Bombs Away!

Predicting Movie Revenue

Juan Morales & Yakir Forman

Goal and Data Source

Goal: to build a model that can predict a movie's revenue

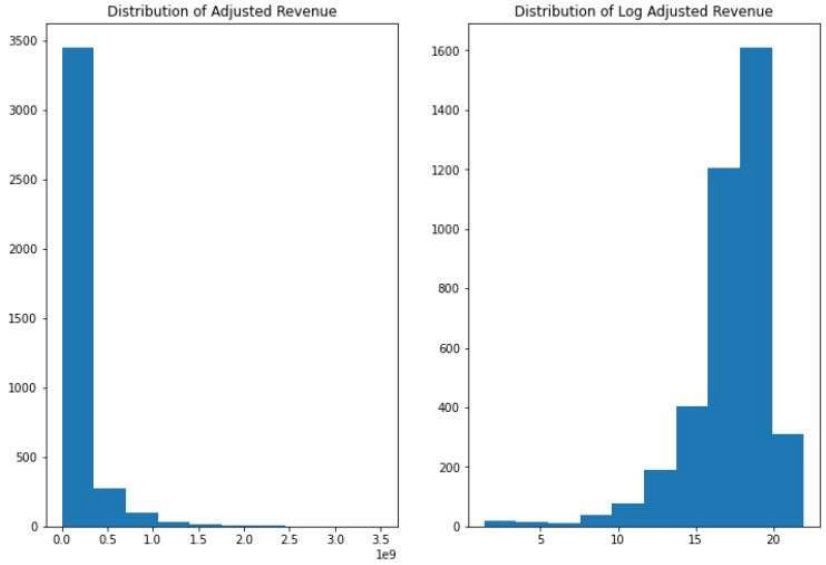
Data: tmdb data of several movies including release date, budget, tagline, director, actors, etc.

Data Cleaning:

- Remove data with missing revenue information
- Ignore factors that would not be known prior to release, e.g., popularity

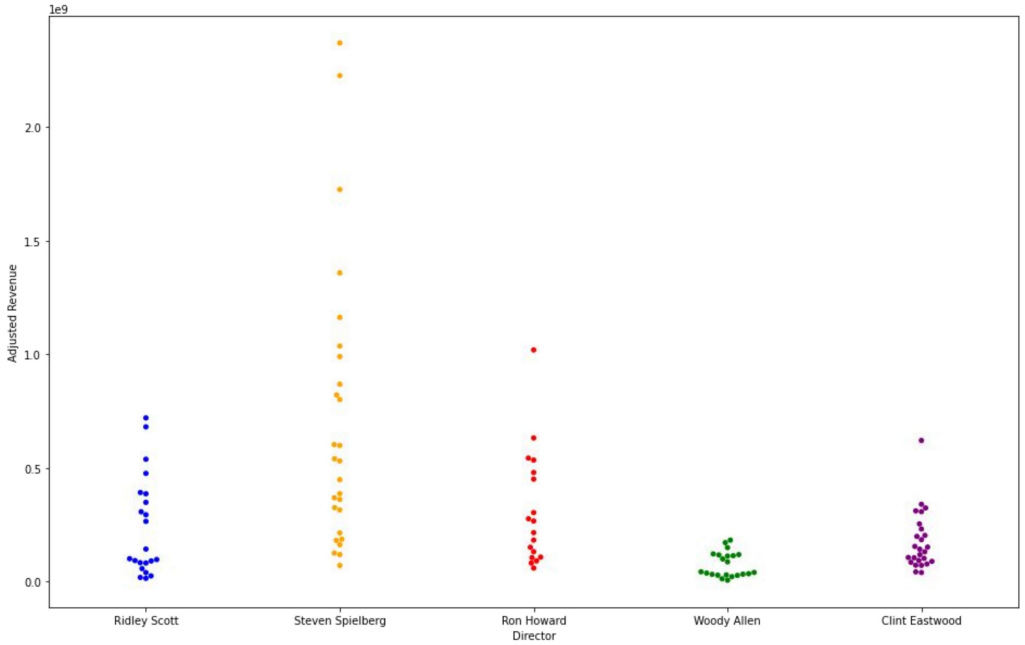
Modeling Approach

- Adjust for inflation
- Explore linear relationships with log revenue, which better matches distribution of revenue
- Log transformations of predictors which also have log distribution, e.g., budget
- One-hot encoding for actors, directors
- TF-IDF to find keywords in taglines
- Lasso for feature selection



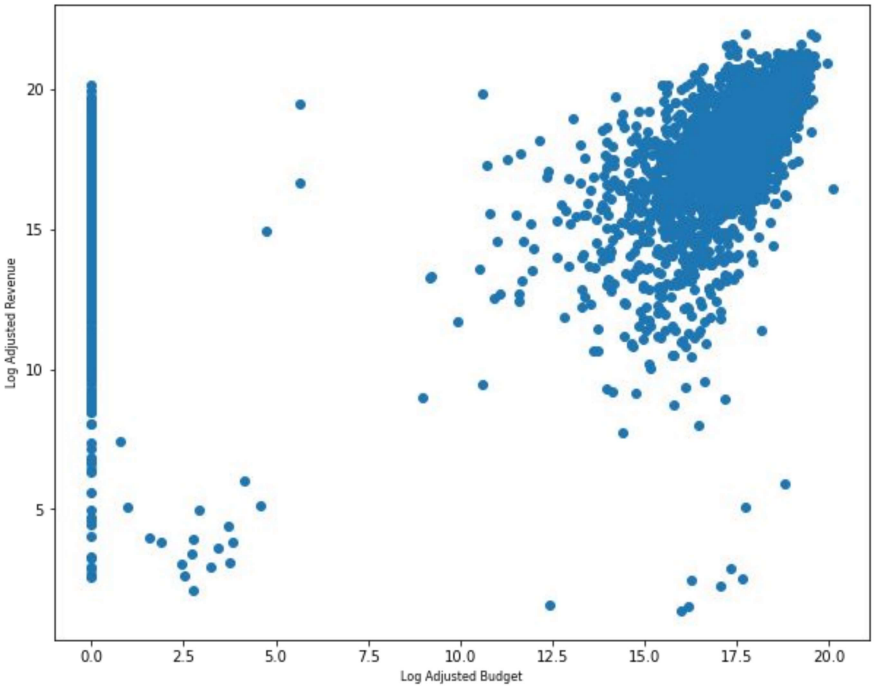
Some Examples from our Exploratory Data Analysis

Top 5 directors – we see that director doesn't strongly predict greater revenue, but a particular director (specifically Steven Spielberg) might be considered.



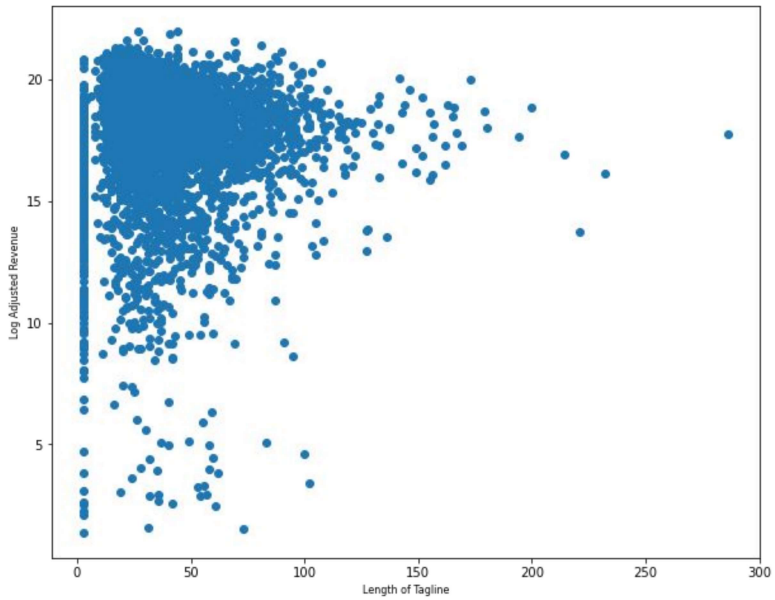
Some Examples from our Exploratory Data Analysis

Budget has the clearest relationship with revenue.



Some Examples from our Exploratory Data Analysis

Length of tagline does not seem to have a significant relationship with revenue. (However, presence or absence of a tagline performs better as a predictor.)



Final Model

The best multiple linear regression model without feature selection performed as follows on the train set:

Avg baseline MSE: 7.18
Avg single linear regression MSE: 5.76
Avg multiple regression MSE: 4.69

With Lasso we were able to select for six features – budget, runtime, Adventure genre, Family genre, release year, and presence of tagline – with MSE 4.83 on the train set

This model had MSE 4.86 on the test set

Note that the lasso coefficient for log budget was 0.85, significantly higher than all others which were below 0.3. (Data were scaled so coefficients are comparable.)

Key Takeaways

- The single most important factor in predicting revenue is budget. A higher-budget film is likely to have higher revenue. Production companies should keep this in mind when determining budgets.
- In addition to budget, some other factors predict a higher revenue: longer runtime, presence of a tagline, and belonging to the Adventure and/or Family genres. These could all be considered when deciding which movie to back and/or what changes to make to a movie or its advertising.
- While considering only these factors performs similarly to considering a wider range of predictors (including the presence of specific actors and directors), and it does perform better than a baseline average, the improvement is only modest. There is still a lot of uncertainty involved.