CAN BOX OFFICE SUCCESS BE PREDICTED BEFORE A MOVIE'S RELEASE?

By Aleksandra Zinoveva

Submitted to

Central European University

Department of Economics and Business

In partial fulfilment of the requirements for the degree of Master of Arts in Economics

Supervisor: Professor Arieda Muço

Budapest, Hungary

2019

Abstract

Due to the motion picture industry being a very risky one with high sunk costs, film studios executives around the world are trying to find better ways to predict whether their films are going to be successful even as early as in the initial stages of development. As only the big players, such as Hollywood studios, have unobstructed access to the real market data, in this thesis I investigate the possibility of predicting a movie's success based on publicly available data on previously released films, which is usually the only information an independent film producer is able to obtain. For this, I fit a Logistic Regression model to the data for 4,727 movies released in 1982-2016. The results show that predicting algorithm does not perform very well, correctly classifying only 62% of films when working with this limited information; however, such factors as movie's budget, facts of belonging to a franchise or being produced with participation of an American company, have proven themselves to be important for distinguishing between potentially successful and unsuccessful films.

Table of contents

List of figures and tables

Introduction

While over 800 films are released every year only in the US and Canada[1], just a few of them manage to break-even and become successful in the theatrical release. And every year both major studios and small production companies are pressured to spend more and more money on production and marketing while having no guarantee of recouping their investments. It is safe to say that the filmmaking industry is and most likely will stay very risky for its participants in the near future. Because of this, filmmakers around the world are looking for ways to better predict the box office earnings of their movies.

Even though a lot of research is done prior to the production of any film, it's still a difficult task to accurately estimate the probability of a film becoming a success. Some industry players, for example, the major Hollywood studios known as the Big Five[2], are able to do this more or less efficiently because they have exclusive access to relatively full and reliable market data due to their unique position on the market (Finney 2010); however, they still prefer to diversify their film "portfolio" because the risk of failure is too high, and only a small fraction of films brings them profit, which in its turn helps big studios cover their losses from the failed movies (Elberse 2013).

Unfortunately, this kind of access is unavailable to the majority of independent film producers. Of course, one can try to improve the odds of a film becoming a success by, for example, carefully choosing directors, producers, and actors, or giving it a wider release. However, such opportunities may be limited for smaller production companies or lower budget movies. But, in any case, after a movie hits the cinemas, it's the audience who determines whether it's going to be a hit (De Vany & Walls 1999).

[1] Based on Box Office Mojo yearly data available here: https://www.boxofficemojo.com/yearly/
[2] They are Universal Pictures, Paramount Pictures, Warner Bros. Pictures, Columbia Pictures and Walt Disney Studios.

Many previous researchers, such as Greenaway & Zetterberg (2012) and Pangarker & Smit (2013), studied how different determinants affect a film's theatrical revenue which is seen as the main indicator of the box office success. However, the size of the revenue does not solely constitute a film's success because there can be cases when several films have approximately the same revenue but completely different budgets and, consequently, different returns on investment, so they could be very different in terms of success and failure as well.

In this thesis, I concentrate on evaluating the possibility of predicting a movie's success before its theatrical release using only publicly available market data. To do this I create a binary outcome variable indicating whether a movie was successful in the box office based on its budget and revenue and fit a Logistic Regression to the data collected on 4,727 movies released in the period between 1982 and 2016. I also explore which of the 39 available independent variables are the most important for distinguishing between successful and unsuccessful films and evaluate how well the assembled model deals with assigning a film to one of these two groups. According to my findings, the described model is not able to distinguish between successful and unsuccessful movies well. However, this is consistent with the conclusions of other authors about the difficulties in predicting a movie's success (DeVany & Walls 1991).

The structure of this thesis is as follows. In Chapter 1, I introduce the main characteristics of the business of film and discuss some prominent works on the determinants of a movie's success. In Chapter 2, I give a description of the data and in Chapter 3 I present the methodology and the results. Finally, in the Conclusion, I briefly summarize my findings.

Chapter 1. Theoretical background and literature review

The entertainment industry is one of the most important ones in the world and it is also one of the biggest ones in the US. Americans spend more than 140 billion hours and more than $280 billion annually on legal forms of entertainment, and around the world this number is at least one trillion dollars (Vogel 2011).

Any sort of entertainment, be it music, movies, literature, is defined by its effect, namely, "a satisfied and happy psychological state"; however, many forms of entertainment are created not because of art for art's sake, but rather for profitability (Vogel 2011, xx). Therefore, they are also the subjects of economic analysis.

The motion picture industry, with its heart located in Hollywood, is a crucial part of the entertainment industry, and its importance for the current world economy is hard to overestimate. Global box office in 2018 surpassed $41.1 billion (that is more than nominal GDP of Tunisia in the same year[3]), continuing to show steady growth compared to the previous years (MPAA 2019a). The same year the film & TV industry in the USA only supported about 2.5 million jobs, paid around $177 billion in total wages and $44 billion to the local businesses across the country, and about three-quarters of the country's population went to the cinema at least once (MPAA 2019b).

The main features usually attributed to this industry are high sunk costs, prototypical production, high unit costs of production, high and increasing marketing costs, high probability of failure and short and very competitive shelf life in the primary market (namely, cinemas) with little to no price differentiation and, at the same time, unpredictability of demand (Finney 2010). This unique combination of characteristics makes the industry extremely risky.

[3] According to International Monetary Fund's estimation. For more information see: https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/index.aspx

It's not a secret that the Big Five studios practically constitute an oligopoly in the US market with considerable barriers to entry. These studios, of course, enjoy their advantageous positions in the market, and they also have symbolic relationships with independent film production companies, as it gives them access to the pool of new talent, fresh ideas, reduction in costs and connection with changing consumer preferences (Finney 2010).

Considering the fact that movies have to compete with other forms of entertainment for the limited leisure time of people, which has been decreasing since 1980 in the US and some European countries, and between each other in the theatres as there are a fixed number of screens in the world and a limited number of timeslots a day, the emergence of this oligopoly is just a reaction to the economic conditions of the business (Finney 2010). Hundreds of movies are getting a theatrical release each year, but only a few of them manage to break-even and make some profit. So, with the uncertainty the film producers are facing, vertical and horizontal integration, portfolio investment, capital concentration, and clustering helps major studios with mitigating the risk (Finney 2010). This sort of integration allows the big studios to implement a so-called tent-pole or blockbuster strategy when one or a few big successful films of a studio are supporting its financial performance and cover the losses incurred from failed movies (Elberse 2013).

Nevertheless, risk mitigation also has its limits. It is predicted by Ampere Analysis that in 2019 over the top (OTT) media services will overtake the theatrical movie grosses for the first time in history. Subscription video on demand already has higher revenues than cinema in the US, and it's a matter of time until it happens in the EU and China as well. Ampere Analysis forecasts OTT to reach as high as $46 billion this year, while worldwide box office revenues are only predicted to be $40 billion (Pennington 2019). So, with the growing market share of digital forms of entertainment, such as online streaming services, film producers are under

greater pressure than ever before. Hence, the ability to predict a film's theatrical performance has even more importance now.

Every studio requires the producer to do thorough market research before a film is made. No financier will ever 'green-light' a film, that is, decide to invest in it and put into production, before getting full information about the film's budget, cast, director, producer, financial plan, pre-sales data, remaining sales estimates and estimated date of delivery (Finney 2010). Yet, a lot of times they invest into films that later fail at the box office.

A lot of researchers have tried to solve the mystery of what has an influence on the movie's financial box office success. Some of them concentrated on some specific factors, such as star power (Elberse 2007) or movie contents (Garcia-del-Barrio & Zarco 2017), while others tried to see the picture more broadly by considering the influence of multiple determinants in their research. For example, Pangarker and Smith (2013) studied the effects of several determinants, such as production costs, genres, critics' rating and others on a film's global box office revenue using, as many others, multiple regression model. They found that the most significant contributor to the box office revenue is production cost. A release by a major production studio also was found to have a positive effect on grosses. The authors also claimed that award nominations had a significant effect on box office revenues; however, it's not clear how this is possible, because the race for the main awards starts late in a year, and most of the films (except the ones which were specifically targeted for, let's say, Oscar nominations) are out of the theatres by then. So, this factor may be more instrumental for boosting DVD or VoD sales, rather than cinema tickets sales. However, there may be exceptions, like in the cases of late-in-a-year releases or special re-releases for awards nominated films.

Greenaway & Zetterberg (2012), in their paper *Success in the Film Industry: What Elements Really Matter in Determining Box-Office Receipts*, concluded that positive reviews and a large number of screens for a premiere positively influence box office performance. Also,

if a movie was released in summer or during the holiday season, it had a better chance at success as well. Additionally, the authors found positive correlations between films being a derivative work (or "based on previous source material", as they put it) or a comedy and its theatrical revenues. At the same time, Greenaway & Zetterberg considered MPAA ratings[4] to have no effect whatsoever.

Nevertheless, while measuring the effects of different factors on theatrical revenues is an important task, the revenues alone don't constitute the financial success of a film in cinemas. Also, for production companies' executives, it is important to have an ability to predict, at least with some accuracy, a movie's box office performance when it hasn't been released or even produced yet, in order to minimize the risks emerging in connection with investing into film production. Therefore, use of such factors as critics' reviews, or continuously changing audience's scores, or other factors that are not possible to foresee before the film's release, is unjustified and literally impossible at this stage.

In the next chapters, I approach the problem by trying to predict whether the film is going to be successful based on both revenues and production budgets using only publicly available information. Despite OLS being the top choice for many researchers, for my analysis I use a Logistic Regression machine learning model to find out if it's possible to make such a prediction and with what accuracy.

[4] For more information on MPAA rating system see: https://www.mpaa.org/film-ratings/

Chapter 2. Data

The data used for the analysis was collected from 2 different sources. Some of it was provided by the *Internet Movie Database (IMDb)*[5], an online database for movies, television, and video games. This dataset includes information on 9 descriptive variables, such as titles, years of release, runtimes, genres, and other information, for 5,676,835 movies.

The rest came from *The Movies Dataset*[6], which includes metadata listed in the Full MovieLens Dataset on over 45,000 movies released before July 2017. This dataset consists of 24 variables, such as budget, revenue, release date, languages, production companies, and others.

Both original datasets were merged together using the common column containing films' unique IMDb IDs. For the purposes of this thesis, I am only looking at films released between 1982 and 2016. The reason for this is that data on films released after 2016 is incomplete in *The Movies Dataset*, and data on films released in 1981 and 1980 is partial and self-selected, as it features only the major releases[7], and, therefore, its inclusion may skew the results. The reliable data for films, specifically grosses and budget, released before 1980 is unavailable. Thus, the final dataset includes 4,727 observations.

Dependent variable

Unlike many researchers, who used a movie's revenue as a dependent variable for analysis of the success of the box office performance (Pangarker & Smit 2013), I use a dummy variable, which takes value 1 if a movie's theatrical release was successful and 0 otherwise. There are many ways to define the box office success, but for the purposes of this thesis I define

[5] IMDb Datasets: https://www.imdb.com/interfaces/
[6] The Movies Dataset on Kaggle: https://www.kaggle.com/rounakbanik/the-movies-dataset
[7] According to BoxOfficeMojo Yearly Box Office data: https://www.boxofficemojo.com/yearly/

it the following way: if a movie's worldwide grosses are at least 250% of the production budget, then it's successful. This number was not chosen randomly.

First, the production company doesn't get the whole sum that a movie takes in from its theatrical run. Usually, the revenues are split between the production company (or the distributor) and the exhibitor according to a film licensing agreement. The structure of the rental fee is usually determined by the greater of two amounts calculated by two methods: *the gross receipt formula* (for a new film, the distributor usually gets around 60-70% in the first week of release, and then this share is gradually decreasing to approximately 30% after 4-7 weeks) or *the adjusted gross receipt formula* (the distributor gets 90% of the box office revenues after deduction of the exhibitor's expenses) (Finney 2010). The information about specific agreements, however, is not available to the public. Therefore, for the purposes of this research, I assume that the production company gets on average 50% of overall box office revenues. According to that, to recoup the production costs a movie should earn at least twice as much as its budget in the cinema, hence it should make 200% of the budget.

Secondly, the company also spends some amount of money on advertising a film. So, it is also very important to recoup that as well. This information is also not disclosed to the public. However, by the rule of a thumb, the advertisement costs are usually calculated as 50% of a production budget (Vogel 2011).

Explanatory variables

To make the model for prediction of the box office success usable in practice I chose to only use parameters available to the decisionmakers before the film's release. Therefore, I selected 39 explanatory variables for my analysis. However, as I have access only to the publicly available aggregated information, my model may turn out to be far from perfect at the actual prediction of the box office performance.

Based on the research of many economists, it's evident that a movie's production ***budget*** plays a big role in the movie's financial success (Pangarker & Smit 2013). One can expect that with the increase in budget we may observe the increase in a movie's grosses, but it may only be true until some point. Huge production costs make it very difficult for a film to break-even; also, studios have to increase their advertisement costs respectively to offset the risk related to the growth in the budget. And there are a lot of examples when big-budget films failed in cinemas: e.g., *John Carter* (2012), *47 Ronin* (2013), *The Lone Ranger* (2013), *Green Lantern* (2011), *Jupiter Ascending* (2015), and many others (Looch 2018). So, even though I already used this variable to construct the dependent one, I consider it to be a good explanatory variable to examine when trying to answer my research question. The data on budgets was gathered from The Movies Dataset and quoted in US dollars.

In some countries, for example Russia (Derkacheva 2018) and Turkey (Smiers 2003), there is a strong opinion that Hollywood is "destroying" competition on movie markets around the world, that the national films don't get enough revenue simply because they are released on the same day or relatively close to the American films. But the issue is clearly more complicated than that. Moreover, one can name multiple reasons why certain movies perform better in the theaters than the others, and it is unlikely "national identity" on its own is one of them. However, there is an opinion that films produced by Hollywood are chosen more often by people around the world compared to the films produced by other countries because of the quality of production and, as Pardo (2007) mentioned, their international nature, the way that the story of a film can resonate with people from different countries. So, theoretically, the fact of a film being produced by or in cooperation with an American production studio can serve as some sort of proxy for these qualities. Therefore, I decided to include a dummy variable (***US_produced***) indicating the involvement of the US companies in the production of movies

into my analysis to check whether this fact can indeed have an effect on a movie's box office performance.

I also decided to check if the original language of a movie influences the odds of becoming successful. Because it's almost impossible to control for every known language in the world, I decided just to control for one – ***English***. As evidence shows, it has a clear dominance among the most popular films released in the recent years (UNESCO 2016).

The other interesting factor to consider is whether a film belongs to a franchise. In the past decades, there has been an increase in the number of various prequels, sequels, and spinoffs being produced. Moreover, about 70% of the #1 movies in the domestic market (the US & Canada) in the last 20 years were sequels[8]. Snow (2016) argued that this may even be a financially safer way to operate in the industry. Of course, this finding is not baseless: if some film is a continuation of an already released successful story, it may count on a pretty large portion of fans of the previously released film to watch it. And, based on the already known reaction to the previously released film and its performance, one can easily try to make a guess about how the sequel movie will perform. This is why I also included a binary variable ***franchise***, which takes the value of 1 if a film belongs to a franchise and 0 otherwise.

The next parameter considered in the model is a movie's runtime (***runtimeMinutes***), which stands for the length of a movie. It's argued in the literature that the longer films tend to have a higher quality than the shorter ones; however, it's better for cinema owners if they are shorter, because this way they could increase the number of films shown within one day (Follows 2017). Also, with the longer films, usually there is a problem with keeping the audience's attention; however, there can be some exceptions, such as recently released *Avengers: Endgame* (2019), which runs for over 3 hours but still managed to hold the #1

[8] Based on yearly box office revenues. For more information see: https://www.boxofficemojo.com/yearly/

position in the box office for several weeks[9]. So, in most of the cases, the companies are being pressured to make their films shorter. Therefore, one can expect the length of a film to have an effect on the film's performance. The data was collected from the IMDb datafile and quoted in minutes.

It is also important to determine if a movie's genre has an effect on its box office performance. It seems obvious that different genres attract different audiences and different amounts of people; for example, an action movie and a documentary may have a significant difference in their earnings. The dataset from IMDb features one or multiple genres for every movie. And the following 22 unique genres were identified by using text analysis tools and the corresponding dummy variables were created: ***action, adventure, animation, biography, comedy, crime, documentary, drama, family, fantasy, history, horror, music, musical, mystery, news, romance, sport, thriller, war, western, sci-fi***.

Finally, as many researchers have found before (e.g., Garcia-Del-Barrio & Zarco 2017; Einav 2007), the timing of a movie's release may also influence its box office performance. Some authors use the proximity to the public holidays as the explanatory variables (Pangarker & Smit 2013); however, for the purposes of this thesis, I will use 12 dummy variables for each ***month of the year***. The date of release information was extracted from Kaggle's The Movies Dataset.

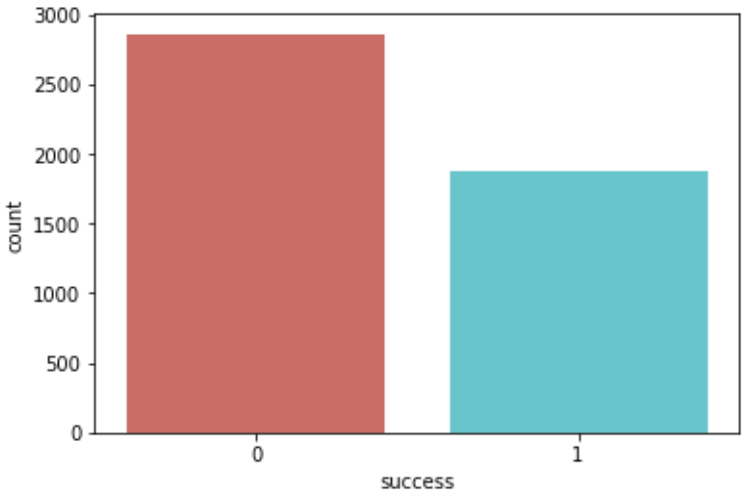
[9] Based on weekly box office data. For more information see: https://www.boxofficemojo.com/weekly/

Chapter 3. Methodology & analysis

Given that my outcome variable is a dummy one and the fact that I want to estimate the probability of any movie becoming successful, it seems reasonable to consider Logistic Regression as a suitable model for my prediction.

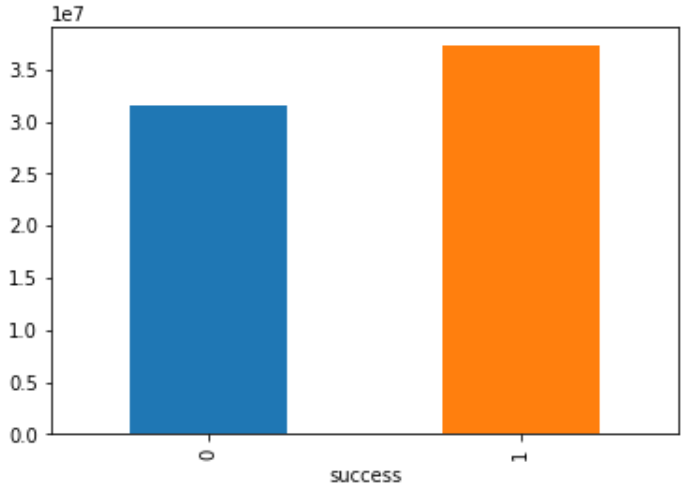
I started my analysis with a large database, containing data on 4,727 films released in the period from 1982 to 2016.

Figure 1. Value count for the outcome (target) variable



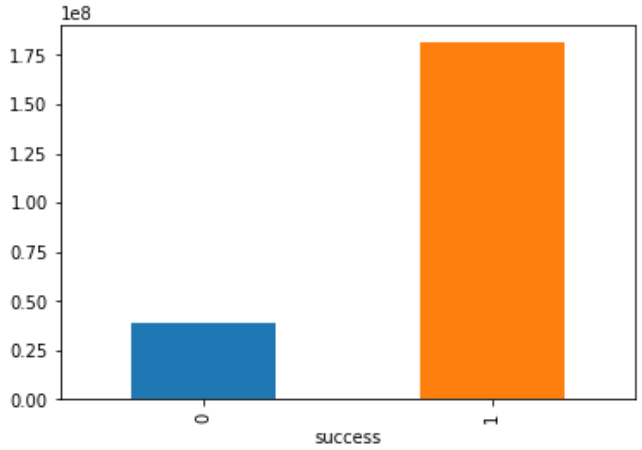
First, I checked whether my data was balanced by looking at the bar plot of the target variable ***success***. The result is in Figure 1. As the plot suggests, the classes are unbalanced: there are 2,859 (60.48%) unsuccessful and 1,868 (39.52%) successful movies in the dataset. That means that for further analysis I would need to balance the data. But before I proceeded with that, I did some more exploration.

Figure 2. Mean budgets by value of the outcome variable



I started by looking at the difference in budgets between successful and unsuccessful films. The bar plot of mean values of production costs by values of the outcome variable ***success*** can be found in Figure 2. As seen in the figure, on average successful movies tend to have higher budgets than the movies that failed in their theatrical run. The mean value for the budget of successful movies in this dataset is $37,232,680, for unsuccessful ones $31,585,900. The difference between these two values is $5,646,780. For comparison, a few low budget films can be produced for that sum.

Figure 3. Mean box office revenues by the outcome variable values



Even though I wasn't using ***Revenue*** as an independent variable for the analysis, it still has an influence on the outcome variable, so it was important to look at the difference between the box office revenues of successful and unsuccessful films. Figure 3 shows a more drastic

difference between successful and unsuccessful movies. On average, a successful movie earned $181,264,900, and the unsuccessful movie only $38,946,180. The difference is $142,318,720, which is more than 3.5 times the mean revenue of an unsuccessful movie.

Figure 4. Histogram of Budgets

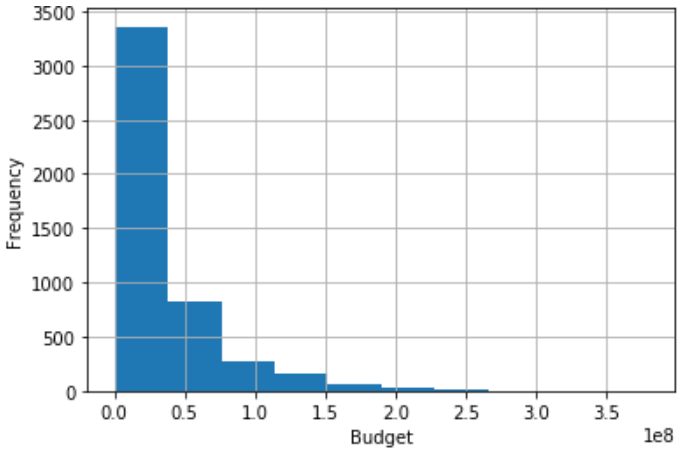
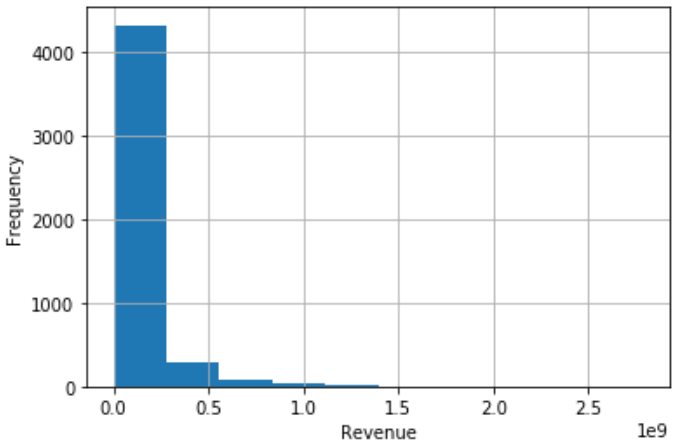
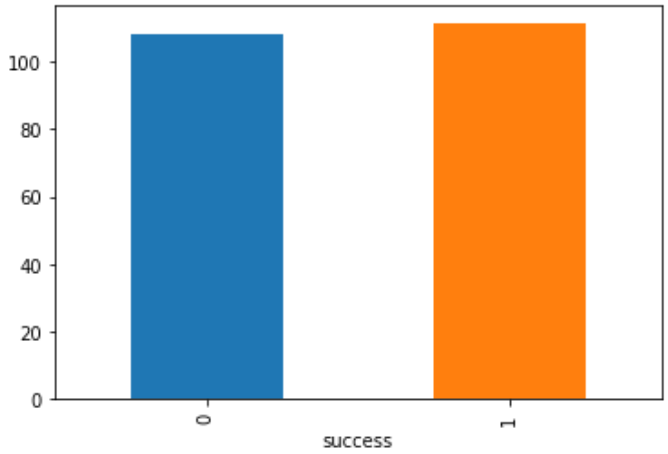


Figure 5. Histogram of Revenues



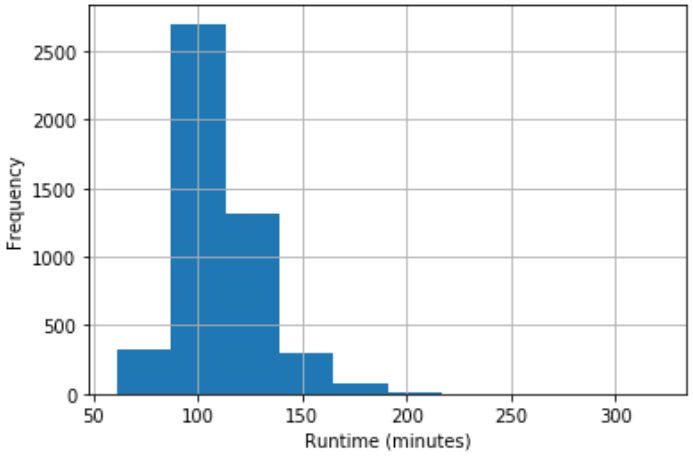
Then I looked at the distributions of both budgets and revenues in the dataset. The histograms are in Figures 4 and 5. As it can be seen in these figures, distributions of budgets and revenues are very similar: they are both extremely skewed to the right, with a lot of small values and a handful of large ones. That means that many films have small production budgets and small box office revenues, and only a small number of films have big budgets and big revenues.

Figure 6. Runtime by values of the outcome variable



The only remaining continuous variable to look at was a film's runtime. Its plot by the successfulness of a movie can be found in Figure 6. We can see that the average runtimes for both categories are approximately the same with a difference of less than 3 minutes. Because of this fact, this variable may not be a good parameter for classification, but I decided not to remove it from the dataset.

Figure 7. Histogram of Runtime



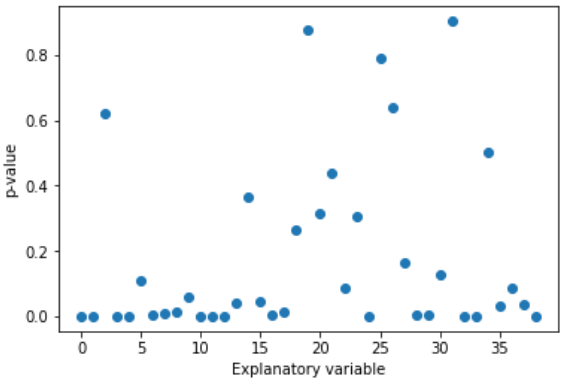
A histogram of Runtime (***runtimeMinutes***) across the dataset can be seen in Figure 7. This histogram is skewed to the right as well. Therefore, it would be reasonable to scale the variables when fitting the Logistic Regression model.

I continued the analysis by dividing the existing dataset into 2 sub-sets, training and testing, using one of the most popular 80:20 proportions and keeping the proportion of successful and failed movies in each sub-set the same as in the complete one. This way I got 3,781 observations in the training set and 946 observations in the testing set.

As I mentioned above, the dataset I used was unbalanced, and because the classification algorithms usually use the majority rule, it could negatively affect the process of classification. Therefore, some transformations were in order. In this thesis I chose to use the technique known as SMOTE, or Synthetic Minority Over-Sampling Technique. The main idea of this technique is that to oversample a minority class, a sample from the dataset is taken, and its k nearest neighbors in feature space are considered. Then to create a synthetic data point, the vector between one of those k neighbors and the current data point is taken. This vector is then multiplied by a random number X which lies between 0, and 1, and this is added to the current data point to create the new, synthetic data point (Chawla et al 2002).

It is important to mention that this oversampling is only used on the training set, while the testing set remains untouched. Before applying SMOTE, my training set contained 1,494 successful movies and 2,287 unsuccessful ones. After oversampling, each of both groups contained 2,287 datapoints, resulting in the total size of training set equal to 4,574. So after that I had a completely balanced sub-set that I could use to train my model on.

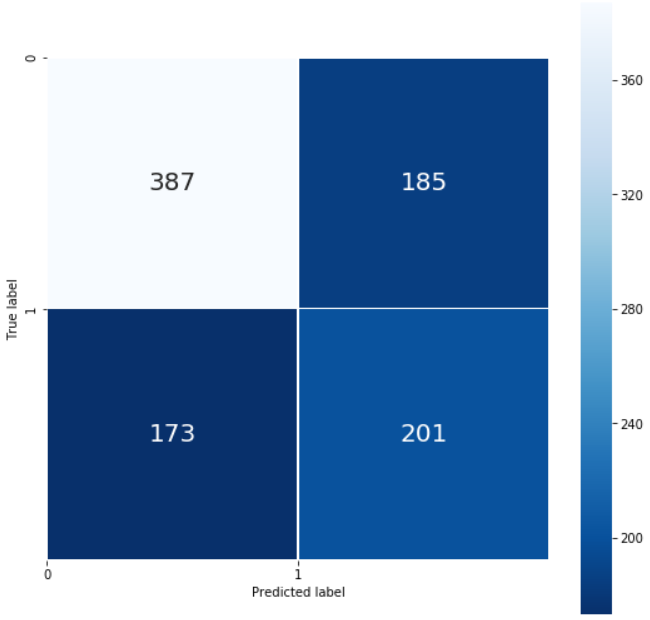
Figure 8. p-values by explanatory variable



As the use of all 39 explanatory variables for classification may not be entirely justified, I decided to leave only 50% of them based on their informativeness. To do that I used the F test for classification (f_classif). That test returned the p-values for all explanatory variables. They can be seen in Figure 8, where the explanatory variables are numerated in order of their appearance in the dataset. In the figure, the smaller p-values indicate more informative variables for classification. These are the variables with the smallest p-values that made through the selection process: ***budget, US_produced, franchise, runtimeMinutes, adventure, animation, biography, crime, documentary, drama, horror, music, war, RM_Aug, RM_Dec, RM_Jul, RM_Jun, RM_May, and RM_Sep***.

Using pipeline with Standard Scaler, I next found the best parameters for my Logistic Regression model via Grid Search and fitted the data. With these parameters, I got the best test score equal to 0.6216 and the best 10-fold cross validation score equal to 0.61.

Figure 9. Confusion matrix



The confusion matrix for this model is shown in Figure 9. From this figure it can be seen that my model's performance was not particularly good: only in 387 instances did it correctly predict an unsuccessful film, while in 185 instances it mistakenly classified them as

successful. Even worse situation can be observed with prediction of the successful movies when 201 films were classified correctly and 173 were misclassified as unsuccessful.

Table 1. Classification report

	precision	recall	f1-score	support
0	0.69	0.68	0.68	572
1	0.52	0.54	0.53	374
micro avg	0.62	0.62	0.62	946
macro avg	0.61	0.61	0.61	946
weighted avg	0.62	0.62	0.62	946

The classification report, shown in Table 1, provides more details. From this table, it can be seen that the weighted average precision in the model's prediction is very low – 62%. At the same time, the model can predict unsuccessful films better than successful ones, for which the prediction is close to being as good as the random assignment of the value. The recall, i.e. the percentage of points classified correctly, is not any better, so it's evident that the classifier works very poorly.

As the next step, I decided to narrow the dataset only to the films released in the last decade, for which full information is available, i.e. 2007-2016; this shrunk the sample size to 2,181 observations. The reason to look at the shorter period of time is, first, that the audience's preferences regarding a movie's genre and contents change over time (for example, if some company was to release a vampire-themed movie right now, it would probably perform worse than if it was released at the time when The Twilight Saga was trending. i.e. 2008-2012). Moreover, the distribution of people's leisure time may be completely different from the one people had in the '80s or '90s, for instance. So, it could be worth trying to decrease the time interval to make the sample more homogeneous in that sense, repeat all previous steps, and see if anything changes in the result.

As it turns out, not many things changed. The shares of successful and unsuccessful films remained approximately the same (40.12% and 59.88% respectively). The average budget for failed movies was still significantly lower than for the successful movies, but the difference

increased to approximately $13.7 million. Also, unsurprisingly, the average revenues for the successful films were still higher than for the unsuccessful ones, and the gap increased to approximately $170 million. At the same time, the difference in the runtimes stayed the same as with the larger sample.

I divided the smaller dataset into 2 sub-sets by the same principle as before and got 1,744 observations in the training set and 437 in the testing set.

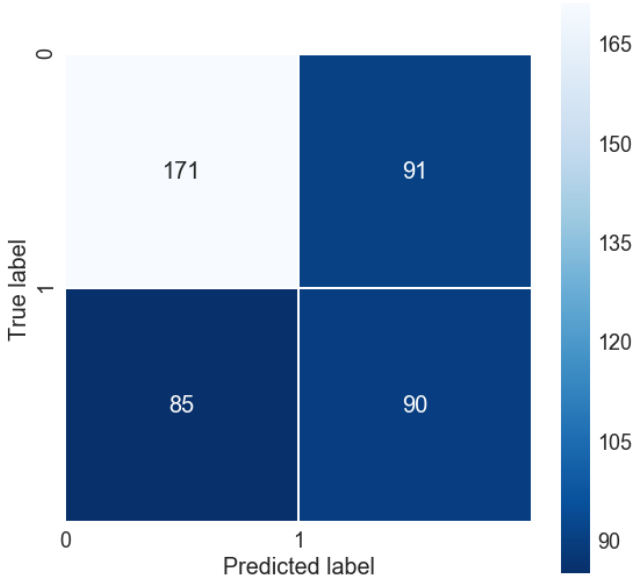
Before using SMOTE, the new training set contained 700 successful movies and 1,044 unsuccessful ones. After oversampling, each of both groups contained 1,044 observations, resulting in the increase of the size of the training set to 2,088 observations.

Again, I decided to leave only 50% of the explanatory variables based on their informativeness for the classification purposes. After performing this kind of feature selection, I was left with only the following explanatory variables that had the smallest p-values: ***budget***, ***US_produced***, ***franchise***, ***runtimeMinutes***, ***adventure***, ***animation***, ***comedy***, ***crime***, ***drama***, ***horror***, ***music***, ***musical***, ***war***, ***sci-fi***, ***RM_Dec***, ***RM_Jan***, ***RM_Jul***, ***RM_Jun*** and ***RM_Sep***. It is important to note that here I've got a different set of variables to the one that was selected for the larger dataset.

Then, using pipeline with Standard Scaler again, finding the best parameters for the model and fitting the data, I got the best test score equal to 0.5973 (lower than before) and the best 10-fold cross validation score equal to 0.62 (a bit higher than before).

The confusion matrix for the smaller sample is presented in Figure 10. From this figure, it is evident that the model was not good at predicting whether the film was going to be a success in this case as well.

Figure 10. Confusion matrix (smaller sample)



The new classification report can be seen in Table 2, where it's evident that the model's performance on a smaller sample is even worse than on a bigger one. Hence, the assumption made earlier about the changes in the audience's preferences and the leisure time distribution was wrong and narrowing down the sample didn't lead to the improvement in classification.

Table 2. Classification report (smaller sample)

	precision	recall	f1-score	support
0	0.67	0.65	0.66	262
1	0.50	0.51	0.51	175
micro avg	0.60	0.60	0.60	437
macro avg	0.58	0.58	0.58	437
weighted avg	0.60	0.60	0.60	437

Conclusion

In this thesis, I predict a movie's success based only on the information available to the public. To do that I fit a Logistic Regression model to the data on 4,727 movies released in the period between 1982 and 2016. However, the resulting model doesn't show a very good performance: the prediction of success was only slightly better than the result of a coin toss would be. And there may be a few reasons for that.

First, the dataset used for this research is far from perfect. The data on budgets is not without the measurement error because it is never reported precisely to the dollar, and in some of the cases only the estimates made by the experts are indicated. The advertisement budgets calculations are very approximate as well and may not reflect the real situation; however, without access to this privileged information it's impossible to come up with numbers reflecting the reality more precisely.

The dataset also didn't have any information about, for example, MPAA ratings that could give more information about the targeted audiences of a film. Information on the presence of the most popular actors or Oscar-nominated directors on board was also missing; however, it could contribute to the explanation of why people chose some films to the other, i.e. contributed to some films' revenues by going to the cinema to see them.

Also, it's very difficult to call the final dataset randomized, as it included only films with present information for theatrical revenue and budget, and usually production companies refrain from disclosing production budget information to the public, especially if their film's box office performance was poor (Garcia-del-Barrio & Zarco 2017).

All in all, it seems like only the executives of the Big Five Hollywood studios have a better chance to predict whether their films will be successful, and not just because they have almost exclusive access to better market data, but also because they have a lot of precise and

complete historical data on a lot of films released on their own, so they can use a better-quality numerical data to base their predictions on. These studios also have a lot of money to spend on preliminary research, and high market power that allows them to involve a lot of professional and talented people on their projects. However, even their predictions sometimes are absolutely wrong.

Nevertheless, the results of this thesis are consistent with the findings of the other researchers. In particular, as DeVany & Walls (1999, 2) concluded, "movies are complex products and the cascade of information among film-goers during the course of a film's run can evolve along so many paths that it is impossible to attribute the success of a movie to individual causal factors". This means that more work is needed to improve the prediction models. And, as long as the film industry stays a very risky and a very expensive one, the studios and the film researchers will continue searching for better ways to predict a film's box office performance, especially before it is released.

Reference list

Box Office Mojo. 2019. Yearly Box Office. http://www.boxofficemojo.com/yearly/ (accessed May 9, 2019).

--- 2019. Weekly Box Office. https://www.boxofficemojo.com/weekly/ (accessed May 25, 2019).

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002): p. 321–357.

Derkacheva, Alena. 2018. Vladimir Medinskiy – ob unichtozhenii rossiyskogo kino "vsemirnoy mashinoy Gollivuda" (Vladimir Medinskiy – about destruction of Russian cinema by "global Hollywood machine"). https://www.the-village.ru/village/city/news-city/333937-hollywood (accessed on May 25, 2019).

De Vany, Arthur and W. David Walls. 1999. Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? *Journal of Cultural Economics* 23(4): 285-318.

Einav, Liran. 2007. Seasonality in the U.S. motion picture industry. *RAND Journal of Economics* 38, no. 1 (Spring 2007): 127-145.

Elberse, Anita. 2007. The Power of Stars: Do Star Actors Drive the Success of Movies? *Journal of Marketing* 71, No. 4 (Oct., 2007): 102-120.

--- 2013. *Blockbusters: Hit-making, Risk-taking, and the Big Business of Entertainment*. New York: Henry Holt and Company.

Finney, Angus. 2010. *The international film business: a market guide beyond Hollywood*. London; New York: Routledge.

Follows, Stephen. 2017. Is a film's length a sign of its quality? https://stephenfollows.com/films-length-sign-quality/ (accessed April 20, 2019).

Garcia-del-Barrio, Pedro and Hugo Zarco. 2017. Do movie contents influence box-office revenues? *Applied Economics* 49:17: 1679-1688.

Greenaway, Melissa and Barrett Zetterberg. 2012. Success in the Film Industry: What Elements Really Matter in Determining Box-Office Receipts. *Science and Social Sciences*. Submission 34.

International Monetary Fund. 2019. World Economic Outlook Database. https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/index.aspx (accessed May 25, 2019).

Looch, Cassam. Big Budget Movies That Didn't Break Even. https://theculturetrip.com/europe/united-kingdom/articles/big-budget-movies-didnt-break-even/ (accessed May 21, 2019).

Motion Pictures Association of America. 2019a. 2018 Theatrical Home Entertainment Market Environment (THEME) Report. https://www.mpaa.org/wp-content/uploads/2019/03/MPAA-THEME-Report-2018.pdf (accessed April 11, 2019).

--- 2019b. The Economic Contribution of the Motion Picture & Television Industry to the United States. https://www.mpaa.org/research-docs/the-economic-contribution-of-the-motion-picture-television-industry-to-the-united-states- 2/ (accessed April 11, 2019).

Pangarker, N.A. and E.v.d.M. Smit. 2013. The determinants of box office performance in the film industry revisited. *South African Journal of Business Management* 44, Issue 3 (Sep 2013): 47-58.

Pardo, Alejandro. 2007. The Europe-Hollywood Coopetition: Cooperation and Competition in the Global Film Industry. *Media Markets Monographs*, no. 8.

Pennington, Adrian. 2019. Streaming vs cinema: What does the future hold for film? https://www.ibc.org/industry-trends/streaming-vs-cinema-what-does-the-future-hold-for-film/3517.article (accessed on May 25, 2019).

Smiers, Joost. 2003. *Arts under pressure: promoting cultural diversity in the age of globalization*. London: Zed Books.

Snow, Shane. 2016. Why Hollywood Makes More Sequels Every Year (Even Though We Like Them Less). https://www.linkedin.com/pulse/why-hollywood-makes-more-sequels-every-year-even-though-shane-snow/ (accessed May 6, 2019).

UNESCO. 2016. Diversity and the film industry: An analysis of the 2014 UIS Survey on Feature Film Statistics. *Information Paper*, No. 29. http://uis.unesco.org/sites/default/files/documents/diversity-and-the-film-industry-an-analysis-of-the-2014-uis-survey-on-feature-film-statistics-2016-en_0.pdf (accessed April 29, 2019).

Vogel, Harold L. 2011. *Entertainment industry economics: a guide for financial analysis*. 8th ed. New York: Cambridge University Press.