Casts Gender and Movie Ratings

OCRUG Hackathon - April 11, 2021

Group 3 - Hollywood Hackers
Mark Jackson, Brian Wang, Joy Chen,
Yiqin Chen, Krystal Maughan

Hypothesis

For **movies** with similar **budget, genre,** and **movie length** in the year range of **1931-2018**, movies where the lead actresses are **female** get lower ratings than the ones where the lead actors are **male.**

Sony emails reveal Jennifer Lawrence paid less than male co-stars

Leak shows that American Hustle's female star was on 7% deal, while male co-stars Christian Bale and Bradley Cooper were on 9% each

Guardian, Dec 13, 2014

Data Cleaning & Prepping

We use **gender**, **budget, genre, runtime,** and **movie year** as our **input,** with **rating** as our predicted value.

Years: (1931-2018). Dataset is not balanced. We cleaned missing values. Used a mix of R and Python and API endpoints to obtain more data.

Movies
(vectorized genres using 1 Hot Encoding)

Ratings
movieId
Ratings (mean)

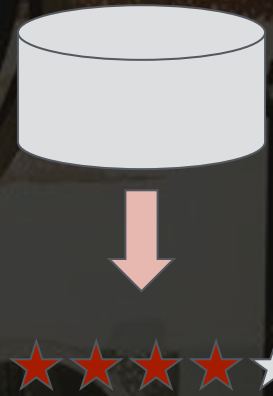
Tags
Id
(long tail)

Cast
1: Female
2: Male
Runtime
Budget

Links
Id
TmId

(265871 x 8)

MovieId,
Release Date,
Budget
Gender
Mean_rating
Runtime
Genre
male/female lead
Year (204 dropped)



Analysis

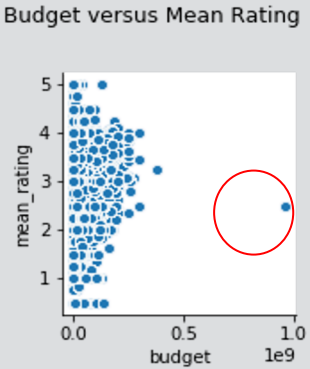


Fig 1: ⅓ of our dataset had a budget of 0. This may or may not be input error. One movie, “Pusher”, had a significantly larger budget.

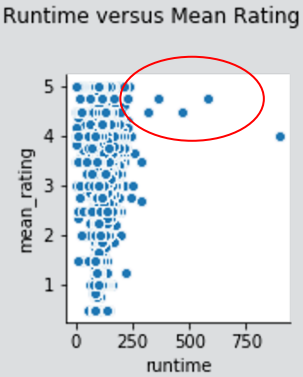


Fig 2: The mean runtime was 110 minutes. There were a few outliers, such as the movie with id #99, a Documentary which is 15 hours long.

Interpretation of Outliers

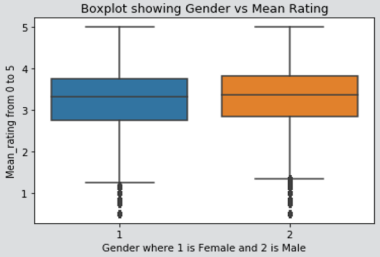
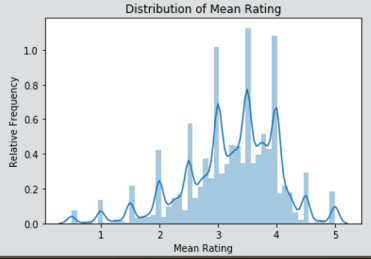


Fig 3: A cursory look at the mean_rating and gender seems to show little difference in the distribution of mean ratings, with male being slightly higher



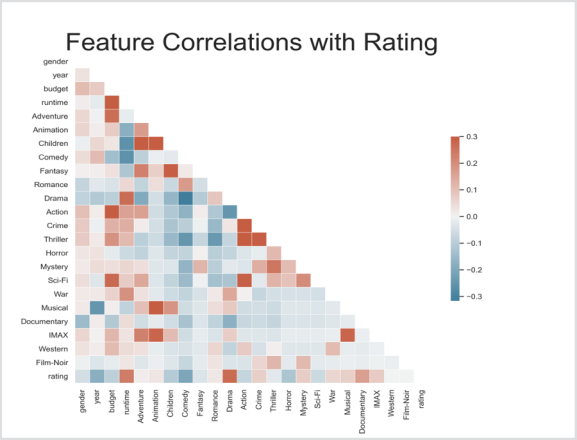


Fig 4 (bottom left): shows that the distribution of Mean Rating is between 3 and 4. Fig 5 (agove) shows the feature correlation with ratings based on gender

Statistically Significant features

Summary

Our Conclusion:
Gender did **not** have a significant impact on ratings when a lead actor is male or female when using XGBoost.

We used XGBoost because it is an Ensemble model (interpretability, feature selective)

MovieLens Small Dataset
- Links.csv
- Ratings.csv
- Movies.csv
- Tags.csv

TMDb Dataset
- Movie_info.csv
- Cast.csv

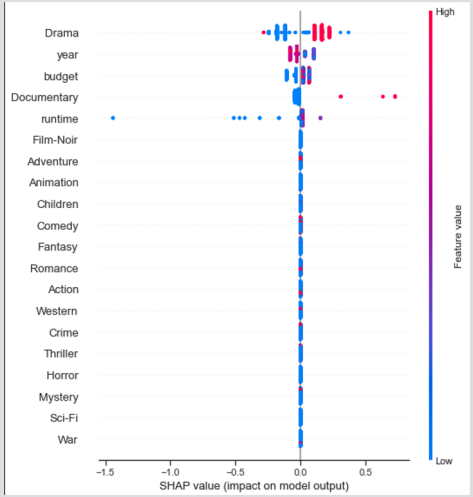


Fig 6 Our SHAP values showed that Drama, year, budget had significant impact on our output for our model

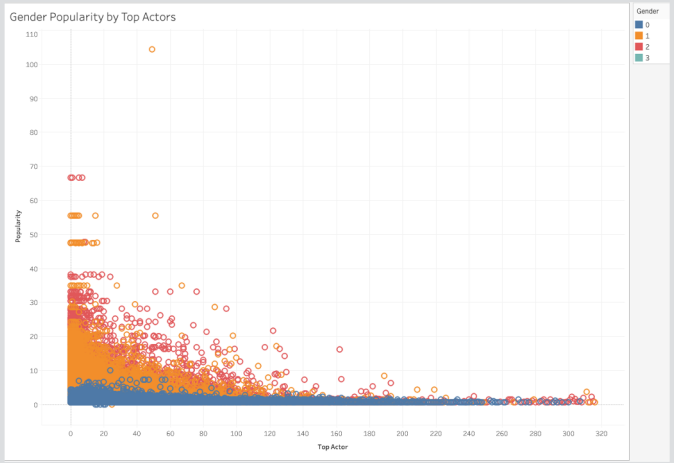


Fig 7: Our conclusion lined up with our data analysis that showed gender popularity by top actors


Thank You