CS591: Data Science with Python
Nisa Gurung & Kristel Tan
Submission Date: December 14, 2016

Movie Gross Profit Analysis: What makes a movie successful?

Abstract

The main goal of this project was to identify trends in movie attributes that correlate with a movie's success in respect to the genre that it belongs in. Movie attributes analyzed include the duration of a movie, its allocated budget, average actor similarity, title length, director hits, and plot keywords. For the purposes of this project, we defined a film's success by how much gross profit it made in the box office, independent of critic reviews and ratings. We define success in this way because audience response tends to be biased and unpredictable whereas gross profit is a more definitive factor. In other words, whether or not a movie is perceived well, it can still be profitable and hence successful for the filmmakers. In addition, we completed separate regression analyses for each genre based upon the assumption that different movie genres do well under different circumstances. By analyzing this data, we expected to be able determine which movie attributes will result in a highly profitable movie launch for each genre and how strongly they contribute to its success.

Introduction and Hypothesis

The primary application for this project sought to benefit and increase competition in two major industries — the entertainment and advertising industries. In the conclusion of this analysis, this project identifies which movie attributes negatively or positively affect a movie release the most and any notable trends we discovered for each genre. Filmmakers, directors, and actors can also determine whether or not they would like to pursue a movie idea or make any adjustments to their planned release with this type of information. This can also be applied to and impact how a film is marketed to their target audience.

The dataset to support this project was initially retrieved from 'kaggle.com'. It was a CSV file of 5,000 movie attributes scraped from IMDB and 'the-numbers.com'. We cleaned this dataset of entries that were missing relevant attributes. Because this step eliminated a significant amount of data, we manually scraped and cleaned more data from IMDB's top box office lists for the past 12 years. We then merged this data back with the original dataset, totaling roughly 6,000 that had all of the attributes we were looking to analyze. We removed movies that were missing column entries because we wanted to ensure that every movie could contribute to every analysis equally.

In regards to our hypothesis, we expected that the attributes that strongly affect a movie release will differ based upon the genre(s) it belongs to. In this study, we analyzed both popular genres and unpopular genres. The popular ones included action, drama, comedy, thriller, romance and the unpopular ones were documentary, musical, and history. Generally speaking, we believed that all movies with a bigger budget would tend to make a greater profit because they can afford more popular actors, a better set, and a quality film crew. In particular, we predicted for example, that romances and comedies short in runtime would do better than longer ones because they do not typically have a complex plot. On the other hand, action, drama, and thriller movies may thrive with a longer duration because their plots tend to have bigger backstories and take more time to unravel. We also anticipated that documentaries, musicals, and history films would be strongly impacted by the number of hits its director has made because although its other attributes such as featured actors

and budget may not be in its favor, a creative director can still positively affect the final product and quality of such productions. Finally, we believed that these types of movies tended to be unpopular because they are usually too long in duration. Therefore, we predicted that films in this category with a shorter duration, would generally be more profitable.

Techniques Applied

For this project, we planned to use two different techniques to help identify movie attributes related to its success. These methods include multiple linear regression and k-means clustering. Multiple linear regression was a natural choice because it gives us an idea of whether or not certain independent variables are significant to any given genre and how strongly they correlate with its gross profit. In particular, we looked at the duration, title length, budget, number of hits the director has produced, and average actor similarity in relation to the gross profit it made. Applying an OLS linear regression function to this data provided us with coefficients and confidence intervals to identify correlations.

Furthermore, we used k-means clustering to analyze a different kind of attribute, which were plot keywords. We chose to focus on evaluating a movie's plot keywords for the top 1,000 profitable movies and least 1,000 profitable movies of our overall dataset. In other words, we clustered these movies by the most similar plot keywords that are associated with them, provided from IMDB's website. Clusters formed based upon words of similar themes and meaning. Individuals points stayed closer to the center of the cluster or strayed away from it depending upon how similar it was to the other plot keywords in that cluster. At a high level, this gave us an idea of what plot keywords formed the largest clusters, which also indicated what storylines movie critics and audiences found most interesting.

Datasets and Experiments

The datasets used in our experiments were manipulated subsets of cleaned data from the master set of movies. Because each analysis technique focused on a different set of movie attributes, we created new, more relevant data frames to help perform the respective experiment. For example, in the multiple linear regression model, we extracted a dataset that only consisted of the movie title, gross profit, genres, duration, calculated title length, budget, calculated director hits, and calculated average actor similarity. In a similar fashion, for the k-means clustering technique, we used a subset that consisted of only movie titles and plot keywords. Using smaller subsets makes it easy to iterate through the data and focus on only what each experiment needs.

As stated in the previous section, our goal with the linear regression model was to identify whether or not there was a strong correlation between the gross profit a movie made and its duration, title length, budget, director hits, and average actor similarity. Before determining this model, we first identified which movies belonged to which genres (i.e. action, drama, comedy, thriller, romance, documentary, musical, and history). We then calculated the number of words in each movie's title to get its title length. We also calculated how many hits a director had by counting how many movies they produced of the top 500 grossing movies overall. An additional step we took was to calculate average actor Minkowski distance similarity based upon how many movies of each genre three featured actors appeared in and took the average of the similarity measures per movie.

Finally, we passed all this data to the statsmodels api OLS function and plotted the partial regression of the results. We completed these steps for each the eight genres, gradually repeated them by removing attributes that did not result as significant, and recorded our observations along the way.

For k-means clustering, we extracted a list of movie titles and their corresponding plot keywords. We used the TfidfVectorizer to obtain a tf-idf matrix. One of the parameters for this vectorizer was the tokenize function, which returned a filtered list of tokens. Then, we evaluated the number of clusters for our k-means plot, passing the vectorized matrix of the keywords and the max number of clusters as 6. After that we ran the k-means algorithm on the tf-idf matrix. Since we did not have numeric values to visualize the results of the k-means, we used Multidimensional Scaling (MDS). We obtained a distance matrix for the tf-idf matrix using cosine similarity and using this we obtained coordinates for visualization from MDS.

Results and Discussions

Multiple Linear Regression

Our results from the multiple linear regression experiment provided some useful insight for each genre. In general, we also noticed that our R-squared values all fell below 50%. Adding more predictors (i.e. movie attributes) to our analysis could possibly increase these numbers, but we believe that this is due to that fact that viewership and turnout for a movie is highly variable. Individuals can choose to see a movie and in turn contribute to its gross profit independent of the attributes that we have observed. Despite this, all results still had low p-values, which indicate a real relationship between the signifiant predictors and gross profit. The coefficients also estimate how strong the trends observed are while the R-squared value represents the variability of the plots around the regression line.

Now we will discuss our specific findings for each of the eight genres and their linear regression results.

Action

As you can see from the OLS Regression Results on the next page, the most significant attributes for action movies are its duration, budget, and director hits. The coefficients for each of these attributes are also positive. This indicates that the longer the duration, the bigger the budget, and the more hits a director has, the higher gross profit an action film will make. We find these observations to make logical sense because popular action movies, especially ones pertaining to war or superheroes, are generally longer and have a hyped reputation with the public from pre-cursor films, novels, and other stories.

Results for Action Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.092
Model:	OLS	Adj. R-squared:	0.090
Method:	Least Squares	F-statistic:	48.73
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	5.29e-30
Time:	13:40:27	Log-Likelihood:	-48239.
No. Observations:	1454	AIC:	9.649e+04
Df Residuals:	1450	BIC:	9.651e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-1.847e+13	9.41e+12	-1.963	0.050	-3.69e+13	-1.08e+10
duration	2.137e+11	8.64e+10	2.475	0.013	4.44e+10	3.83e+11
budget	4.66e+04	1.35e+04	3.464	0.001	2.02e+04	7.3e+04
director_hits	1.214e+13	1.27e+12	9.566	0.000	9.65e+12	1.46e+13

Omnibus:	1711.021	Durbin-Watson:	1.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	203930.002
Skew:	5.956	Prob(JB):	0.00
Kurtosis:	59.782	Cond. No.	8.53e+08

Figure 1a. OLS Regression Results for action movies.

Drama and Thriller

Dramas and thrillers interestingly had the identical results. Their most significant predictors for gross profit success were duration, title length, director hits, and average actor similarity. Again, all coefficients were positive, implying that the longer the movie and the title, the more hits a director has, and the more similar the movie's featured actors are, the higher gross profit a drama or thriller film will make. Of these findings, we cannot exactly pinpoint why title length contributes to these types of films' gross profit, but perhaps this means that there has simply been a trend in successful dramas and thrillers where their titles have had more words in the past few years.

Results for Drama Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.084
Model:	OLS	Adj. R-squared:	0.083
Method:	Least Squares	F-statistic:	77.51
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	5.60e-63
Time:	13:41:15	Log-Likelihood:	-1.0914e+05
No. Observations:	3373	AIC:	2.183e+05
Df Residuals:	3368	BIC:	2.183e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-8.231e+12	2.69e+12	-3.064	0.002	-1.35e+13 -2.96e+12
duration	8.328e+10	2.33e+10	3.579	0.000	3.77e+10 1.29e+11
title_len	8.292e+11	3.28e+11	2.529	0.011	1.86e+11 1.47e+12
director_hits	6.631e+12	4.54e+11	14.615	0.000	5.74e+12 7.52e+12
actors_similarity	8.026e+10	3.65e+10	2.201	0.028	8.75e+09 1.52e+11

Omnibus:	4416.474	Durbin-Watson:	1.042
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1029590.225
Skew:	7.201	Prob(JB):	0.00
Kurtosis:	87.371	Cond. No.	656.

Figure 1b. OLS Regression Results for drama movies.

Results for Thriller Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.104
Model:	OLS	Adj. R-squared:	0.102
Method:	Least Squares	F-statistic:	51.94
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	1.91e-41
Time:	13:41:18	Log-Likelihood:	-58969.
No. Observations:	1804	AIC:	1.179e+05
Df Residuals:	1799	BIC:	1.180e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-2.41e+12	2.01e+12	-1.201	0.230	-6.35e+12 1.53e+12
title_len	1.678e+12	6.33e+11	2.651	0.008	4.37e+11 2.92e+12
budget	4.383e+04	1.05e+04	4.188	0.000	2.33e+04 6.44e+04
director_hits	8.604e+12	7.66e+11	11.231	0.000	7.1e+12 1.01e+13
actors_similarity	1.718e+11	5.92e+10	2.902	0.004	5.57e+10 2.88e+11

Omnibus:	2217.125	Durbin-Watson:	1.349
Prob(Omnibus):	0.000	Jarque-Bera (JB):	362090.393
Skew:	6.382	Prob(JB):	0.00
Kurtosis:	71.222	Cond. No.	2.24e+08

Figure 1c. OLS Regression Results for thriller movies.

Comedy

The most significant predictors for comedies were budget and director hits. In theory and as we hypothesized, a bigger budget should do more good than harm for any movie as long as it is managed well, so this is not a surprising factor. We also believe that the number of director hits had a significant impact on the response variable because good comedy heavily depends on the right timing and execution, which are often determined by the director. Many comedies of the same type of jokes and vibe are also often directed by the same person as a sequel or follow-up to a previous release.

Results for Comedy Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.108
Model:	OLS	Adj. R-squared:	0.108
Method:	Least Squares	F-statistic:	145.4
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	2.63e-60
Time:	13:41:45	Log-Likelihood:	-78033.
No. Observations:	2392	AIC:	1.561e+05
Df Residuals:	2389	BIC:	1.561e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	1.983e+12	9.59e+11	2.067	0.039	1.02e+11 3.86e+12
budget	1.372e+05	1.53e+04	8.975	0.000	1.07e+05 1.67e+05
director_hits	7.985e+12	6.62e+11	12.059	0.000	6.69e+12 9.28e+12

Omnibus:	2295.188	Durbin-Watson:	0.792
Prob(Omnibus):	0.000	Jarque-Bera (JB):	115280.918
Skew:	4.563	Prob(JB):	0.00
Kurtosis:	35.763	Cond. No.	8.32e+07

Figure 1d. OLS Regression Results for comedy movies.

Romance, Musical, and History

We observed that the number of director hits had the most significance to a romance's and musical's gross profit. Similar to comedies, we believe that this result makes sense because directors can play a large part in the casting of movies. Casting is arguably important for any genre, but particularly important for romances because the main characters need to have a chemistry that the audience believes. In regards to musicals and history movies, some of the most popular films of these types have also been based upon true stories, novels, and playwrights, all of which take a good director to convey the original story.

Results for Romance Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.099
Model:	OLS	Adj. R-squared:	0.098
Method:	Least Squares	F-statistic:	160.5
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	5.63e-35
Time:	13:42:39	Log-Likelihood:	-47607.
No. Observations:	1468	AIC:	9.522e+04
Df Residuals:	1466	BIC:	9.523e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	3.993e+12	8.75e+11	4.565	0.000	2.28e+12 5.71e+12
director_hits	9.525e+12	7.52e+11	12.668	0.000	8.05e+12 1.1e+13

Omnibus:	1705.434	Durbin-Watson:	0.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	182883.159
Skew:	5.868	Prob(JB):	0.00
Kurtosis:	56.406	Cond. No.	1.73

Figure 1e. OLS Regression Results for romance movies.

Results for Musical Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.105
Model:	OLS	Adj. R-squared:	0.098
Method:	Least Squares	F-statistic:	16.37
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	8.57e-05
Time:	13:43:01	Log-Likelihood:	-4665.3
No. Observations:	142	AIC:	9335.
Df Residuals:	140	BIC:	9341.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	4.811e+12	4.56e+12	1.056	0.293	-4.2e+12 1.38e+13
director_hits	1.726e+13	4.27e+12	4.046	0.000	8.83e+12 2.57e+13

Omnibus:	188.067	Durbin-Watson:	1.602
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8376.829
Skew:	5.155	Prob(JB):	0.00
Kurtosis:	39.187	Cond. No.	1.87

Figure 1f. OLS Regression Results for musical movies.

Results for History Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.054
Model:	OLS	Adj. R-squared:	0.050
Method:	Least Squares	F-statistic:	14.96
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	0.000138
Time:	13:43:13	Log-Likelihood:	-8681.9
No. Observations:	266	AIC:	1.737e+04
Df Residuals:	264	BIC:	1.738e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	6.911e+12	2.61e+12	2.653	0.008	1.78e+12 1.2e+13
director_hits	5.84e+12	1.51e+12	3.868	0.000	2.87e+12 8.81e+12

Omnibus:	307.268	Durbin-Watson:	1.454
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13789.914
Skew:	5.006	Prob(JB):	0.00
Kurtosis:	36.822	Cond. No.	2.25

Figure 1g. OLS Regression Results for history movies.

Documentary

The OLS Regression Results proved that none of the predictors were significant for documentary films because all of the confidence intervals contained 0. We believe that this occurred because documentaries fell under the unpopular movie genres and have a high variability in terms of viewership. In other words, most individuals only watch documentaries of topics that they are interested in, independent of who directed it or what actors are featured in it. This high variability contributes to insignificant predictors.

Results for Documentary Movies

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.176
Model:	OLS	Adj. R-squared:	-0.118
Method:	Least Squares	F-statistic:	0.5982
Date:	Mon, 12 Dec 2016	Prob (F-statistic):	0.702
Time:	13:42:47	Log-Likelihood:	-373.64
No. Observations:	20	AIC:	759.3
Df Residuals:	14	BIC:	765.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-4.769e+06	6.54e+07	-0.073	0.943	-1.45e+08 1.35e+08
duration	2.175e+05	4.64e+05	0.468	0.647	-7.78e+05 1.21e+06
title_len	-1.388e+05	6.91e+06	-0.020	0.984	-1.5e+07 1.47e+07
budget	1.1618	1.078	1.077	0.300	-1.151 3.475
director_hits	1.005e+07	1.25e+07	0.806	0.434	-1.67e+07 3.68e+07
actors_similarity	-3.449e+05	1.21e+06	-0.285	0.780	-2.94e+06 2.25e+06

Omnibus:	12.801	Durbin-Watson:	1.883
Prob(Omnibus):	0.002	Jarque-Bera (JB):	10.592
Skew:	1.495	Prob(JB):	0.00501
Kurtosis:	4.940	Cond. No.	8.23e+07

Figure 1h. OLS Regression Results for documentary movies.

Results and Discussions — *K-means Clustering*

Our first k-means clustering analysis on the top 1000 grossing movies led to the following cluster visualization on the next page. To reiterate, the clusters were formed based upon plot keyword cosine similarity. In other words, the closer a point is to another point, the more similar their respective movies are based upon their plot keywords. In particular, we found that for the top 1000 grossing movies the overarching themes of the six clusters generated were police, love and relationships, sex, death and violence, military, and teenage related.

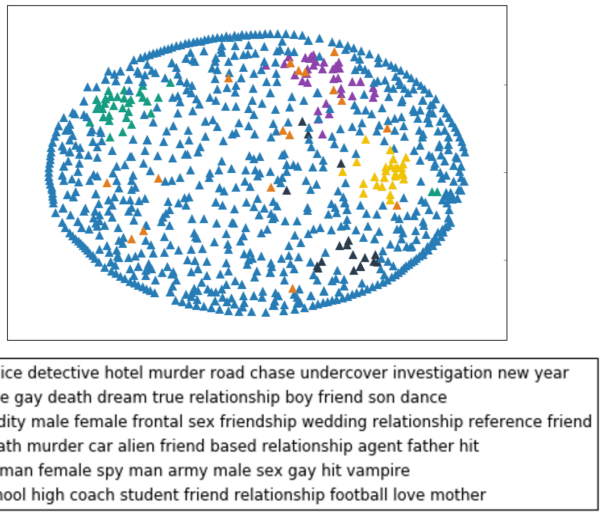

Figure 2a. K-means clustering visualization for the top 1000 grossing movies.

Note that these themes are not the most accurate representations of each cluster, but rather the most important observation to take back from this is that movies of the blue plot keywords, according to the legend, appear to be the most reoccurring of the top 1000 grossing movies. They are also quite similar to movies of the other colored plot keywords because of how interspersed they are. In addition, we observed that the orange plot keywords seem to be the most varied and perhaps least reoccurring. Based upon this result, we can confidently say that the top 1000 grossing movies had plots related to death and murder.

The second k-means clustering analysis of the least 1000 grossing movies led to the following cluster visualization. We found that it was rather difficult to identify the overarching themes for each cluster because the plot keywords inside of them appeared only somewhat related in terms of meaning. The most reoccurring and common plot keywords, however, were of the green color according to the legend. Similar to the blue plots of the top grossing movies, the green ones here were the most interspersed, indicating that all movies of other clusters are fairly similar to it. We can also

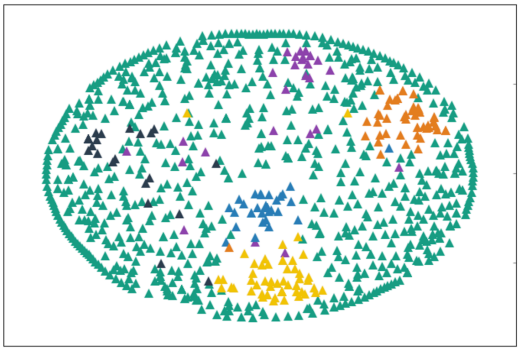

Figure 2b. K-means clustering visualization for the least 1000 grossing movies.

conclude that because the words in each cluster are very varied in meaning, the plots of the least grossing movies were not very focused. This could arguably be the very reason that they did not make as much money.

Conclusion and Future Work

As we initially hypothesized, our results have proven that each film genre has different significant predictors with most significant predictors being within what is generally expected of the filmmaking process. All of the coefficients from our linear regression analyses turned out to be positive and because the majority of the movie attributes we were observing had positive connotations, an increase in these attributes exhibited a general trend of also promoting high gross profit. The k-means clustering analysis provided some additional interesting insight in regards to the plot themes of top grossing movies and least grossing movies. They cannot guarantee that a movie of the clustered plot themes will succeed, but they give an idea of what separates a high grossing movie from a low grossing one. Based on the trends observed, high grossing movies tend to have plots related to death and murder and are more focused than low grossing movies.

Nevertheless, the analysis provided from this project is by no means perfect given the data and time constraints provided. As previously mentioned, movie viewership is inherently variable. Therefore, some future improvements that could be made to our linear regression model would be to analyze a larger set of movie attributes, gather even more data for the overall dataset, and perhaps look into different models for correlations. With improvements such as these and hopefully more trends discovered, the project can also be expanded to make confident predictions for how much a movie will gross within its genre or what attributes a movie release may be lacking or investing too much in.