Interpreting the impact of lead actor gender and age on movie gross

Hannah Lewis

Metis, Linear Regression & Web Scraping Module

August 6, 2021

Overview

Motivation

- Female protagonists—particularly, female protagonists portrayed by actors over the age of ~35—are significantly underrepresented in movies

Objectives

- Determine what (if any) impact these differences in actor demographics (i.e., gender and age) have on the lifetime gross of a movie

Results

- Lifetime movie gross is strongly dependent on how it performs during its opening weekend, the budget, its reception by audiences, and its genre

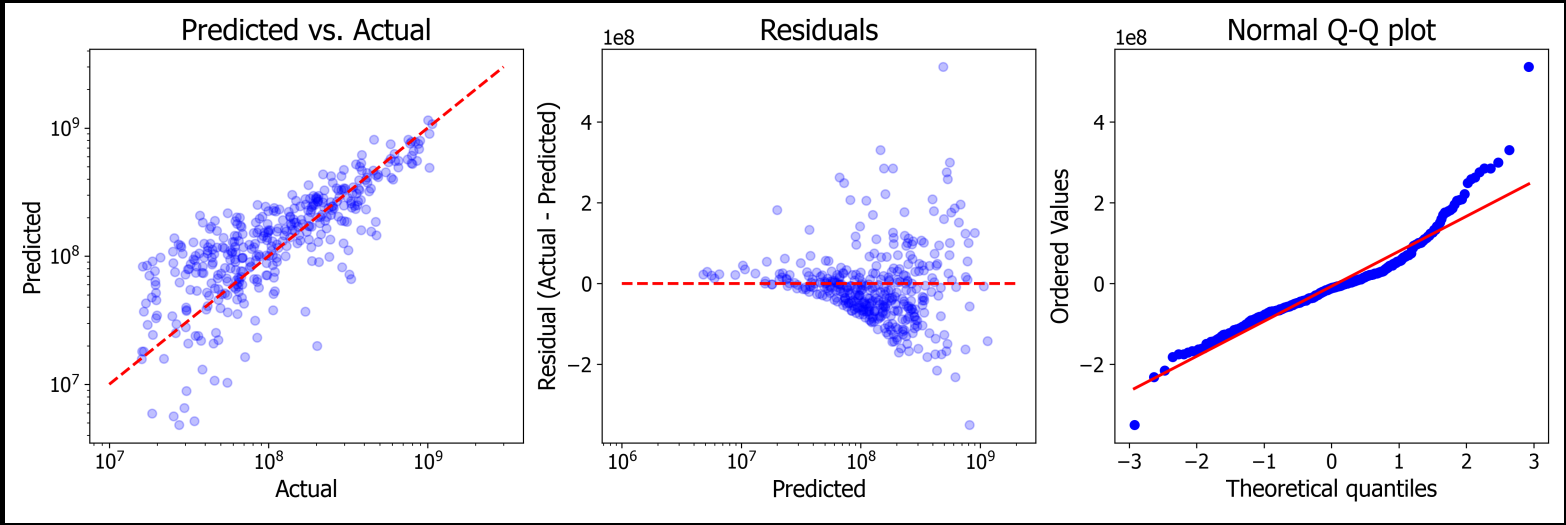
Data & Methods

Target: worldwide lifetime gross
Features: **domestic opening gross**, **budget**, **studio**, MPAA rating, run time, **genre**, lead name, **lead age**, lead height, **lead gender**, **Rotten Tomatoes audience** and **Tomatometer scores**, international movie?

- Baseline models trained on domestic opening gross (and its square), budget, and Rotten Tomatoes scores as features
- Select ridge regression model (validation $R^2 \sim 0.720$, MAE ~ \$68.4M)

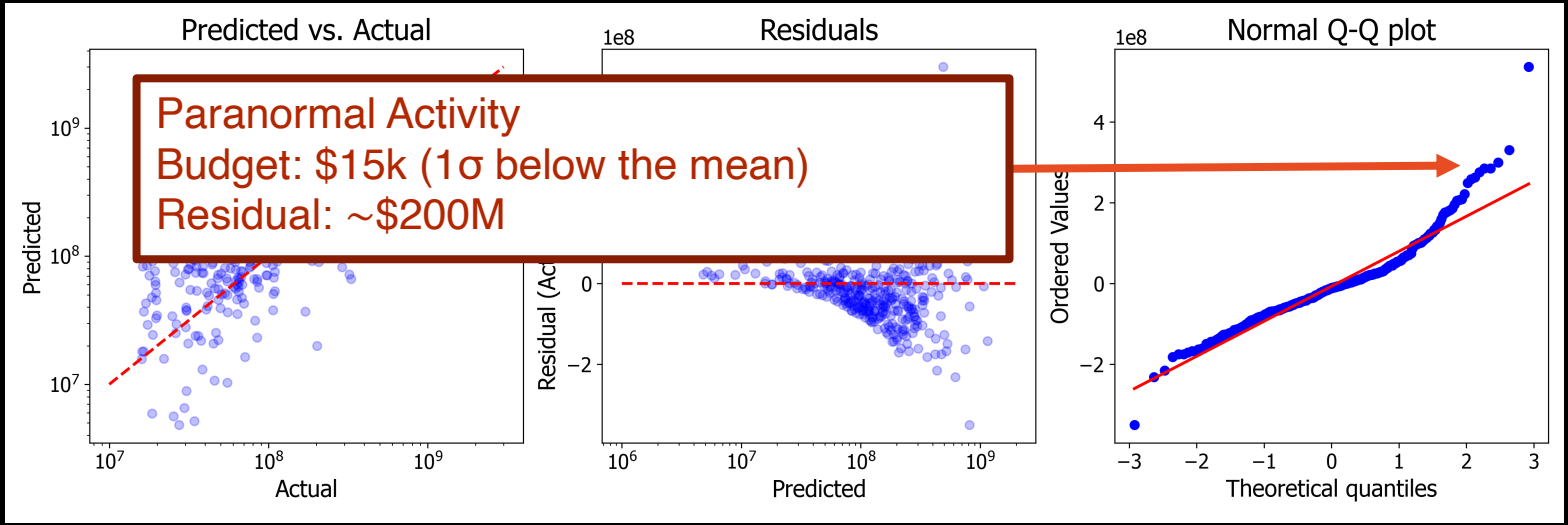
- Expanded model includes lead actor gender and age as features, as well as dummy variables for the studio that produced the movie and the genre of the movie

Results: model performance



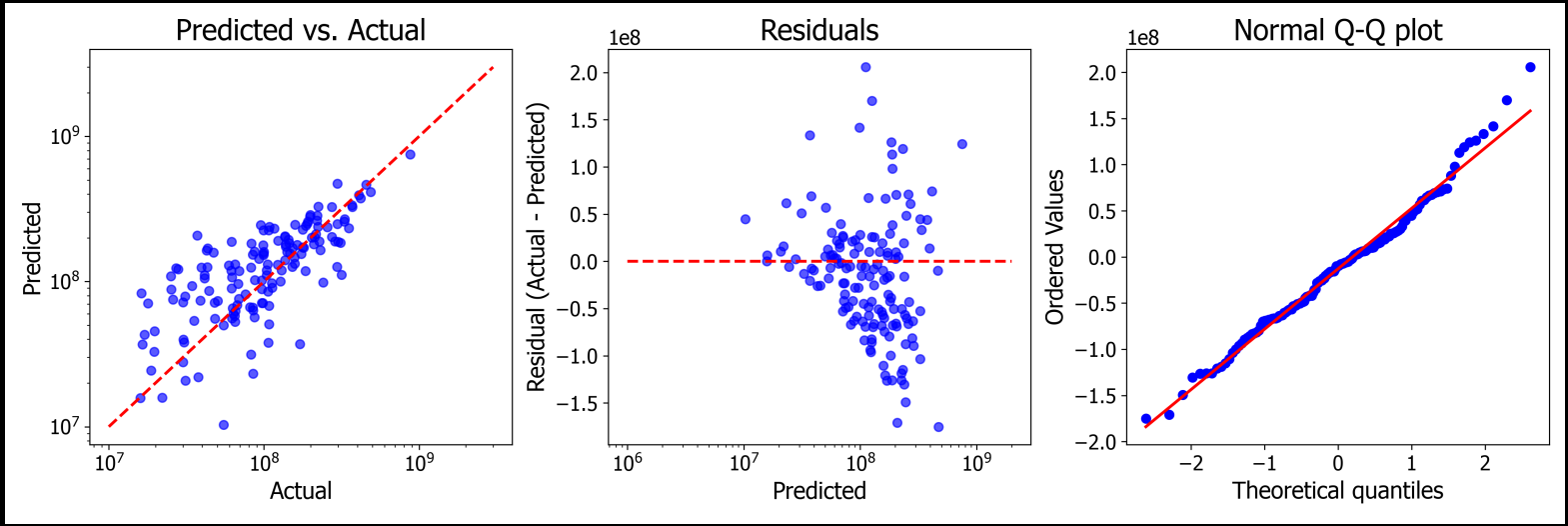
- Model does not generalize well for outliers in budget
- For any budget ($15k up to 300M), MAE = $63.4M

Results: model performance



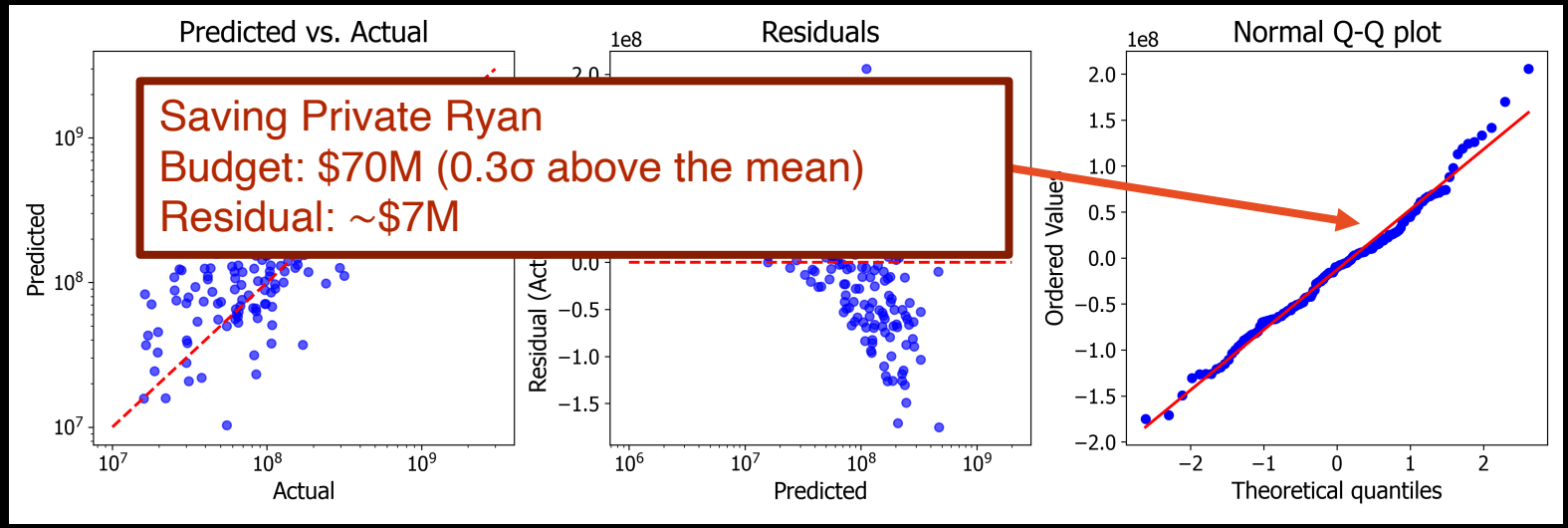
- Model does not generalize well for outliers in budget
- For any budget ($15k up to 300M), MAE = $63.4M

Results: model performance



- Model works best for movies with budgets within ±0.5 standard deviations of the mean ($30M up to $80M), MAE = $51.0M

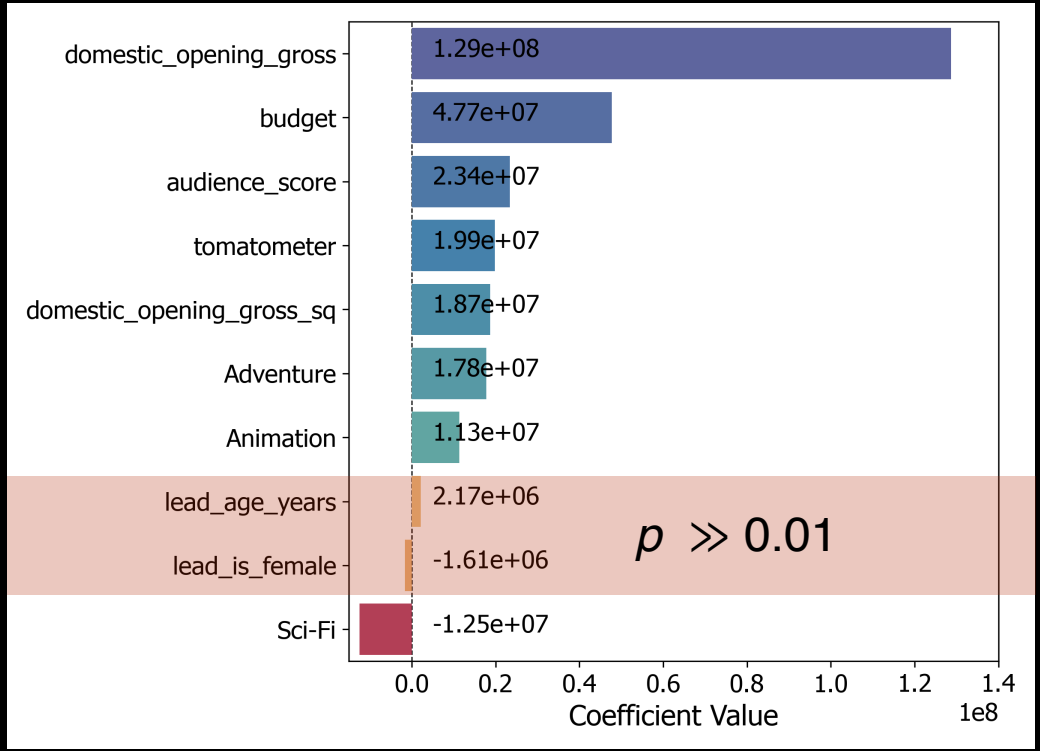
Results: model performance



- Model works best for movies with budgets within ±0.5 standard deviations of the mean ($30M up to $80M), MAE = $51.0M

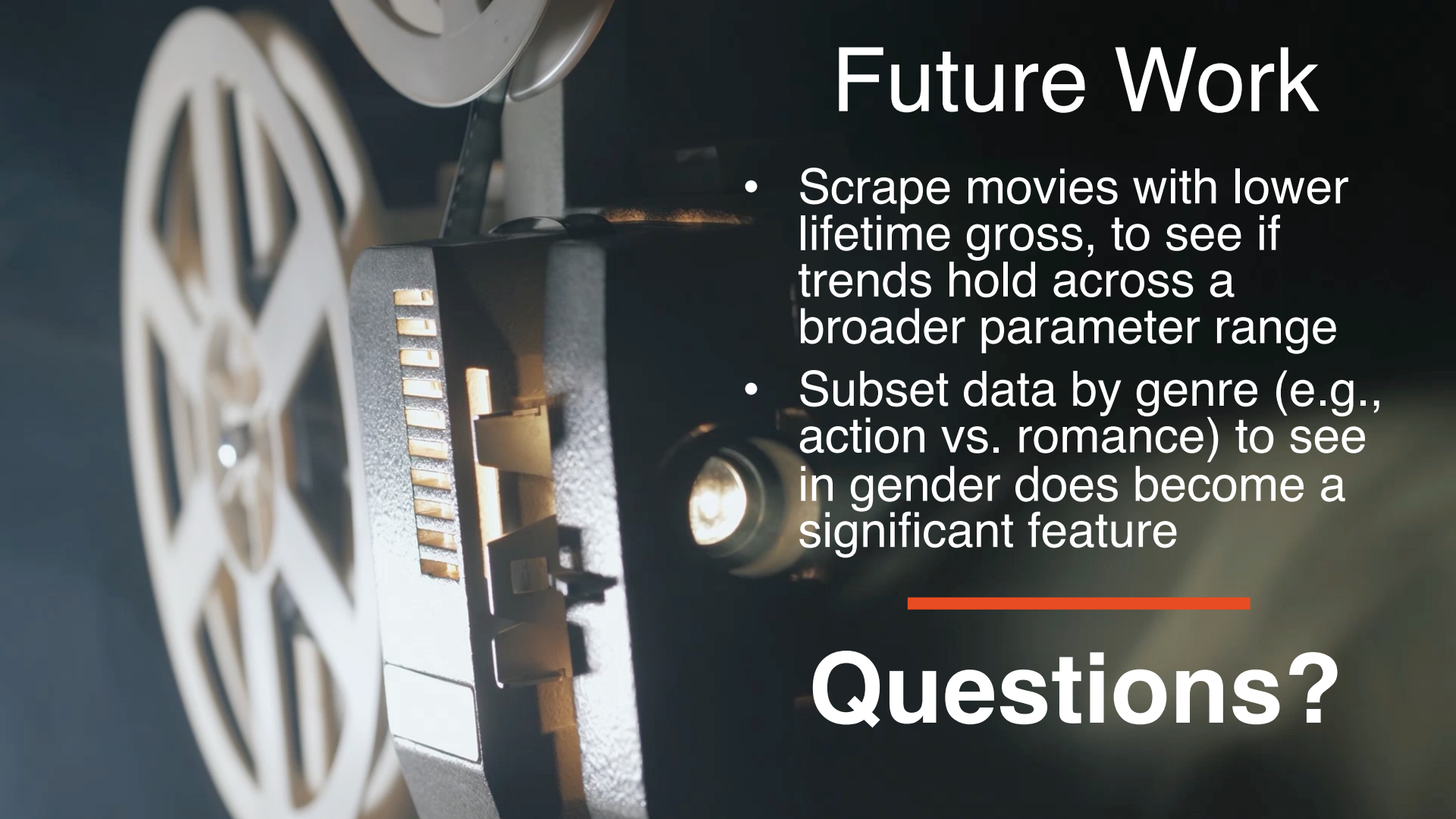
Results: significant features

- Significant coefficients
  - $p < 0.01$
  - Reject the null hypothesis that $\beta_i = 0$
- In the final model, lead actor age and gender **do not** have significant coefficients



Conclusions & Recommendations

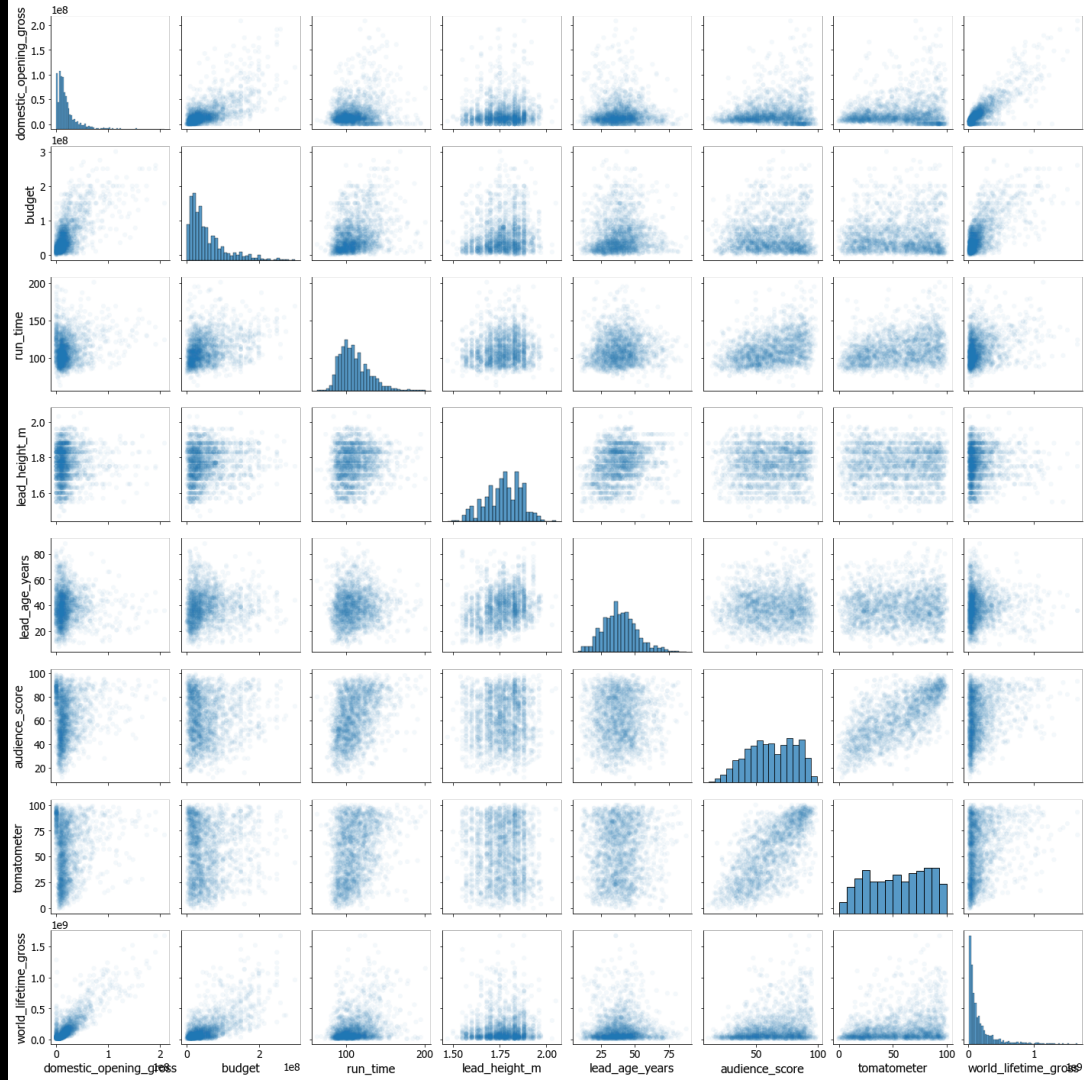
- Lead actor gender and age are not among the most important features that determine movie gross
- Studios should produce a larger fraction of movies that are led by a more gender and age-diverse range of actors

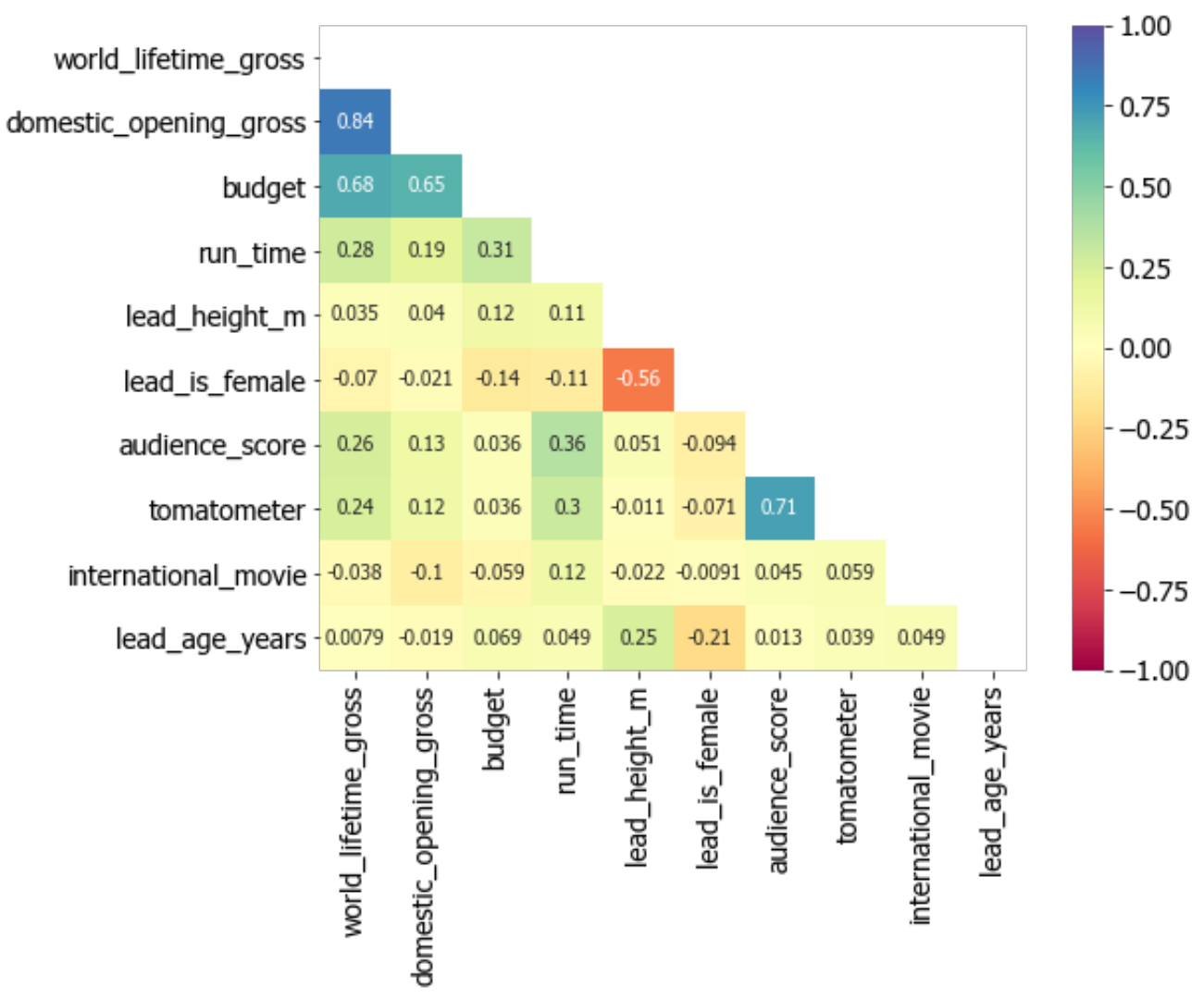


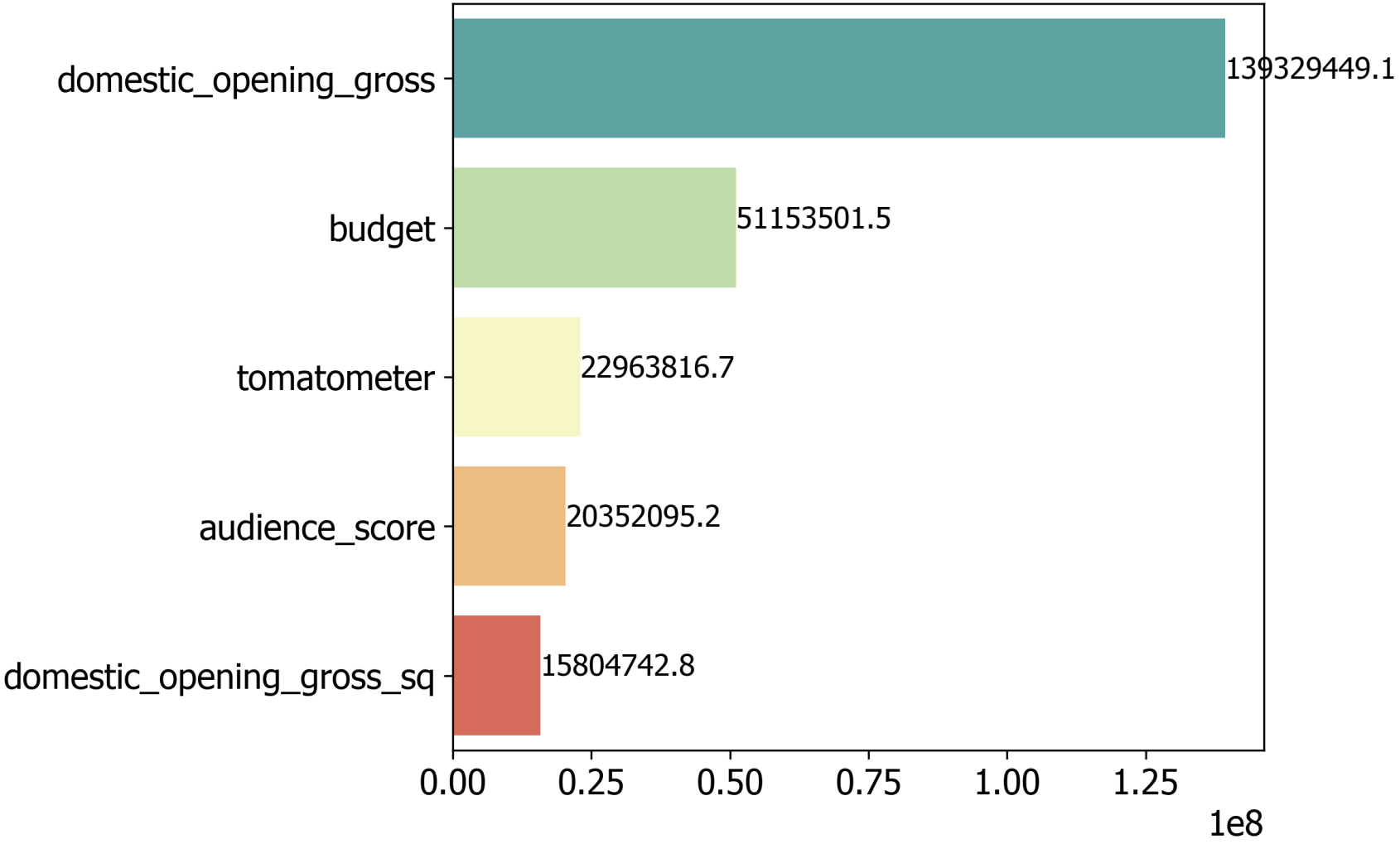
Future Work

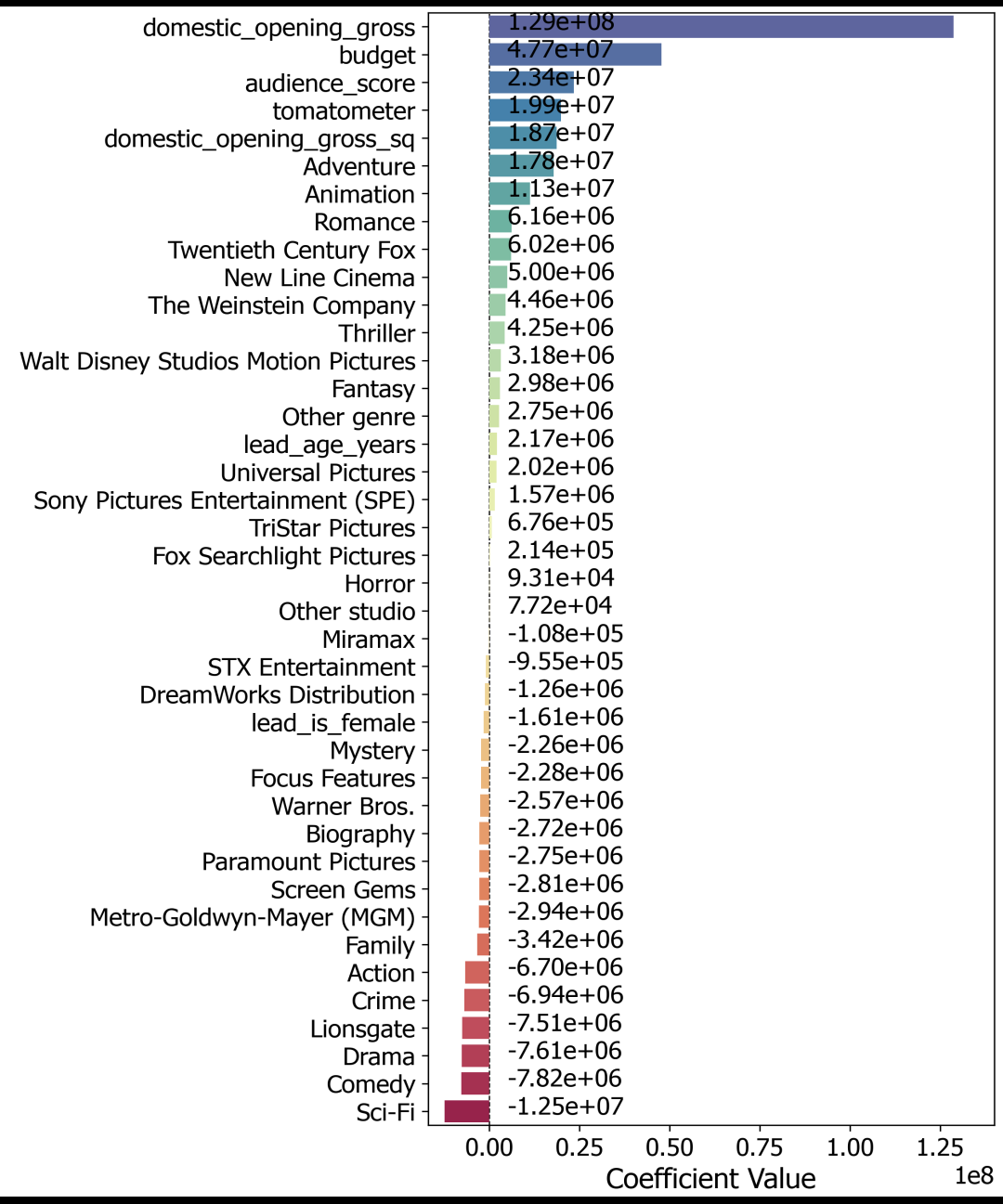
- Scrape movies with lower lifetime gross, to see if trends hold across a broader parameter range
- Subset data by genre (e.g., action vs. romance) to see in gender does become a significant feature

Questions?


domestic_opening_gross
budget
run_time
lead_height_m
lead_age_years
audience_score
tomatometer
world_lifetime_gross
1e8
1e9


world_lifetime_gross
domestic_opening_gross
budget
run_time
lead_height_m
lead_is_female
audience_score
tomatometer
international_movie
lead_age_years
0.84
0.68 0.65
0.28 0.19 0.31
0.035 0.04 0.12 0.11
-0.07 -0.021 -0.14 -0.11 -0.56
0.26 0.13 0.036 0.36 0.051 -0.094
0.24 0.12 0.036 0.3 -0.011 -0.071 0.71
-0.038 -0.1 -0.059 0.12 -0.022 -0.0091 0.045 0.059
0.0079 -0.019 0.069 0.049 0.25 -0.21 0.013 0.039 0.049
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


domestic_opening_gross
budget
tomatometer
audience_score
domestic_opening_gross_sq
139329449.1
51153501.5
22963816.7
20352095.2
15804742.8
0.00
0.25
0.50
0.75
1.00
1.25
1e8


domestic_opening_gross
budget
audience_score
tomatometer
domestic_opening_gross_sq
Adventure
Animation
Romance
Twentieth Century Fox
New Line Cinema
The Weinstein Company
Thriller
Walt Disney Studios Motion Pictures
Fantasy
Other genre
lead_age_years
Universal Pictures
Sony Pictures Entertainment (SPE)
TriStar Pictures
Fox Searchlight Pictures
Horror
Other studio
Miramax
STX Entertainment
DreamWorks Distribution
lead_is_female
Mystery
Focus Features
Warner Bros.
Biography
Paramount Pictures
Screen Gems
Metro-Goldwyn-Mayer (MGM)
Family
Action
Crime
Lionsgate
Drama
Comedy
Sci-Fi
1.29e+08
4.77e+07
2.34e+07
1.99e+07
1.87e+07
1.78e+07
1.13e+07
6.16e+06
6.02e+06
5.00e+06
4.46e+06
4.25e+06
3.18e+06
2.98e+06
2.75e+06
2.17e+06
2.02e+06
1.57e+06
6.76e+05
2.14e+05
9.31e+04
7.72e+04
-1.08e+05
-9.55e+05
-1.26e+06
-1.61e+06
-2.26e+06
-2.28e+06
-2.57e+06
-2.72e+06
-2.75e+06
-2.81e+06
-2.94e+06
-3.42e+06
-6.70e+06
-6.94e+06
-7.51e+06
-7.61e+06
-7.82e+06
-1.25e+07
0.00
0.25
0.50
0.75
1.00
1.25
Coefficient Value
1e8