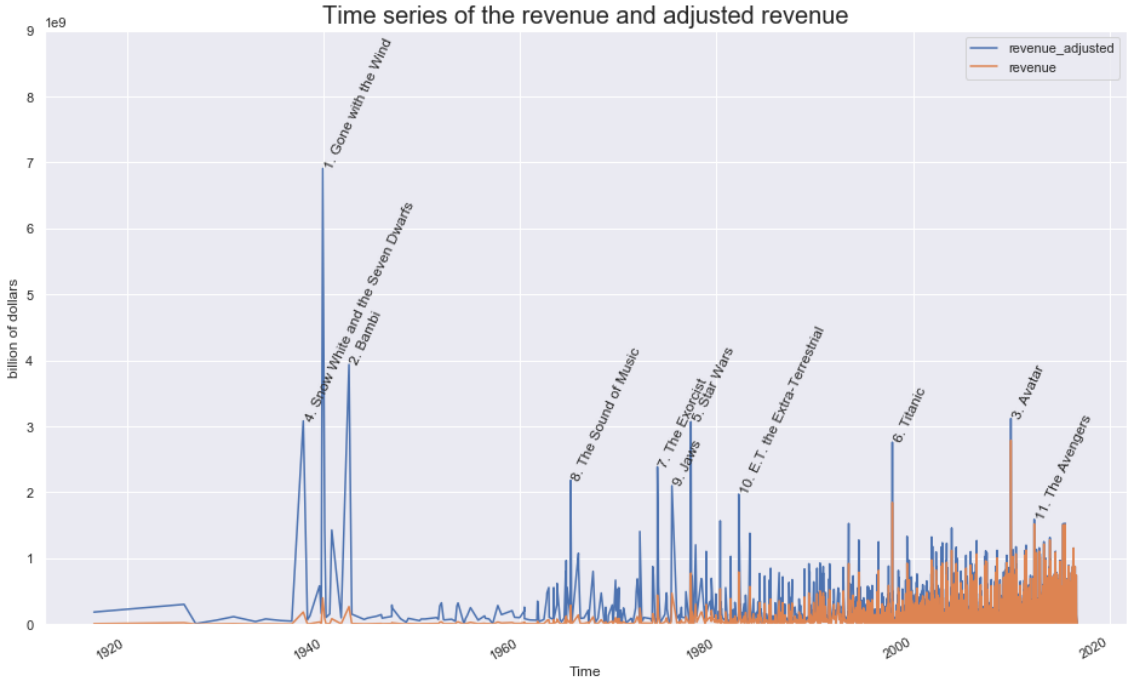

Figure 4 – Time series of the revenue generated during 1990-2015 (the blue curve is adjusted to inflation while the orange is not)

We are surprised to find out that the movie Gone With the Wind (1939) became the highest grossing movie with almost 7 billion dollars followed by the more surprising Walt Disney movie Bambi (1942) with 4 billions while the known Avatar is (only) at 3 billions. So how can we explain these differences ? Well actually what is not shown in this graph is that the re-releases of the movie. Gone with the Wind was a big hit at the time of its release but only made 500 millions in current dollars, it was then re-released more than 8 times which explains its success. The takeaway message here is that movies don't only generate money at their release but continue also to do so years after that.

Furthermore, it would be interesting to consider the revenues of the films according to the budget with adjustment for inflation (Figure 5), this will allow us to see the net result of each film. If the income is higher than the budget, we speak of a profit, whereas if it is lower than the budget, we speak of a deficit. The dashed line represents the case where the income is equal to the budget, so all the films below this curve report a loss, and all the films above this curve report a gain, so the further the film is from this curve, the greater its gain/loss.

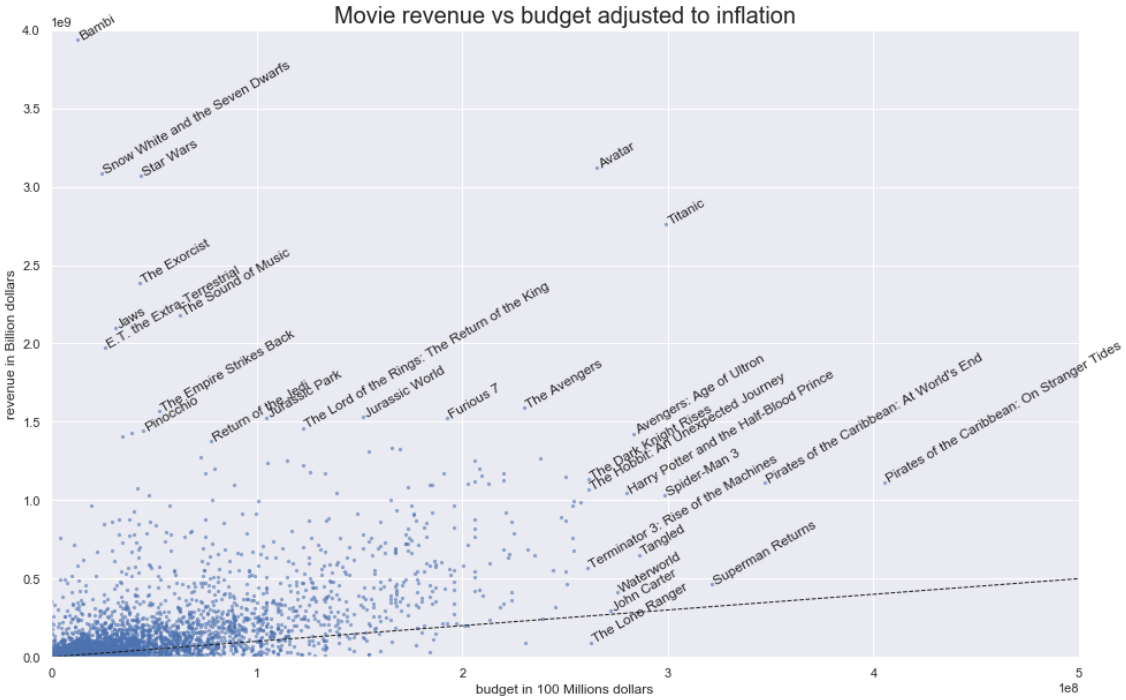

Figure 5 – Film revenues versus budget with adjustment for inflation (the dashed line represents the case where the budget is equal to the revenues)

In addition, it can be seen from the graph that the majority of films generate a profitable result (above the dashed line) and in general, the higher the budget, the higher the revenue, in accordance with the correlation matrix (figure 1). Nevertheless, there are films that have a very big budget such as *The lone ranger*, for example, but are suffering a loss and are far from breaking even, furthermore there are many films that have a small budget but generate huge revenues such as the films of *bambi*, *Snow white and the seven dwarves*, these latter can be considered as outliers due to their low numbers compared to the rest of the dataset.

Furthermore, an interesting aspect to study would be the relationship between a film's popularity with the public and its respective budget (figure 6). The popularity is directly calculated by TMDB by taking into account several factors which are: The number of votes for the day, the number of views for the day, the number of users who marked it as a "favorite"

for the day, the number of users who added it to their watchlist for the day, the release date, the total number of votes, the score of the previous days.

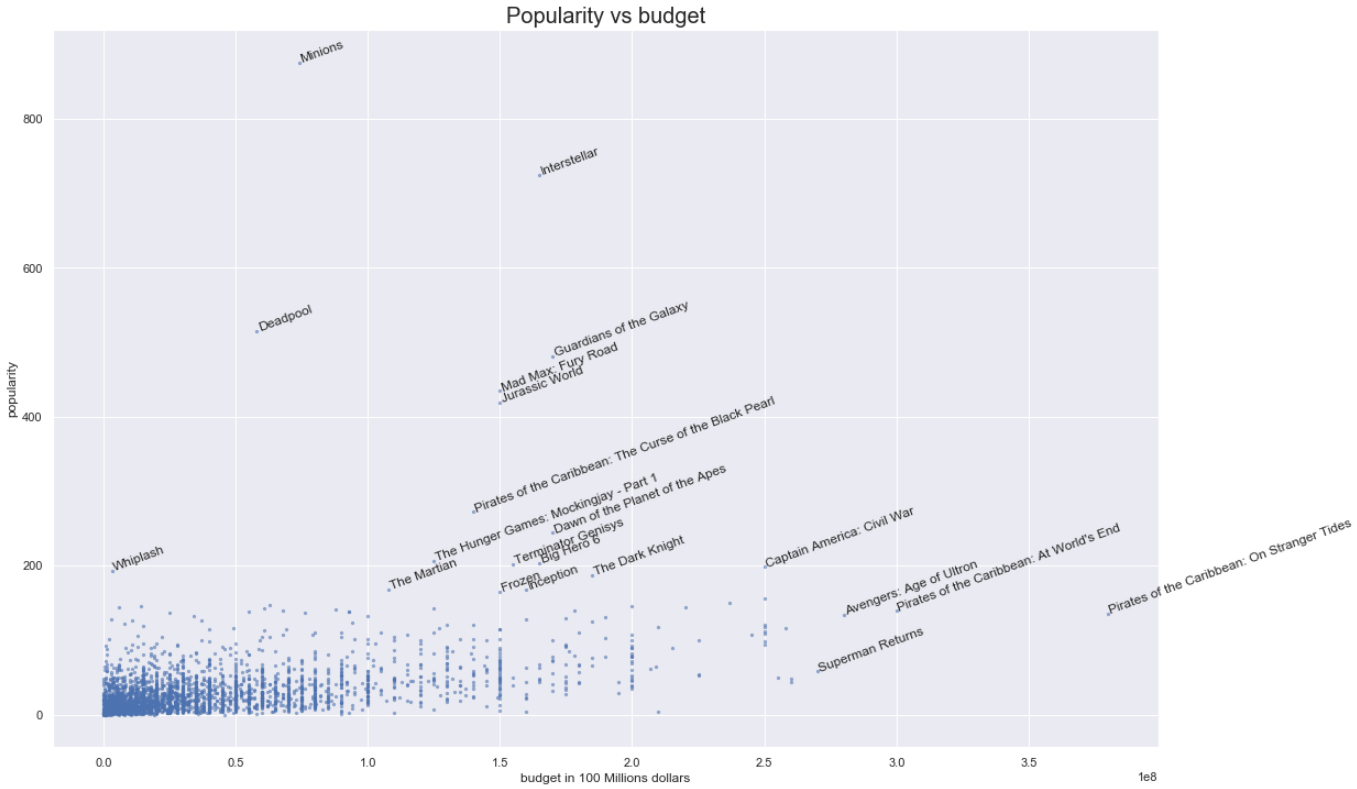

Figure 6 – Film popularity (TMDB) versus budget

On the one hand, we can see on the graph that the majority of films are in the bottom left corner (low budget, low popularity), on the other hand some films have a medium budget but have had a great success with the audience as is the case for *The Minions* or*Interstellar*, while some films have a huge budget but have not been very successful with consumers as is the case for *Pirate Of The Carribean : On Stranger Tides* or *Superman Returns*, therefore all these films can be considered as outliers. Also, another important thing to see is that these outliers are different from the figure 5 (revenue vs. budget) which shows that a film that is very popular with consumers is not necessarily the one that generates the most money, contrary to what one might think.

Relation between average rating and Revenue : During the EDA we wanted to find if there is any relation between the average rating given by viewers and the revenues of the movie, but this relation is affected by another criteria which is the number of rating shown in the figure below :

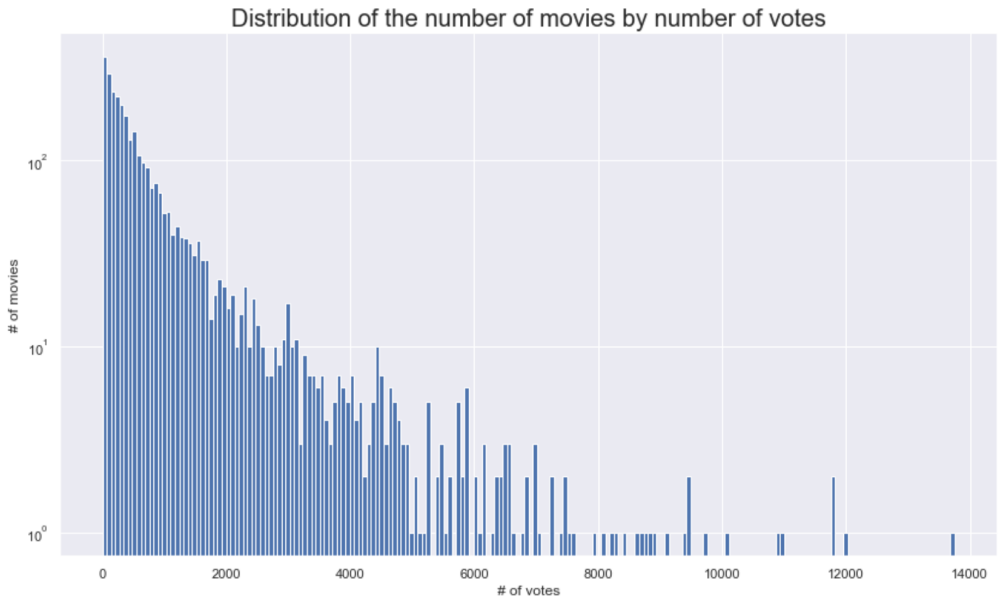

Figure 7 – Distribution of the number of movies by the number of votes)

As we can see in the figures that there are about 160 movies have no rating. The next step consists of dropping the movies without any viewers rating and analysing the relation between revenues and rating grouped by decade as we can see in the figure below:

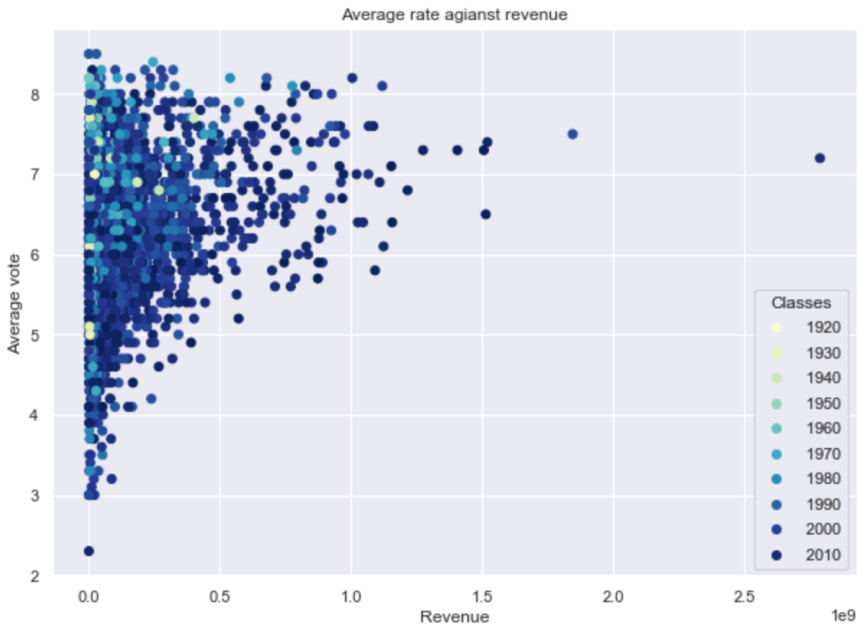

Figure 8 – Average rate vs revenue

It can be seen that for the majority of movies the revenues increases when the rating increase and we can observe one outlier with rating of 7.2 and a revenue about 2.8 billion dollars.

Production companies : In the data set we found 3539 production companies, but there is a big number of these companies that have only one movie, so we chose to work with the top 20 companies based on the number of movies produced, these 20 companies are kept for the model derivation. The figure below shows the number of movies produced and the cumulated revenues of the 20 top companies.

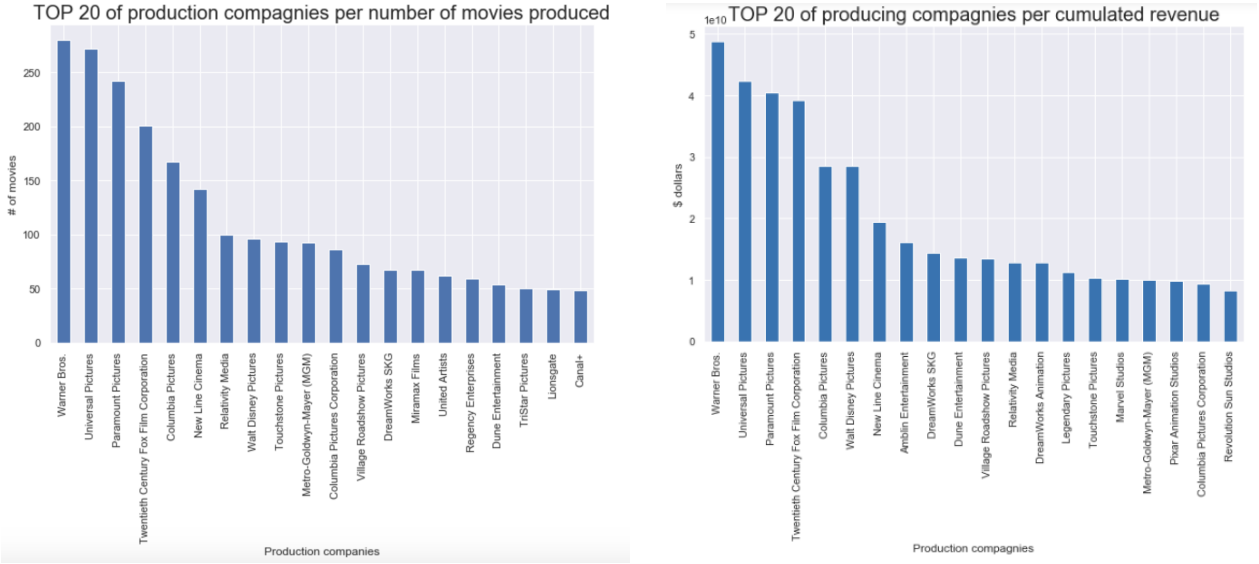

Figure 9 – Top 20 companies by number of movies produced and cumulated revenue

Actors and crew members influence on the movies revenue and rating: The next step is to check whether the actors affect the movies revenues or not, and to do so we need to modify the data set by adding a new features for the actors gender to find out its impact on the earnings as shown in the figure below :

Figure 10 – Mean revenue by gender

We can not conclude anything from this fact because the movies industry is dominated by men, according to the data set there are 16330 actor and 8326 actress. Another feature that can be mentioned is the number of actors per movie and its impact on revenues and ratings as shown below:

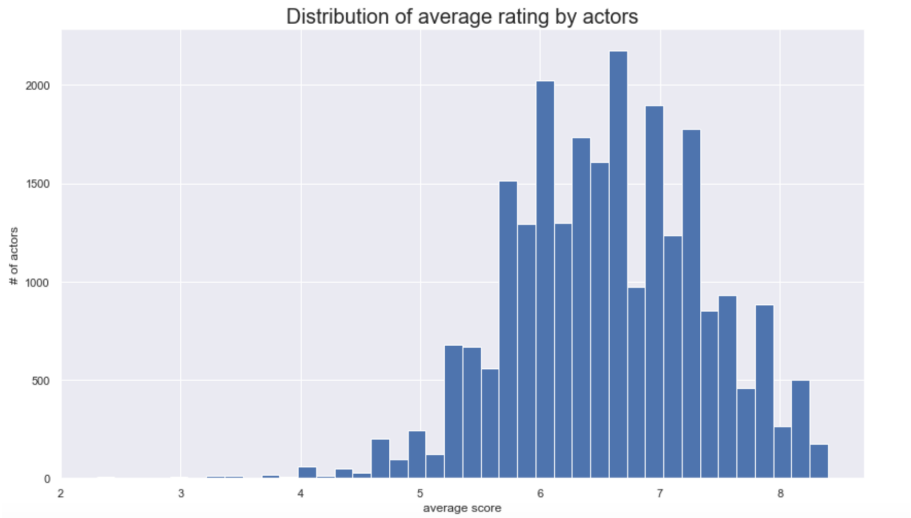

Figure 11 – Average rate vs revenue

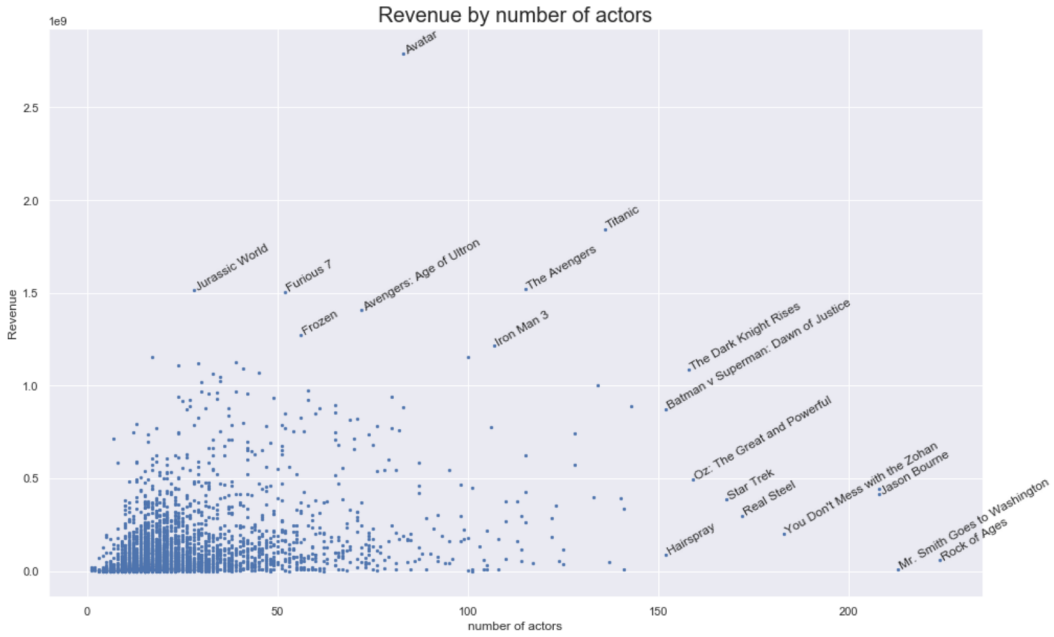

Figure 12 – Mean revenue by gender

As we can in fig.12 see above that the number of actors affects the rating of movies especially when the number of actors is around 6 and 7 main actors and according to fig.11 in most of the cases the revenue increases when the number of actors increases and we can observe that there are three outliers Avatar, Mr. Smith goes to Washington and Rock of ages. Then we tried to find out if the appearance of certain names of actors in the cast will effect the revenues, so we chose the list of 500 actors that appear the most. We found that there is a difference between the average revenue and the number of appearance, for example Samuel L. Jackson has the biggest number of movies but Stan Lee has the most important average revenue see figure .

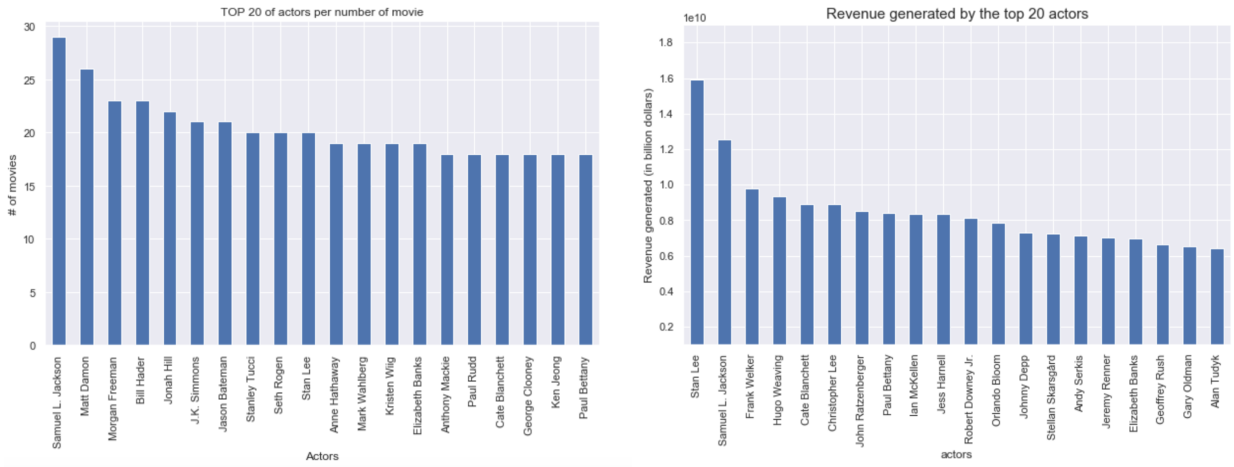

Figure 13 – Number of movies with the top 20 actors and the sum of revenues generated by top 20 actors

For the crew members we started by creating new features: crew member name ,job,department and gender. The main information that we can get from this analysis is the list of crew members and their jobs with the most important revenues as shown in the table below :

Crew member job	**Cruw member name**	**Revenue [Billion dollars]**
Compositors	Brian N. Bentley	15.70
Orchestrator	Kevin Kaska	15.20
Original Music Composer	Hans Zimmer	14.61
Executive Producer	Stan Lee	14.40
Casting	Sarah Finn	1.221
Original Music Composer	John Williams	11.37
Producer	Kevin Feige	9.97
Foley	Dan O'Connell	9.83
Original Music Composer	Danny Elfman	9.64
Executive Producer	Louis D'Esposito	9.51

Table 1

3 Model derivation

3.1 Data Preprocessing

To predict movies revenues we need first to make some changes on the initial data set so it can be used by the chosen algorithms, after data cleaning we did further data preprocessing such as:

Restriction of the actors and production companies features: As mentioned in the previous part we took into consideration only the 500 most appeared actors and the top 20 production companies.

Creating dummy variables: Since our data set cannot be used by our algorithms, we use one-hot encoding method and create the different categorical variables from the features original language, country of production,production companies, actors, crew names and genre.

Unused features: Before starting the revenue and rating prediction we drop the features revenue, number of votes, Rating average and popularity because in the first place we will get these information after the release of the movie and what we are trying to do here is to predict whether the movie is successful or not also if a movie have an important revenue it means that it's popular which leads to an increase in the number of votes and rating.

Standardization: As a last step before finding the right model to fit our data for prediction we will standardize our data so that all features will have same weight and same effect on the prediction.

Naturally, we will divide our data into two sub-samples : one is used to train the model and the other to test the model using prediction.

3.2 Prediction model

We will try to derive a model to predict movies revenues and movies ratings. We will be working with two kinds of datasets :

- The films dataset without any features about crew members or cast.
- The films dataset plus new columns of crew and cast. (1570 features)

We will try to fit 3 different models : Ordinary Least Squares (OLS) regression, Ridge regression and XGboost, and try to find the most suitable by evaluating their performance.

Regression (OLS) treats all variables equally, whereas the ridge regression model is a regression model that has regularization and can rescale its weights by a hyperparameter alpha. We tune alpha to have the best score for this model. This said, we will use iterative feature reduction based on statistical significance when fitted on the data with crew and cast by the OLS model as this will reduce the number of parameters the model have to estimate and gauge the importance of the features.

The XGboost (eXtreme Gradient Boosting) model is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Indeed, such as random forest XGboost leverage the power of boosting but with the difference that now errors are minimized by gradient descent. It achieves superior results by mixing system optimization like parallelization and tree pruning with algorithmic enhancement such as regularization and cross-validation.

The metric we will be using for evaluating the different models will be the adjusted r squared as it is a good evaluation for regression models, it compares the descriptive power of the models. This metric compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. The baseline will be predicting all movies revenues as the mean of revenues in the training set. The best score you can get is 1 and the worst is minus infinity.

3.3 Prediction

3.3.1 Revenue prediction

Our baseline will be the mean predictor with a score of -0.004.

	OLS	**Ridge regression**	**XGboost**
Without Cast score	-0.458	0.513	0.595
With Cast score	-3106.53	-0.431	0.559

The OLS model performs very poorly on this kind of data as it has too many features. So we will use iterative feature reduction based on significance. We will compute the different p values of the features and retain only those below 0.05. This method uncovers how features affect the revenue, we discovered that actors like Mickie McGowan, Bob Bergen and David Oyelowo as well as crew members like Justine Baddeley, Broderick Johnson and Janusz Kamiński are people you don't really want in your movie. And the most advantageous features (the money makers) are the budget and number of actors. However this method didn't make our model better than our baseline, and performed very poorly. In this framework we can conclude that XGboost outperforms both Ridge Regression and OLS.

When we run XGboost with the data with cast, it dropped these feautures : 'Action', 'Canada', 'Canal+', 'China', 'Dune Entertainment', 'France', 'Horror', 'Italy', 'Japan', 'Miramax Films', 'Music', 'Spain', 'United States of America', 'Village oadshow Pictures', 'War', 'independent film', 'love', 'month' when run on the data with cast, so these parameters don't play a big role in the revenue of films. We can say that the countries of production in this set don't have a big influence on the revenue of films when we take into consideration the cast.

For the data without cast, we can see by Figure 14 the most important features: Budget, runtime, day of the year and year, Universal pictures and so on. Thus, the budget has a big influence on the success of a movie and its revenue which is logical. The runtime of a movie plays a role too, as if it is a short movie it wouldn't be so interesting and if it's too long it might get boring. We observe that we have some of the big production companies too, Universal Pictures, Twentieth Century Fox Film Corporation, Walt Disney Pictures and Warner Bros, as their movies tend to be most successful.

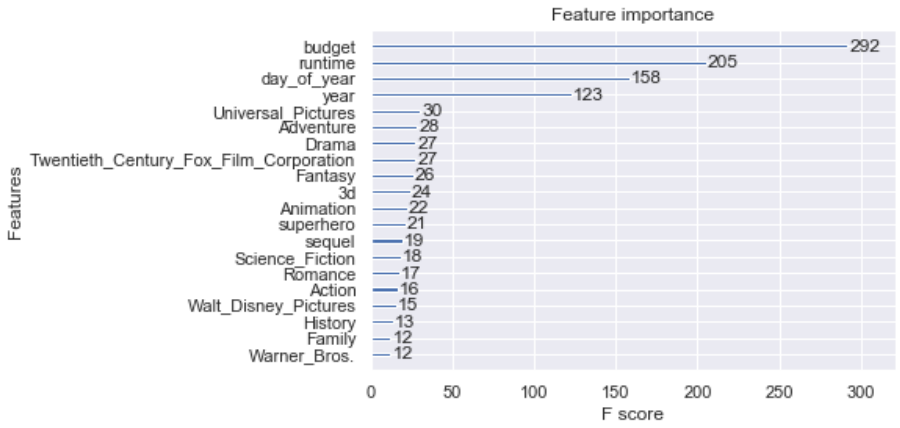

Figure 14 – Data without cast: Most important features by XGboost, the model retained 60 features.

In the second case (Figure 15), for the data with the cast features, we can see a change in the importance of features. Budget is still the most important, as it has a high correlation with revenue as seen in the EDA this can be expected. We still have runtime as second most important, and number of actors as a new feature. Taking into acount the cast members we can say that the budget and runtime of the film still has big importance and that the number of actors and the crew members plays a role too in the success and revenues of the movie. The big firms as Walt Disney Pictures still have significant weight in the prediction as well. Some Crew members have weight on the revenues too, these may be Executive Producers, special effects engineers, compositors or Orchestrators.
As a final note, we can dive more under the hood of XGboost and look at its decision tree (Figure 16). We can see how it uses the significance of the feature for the predictions.

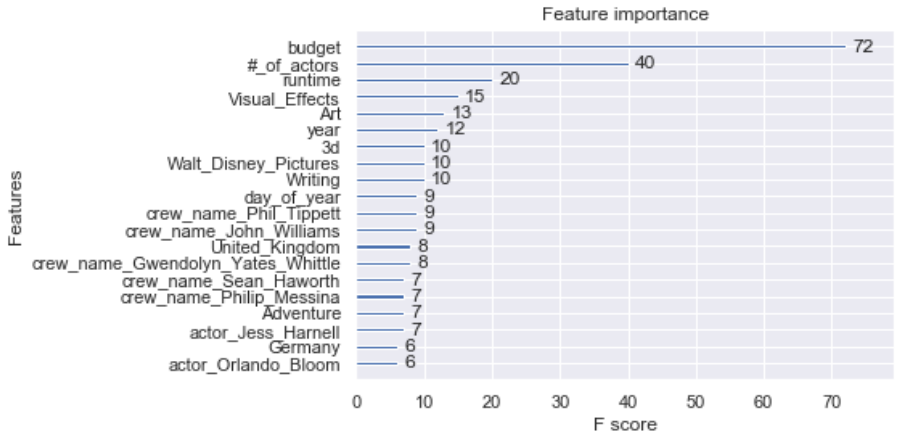

Figure 15 – Data with cast: Most important features by XGboost, it retained 183 features

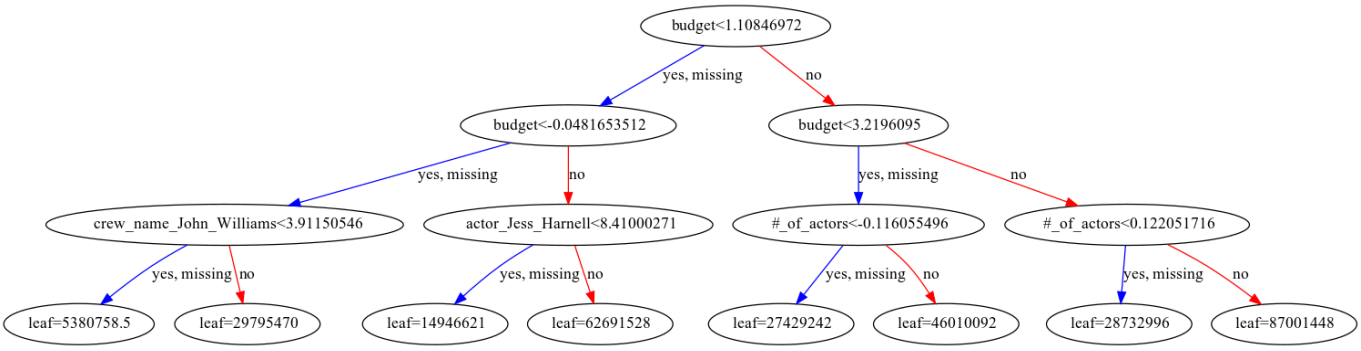

Figure 16 – Decision tree of XGboost

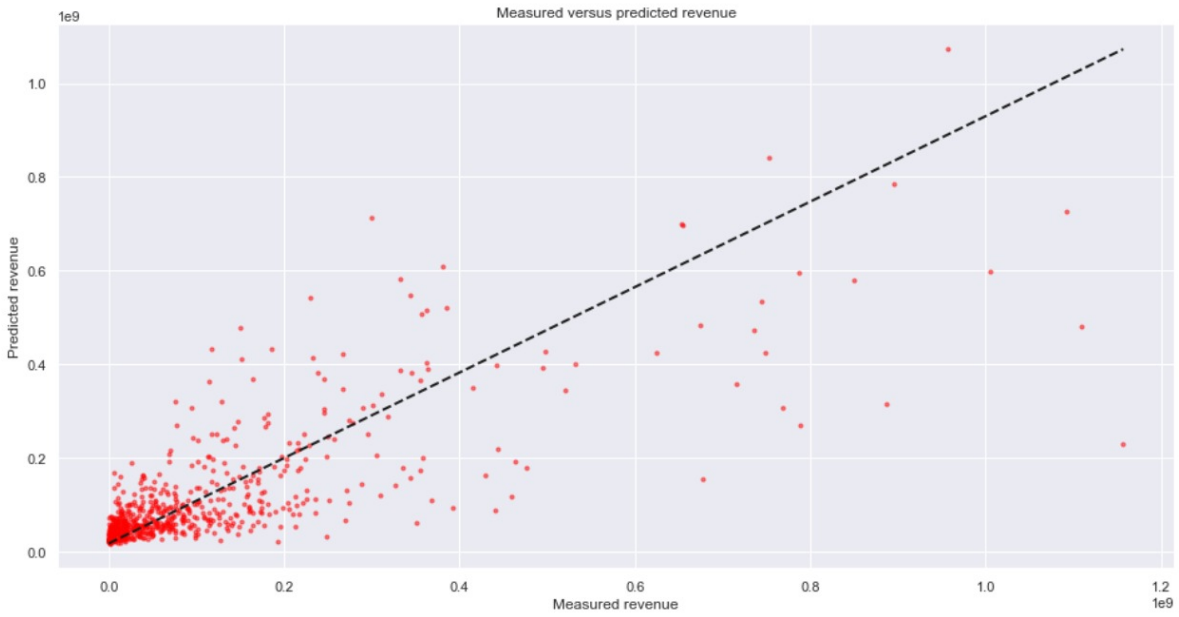

Figure 17 – Measured versus predicted revenue

3.3.2 Rating prediction

Our baseline will be the mean predictor with a score of -0.001.

	OLS	**Ridge regression**	**XGboost**
Without Cast score	-55.94	0.213	0.305
With Cast score	-286493.7	-0.962	0.33

For predicting the ratings, the best model to use is the XGboost model. We can see that it has the best score. Again, the OLS model is performing poorly and this is expected since it is not suited for this kind of regression problems. The ridge regression's score is better but it's outclassed by the baseline so we don't consider it.

For the Rating predictions without cast, we can see by Figure 18 the most important features: runtime, day of year, year ,budget, Action, Comedy, Thriller... And we find Universal pictures too. Again, we notice the importance of the budget on the ratings. However, the most important feature is the runtime now. We are trying to predict the ratings, so it is only normal that the runtime will be the most significant feature as explained before a long movie can get boring and a short one can be vain. The day of year significance can be inferred by social trends and internet hype over some movies, for example movies during the holidays are quite similair and therefore their reviews can be predicted more easily. The presence of

many different genres in the figure can be explained by the fact that each genre has its own fanbase and therefore their own distinct reviewing behaviour.

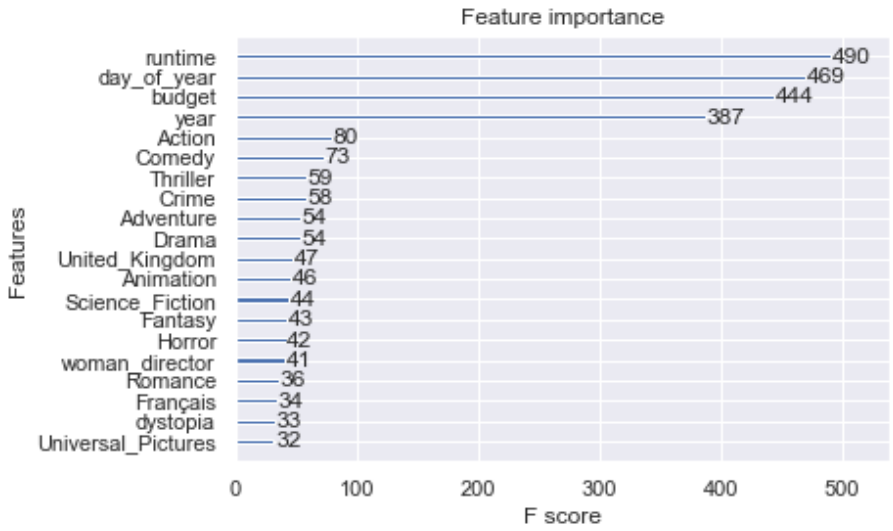

Figure 18 – Data without cast: Most important features by XGboost, it retained 74 features.

When adding the cast features to the data we can see a slight change in the importance of different features (Figure 19). The budget is the most important again, followed by the runtime and the number of actors. By now, we can conclude that the budget plays the most significant role in the success of a movie. The year and day of the year both affect the rating, and it can be reasoned as before that trends and holidays are the cause behind this result. What's also interesting is that the number of actors has an effect on the ratings too.
The decision tree of XGboost model fitted to this data is depicted in (Figure 20).

Figure 19 – Data with cast: Most important features by XGboost, it retained 164 features.

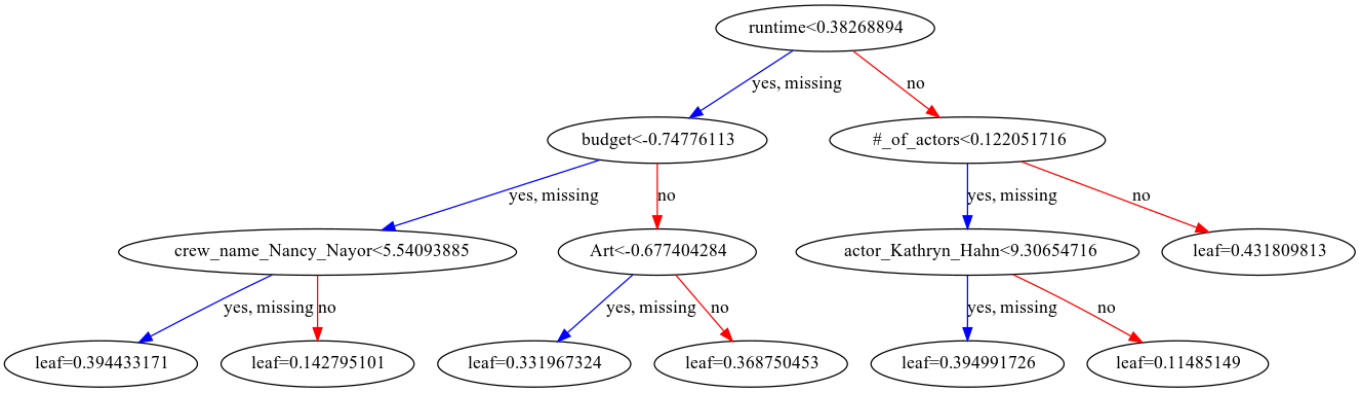

Figure 20 – Decision tree of XGboost

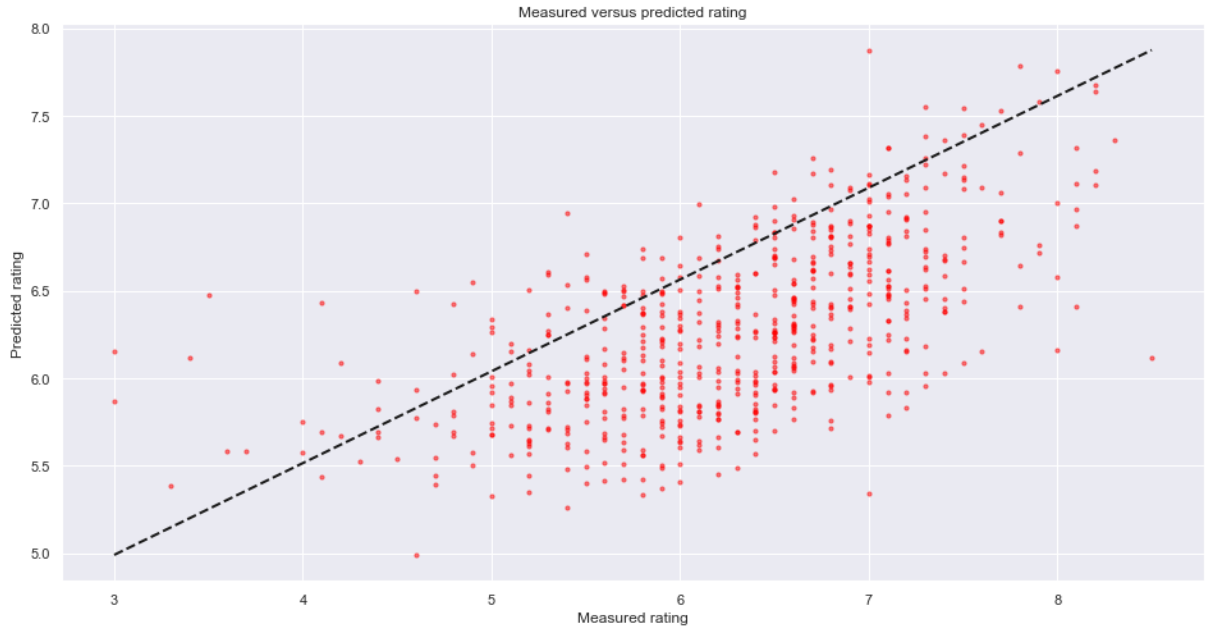

Figure 21 – Measured versus predicted rating

4 Conclusion

Working on the TMDb movies dataset gave us the opportunity to conduct studies and analysis on real life data about movies. From data cleaning, Exploratory Data Analysis to model derivation we investigated different aspects of movie production and success. We were faced with raw unprocessed data, yet derived a way to process and clean it, and were able to visualize the relation between the movies characteristics and what measures their success and in the process we learned about what makes a movie successful. For instance we analysed how important the budget is for deciding the movie success, obviously this comes as no surprise but here we were able to gauge its importance via statistical metrics. We experimented with ML models to find the best suitable one for our data and goals. Our goal from the beginning of this project was to derive a good enough model to predict the success of a movie beforehand. We may need to try to fit more models and refine our approach to get better results. We hope that in the future the film industry will rely more and more on Data Analysis and Data Science in general for improving the competitiveness of the market and reaching new levels of film production.

5 Reference

[1] https://www.themoviedb.org/about
[2] https://www.themoviedb.org/faq/general
[3] https://www.kaggle.com/tmdb/tmdb-movie-metadata