Abstract

The goal of this brief is to summarize an attempt to predict the goss total return of a movie based on other aspects of the movie available from the website: http://www.boxofficeguru.com/film.htm.

The gross total can be calculated as

$$\begin{aligned}\text{Gross Total} &= \text{Gross Opening / 1st Week \%} \\ &= \text{Theaters Widest} \cdot \text{Per Theatre Revenue}\end{aligned}$$

For the purpose of this short project we will therefore assume that '1st Week %' and 'Per Theater Revenue' are unavailable. As is convention for positive data, we will analyze 'LogTotal' := log('Gross Total') in terms of the remaining variables.

Since the data has few covariates which each have intuitive meaning, we opt for a linear regression approach. Firstly, we will handle the categorical variable 'Distributor' which contributes to the design matrix in the form of indicator columns (automatically handled by the patsy method dmatrices). Next we looked at the seasonal variation of returns, i.e., 'LogTotal' as a function of the day of the year, simply called 'Day' in our code, and noticed a nonlinear fluctuation so the natural transformation of 'Day' was a bspline. The original 'Gross Total' variable seemed linear in the number of days the movie played in theaters, 'DaysPlayed', so we used used log('DaysPlayed') as still another covariate. The same holds for 'OpeningGross'. We then discarded the number of opening theaters, 'OpeningThtrs', because it 1.) had a correlation of 0.9 with the maximum number of theaters the movie played in, 'MaxThtrs', and 2.) there was no trend in revenue for movies with few small vs. large 'OpeningThtrs'; the partial R^2 for 'LogTotal' with respect to 'OpeningThtrs' was 0.02.

Of the selected covariates, there was essentially a negligible amount of missing data. Observations missing any of these attributes however were dropped before the regression analysis began. We used the error metrics given by the mean and median of the following vector of errors:

$$\text{Error} = \frac{|\text{Total} - \text{Fitted}|}{\text{Total}}$$

to assess quality of prediction. We were able to achieve a median error of 18% and R^2 of 0.97 for the testing data (titles beginning with 'D'). See the next page for a plot. We were also able to verify that the testing residuals were essentially normally distributed.

Testing Data Results ('D' titled movies)

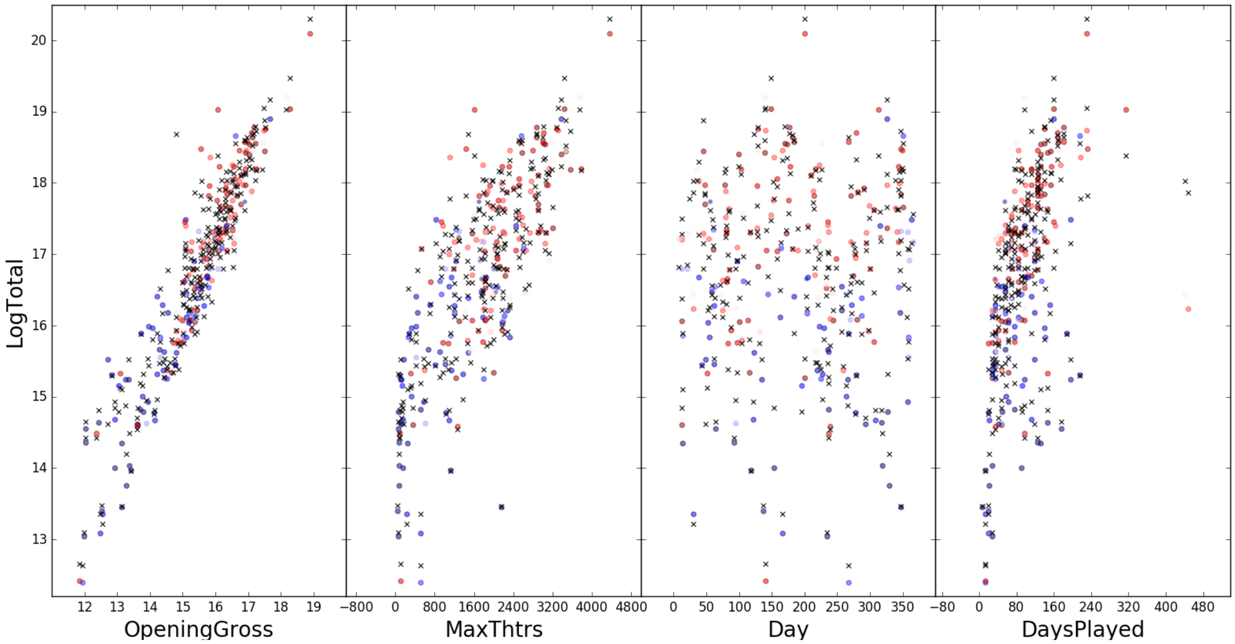

Figure 1: Fitted values in black. One outlier in the DaysPlayed graph has been removed for clarity. Color is assigned by movie distributor. The red hue increases with the distributor's average return.