MGT-415 - DATA SCIENCE IN PRACTICE

THE PERFECT MOVIE RECIPE

HIEN LÊ, ZAFER KOCAOGLU, FRANCESCO MAIZZA,
ANITA MEZZETTI, NATALIIA SURIANINOVA

SPRING SEMESTER 2020

INTRODUCTION

For a long time, movies have made up one of the most dominant aspects of the entertainment and leisure industry, its global value being estimated at $136 million dollars in 2018 [7]. This not only tells us, the audience, that the film industry has become an indispensable part of our lives, but brings up along with it a series of puzzles to the people on the production line. Specifically, from a business perspective, what can be done to help this industry thrive and flourish with the continual changes in other aspects of life? While this big question requires a well-rounded analysis from various socio-economic and psychological perspectives, it can be looked at through the technical lens of big data and data science, in which patterns, if there are any, can be explored and from there conclusions drawn.

With this in mind, the current study sought to find out whether some already-available features could help determine the success of a movie, whereby success is defined by a movie's IMDb score and its gross-budget ratio - signifying its critical and commercial success, respectively. In this report, we will first explain in details our investigation into the trends in the cinema industry and how they could potentially later affect some of our findings, and then go in to the technicalities which include our choice of data (the Internet Movie Database data set obtained from data.world) to study as well as our initial findings of this data. We will then demonstrate our attempts in constructing several models that help predict the success indices, in order to find out whether the features provided for a film can actually forecast success, or lack thereof. Our implementation of these models helps answer yet another important question: which algorithm/model can help best predict a movie's IMDb score and/or its gross-budget ratio? Or rather, which algorithm should businesses apply to predict the success of their upcoming product?

1

THE MOVIE INDUSTRY

This chapter presents the background of our study, which centres around the movie industry in today's world. The industry has, since its first appearance, been accompanied by numerous changes and movements in the way not only producers but also the audience motivate and dictate the different modes of entertainment consumption. This dynamic is also the reason why we are interested in investigating this subject and seeking to see if there are factors that come into play when it comes to the success of a movie.

1.1 TRENDS IN THE MOVIE INDUSTRY

In order to be attractive to their audience, the cinema industry has been in a continuous path of innovation and change. The last decade is a clear example of how this business is able to radically transform itself. Indeed, in less than 10 years, while we have seen many rises and falls of digital technology, we have observed a revolution started by Netflix and its streaming platform, which in few years has destroyed the DVDs industry, and transformed the way people watch films as well as the way in which studios distribute their movies. This trend has also led new streaming services from Apple, Disney, Warner Bros and other major studios to become available soon [2]. Nevertheless, while the industry is trying to innovate itself from a distributional point of view, we cannot say the same about the contents of the movies. Indeed, creating a blockbuster movie is a business that thrives on the stories that are already in our pop culture, and as a matter of fact, in the last decade, films based on pre-existing ideas (remakes, adaptations, sequels or reboots) have resulted in far better revenue performances than original stories: of the top 30 most profitable films of the last 10 years, only one movie – Frozen – was an original story, not a sequel, reboot, or adaptation [13].

However, notwithstanding the fact that the stories are not new, authors have had to adapt their story line to a society that is radically different from the society in which the story was initially exposed. In fact, today's audience is more multi-ethnic and more conscious about the representations of various demographics, requiring films to reflect this cultural change, and thus resulting in fewer movies with a white male as main character and more with space for actors with diverse backgrounds, and a better LGBTQ representation. With this in mind and with a goal of maximising revenues, the movie industry is developing stories that give voice to marginalised groups, offering them an image to which they can relate. Increased representation in movies helps everyone leave the theatre feeling inspired and more than willing to pay for the next film.

All the above explain the reason why we are still amazed by the adventures and stories that the films narrate from 100 years. As a matter of fact, one of the most wonderful things about cinema is its ability capture from the reality the right topics and issues for inspiring and evoking emotions in audiences.

On the other hand, society has changed also in the way in which people communicate and spread information. Social networks have given each of us the unique opportunity to be in touch with the entire world. At the same time, they also allow firms, in particular major film studios, to collect real-time data, which are essential for the profitability of the movie. This has actually affected the outcomes of some movies. Two examples that have reached the headlines are *Sonic The Hedgehog* and *Suicide Squad*. For the former, after negative reactions on different social networks to the initial design of its title character, Paramount Pictures decided to completely reanimate it, thus delaying its release for months. While for the latter, it has been reported that 2016's Suicide Squad was significantly modified when social media users responded enthusiastically to the comic elements of the first trailer (originally, the film was a lot darker and was changed to have a lighter tone) [11].

From these trends, it is clear how important it is for the film industry to be constantly in touch with their audience, so as to determine the right characteristics for generating high and stable profits and the technological factors that are more likely to be positively accepted by the audience. It is also with this motivation in mind that this study decided to look into certain relevant factors to see whether they can help predict a movie's success, and to which extent these determinants actually "work". With this, it is important to define the notion of "success" for a movie, while looking at the data that we have available.

1.2 THE STUDY

The current study sought to find out, from the business perspective of a movie producer, ways in which one can maximise the critical as well as commercial success of a product based on a set of given information (features) related to the movie in preproduction. This way, businesses can actually make the decision of whether or not to go forward with producing a film, or at least change/revert a number of decisions on the film's characteristics. Furthermore, selecting the relevant features that actually bear significance is without doubt a daunting task, this study will also tackle this. In short, we would like to get close to singling out a "recipe" for a successful movie, through thorough data analyses and prediction modelling.

The data set that will be used throughout the study is from the Internet Movie Database (IMDb) [1]. It has been retrieved online from data.world [21]. It contains 28 features with 5043 instances of movie, spanning across 100 years, from 1916 to 2016, and covering movies from in 66 countries. A description of these 28 characteristics can be found in Chapter 2.1.

IMDb ratings are based on the votes cast by registered users (from 1 to 10) Individual votes are then aggregated and summarised as a single IMDb rating, visible on the title's main page. Users can vote any on every released title in the database. While they can vote as many times as they would like, any new vote on the same title will overwrite the previous one, so each vote is per title per user. Various filters are applied to the raw data in order to eliminate and reduce attempts at votes stuffing by people seeking to inflate

1 IMDb is an online movie database that provides information about films, television programs as well as their cast, crew but also allows studios to collect first hand impressions thanks to fan and critical reviews, and ratings. Originally a fan-operated website, the database is owned and operated by IMDb.com, Inc., a subsidiary of Amazon [10].

the current rating of a movie. For this reason, IMDb publishes weighted vote averages rather than raw data averages. In order to ensure that the rating mechanism remains effective, IMDb does not disclose the exact method used to generate the rating [8]. The IMDb rating, though may be flawed, is a good indicator of a movie's critical success, and therefore a variable that producers and studios should pay attention to.

The other columns in the data set that drew our attention are the budget and the gross revenues, as these can say something about a movie's profitability. We therefore took the gross-budget ratio as another index for the success of a movie, next to the IMDb ratings. In short, we have the IMDb ratings as a variable that signifies a movie's critical success, while the gross-budget ratio was selected to be one that represents its commercial outcome.

2

DATA WRANGLING

Data wrangling refers to the process of cleaning, restructuring and enriching the raw data available into a more usable format. The original data set contains some features which are not relevant, which we need to identify and establish if we can delete them. Apart from that, data wrangling also takes care of filling or removing missing values and deleting duplicates.

2.1 FEATURES ANALYSIS

As the first step of data wrangling, we analysed our features to understand which of them are actually useful. Table 1 describes all the attributes of the data set.

2.1.1 *colour*

colour is a binary feature, i.e. a movie is either *Black and White* or *colour*. As Figure 1 shows, after 1970, most of the movies are full-colour. It is also likely that nowadays the movie industry mostly produces *colour* films. Given this reasoning, we can deduce that such feature does not influence our analysis and can therefore disregard it.

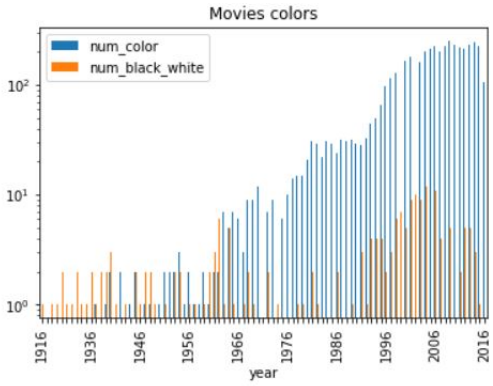


Figure 1: Movies' colour

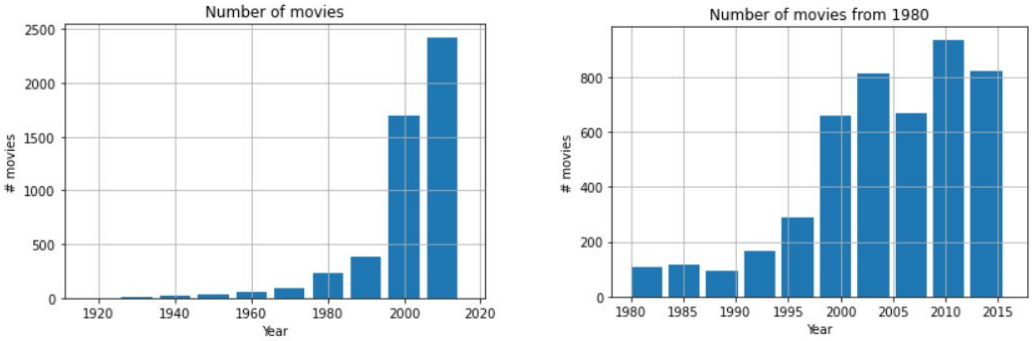
2.1.2 *Year*

Viewers' tastes change gradually over the year, and it is thus dangerous to base decisions on data which are out-dated. It can be observed that a very large number of the movies in the dataset were filmed after 1980 (See Figure 2a). Therefore, to be consistent with the current market, we decided to ignore motion pictures which were produced before 1980.

Variable Name	description
movie_title	Movie's title
duration	Duration in minutes
director_name	Name of the Director of the Movie
director_facebook_likes	Number of likes of the Director on his Facebook Page
actor_1_name	Primary actor starring in the movie
actor_1_facebook_likes	Number of likes of the Actor_1 on his/her Facebook Page
actor_2_name	Other actor starring in the movie
actor_2_facebook_likes	Number of likes of the Actor_2 on his/her Facebook Page
actor_3_name	Other actor starring in the movie
actor_3_facebook_likes	Number of likes of the Actor_3 on his/her Facebook Page
num_user_for_reviews	Number of users who gave a review
num_critic_for_reviews	Number of critical reviews on imdb
num_voted_users	Number of people who voted for the movie
cast_total_facebook_likes	Total number of facebook likes of the entire cast of the movie
movie_facebook_likes	Number of Facebook likes in the movie page
plot_keywords	Keywords describing the movie plot
facenumber_in_poster	Number of the actor who featured in the movie poster
colour	Film colourization. 'Black and White' or 'colour'
genres	Film categorization
title_year	The year in which the movie is released
language	language
country	Country where the movie is produced
content_rating	Content rating of the movie
aspect_ratio	Aspect ratio the movie was made in
movie_imdb_link	IMDB link of the movie
gross	Gross earnings of the movie in Dollars
budget	Budget of the movie in Dollars
imdb_score	IMDB Score of the movie on IMDB

Table 1: Features' description

We can also notice that the number of films quickly increased after 1995. This is due to long period of economic stability that has characterised the 90s and introduced cinema to one's everyday life. Indeed, this increase in the number of films remained even during the Dot-com bubble and the 2008's crisis. [14]



(a) Movies for year

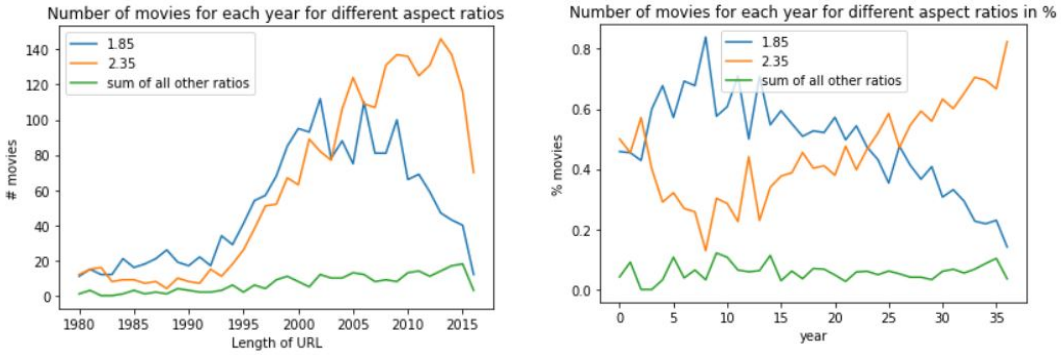
(b) Movies for year after 1980

Figure 2: Movies' Year

2.1.3 *Aspect Ratio*

Movies' aspect ratio is a factor that has changed significantly over the years [1]. At the beginning, although several formats were available, most silent films were made in the 1.33 ratio. When sound was integrated, the analogue audio track, which runs alongside of the film, reduced the visual width to around 1.2 ratio. Once World War II ended, consumers' demand for TVs skyrocketed and by 1954 over half of the households in the US had a television. Consumers were conveniently staying at home with their new TVs, many of which had an AR of 4:3 or 1.33.

Therefore, there are many ratios which have been used, but two of them, 1.85 and 2.35, have been the most common since 1980. If we look at Figure 3a, it may seem that this is not true, this is, however, due to the fact that the number of total movies produced (considering the ones in our database) has decreased. Indeed, if we look at the percentages in Figure 3b, we notice that in the last years there is a steep rise in



(a) Aspect ratio through the years

(b) Aspect ratio through the years in percentage

Figure 3: Movies' Aspect Ratio

movies with a 2.35 ratio [1]. In particular, this is the aspect ratio which has been chosen as standard ratio for films, because it creates the full cinematic effect [20]. For this reason, we can argue that nowadays film makers all prefer the 2.35 ratio, which allows the best vision quality and has become an industry standard, and hence there is no need to take aspect ratio into consideration for our analysis.

2.1.4 *Country*

Figure 4 shows that most of the movies were produced in the US. In fact, it is also likely that the hypothetical movie company for which we are doing this analysis is American. If we would like to consider an European or Asian company, we should use another database, which better reflects customers' tastes. In this case, we have a big majority of US movies and, consequently, we decide to mainly consider American viewers. Furthermore, one issue is that we do not know for movies which are not American in which currencies their budget and their gross are. Theoretically, reading the information of our databases, the prices should be all in dollars. However, some movies (e.g the Japan film 'The Host') are definitely presented in their local currencies.

The first part of the Chapter 4 relies on the gross-budget ratio, defined in Section 3.4. This variable is a fraction of the gross and the budget so, if they are both in the same currency, this should not a problem. After that, for the last part of the EDA, we drop all movies which are not produced in the US to see if our results change. Results are explained in Section 4.3.5. This way we can avoid potential dangerous currency-related mistakes.

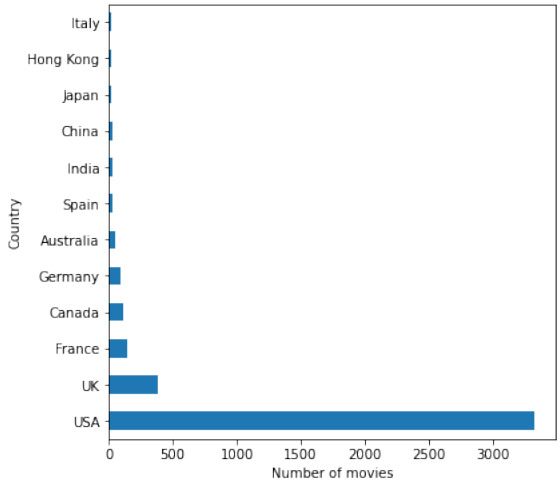


Figure 4: Movies' country

2.2 DROP DUPLICATES AND KEY VALUES

In Section 2.3, we will merge different datasets to fill missing values. But first, in order to get there, we need to find a singular identifier for each movie, or in other words, its

1 2.35 AR represents the "Widescreen format". This wide aspect ratio provides a much more cinematic look and used mostly in films.

key. This value can be taken from the IMDb id - the identification code of the film on the IMDb website IMDB website. We then extract this id from the *movie_imdb_link* feature, for instance from
http://www.imdb.com/title/tt0499549/?ref_=fn_tt_tt_1 we would get *tt0499549*. After this, we add a column to the dataset containing all the *imdb_id* and we drop the *movie_imdb_link* feature.

In the data-frame, some movies are present in duplicates. We thus had to delete them in order to have only one row for each movie. Instead of simply dropping duplicates, we first sorted the movies by the number of NaN values that they have. For instance, if the film x has two copies, one with only the director missing and one with also the genre not specified, we drop the second one. This way we keep only the first appearance and we are sure to drop the less significant copy.

2.3 FILLING NAN VALUES

Table 2: Number of NaN values for each feature before the filling process.

FEATURE	# NaN VALUES		
	director_name	0	
num_critic_for_reviews	40	facenumber_in_poster	13
duration	12	plot_keywords	138
director_facebook_likes	0	num_user_for_reviews	15
actor_3_facebook_likes	16	language	5
actor_2_name	10	country	1
actor_1_facebook_likes	7	content_rating	253
gross	604	budget	369
genres	0	title_year	0
actor_1_name	7	actor_2_facebook_likes	10
movie_title	0	imdb_score	0
num_voted_users	0	movie_facebook_likes	0
cast_total_facebook_likes	0	director_name	0

Table 2 presents the number of NaN values for each attribute. As can be seen, there are features that have a rather high number of NaN values. The goal of this part is therefore to explain our approach in reducing these numbers, in particular the NaNs in *gross* and *budget*, as they are crucial features. In order to do that, we use two different datasets: one from the Wikipedia website and one which uses TMDb data.

2.3.1 *The Wikipedia dataset*

Unlike the base data set, we do not directly download the data we need, but use the wikidata query tool Wikidata Query Service. This tool, written in a semantic query language for databases, allows to query directly against the Wikidata data set. The reason

why we do not only use the Wikipedia data is that we do not consider it a fully reliable source. Moreover, even if it is really rich of information, it is better to use it only as a complementary source.

Using Wikidate Query, we downloaded for each movie the IMDB_id, the gross, and the cost. We then merged this data set with the original one and filled the missing values with the Wikipedia ones where possible. It should be noted that this method did not result in a perfect data set, because Wikipedia is not able to completely fill in the lack of date, but we nevertheless considerably improved the data set.

2.3.2 *The TMDb dataset*

In order to have as much information as possible, we repeated the same steps with the TMDb dataset [9], which contains the IMDB id as well. After this procedure, the data set became substantially more exhaustive. In particular, the number of missing *gross* entries has decreased from 604 to 577 and *budget* entries from 369 to 171.

2.4 REMOVING NAN VALUES

To conclude the Data Wrangling part, we filled in the missing (NaN) values using the data sources described above. Apart from the budget, which requires further examination (Section 3.3), numerical features' missing values have been substituted with the median of the whole sample. While for descriptive columns, as actors' and directors' names, we used a dummy string which marked the missing data.

3

FEATURE ENGINEERING

If feature engineering is done correctly, it can increase the predictive power of ML algorithms. Feature engineering is the process which modifies existing features and creates additional and synthetic features that make machine learning algorithms perform better. This step allows to create attributes from raw data that help facilitate the machine learning process and it directly influences the predictive models as well as their performances. It is worth noting that we do not consider removing duplicates, handling missing values, or fixing mis-labeled classes to be feature engineering. We have put these under the previous section Data Wrangling (Chapter 2).

3.1 ONE HOT ENCODING

In a data set, there can be categorical and numeric variables. In the case of categorical attributes, they contain label values instead of numbers and the number of possible values is often limited to a fixed number [12]. This is also what is observed in the *genre* feature: each movie has one or more genres. The number of movies for each genre is shown in Figure 5.

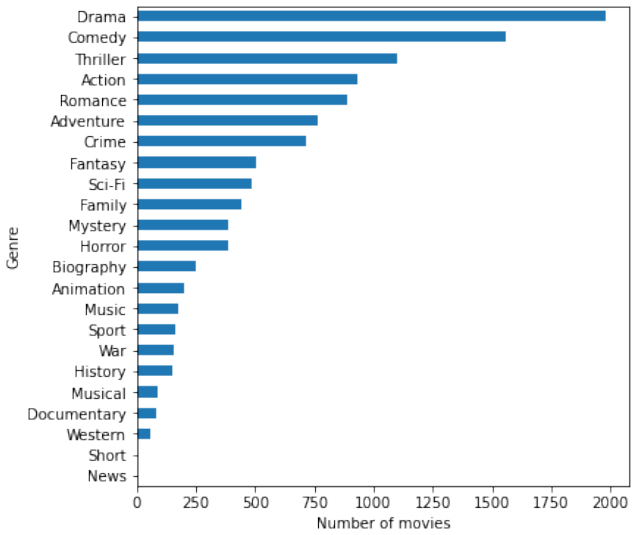
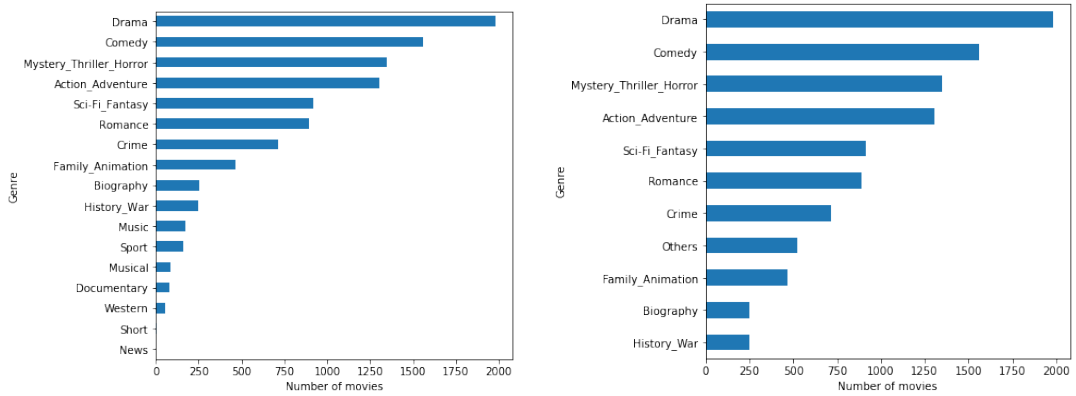


Figure 5: Number of movies for each genre

Some algorithms work with categorical data, but most work exclusively or at least better with numerical values. This means we would have to convert categorical features into numerical ones. Usually, there are two ways to do that: we can transform each label into an integer (1,2,3...) or into a binary (0 or 1). The only benefit of the integer transformation is that we can still have one column to describe the original values (for example ['drama','comedy','drama','war'] becomes [0,1,0,2]). While with the binary representa-

tion, there is one column for each possible label and each element of the data-frame (each movie in our case) has *0* for columns which so not contain its genre and *1* for columns which do. We decided to use the second method, which is often called one-hot encoding in literature. Our choice comes from the fact that if we had used the integers, we would have given an order to these genres. A high or low number would have different meaning in some algorithms of our analysis, while we want that each label is equally important. Moreover, one movie can have more genres, thus the integer conversion would have been hard to use.

Before proceeding with the genres encoding, we reduced the number of labels. This is an important step of feature engineering, as an effective reduction in the number of features can help prevent over-fitting and therefore boost the performance of the prediction algorithms. In order to do that, we regrouped genres which are similar: *Mystery* with *Thriller* and *Horror*, *Sci-Fi* with *Fantasy*, *Family* with *Animaton*, *Action* with *Adventure*, and *History* with *War*. Figure 6a conveys that the number of labels is already reduced enough to make data more usable. This is, however, still not good enough: keeping genres representing too few movies in each would result in unnecessary imbalance. Therefore, we classified as *Others* for all the genres with less than 250 movies. At this point, as described in Figure 6b, we could proceed with one-hot encoding, creating one column for each genre.



(a) Genres after regrouping.

(b) Genres after the *Others* creation.

Figure 6: Movies' Aspect Ratio

3.2 RATING

The data set includes, for each movie, some features about the directors and the three main actors. Apart from their names, we do not have any useful numerical feature directly available but their Facebook likes. The reason why we hesitated in using this column is that it can give unreasonable results: we theorised that some of the important directors of our data set produced their most famous movies before Facebook became popular, and the same can be applied for the actors.

To replace the Facebook likes as the index of a director's popularity, we used quantitative information from their movies. Keeping in mind that the director is likely the one with the most significant impact on the outcome of a movie, we decided to use also their movies' IMDb score to rank them. In particular, for each director, we calculated

how many films he/she has lead and the mean of IMDb scores of his/her movies. After this, we transformed each variable into a number between zero and one divided by the maximum (we did not normalise because their distributions are not similar to the normal distribution), and computed the mean, for each director, of these two numerical features. This allows to give the same 'weight' to each attribute. Now, we have managed to attain a numerical 'index' for each director, which can be easily converted into an integer ranking score between 1 and 10.

Regarding the actors, we followed a similar procedure, using as numerical features the number of movies in which they played and their Facebook likes. The reason why we decided not to directly use the Facebook likes in the analysis is that, even if the social media popularity is important for an actor, it may not be representative of their acting capacity. We also considered the log-distribution for Facebook likes, because values are strongly skewed (see Figure 7 for actor 1; the distribution is very similar for the others as well), and finished by extracting an integer score.

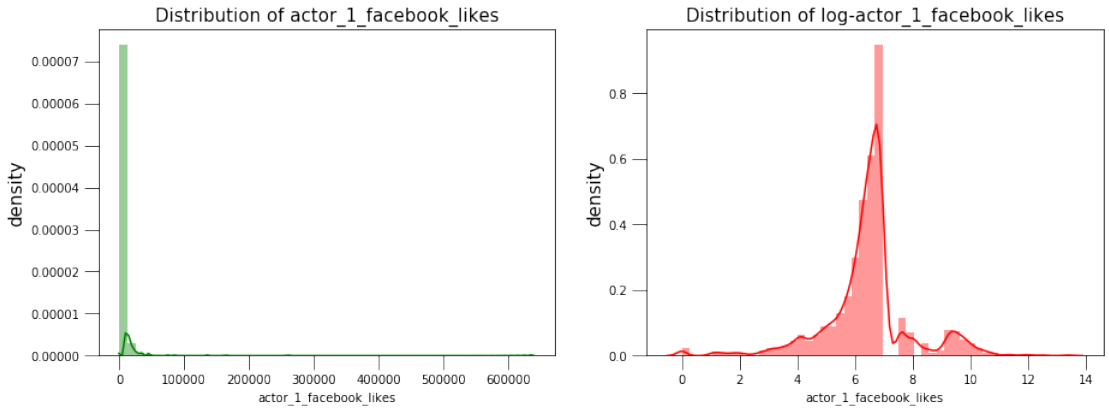


Figure 7: Probability distribution of Main Actor Facebook likes.

3.2.1 *Rating Recent Movies*

After we have ranked all movies, assigned to each film a numerical value for the director and one for each of the three most important characters, we focused on more recent pictures. In particular, we selected only movies produced after *2000*. The reason why we did this is that we wanted to conduct an analysis which is as relevant and as up-to-date as possible. Therefore, we needed to consider only the last few years. In this way, we can use social media influence also for the directors and we can study the impact of having an influential director or protagonist nowadays.

Studying recent movies, we sought to find the most prominent director as before, including also the mean of his/her movies' Facebook likes, the result being that *Christopher Nolan* is the most prominent director. The same was then done with the protagonist with the same feature as the previous actors' ranking and we found that, so far, *Johnny Depp* has been the most "famous" actor of this century. We will use these results in the exploratory data analysis (Chapter 4).

3.3 MISSING BUDGET VALUES

At this point there are 269 missing values in the budget feature which have to be filled in order to ensure a reliable analysis. This is especially necessary in calculating the gross budget ratio, for which we cannot have zero values as the denominator.

We tried to assign to each movie a budget which is as similar as possible to the one of movies with similar characteristics . For this purpose, we used the actors' ranks found in Section 3.2: we calculated the weighted average (50% actor_1, 30% actor_2 and 20% actor_3) to obtain the cast_rank.

Regarding genres, we know that some genres require a bigger investment. From an analysis which shows the median budget of movies for different genres between 2014 and 2018 [4] [1], we found that some genres are more expensive, to which we gave a bigger budget weight. We used this analysis because if our company were interested in investing now or in the near future, an analysis which focuses on the last years would be more trustworthy.

Therefore, we have managed assign a rank, called genre rank, to each movie, according to its genres. This number is between 0 and 1 and assigns a possible level of costs linked to the genre. To be more clear: a film with an high genre rank (an adventure movie for example) is more likely to to cost more. This same reasoning can be applied to the director and the cast.

At this stage, we have three ranks (cast, director and genre) for each movie. We calculate the mean and we converted the result into an integer between 1 and 20. Then, we grouped our database according to this value (20 groups). For each group, we computed the median of the movies' budgets (not considering zero values) and replaced the missing values with the median of the group to which they belong. We have thus complete our column using a specific analysis and without creating outliers.

3.4 GROSS BUDGET RATIO

In order to have a proxy for profitability, it was necessary to create a new variable: the ratio between gross revenues and budget. This new variable will shows how many dollars is the film is able to generate for each dollar of budget. However, this variable is only a proxy since based on gross revenues and not on after tax revenues. Hence, it is worth keeping in mind that will likely overestimate the actual profits.

1 We do not plot this Figure with our data because we desire to have an analysis which is as new as possible and in our database there are not any movies after 2016.

4

EXPLORATORY DATA ANALYSIS

In this chapter will critically investigate and explore the data set. The goal of this examination is to identify patterns, spot anomalies, and verify certain assumptions with the help of summary statistics and graphical representations.

4.1 UNIVARIATE ANALYSIS

The exploratory data analysis phase starts from a univariate analysis of the budget and gross earning in dollars of the movies. Due to the heavily right-skewed distribution of the budget and gross earning, it was recommendable to consider the log transformed values of these variables.

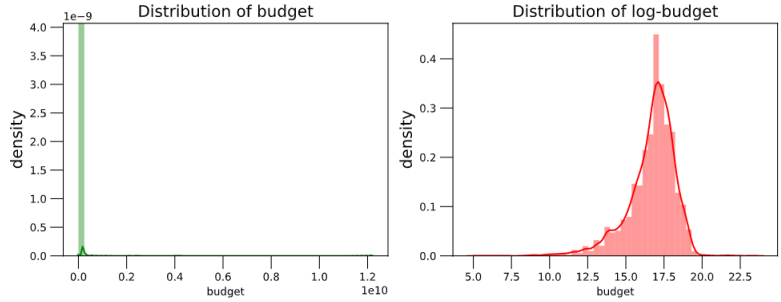


Figure 8: Budget's Distribution

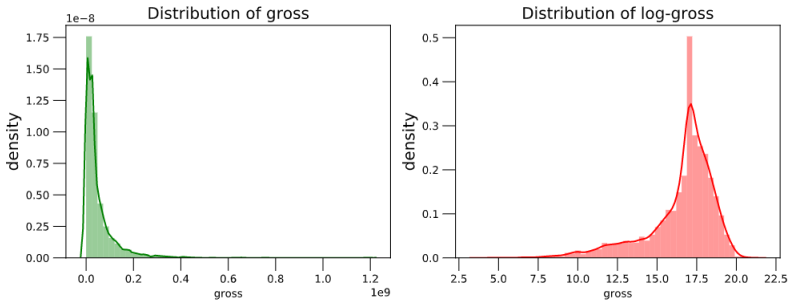


Figure 9: Gross's Distribution

From the the distributions of the logs of the variables in Figure 9, two points are clear:

1. The gross earnings are on average higher than the budget, meaning that on average movies are able to generates profits. However, it is worth noting that this data can be affected by a "survivorship bias"[1] since films that show a negative earnings

1 Survivorship bias occurs when an analyst calculates the performance results of a group of investments, using only the survivors at the end of the period and excluding those that no longer exist[18]

prospects are often not finished in order to avoid the costs of distribution to the public [15]. Another aspect that has to be taken into account is the fact that even if gross earnings are higher than the budget, the profit can be negative due to the impact of taxes or fluctuations of foreign exchange rates.

2. Gross earnings have a higher variance compared to the budget. This, coupled with a heavy left tail, means that gross earnings are more prone to being lower than average than the budget. This also means that a significant portion of the movies in the data set generate negative profits.

The above-mentioned results are even more evident when looking at Table 3 that shows the main statistics of gross earnings, budget and profitability ratio defined as ratio between gross earning and budget. This ratio, as has been previously mentioned, is a proxy for the actual profitability of the production of a film.

	gross	budget	profits	gross_budget_ratio
count	3.966000e+03	3.966000e+03	3.966000e+03	3966.000000
mean	4.785363e+07	4.322898e+07	4.624658e+06	8.689927
std	6.876485e+07	2.193372e+08	2.233179e+08	130.490720
min	6.180800e+01	2.180000e+02	-1.221330e+10	0.000017
25%	7.446117e+06	9.000000e+06	-1.103094e+07	0.446382
50%	2.500307e+07	2.380000e+07	1.611380e+06	1.108586
75%	5.838524e+07	4.867173e+07	2.320307e+07	2.372303
max	1.200264e+09	1.221550e+10	1.174264e+09	7194.485533

Table 3: summary statistic of relevant variables

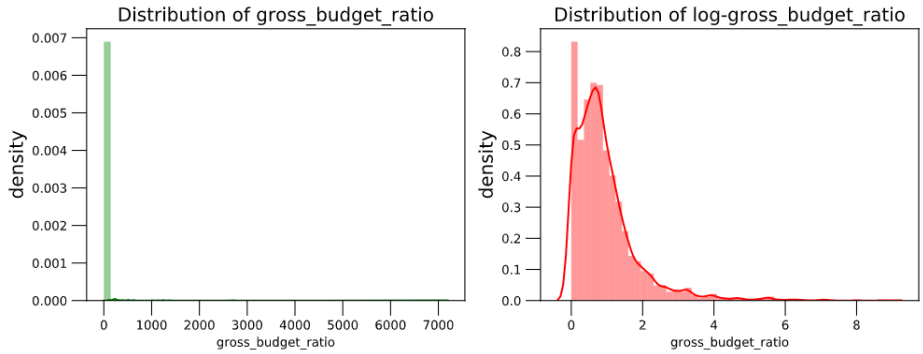
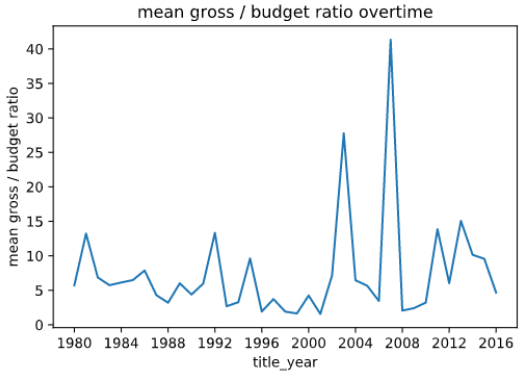


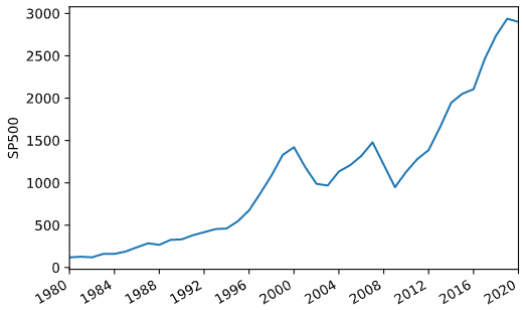
Figure 10: Gross budget ratio's Distribution

It can be deduced from the summary statistics of gross-budget ratio the that there many significant outliers that result in a mean that is 8 times the value of the median of this distribution, causing a distribution that is strongly skewed to the right. We can see how this rate depends on the financial conditions of the US economy by comparing this ratio and the SP500 index in Figure 12.

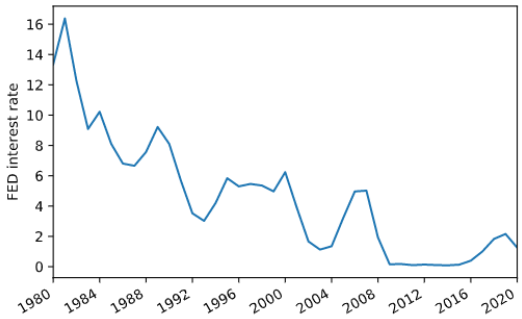
One may notice that the peaks in the profitability rate are correlated with peaks in the stock market. This indicates that when the stock market shows abnormal returns (i.e the



(a) Mean gross/budget ratio 1



(b) SP 500 2



(c) Federal reserve Funds rates

Figure 11: Gross budget ratio's compered with Federal reserve Funds rates and SP500

economy is performing well), movies are also more profitable. This in turn points to the fact that movies are a pro-cyclical investment [2].

However, the validity of this result is compromised when looking to the Federal Reserve Funds rates [6], or in other words, the rate that determine the cost of money for a Bank in the US. As a matter of fact, the peaks in profitability coincide with periods of high interest rates. This means that the higher cost for financing a movie cut off from the market movies with a lower expected profitability. Indeed, it can be seen that the two major peaks in profitability in 2003 and 2007 are correlated with a lower number of films produced compared with the previous year. The same cannot be said about the peak in 1998 or the peaks in 2015 and 2013, meaning that the positive effect on returns due to the positive economic performances overweight the higher financing cost.

4.2 BIVARIATE ANALYSIS

As one of the main indicators of the success in terms of critics and public opinion is the IMDb rating, we expect a strong correlation between the profitability of a movie and its rating. However, from Figure 12 that shows a pairplot and a heatmap in which the outliers of the profitability ratio are removed, it is evident that there is not as strong of a correlation as we had expected. This non-correlation is even more clearly demonstrated in Figure 13, which shows no positive trends between IMDb scores a profitability. More-

2 Procyclic refers to a condition of a positive correlation between the value of a good, a service, or an economic indicator and the overall state of the economy [5]

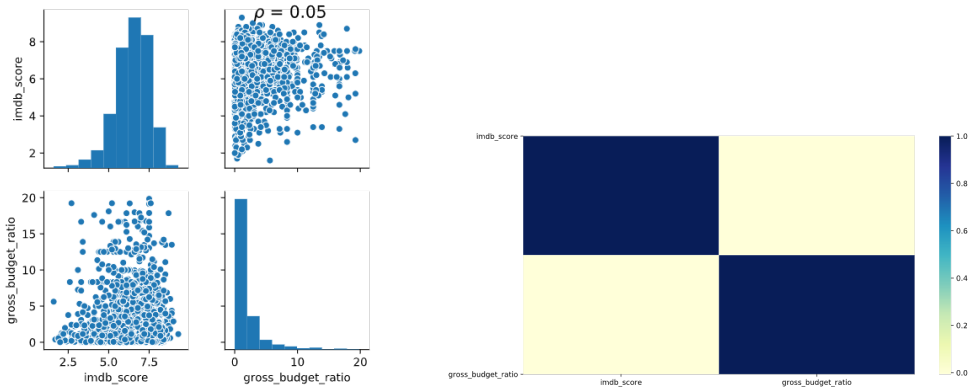


Figure 12: Correlation IMDb rating and gross/budget ratio

over, this figure also indicates the IMDb score has the same mean for the different levels of profitability.

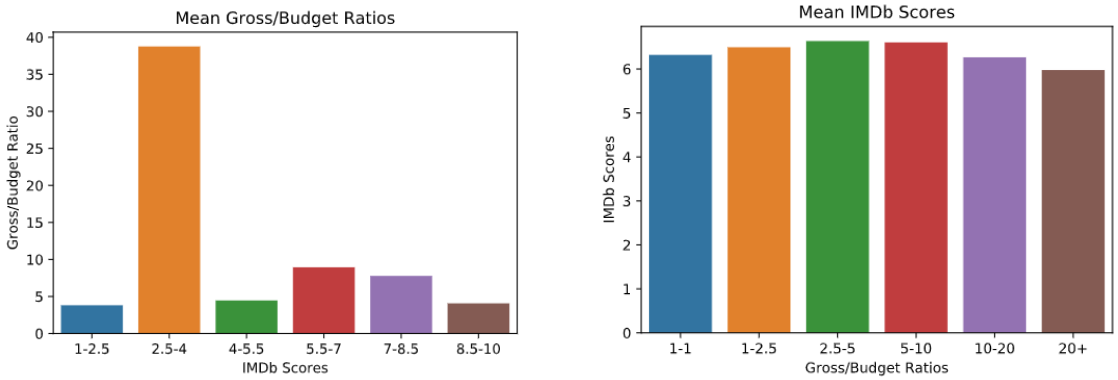


Figure 13: Joint distribution IMDb rating and earnings budget ratio

4.3 MULTIVARIATE ANALYSIS

4.3.1 *The influence of cast and director on profitability*

One of the most important parts of the cost of a film is the salary of the cast and the director [17]. For this reason, one may want to see if there is any trend between gross-budget ratio and rank of the cast and the director. Ranks are determined using the sorting algorithm presented in Chapter 3.

From Figure 14, we can see that on average the rank of the director and cast does not have an impact on profitability. However, it is still worth checking if this is a trend that is valid also for top directors and actors since they make up a significant part of the budget of a blockbuster movie.

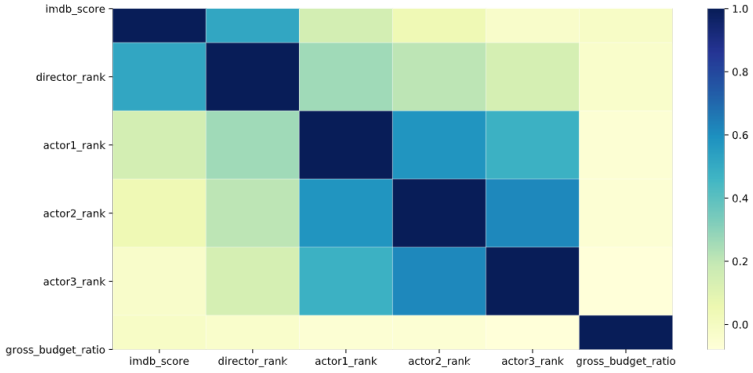


Figure 14: Correlation heat-map rank cast and director with profitability

4.3.2 *The influence of social media on profitability*

One of the phenomena that have greatly influenced our society in the last decade is the development of online social networks. It is therefore important to find out whether the Face book presence of the cast and the director has an impact on the success of a movie. We attempted to see such a correlation through a heat map that takes into account only the movies produced after the 2009, the year in which Facebook reached worldwide recognition.

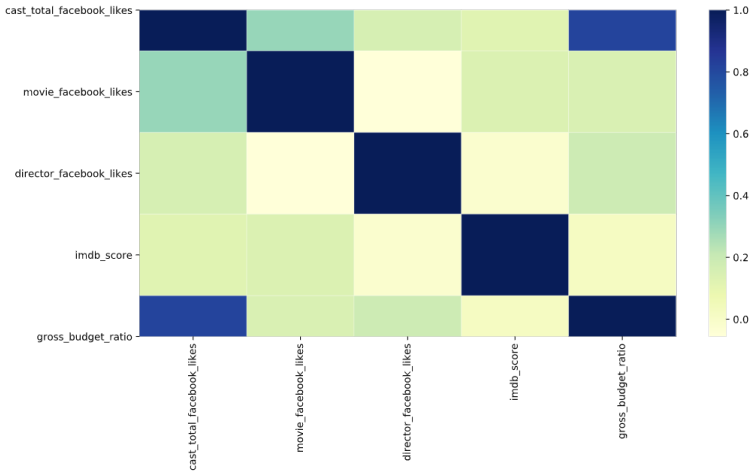


Figure 15: Correlation heat-map presence on social networks and profitability

From Figure 15 , we can see a strong correlation between the total Facebook likes of the cast and the profitability index, which highlights a positive trend between these two variables. It is also interesting to see that these data support the conclusion of a lack of a positive trend between IMDb score and profitability, since the total Facebook likes of the cast do not bear any correlation with the IMDb rating.

From Figure 16, we can see that for returns higher than 2.5 of the earnings/ budget ratio (i.e above the 75% of the entire sample), there is a strong positive trend between this ratio and the cast's Facebook likes. We can thus conclude the existence of a trend, but not necessarily a cause effect relationship since it can be very likely that the high Facebook likes of the cast has helped the film reach an greater number of audience and

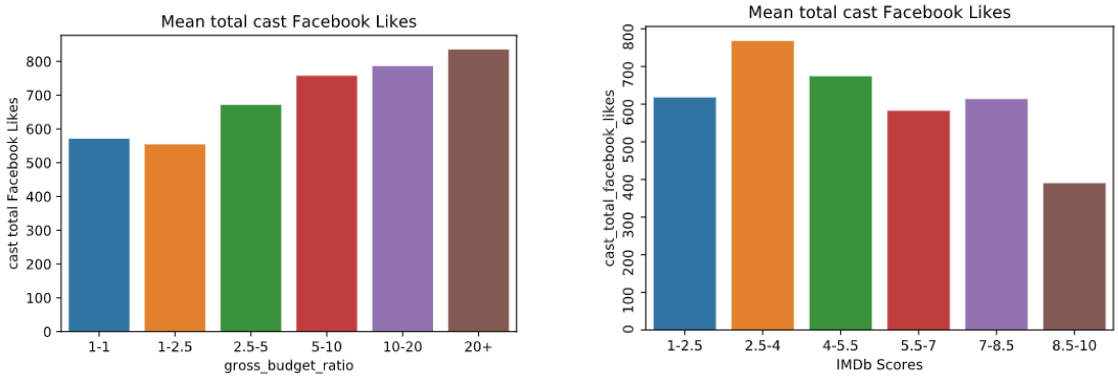


Figure 16: Mean total Facebook likes for groups of earnings budget ratio and IMDb scores

hence the higher profit, or otherwise, it was the success of a film that these actors more famous on social media. Nevertheless, due to the low correlation with the IMDb rating, we are more inclined towards the first hypothesis.

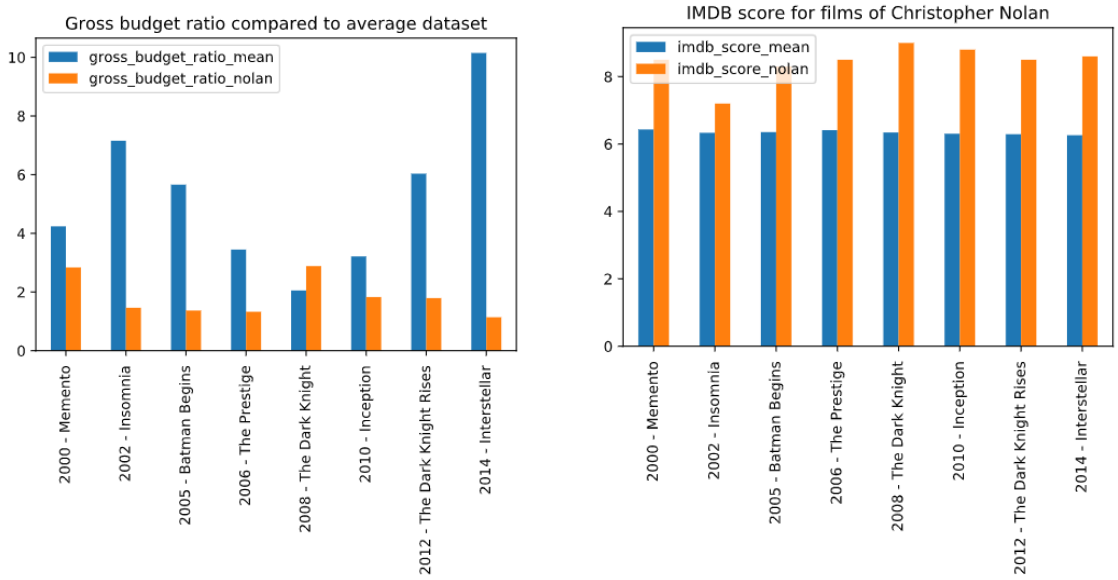
4.3.3 *The prominent directors*

In order to see if the correlation results are only prevalent among the average films or whether they are characteristic even for the most acclaimed films, we conducted an analysis for films from the top directors. In order to see the impact of social media, we chose a director that is the top ranker an considered only the films produced in this millennium.

The director with the highest rank determined using the sorting algorithm presented in Chapter 3 is Christopher Nolan. From Figure 17, Nolan's movie career shows a constant high above the mean (6.44379) IMDb rating, meaning that all of his films have been well accepted by the critics and the public. However, their profitability has not demonstrated the same pattern, since only two movies are in the the top 75% for profitability and only one has a gross budget ratio higher than the average for that year. This example offers a strong support to the theory that there does not exist a correlation between IMDb rating and the profitability of a movie. Nolan's first film "*Memento*" has not only been one that offered the general public and the critics a glimpse of Nolan's talent, but it has been the most successful in terms of rate of profits, making the director an attractive asset for many cinematographic studios.

The same profitability has been replicated by only another movie "*The Dark Knight*". It should be noted that the latter was released in 2008 and had therefore been financed in a period (2006-2007) of high interest rates and positive positive economic performances. Consequently, the success of this film may have been influenced by external macroeconomic variables that have resulted not only in a higher number of audience but also in more studio's decisions aimed towards developing a block-buster film, limiting the power of the director.

On the other hand, from the simple analysis of the earning-budget ratio, it seems that Nolan has not been able to consistently generate above-average profits from his films. From Figure 18, it is clear that, in absolute terms, the profits of Nolan's first film are only



(a) gross/budget ratio

(b) IMDb Score

Figure 17: Trends in Nolan's movie career

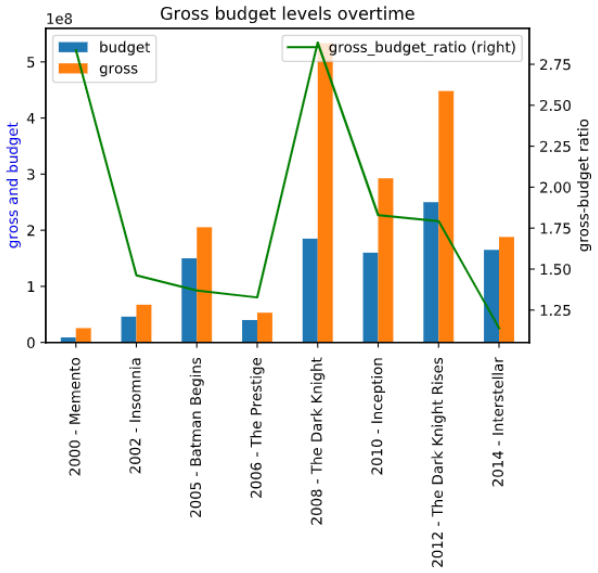


Figure 18: Gross budget levels overtime

a fraction of the figures that he has been able to generate with a higher budget. Moreover, even with different budget constraints, Nolan has been able to generate positive returns from his other projects, manifesting impressive talent even without a costly cast or advanced special effects.

If, from the analysis above, there are proofs supporting the general results obtained in the EDA for the entire sample, the same cannot be said for the correlation of profitability and cast's Facebook likes.

It should be noted that the data on the cast's Facebook likes are about the level of likes reached by the cast in 2016. As can be seen from Figure 19, which shows the Facebook

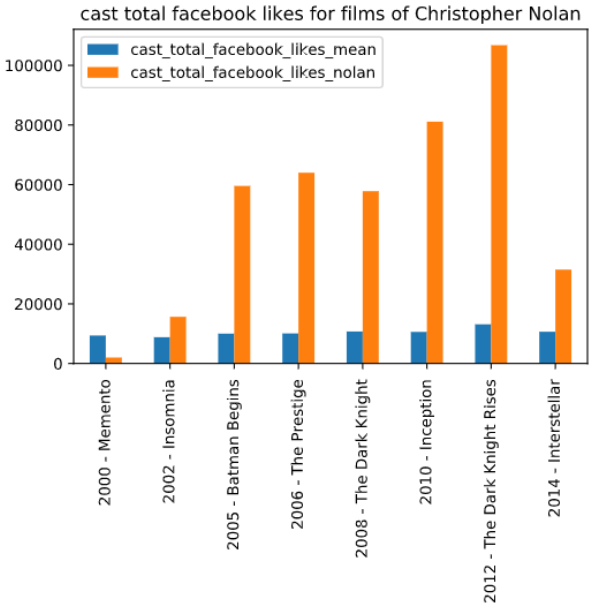


Figure 19: Total Facebook likes of the Nolan cast

likes of each single film, the number of likes does not show a particular pattern with the profitability and the IMDb score.

4.3.4 *The prominent actor*

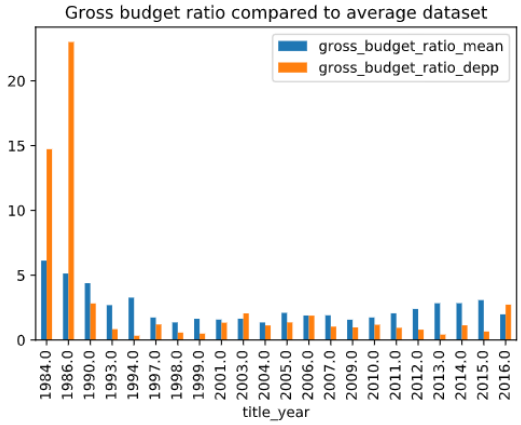
The same analysis that was done for the directors could be applied to the actors. The actor with the highest rank, determined using the sorting algorithm in section Chapter 3, is *Johnny Depp*. In Figure 20, Depp's career has observed a gradual decrease in IMDb ratings. The profitability has not got the same pattern, since only in 4 certain years was the gross budget ratio higher than the average for that year. However, is important to note that all his films have a gross budget ratio above one, meaning that they are profitable.

This again consolidates the hypothesis that there is not a correlation between IMDb rating and the profitability of a movie.

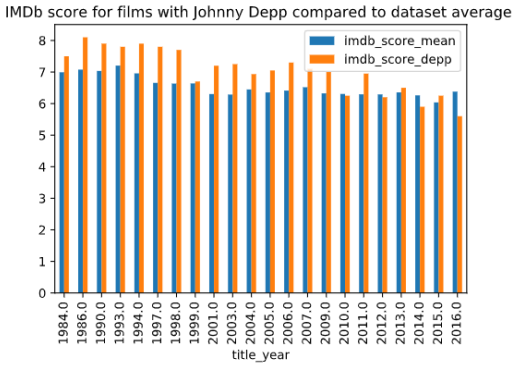
In Figure 21, which indicates the Facebook likes of each of Depp's films, we can see that the number of likes does not bear any significant correlation with the profitability and the IMDb score. In particular, it is important to notice that even films produced before the 2009 showed a high cast total Facebook likes. This means that the above mean value of the Facebook likes is due to external factors correlated with the number of likes.

4.3.5 *Results in the US*

The results above are about films produced worldwide. However, a large majority of the film in our sample are produced in the US, where the major studios are located. In particular it can be seen that the same trends identified in this part are valid if we consider only the US sample.

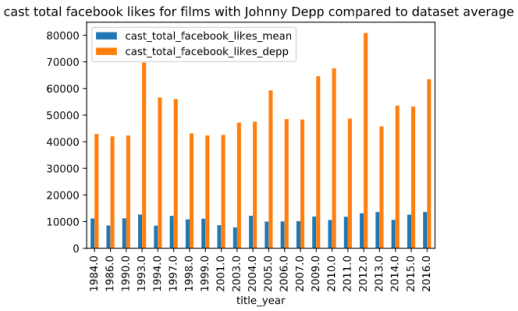


(a) gross/budget ratio 1

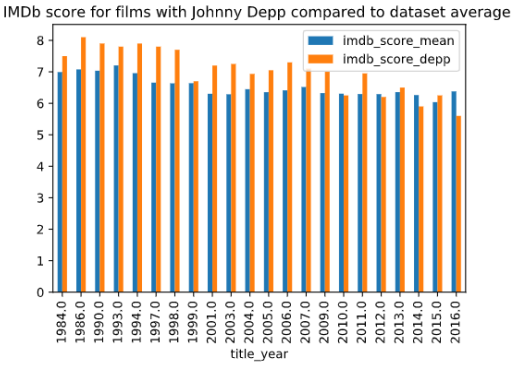


(b) IMDb Score

Figure 20: Trends in the Depp's movie career



(a) Total Facebook likes of the cast 1



(b) IMDb Score

Figure 21: Trends in the Depp's movie career

4.4 KEYWORDS ANALYSIS

After having analysed the statistical characteristics, we attempt to see if there are any patterns in the genres of the films and, in particular, if there were certain keywords describing the movie plot that would systematically generate higher profitability. In order to narrow down our analysis, we first determine which genres generates a higher profitability index on average.

We can see from Figure 22 that the top genre for profitability is *Biography*.
Then from Table 4, we can see that, compared to any other words, when the words "high" and "concerts" are used for describing the plot of a biographical film, this movie gets an higher profitability.

The same is valid for many top 10 words for profitability of films in this genre. This means that, for the biographical genre, the visible topics show a stronger correlation with profitability rather than with the budget. Moreover, we can conclude that biographical movies, whose plots contain "high" and "concerts", have on average the best performance.

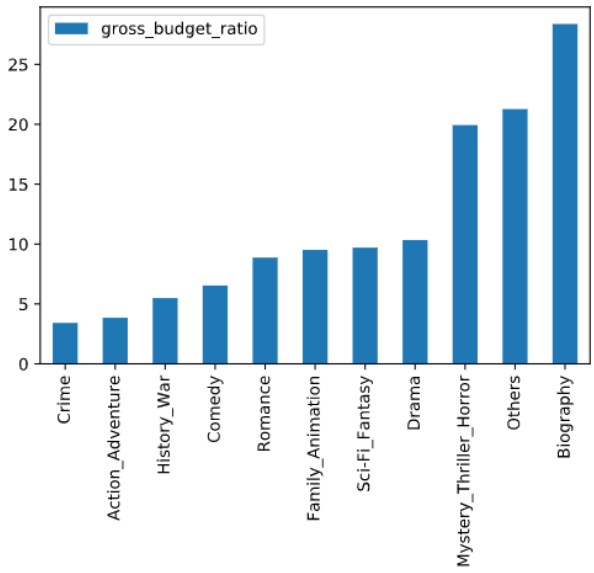


Figure 22: Average profit index for genres

word	profits
high	3.33667
concert	3.07858
character	2.6856
death	2.58274
mother	2.57069
school	2.46829
rights	2.41461
star	2.33207
u.s.	2.17649
music	2.12248

word	mean_gross
u.s.	9.767e+07
death	9.21668e+07
high	8.45289e+07
concert	8.04286e+07
politics	7.85562e+07
racism	6.89475e+07
school	6.51011e+07
rights	6.21763e+07
star	6.1828e+07
president	6.17316e+07

word	mean_budget
king	5.35e+07
boxing	5.25e+07
fbi	4.87143e+07
boxer	4.75e+07
protagonist	4.7432e+07
u.s.	4.4875e+07
president	4.43333e+07
african	4.42857e+07
the	4.34453e+07
politics	3.825e+07

Table 4: Words associated with Biographical films

Since Biography cover only the 7.5% of the sample, we considered also the most common genre : *Drama* (Figure 6b). From Table 5, we see that also for films in this genre the words associated with the highest mean profitability are not the ones that generate the highest mean budget.

This means that also for the drama genre, which makes up a significant part of the sample, the topics treated show a stronger correlation with profitability rather than the initial budget. In particular, drama movies that contain "independence" in the plot have the best performance on average.

PCA After this, we tried to conduct a Principal Component Analysis (PCA) of the keywords, where each word characterises an observation and each genre as a variable. PCA is a statistical method that makes use of an orthogonal transformation to summarise and to visualise the information in a data set containing observations described by multiple inter-correlated variables. PCA is used to extract the important information from

word	profits
independent	11.3073
toilet	6.14834
uncle	4.79498
unrequited	4.5094
mississippi	4.4002
psychic	4.26803
homosexuality	4.00307
hop	3.92504
hip	3.92504
ballet	3.78198

word	mean_gross
ring	1.72922e+08
epic	1.56343e+08
setting	1.53772e+08
earth	1.48001e+08
space	1.42776e+08
vampire	1.29284e+08
middle	1.29266e+08
werewolf	1.2463e+08
king	1.17385e+08
sniper	1.16448e+08

heightword	mean_budget
river	1.53375e+09
oral	8.427e+08
pregnant	7.18991e+08
daughter	4.45443e+08
slur	3.86429e+08
bus	3.75343e+08
christian	2.42868e+08
jewish	2.39924e+08
lesbian	2.31562e+08
casino	1.735e+08

Table 5: Words associated with Drama films

a multivariate data table and to express this information as a set of few new variables called principal components. These new variables correspond to a linear combination of the original ones and the number of principal components is less than or equal to the number of original variables [16]. The main goal of PCA is to reduce the number of dimensions, but another popular application of PCA is in visualising higher dimensional data. Therefore, PCA can organise observed data into meaningful structures. What we hoped to obtain was a plot where we could see distinguishable clusters (visualise in 2D a problem in n-dimension where n is the number of genres), where each cluster represents a keywords' group. Our result is quite a mixed set of points, where the genres greatly overlap. However, the two principal components are well identifiable [3]..

3 See the code on Github (Appendix A) for more details

5 PREDICTION MODELS AND THEIR PERFORMANCES

In this chapter, we demonstrate our attempts in solving the problem of predicting a film's success through five main algorithms - linear regression, logistic regression, decision trees, k-nearest neighbour, and neural network. After detailing the procedures as well as the selected parameters, we present and explain the results that we obtained from these algorithms with the relevant metrics.

5.1 FEATURE SELECTION

Before diving into the details of each model, one shall not overlook a very important step, which involves the selection of features. All models that were experimented with in this study made use of some basic features in the dataset, the selection of which was based on the findings in the feature engineering process (Chapter 3). Since data available before predicting differs, the sets of features used in gross-budget ratio and IMDb score also differ from each other.

5.1.1 *Features Used for Gross-Budget Ratio Prediction*

During the prediction of gross-budget ratio, we only used features available to movie producers and investors *before* filming the movie. Since IMDb score, movie Facebook likes and similar features are only available after the release of movie, we left those features out of our models. 26 features which are available before release were thus used in gross-budget ratio prediction can be seen in table 6.

Duration	Actor 1 FB Likes	Actor 2 FB Likes	Actor 3 FB Likes
Language	Director FB Likes	Cast Total FB Likes	Number of Faces on Poster
Country	Content Rating	Budget	Biography
Comedy	Crime	Drama	History - War
Romance	Mystery - Thriller - Horror	Sci-Fi - Fantasy	Family - Animation
Others	Action - Adventure	Director Rank	Actor 1 Rank
Actor 2 Rank	Actor 3 Rank		

Table 6: Features Used in Gross-Budget Ratio Prediction

5.1.2 *Features Used for IMDb Score Prediction*

For IMDb Score prediction, we used features in table 7. We assumed that the movie has been already filmed and we have all the data connected to the social response. We also

included the rankings of actors and director due to the their impact on social media reaction. We didn't include gross variable since IMDb score has an effect on it.

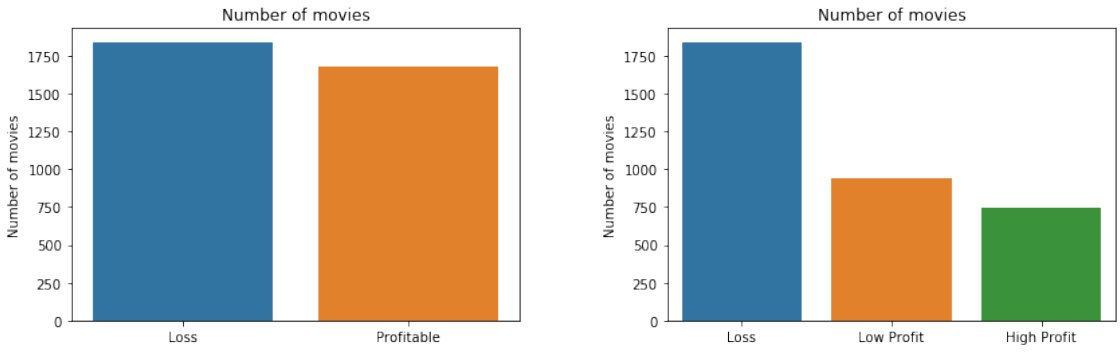
Duration	Budget	Director rank	Actor 1 rank
Actor 2 rank	Actor 3 rank	Movie FB likes	Cast Total FB likes
Biography	Comedy	Crime	Drama
Romance	Mystery - Thriller - Horror	Sci-Fi - Fantasy	Family - Animation
Action - Adventure	History - War	Others	Number of Faces on poster

Table 7: Features Used in IMDb score Prediction

5.2 REGRESSION OR CLASSIFICATION? OR BOTH?

Due to the nature of the problem, which involves predictions of two dependent variables, the IMDb score and the gross-budget ratio, we face the question of which direction to choose for each - regression or classification? This concern arose as both regression and classification can be applied for these two variables, depending on the way we treat data: even though these variables are both continuous, we can group them into multiple bins that bear qualitative values. For example, a film with a gross-budget ratio smaller than one can be regarded as unprofitable and films in this group shall be labelled 0, similarly movies which are profitable (gross-budget ratio greater than one) and which are very profitable (with an arbitrarily chosen gross-budget ratio) should be labelled as 1 and 2 respectively.

The continuity of the labels taken into account, we proceeded to establish for both variables regression models of the three algorithms presented in section 5.6. We later however built also classification models for the prediction of the gross-budget ratio, due to poor observed performance of the regression models, of which details will also be examined extensively in the following section.



(a) 2 Class (b) 3 Class

Figure 23: Gross-Budget Ratio Classification

It is also worth noting that for this classification problem, we first tried categorising the data into two simple classes, profitable and unprofitable, and later three classes, due to a significant number of films that possess extraordinarily high gross-budget ratios.

Figure 24 demonstrates the distribution of the data with regards to the gross-budget ratio after the extreme outliers (gross-budget ratio greater than 5) have been disregarded.

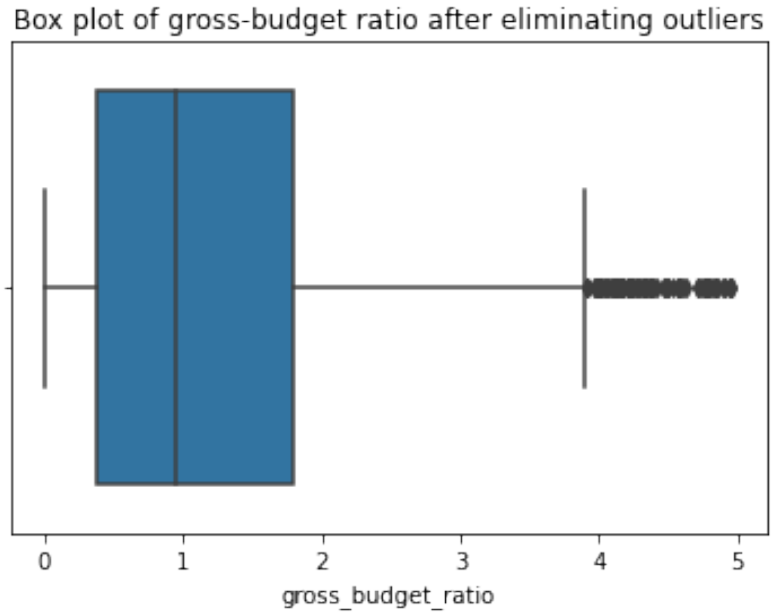


Figure 24: Distribution of gross-buget ratio

5.3 THE PROBLEM OF DATA IMBALANCE IN CLASSIFICATION

As can be seen from the Figure 23, when separated to three classes, there is a problem of data imbalance. Data imbalance means that some classes are not equally represented in the classification problem and might cause an under-representation of minority classes in some algorithms. We can combat this problem by just simply over-sampling the minority class(es). There are also many other available methods to perform over-sampling that are more sophisticated. For example, in some places we applied Synthetic Minority Oversampling Technique (SMOTE) which creates new minority samples within the nearest area of existing minority samples.

5.4 DATA PREPARATION FOR MODELS AND OPTIMIZATION OF MODEL PARAMETERS

5.4.1 *Scaling of Features*

Some algorithms, such as KNN, are scale variant. It means that if one feature has smaller values and other have larger values, it might affect the prediction results. Thus, we scaled the data when it is necessary for the applied model. Since our features are not normally distributed, we used Min-Max Scaler because it is more robust if the distribution of the data is unknown.

5.4.2 *Log-transformation of Gross-Budget Ratios*

In some of the implementations presented in the next section, we took the log-transformed value of the gross-budget ratios instead of their real values. This is due to the heavily-right-skewed distribution of this variable (even after the elimination of extreme outliers, see Figure 24), which can significantly disturb the learning process of some algorithms such as neural networks. Figure 25 shows the distribution of the log-transformed values, which gives a more well-zoomed-in picture of this variable.

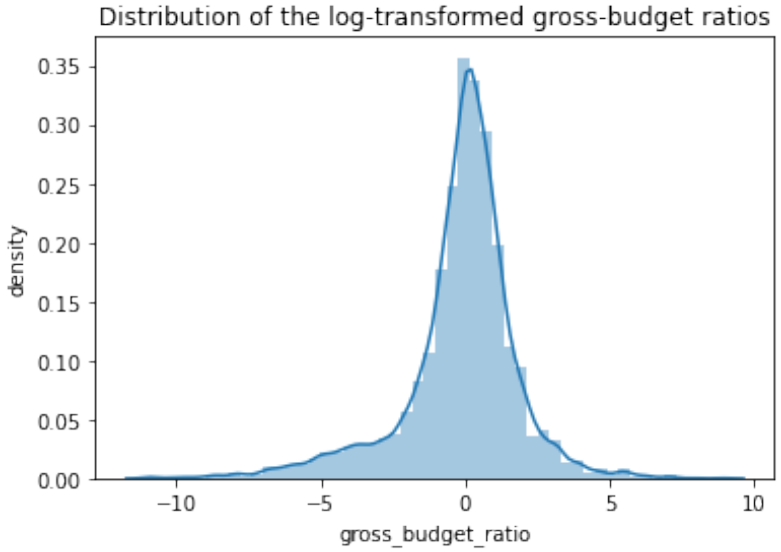


Figure 25: Distribution of log(gross-budget ratios)

5.4.3 *Hyper-parameter Optimization*

With the algorithms that made use of the `sklearn` library, we used the GridSearchCV algorithm to optimize hyper-parameters of the models from given set of values. This algorithm also helps us to prevent over-fitting and under-fitting by applying K-Fold cross validation.

5.5 SOME EXPERIMENTAL TWEAKS

Having experienced certain issues during the modelling process given in the next part, we started trying out some alternatives in terms of tweaking the data as well as the models. Some of these experiments will be presented in this section.

5.5.1 *IMDb Score Prediction*

During the creation of the models for IMDb score prediction, we were considering multiple combinations of variables and timings. Firstly, since IMDb score projects in particular response of the social media, we should consider the influence of the Facebook likes. That is why we tested models with only films with likes and films without likes. Also,

knowing that specific genre also effects the audience, and hence the ability to show the response (small kids and old people are tend to have FB accounts less than teenagers and middle-age people), we run model without consideration of genre.

Secondly, we should take into account that films released before 2009, when Facebook become a popular social network, have affected less from the FB likes. So, we ran model with the same variables presented in Table 7, but only with films released after 2009. As can be seen in Figure 26, movies after 2009 have more FB likes in average.

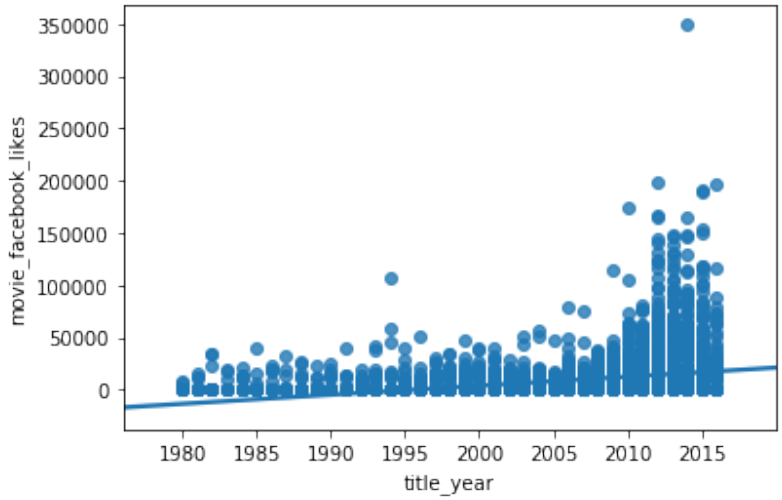


Figure 26: Distribution of FB likes throughout the years

Thirdly, for the skewed features, in our case it is budget, we used log transformations of them. For all these alternative combinations, we applied algorithms to the different feature combinations mentioned in 5.6. More about the combination of variables can be seen in Appendix A Table 13.

5.5.2 *Gross-Budget Ratio Prediction*

While creating models for gross-budget ratio prediction, we tried a couple of different combinations. First of all, we applied the models on all movies and following that, on only U.S. made movies. Within each group, we applied algorithms to the different feature combinations based on the results of Chapter 3. Lastly, for the skewed features, we used log versions of them.

5.6 THE MODELS

This section presents the number of models that we tried out in the prediction phase of the study. In each of the models, we detail our choice of parameters, the issues that we ran into during the process, as well as some figures to demonstrate these models' performance.

5.6.1 *Linear Models*

- Ordinary Least Squares (OLS):
  The OLS method finds the unbiased coefficients that best fit the data given. So, we can't say which variable is more important, we just find the coefficients for our fitted data.

  We used OLS as the first attempt of the fitting due to it's simplicity : no parameters, small time of the execution. Despite the OLS doesn't give the best performance, it is convenient way to detect any changes in model.

  This method is only applied to the IMDb Score prediction and we used MSE and R^2 score as our metrics . They were chosen for their simplicity to optimise and have double view on the performance of the model. Figure 27 shows the best performing OLS model for the IMDb prediction. Even though it is not the overall best performing model, it still gives promising results.

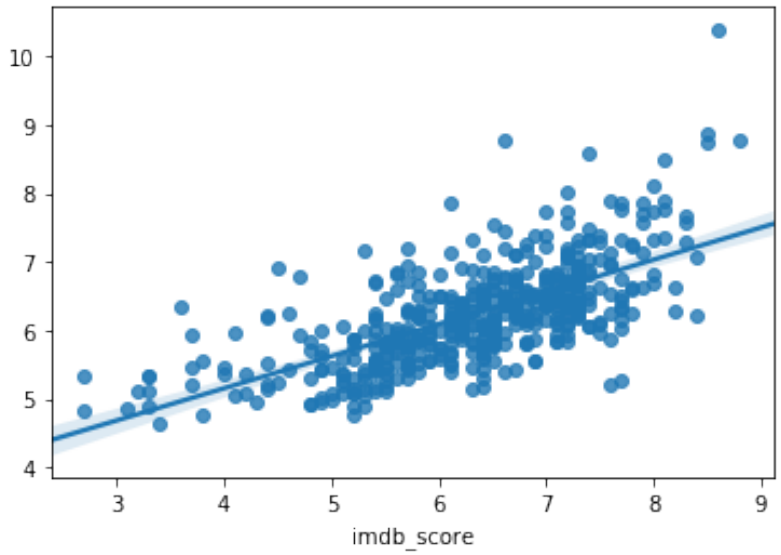


Figure 27: Scatter plot of best performing OLS model for regression of IMDb scores

- Ridge Regression:
  Ridge regression is a regularized linear regression method, which works better if features have multicollinearity (correlation between features). It decreases the coefficients of the features that are less significant, thus, reduces variation. Algorithm has alpha value as hyper-parameter and as alpha increases model complexity reduces. Thus, we tried to optimize alpha value so that we won't over or under-fit our model.

  We used ridge regression for both IMDb score and gross-budget ratio. Figures 28 and 29 shows the results of best performing ridge models. We can see that ridge regression performs relatively good for IMDb score prediction and performs poorly at gros-budget ratio prediction.

- Lasso Regression:
  Lasso regression is another regularized linear regression algorithm, which is similar to Ridge regression. Difference between Ridge and Lasso is that, if a feature

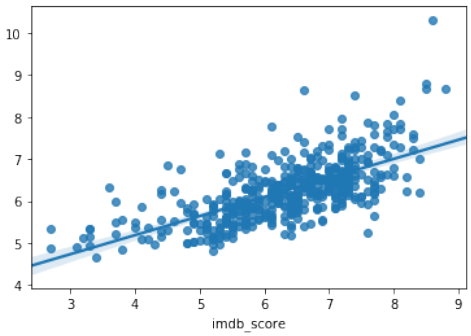


Figure 28: Ridge regression scatter plot for IMDb scores

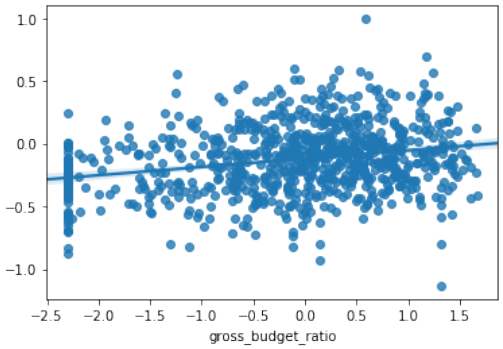


Figure 29: Ridge regression scatter plot for gross-budget ratios

is not significant Lasso makes it's coefficient 0. Hence, Lasso creates models with fewer features. Similarly to ridge, lasso also takes alpha as hyper-parameter and acts in a similar manner. As alpha increases model becomes less complex. Hence, it should be optimized to overcome under and over-fitting.

Lasso regression is used only for IMDb score prediction. Results of best performing model can be seen in Figure 30. This model performs relatively poorly compared to OLS and Ridge regression.

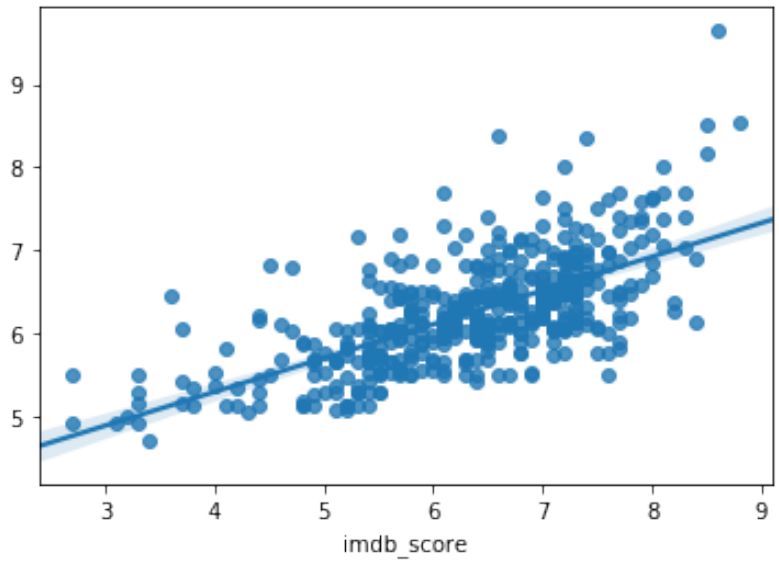


Figure 30: Lasso regression scatter plot for IMDb scores

- ElasticNet Regression:

ElasticNet regression is another regularized linear regression model, it combines feature elimination property of Lasso and feature coefficient reduction property of from the Ridge model to improve model prediction. It also has alpha as hyper-parameter and effect of alpha is similar to ridge and lasso.

In the figure 31, results of IMDb score prediction with ElasticNet are given. This model also performs poorly relative to OLS and Ridge.

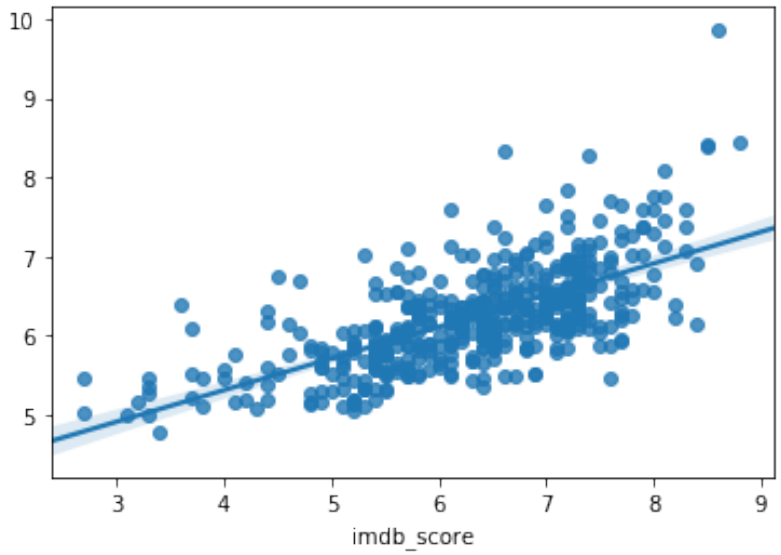


Figure 31: ElasticNet regression scatter plot for IMDb scores

5.6.2 *Logistic Regression*

Logistic regression is a classification algorithm. It is a supervised learning method and it can be used for binary and multi-class classification. It tries to predict probability of a given set of features belongs to predetermined classes.

In this model, we try to optimize hyper-parameter 'C', where lower C value means higher regularization (might lead to underfit) and higher value of C means less regularization (might cause overfit). One important thing to mention here is that, we are optimizing hyper-parameters based on 3 metrics, specificity score and accuracy score and recall score. Specificity score is the ratio of predicting a loss correctly given that movie is loss project. In the Figure 32, confusion matrix of best performing logistic regression can be seen.

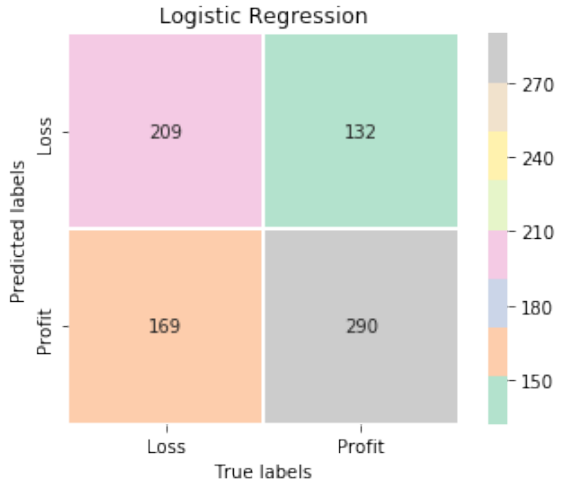


Figure 32: Logistic regression confusion matrix for gross-budget ratio

5.6.3 *Decision Tree (and its variations)*

Decision tree is a supervised learning algorithm which can be used for both classification and regression. It tries to solve the problem by using a tree representation. In the tree, each internal node of the tree corresponds to a feature condition, and each leaf node corresponds to a class label (classification) or a value (regression). Decision tree and its variations/extensions, such as Random Forest, are known for being highly interpretable with its reasoning as well as its training mechanisms.

We used this method for both IMDb score and gross-budget ratio predictions and for latter we applied regression and classification. Best regression results with decision tree can be seen in Figures 33 and 34. As can be seen decision tree regressors perform poorly.

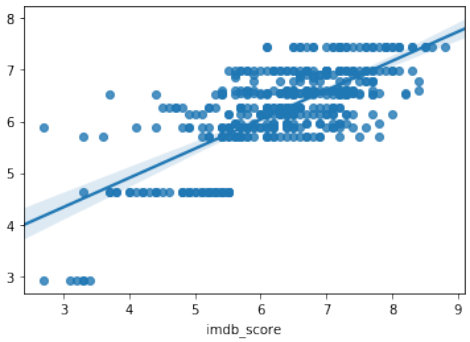


Figure 33: Decision Tree regression scatter plot for IMDb scores

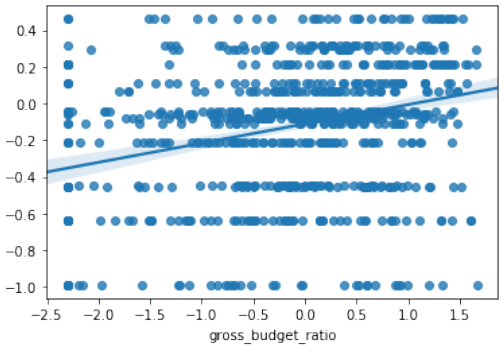


Figure 34: Decision Tree regression scatter plot for gross-budget ratios

For the gross-budget ratio classification, confusion matrix is shown in Figure 35. It works better than logistic regression with respect to metrics mentioned in 5.6.2.

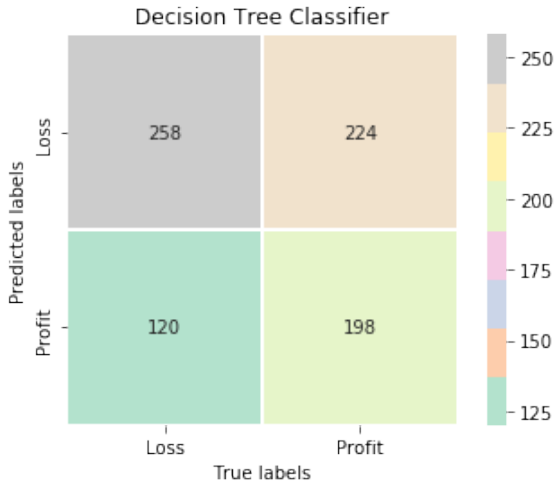


Figure 35: Decision tree classifier confusion matrix for gross-budget ratio

5.6.3.1 *Random Forest*

Compared to the decision tree algorithm using full data set, Random Forest randomly selects observations and features and creates many decision trees. Following that, aver-

ages of those decision tree results are taken to predict classes (classification) or values (regression).

Similarly to decision tree, this method is used for both IMDb score and gross-budget ratio predictions. Results of the regression models are given in Figures 36 and 37. Both regression models are performing best among the models we tried. However, gross-budget ratio prediction still not good enough to be used in business.

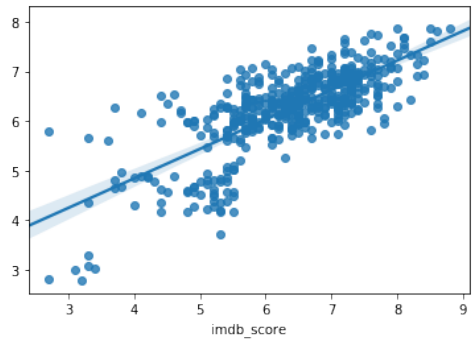


Figure 36: Random forest regression scatter plot for IMDb scores

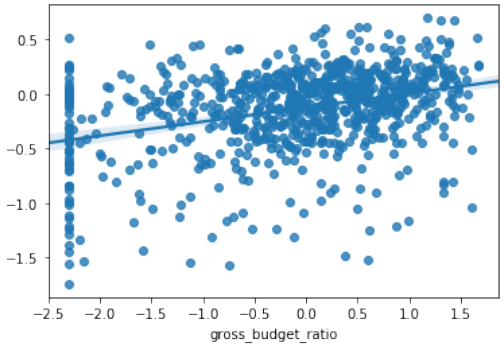


Figure 37: Random forest regression scatter plot for gross-budget ratios

For the gross-budget ratio classification, confusion matrix is shown in Figure 38. Since it's specificity and recall score is relatively high at the same time, we can say that it is the best performing classification method with respect to metrics mentioned in 5.6.2. In section 5.7, there will be in-depth analysis for Random Forest methods.

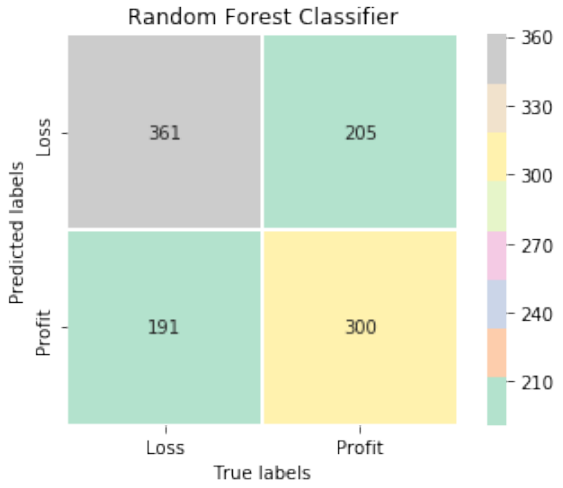


Figure 38: Random forest classifier confusion matrix for gross-budget ratio

5.6.3.2 *XGBoost*

In Gradient Boosting each new tree is fit on a modified version of the original data set. This means that, it tries to improve model iteratively adding new trees. While adding new tree, it increases the weights of those observations that are difficult to predict and lower the weights for those that are easy to predict. We applied this method only to the best performing combination of data as an extra.

Since it takes a lot of computational time, we used this model only on the best combination for gross-budget ratio prediction. In the Figures 39 and 40, results of XGBoost method for gross-budget ratio classification and regression can be found. After the Random Forest and KNN, XGBoost is the third best performing model.

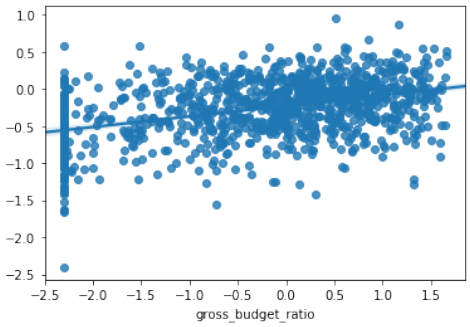


Figure 39: XGBoost regression scatter plot for gross-budget ratios

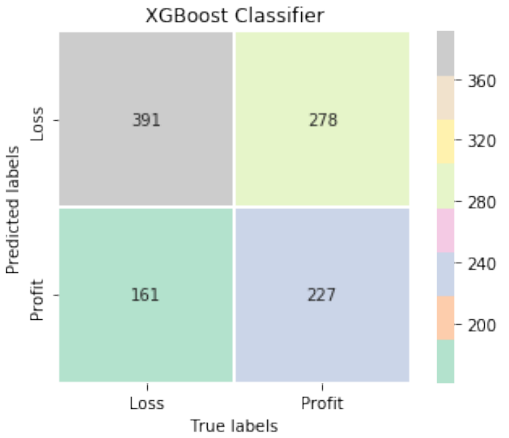


Figure 40: XGBoost classifier confusion matrix for gross-budget ratio

5.6.4 *K-Nearest Neighbor*

K-Nearest Neighbor algorithm predicts the new data based on the similarity function calculated with k nearest observations. It can be used in both classification and regression.

In this method, we try to optimize number of neighbours (k). As k gets closer to one model memorizes training data and over-fits and as k increases there is a chance of under-fitting. As can be seen in Figure 41, KNN is the best performer for classification after random forest. Even though KNN predicts more losses in total (by 33 more correct losses), it performs poorly at predicting number of profitable movies correctly compared random forest (by 71 less correct profitable movies). Thus, we conclude that random forest is performing better for our metrics.

5.6.5 *Neural Networks*

Neural Network (or Artificial Neural Network) algorithms have gained attentions in the past decades due to its robustness and the multiple libraries that aid the development of neural-network-based prediction applications. The use of these algorithms is often, however, scrutinised for its black-box nature: it is not an easy task to figure out what is going on between the layers of such a network. With the current problem as well as the study's purpose, we seek to be able to thoroughly and accurately explain to businesses the technical sides and thus a parsimonious model is preferred. The networks used in this study are therefore relatively simple in their composition as well as their optimisation techniques, minimising as much as possible the hyper-parameter tuning process.

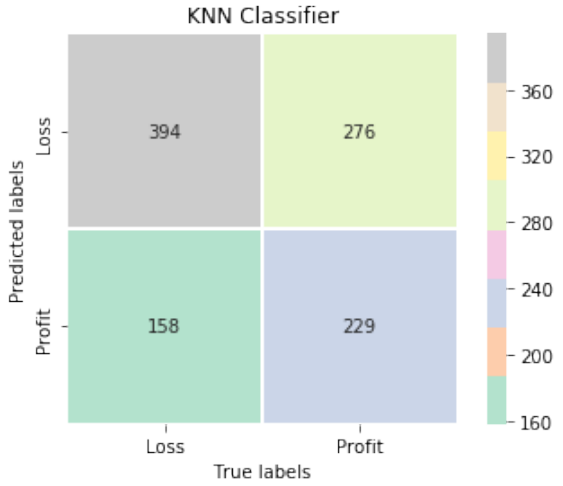


Figure 41: KNN classifier confusion matrix for gross-budget ratio

Four main neural networks were used in this study, one for regressing the IMDb scores and three others for regressing as well as classifying the gross-budget ratios (profitability). Table 8 presents an overview of these architectures, which were decided based on the preliminary results of several training trials (each containing 500 epochs). These may not yet be the optimal choices, but they stood out as the stable ones after the numerous attempts. For classification of the three-class profitability problem, an additional simple data re-sampling procedure was applied next to the without-sampling version, in order to bring the number of class 1 (profitable) and class 2 (very profitable) to the same quantity as class 0 (not profitable).

Unlike the other models that were presented above, we did not train the neural network models with each different set of features or each feature individually. This practice would not make much sense in the context of neural networks, since not only does it create very sparse and over-fitting networks, but this non-linear algorithm's strength is already in it being highly adaptable and agile in assigning weights to the features.

As indicated in Table 8 itself, instead of regressing the real gross-budget ratio, the implementation resorted to using the log-transformed values of this variable. This is again due to the heavily skewed distribution of the gross-budget ratios which was mentioned in section 5.4.2.

While the network used for predicting IMDb scores showed promising results, the other three networks did not attain the same success. In terms of regression, Figure 42 and 43, which contain the scatter plots of the real and predicted values of the two variables, demonstrates clearly this difference in performance. The results for classification of gross-budget ratios were also not the best, with accuracy fluctuating between 50% to 60%, and recalls/precisions extremely unaligned between the classes.

One problem that was often encountered through the various training trials of the three last networks was the presence of over-fitting. This is why parsimonious networks with as few hidden layers and neurons were chosen, and why drop-out was utilised in the case of data re-sampling for classifying profitability. These attempts still did not help boost the performance on the test set, as can be seen in the classification report in Table 9, conveying the fact that predicting profitability is, indeed, a very difficult problem that requires a more in-depth revision in not only the algorithms but the data set itself.

Network Architecture	Regressing IMDb Score	Regressing log(Gross-budget Ratio)	Classifying Gross-budget Ratio
# hidden layers	1	1	1
# neurons in hidden layer(s)	32	32	32
Activation function	selu	selu	sigmoid and softmax
Optimiser	Adam	Adam	Adam
Loss function	MSE	MSE	categorical cross-entropy
Miscellaneous			used drop-out to minimise bias

Table 8: The Neural Network Architectures and their Parameters

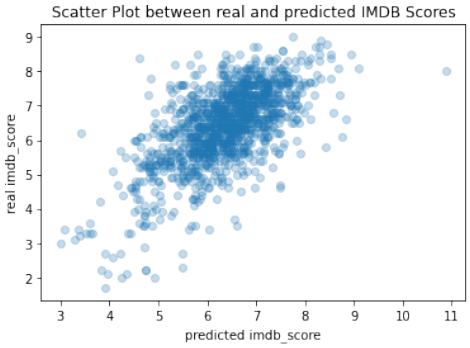


Figure 42: Scatter plot for regression of IMDb scores

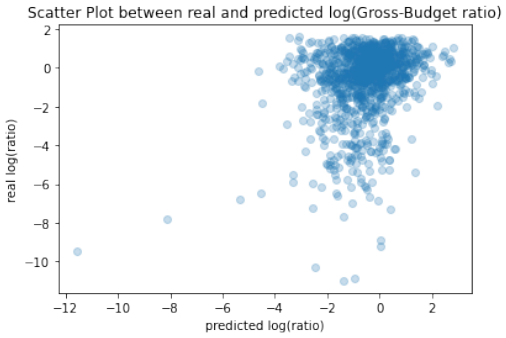


Figure 43: Scatter plot for regression of gross-budget ratios

Moreover, the EDA phase of this study has already demonstrated the non-correlation between the IMDb scores and the profitability, meaning that the simple network that succeeded in predicting IMDb scores is not the solution for the profitability counterpart.

Metric	Binary clf	3-class clf with resampling	3-class clf w/o resampling
Accuracy	0.6	0.5	0.53
Precision (for class 0 / 1 (/ 2))	0.63 / 0.56	0.59 / 0.35 / 0.43	0.76 / 0.24 / 0.22
Recall (for class 0 / 1 (/ 2))	0.64 / 0.55	0.67 / 0.33 / 0.35	0.61 / 0.29 / 0.42
F1 score (for class 0 / 1 (/ 2))	0.64 / 0.56	0.62 / 0.34 / 0.39	0.67 / 0.26 / 0.29

Table 9: Results of Neural Networks for Classification

5.7 PERFORMANCE RESULTS AND INTERPRETATIONS

This section presents the detailed results of the models above through the relevant regression and classification metrics. These metrics, combined with the other observations in the previous section, will serve as the basis for our discussion of the algorithms as well as the implications that we see for film businesses in the next chapter.

5.7.1 *IMDb Score Prediction Results*

The best predictions for each of the combinations from Appendix A Table 13 was achieved by Random Forest model. The best result was achieved by combination of variables 6 which can be seen in Table 10. Other results for all combinations can bee seen in Appendix A Table 14. The lowest MSE score and highest R^2 were shown by Random Forest, graph for this model can be seen in Figure 44. All the regression models achieved low MSEs on the test set (given that the range of the IMDb score is from roughly 0 to 10), and the reasonable R^2 scores across the models, indicating that the variance of the data was well learnt. The success of Random Forest could be attributed to the algorithm's concept of bootstrap aggregating, which helped reducing variance in prediction to get higher predictive power.

Out of these, the Neural Network regression model performed the worst. This could be due to the simplicity of the network architecture, and the fact that the hyper-parameters were not extremely well tuned. We also stopped the training after 500 epochs, after seeing that the loss on the validation set has reached a reasonable low. From the scatter plot Figure 42, one can not only see the "fat" concentration around a regression line that signals an amount of noise, but also some outliers that were not learnt by the algorithm. All in all, the numerous-parameter nature of the neural networks made it easy to overlook certain problems in the training process. We must, however, keep in mind that the inclusion of the neural networks' algorithm in the study was not without the aim of being able to explain to businesses what is going on behind the scene, and thus we sought to keep the architectures relatively minimal. This could very well be the compromise that this algorithm is facing in this prediction problem.

Metric	OLS	Ridge	Lasso	Elastic Net	Decision Tree	Random Forest	Neural Network
MSE	0.64	0.63	0.65	0.65	0.57	0.5	0.77
R^2	0.44	0.45	0.43	0.44	0.51	0.56	0.4

Table 10: Results of the model for IMDb Score prediction

Also for Random Forest, we analysed feature importances, as shown in Figure 45. From that we can see that the most important features for that regression are director rank and movie FB likes. The importance of the FB likes can be explained by the fact that they are also indication of the viewers attitude. IMDb score and FB likes are highly correlated, so for prediction it is relevant to use FB likes as another pattern of the response.

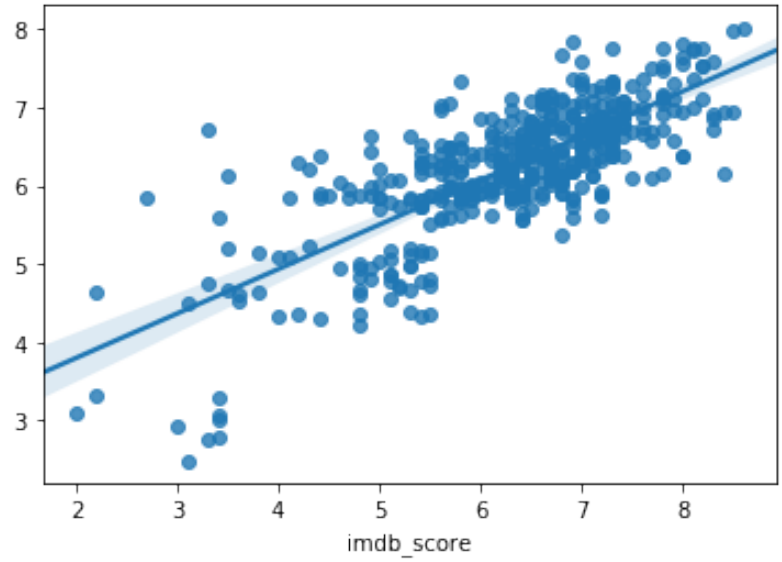


Figure 44: Prediction of the IMDb Score using Random Forest

For director rank, we can say that it influences the IMDb score due to expectations of the viewers. Since the high rank directors, in most cases , create better movies, they arise bigger response in social media and have higher ratings for their films.

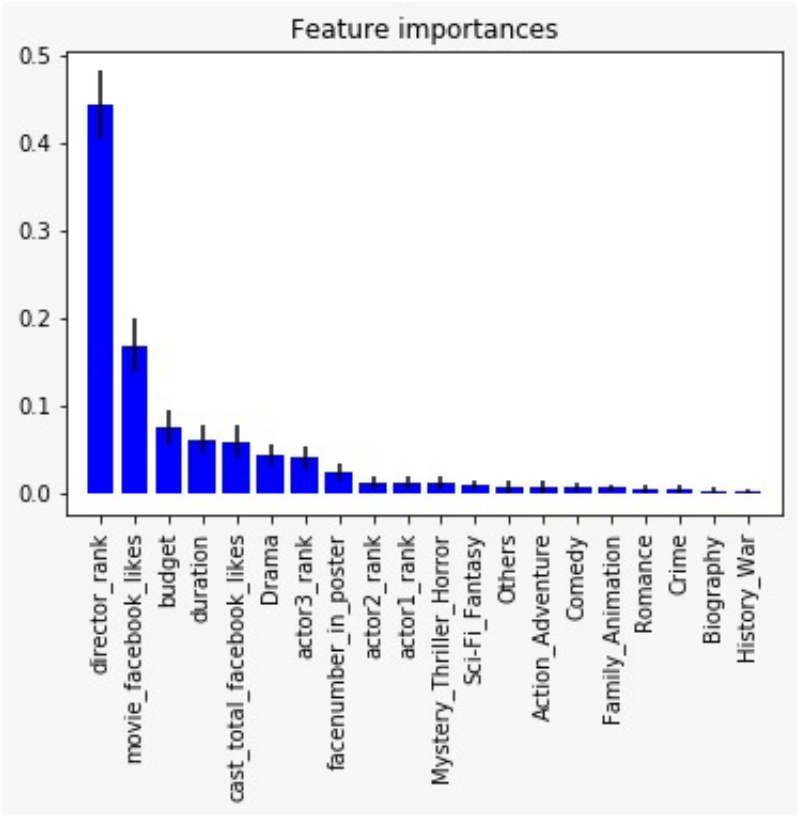


Figure 45: Feature importance of the IMDb Score using Random Forest

5.7.2 *Gross-Budget Ratio Prediction Results*

As mentioned before, for the gross-budget ratio prediction, we first applied regression models and obtained results shown in section 5.7.2.1. After observing that regression models are not working satisfactorily, we decided to apply classification models in 2 ways, first we divided output into 3 groups and then tried again with dividing output in 2 groups.

5.7.2.1 *Regression Results*

In total, we evaluated 37 different regression models based on mean squared error score and results of the models can be seen at Appendix A Table 15. From 37 different models, the best performing model is the random forest regressor performed on the only U.S. made movies. Model includes both Facebook likes and ranks calculated by our group for directors and actors, eliminates some of the non-significant features and takes logarithm of highly skewed features. With the help of the following figures, in-depth analysis of the random forest regressor for the gross-budget ratio can be conducted.

Parameter or Metric	Value
Max Depth	10
Number of Trees	200
MSE	0.903

Table 11: Best gross-budget ratio regressor model parameters and metric

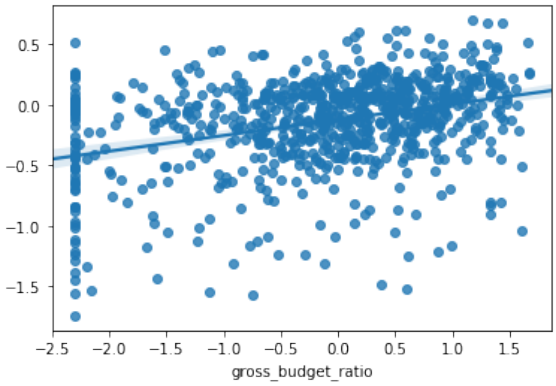


Figure 46: Best model: Random forest regression scatter plot for gross-budget ratio

As can be seen from the table 11 and Figure 46, even the best performing regression model for the gross-budget ratio prediction is not giving promising results. Mean squared error of 0.903 tells that in average there is a difference of 1 in prediction. This means that, mean error equals to money invested in the production of the movie (budget). Hence, we can conclude that, with the available data it is not logical to apply regression models to predict gross-budget ratio. Following that, we decided to try classification methods.

5.7.2.2 *Classification Results*

- Multi-Class (3 Classes) Classification
  During classification of gross-budget ratio, we first tried separating output into 3 classes such as loss, low profit and high profit. We applied classification models for 3 class prediction with the baseline models from `sklearn` based on accuracy score metric and obtained the results (48 results) given at Appendix A Table 16. Accuracy scores of 3 class prediction are between 40-55%, which are worse or slightly better than naive guessing of predicting everything as loss which is around 50% of the data. The same low results were also observed for neural networks, as can be seen in Table 9. Hence, we concluded that it is not a good prediction that can be used in movie industry by investors.

- Binary Classification
  Lastly, we decided to make binary classification and divided output data into 2 classes; loss and profit. Compared to 3-class prediction, here we had rather balanced data between the classes. The results (48 results) of binary classification can be seen in Table 9 and Appendix A Table 17. We decided the best model based on 3 different metrics such as accuracy, specificity and recall score. Specificity score gives the power of predicting losses and recall score gives the power of predicting profitable movies.
  Similarly to regression, best performing model for classification is also random forest. With the help of the following figures, in-depth analysis of the random forest classifier for the gross-budget ratio can be carried out.

Parameter or Metric	Value
Max Depth	10
Number of Trees	200
Accuracy	0.625
Specificity	0.654
Recall	0.594

Table 12: Best gross-budget ratio classifier model parameters and metrics

From Table 12 and Figures 47 & 48, we can see that this models works relatively better than the models in the previous sections. ROC curve shows that it works better than random guessing and from the confusion matrix we can see that it predict 65% of losses. Since, it can predict 2/3 of the movies that is gonna cause loss, we can suggest this model to the risk-averse companies. This model wouldn't be attractive for risk-loving movie producers because it predicts 40% of the profitable movies as a loss and risk loving companies might lose good investment opportunities.

We can further investigate the model by looking at the feature importance values given in the Figure 49. As can be predicted, the most important feature is the Budget. Following that facebook likes of actor 3 comes and it is pretty weird. However, if the facebook likes of actor 3 is higher than it might means that cast of the movie are all well-known actors. Hence, it might achieve higher gross-budget ratios.

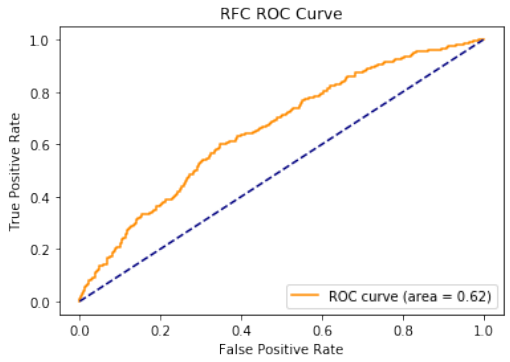


Figure 47: Best model: Random forest classifier ROC Curve

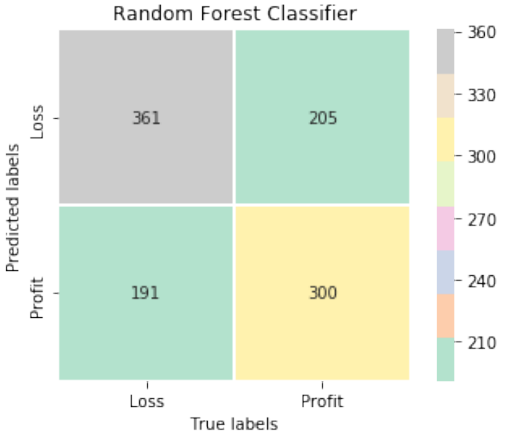


Figure 48: Best model: Random forest classifier confusion matrix for gross-budget ratio

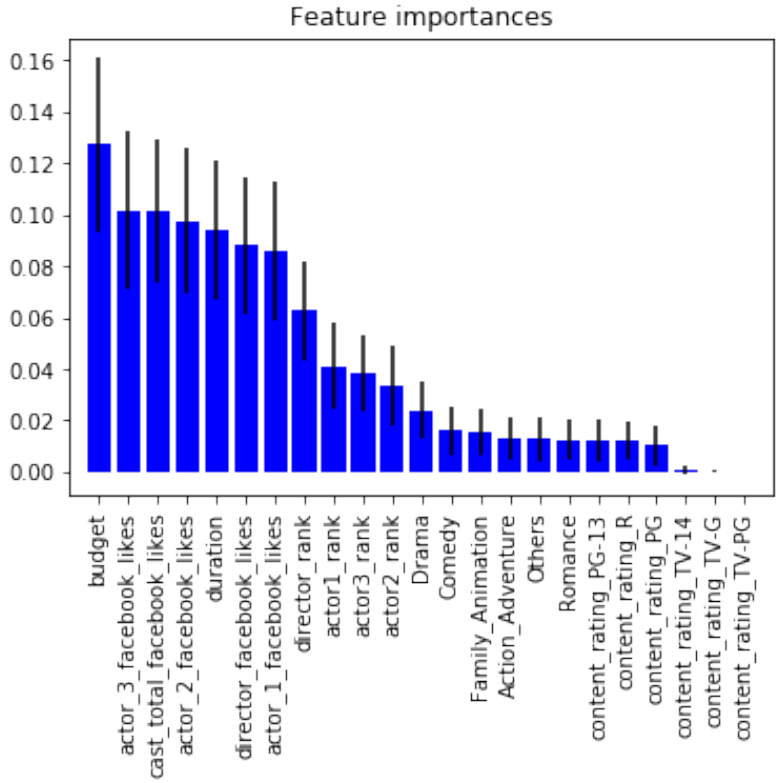


Figure 49: Best model: Feature importance

As a last remark, we can conclude that with the available data before the release of the movie, it is hard to correctly predict profitability of the movie. Movie producers can conduct surveys to collect more data from the movie viewers and may be increase their accuracy score for predicting profitability.

6

IMPLICATIONS FOR BUSINESSES

Having extensively utilised the data set of IMDb to explore the world of cinema and analysed it from the many different angles, the current study has deducted a few useful observations and lessons that one can potentially convey to those working on the side of movie production. While they do not guarantee an extremely strong generalisation, they are nevertheless valuable assets when looking from the business perspective. These implications for producers/studios/film-makers can be summarised in the following points:

- In terms of relevant features, the cinema industry should not focus on variables which can be taken as today's standards, like a film's colour and aspect ratio.
- In order to include all influential attributes in the different algorithms, categorical variables must be transformed into numerical ones. For example, the business can use one-hot encoding for genres and some methods to rank the movies' most impactful individuals (e.g director and main actors).
- Businesses should always keep in mind that audience's tastes change really quickly, so it is important not to found the analysis on out-dated attributes. The producers should always consult the latest trends and developments.
- In terms of prediction models, businesses can make good use of the Random Forest algorithm, as this model is not only intuitive but it has proven to also be robust in its performance, potentially due to its ability to reduce variance in training.
- For the gross-budget ratio predictions, movie producers can use the best performing binary classification method if they are risk-averse. Otherwise, if they are risk loving, we would not suggest using it because it predicts 40% of profitable movies as loss.
- All in all, the poor performance observed in predicting profitability has shown that this variable may be unpredictable or more features/more engineering need to come into play to make this task possible!

Putting everything into perspective, it is important to realise that the cinema industry is going to face major changes in the next months due the COVID-19 outbreak. Indeed, the Coronavirus is having a devastating effect for film business as box offices worldwide face losing billions, and stoppages in filming have left thousands in a mostly freelance industry without work [19]. Moreover, the closure of theatres is driving studios to change the way in which they reach the spectators. In the absence of theatrical release, films are being fast-tracked to streaming. Universal was the first studio to take these steps, announcing that it will make movies available at home on its streaming platform [3]. Hence, streaming platforms will not be a competitor for the cinema industry but actually the best way for overcoming this crisis. For this reason, one of the main limits of the

data set analysed is the lack of data on series and films offered on streaming platforms which would have allowed to offer insights into the way in which studios could face their transition from offline to online.

Nevertheless, it is important to remind ourselves that many events have been predicted to kill the film industry: the 1918 influenza epidemic, the second world war, the invention of television, the invention of VCRs, the invention of the internet [3]. And yet the appetite for movies, like a phoenix, rises, every time stronger than before. This is because, especially in times of political trouble and uncertainty about the future, people develop an urge to escape from the reality, and now more than ever, they want to keep themselves entertained.

[1] John L. Berger. *A brief history of Widescreen format*. URL: https://www.widescreen.org/widescreen_history.shtml.

[2] Kyle Buchanan. "How Will Movies Survive the Next 10 Years?"" In: *The New York Times* (2019).

[3] Kristin M Burke. *Reports of the death of the film industry have been greatly exaggerated*. Apr. 2020. URL: https://www.theguardian.com/film/2020/apr/14/covid-19-killed-the-film-industry-hollywood-coronavirus.

[4] Stephen Follows. *How much does the average movie cost to make? (July 2019)*. URL: https://stephenfollows.com/how-much-does-the-average-movie-cost-to-make/.

[5] Akhilesh Ganti. *Procyclic Definition*. Jan. 2020. URL: https://www.investopedia.com/terms/p/procyclical.asp.

[6] Board of Governors of the Federal Reserve System (US). *Effective Federal Funds Rate [DFF]*. May 2020. URL: https://fred.stlouisfed.org/series/DFF.

[7] IBISWorld. *Global Movie Production & Distribution Industry - Market Research Report*. URL: https://www.ibisworld.com/global/market-research-reports/global-movie-production-distribution-industry/.

[8] IMDb. *IMDb Rating FAQ*. URL: https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#.

[9] Kaggle. *TMDBs data-set*. URL: https://www.kaggle.com/tmdb/tmdb-movie-metadata.

[10] Robert Lewis. *IMDb*. Oct. 2017. URL: https://www.britannica.com/topic/IMDb.

[11] James Luxford. "The Top Movie Trends Of The 2020s." In: *Amex Essentials* (Mar. 2020).

[12] Machine Learning Mastery. *Why One-Hot Encode Data in Machine Learning?* URL: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.

[13] MPA. "2019 THEME Report." In: *Motion Picture Association* (Mar. 2020).

[14] Rebecca Nicholson. *From Clueless to Tarantino: why the 90s was Hollywood's fairytale decade*. Sept. 2017. URL: https://www.theguardian.com/film/2017/sep/29/from-clueless-to-tarantino-why-the-90s-was-hollywoods-fairytale-decade.

[15] Stephen Saito. *Nine Great Unrealized Films from Scott, Nolan, Scorsese and Other Notable Contemporary Filmmakers*. June 2012. URL: http://moveablefest.com/nine-great-unrealized-films-from-great-contemporary-filmmakers/.

[16] Elite Screens. *Understanding Aspect Ratio*. URL: https://elitescreens.com/front/front/cms/slug/understanding-aspect-ratio.

[17] Parlay Studios. *Feature Film Budget Breakdown - Average Cost of Films: Parlay Studios*. Nov. 2017. URL: https://parlaystudios.com/blog/feature-film-budget-breakdown/.

[18] *what is 'survivorship bias' and why does it matter?* Mar. 2016. URL: https://www.vanguard.co.uk/documents/adv/literature/survivorship-bias.pdf.

[19] Kate Whiting. *This is how coronavirus has changed the film and TV industry*. URL: https://www.weforum.org/agenda/2020/05/covid-19-coronavirus-tv-film-industry/.

[20] Wikipedia. *Principal component analysis*. URL: https://en.wikipedia.org/wiki/Principal_component_analysis.

[21] Data World. *IMDB 5000 Movie Dataset - dataset by data-society*. Nov. 2016. URL: https://data.world/data-society/imdb-5000-movie-dataset.

APPENDIX

A

APPENDIX

All the code for this project can be found in the following link: https://github.com/anitamezzetti/EPFL_Data_science_in_practice.git. In particular, at link: https://github.com/anitamezzetti/EPFL_Data_science_in_practice/tree/master/Project/data you will find the data we used. *movie_metadata.csv* is the main dataset, the one we imported at the beginning. *wiki_data.csv* and *tmdb_movies_data.csv* are used in Section 2.3 to fill missing values. In Chapter 5 we imported *data_regression.csv* and *data_regression_onlyUS.csv*, two datasets which have been created in previous steps of our analysis.

A.1 IMDB SCORE PREDICTION

Name	Variables	Timing
1) Basic model	Duration, Budget, Genre, Director rank, Actors' ranks, Movie FB likes, Cast FB likes, Face Number in poster	1980-2019
2) No genre	Duration, Budget, Director rank, Actors' ranks, Movie FB likes, Cast FB likes,Face Number in poster	1980-2019
3) Log budget	Duration, Budget, Director rank, Actors' ranks, Movie FB likes, Cast FB likes,Face Number in poster	1980-2019
4) Only films with likes	Duration, Budget, Genre, Director rank, Actors' ranks, Movie FB likes, Cast FB likes,Face Number in poster	1980-2019
5) Without likes	Duration, Budget, Genre, Director rank, Actors' ranks,Face Number in poster	1980-2019
6) Films from 2009	Duration, Budget, Genre, Director rank, Actors' ranks, Movie FB likes, Cast FB likes,Face Number in poster	2009-2019

Table 13: The combinations of variables for IMDb score prediction

Name	OLS		Ridge		Lasso		Elastic Net		Decision Tree		Random Forest	
Basic model	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
	0.82	0.28	0.74	0.35	0.74	0.34	0.74	0.34	0.59	0.48	0.5	0.56
No genre	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
	0.84	0.32	0.84	0.32	0.89	0.27	0.89	0.28	0.69	0.44	0.64	0.48
Log Budget	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
	0.72	0.36	0.72	0.36	0.75	0.33	0.74	0.34	0.7	0.38	0.58	0.48
Only films with likes	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
	0.8	0.44	0.8	0.44	0.83	0.41	0.82	0.42	0.66	0.53	0.54	0.62
Without likes	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
	0.73	0.38	0.73	0.38	0.76	0.35	0.75	0.36	0.67	0.43	0.59	0.5
Films from 2009	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
	0.64	0.44	0.63	0.45	0.65	0.43	0.65	0.44	0.57	0.51	0.5	0.56

Table 14: Results of the models for IMDb Score prediction

A.2 GROSS-BUDGET RATIO PREDICTION

Ranks+ FB likes				
All movies				
Model	Ridge regression	Decision tree	Random forest	XGBoost
Metrics	MSE	MSE	MSE	MSE
1.1.1)	1.32	1.29	1.22	
1.1.2)	1.18	1.25	1.14	1.09
Only US movies				
1.2.1)	1.14	1.17	1.11	
1.2.2)	0.96	0.93	0.9	
Only FB likes				
All movies				
Model	Ridge regression	Decision tree	Random forest	XGBoost
Metrics	MSE	MSE	MSE	MSE
2.1.1)	1.35	1.33	1.22	
2.1.2)	1.21	1.24	1.18	
Only US movies				
2.2.1)	1.17	1.2	1.14	
2.2.2)	0.97	0.98	0.94	
Only Ranks				
All movies				
Model	Ridge regression	Decision tree	Random forest	XGBoost
Metrics	MSE	MSE	MSE	MSE
3.1.1)	1.32	1.28	1.21	
3.1.2)	1.19	1.22	1.12	
Only US movies				
3.2.1)	1.13	1.17	1.1	
3.2.2)	0.96	0.94	0.9	

Table 15: Results for the Regression Models for the Gross-Budget Ratio Prediction

Ranks+ FB likes				
All movies				
Model	Logistic regression	Decision tree	Random forest	KNN
	Accuracy score	Accuracy score	Accuracy score	Accuracy score
1.1.1)	0.47	0.52	0.54	0.52
1.1.2)	0.47	0.52	0.53	0.5
Only US movies				
1.2.1)	0.38	0.5	0.48	0.48
1.2.2)	0.46	0.5	0.49	0.46
Only FB likes				
All movies				
Model	Logistic regression	Decision tree	Random forest	KNN
	Accuracy score	Accuracy score	Accuracy score	Accuracy score
2.1.1)	0.42	0.53	0.52	0.51
2.1.2)	0.45	0.52	0.53	0.51
Only US movies				
2.2.1)	0.4	0.5	0.51	0.49
2.2.2)	0.43	0.49	0.5	0.45
Only Ranks				
All movies				
Model	Logistic regression	Decision tree	Random forest	KNN
	Accuracy score	Accuracy score	Accuracy score	Accuracy score
3.1.1)	0.43	0.52	0.54	0.51
3.1.2)	0.46	0.52	0.52	0.52
Only US movies				
3.2.1)	0.39	0.5	0.5	0.48
3.2.2)	0.45	0.51	0.5	0.46

Table 16: Results for the Multi-Class Classification Models for the Gross-Budget Ratio Prediction

Ranks+ FB likes												
All movies												
Model	Logistic regression			Decision tree			Random forest			KNN		
	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score
1.1.1)	0.86	0.53	0.18	0.54	0.55	0.57	0.65	0.59	0.53	0.69	0.58	46
1.1.2)	1	0.52	0.00	0.54	0.55	0.57	0.65	0.63	0.59	0.71	0.59	0.45
Only US movies												
1.2.1)	0.55	0.62	0.68	0.68	0.57	0.47	0.55	0.62	0.68	0.55	0.55	0.55
1.2.2)	0.51	0.62	0.71	0.68	0.57	0.48	0.53	0.6	0.66	0.56	0.55	0.53
Only FB likes												
All movies												
Model	Logistic regression			Decision tree			Random forest			KNN		
	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score
2.1.1)	0.87	0.53	0.17	0.54	0.55	0.57	0.64	0.6	0.54	0.69	0.58	0.46
2.1.2)	1	0.52	0.00	0.54	0.55	0.57	0.66	0.61	0.55	0.66	0.56	0.45
Only US movies												
2.2.1)	0.52	0.6	0.66	0.55	0.54	0.53	0.53	0.6	0.65	0.57	0.56	0.54
2.2.2)	0.44	0.58	0.70	0.57	0.54	0.51	0.52	0.58	0.64	0.55	0.55	0.55
Only Ranks												
All movies												
Model	Logistic regression			Decision tree			Random forest			KNN		
	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score	Specificity score	Accuracy score	Recall score
3.1.1)	0.78	0.52	0.24	0.53	0.55	0.57	0.66	0.62	0.58	0.7	0.58	0.46
3.1.2)	1	0.52	0.00	0.53	0.55	0.57	0.65	0.61	0.58	0.69	0.56	0.43
Only US movies												
3.2.1)	0.54	0.62	0.70	0.52	0.54	0.55	0.53	0.62	0.70	0.53	0.55	0.57
3.2.2)	0.47	0.59	0.69	0.52	0.55	0.58	0.58	0.62	0.66	0.61	0.56	0.51

Table 17: Results for the Binary Classification Models for the Gross-Budget Ratio Prediction