Predicting Movie Box Office Sales

Project Luther - Weeks 2 and 3 of the Metis Data Science Bootcamp

By Steven Bierer October 15, 2018

Location Scouting [Overview]:

Our second project, the first to be tackled individually, was focused on scraping data from the internet and creating a predictive model using linear regression techniques. The choice of data set was open-ended, and I chose to analyze movies for which online data sources are plentiful. Also, I really like movies, and I consider myself having a fair amount of "domain knowledge" in the field.

Movies can be very expensive to produce, and their box office success is far from guaranteed, even when the writing, acting, and directing are excellent. That rule is particularly critical to films that aren't granted huge production and marketing budgets, making it hard to determine in advance how many people will know about the movies, much less pay for tickets to see them in theaters. Studios and production companies may be less inclined to make quirky romantic comedies or moody crime dramas - hardly summer block-buster material - if they continually lose money on them. For this reason, an ability to determine how well a lower-budget film will perform at the box office would be quite useful for the financial backers of these relatively risky investments.

The approach of this study was to apply linear regression techniques to movie information taken from the well-known Internet Movie Database (IMDb.com). The following tools and sources were used in the analysis:

Sources: IMDb.com, starting from their "Advanced Title Search" page dfdf (https://www.imdb.com/search/title/form).

Tools: 1) Beautiful Soup for information searching and extraction from IMDb.
2) Python and the standard libraries numpy, pandas, matplotlib, and seaboard for routine data manipulation and graphical display.
3) Python modules sklearn and statsmodel for regression modeling.

4) The code editor Spyder was used to develop a module of functions for data scraping. This marks the first time an editor other than Jupyter Notebooks was incorporated into the work flow.

Concepts Applied: Linear regression, regularization, cross-validation, normality testing, mean-squared error, R-squared statistic, F-statistic.

Casting Call [Features and Data Filtering]:

Only feature films were targeted for this analysis, covering the years 2000 to 2018 (the week of October 8th). The following types of data were obtained from the data base:

Data Element	Type	Transformation
Gross Box Office Sales	Numerical	Box-Cox
Opening Weekend Sales	Numerical	Binned, Smoothed (for time-series analysis)
Budget	Numerical	Box-Cox
Release Date	Numerical	Time Filtering and Averaging
Genre: Family, Comedy, etc	Categorical	Dummy Variable (0 or 1 by category)
MPAA Rating: G, PG, etc	Categorical	Rank Score (0 to 4)
Language: English or not	Categorical	Dummy Variable (0 or 1)
User Rating Score	Numerical	None (only used for error analysis)
User Rating Count	Numerical	None (only used for error analysis)

Information from a total of 7027 feature films was extracted. As seen in Fig. 1 below, there are a lot of low-quality films in this set in terms of both budget and ticket sales. Movies with no budget data, or budgets less than $5 million, were excluded as most of these are screened in an extremely limited number of theaters. That filtering narrowed down the number of films to 2693 for most of the analyses conducted. Required transformations of the variables are described below.

Script Reading [Data Assessment and Feature Transformation]:

Both budget and gross box office tickets sales were highly skewed toward values below $10 million. This asymmetry is evident in the histograms of the joint-plot graph shown in Figure 1. For this reason, both variables were transformed to impose greater normality on their distributions, as described in the next section.

Fig. 1 also demonstrates a substantial correlation between budget and total sales. Across all movies with both budget and box office data, the correlation coefficient was 0.72.

Correlation coefficients among all the variables were calculated. Excluding opening weekend box office sales, user ratings, and number of user ratings - none of which were directly used in the predictive modeling - only one feature pair aside from the budget/total sales pair had a correlation coefficient over .40: MPAA rating and Release Year exhibited a correlation coefficient of 0.47.

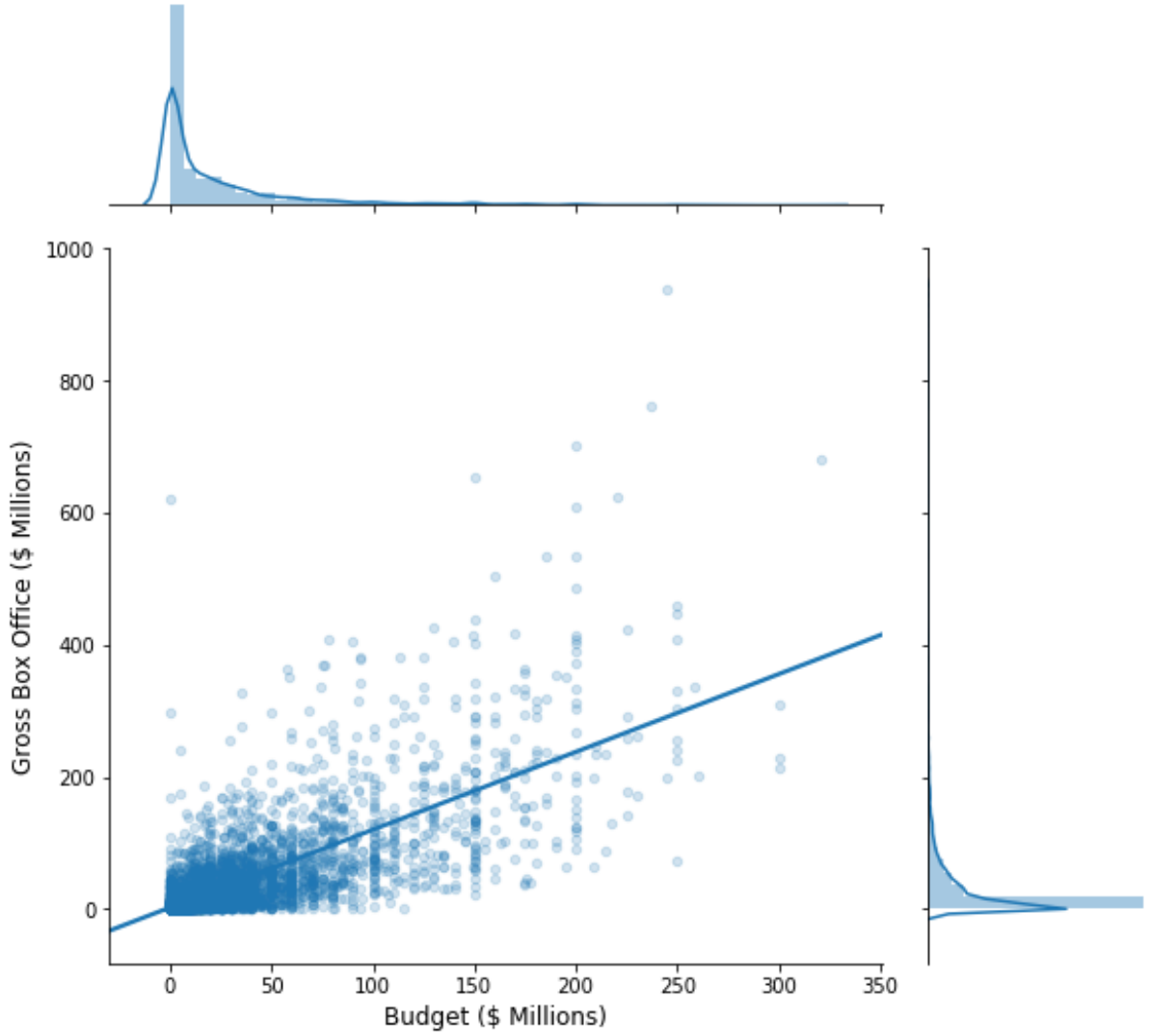

Figure 1. Scatter plot and histograms of total U.S. box office receipts versus movie budget, for all films with both values > $0 (n = 5726). Solid line defines the best linear fit to the data.

For analyzing trends over time and implementing the linear regression, as

discussed below, the movie data was separated into two groups. "Low-budget" movies (n = 1825) were defined as having budgets over $5 million but less than $50 million; "high-budget" movies (n = 868) had budgets over $50 million.

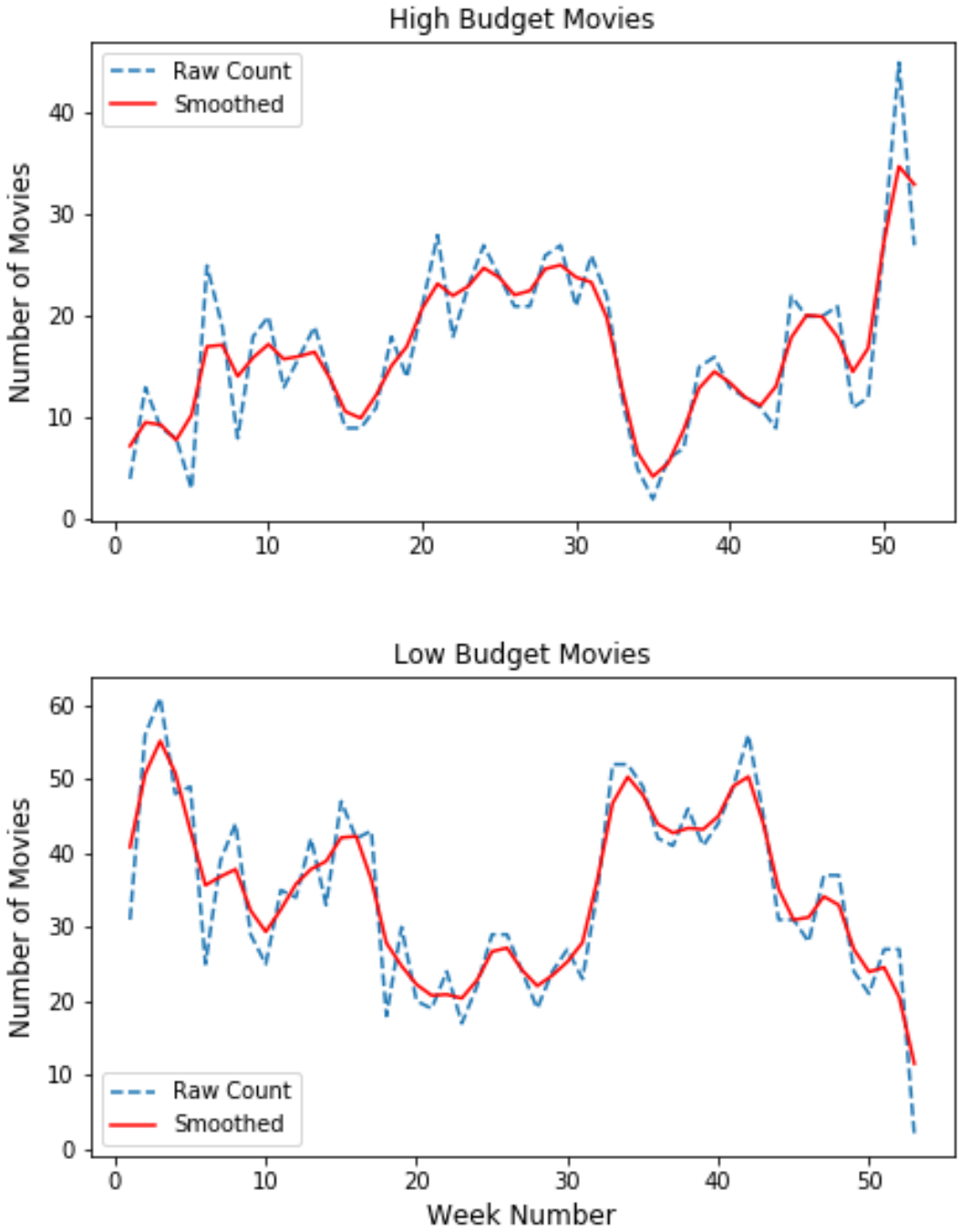

Figure 2. Time series for depicting the number of high-budget (top panel) and low-budget movies (bottom panel), averaged by week across all years of data (blue dotted lines). From the

smoothed values (red lines), two normalized scores were created: “Seasonal High” and “Seasonal Low”.

The numbers of movies released during every week of the year were accumulated over the years in the data set, separately for the low- and high-budget subsets. The resulting time series, shown in Figure 2, represent seasonal trends in film release. Interestingly, the major peaks and valleys in the traces are nearly opposite for the low- and high-budget movies. After smoothing with a 3-point hamming window, each time series was normalized by subtracting its mean and dividing by the standard deviation. This process yielded “Seasonal Low” and “Seasonal High” scoring metrics that were used as additional features for linear regression.

A similar procedure was applied to opening weekend box office sales for high-budget movies, although in this case the values were summed for all weeks in the data range rather than binning across years. The rationale was that such a metric would quantify the “competition” from other movies that are performing successfully in theaters. However, regression coefficients for this feature in first- and second-order models were not significant and were thus excluded from the final modeling procedure.

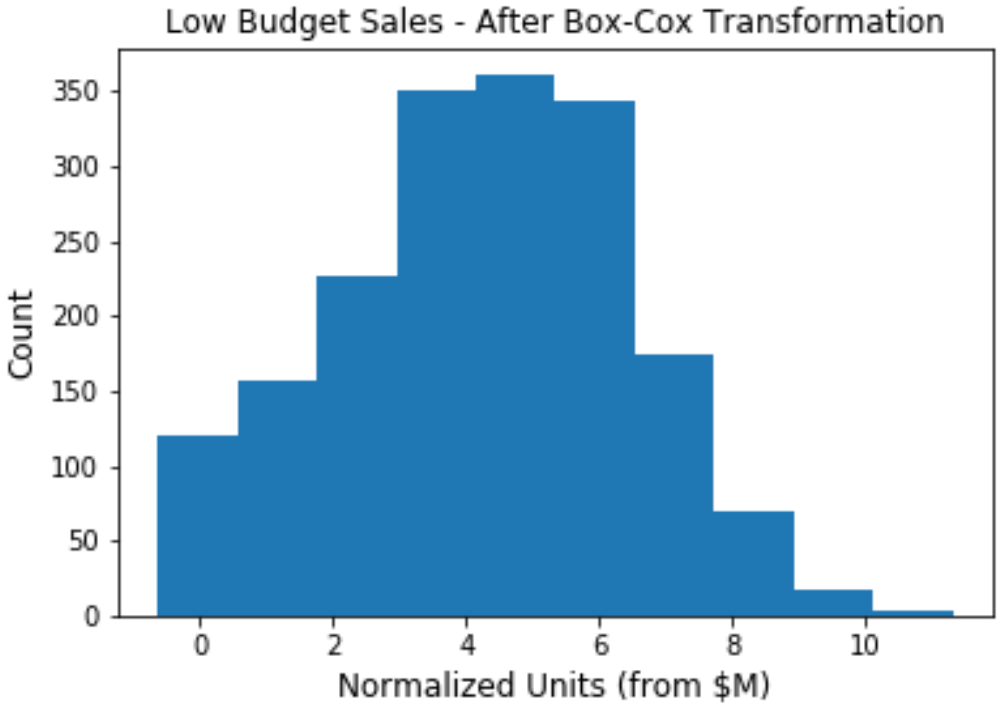

Figure 3. Distribution of total box office sales for low-budget movies following transformation using the Box-Cox method

(power parameter = .436). The distribution is notably less skewed that before transformation (compare to top histogram in Fig. 1).

Given the large heteroskedasticity apparent in the budget and box office sales data (Fig. 1), these variables were nonlinearly scaled using the Box-Cox transformation to give a more gaussian-shaped distribution of values. The transformed distribution for budget, for the low-budget subset of movies, is shown in Figure 3 as an example of the outcome of this procedure. Data conditioned in this manner are more appropriate for the least squares regression techniques used in the present analysis, which assume a normal distribution of residual errors.

Finally, the categorical data used as regression features were transformed to binary (0 or 1) values for each class in the category. The genres chosen were "Action" (including movies tagged as "Adventure" and "Thriller"), "Comedy", "Family" (also including "Animated"), and "Horror". Language was transformed to a feature "English" describing whether a film was primarily spoken in English (1) or not (0). MPAA ratings were transformed to a scale increasing with the size of the possible audience (NC-17 or unrated = 0, R = 1, PG-13 = 2, PG = 3, G = 4).

The Set Piece: [Regression Analysis]

The first stage in the regression analysis was to evaluate how well a simple first-order ordinary least squares (OLS) model performed on the data. As expected, preliminary OLS modeling for the high-budget data resulted in reasonably good fits using all the data (Figure 4, top panel). The R-squared and adjusted R-squared values were 0.368 and 0.361, respectively; the F-statistic was 50.0 ($p<<0$), indicating very high significance of model parameters. Significant features (based on the t-test) were Budget, Release Year, and English with positive coefficients and Seasonal Low and Horror with negative coefficients, which makes intuitive sense. OLS model fits for low-budget data, on the other hand (Fig. 4, bottom panel), was not as good. R-squared and adjusted R-squared were 0.221 and 0.216, respectively, with F-statistic = 51.4 ($p<<0$). Significant positive coefficients were Budget, Release Year, Seasonal Low and High, Comedy, Family, and English.

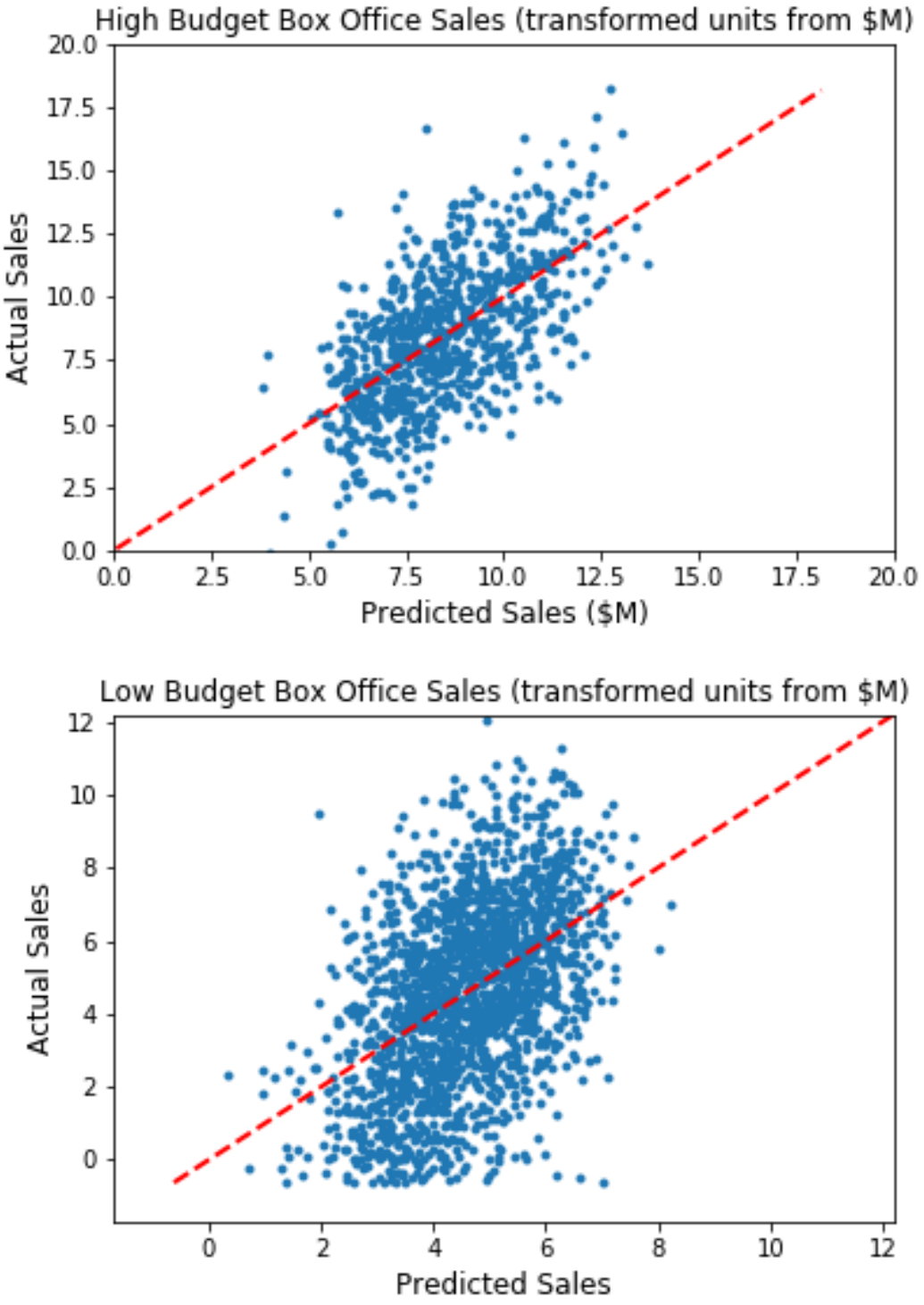

Figure 4. First-order linear regression results. Scatter plots depict actual versus predicted box office sales for high-budget (top panel) and low-budget films (bottom panel), after the power transformation. R-squared values are .368 and .221,

respectively. Red lines indicate a 1:1 ratio of the actual and predicated values.

In the next stage, K-folds cross-validation was used to assess the addition of higher-order polynomial terms to the model. For this procedure, the data was split into four blocks for repeated train/test scoring based on the mean squared error, and the order of the polynomial transformation of the features was stepped across several values. Results for the low-budget data are shown in Figure 5 for orders 0 to 5. As is typical, the error for all the data (green line) decreased steadily with model order, indicating overtraining. For the cross-validated output (blue line) averaged over consecutive subsets of the data not used for training, MSE levels were lowest at orders 1 and 2. This outcome suggests using second-order polynomials in the next stage of regularization in which the number of model features is narrowed down.

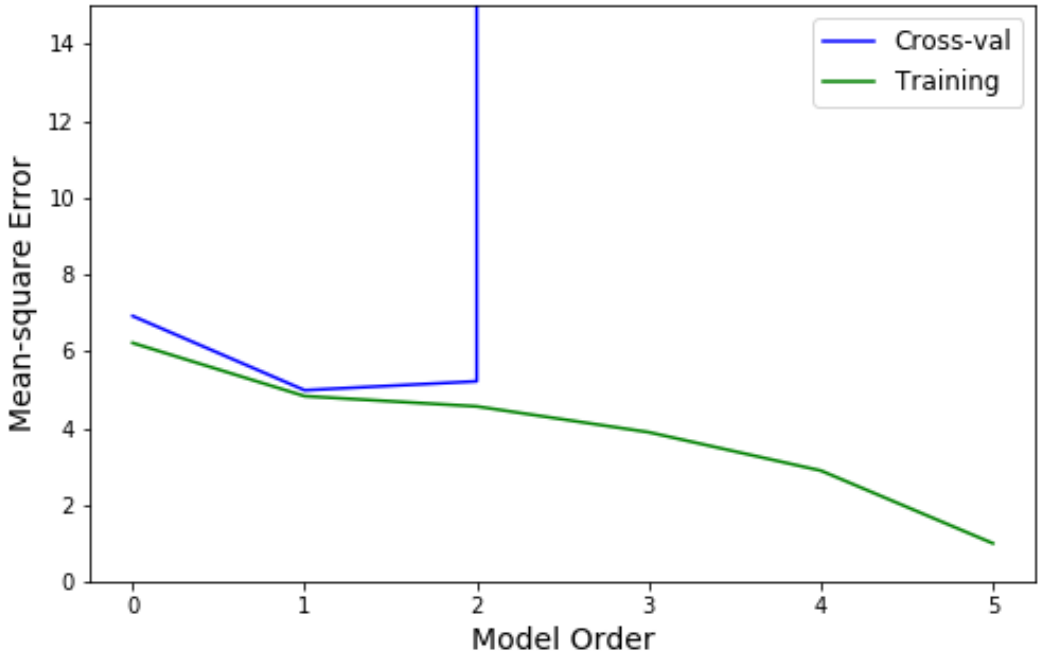

Figure 5. Cross-validation results for low-budget data. The MSE trajectory of the cross-validation data suggests that a closer examination of features using a 2nd-order model is warranted.

Lasso regularization was used to penalize features contributing little to reducing mean square error in the prediction. Using a cross-validation process on 70% of the low-budget data (30% was reserved for testing), the penalty factor, lambda, was varied over a wide range and the MSE and R-squared value calculated at each step.

The results of this process are presented in Figure 6. MSE in the top panel for the final test data decreases to a minimum at an alpha of 10^{-2} for the test subset of data. A comparable optimum is seen for the R-squared value in the lower panel. At this optimal lambda value, the R-squared value was 0.194 in transformed units.

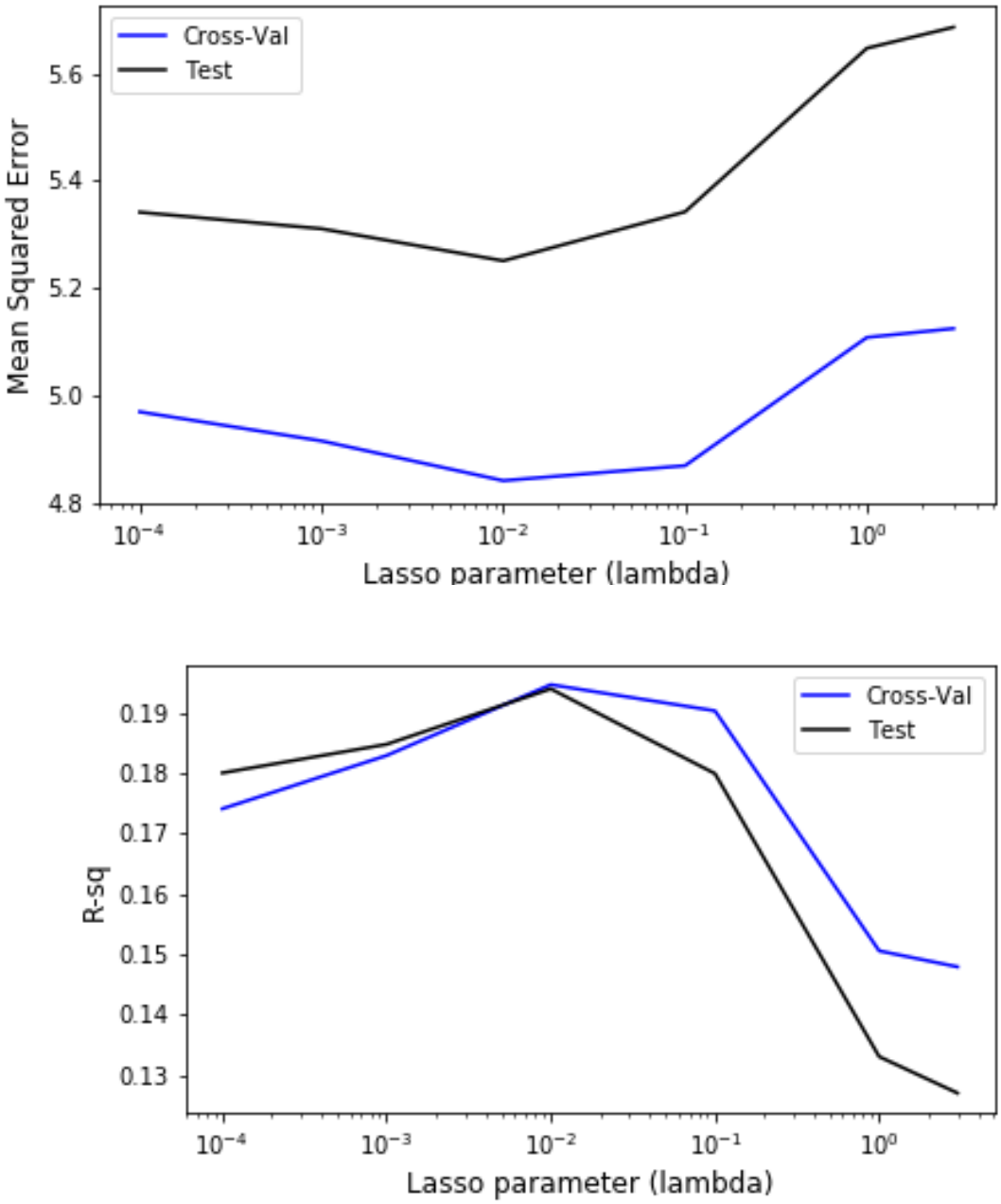

Figure 6. Cross-validation results for low-budget data using Lasso regularization. Both panels, for MSE (top) and R-

squared (bottom), indicate that a lambda value of 10^{-2} is optimal.

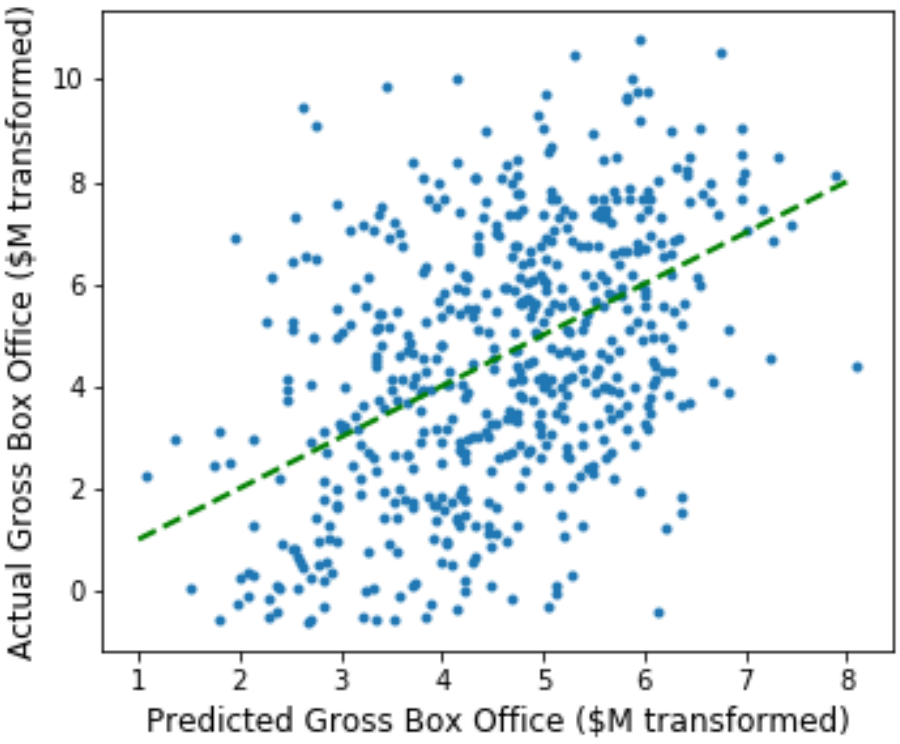

Figure 7. Actual versus predicated gross box office sales (Box Cox transformed units) for the test subset of low-budget movies. Green line marks a 1:1 ratio for a perfect prediction.

Figure 7 depicts the actual versus predicted low-budget box office values using the test data only. The high degree of scatter in the points reflects the low R-squared value.

Wrap Up: [Conclusions]

While the box office sales for movies with high-end budgets (> $50 million) could be fairly well predicted given the feature set, the prediction for low-budget movies was rather poor. The wide separation of the training and testing results for mean squared error in Figure 6 indicates that the data may have been under fit. Adding additional data elements for each movie as features would likely have improved the fitting. Indeed, during preliminary analysis of the data, the addition of post-hoc

movie data like user ratings substantially improved the ability of the linear regression to predict box office sales, even for low-budget films.

One type of data that may have improved performance of the model is a better "competition" score. A week-by-week time series was developed for this purpose, but it was based on only the total opening week sales of high-budget movies. Perhaps a better metric would have considered genre-specific movies, such that a low-budget family-friendly comedy would have more competition from other family-friendly comedies, under the assumption that movie-goers can only see so many movies over the course of a few weeks.