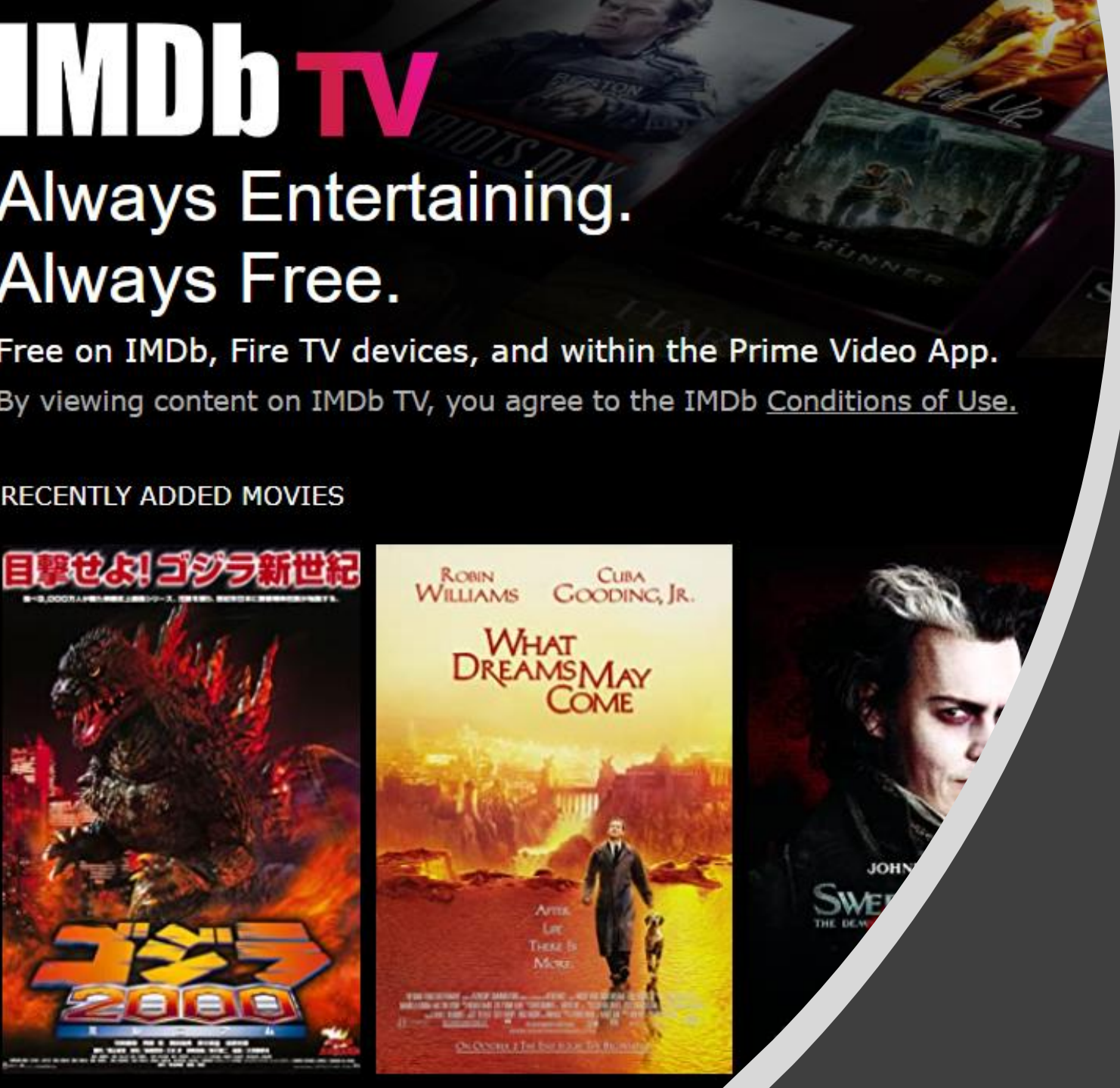
Predictability of TV Show Ratings



Objective:

Use information from IMDB website to predict TV show ratings

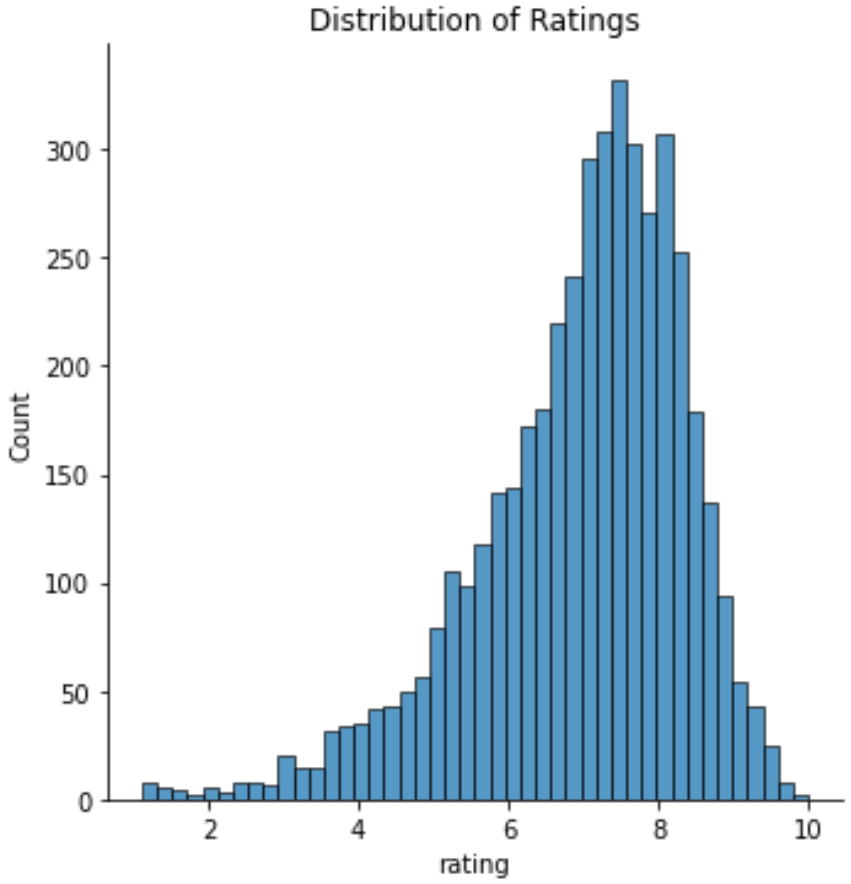
Methodology

- Data:
  - 4499 tv shows
  - Features – year released, number of reviews, genre, certificate, number of episodes, actors, network, runtime, and awards
  - Target – Rating

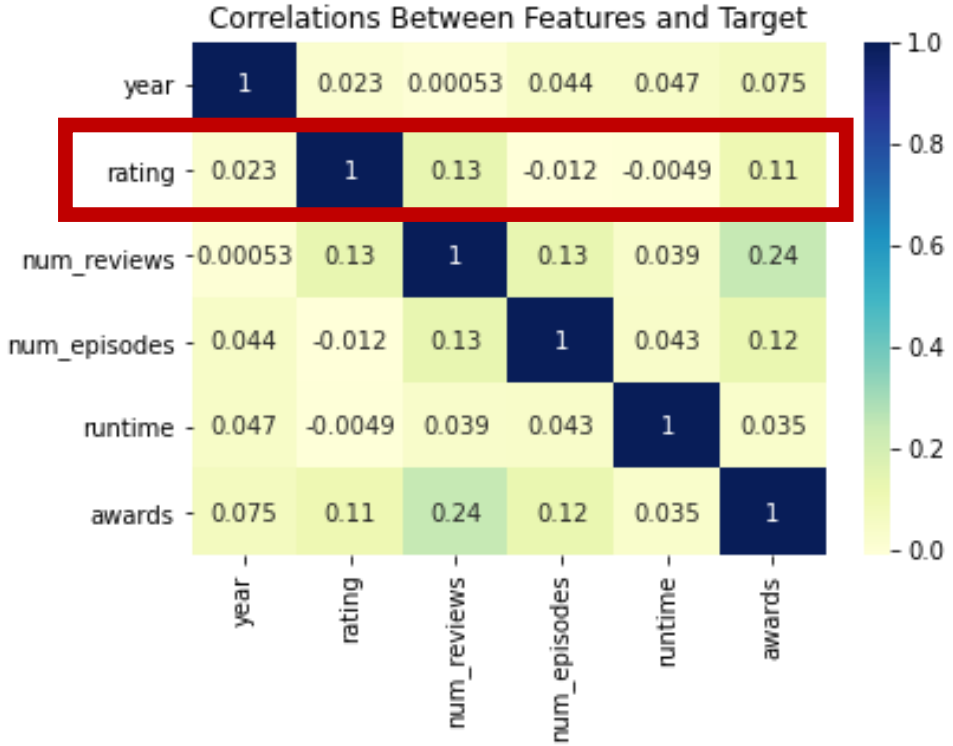
- Tools:
  - Python
  - BeautifulSoup
  - Statsmodels
  - Scikit Learn
  - Pandas
  - Numpy

- Models:
  - Ordinary Least Squares (OLS)
  - Ridge
  - Lasso

Data Analysis



Target distribution is left skewed



No strong correlations between numeric features and target

Baseline Model

Dep. Variable:	rating	**R-squared:**	0.026
Model:	OLS	**Adj. R-squared:**	0.025
Method:	Least Squares	**F-statistic:**	23.69
Date:	Wed, 14 Apr 2021	**Prob (F-statistic):**	1.37e-23
Time:	16:38:28	**Log-Likelihood:**	-7747.4
No. Observations:	4499	**AIC:**	1.551e+04
Df Residuals:	4493	**BIC:**	1.555e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-14.5869	17.400	-0.838	0.402	-48.699	19.526
year	0.0106	0.009	1.233	0.218	-0.006	0.028
awards	0.2473	0.044	5.620	0.000	0.161	0.334
num_episodes	-0.0002	7.89e-05	-2.554	0.011	-0.000	-4.69e-05
runtime	-0.0005	0.001	-0.791	0.429	-0.002	0.001
num_reviews	5.463e-06	7.19e-07	7.601	0.000	4.05e-06	6.87e-06

R-Squared = 0.026

Validation and Testing Schemes

- Used cross validation with Kfolds=5

Simple Linear Regression	Ridge	Lasso
Train Mean CV R^2 = 0.117	Train Mean CV R^2 = 0.117	Train Mean CV R^2 = 0.110
Validation Mean CV R^2 = 0.056	Validation Mean CV R^2 = 0.056	Validation Mean CV R^2 = 0.061

Conclusion: All three models are overfit

Model Refinement - LassoCV

- Removed 2 features – actors and network
- Added quadratic terms to 2 features – number of reviews and runtime

Final R^2 value = 0.084

MAE = 1.03

RMSE = 1.35

Future Work

Sitcoms	The Office		The Big Bang Theory		Arrested Development		Scrubs		South Park	
Regression Model Evaluation	R^2	**RMS**	R^2	**RMS**	R^2	**RMS**	R^2	**RMS**	R^2	**RMS**
Linear	-0.505	0.337	-0.14	0.185	-4.16	1.082	-1.85	0.323	-1.16	0.531
K Nearest Neighbors	**0.398**	**0.135**	**0.176**	**0.134**	-0.17	0.245	0.043	0.105	-0.53	0.37
Stochastic Gradient Descent	-24.145	5.636	-7.221	1.339	-32.28	6.966	-10.234	1.24	-18.71	4.8
Decision Tree	0.147	0.191	0.17	0.134	-0.271	0.266	-0.274	0.147	-1.75	0.67
Neural Network	0.321	0.15	-0.051	0.171	-0.415	0.296	-0.248	0.138	**-0.24**	**0.3**
Decision Forest	0.33	0.15	0.173	0.135	**0.40**	**0.126**	**0.163**	**0.0921**	-0.4	0.34

TABLE II
PERFORMANCE OF MACHINE LEARNING BASED REGRESSION MODELS ON SITCOM DATASETS

- Use more advance models
- Work with networks to obtain more data about each show that aren't listed on IMDB

Appendix

Lasso Regularization
Train Mean CV R^2 = 0.077
Validation Mean CV R^2 = 0.082
Test Mean CV R^2 = 0.084

- Forecasting the Success of Television Series using Machine Learning ‘https://arxiv.org/pdf/1910.12589.pdf’