INVISIBLE VARIABLES

Ridley Arp 11-9-21

INTRODUCTION

Data Used:

- Top 1000 Highest Grossing Movies of All Time List (Domestic) - BoxOfficMojo.com
- Actor information and traits - IMDB.com

Tools Utilized:

- Python BeautifulSoup and Requests to scrape
- Python Numpy and Pandas for data manipulation
- Python Scikit-learn and Statsmodels for modeling
- Python Seaborn and Matplotlib for visualizations

METHODOLOGY

- **Data Collection and Cleaning**
  - **Raw data was scraped via a Requests and BeautifulSoup pipeline and stored in a local CSV file accessible via Pandas**
- **Feature Engineering**
  - **Utilized available data to synthesize additional datapoints desirable in our analysis**
- **Modeling and Selection**
  - **Modeled regressions on a 60/20/20 split, narrowing traits to three highest-impact**

RESULTS

HIGHEST PERFORMING MODEL

	coef	std err	t	P>|t|
const	1.605e+08	2.29e+07	7.006	0.000
x_1	-9.59e+06	7.63e+06	-1.256	0.209
age_at_release_1	-6.945e+07	3.72e+07	-1.867	0.062
frequency_at_release_1	-3.704e+07	2.04e+07	-1.812	0.070
x_2	9.216e+05	6.16e+06	0.150	0.881
age_at_release_2	-3.55e+06	3.22e+07	-0.110	0.912
frequency_at_release_2	5.819e+07	1.9e+07	3.064	0.002
x_factor	-4.334e+06	4.37e+06	-0.991	0.322
value_at_release_1	1.507e+07	3.87e+07	0.390	0.697
value_at_release_2	5.454e+07	3.51e+07	1.555	0.120

R-squared:	0.023
Adj. R-squared:	0.015

Even with many features to check against, these elements combined can only account for a small overall effect in the BO results.

However, many of the most likely contenders for worth such as value at release were some of the least significant contributors.

HIGHEST EFFECT FEATURES

	coef	std err	t	P>|t|
const	1.806e+08	1.16e+07	15.504	0.000
age_at_release_1	-4.757e+07	2.42e+07	-1.968	0.049
frequency_at_release_1	-2.875e+07	1.99e+07	-1.443	0.149
frequency_at_release_2	6.12e+07	1.89e+07	3.230	0.001

R-squared:	0.015
Adj. R-squared:	0.012

Prior experience of the primary supporting actor still holds the highest statistical significance even with significantly reduced features with overall accuracy relatively unchanged.

INTERPRETATION

- Main Features of Interest
  - Given the retained accuracy of the three feature model, our three highest contributing factors include the prior experience of the main and supporting actor, as well as the age of the primary actor
    - For each film the secondary actor has been in prior there is an average increase of $61,200,000 +- $18,900,000 to the expected BO outcome
    - For each film the primary actor has been in prior there is an average DECREASE of $28,750,000 +- $19,900,000 to the expected BO outcome
    - For each year older the primary actor is at time of release there is an average DECREASE of $47,570,000 +- $24,200,000

ACTIONABLES

- Given the prior information, notable trends:
  - The highest grossing films consistently star younger leads
  - These leads succeed more frequently when backed with an experienced main supporting character
  - The most experienced lead actors do not guarantee better results, as none of the top box office contenders had highly seasoned leads.

FURTHER ACTION AND NOTEABLES

- **Of the included films:**
  - **48% were cast M/F**
  - **47% were cast M/M**
  - **5% were cast F/F**
- **Of the 100 oldest films on the list:**
  - **73% were cast M/F**
  - **27% were cast M/M**
  - **0% were cast F/F**

Not enough data to thoroughly explore effects of gender on films, but there seems to be promising indications that a more diverse cast tends to have a longer lasting appeal.