Predicting Movies’ Revenue

Milestone Report

Ali Naji

Table of Contents

Project Proposal

Movie industries lose a lot of money every year on movies that do not create profit. According to a podcast, about 80% or more of Hollywood movies fail to turn a profit. One example is 'Gone in 60 Seconds' starring Nicholas Cage and Angelina Jolie where the movie grossed to $240 million but when marketing costs and movie budet were minused, it turned out the production company had lost $212 millions. This creates a problem for movie production companies, and having a sense of profitability is essential in the company's long term survival.

A production company can greatly benefit from their past experiences in movies' successes and failures. It turns out that gauging movie's profitability is too complex for humans to understand as there are many features that a movie can have that play greatly in the prediction of success. Nevertheless, those companies can greatly benefit from modern machine learning models that can handle complex features and give out great prediction results. Therefore, employing historical movies' data and using them in ML algorithms to predict success of the movie could save companies millions of dollars on rather unsuccessful movies.

For this task, we need the maximum amount of information about every movie in the relatively not too distant past. The kind of information include a mixture of movie genre, release year, starring actor, director and other info. But most importantly, we need the revenue and budget of the movie in order to determine the net revenue of the movie (gross - budget) and hence, determine if the movie is profitable at all. Other information like critic reviews, although could be quite useful in determining success of the movie, will not be helpful/applicable for future scenarios where there are not critic reviews yet. For these reasons, the data we're going to use for this task is Daniel Grijalva's Kaggle 'Movie Industry Three decades of movies' dataset. The dataset contains 15 columns (14 features and 1 target) and 6820 rows. It contains many indicators to do the task with numerical, date, and categorical string data types. It is relatively clean and contains relatively low number of 'Nans' (unknown values) which makes it easier to wrangle. The column features are all unique, and from a human perspective, are good indicators of movie's performance. The dimensionality of the dataset makes a great deal of sense as well. Many string columns will need to encoded to make them useful in the regression task. Other datasets that contain other info about each movie, like movie franchise, popularity, and remakes could be joined later to improve predictability as needed.

For this problem, we need a regression model that would estimate revenue of the movie and a control statement that would determine if it is greater than zero hence profitable movie. We are going wrangle and explore the data to get it ready for processing. Then, we're going to split the movies to train and test sets, train ML models on the train set, and test them on the test set. As for the ML models that will be used for this task, a mixture of algorithms will be used and their performances/accuracies will be compared to each other and the best performer will be chosen and fine tuned for optimal performance. We will also manipulate the data as needed to achieve best results.

Data Cleaning

The very first step is to import necessary libraries to wrangle the data. For example, we need to import pandas and numpy with their standard aliases. Others will be imported later as needed We then load the dataset as pandas dataframe to handle it. The next step is to inspect the dataset especially the first few rows, the column labels and their data types. This is important for the upcoming steps to have a good idea about the dataset. It helps identify, and ultimately drop, useless columns such as ID, codes, etc... This can be done using .head(), .describe(), and .info() methods of the dataframe to inspect it, .columns for column labels, and .dtypes for data types of columns. In this step, release date column (Year-Month-Day format) will be split to separate year, month, and day numerical columns so that the date could be inputted inside ML model(s)

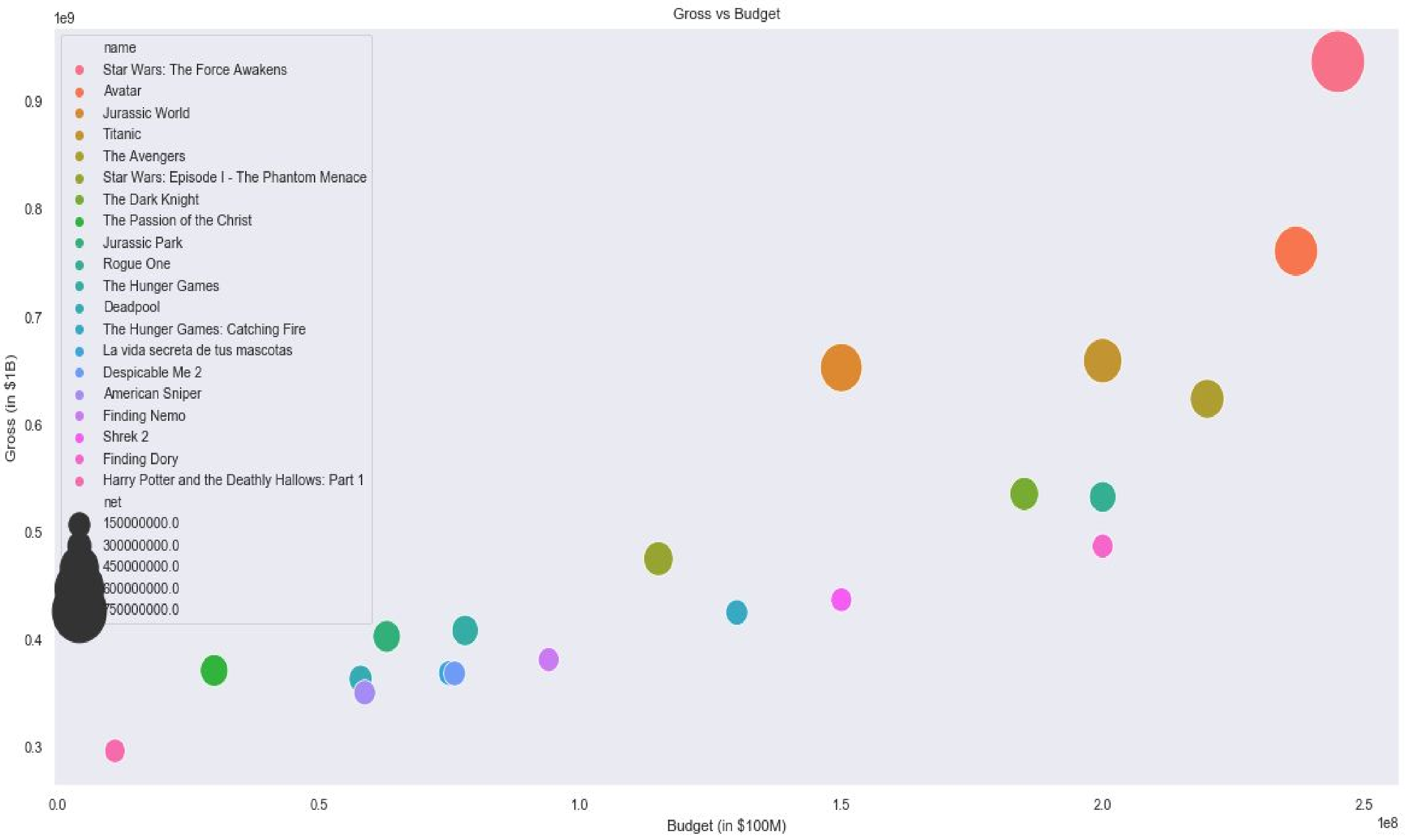
Next, we need to inspect for nulls and missing values in the dataset. In our case, there were no nulls although there were some zeros in the budget column. Going back to the dataset description, it was mentioned that the dataset had 0 budget for films for which the budget was unknown. In that case, we had to replace the zero values with some better representative values. There were about 2200 movies with missing budget. Because budget may be a critical variable in estimating gross, we will filter out those movies instead of substituting their budget with a statistic. Later, we will get back to those movies, estimate their budgets by other means, and append them back to the original dataframe

In many numeric columns, there may be significant number of outliers that could be handled in order not to disrupt plotting and visualization later. Those could be handled early on by creating histograms of numerical data columns and removing outliers below or beyond 95% percentile of their respective columns, if needed.

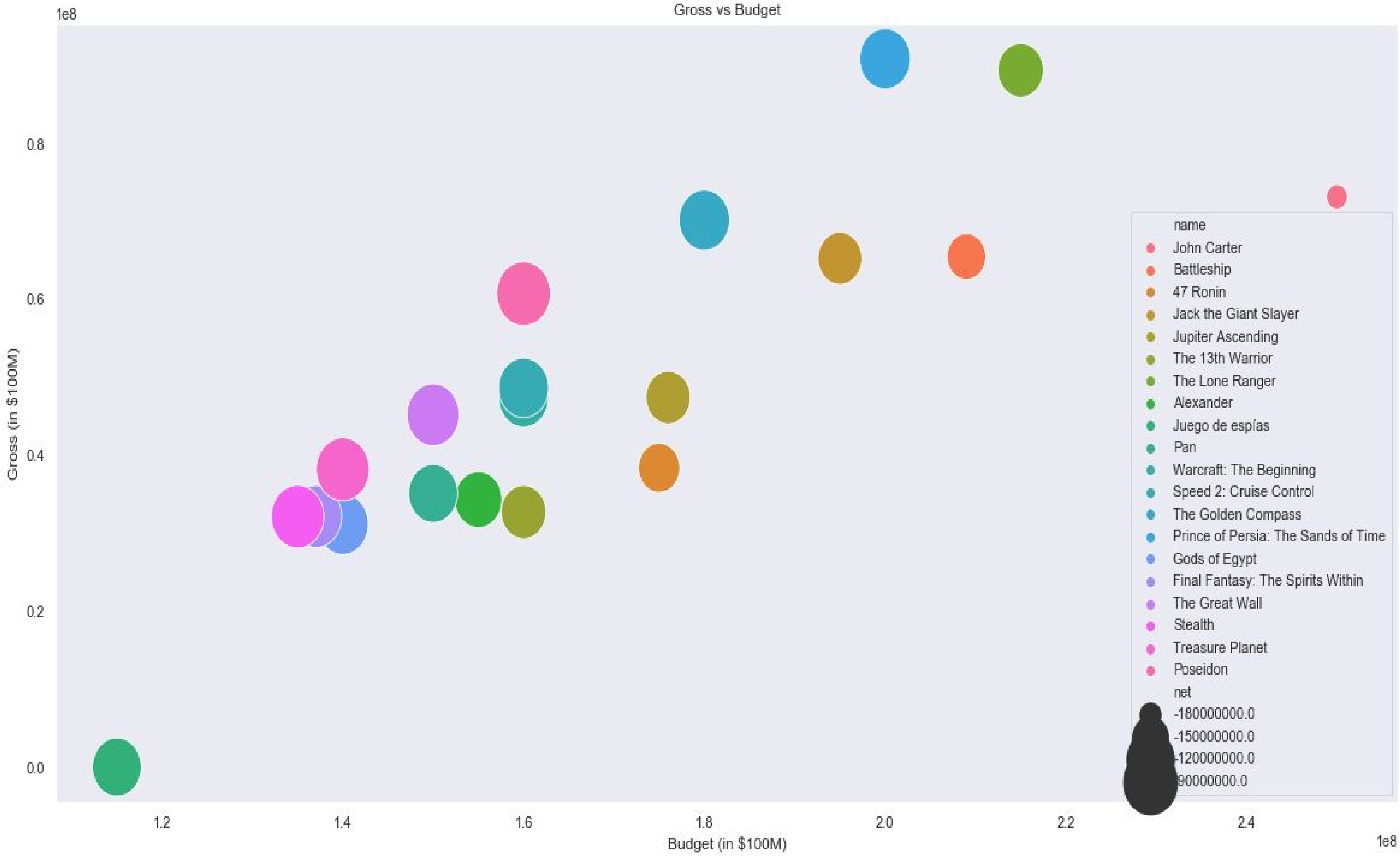
Data Story

What Movies Make Great Profits?

Movie fans have increased dramatically in the past few decades dramatically increasing demand for movies. But what kind of movies do those fans like most? And what kind of movies are most successful? The following graph shows the world’s most profitable movies from 1986 to 2016 in bubbles, the bigger the bubble the more profits the movie made. The x-axis is the budget in $100M while the y-axis represents the gross in $1B. The famous action & sci-fi movie, *Star Wars: The Force Awakens* (2015), topped the list of greatest profits (net revenue, gross - budget) with net revenue of about $691M. *Avatar* had the second greatest success with net revenue of about $523M. Other films had net revenues between $285M to $500M. The common theme between most of these movies, 85% of them, is that they were released in the 21st century and in the US. 55% of them had PG-13 rating and 50% of them had action genre.



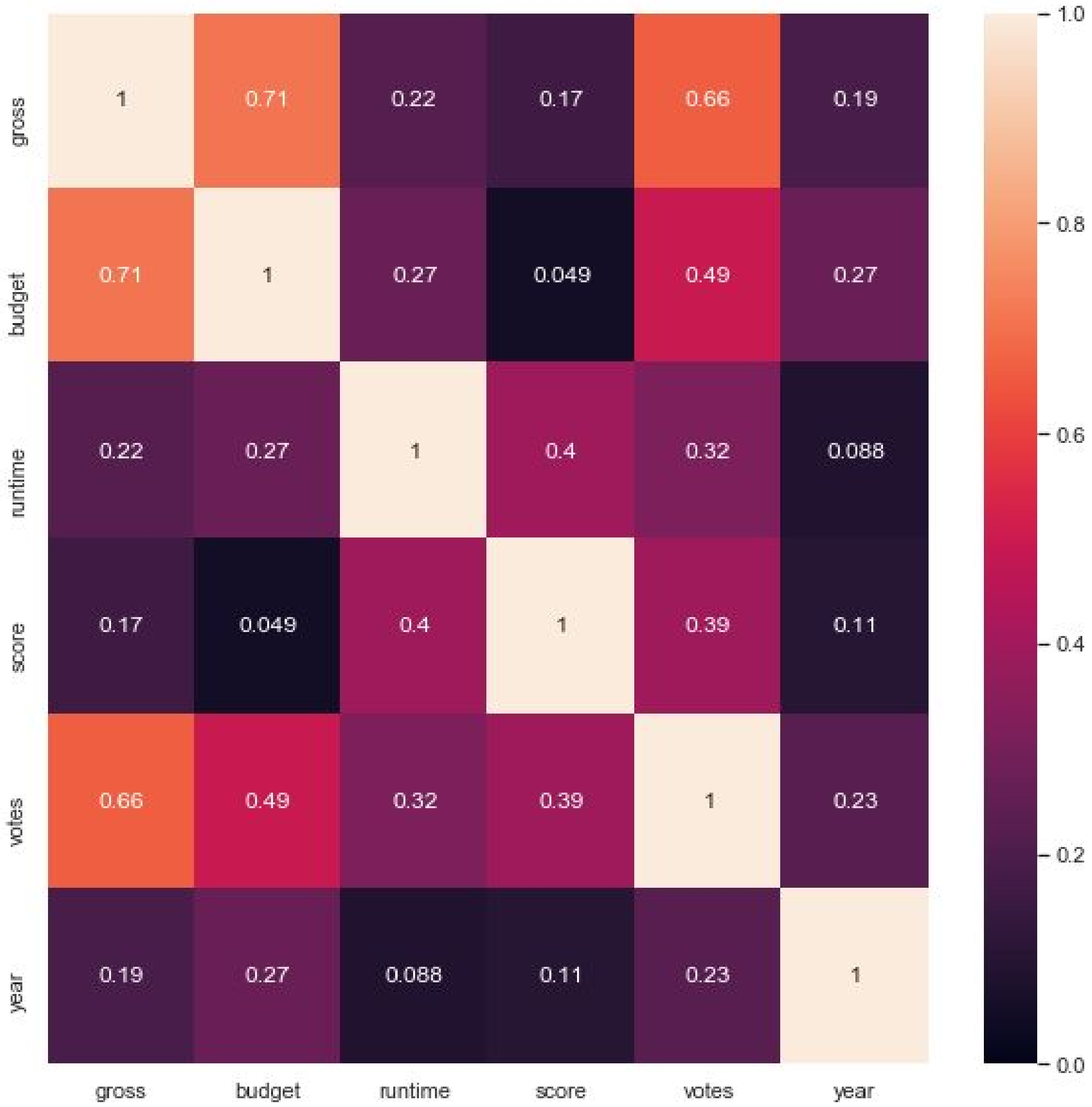
Does that mean that modern movies with PG-13 rating and action genre will most likely be successful? Not really. Looking at the worst performers during that same period in the below similar graph, we can see that action movie *John Carter* (2012) had net revenue of about $-177M (that's $177M lost) which also had PG-13 rating and was released in the US. That applies to many of the other worst performers list.



So is it possible to predict movie success based on its info? It turns out that the answer to this is complex and additional information about the movie context, its starring actor, director, and writer are needed to sufficiently answer this. In the next few sections, we will explore the data and find any correlations between different features. Later, more sophisticated methods will be used to reveal secrets behind successful movies.

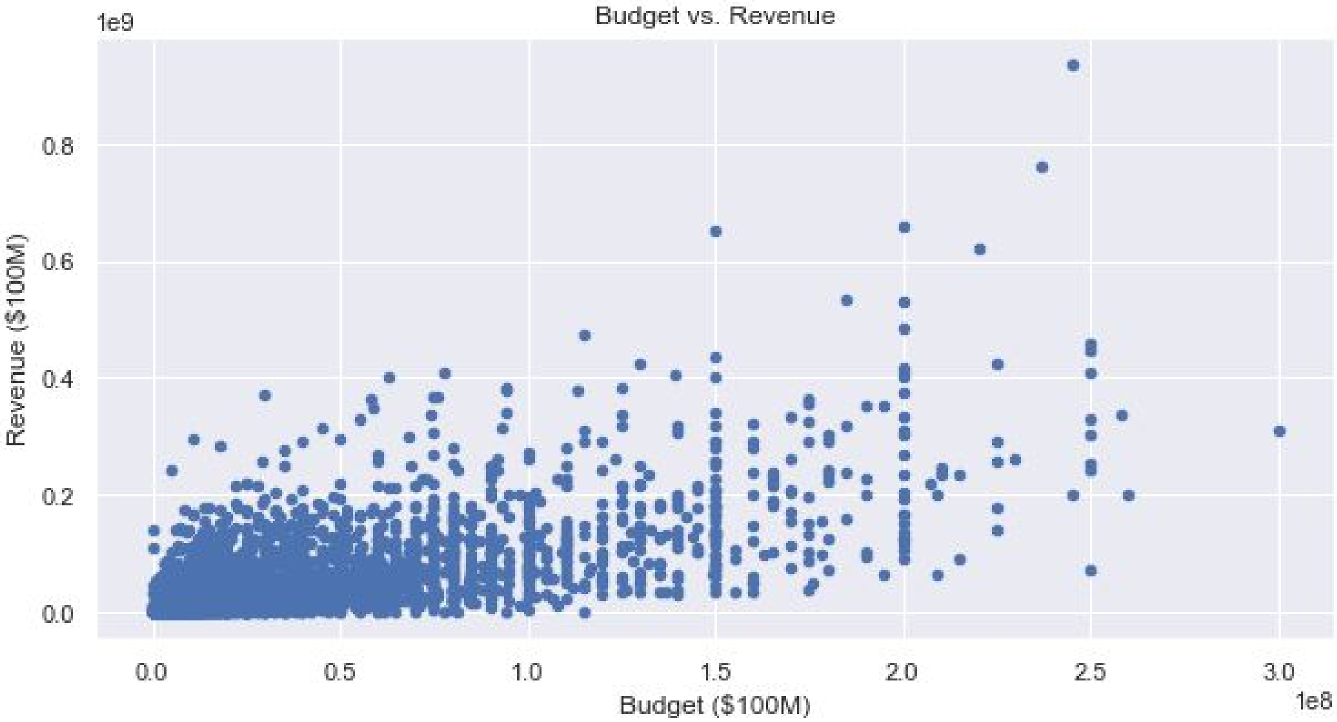
Data Exploration Findings

After visualizing relations between several string columns and target variable, revenue of movie, many of them had weak or mild correlations with the target. However, finding correlations between numerical columns using a correlation matrix have revealed interesting results as shown in the figure below



This shows an interesting result, that there is a strong correlation between budget and revenue (0.71 pearson correlation).

When creating a scatter plot between budget vs. revenue we get the following plot:



This shows a roughly linear trend. In order to verify that this is always the case between the two, we need to do a hypothesis test where fraction of pearson correlations in permuted replicates are at least as extreme as observed one to the total number of replicates, which is p_value. If that p_value < 0.05, we reject the null hypothesis as there is statistically significant correlation

H0: No significant correlation between Budget and Revenue

H1: There is significant correlation between them

After conducting the test, we got a p_value of 0.0 which allows us to reject the null hypothesis and conclude that there is a statistically significant correlation between budget and revenue. This is a primary result that will help us create models to estimate revenue based on budget.

Next Steps

Next steps include encoding string columns such that they will be useful to ML models down the line. After creating ML models, we will test their accuracies and determine if there needs to be additional data to support the regression task. If needed, those will be joined on the movie name and will include information like movie franchise, remakes, and publicity.