BLL514E What Makes a Movie Successful

Ahmet Drobi
Informatics Institute
Istanbul Technical Univeristy
Istanbul, Turkey
nawar.droubi@gmail.com

***Abstract*—In this study, we present a statistical analysis about a movies data set composed of around 8,000 movies meta-data, which includes: the title of the movie; the budget, the revenue, run-time, the Motion Picture Association of America rating; genre of the movie; number of top 100 actors according to IMDb that were present in the Movie; languages that the movie was released in; countries in which the movie was released, Meta-score of the movie; IMDb rating, number of IMDb votes on the movie; production company and year of release. The data-set contains some missing values, which we had to clean, then we did some feature engineering to prepare the data-set for analysis and for creating the machine learning model. The statistical analysis of the data-set was for the aim of testing 9 hypothesis with the main aim of the project being finding what effects the revenue of a movie.**

I. Introduction

The project aims at providing an explanation to the "success" of a movie. The term "successful movie" in this project refers to a movie that produced revenue equal to double its budget. Furthermore, a machine learning model was created to try to predict if a movie with a given set of features would be successful or not under the same criteria. Various works have been done to try to answer this question. One important research paper done in this subject, and from which the idea of this project spawned is "How to Make a Successful Movie: Factor Analysis from both Financial and Critical Perspectives" [1]. However, this research paper had two shortcomings: First, it puts too much emphasis on the rating on the movie citing it as a part of the criteria that decides if a movie is successful or not; secondly: it only considers movies that were produced in Hollywood. Our work will rectify those shortcomings by first shifting ratings to be a factor in the success of the movie, not the definition of said success. and then by also including movies produced outside of Hollywood (Like Bollywood for example). We believe that not just the rating of the movie that define its success, but also in what language it was released, what budget did it has, and how many "famous" actors were in the movie. To achieve our goal, first, we did basic feature engineering techniques on some of the features of the data-set that we collected, then use some well known statistical methods to test 9 hypothesises, each of which tries to explore one aspect of the "success of a movie" formula. Finally based on our findings in the statistical analysis we will create a machine learning model that tries to predict if a movie is successful based on a set of features.

II. Methodology

The methodology used in creating/collecting the data-set was random sampling from a collection of top 10,00 movies by revenue as of the 5^{th} of May, 2019. To test our main research hypothesis we created and tested 9 hypothesises in the aim of finding out which features of the data-set effected the movie revenue in order to include them in the ML model.

A. Movie Dataset

In order to create the movie data-set we used The Movie Database API [2] to get a random list of 10,000 movies from the top 100,000 by revenue, then we used OMDb API [3] to collect the meta-data for the movies. The meta-data we used in this study are: 1) Title 2) Budget 3) Revenue 4) Run-time 5) Motion Picture Association of America rating 6) Genre 7) Top 100 Actors in Movie 8) Language 9) Country 10) Meta-score 11) IMDb Rating 12) IMDb Votes 13) Production Company 14) Year of Release

B. Data Processing

The original data-set contained total of 10,000 movies, after filtering the data-set it reduced to 8,000 movie. We found a number of problems in the following features:
1) The Year column : was missing a lot of values from the OMDB API, where the one got from The Movie Database was complete, so we dropped the original column and extracted only the year from the date column which we got from TMD, then renamed it into "Year of Release" to better reflect the actual content of the column.
2) The Country column: was treated in a similar way to Language column. So, for movies with this value missing we assigned the value USA
3)The IMDB Rating and Votes columns: Zero has been assigned to the missing values, in purpose of protect the data-set from the missing value in this column.
4) The Actors column: This column had the name of the actors in that film. The main goal of this column is to see how many famous actor a movie include. To achieve this goal we gathered the top 100 famous actor and we checked each movie how many famous does it have.
5) The Production column: The missing values in this columns has been changed from null to others.
6) The Run-time column: The mean of this column has been assigned to the missing value.
7) The Genre column: To be able to use label encoding on

this column we choose the first value in this column which originally had multiple values.

III. Experiment Results

A. Statistical Analysis

First we found the statistical data for the numerical features as follows:
1) For IMDb Votes the mean is: 79283.1397 vote
2) For Production Company the mode is ”Others” which indicates that no one production company had their movie be in top 100,000 movie by revenue.
3) For Revenue the mean is: 7.099426e+07.

B. Hypothesis Testing

Using the correlation heat diagram we noticed the features that had a likely significant relation to revenue. Those features were: the budget and IMDb votes. Furthermore, we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result):

1) H0: The lower the number of famous actors in a movie, the lower the revenue is.
H1: The higher the number of famous actors in a movie, the higher the revenue is.
Method of testing: using bar plot 3 and checking the Peterson’s Correlation Coefficient which was 0.187.
Result: We rejected the alternative hypothesis.

2) H0: The budget of a movie doesn’t effect the number of famous actors in it.
H1: The budget of a movie does effect the number of actors in it.
Method of testing: the Peterson’s Correlation Coefficient which was 0.141.
Result: We rejected the alternative hypothesis.

3) H0: If a movie’s genre is Action, it has a lower average revenue, than if it was of another genre.
H1: If a movie’s genre is Action, it has a higher average revenue, than if it was of another genre.
Method of testing: calculating the overall mean, and comparing it with the mean of action genre movies.
Result: The alternative hypothesis passed and we reject the null hypothesis

4) H0: If a movie is released outside of USA, it’s revenue will be lower than revenue if released in USA only.
H1: If a movie is released outside USA, it’s revenue will be higher than revenue if released in USA only.
Method of testing: the Mann-Whitney U test with p-value = 2.44e-10
Result: We reject the alternative hypothesis

5) H0: The Higher the budget of a movie the Lower the revenue is.
H1: The Higher the budget of a movie the Higher the revenue is.
Method of testing: the Peterson’s Correlation Coefficient which was 0.51.
Result: The alternative hypothesis passed and we reject the null hypothesis

6) H0: The Higher number of famous actors in a movie, the Lower its rating is.
H1: The Higher the number of famous actors in a movie, the Higher its rating is.
Method of testing: bar plot2
Result: The alternative hypothesis passed and the null hypothesis failed.
7) H0: The Higher the number of people who voted for a movie, the higher its rating is.
H1: The Higher the number of people who voted for a movie, the lower its rating is.
Method of testing: the Peterson’s Correlation Coefficient which was 0.34.
Result: We reject the alternative hypothesis.

8) H0: If a movie is produced by others, then it will have higher than average revenue.
H1: If a movie is produced by others, then it will have lower than average revenue.
Method of testing: the Mann-Whitney U test with p-value = 2.44e-10
Result: we reject the alternative hypothesis

9) H0: If a movie has more than 1 award, it has a lower revenue.
H1: If a movie has more than 1 award, it will have a higher revenue.
Method of testing: using bar plot5 and checking the Peterson’s Correlation Coefficient which was 0.04.
Result: We reject the alternative hypothesis

C. Machine Learning Model

We tried in this part to create a ML model that could classify - with acceptable accuracy - movies into successful and successful movies, and for it to predict if the movie will be so. The criterion chosen to whether a movie is successful or not if it’s value is revenue is double it’s budget. We first split the data-set into training and testing data-sets (.7, .3 respectively). We used Logistic Regression to achieve our goal. The label (y) was the column ”IsSuccessful” which indicate, as the name states, whether a movie is successful or not. The features used for final model (after multiple tries) were budget and imdbVotes. As you can see in the following two tables the model had an accuracy of 0.77 and a weighted average of 0.76 both of which falls within acceptable accuracy rates for this project. To better understand the TPR and FPR we can check the ROC digram 6 which gives a better representation of the accuracy of our model.

Confusion Matrix		Target	
		True	False
Model	True	497	366
	False	173	1318

	Precision	Recall	F1-score	Support
0.0	0.74	0.58	0.65	863
1.0	0.78	0.88	0.83	1491
Accuracy			0.77	2354
Macro avg	0.76	0.73	0.74	2354
Weighted avg	0.77	0.77	0.76	2354

IV. Figures

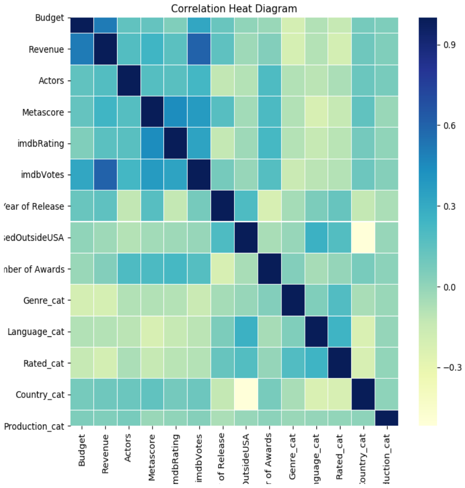

Fig. 1: Data-set Correlation Diagram

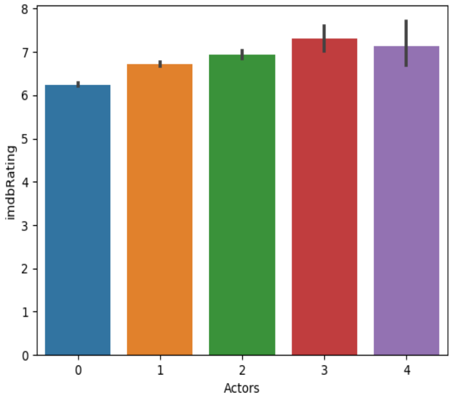

Fig. 2: Bar diagram of IMDB rating vs Actors

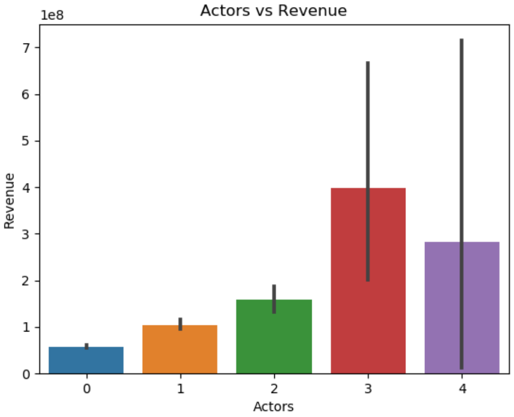

Fig. 3: Bar diagram of Actors vs Revenue

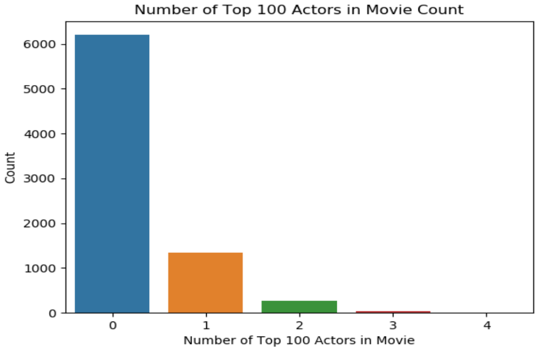

Fig. 4: Movies count for each number of famous actors

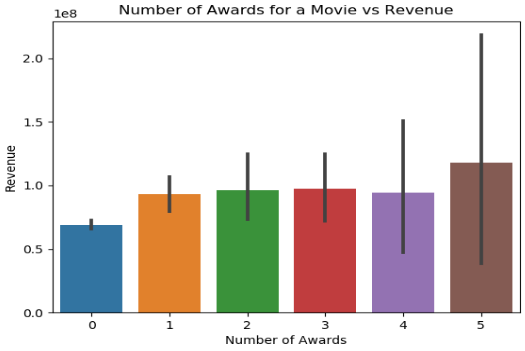

Fig. 5: Bar diagram of the number of awards a movie got vs revenue

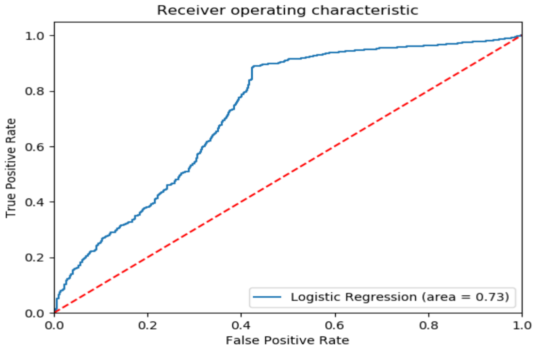

Fig. 6: ROC of the ML model

V. Conclusion

We set out in this study to answer the question of what affects the revenue of a movie and to build a ML model to classify/predict whether a movie is successful or not. We found out that the budget and the number of IMDB-votes of a movie are the main effecting forces on its revenue. The budget in this case is predictable, however, the IMDB-votes could indicate that the more successful the movie is the more the audience are inclined To vote for it on IMDB – needs further studying. For the ML model we used Logistic Regression to create a model that can classify/predict whether a movie is successful with an accuracy of 0.73 using 2 features which are budget and IMDB-votes.

References

[1] "How to Make a Successful Movie: Factor Analysis from both Financial and Critical Perspectives," Information in Contemporary Society, Mar. 2019, doi: 10.1007/978-3-030-15742-563.

[2] T. M. Database, "The movie database api," https://www.themoviedb.org/, May 2019.

[3] B. Fritz, "Omdb api,"https://www.omdbapi.com/, May 2019.