

# Explaining IMDb Ratings by Actor and Director Characteristics

Charissa Rentier  
2018-02-02



# The Question

Can we explain movie rating by actor and director characteristics, leaving out details about the contents of the movie (genre, topic, title, etc.)?

# The More Interesting Question

In particular, how does actor **diversity** fit in?

# The Data

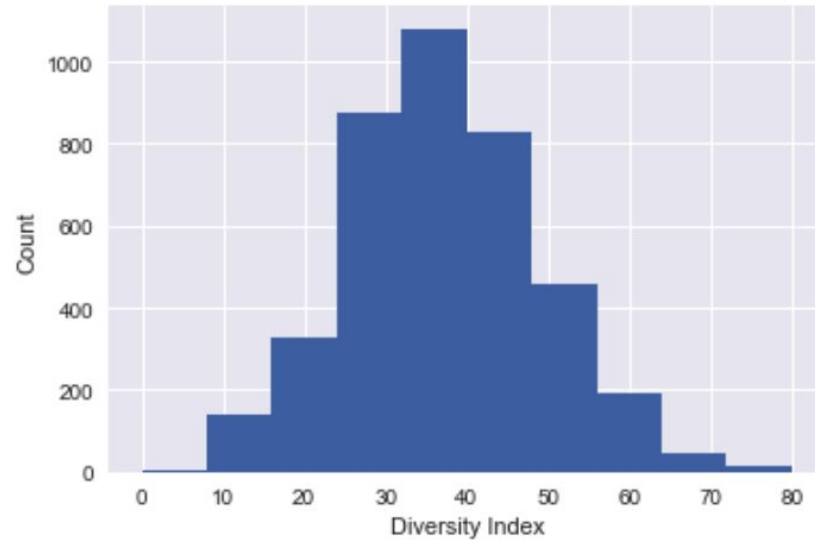
- Source: IMDb
- Selection: United States, 1990 - 2018, > 500 Votes
- Actor Characteristics
  - Acting Roles and Other Roles
  - Media Presence
  - Award Wins and Nominations
  - *Gender*
  - *Birth Country*
  - *Age at Time of Release*
- Director Characteristics
  - Directing Roles and Other Roles
  - Media Presence
  - Award Wins and Nominations

# The Diversity Index

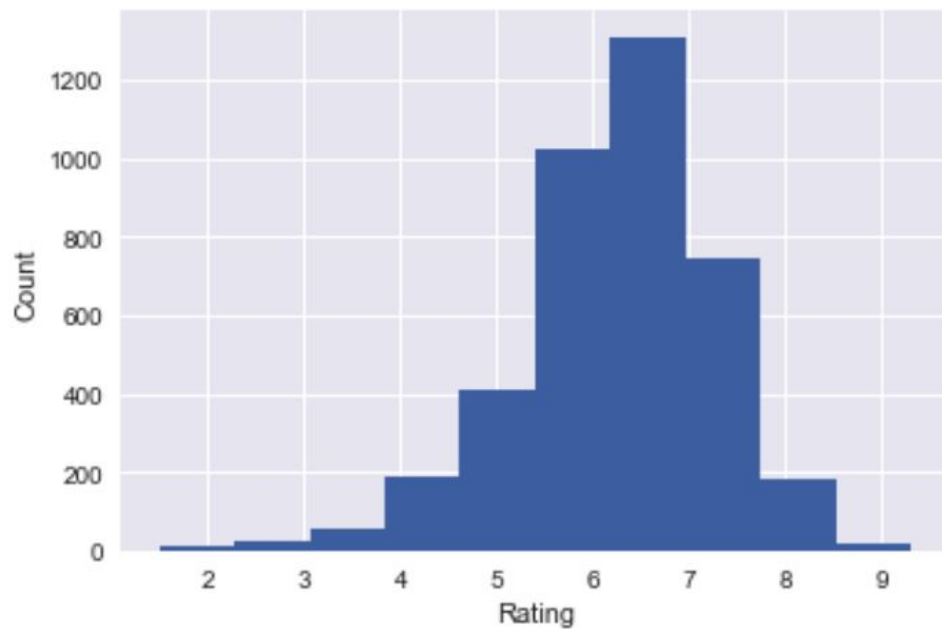
Composite of:

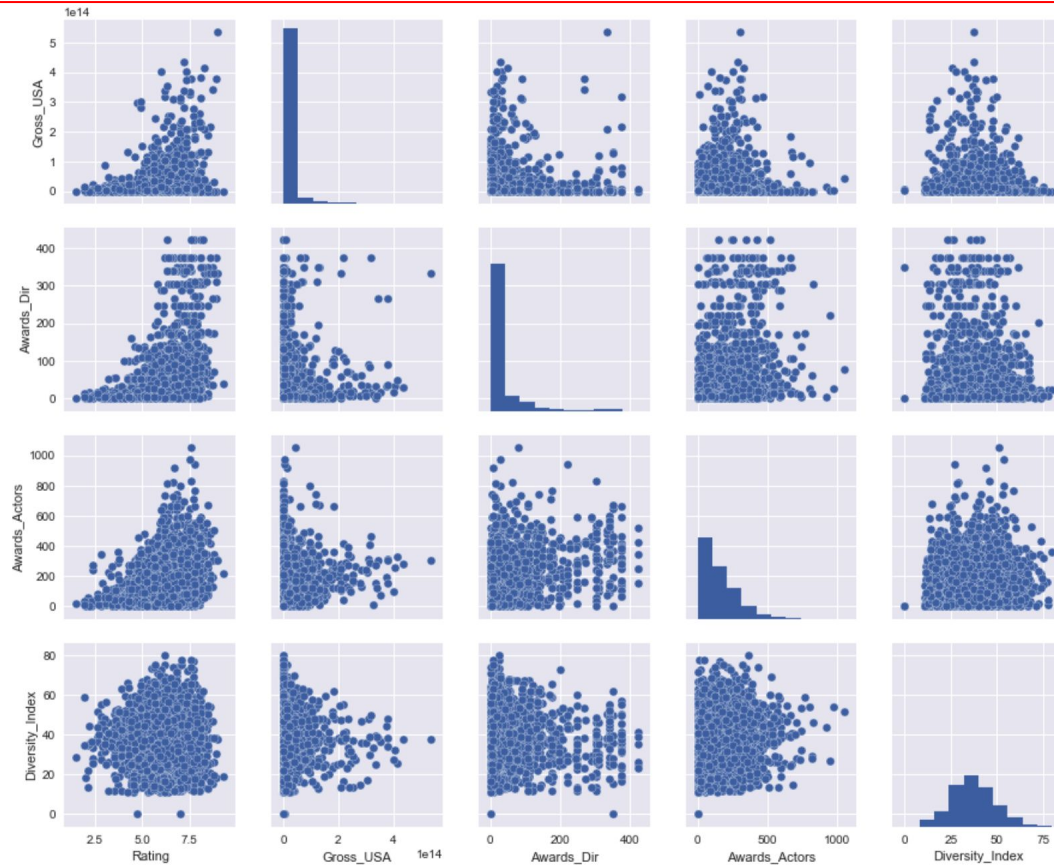
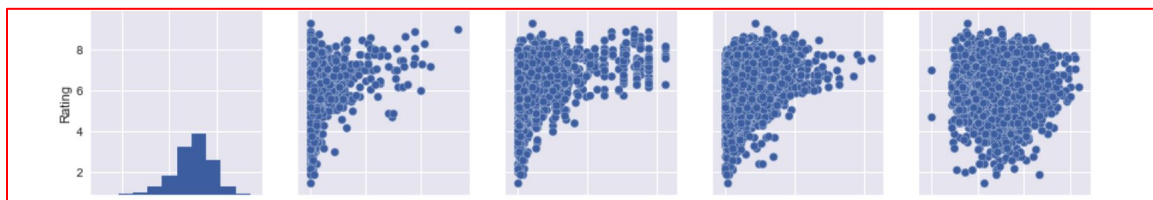
- Share of Female Actors
- Difference in Birth Countries among Actors
- Age Spread among Actors

# The Diversity Index (spread)



# The Ratings







# Feature and Model Selection

OLS Regression Results

<b>Dep. Variable:</b>	Rating	<b>R-squared:</b>	0.215
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.212
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	86.44
<b>Date:</b>	Fri, 02 Feb 2018	<b>Prob (F-statistic):</b>	6.99e-158
<b>Time:</b>	02:21:23	<b>Log-Likelihood:</b>	-4166.2
<b>No. Observations:</b>	3169	<b>AIC:</b>	8354.
<b>Df Residuals:</b>	3158	<b>BIC:</b>	8421.
<b>Df Model:</b>	10		
<b>Covariance Type:</b>	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
<b>const</b>	6.0131	0.065	91.915	0.000	5.885	6.141
<b>Gross_USA</b>	2.121e-15	4.19e-16	5.058	0.000	1.3e-15	2.94e-15
<b>DirRoles_Dir</b>	-0.0047	0.001	-4.992	0.000	-0.007	-0.003
<b>Pub_Dir</b>	-0.0054	0.001	-6.834	0.000	-0.007	-0.004
<b>TVApp_Dir</b>	0.0019	0.000	4.375	0.000	0.001	0.003
<b>Awards_Dir</b>	0.0049	0.000	12.488	0.000	0.004	0.006
<b>Awards_Actors</b>	0.0028	0.000	14.106	0.000	0.002	0.003
<b>OthRoles_Actors</b>	-0.0019	0.001	-2.231	0.026	-0.004	-0.000
<b>TVApp_Actors</b>	-0.0003	0.000	-2.188	0.029	-0.001	-3.56e-05
<b>Pub_Actors</b>	-0.0006	0.000	-3.295	0.001	-0.001	-0.000
<b>Diversity_Index</b>	-0.0032	0.001	-2.213	0.027	-0.006	-0.000
<b>Omnibus:</b>	311.551	<b>Durbin-Watson:</b>	2.006			
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	488.771			
<b>Skew:</b>	-0.721	<b>Prob(JB):</b>	7.32e-107			
<b>Kurtosis:</b>	4.273	<b>Cond. No.</b>	1.66e+14			

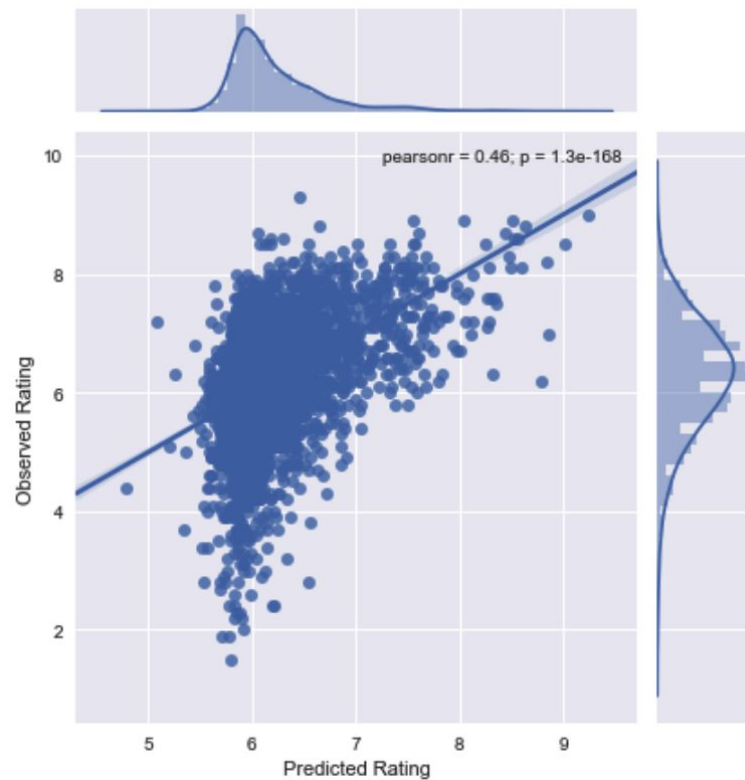
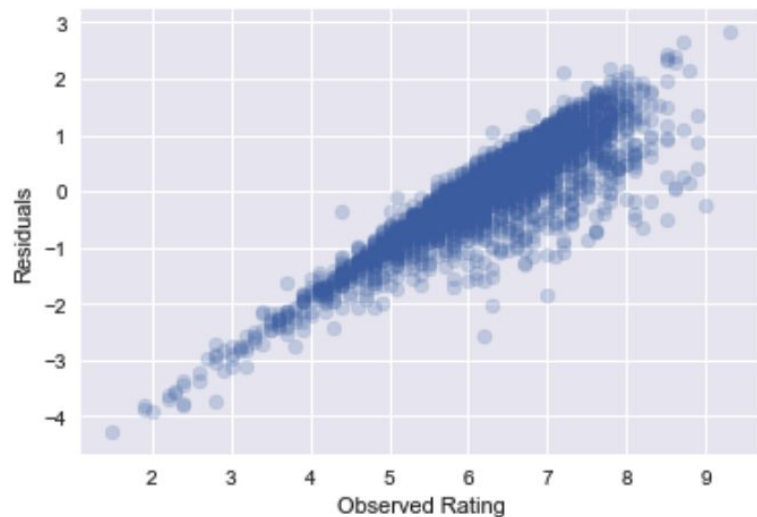
# Feature and Model Selection

Things to note:

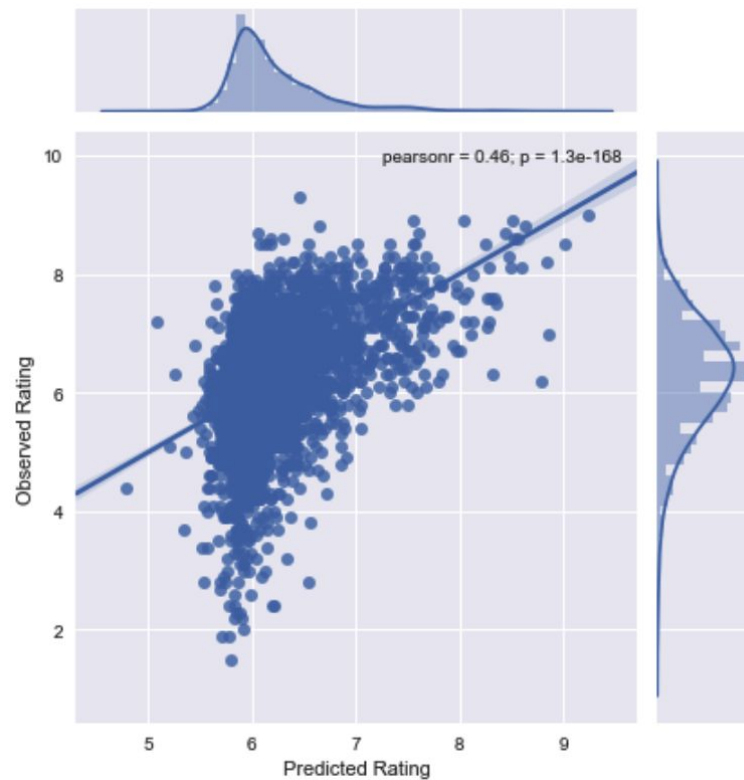
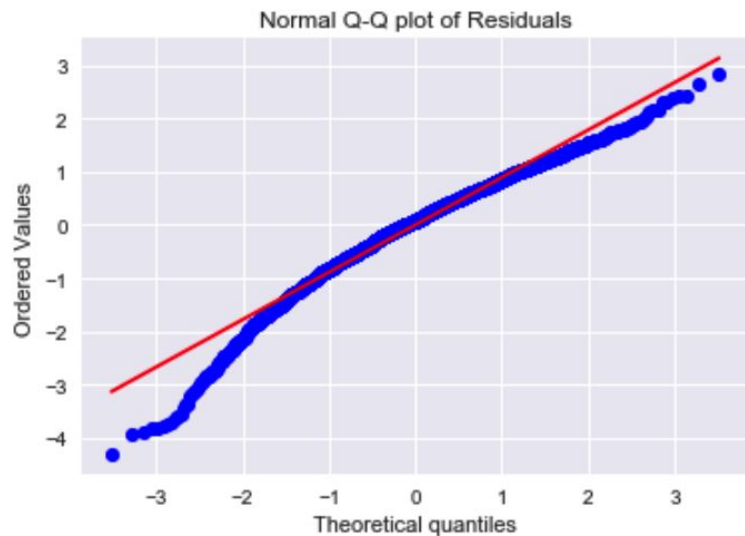
- Negative Coefficient for Diversity
- Positive Coefficients for Actor and Director Awards

	coef	std err	t	P> t	[0.025	0.975]
const	6.0131	0.065	91.915	0.000	5.885	6.141
Gross_USA	2.121e-15	4.19e-16	5.058	0.000	1.3e-15	2.94e-15
DirRoles_Dir	-0.0047	0.001	-4.992	0.000	-0.007	-0.003
Pub_Dir	-0.0054	0.001	-6.834	0.000	-0.007	-0.004
TVApp_Dir	0.0019	0.000	4.375	0.000	0.001	0.003
Awards_Dir	0.0049	0.000	12.488	0.000	0.004	0.006
Awards_Actors	0.0028	0.000	14.106	0.000	0.002	0.003
OthRoles_Actors	-0.0019	0.001	-2.231	0.026	-0.004	-0.000
TVApp_Actors	-0.0003	0.000	-2.188	0.029	-0.001	-3.56e-05
Pub_Actors	-0.0006	0.000	-3.295	0.001	-0.001	-0.000
Diversity_Index	-0.0032	0.001	-2.213	0.027	-0.006	-0.000
Omnibus:	311.551	Durbin-Watson:		2.006		
Prob(Omnibus):	0.000	Jarque-Bera (JB):		488.771		
Skew:	-0.721	Prob(JB):		7.32e-107		
Kurtosis:	4.273	Cond. No.		1.66e+14		

# Feature and Model Selection



# Feature and Model Selection



# The Preliminary Results

Simple regression mean CV  $R^2$ : 0.210  $\pm$  0.010

Degree 2 polynomial mean CV  $R^2$ : 0.010  $\pm$  0.025

Ridge mean CV  $R^2$ : 0.210  $\pm$  0.010

Lasso mean CV  $R^2$ : 0.210  $\pm$  0.010

Simple Linear Model Coefficients:

```
[('Gross_USA', 1.9457014843793962e-15),  
 ('DirRoles_Dir', -0.0046025943342359165),  
 ('Pub_Dir', -0.0051318569513767367),  
 ('TVApp_Dir', 0.0018858125569139862),  
 ('Awards_Dir', 0.0048302098137087343),  
 ('Awards_Actors', 0.0026360314584318147),  
 ('OthRoles_Actors', -0.0022416376375657956),  
 ('TVApp_Actors', -0.00026954215178174494),  
 ('Pub_Actors', -0.00047978096967295377),  
 ('Diversity_Index', -0.0037316156817060854)]
```

# The Preliminary Results

Simple regression mean CV  $R^2$ : 0.210  $\pm$  0.010

Degree 2 polynomial mean CV  $R^2$ : 0.010  $\pm$  0.025

Ridge mean CV  $R^2$ : 0.210  $\pm$  0.010

Lasso mean CV  $R^2$ : 0.210  $\pm$  0.010

Random Forest mean CV  $R^2$ : 0.243  $\pm$  0.014

Gradient Boosted mean CV  $R^2$ : 0.265  $\pm$  0.022

Gradient Boosted Model Feature Importances:

```
[('Gross_USA', 0.19255294153249228),  
 ('DirRoles_Dir', 0.079932336495808068),  
 ('Pub_Dir', 0.028005357387981666),  
 ('TVApp_Dir', 0.094659429971916884),  
 ('Awards_Dir', 0.08653269907407711),  
 ('Awards_Actors', 0.11877634273199668),  
 ('OthRoles_Actors', 0.076928030613881301),  
 ('TVApp_Actors', 0.11914705928408131),  
 ('Pub_Actors', 0.11070117659767953),  
 ('Diversity_Index', 0.092764626310084947)]
```

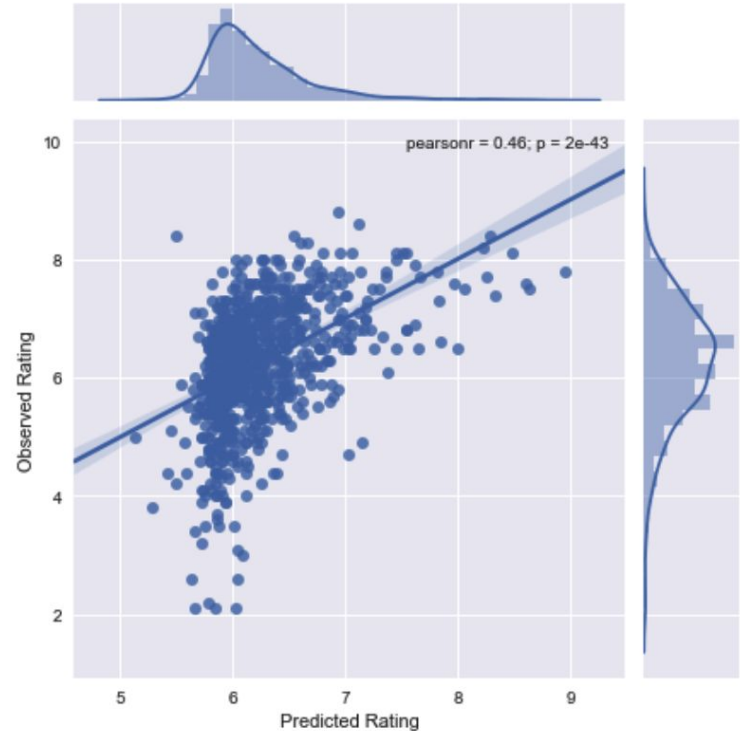
# The Test Results

Simple regression test  $R^2$ : 0.215

Ridge regression test  $R^2$ : 0.215

Lasso regression test  $R^2$ : 0.215

Model fit of Simple Linear Regression



# The Test Results

Simple regression test  $R^2$ : 0.215

Ridge regression test  $R^2$ : 0.215

Lasso regression test  $R^2$ : 0.215

Simple Linear Model Coefficients:

```
[('Gross_USA', 2.1210507296905302e-15),  
 ('DirRoles_Dir', -0.0047484679075308718),  
 ('Pub_Dir', -0.0054471448693569268),  
 ('TVApp_Dir', 0.0018841382639200978),  
 ('Awards_Dir', 0.0048767636946479278),  
 ('Awards_Actors', 0.0027523972619924904),  
 ('OthRoles_Actors', -0.0018672352763784588),  
 ('TVApp_Actors', -0.00034336601015047844),  
 ('Pub_Actors', -0.00055619490342404508),  
 ('Diversity_Index', -0.0031503309350486676)]
```



# The Test Results

Simple regression test  $R^2$ : 0.215

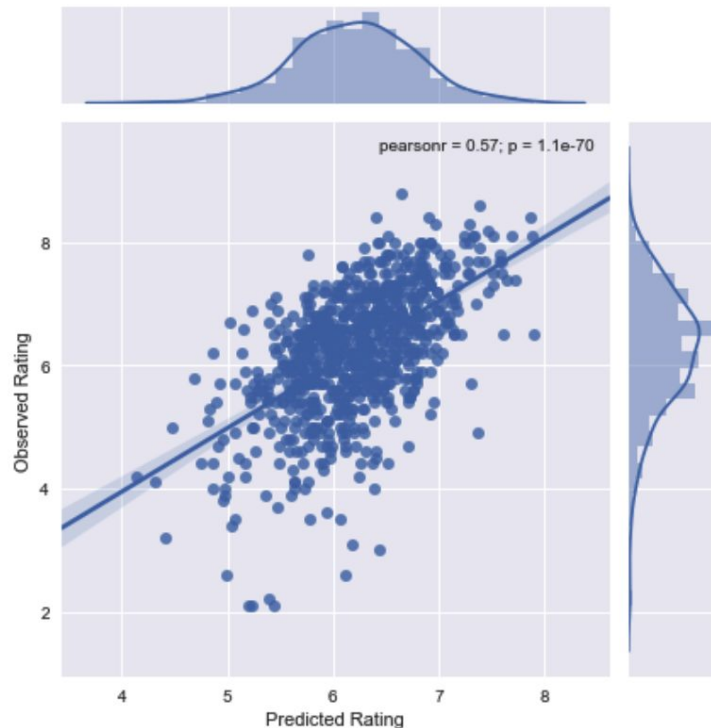
Ridge regression test  $R^2$ : 0.215

Lasso regression test  $R^2$ : 0.215

Random Forest regression test  $R^2$ : 0.284

Gradient Boosted regression test  $R^2$ : 0.329

Model fit of Gradient Boosted Regression



# The Test Results

Simple regression test  $R^2$ : 0.215

Ridge regression test  $R^2$ : 0.215

Lasso regression test  $R^2$ : 0.215

Random Forest regression test  $R^2$ : 0.284

Gradient Boosted regression test  $R^2$ : 0.329

Gradient Boosted Model Feature Importances:

```
[('Gross_USA', 0.18365443249343977),  
 ('DirRoles_Dir', 0.080877592725099923),  
 ('Pub_Dir', 0.024577300248119757),  
 ('TVApp_Dir', 0.098792045492534467),  
 ('Awards_Dir', 0.088853385435691129),  
 ('Awards_Actors', 0.13507694080267124),  
 ('OthRoles_Actors', 0.070227800575368773),  
 ('TVApp_Actors', 0.11901628522757654),  
 ('Pub_Actors', 0.11422522028274802),  
 ('Diversity_Index', 0.084698996716750449)]
```

# Limitations

- Non-normally distributed errors
- Skewed distribution of features

# Further Research

- Additional Feature Transformations
- Additional Information
  - Maybe it was too optimistic to want to predict Rating on Actor and Director Characteristics
- Expand Timeframe
  - Scraped data for ~30k movies, there's more data there!