Predicting Film Profitability with Multiple Regression Summarization

Demond Love

Statistical/Hypothetical Question
The central statistical question that I am exploring is if it is possible to predict the profitability of a film, with a high degree of certainty, using multiple regression?

I hypothesize that we will be able to create a strong, yet incomplete model from the data in the Kaggle TMDB 5000 Movie dataset.

Outcome of the Exploratory Data Analysis
During my analysis, I used correlation analysis to analyze the relationship between the variables given and regression analysis to build a model to predict the financial success of films. Correlation analysis helped to explore the form, direction, and strength of the relationship between variables. Whereas regression analysis will be used to provide a best fit for film profit, given it's budget and quality.

With this analysis, we can confirm my initial hypothesis that we create a strong, yet incomplete model from the data in the Kaggle TMDB 5000 Movie dataset. This is evidenced by the multiple regression model's R-squared value. This value provides a measure of how well the model as a whole explains the values of the dependent variable. Therefore, our model explains over 36% of the variation in a film's profit. Since models with more features always explain more variation, the adjusted R-squared value corrects R-squared by penalizing models with a large number of independent variables. For our model, R-squared and adjusted R-squared values are pretty much the same. As shown by the F-statistic and it's p-value, this finding is statistically significant, which means that we can reject the null hypothesis that we cannot create a model to predict the profitability of a film using multiple regression.

This analysis allows us to conclude that that runtime and quality, denoted by it's average rating, don't have a meaningful effect on the film's financial success, but the film's budget does have a significant impact. Therefore, it can be said that to maximize ones profitability, one must make a significant investment in terms of budget. However, if you are to make a smart investment, then more must be known regarding the film in question, since our model only predicts a mediocre amount of the variability in the film's financial success.

What do you feel was missed during the analysis?
The number one thing missed during this analysis is a time series analysis based on when the film was released. It is widely known that the biggest movies come out as either summer blockbusters or Christmas films. I would have liked to have tested this theory.

Were there any variables you felt count have helped in the analysis?
Although it would be difficult to do so, the ability to quantify the financial viability of movie stars in a film would have a significant impact on the model. Along with factors such as marketing campaign, trailer quality, and the film's subject matter.

Were there any assumptions made you felt were incorrect?
Yes, I believe my assumptions that runtime and vote_average being significant impacts were wrong. Although their hypothesis testing returned statistically significant test statistics, neither displayed a meaningful real-world effect size.

What challenges did you face, what did you not fully understand?
The number one challenge I faced was working with inflation rates and date data. I wanted to complete the time series analysis discussed in the 'What do you feel was missed during the analysis?' section. However, I couldn't complete this because I didn't know how to adjust my budget and revenue data based on inflation. Completing this analysis without doing so would have been incorrect, since my data ranged from 1917 through to 2016.