



Understanding the linkage between Gender Diversity and Hollywood Box Office Outcomes

Mallory Banks, METIS: Regression Analysis
Project 2

Introduction: A Study of Gender Diversity in Hollywood

Problem Statement and Objectives:

- The question around this analysis is understanding the link between diversity of movie casts and production to the overall financial outcome for the movie.
- Hypothetically, the client could be the movie industry in at least the United States. In my mind, this analysis is commissioned by “The Academy” from my boutique analytics firm as a reactionary action to the “Oscars So White” controversy of 2015. This analysis would have been done in 2020 , but with the Covid 19 Pandemic, it was pushed to 2021 and will help to illuminate trends around diversity (in front of and behind) and how it impacts movie success as well as other important features and factors of film. Because of the lack of Racial and Ethnic Data cataloged in the film industry, we’d begin with Gender Diversity with intent to build further





Methodology Part 1

Data:

- There were two major sources of data for this project, one obtained through the webscraping of boxofficemojo.com for the top 200 highest grossing movies of 11 years and the other through research into Gender Diversity for Actors, Directors, Producers, and writers for films in the US and Canada for the same time period. For the purposes of this project, this dataset was subsetting on only for American made films for the specified time period of **2009-2018**.

Tools:

- Tools used included sqlite3, jupyter notebooks running Python and using packages beautifulsoup, scikitlearn, pandas, matplotlib, and seaborn.

Metrics:

- The Target metric for this Analysis was Domestic Gross (In Millions). Additionally because of the limitations of the dataset, Non Binary and Gender Non Conforming People are not coded here



Methodology

Data Collection: Data from Top 200 Highest Grossing Movies Scraped from boxofficemojo from 2009 to 2018. This was then joined to a dataset I sourced from another GitHub user documenting actor, producer, writer, and director diversity for these films. Data was then wrangled and cleaned.

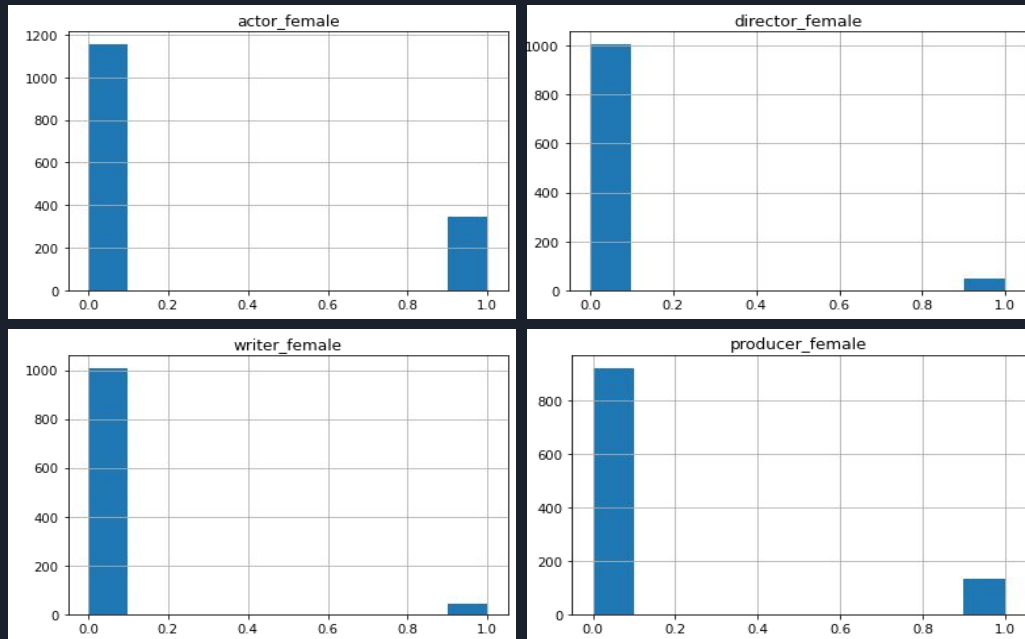
Initial Regression: Through OLS (ordinary least squares) regression identify the fundamental drivers (Budget, Release Size, Opening, Producer Gender etc.) of box office outcomes.

Refined Regression: Utilize Elastic Net Regression Models to isolate drivers of box office outcome and remove “noise” variables to create a stronger, more reliable model.

Pinpoint Areas of Focus: the Goal here is not necessarily to predict Box Office Out Comes, but to Understand which, if any, variables relating to gender are influential in Box Office Outcomes for Hollywood Films

The Distribution of Gender Diversity in Films Released from 2009-2018 is limited in front of the Camera and Behind

When Looking at the Top-Billed Lead Actors, Lead Producers, Main Writers, and Directors We See a High Skew for Male Gender Identity in All of These Roles



Available Features for Prediction for Domestic Gross

(* connotes initial influence on Predicting box office success)

| Variable | Description |
|------------------|---|
| Theater_Num* | Number of Theaters Movies Was Released to |
| Inter_Gr* | International Gross |
| Length* | The Length of Time the Movie Remained in Theaters |
| Open_Num* | The Opening Gross of the Film |
| Budget* | The Budget of the Film |
| runtime* | The length of Film Time |
| director_female | Binary Dummy Variable based on Director Gender Info |
| actor_female* | Binary Dummy Variable based on Actor Gender Info |
| producer_female* | Binary Dummy Variable based on Producer Gender Info |
| writer_female | Binary Dummy Variable based on Writer Gender Info |

VIF From First OLS - Further Variable Isolation

Problematic

| | feature | VIF |
|---|-----------------|-----------|
| 0 | Theater_Num | 13.854880 |
| 1 | Inter_Gr | 4.643573 |
| 2 | Length | 5.760708 |
| 3 | Open_Num | 5.099832 |
| 4 | Budget | 5.646426 |
| 5 | runtime | 12.601932 |
| 6 | director_female | 1.177477 |
| 7 | writer_female | 1.184299 |
| 8 | producer_female | 1.178359 |
| 9 | actor_female | 1.388489 |

Much Better after dropping
Several Interacting Variables -
Minimal Regression Problem

| | feature | VIF |
|---|-----------------|----------|
| 0 | Length | 2.206347 |
| 1 | Budget | 3.785082 |
| 2 | actor_female | 1.262689 |
| 3 | producer_female | 1.162903 |
| 4 | Open_Num | 3.162333 |

OLS Regression 2

Dep. Variable: Domestic_Gr **R-squared:** 0.815

Model: OLS **Adj. R-squared:** 0.814

Method: Least Squares **F-statistic:** 922.9

Date: Wed, 23 Feb 2022 **Prob (F-statistic):** 0.00

Time: 16:04:29 **Log-Likelihood:** -5303.4


No. Observations: 1053 **AIC:** 1.062e+04

Df Residuals: 1047 **BIC:** 1.065e+04

Df Model: 5

Covariance Type: nonrobust

| | coef | std err | t | P> t | [0.025 | 0.975] |
|-----------------|----------|-------------------|------------|-------|---------|--------|
| Intercept | -13.9426 | 3.189 | -4.372 | 0.000 | -20.200 | -7.685 |
| Length | 0.5041 | 0.067 | 7.478 | 0.000 | 0.372 | 0.636 |
| Open_Num | 0.0003 | 5.65e-06 | 45.327 | 0.000 | 0.000 | 0.000 |
| Budget | 0.1673 | 0.029 | 5.726 | 0.000 | 0.110 | 0.225 |
| actor_female | 5.9338 | 2.853 | 2.080 | 0.038 | 0.336 | 11.532 |
| producer_female | 7.4180 | 3.509 | 2.114 | 0.035 | 0.533 | 14.303 |
| Omnibus: | 1118.464 | Durbin-Watson: | 1.809 | | | |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 105854.328 | | | |
| Skew: | 4.925 | Prob(JB): | 0.00 | | | |
| Kurtosis: | 51.121 | Cond. No. | 1.12e+06 | | | |



Simple Cross Validation for Linear Regression With Target Variables

Initial Linear Regression Model severely overfit using r^2 metric

R^2 Train ->

0.8362516820751223

R^2 Test ->

0.759985840473083

After Removing “actor_female” as interacting Variable and Re-Testing
Other Gross Metrics New Linear Model Emerged with “Inter_Gr” Added

R^2 Train ->

0.8624789434939415

R^2 Test ->

0.8718353997907384



Conclusions

Conclusions

- When Looking at Impactful Predictive Levers for Box Office Success (Domestically) for Movies released 2009-2018, we were able to see movies where women were the Top Billed Producer contributing significantly to the success of the Film from a Domestic Box Office Perspective
- However, the overall presence of Lead Actresses and Executive producers is woefully small. Smaller still are those women leading Director and Writer's Rooms.

Recommendations

- Creation of Task force to understand how to increase Women in Leading Roles in front of and behind the camera. This is not just a "Good Will" experiment but also one that has the potential to influence the Film Industry Bottom Line. With folks slower to get to the theater and the popularity of streaming, focusing on Diversity in production can be a way to encourage movie going experiences
- Additionally, data rigor around accounting for and recording this data should be a priority as we continue to focus on multiple dimensions of diversity



Future Work

- This work would be further fleshed out and improved with the inclusion of Racial and Ethnic Diversity as well as LGBTQ+ representation.
- This data is not well documented or available to the public, once data efficacy issues are address the more robust model may be created to understand the effect of multiple dimensions of diversity on Box Office Success (Ex: What would it look like to re-view drivers as a function of Intersectional Identity?)
- Is there any data on movie goer gender?
 - How might we understand Genre as a function of Movie Goer attendance? Are Female Movie goers more or less likely to see a Sci/Fi or Action film as other gender identities for example?

Heteroskedasticity Problems - Plotted Residuals not Eased by log, square transformations

