# Predicting the Domestic Total Gross of Movies with Linear Regression

**By Alex Smith**

### Overview: Problem and Background

The movie industry is a multi-billion dollar industry that has had consistent and substantial growth in the past couple decades. According to statista.com, 2.8 billion dollars was spent in box offices in 1980 and by 2016 that amount has nearly quadrupled to 11.37 billion dollars. I was curious to see if the gross of a movie could be predicted well, before it's release, so I posed the following question: can we predict the total domestic gross of a movie using a linear regression model? The model I built is useful for a movie production studio looking to get an estimate of the domestic gross for a particular movie, or to an executive producer deciding which movies to invest in, with only certain information available.

### Data

The data I used to build my linear regression model was mainly scraped from Box Office Mojo (boxofficemojo.com) using BeautifulSoup. For about 20 genres, I scraped the title and links from each genre page. For each title and link pair, I then scraped the following features: title, director, budget, runtime, genre(s), rating, release date, and domestic total gross. After that, I scraped a page from Wikipedia to get a list of directors that had won a Golden Globe for directing. I used this list to engineer a feature called winning director, which identified if a director for a particular movie had previously won a Golden Globe for directing. I also created the feature season from the month of the release date, and split that into the dummy variables winter, spring, summer, and fall. Additionally, I logged the domestic total gross (my target variable), and budget, as both were high in magnitude, in the millions. The final features that I chose to use in my model are: logged domestic total gross, logged budget, runtime, rating (G, PG, PG-13, R, and NC-17), release season (winter, spring, summer, and fall), release year, and winning director.

### Modeling

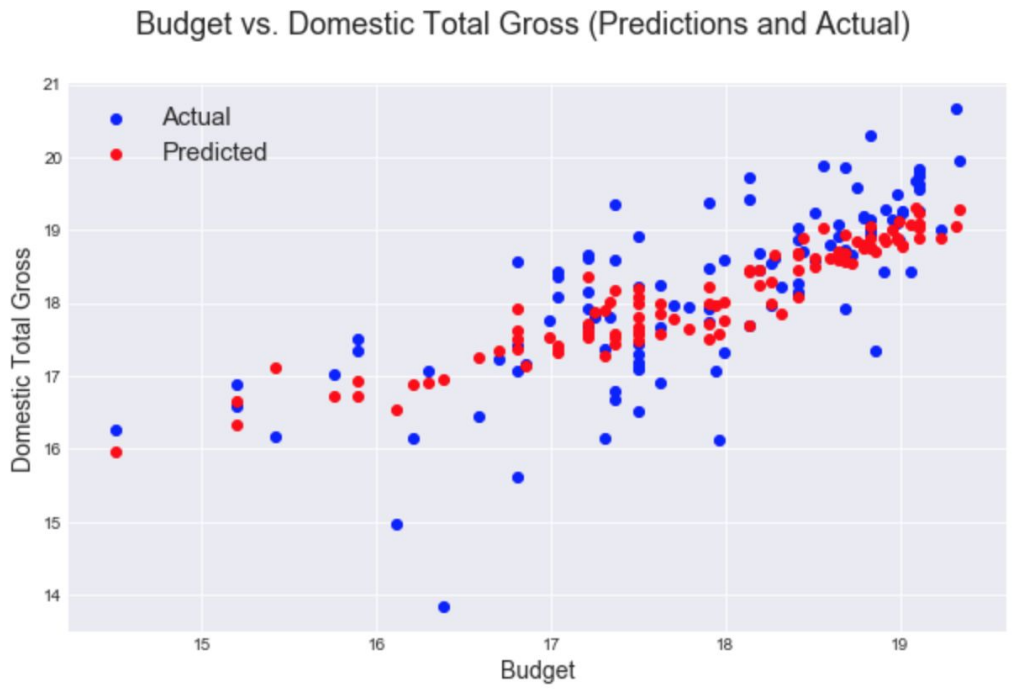
Once I had my dataset cleaned and ready to model, I looked at the correlations between features and made a scatter plot of each feature against my target variable, domestic total gross. Based on these plots, combined with the knowledge of the scale of budget and domestic total gross (both are mainly in the millions), I decided to log both variables. The two base models that I built were identical linear regression models except the first used the non-logged budget as a feature and the non-logged domestic total gross as the target variable. Meanwhile,

the second model used the logged budget as a feature and the logged domestic total gross as the target variable.

I decided to use R-squared as the performance metric of comparison between my models because R-squared is the proportion of variance explained and gives me a good sense of how predictive my model is. Also, it is the default metric of Scikit-Learn, the package I used to build my models.

While I would have like to do a 60-20-20 train-validate-holdout split, I only had about 550 rows of data so this was not possible. When I tried it, my R-squared was lowered by 0.10. Instead, I did an 80-20 train-test split. I used 5-fold cross-validation with both models after comparing the results with 2-10 folds.

After training my two models, I compared the R-squared for both (first = 0.407, second = 0.411) and chose to move forward with the second model (logged budget and dtg) into Lasso / L1 regularization. First, I created a baseline Lasso model with an arbitrary alpha of 0.01 before doing a grid search for the hyper-parameters of alpha and normalize. Using the results of the search, I created another Lasso model with the optimized hyper-parameters. This model slightly improved the R-squared up to 0.447. The optimized Lasso model gave me 0 coefficients for the features year, spring, and PG-13, and the highest coefficients for the features NC-16, log_budget, and R, when normalized. Interestingly, the model performed better on the test data compared to the train data, leading me to believe there may be some underfitting. My model's predictions vs. the actual values for budget vs. domestic total gross can be viewed below.



**Results**

With my Lasso model optimized, I calculated the RMSE and compensated for the logging of domestic total gross earlier in my process. The RMSE of my final model was $14,850, not too bad considering the numbers my model deals with are in the millions. The R-squared of my final model was 0.447 for the training data and 0.609 for the testing data. I plotted the residuals (below) for my model and there was no clear heteroskedasticity.

My results led me to a few thoughts moving forward. First of all, I would like to delve deeper into the possibly underfitting that is occurring with my model. Additionally, I would like to scrape and incorporate more data into my model to create a validate and holdout test set, and to increase the predictive ability of my model. I would also like to integrate genre, composer, actor(s) and actress(es), and social media likes or mentions.

**Challenges**

Originally I was interested in building a linear regression model to predict the domestic total gross of exclusively for Sci-Fi movies including various Sci-Fi subgenres. However, there was not enough data available for this objective, as there has only been a certain number of Sci-Fi

movies made. Because of this, I broadened my objective to include other genres as well, incorporating approximately 20 genres total.

Additionally, I originally intended to incorporate the feature genre into my models, but time and scope of the project did not allow for that in this iteration. I will explore this in the future.