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 (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, normal distribution, 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	Time-Series
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 the graphs below, there is a lot of junk in this set. 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 to 2693.

Script Reading [Initial Data Assessment]:

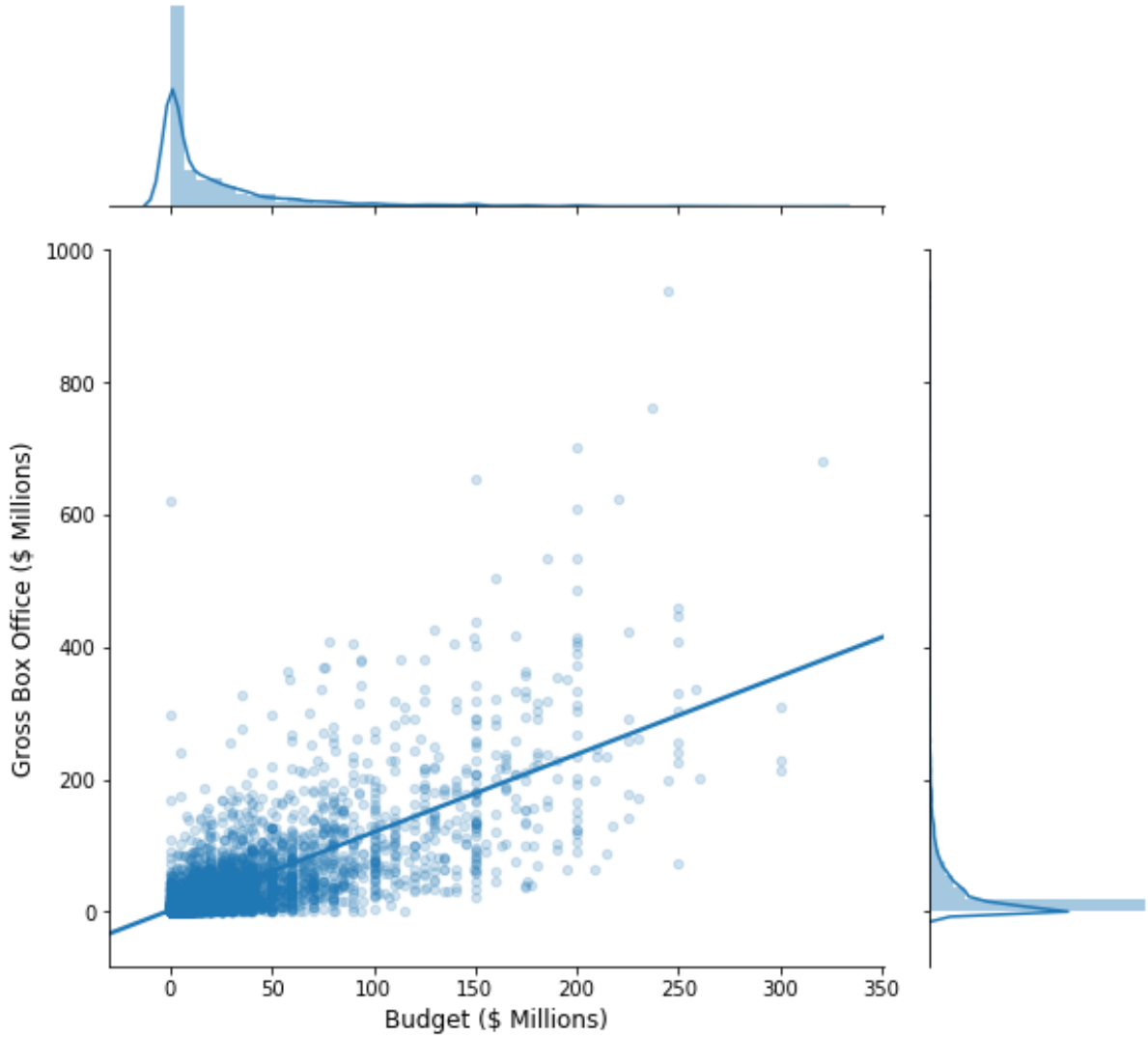

Figure 1. Scatter plot of total U.S. box office receipts versus movie budget.

1b) Histograms of sales and budget; what are good cut-off values for no-, low-,and high-budget movies?

- TOTAL SALES: such a skew! Even when narrowing down to < 20 million

df_lowbudget (N = 1825), df_highbudget (N = 868), df_nobudget (N = 3129 … explore if time)

No correlations

Box-cox

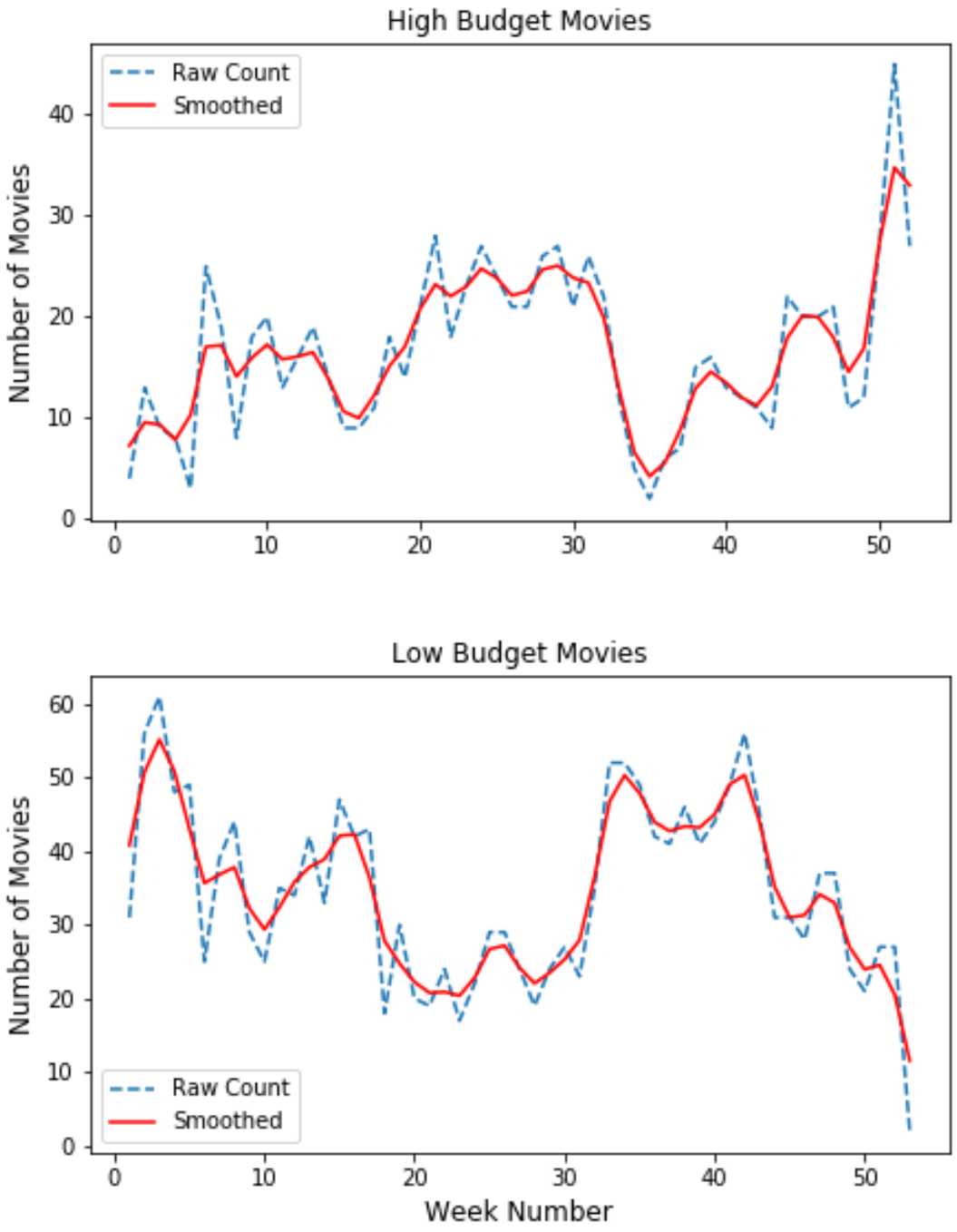

Figure 2. Time-series for high-budget (top panel) and low-budget number of movies (bottom panel), averaged by week across all years of data (blue dotted lines). From the smoothed values (red lines), two normalized

scores were created: Seasonal High and Seasonal Low.

The Set Piece: [Regression Analysis]