Project 2 Summary

Predicting a Movie Sequel's Box Office Return

Charlie Lew

Abstract

The purpose of this project was to leverage tools such as web scraping in conjunction with regression in python to predict the domestic gross office returns of a movie sequel. For this work, only the direct sequel is considered as opposed to a third or more movies in a particular franchise to keep the objective focused. Based on the preliminary findings, the returns of a sequel follow a near one-to-one correlation with the movie's original provided the original's box office numbers were above USD 50 million. Anything below it is more challenging to predict based on a linear regression. Next steps on how to improve the predictive model will be to include more categorical features as well as expand the current number of movies in the database.

Introduction

Two of my favorite sequels of all time are George Lucas' *Star Wars: The Empire Strikes Back* and James Cameron's *Aliens*. There are there are other great ones but the mentioned two are the ones at least to me that have stood the test of time. Both have great visual effects, action and most important of all, good story telling that transcends across cultures. However, this is what the audience sees and appreciates but to a movie studio that creates and produces sequels, one cannot just rely on a feel-good story alone to green light a project. Investment in a sequel has to be sound and ensure that a studio not only recovers the cost but makes a healthy profit. I have often wondered what decisions do executives make or at the very least can one predict the box office returns of a sequel beforehand which could therefore greenlight a project? Hence, the aim of this project is to leverage data analytics to predict the box office return of a movie sequel.

Methodology, Data & Tools

To keep the objective focused, only the immediate sequel of an original is considered. For example, the sequel to Superman will be Superman II but Superman III and IV are not included. Furthermore, any movie that is direct to DVD or video is excluded. Since the objective is to predict box office returns, only movies that were released to cinemas are considered. Knowing the initial parameters, a combination of Python and Beautiful Soup was used to scrap data exclusively from International Movie Database (IMDB). Another website called Box Office

Mojo was also sourced but was only used as a cross-reference to check and in some cases augment the missing data from IMDB. As a start, features that were used for this initial study are shown in Table 1 below

Table 1: Features used for this study

Features	Dimension	Type
Gross Domestic Box Office	USD	Integer
IMDB Movie Score	None	Float
Runtime	Minutes	Integer
Release Month	Month	Integer
Days between Original and Sequel Release	Days	Integer

With the exception of Release Month, all other features are continuous. After the scraping and cleaning were performed, a total of 270 movies were gathered for the analysis. To gain a potential understanding as to how the box office returns for a sequel compares to its original, Figure 1 shows the comparison between the sequel box office returns versus the original

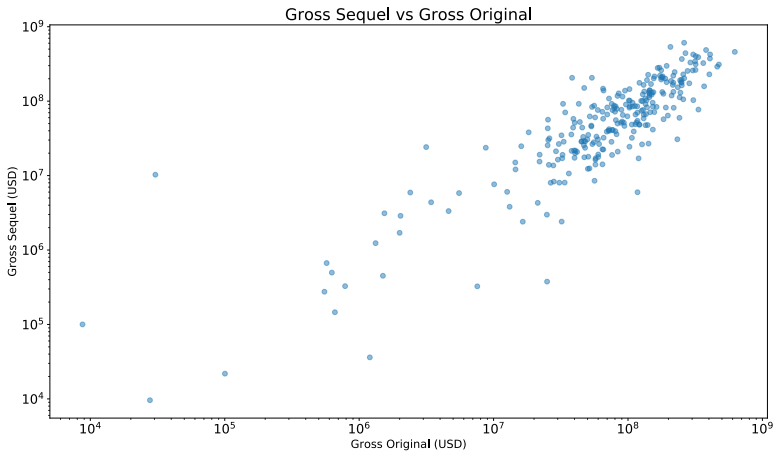
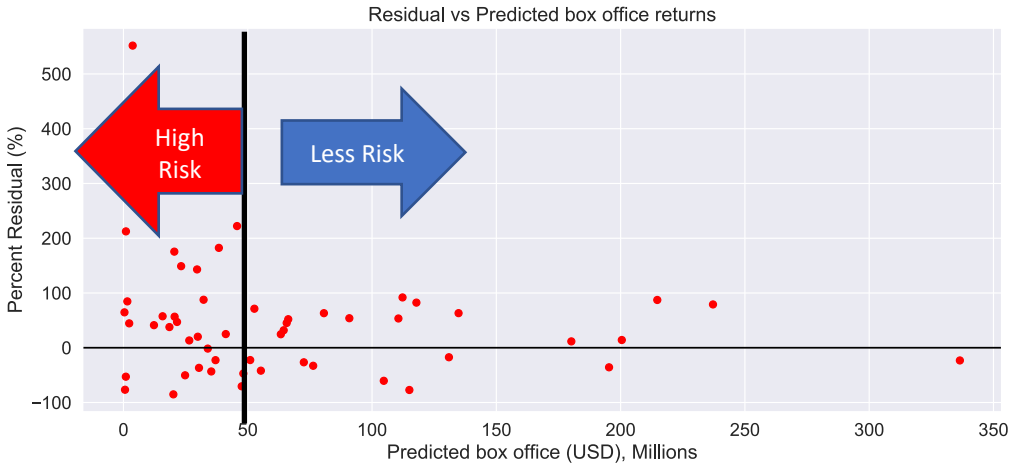


Figure 1: Gross sequel vs original

From Figure 1, datapoints that are clustered over USD 50 million appear to have linear relationship with a near one-to-one ratio. Anything below USD 50 million has an uneven spread. However, this gives us an idea of at least where to start when constructing our model. As a first step in predicting the box office returns for a sequel, an Ordinary Least Squares or OLS method using Linear Regression will be used. The package that utilizes this is called Scikit-Learn which is an external Python library. In order to properly test our linear regression, 80% of the data is used for training whereas 20% is used for the prediction. Hence, an 80-20 split.

Preliminary Results

Figure 2 below shows residual plot of the predicted box office returns for a sequel. The residuals are plotted in terms percentage whereby the actual values are subtracted with the predicted numbers divided by the predicted numbers.



Due to the limited number of observations and features the calculated R-squared of the training and predicted as well as the root mean squared error are

- *R^2 Score Training = 0.74*
- *R^2 Score Predicted = 0.78*
- *RMSE = $54.5 million*

The scores overall need improvement the model can only predict at best for 75% of the data that is collected. Missing features which are categorical such as cast, genre, studios and historical fact such as cult classic for example may improve the model further. Nonetheless, Figure 2 based on the current logistics, show that if the original made approximately 50 million dollars and above, then there is a high probability that it will fall between -100 and 100 percent performance. However, if the original made less than 50 million, then there is a high risk that it can be smash success or total failure. As a test, the movie *Maleficent* made USD 241 Million at the domestic box office. Its upcoming sequel, *Maleficent: The Mistress of Evil*, will be released on October 18th, 2019. Based on this study, the predicted box office return is expected to under-perform slightly making USD 226 Million.

Future Work

Predicting the box return of a sequel is no easy task. Entire departments within a movie studio are dedicated in studying this. In order to improve the current simplistic model, more movies need to be added to the list as well as the additional features. Another step to perform is to normalize the data for better accuracy as well. Once these are included, perhaps alternative regression methods can be used without penalizing accuracy.