Mariya Graff

December 15th, 2021

IMBD Movie Rating

A linear regression model to describe and predict the IMBD rating for a new movie.

Contents

1 Why IMBD rating?
EDA and features

2 Base model

3 Progress (VIF, feature engineering)

4 Final model
"Power of the Dog" test

1

Why IMBD rating?

Gauge of public reception; discovery tool

325M monthly visits

Gauge of public reception

Movie discovery top rated lists

31st top visited websites in the US

3rd top visited website in TV, Movie and Streaming in the world

Suggestion algorithm

1 Data Sources & Tools

Web-scraped

BeautifulSoup, Requests

IMDb

Imported

Pandas

kaggle

Other tools: *Python, Matplotlib, Seaborn, Numpy, Statsmodels, sklearn*

1 Cleaning and EDA

From this

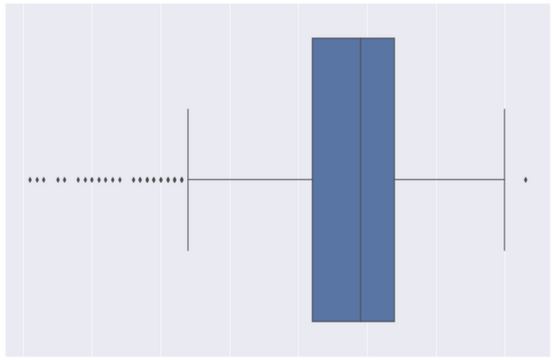
```
[{'title': 'Red Notice',
  'year': '2021',
  'certificate': 'PG-13',
  'runtime_min': '118',
  'genre': 'Action',
  'sub_genre': ['Action', ' Comedy', ' Crime'],
  'rating': '6.4',
  'votes': '133157',
  'metascore': '39',
  'gross': '133,157'},
 {'title': 'Ghostbusters: Afterlife',
  'year': '2021',
  'certificate': 'PG-13',
  'runtime_min': '124',
  'genre': 'Adventure',
  'sub_genre': ['Adventure', ' Comedy', ' Fantasy'
  'rating': '7.7',
  'votes': '29062',
  'metascore': '59',
```

To this

rating	votes	metascore	gross_mil	director	writer	star	company	budget
7.6	516213	63	285.76	Chris Columbus	John Hughes	Macaulay Culkin	Hughes Entertainment	
7.9	551997	82	165.36	Rian Johnson	Rian Johnson	Daniel Craig	Lionsgate	
7.6	175550	78	7.99	Richard Linklater	Richard Linklater	Jason London	Gramercy Pictures (I)	
7.6	652796	83	142.50	Quentin Tarantino	Quentin Tarantino	Leonardo DiCaprio	Columbia Pictures	
7.0	190141	75	127.81	Kevin Lima	Bill Kelly	Amy Adams	Walt Disney Pictures	

Things like...

- Removed duplicates and NaN values
- Scaled data
- Removed Outliers



1

... arrived at:

2112 data points/rows
28 features/columns

Rating	Runtime (min)	Number of Votes	Metascore	Critic rating	Budget (mil)
MPAA rating →	G	PG	PG-13	R	
Genres →	Action	Adventure	Animation	Biography	Biography
	Comedy	Crime	Drama	Horror	
Directors →	Christopher Nolan	Martin Scorsese	Paul T. Anderson	Peter Jackson	Quentin Tarantino
Most frequent directors with highest avg rating	Sam Mendes	Steven Spielberg	Wes Anderson	David Fincher	James Cameron
	Other				

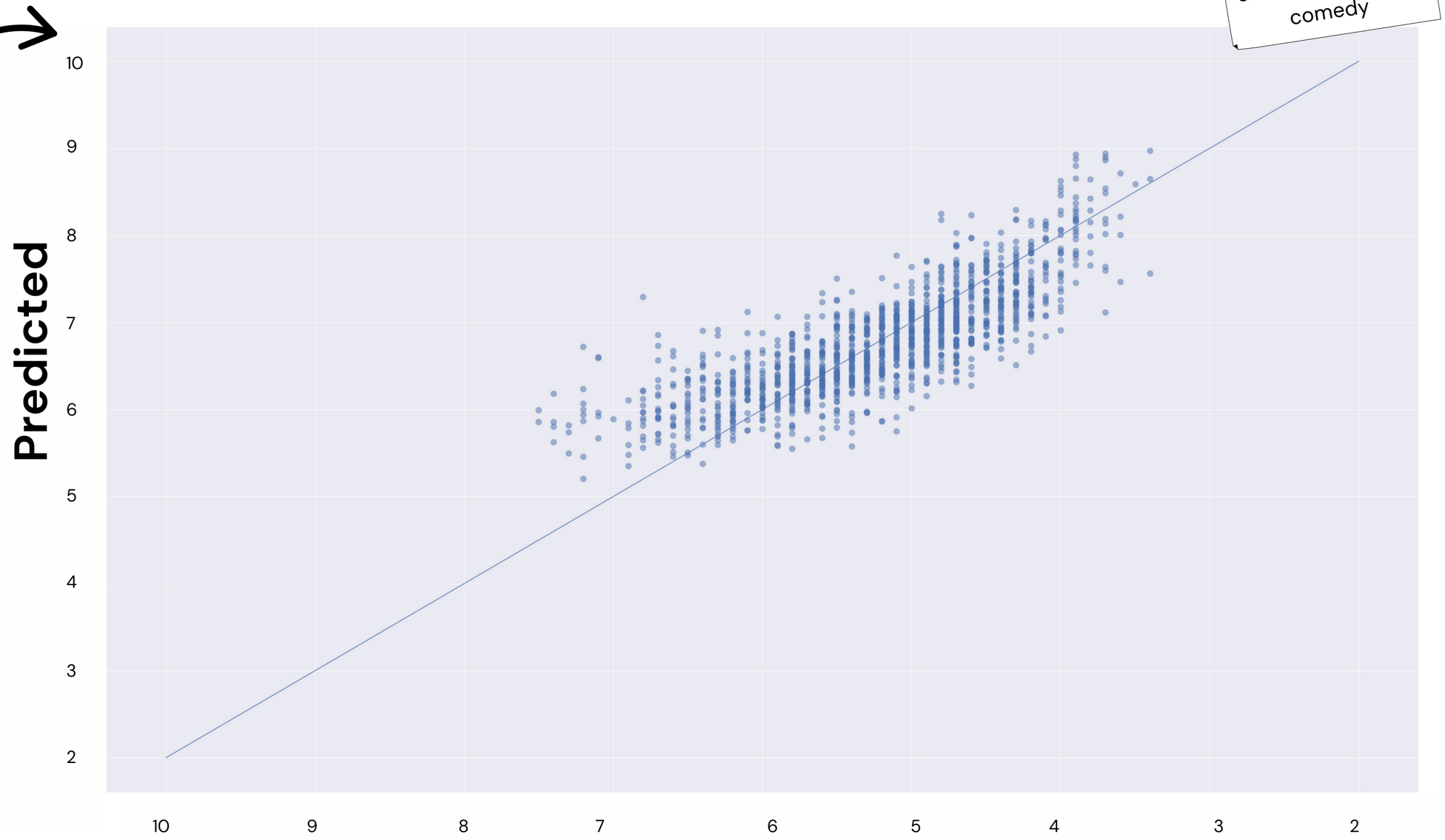
2 Base Model: R2 training of 0.641

Reg R2: 0.641
Ridge R2: 0.641
Lasso R2: 0.639

Reg RMSE: 0.494
Ridge RMSE: 0.494
Lasso RMSE: 0.496

KFold cross-validation

Comparing to the most general combo of G-rated, comedy

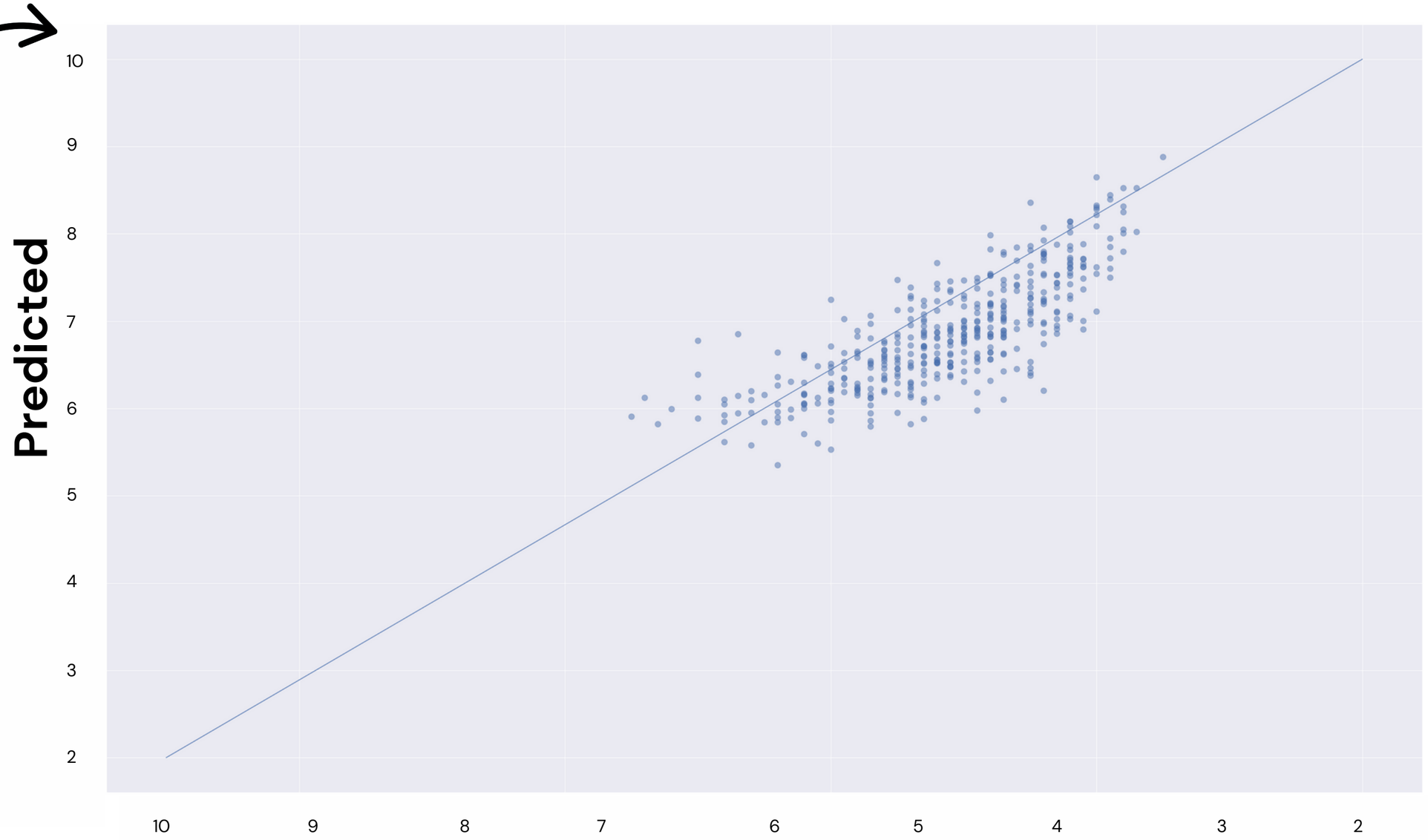


2 Base Model: R2 test of 0.61

Reg R2: 0.61

Ridge R2: 0.597

Lasso R2: 0.599



3

Let's try to do better...

3

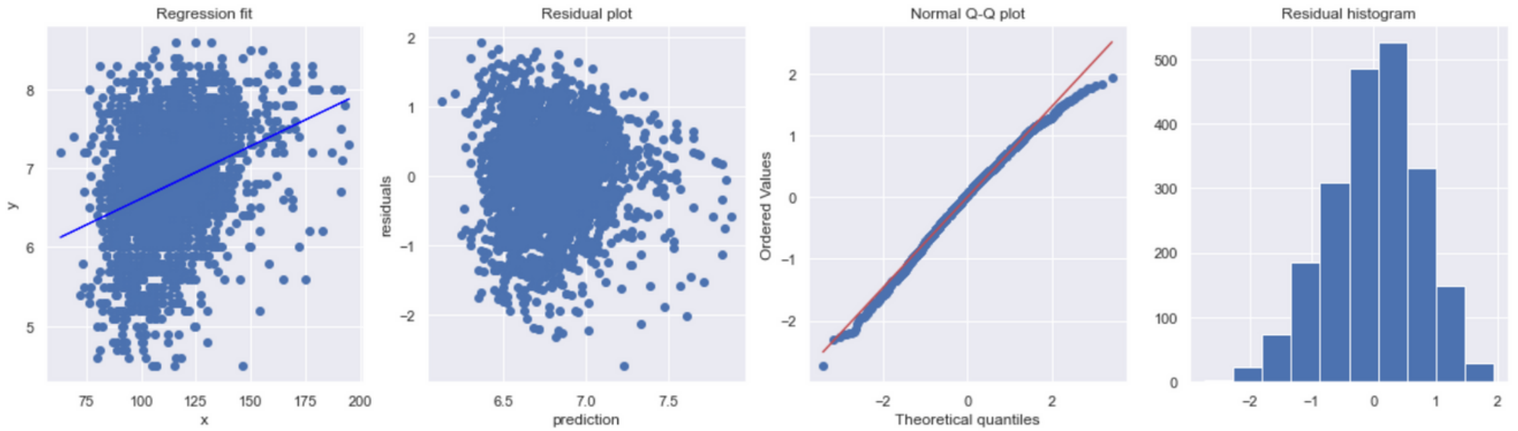
VIF

VIF > 2 indicate a problem...

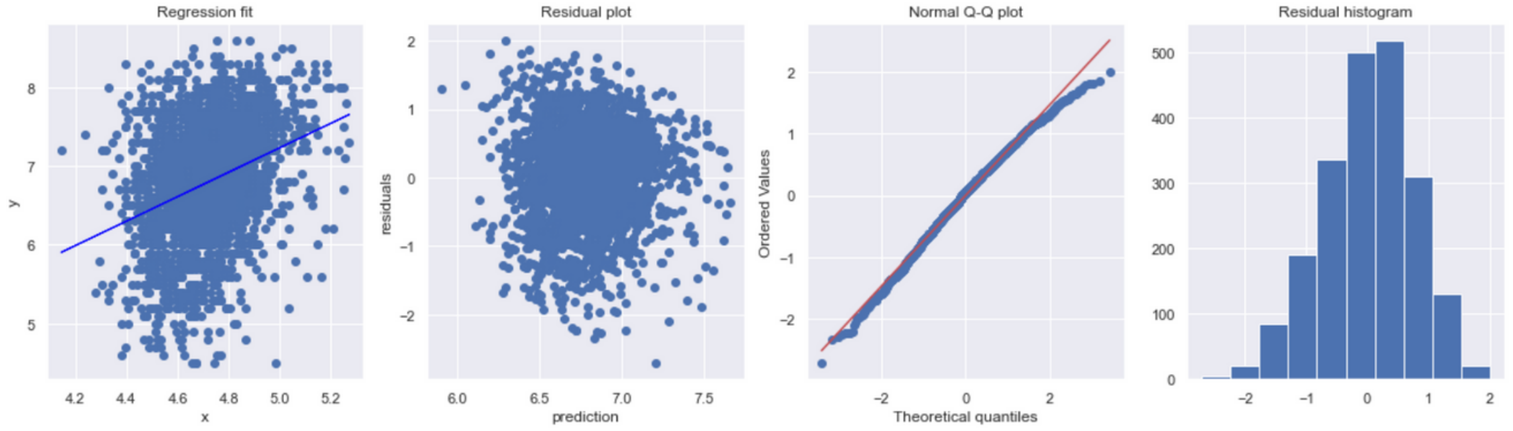
runtime_min	1.714873
votes_norm	1.558916
metascore	1.478680
budget_min	[illegible]
PG	6.324398
PG-13	10.173952
R	11.058830
Action	2.147556
Adventure	1.435907
Animation	2.306013
Biography	1.567688

Makes sense...
Genres are pretty closely linked to MPAA ratings...
So let's combine them!

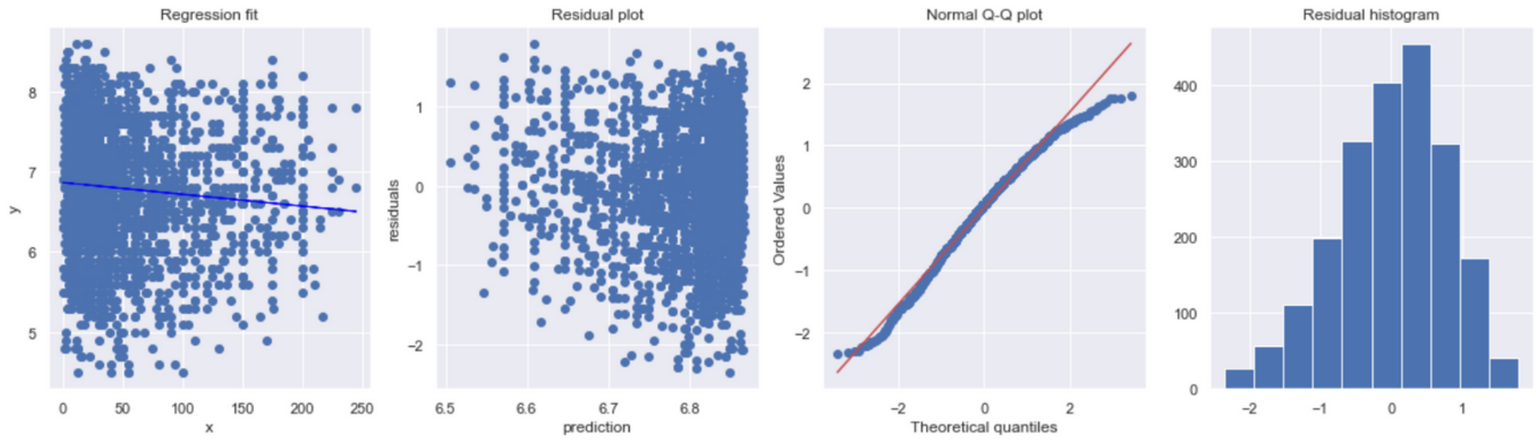
3 Transformation/fit analysis



Let's also take the natural log of **Runtime**



3 Transformation/fit analysis



Budget doesn't show a strong linear relationship, BUT it does contribute to our R2, so we'll keep it for the time being

3 Model 2

Although our R2 dropped, our RMSE has decreased and our test fit is comparably similar.

Base model		Model 2
Train		**Train**
Reg R2: 0.641 Ridge R2: 0.641 Lasso R2: 0.639		Reg R2: 0.633 Ridge R2: 0.633 Lasso R2: 0.631
Reg RMSE: 0.494 Ridge RMSE: 0.494 Lasso RMSE: 0.496		Reg RMSE: 0.476 Ridge RMSE: 0.477 Lasso RMSE: 0.479
Test		**Test**
Reg R2: 0.61 Ridge R2: 0.597 Lasso R2: 0.599		Reg R2: 0.616 Ridge R2: 0.597 Lasso R2: 0.593

3

VIF

variables	VIF
votes_norm	1.506289
metascore	1.401181
budget_mil	1.862259
PG Animation	1.425447
PG-13 Action	1.606551
PG-13 Comedy	1.175599
PG-13 Drama	1.139645
R Action	1.282247
R Biography	1.195993
R Comedy	1.226174
R Crime	1.178101
R Drama	1.278362
David Fincher	1.027921
Paul Thomas Anderson	1.033238
Peter Jackson	1.015052
Quentin Tarantino	1.038955
Sam Mendes	1.013539
Wes Anderson	1.013629
runtime_log	1.468658

Our VIF's are looking healthy, which means we can trust our coefficients much more now!

4

Let's try to do better with:

An interaction variable

Average rating between each director, writer and star combo.

Decreasing dimensionality

13 features, based on T-values and p values:

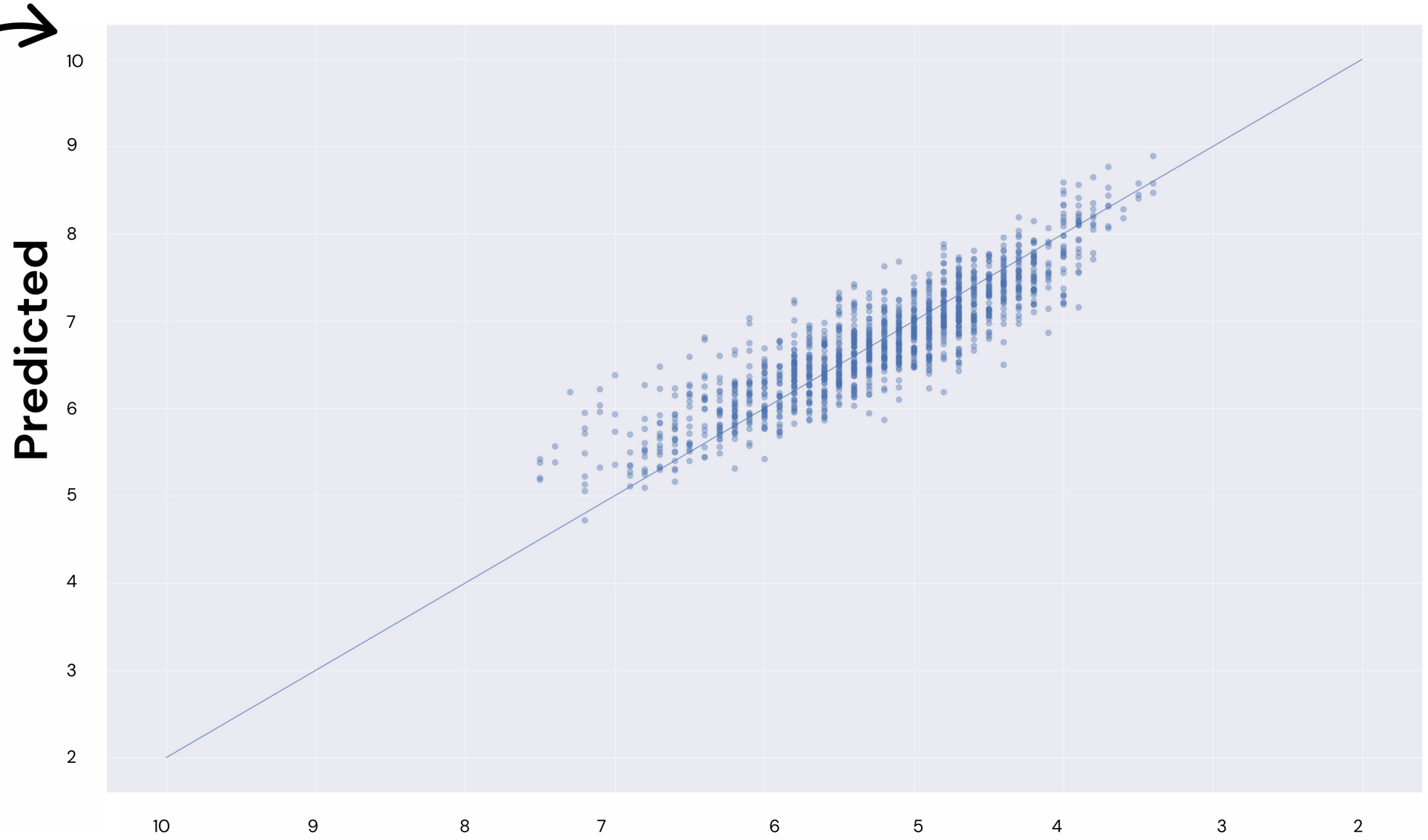
votes, metascore, budget_mil, PG Animation, PG_13 Action, PG-13 Comedy, R Comedy, R Drama, Paul Thomas Andrerson, Peter Jackson, Quentin Tarantino, runtime_log, dir_writer_star_mean

4 Final Model: R2 training of 0.783

Reg R2: 0.783
Ridge R2: 0.783
Lasso R2: 0.781

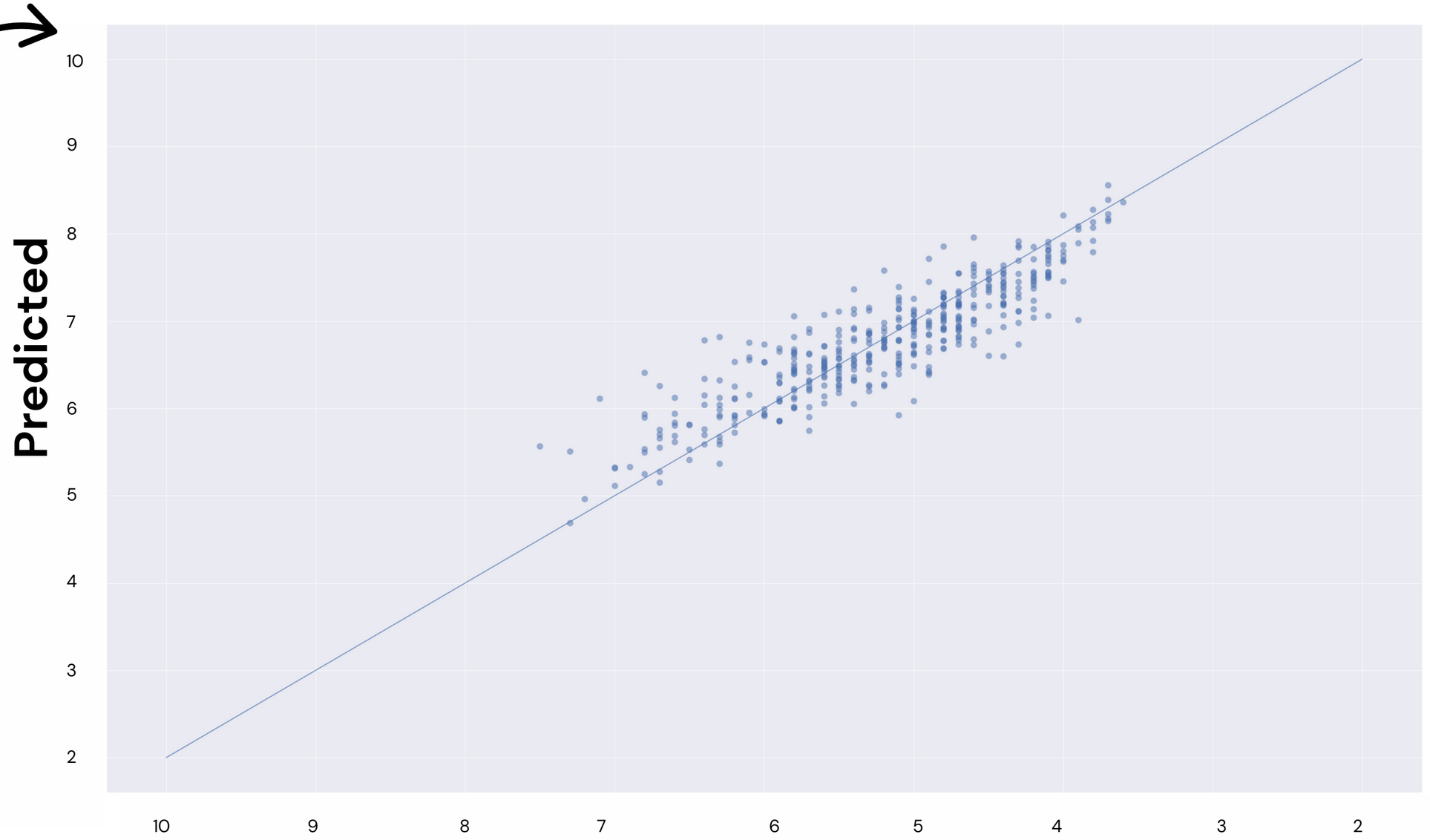
Reg RMSE: 0.372
Ridge RMSE: 0.373
Lasso RMSE: 0.379

KFold cross-validation



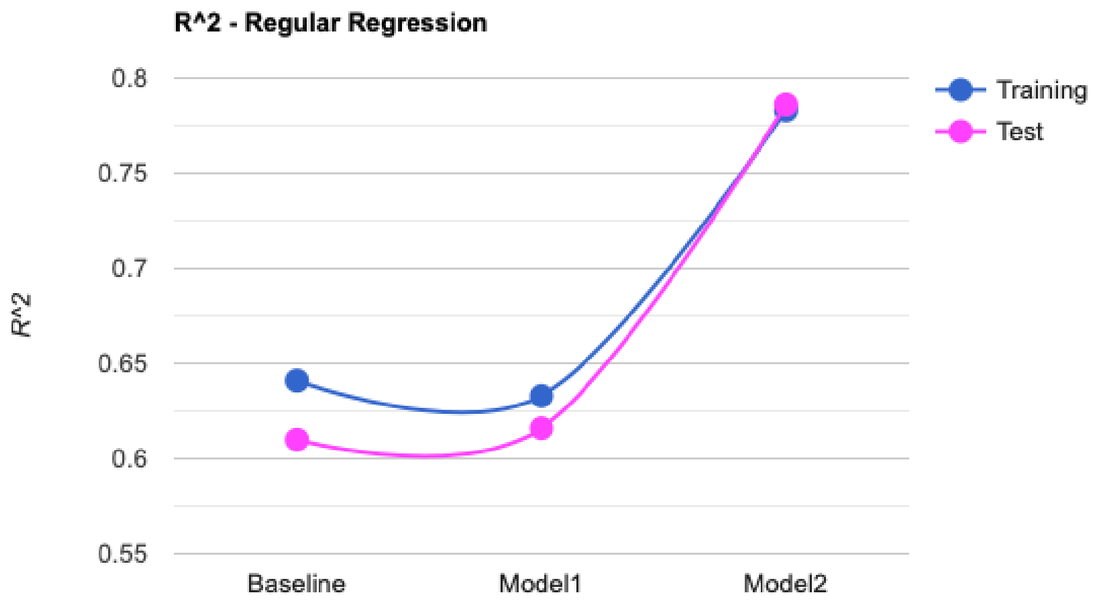
4 Final Model: R2 test of 0.787

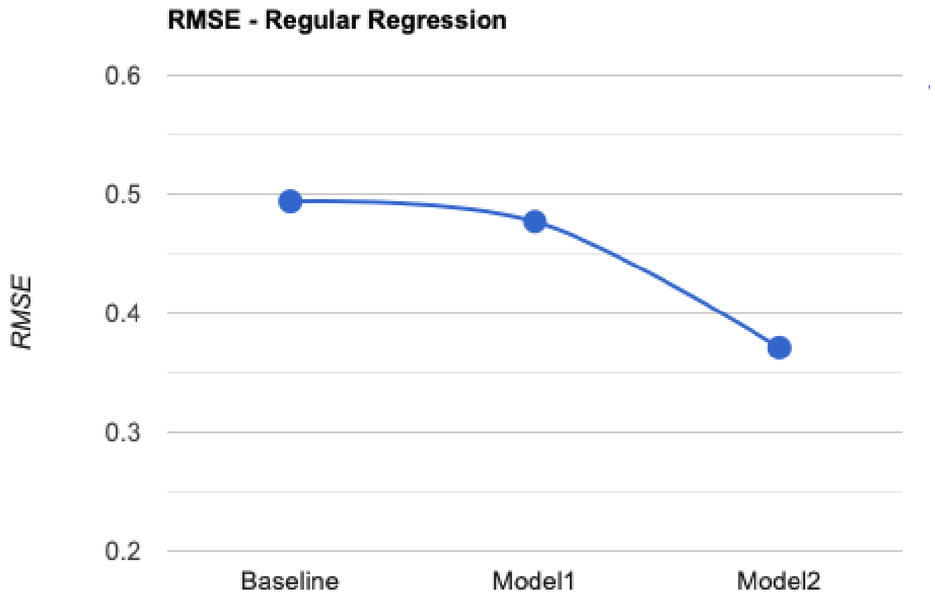
Reg R2: 0.787
Ridge R2: 0.785
Lasso R2: 0.786



4

R2's and RMSE's progress





4

Model comparison

Base Model	Model 2	Final Model
Train	**Train**	**Train**
Reg R2: 0.641 Ridge R2: 0.641 Lasso R2: 0.639	Reg R2: 0.633 Ridge R2: 0.633 Lasso R2: 0.631	Reg R2: 0.783 Ridge R2: 0.783 Lasso R2: 0.781
Reg RMSE: 0.494 Ridge RMSE: 0.494 Lasso RMSE: 0.496	Reg RMSE: 0.476 Ridge RMSE: 0.477 Lasso RMSE: 0.479	Reg RMSE: 0.372 Ridge RMSE: 0.373 Lasso RMSE: 0.379
Test	**Test**	**Test**
Reg R2: 0.61 Ridge R2: 0.597 Lasso R2: 0.599	Reg R2: 0.616 Ridge R2: 0.597 Lasso R2: 0.593	Reg R2: 0.786 Ridge R2: 0.785 Lasso R2: 0.786

4 Interpretation

votes_norm	0.0012
metascore	0.0118
runtime_log	0.4201
dir_writer_star_mean	0.7240
PG Animation	0.1807
Peter Jackson	0.1318

Votes: 1,000 increase is 0.12 increase in rating

Metascore: 10 point increase is 0.118 increase in rating

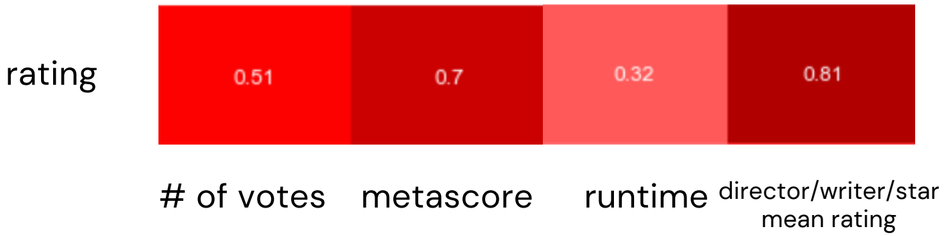
Runtime: 1% increase is 0.4201 increase in rating.

Director/writer/star mean rating: One point increase is 0.73 increase in rating.

PG Animation: this combo is 0.18 increase in rating.

Peter Jackson: this director is 0.13 increase in rating.

The strongest correlations:



4

"The Power of The Dog" test

Predicted

7.4

Actual

7.0

Thank you!