Predicting a Movie's Revenue Before its Release with IMDB Bot

BROWN DATA SCIENCE

Xiling Zhang[1], Ben Foulon[2], Yuchen Hua[3], and Yihao Zhou[4]

[1]Department of Computer Science, Brown University; [2]School of Engineering, Brown University;

[3]Department of Physics, Brown University;

BROWN Computer Science

Introduction

- Making movies is very expensive. While a good movie can be a financial boon, a box office flop can devastate investor earnings and cripple smaller studios
  - In an era of increased competition from streaming services such as Netflix and of reduced work activity due to COVID-19, making movies is harder than ever
- Being able to predict movie revenues *a priori* is thus a lucrative endeavor (and, for smaller studios, key to survival)
- There is a wealth of publicly available data on movies and customer opinions (expressed in Tweets, for example) that have the potential to be predictive of box office success
- Movie investors thus have the means (data sources), motive (profits), and opportunity (data analytics) to predict a movie's revenue before its release
- With this in mind, we introduce IMDB Bot, a machine learning (ML) model that uses movie information from IMDB Bot and director and writer popularity from Tweet Sentiment Analysis to predict a movie's revenue
- Our model, using Random Forests, has an R^2 of 0.55 and a mean absolute error (MAE) of \$14.6 million. Given that revenues are usually in the 10s-100s of millions, the model can reasonably say whether a movie is likely to boom or flop

Data and Methodology

- **Movie Data**
  - From the IMDB database, we obtained the title, release date, genres, director, writer, actors, countries, and runtime of movies released from 2009 to 2019
  - We used web scraping to get each movie revenues, which were not listed in the IMDB database
- **Data Details**
  - Genres, director, writer, actors, and countries were used variably as categorical features. These features were encoded into binary presence/absence variables based on their frequency in the data using Sklearn's CountVectorizer
  - Budget information was too sparse and therefore wasn't considered as a feature
- **Sentiment Analysis**
  - Tweets about each movie's director and writer were found with Twitter's API via Python's Tweepy package
  - Sentiment analysis was performed on the collected tweets using the TextBlob package in Python
  - Values can range from -1 (hate) to 1 (love)
  - We input (1) tweet count and (2) average sentiment analysis for each movie's writer and director into our data
- **Training Details**
  - 80/20 train/test split and 10-fold cross validation

Model Selection

- Random Forest (RF) and Gradient Boosting (GB) models performed the best
- RF selected as the ML approach for our model
- L_2-regulated (Ridge) linear regression does much worse than the more complex RF and GB models, suggesting that correlations between the features and revenue are non-linear
- Sklearn's DummyRegressor (strategy=mean) was used as a baseline. As expected, all models capture more correlation than the DummyRegressor

Model Type	R^2	MAE (Millions USD)
Random Forest (RF)	0.554	14.60
Gradient Boosting (GB)	0.552	15.81
Ridge Regression	0.361	25.29
Decision Tree	0.107	17.77
Dummy Regressor (baseline)	-0.031	17.13

Table 1: Models and their performance. 30 features max used with CountVectorizer. All categorical features included.

Results and Analysis

Claim #1: Using a 30-feature maximum for count vectorizing categorical variables gives the best balance between R^2 and MAE

Max Features	R^2	MAE (Millions USD)
10	0.542	14.97
20	0.555	14.67
30	***0.554***	***14.60***
40	0.550	14.50
50	0.548	14.49
100	0.544	14.38

Table 2: The model's performance based on the max number of features used in CountVectorizer. All category columns were used. RF with a 30-feature limit selected as the IMDBot model.

Claim #2: Genres play an important role in movie revenues

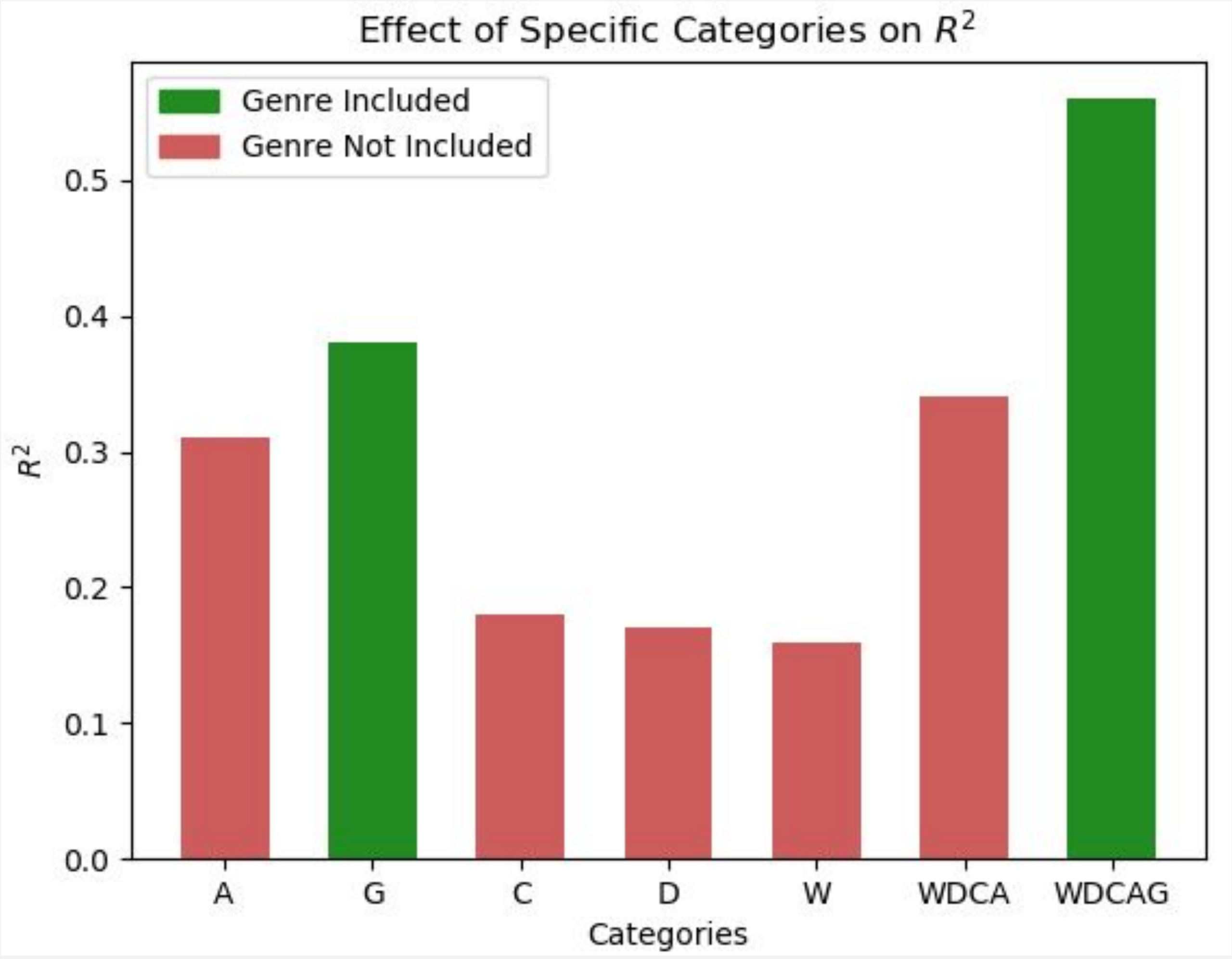

Figure 1: R^2 values based on category variable inclusion. Category labels refer to Writers (W), Directors (D), Countries (C), Actors (A), and Genres (G).

Claim #3: IMDBot works best with sci-fi, action, animation and comedy genres

- The results are not solely count-dependent: Drama, Romance, and Documentary have high counts but low R^2 values, while Sci-Fi has a low count but a high R^2.
- There are some anecdotal cases of related genres having similar R^2 values (Sci-Fi and Fantasy, Horror and Crime).

Genre	Count	R^2	MAE (Millions USD)
Adventure	1772	0.646	66.7
Sci-Fi	579	0.632	64.4
Action	3034	0.614	40.9
Animation	1125	0.559	39.5
Comedy	7727	0.553	14.9
Fantasy	864	0.461	49.9
Mystery	1069	0.341	17.9
Thriller	2544	0.278	19.0
Horror	1582	0.216	15.3
Crime	2051	0.210	15.4
Romance	2993	0.195	11.0
Drama	11397	0.137	11.8
Documentary	2820	-1.711	16.6
Musical	206	-21.467	16.4

Table 3: Selected genres and their effects on the model.

Model Prediction Analysis

- In general, as movie revenue increases, our model's absolute error on that movie's predicted revenue also increases.
- However, errors for most movies are small relative to their revenue, especially for movies making less than \$500 million (which is most movies)
- This suggests that the model can estimate pretty well which movies will be hits or flops. After all, a movie making \$200 +/- \$14.6 million is good at either extreme, whereas a movie making only \$20 +/- \$14.6 million is bad at either extreme
- Figure 3 illustrates this point: the overlap between predicted and actual revenue is substantial, and there are very few qualitative misses. For example, movie 310 is estimated to make ~\$550 million instead of \$1.2 billion; this is a big difference for bookkeeping, but qualitatively the model correctly predicts that the movie will be a hit.

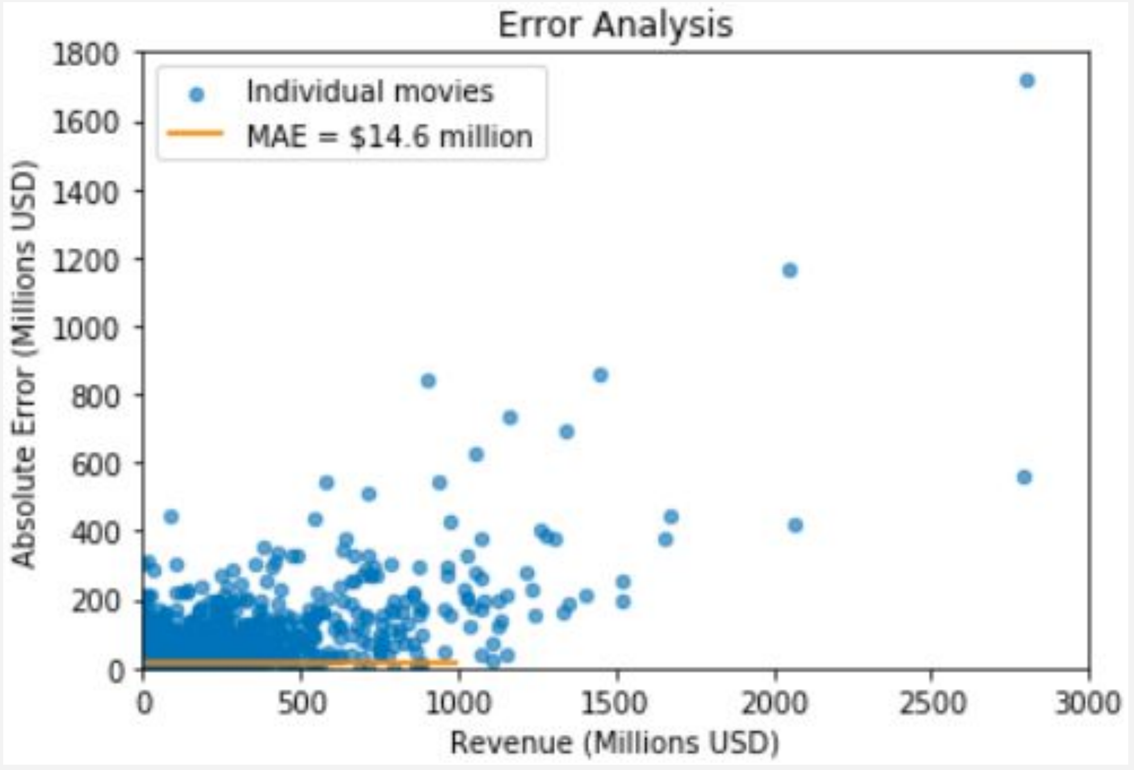

Figure 2: Comparing the model's prediction absolute error to the movie's actual revenue for each movie.

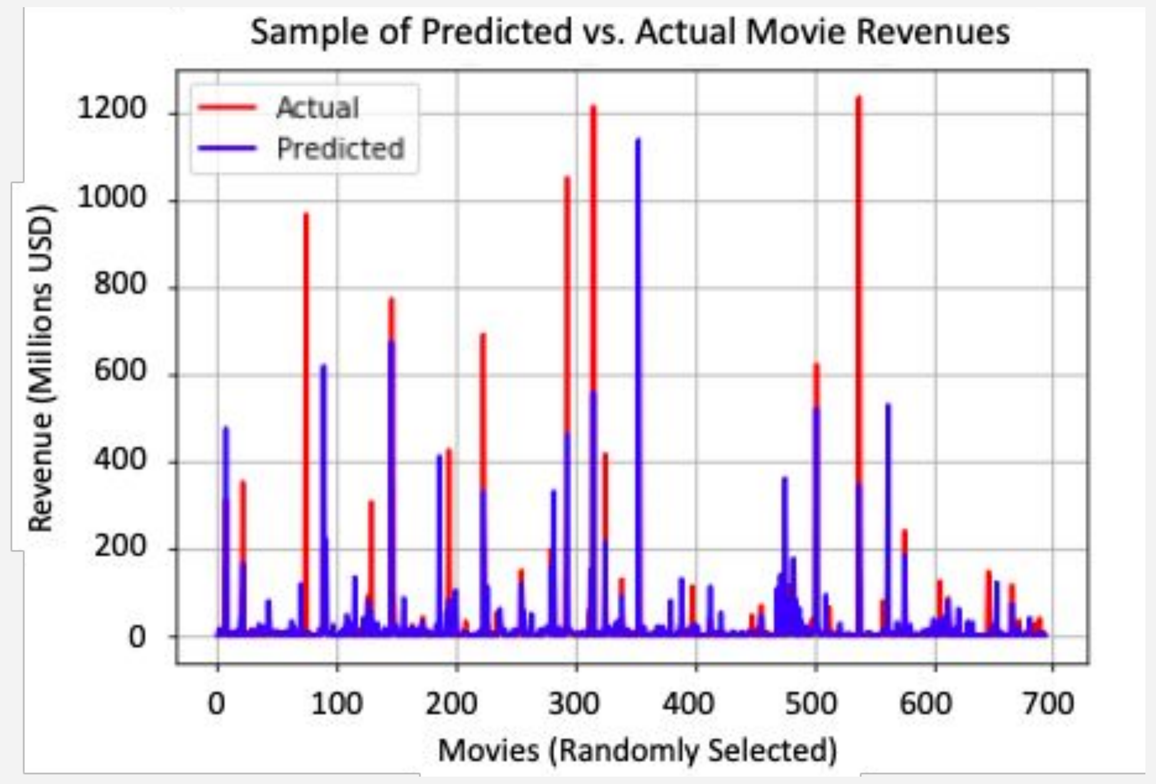

Figure 3: Comparing predicted and actual movie revenues for randomly selected movies in our dataset.

Discussion and Next Steps

- Conclusions
  - Although the model cannot pinpoint exact revenues, the model captures a lot of correlation and can reasonably predict whether a movie will be a box office hit or flop
  - Genres are important features in the model.
    - In particular, the *adventure, sci-fi, action, animation and comedy* genres correlated much better with revenue than the other features
    - As seen in Figure 1, categorical variable inclusion had a much greater effect on model performance than the collective contribution from numerical variables
- Future directions
  - Use a Neural Network to try to model the evident non-linear correlations better
  - Find a more methodical way to incorporate actor information, and collect SA on their tweets. There are so many actors in the data that incorporating them presents challenges.
  - Collect information on budget and incorporate into the model
  - Explicitly reformulate the model into a classifier of low, mid, and high revenue movies