Identifying Features of Profitable Sci-Fi Films

Juliana McCausland

The goal of this project was to investigate the features that contribute to higher revenues for science fiction films. I scraped data from various websites, ending up with information about the revenues, actors, directors, runtime, MPAA ratings, release dates, and plot keywords for 1700 films. I originally wanted to see if there was a relationship between the number of 'tropes' (which I extracted from plot keywords in IMDb) in the film and revenue. I was hoping to see more defined linear relationships between the features and the target (revenue), but I was still able to identify some of the important features using random forest regression. I started with pretty low R squared scores, and after some engineering, cleaning, and scaling, ended up with a score up to 0.76. Overall, my data indicated that the actors and directors are the most important features when considering higher revenues for sci-fi films.

Design

Recent research has shown that science fiction can be both healing post-catastrophe and helpful as a form of mental preparedness pre-catastrophe. The backstory is that film production companies would like to explore this idea, given the current state of the world, while bringing in maximum revenue. These companies would like to know what features in particular make for the most profitable films, and they hypothesized that the number of tropes related to the topics of healing and preparedness might be significant. This analysis serves to explore this hypothesis and identify other important features of science fiction films that increase revenue.

Data

I scraped most of my data from IMDb and Box Office Mojo. I also scraped data for the highest-grossing actors and highest-grossing directors from TheNumbers. I started with 1700 movies, but that was cut down substantially through the cleaning process. Much of my data was categorical, which I converted to dummies manually. I avoided relying on data like IMDb ratings and opening weekend revenues because these would not be reflective of the features inherent to the films that lead to success.

Algorithms

I used the methods learned for scraping (with BeautifulSoup), testing, validation, and cross validation. I performed an initial linear regression after identifying features that seemed most relevant (by looking at correlations). I performed a simple validation/test and then scaled to adjust for the coefficients. I then created interaction variables by multiplying the runtime and the distinct-value/binary features. I re-trained (at which point my R squared was about 0.58), and saw that my coefficients had all decreased (they were originally very high). I then did random forest regression, and saw my R squared increase to 0.78.

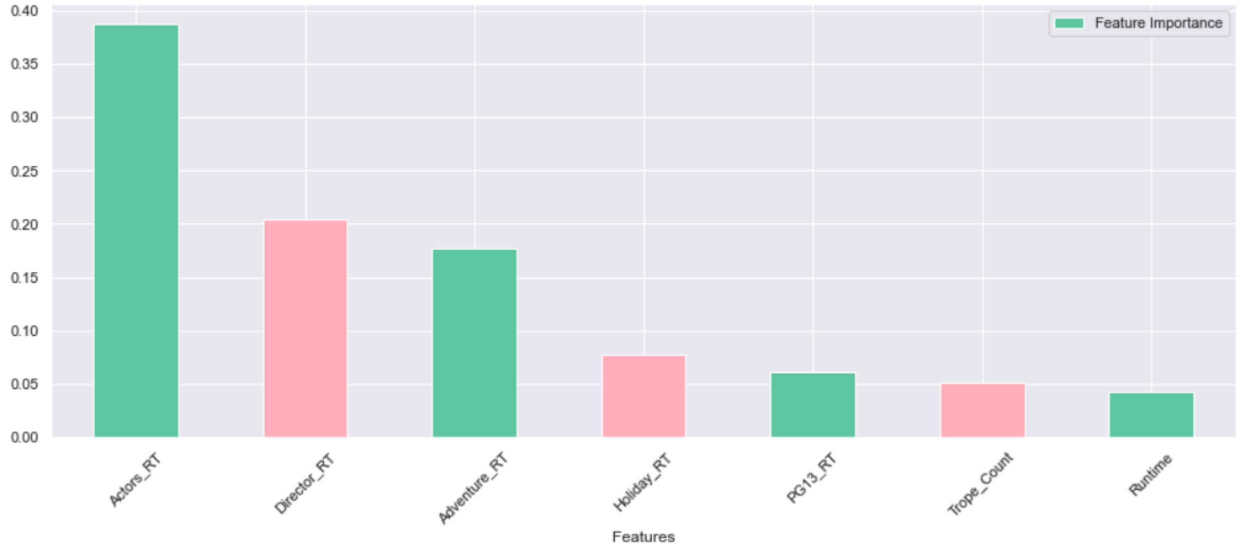
I used random forest's feature_importances_ attribute to identify the most important features of the model. The two most important were actors and directors. Tropes were among the most important features, but they had a very low correlation to begin with and did not seem too significant overall.

I then did cross validation simply for due diligence, using kfolds for ridge regression, lasso regression, and standard linear regression. Ridge seemed to be slightly better than the others.

Tools

- Jupyter Notebooks, Pandas, NumPy, Sklearn, Seaborn, Matplotlib, BeautifulSoup, Random Forest

Communication


🟩 Feature Importance
Actors_RT
Director_RT
Adventure_RT
Holiday_RT
PG13_RT
Trope_Count
Runtime
Features