Predicting Animated Movie Revenue

GitHub Repo: https://github.com/smyansengupta/ds3000project

Andy Shen

Jeffrey Krapf

Jason Yu

Reagan White

Abstract

This paper aims to analyze the factors influencing the prediction of an animated movie's box office revenue. The research utilizes historical data on animated films from 2000 to the present, sourced from Wikipedia, and focuses on predicting box office revenue based on production budgets. The dataset was meticulously cleaned by removing unnecessary columns and converting relevant information into Python-compatible formats. The findings indicate that while it is feasible to predict and analyze box office revenue based on budget for lower-budget animated films, the accuracy diminishes as budgets increase. For higher-budget productions, additional factors significantly impact a movie's success or failure, suggesting that incorporating these variables would enhance predictive accuracy.

Introduction

For our project we explored the factors that influence an animated movie's box office performance. Specifically, we investigated whether it's possible to predict box office earnings based on the film's budget. Understanding these aspects can offer insights into what makes an animated movie financially successful and could help to forecast its potential performance. An article we came across that ran their own linear regression model mentioned that while a larger budget is strongly linked to a film's success, other factors like distribution and critical reviews also carry significant weight. Interestingly, shorter movie titles often perform better, and sequels tend to dominate at the box office. They focused specifically on the top animated movies of all time; however, we will instead focus on releases from the 2000s to the present to see if our predictions will be accurate for a larger number of movies and variety of different styles. To account for the changing meta in films over the years we also decided to see if the date the movie was released had any impact on the box office revenue, whilst still keeping primary focus on our original factor of the movie's budget.

Method

The first ML method we chose was a simple linear regression to predict the revenue based on the budget of the film and a multiple linear regression model. We decided a linear regression for this would make the most sense because we were predicting numerical values by using another set of numerical values. It further made sense because we were only focusing on two factors, budget and revenue. Our main assumption when using this method was that the budget would influence the expected revenue of the film, and we would get a general idea of how the two are related.

The second method we chose was a multiple linear regression model, to account for the shift in film revenue over time. In this model, we kept our original budget versus box office revenue, however we also had the movie's release date as a second factor. By having these 2 models, we would get a good general idea of the relation between budget and revenue but also have a separate model to account for any potential biases caused by what was happening at the time the movie was released.

However, these models led to some pitfalls, being that there are a lot of differences between animated films that were not accounted for in our models. Such factors include genre, medium, as well as all the different animated studios that are operated across the world. These methods did not consider these variables which play into the success of a film and therefore impacted the correctness of this model.

Results

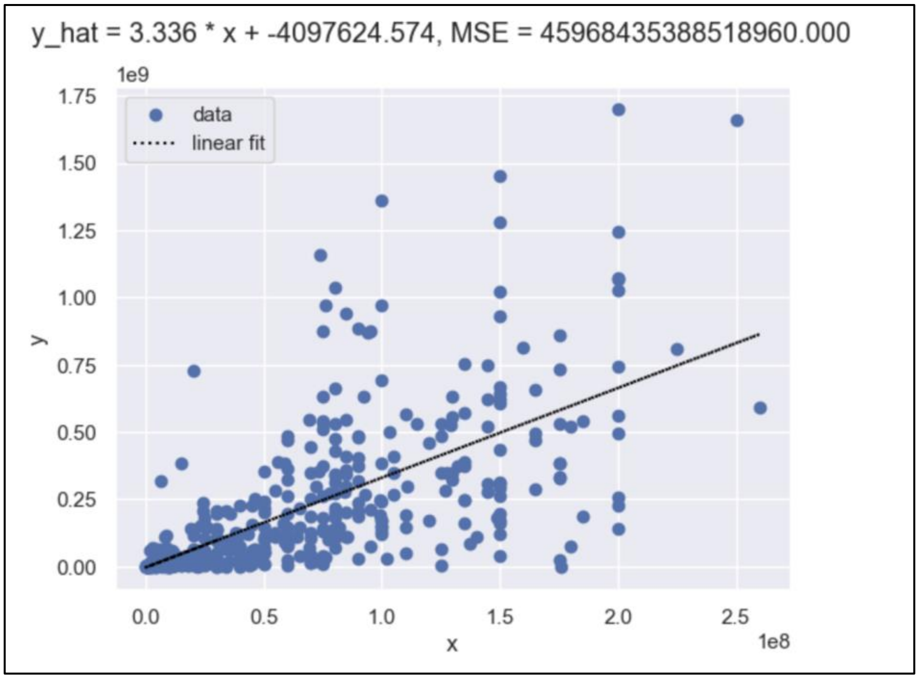

Figure 1. Animated movie budget vs. revenue with line of best fit

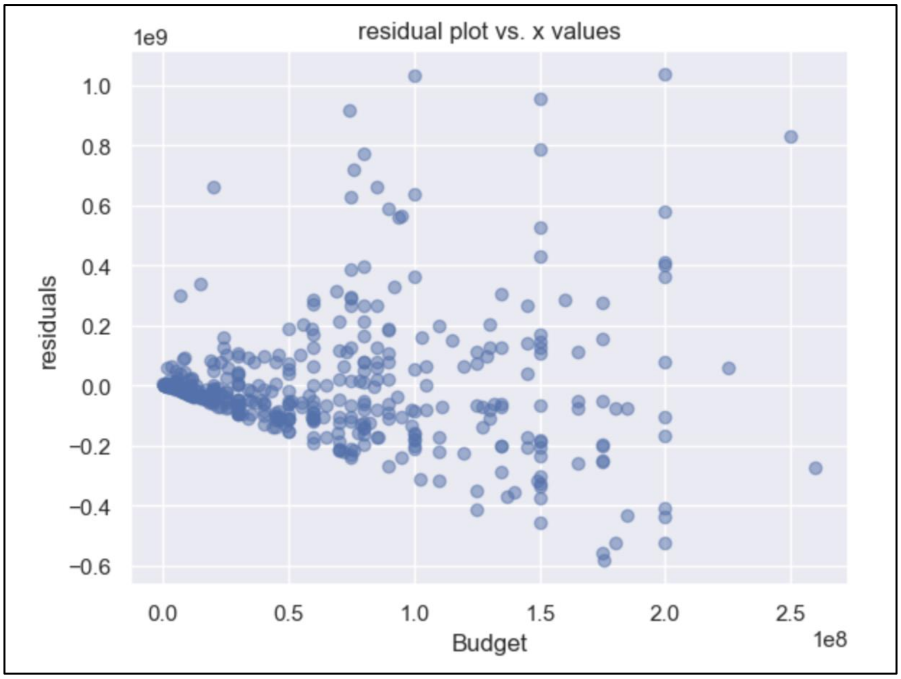

Figure 2. Animated movies budget vs. residuals

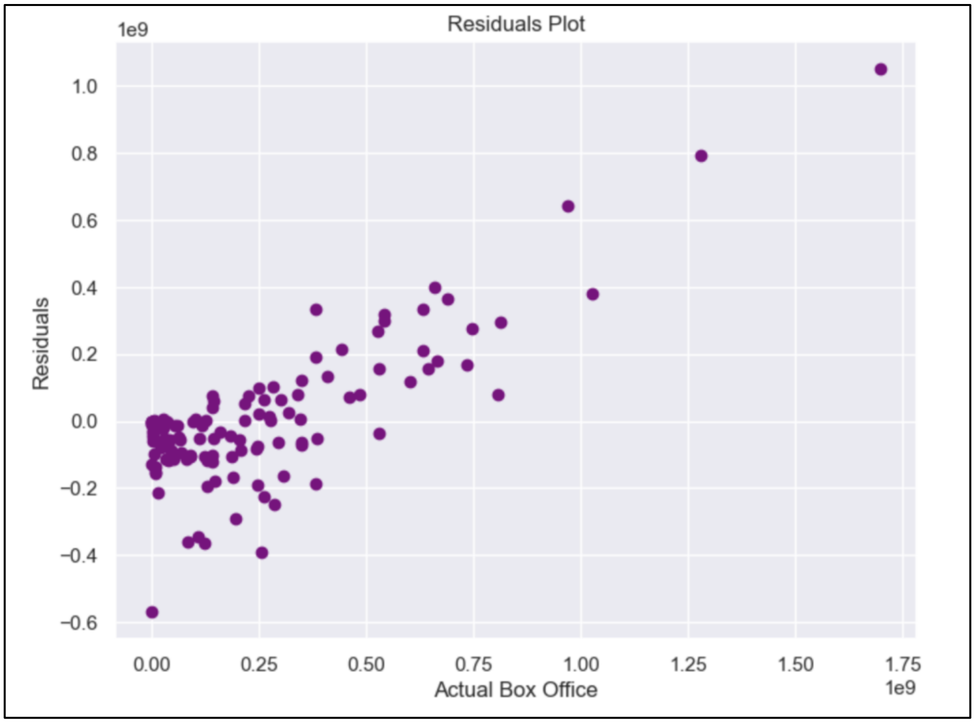

Figure 3. Animated movies residuals plot

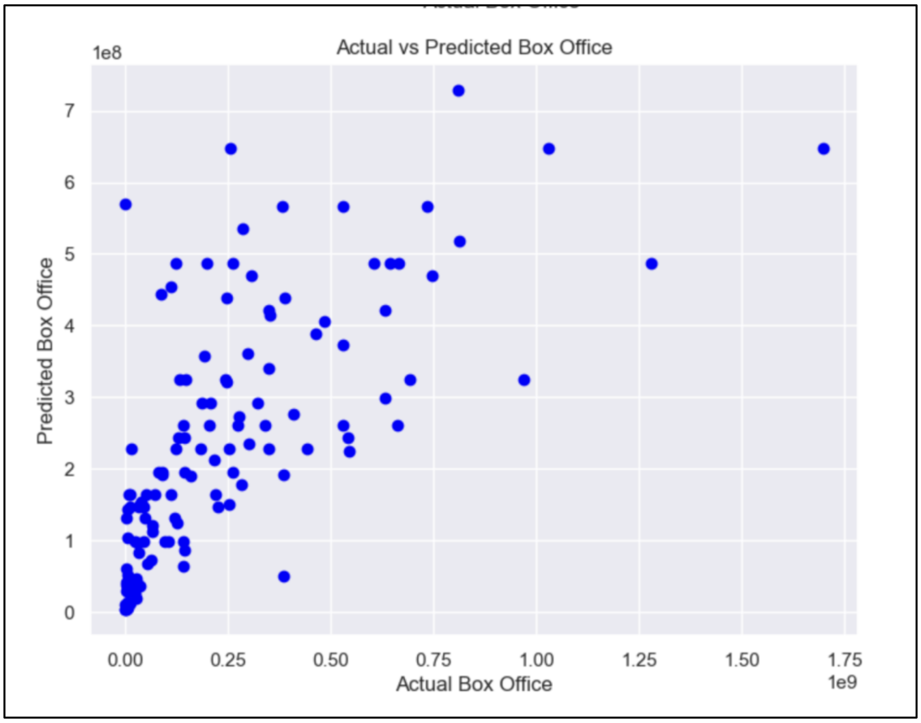

Figure 4. Animated movie actual vs. predicted box office revenue

Discussion

Through our models we found solutions to our question of if we can use the budget to predict revenue. We determined that for smaller budget films it is possible to use budget to predict the box office revenue but that for larger budget films it becomes more difficult to predict. This can't exactly be taken at face value though because there were still some films that did not match the prediction in the smaller budget range as well as the relativity of the model could be skewed because the revenue amounts are so large for our intervals. We believe that an action that could be taken as a result of this analysis is that movies with larger budgets should still focus on creating a good film technically and have good marketing because the film could still do poorly in the box office although it has a larger budget than some of the other films we analyzed. Another action in relation to our model directly is that we could try to divide the films into sections based off of their medium to see if we get more accurate results when predicting revenue based on the budget. If medium plays into the success of a movie we would most likely be able to see some trend of better performing movies in certain mediums and possibly more accurate predictions when medium is taken into consideration. An unexpected question that arose was why animated movies with larger budgets had such diverse results for their revenue. In the future if we analyzed the films with the highest budgets to see what they had in common and where they differed, we could get more insight into what makes a film successful.

In our second model, we used a multiple linear regression model, which had the movie's release date and the budget as the 2 factors on which the movie's box office revenue was predicted. The residual plots for this model show a consistent minor undervaluing for box office revenues below 250 million USD, with a typical undervaluing of approximately 100 million USD. It also displayed a consistent overvaluing for box offices above 500 million USD, getting more extreme as the box office revenue increased. For movies with box office revenues between 250 and 500 million USD, the residual plot was random yet centered at 0. Overall, this model seemed a bit stronger than our first, but primarily for movies with box office revenues between 250-500 million USD. It again reinforced the answer to our original question, being that generally small budgets have a stronger positive linear relationship to box office revenue, but that the relationship weakens as both budgets and box office revenues increase. Unfortunately, this model also has a majority of the same downsides as our first model, with no accountability for other variables in the movies, such as the genre or medium of the film. Because the residual plot seems to be linear, rather than random as one might expect, it created the unexpected question: "Does the movie's release date actually have a linear relationship with the box office revenue?" Since our first model seemed to pass the expectations for linearity and constant variance, it would seem more likely that the release date would be the variable that affected the model's residuals more in a non-random way. In the future, it could be better to group the movies by release date and then run simple linear regressions on each of those groups. It would then show the trend of how the relationship changed over time, instead of being an actual variable throughout the model and causing non-random residuals.