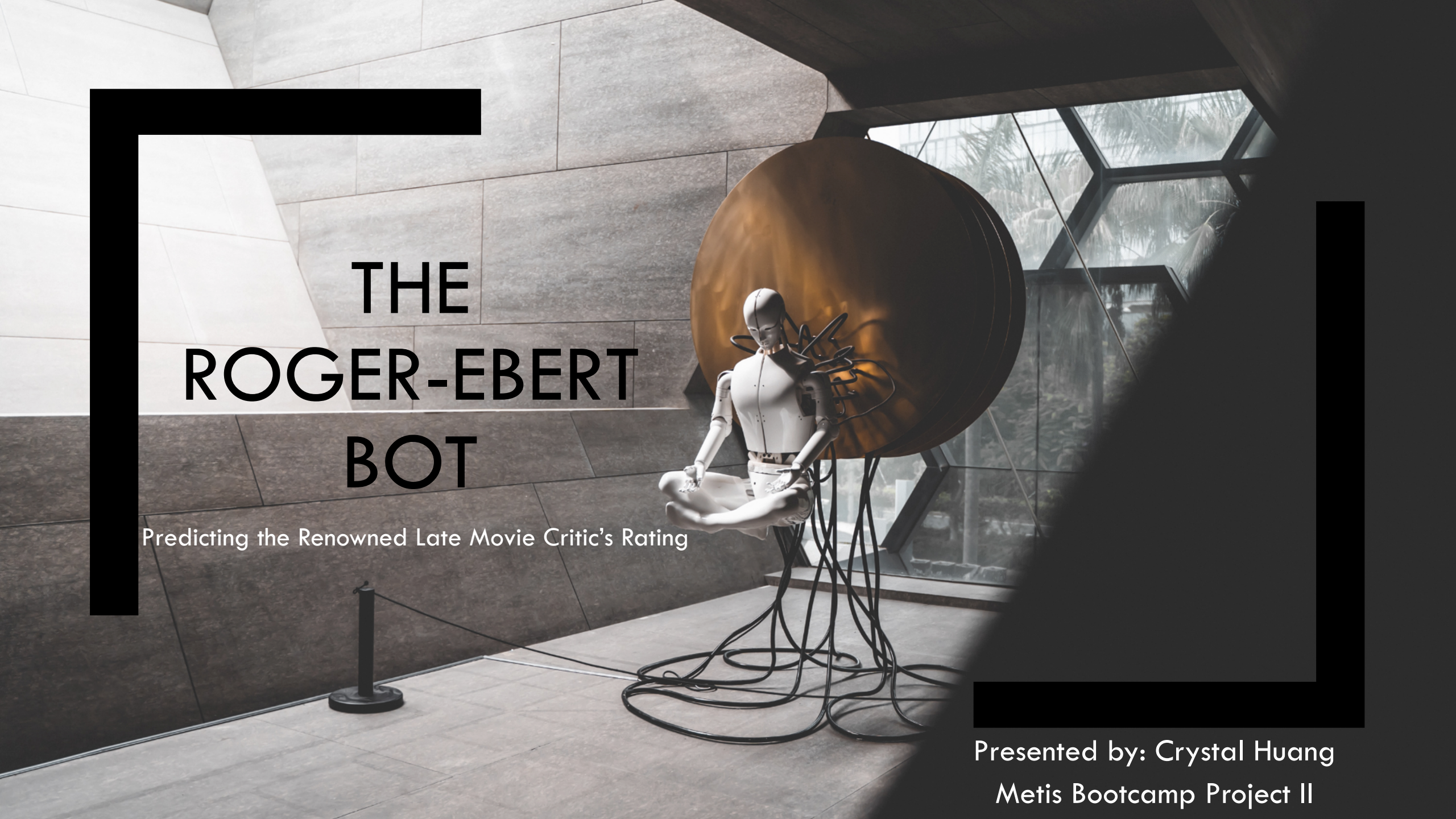


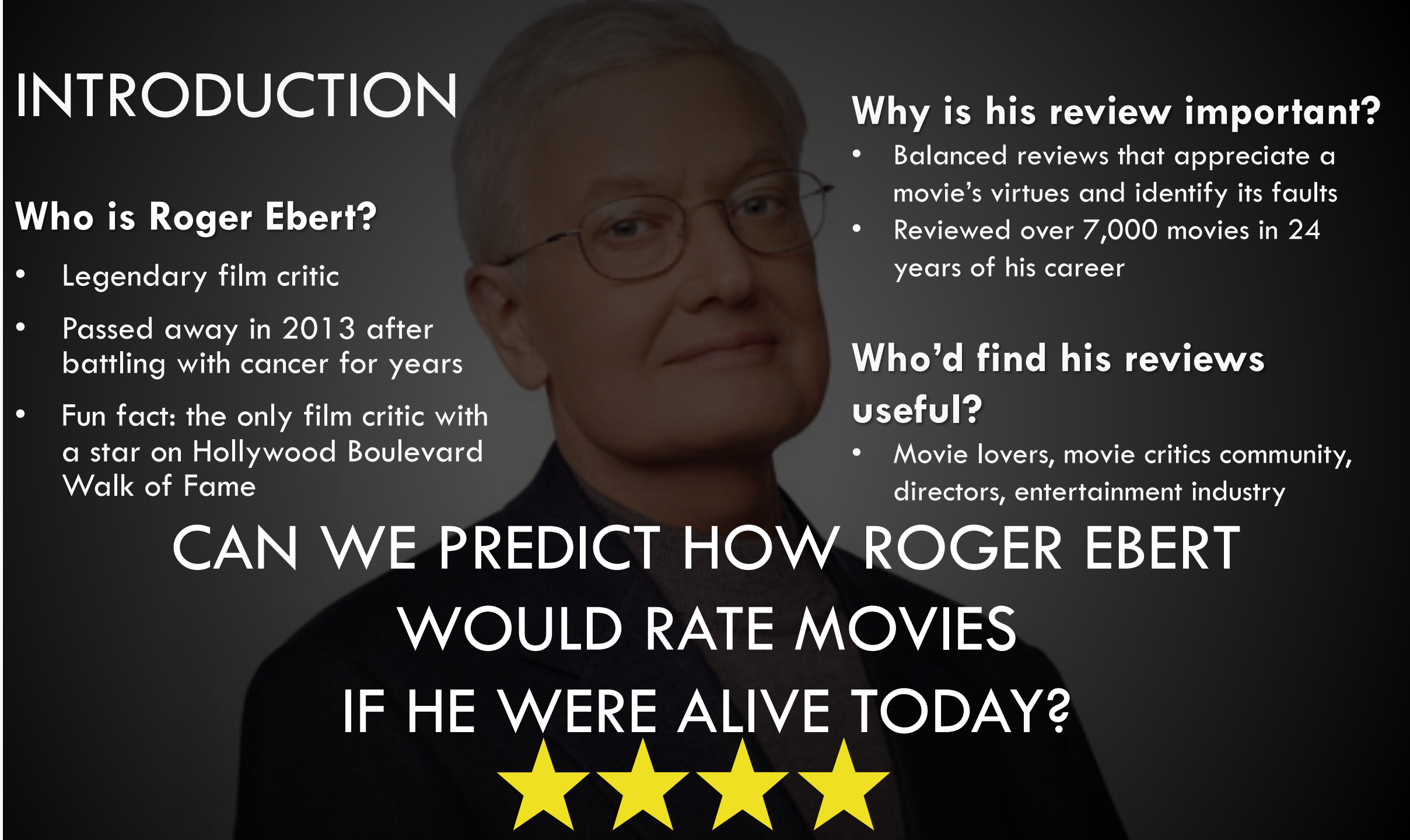
THE ROGER-EBERT BOT

Predicting the Renowned Late Movie Critic's Rating

Presented by: Crystal Huang

Metis Bootcamp Project II

INTRODUCTION



Who is Roger Ebert?

- Legendary film critic
- Passed away in 2013 after battling with cancer for years
- Fun fact: the only film critic with a star on Hollywood Boulevard Walk of Fame

Why is his review important?

- Balanced reviews that appreciate a movie's virtues and identify its faults
- Reviewed over 7,000 movies in 24 years of his career

Who'd find his reviews useful?

- Movie lovers, movie critics community, directors, entertainment industry

CAN WE PREDICT HOW ROGER EBERT WOULD RATE MOVIES IF HE WERE ALIVE TODAY?

★★★★

TOOLS

Language

Web-scraping

Data manipulation

Plot graphs

Modeling & Testing

METHODOLOGY

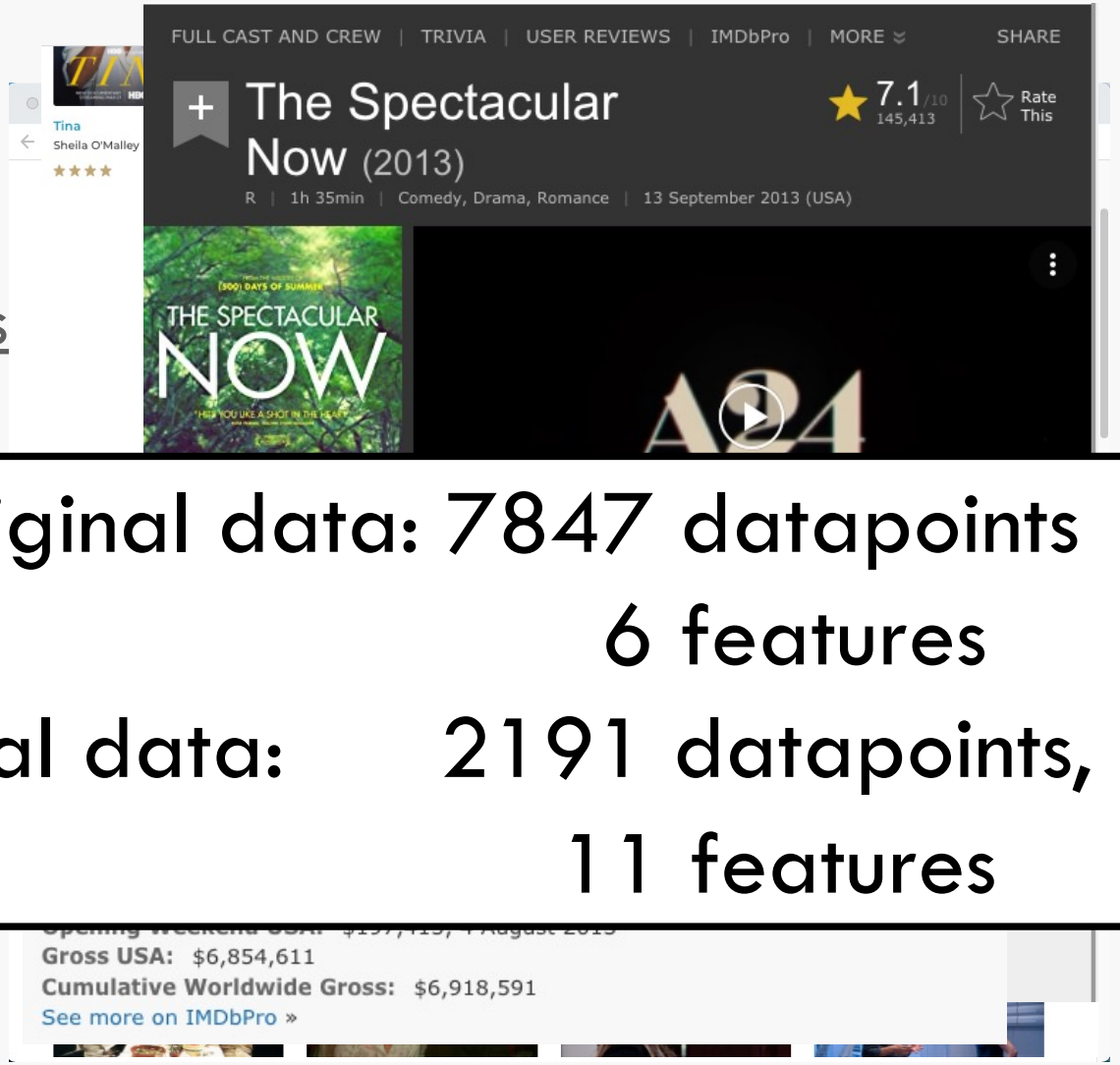
Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)

PRIMARY:

- Roger Ebert Reviews Website: https://www.rogerebert.com/reviews

SECONDARY:

- MovieLens movie dataset from Kaggle (IMDb ID)
- IMDb.com
  – scrap more features

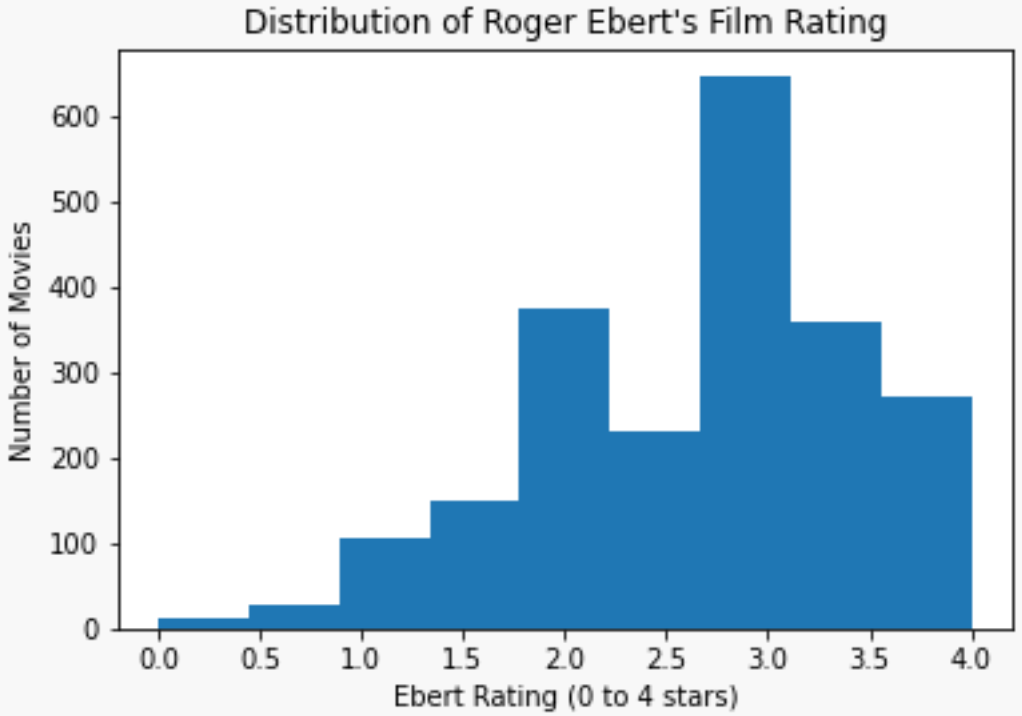


- Original data: 7847 datapoints, 6 features
- Final data: 2191 datapoints, 11 features

METHODOLOGY

Data Acquisition & Cleaning → **EDA, Feature Engineering & Selection** → Train & Evaluate Model (Model Selection)

TARGET: Ebert Rating



Least Amount of Ratings

Most Amount of Ratings

METHODOLOGY

Data Acquisition & Cleaning → **EDA, Feature Engineering & Selection** → Train & Evaluate Model (Model Selection)

TARGET: Ebert Rating	Year	MPAA Rating	Runtime
Genre	Sub-genre	MovieLens Rating	IMDb Rating
Budget	Worldwide Gross	Domestic Gross	Opening Week Gross

METHODOLOGY

Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)

TARGET: Ebert Rating	Year	MPAA Rating	Runtime
Genre	Sub-genre	MovieLens Rating	IMDb Rating
Budget	Worldwide Gross	Domestic Gross	Opening Week Gross
Genre dummies	Sub-genre dummies	MPAA Rating dummies	Opening Gross Proportion

METHODOLOGY

Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)

TARGET: Ebert Rating	Year	MPAA Rating	Runtime
Genre	Sub-genre	MovieLens Rating	IMDb Rating
Budget	Worldwide Gross	Domestic Gross	Opening Week Gross
Genre dummies	Sub-genre dummies	MPAA Rating dummies	Opening Gross Proportion

65 Features

METHODOLOGY

Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)

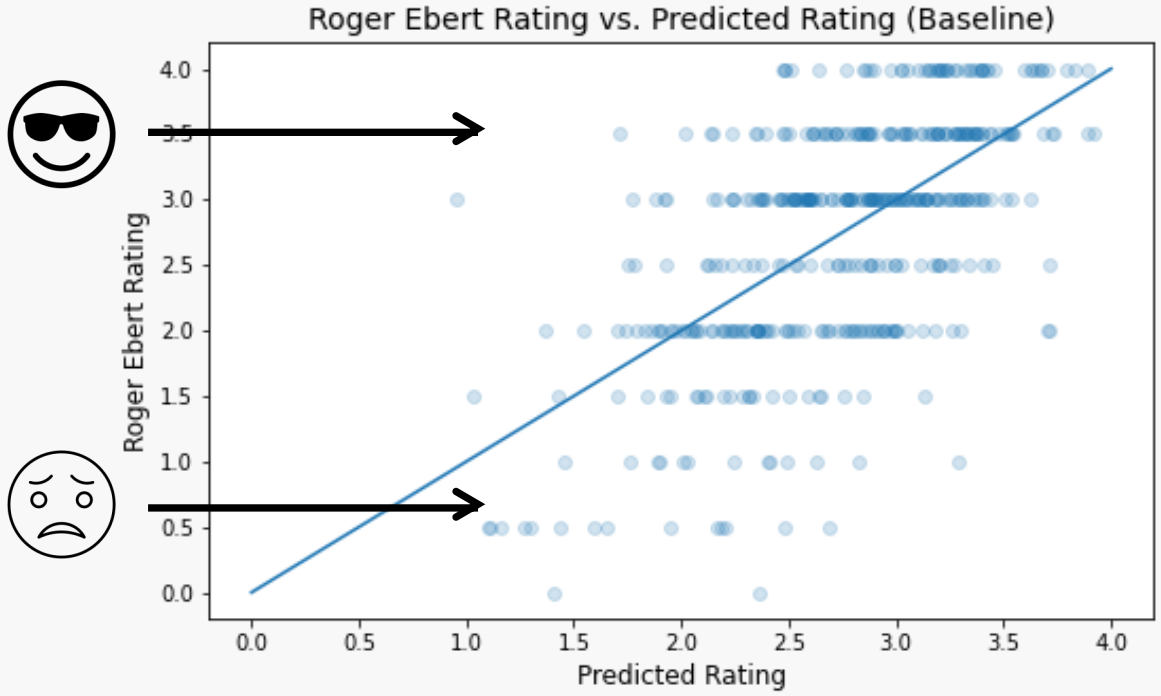
BASELINE MODEL

EBERT_RATING		R-squared:	0.382
OLS		Adj. R-squared:	0.379
Least Squares		F-statistic:	
Date:	Tue, 13 Apr 2021	Prob (F-statistic):	2.37e-220
Time:	11:59:05	Log-Likelihood:	-2307.0
No. Observations:	2189	AIC:	4634.
Df Residuals:	2179	BIC:	4691.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err		P>|t|	[0.025	0.975]
const	-0.3164	0.159	-1.99	0.046	0.627	-0.005
MOVIELENS_RATING	0.6912	0.082	8.46	0.000	0.531	0.851
IMDB_RATING	0.1379	0.044	3.16	0.002	0.052	0.223
RUNTIME	0.0024	0.001	2.86	0.004	0.001	0.004
MPAA_R	-0.1555	0.066	-2.35	0.019	0.285	-0.026
OPEN_PROPORTION	-0.7848	0.125	-6.27	0.000	1.030	-0.540
GENRE_Comedy	-0.2646	0.066	-4.01	0.000	0.394	-0.136
MPAA_PG-13	-0.2265	0.068	-3.33	0.001	0.360	-0.093
GENRE_Science Fiction	-0.1522	0.051	-2.97	0.003	0.252	-0.052
MPAA_PG	-0.1505	0.072	-2.09	0.036	0.291	-0.009

Omnibus:	93.423	Durbin-Watson:	2.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	104.986
Skew:	-0.513	Prob(JB):	1.59e-23
Kurtosis:	3.314	Cond. No.	1.34e+03

9 features with p-values <0.05



CAN I IMPROVE MY MODEL?

METHODOLOGY

Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)

Cross Validation on Each Model (Baseline)			
Model Type	**R^2 (Validation)**	**RSME**	**MAE**
Simple Linear Regression	0.368 +- 0.010	0.682	0.561
Ridge	0.368 +- 0.010	0.682	0.561
LASSO	0.368 +- 0.009	0.684	0.563
Polynomial	0.346 +- 0.017	0.691	0.570

- R-squared values are **close and low** between models and between train and validation sets (except for Polynomial)
- Suggests that the models may be **underfitting**

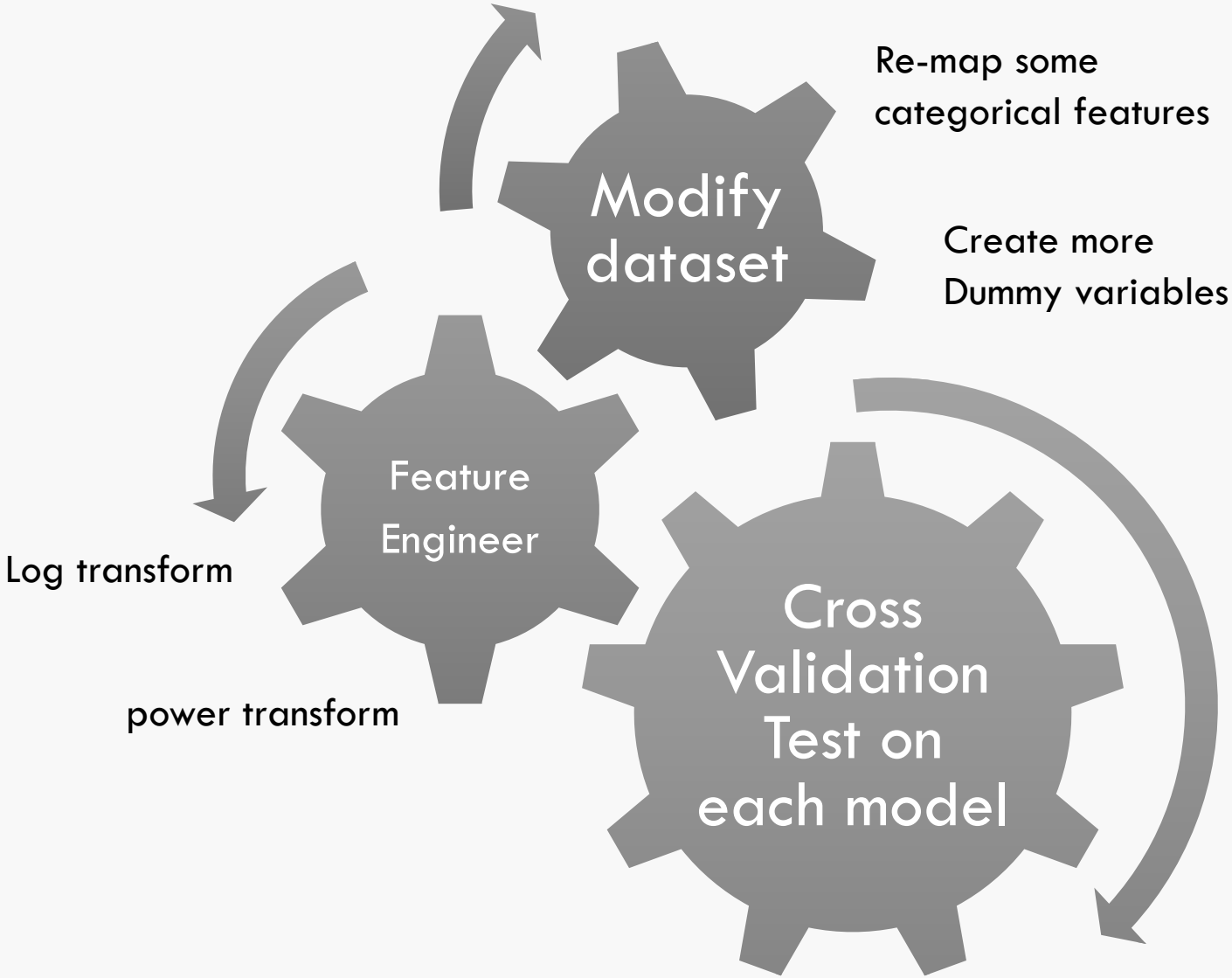
Cross Validation on Each Model (Baseline)		
Model Type	**R^2 (Validation)**	**R^2 (Train)**
Simple Linear Regression	0.368 +- 0.010	0.384 +- 0.002
Ridge	0.368 +- 0.010	0.384 +- 0.002
LASSO	0.368 +- 0.009	0.384 +- 0.002
Polynomial	0.346 +- 0.017	0.408 +- 0.003

What Next?

- **Increase Complexity** by
  - Adding more features to the dataset
  - Feature Engineering

METHODOLOGY

Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)



METHODOLOGY

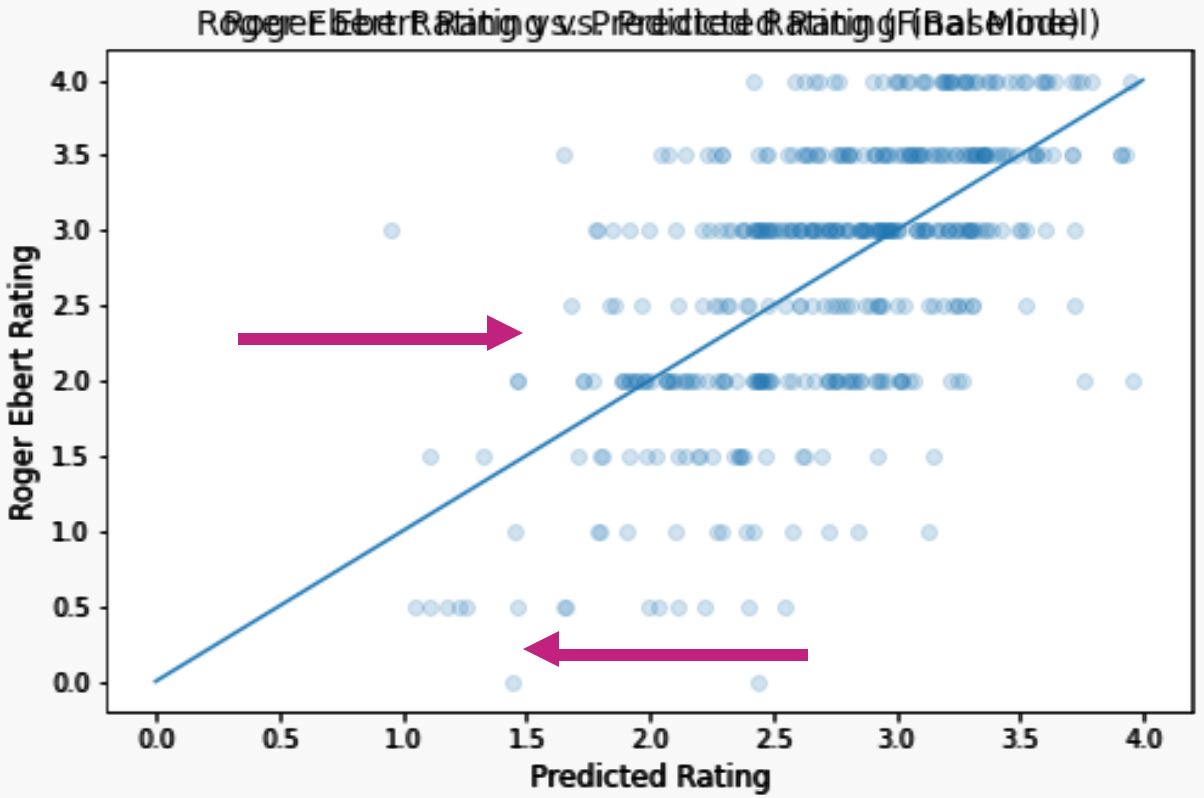
Data Acquisition & Cleaning → EDA, Feature Engineering & Selection → Train & Evaluate Model (Model Selection)

Cross Validation Test on Each Model (Final)			
Model Type	**R^2**	**RSME**	**MAE**
Simple Linear Regression	0.3708	0.707	0.560
Ridge	0.3703	0.708	0.560
LASSO	0.3711	0.707	0.559
Polynomial	-4077.74	56.97	9.703
Polynomial-Ridge	0.2611	0.767	0.604
Polynomial-LASSO	0.3454	0.722	0.574

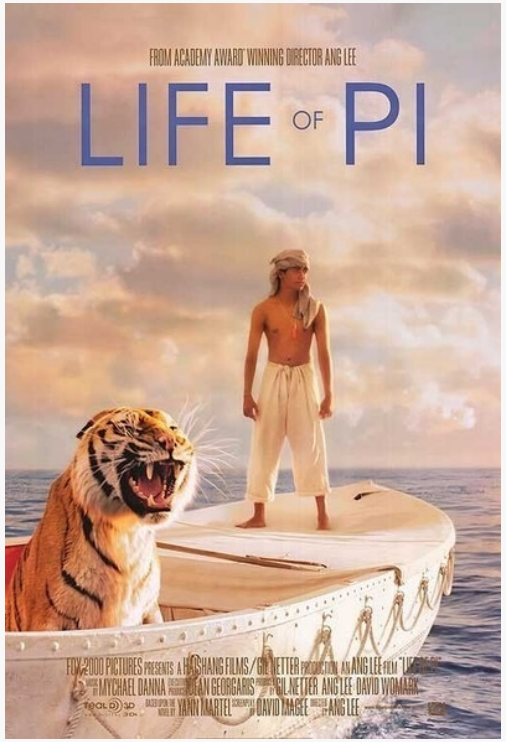
RESULTS – MODEL PERFORMANCE

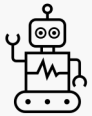
	Baseline	Final Model
R^2	0.382	0.396
MAE	0.563	0.55

Prediction is off by about 0.5 star

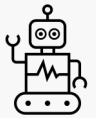


RESULTS –
MODEL PREDICTION (VS. ACTUAL RATINGS)




3.6


4.0

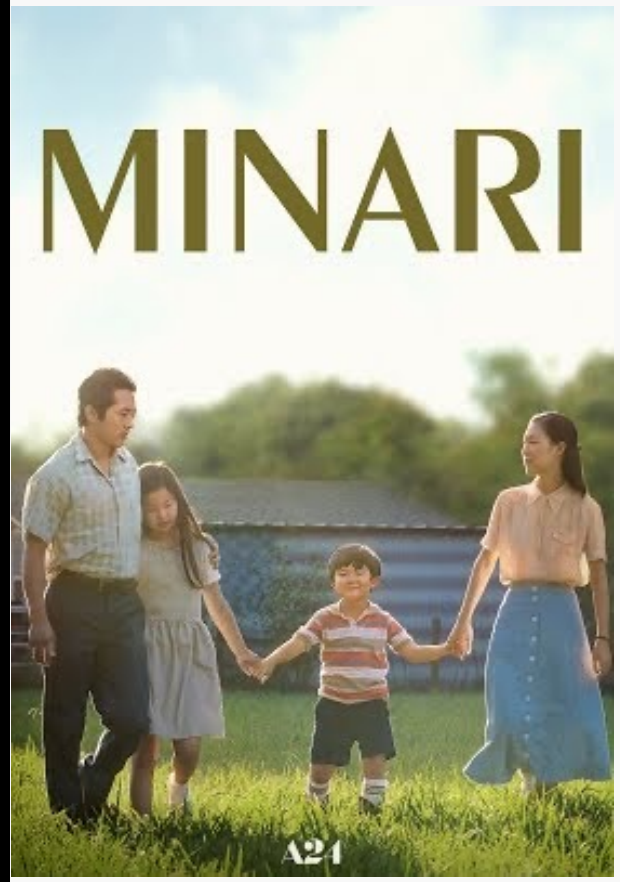

2.5


2.0


1.7


1.0

RESULTS – MODEL PREDICTION



3.6

2.6

3.4

3.5

INSIGHTS

- Linear Regression may not be the best prediction model for this dataset
- R-squared values are typically lower than 50% for predicting human behavior. Humans are simply harder to predict than, say, physical processes. [1]

FUTURE WORK

- Non-linear prediction model
  - *Tree-based model*
- More datapoints –
  - *filling missing values in original dataset*
  - *web-scraping more features (i.e. directors, writer, actors/actress)*
- Flask app – deploy the prediction model


THANK YOU
Questions?
crystal-ctrl
crystal-huang-ds

APPENDIX 1: ARTICLE CITED

- How to Interpret R-squared in Regression Analysis by Jim Frost
  - *https://statisticsbyjim.com/regression/interpret-r-squared-regression*

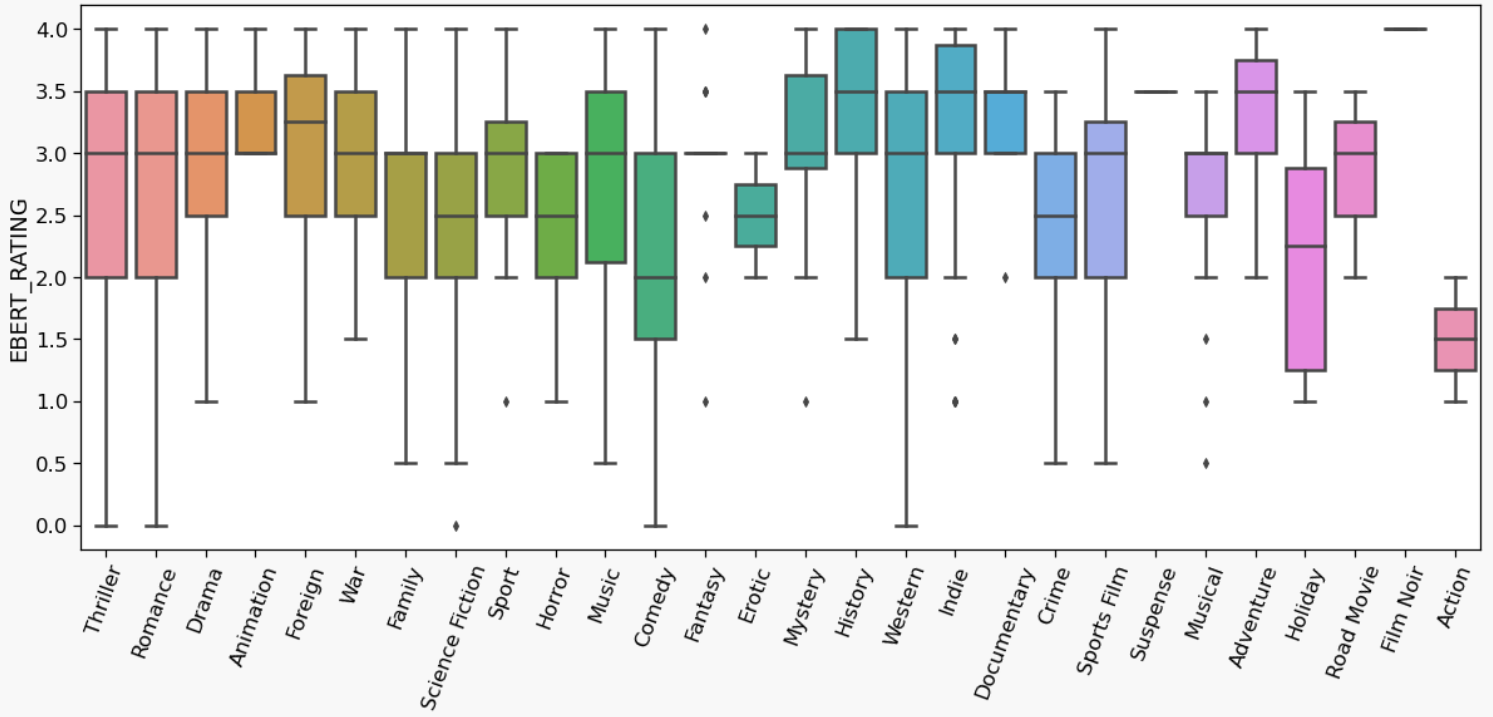
ARE LOW R-SQUARED VALUES INHERENTLY BAD?

No! There are two major reasons why it can be just fine to have low R-squared values.

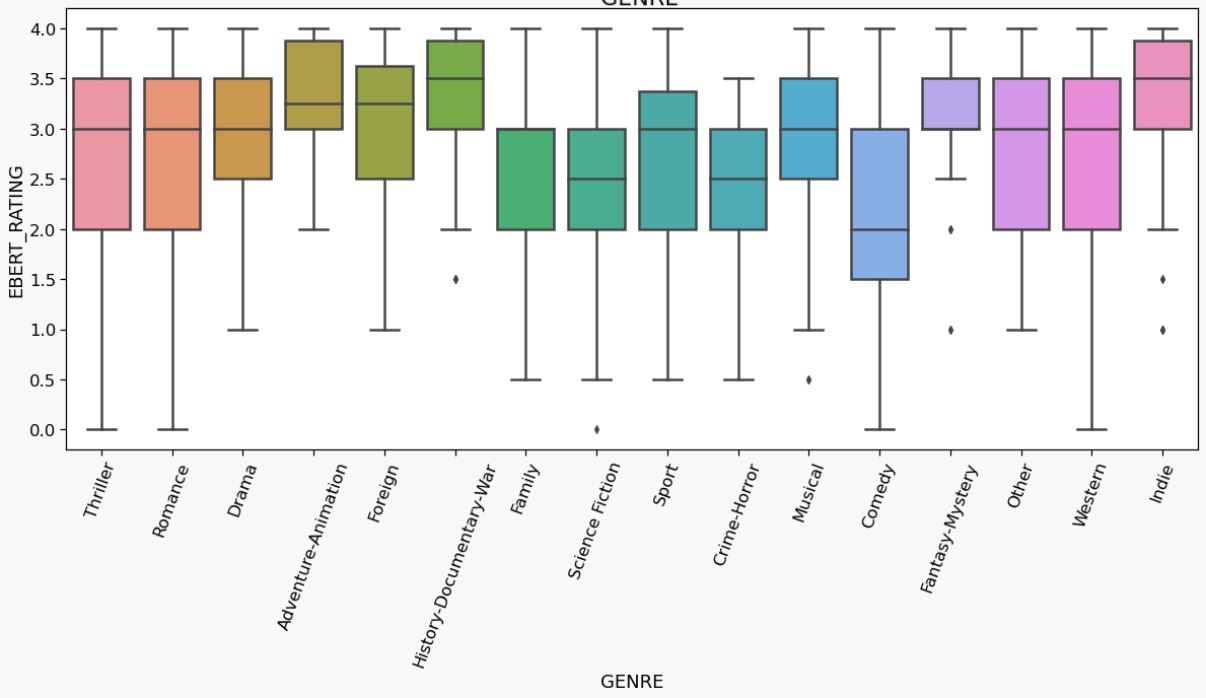
In some fields, it is entirely expected that your R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes.

APPENDIX 2

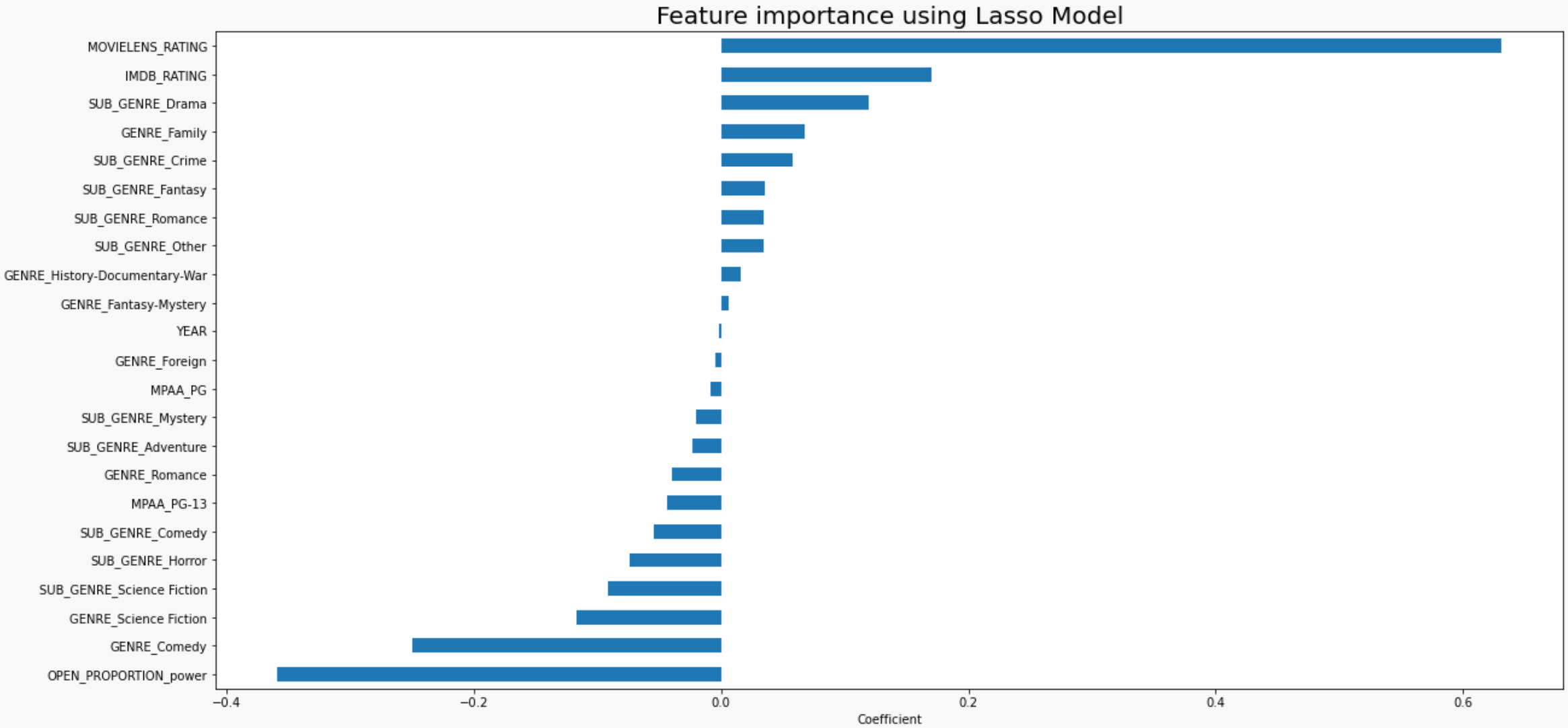
Before



After



APPENDIX 3


Feature importance using Lasso Model
MOVIELENS_RATING
IMDB_RATING
SUB_GENRE_Drama
GENRE_Family
SUB_GENRE_Crime
SUB_GENRE_Fantasy
SUB_GENRE_Romance
SUB_GENRE_Other
GENRE_History-Documentary-War
GENRE_Fantasy-Mystery
YEAR
GENRE_Foreign
MPAA_PG
SUB_GENRE_Mystery
SUB_GENRE_Adventure
GENRE_Romance
MPAA_PG-13
SUB_GENRE_Comedy
SUB_GENRE_Horror
SUB_GENRE_Science Fiction
GENRE_Science Fiction
GENRE_Comedy
OPEN_PROPORTION_power
−0.4
−0.2
0.0
0.2
0.4
0.6
Coefficient

APPENDIX 4-

Assumption 1: Regression is linear in parameters and correctly specified