Predicting Video Popularity on YouTube

Using Regression Analysis to Predict Interactions on Digital Media

Intro

- Video ads for small and large brands
- Pay-per-view
- “Brand-Safe”?
- Video Popularity = Ad Engagement

How to predict popularity?

Methodology

> Data

YouTube Video Metadata (How-To & Tutorials Category)

> Algorithm

Linear Regression

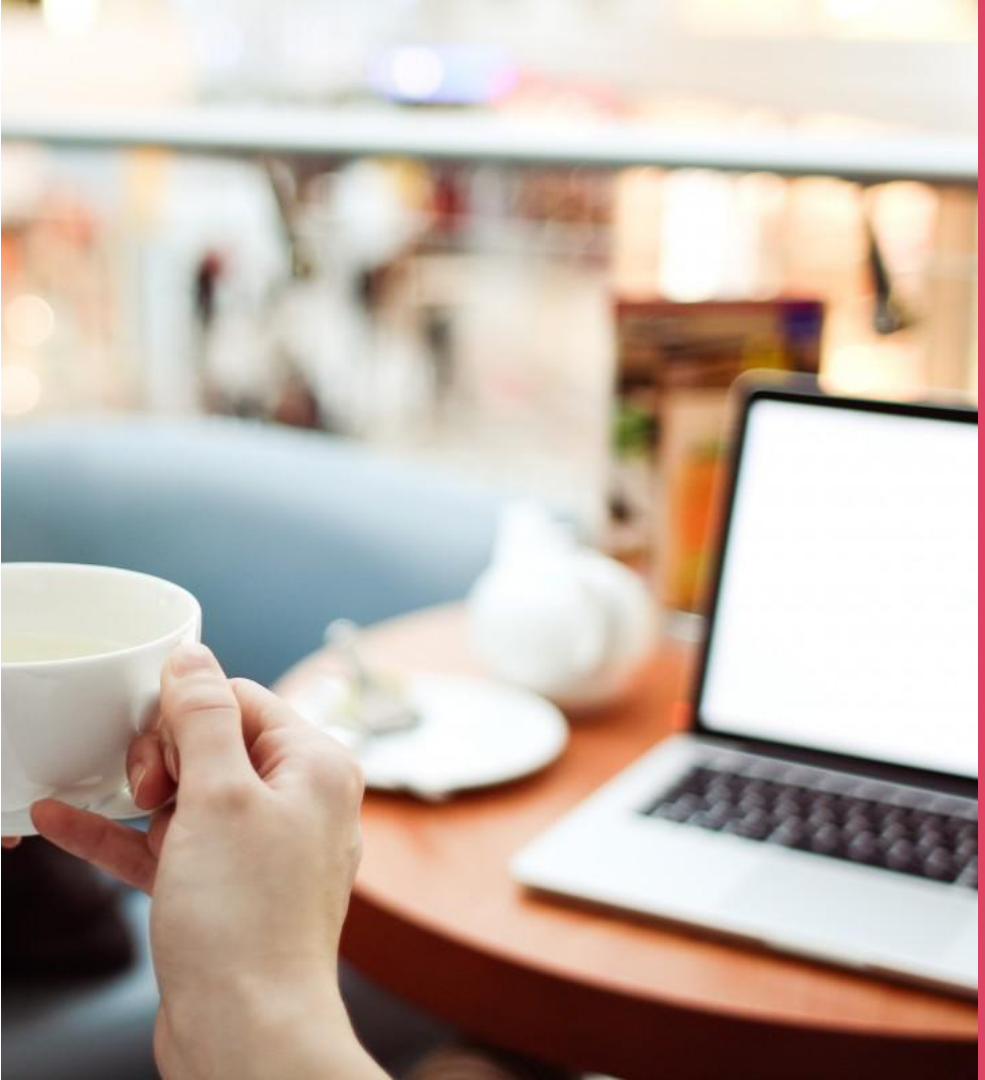
Gather & Clean Data → Transform Data & Select Features → Train & Evaluate Model

TOOLS:

pandas

scikit learn

matplotlib



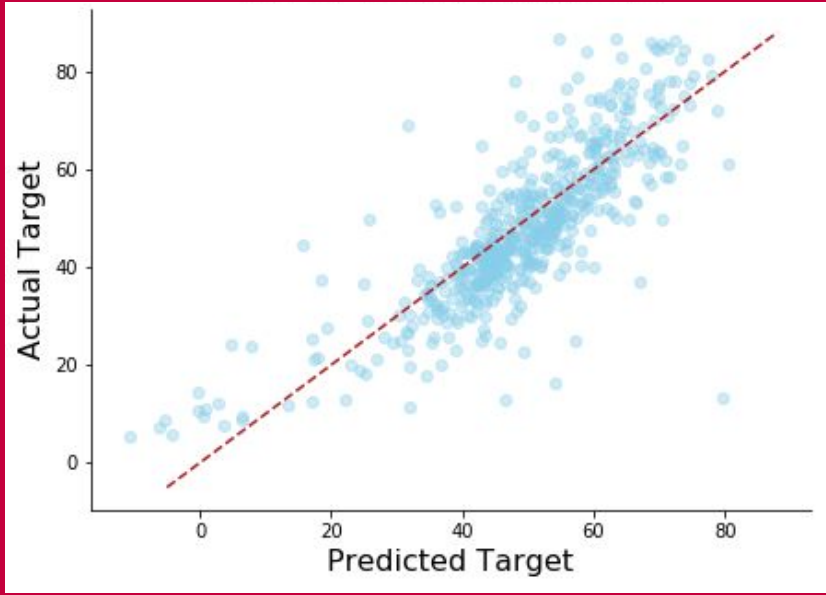
Feature & Model Selection

OLS Model	Age	Title Length
Last Likes	Duration	Title Score
10th Likes	Family Friendly	Tags

Results

MODEL PERFORMANCE

View Count Test Predictions



> Mean Absolute Error

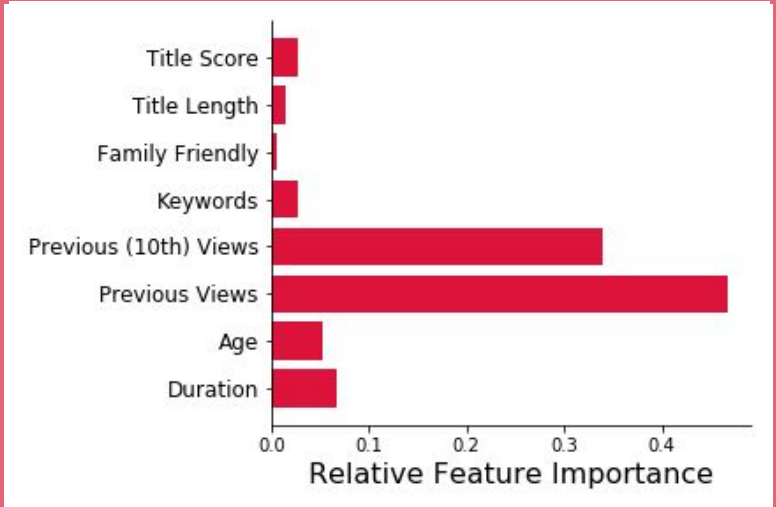
190,022 Views

(Average Error)

Results

FEATURE IMPORTANCE

Views Prediction Features



> 1 view (latest video)→

2 views on the next

> 1 view (10th newest video)→

1 view on the next

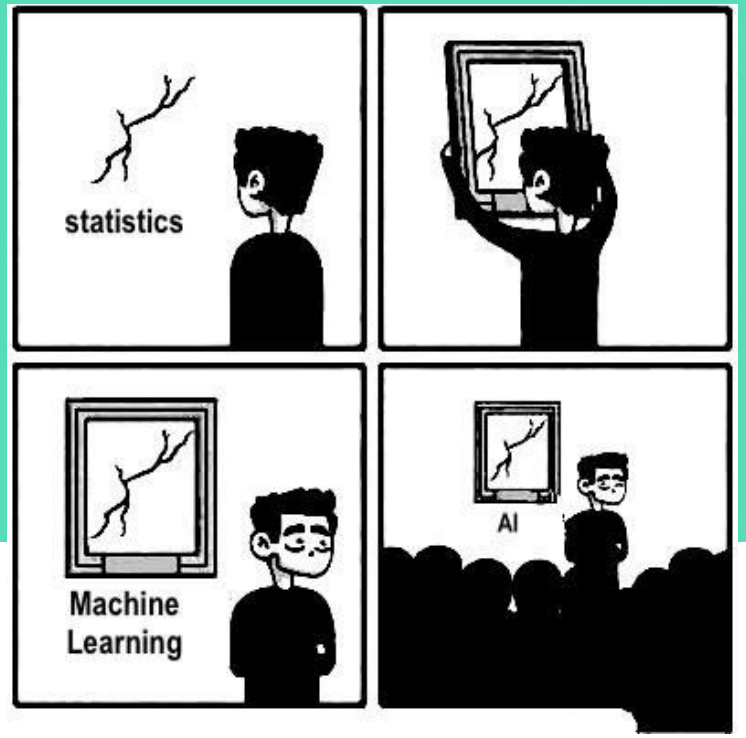
In Conclusion...

- Success breeds success
- Majority of views within first week
- Identify up-and-coming channels

Future Work

- Early View Data
- Growth Statistics
- Other platforms

Questions?



APPENDIX

SOURCES

https://mediakix.com/blog/most-popular-youtube-videos/#gs.cd44w2

https://www.youtube.com/ads/

https://www.thedrum.com/news/2019/03/05/google-says-youtube-might-never-be-100-brand-safe

https://www.kaggle.com/datasnaek/youtube-new

https://support.google.com/youtube/answer/2454017?hl=en

https://www.businessinsider.com/millennials-skip-youtube-ads-and-thats-ok-2017-1

https://www.thinkwithgoogle.com/consumer-insights/online-video-shopping/

https://pdfs.semanticscholar.org/7dad/e77c5a6c58ec2543ea10ed499395957fbcf4.pdf

Summary Statistics

OLS Regression Results			
Dep. Variable:	y	**R-squared:**	0.680
Model:	OLS	**Adj. R-squared:**	0.679
Method:	Least Squares	**F-statistic:**	584.4
Date:	Wed, 09 Oct 2019	**Prob (F-statistic):**	0.00
Time:	17:25:48	**Log-Likelihood:**	-8045.2
No. Observations:	2210	**AIC:**	1.611e+04
Df Residuals:	2201	**BIC:**	1.616e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	49.3531	0.197	251.110	0.000	48.968	49.739
duration	1.1383	0.208	5.470	0.000	0.730	1.546
age	0.8994	0.202	4.456	0.000	0.504	1.295
log_prev_views	8.0288	0.313	25.655	0.000	7.415	8.642
log_prev10_views	5.8254	0.312	18.680	0.000	5.214	6.437
tags	-0.4586	0.199	-2.304	0.021	-0.849	-0.068
family_friendly	-0.0958	0.197	-0.485	0.627	-0.483	0.291
title_length	0.2437	0.201	1.213	0.225	-0.150	0.638
title_score	-0.4798	0.198	-2.429	0.015	-0.867	-0.092

Omnibus:	303.051	**Durbin-Watson:**	2.032
Prob (Omnibus):	0.000	**Jarque-Bera (JB):**	1067.922
Skew:	0.660	**Prob(JB):**	1.27e-232
Kurtosis:	6.139	**Cond. No.**	2.99

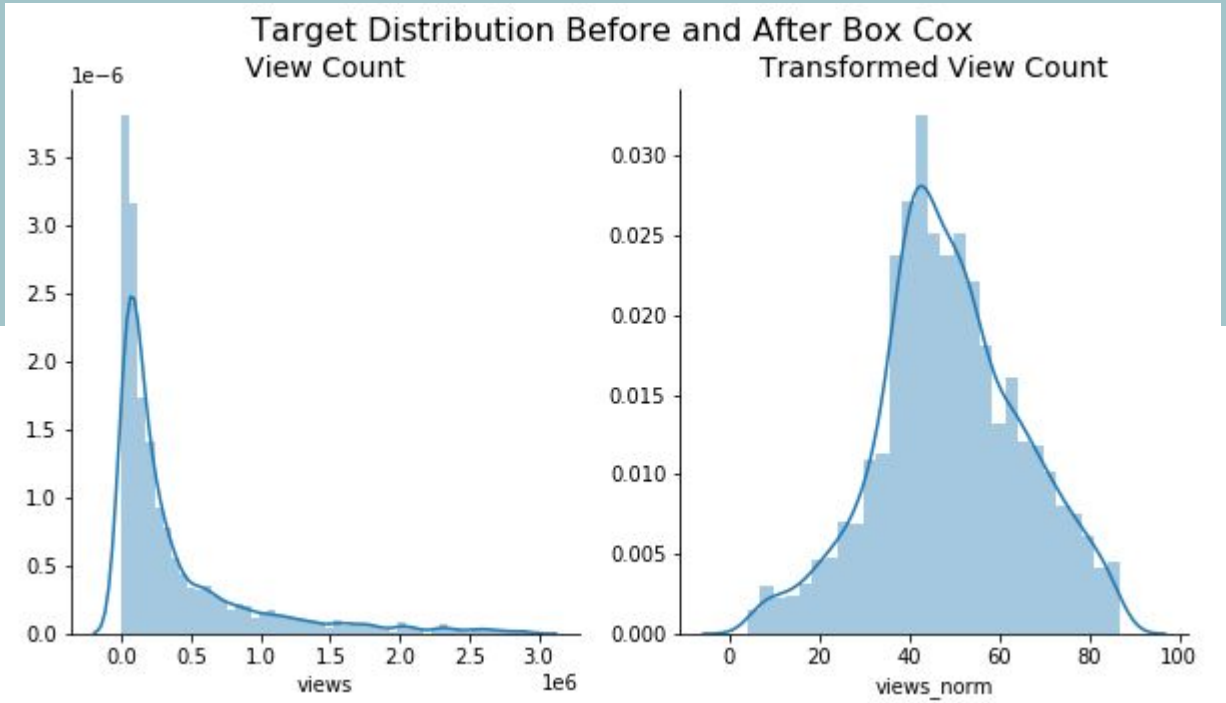
Model Comparison

Linear Regression Model Performance Comparison				
Model Type	R^2 (in-sample)	R^2 (out-of-sample)	Mean Absolute Error	Root Mean Squared Error
OLS	0.6725	0.6464	190,022.26	379,676.63
Poly Feat	0.6935	0.6857	192,563.06	401,409.30
LASSO	0.6799	0.6464	190,010.94	379,686.28
Ridge	0.6799	0.6475	189,772.55	379,917.50
Elastic Net	0.6799	0.6475	189,757.81	379,865.06

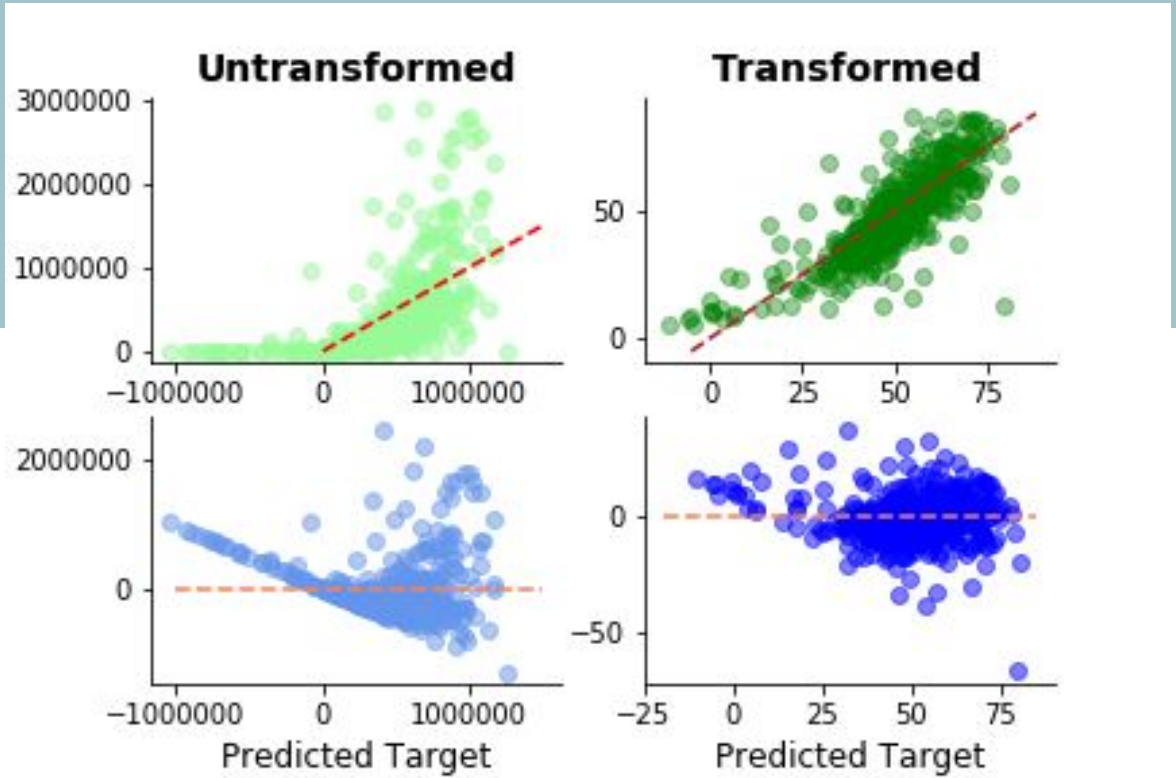
Regularization Cross-Validation Tuning Results

Tuned Regularization Hyperparameters		
Model	**Alpha (Lambda)**	**l1_ratio**
LASSO	0.00296	--
Ridge	32.92526	--
Elastic Net	0.015204	0.1

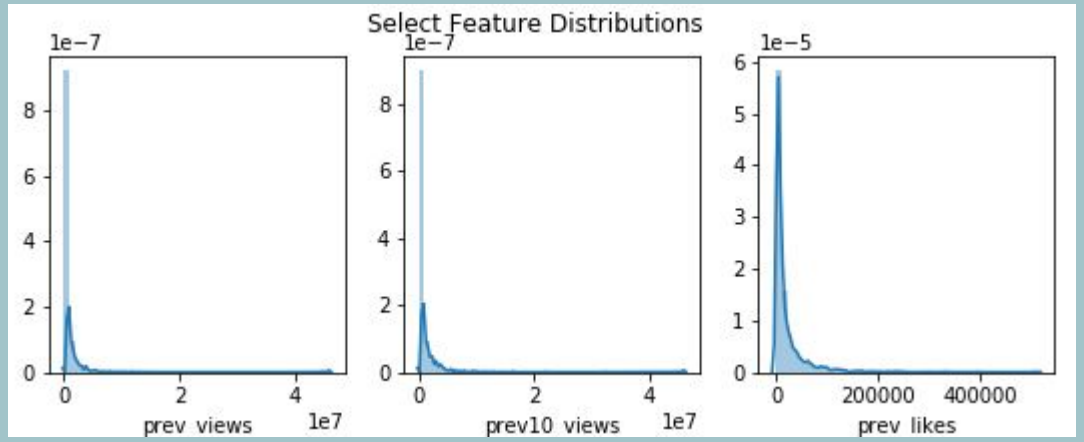
Box Cox

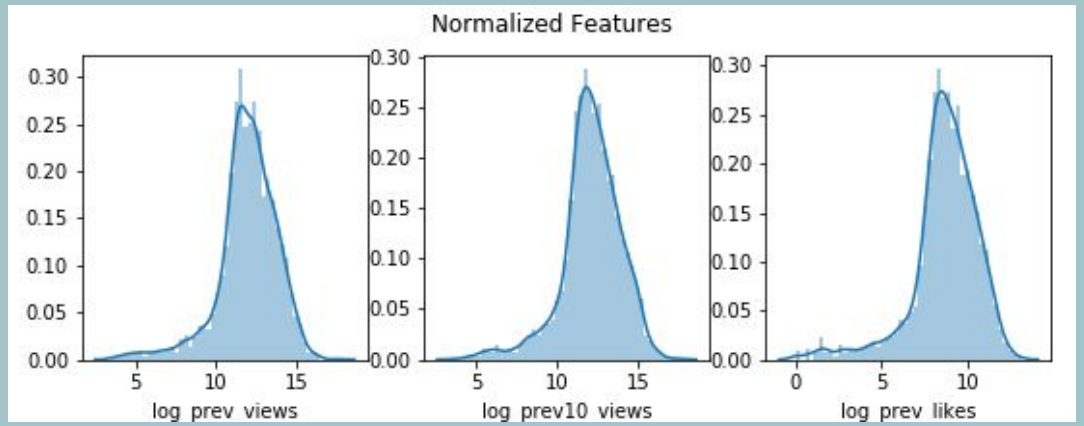

Target Distribution Before and After Box Cox
View Count
1e−6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
views
Transformed View Count
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
views_norm

Base Case vs Box Cox



FEATURE ENGINEERING: Log Transforms of Skewed Features





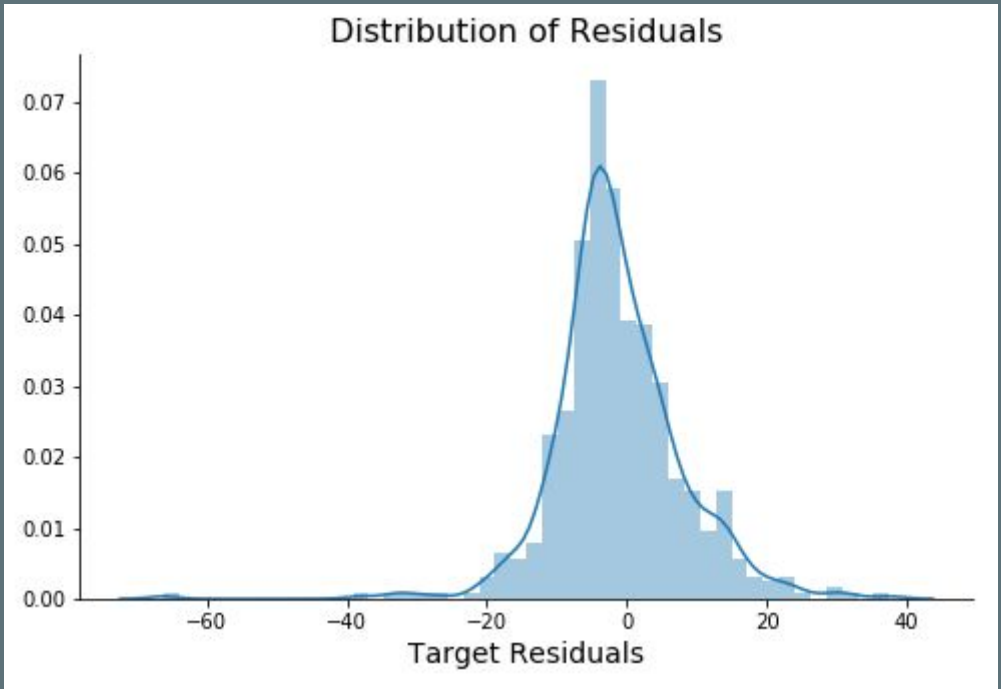
Assumption of Linear Regression #1:

Regression is linear in parameters & correctly specified

	coef
intercept	49.3531
duration	1.1383
age	0.8994
log_prev_views	8.0288
log_prev10_views	5.8254
tags	-0.4586
family_friendly	-0.0958
title_length	0.2437
title_compound_score	-0.4798

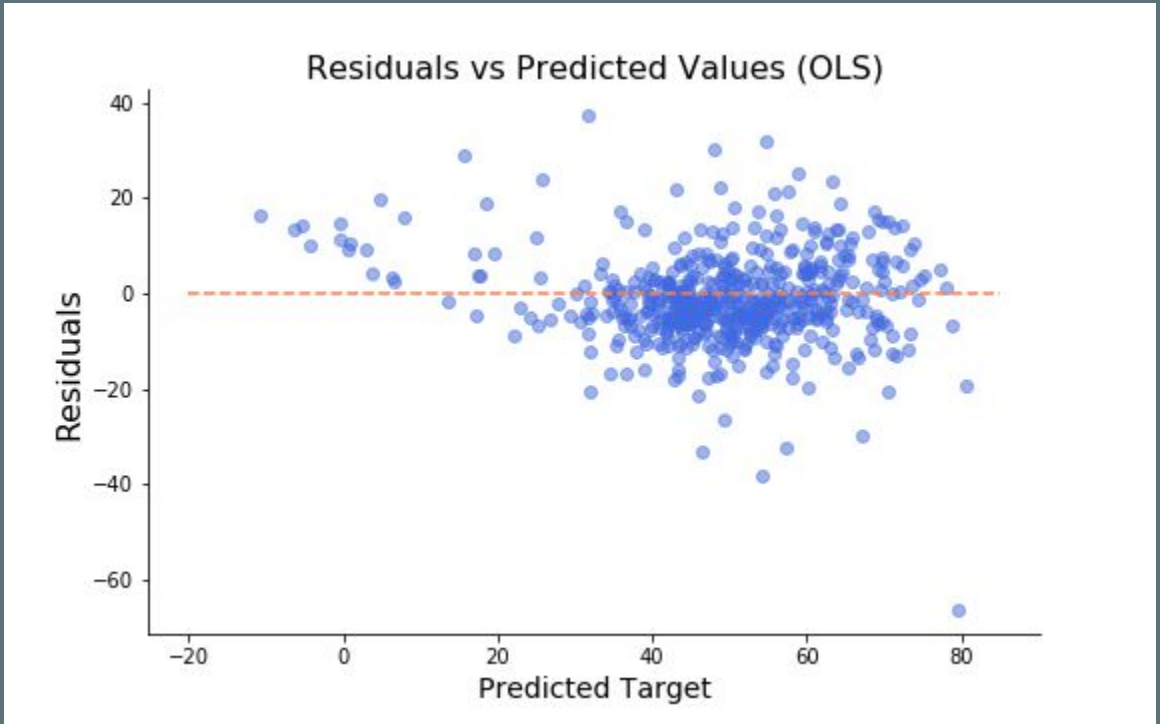
Assumption of Linear Regression #2:

Error terms are normally distributed and zero population mean



Assumption of Linear Regression #3:

The error term has constant variance (homoscedastic)



Assumption of Linear Regression #4:

Errors are uncorrelated across observations

OLS Regression Results			
Omnibus:	303.051	**Durbin-Watson:**	2.032
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1067.922
Skew:	0.660	**Prob(JB):**	1.27e-232
Kurtosis:	6.139	**Cond. No.**	2.99

Assumption of Linear Regression #5:

No independent variable is a perfect linear function of any other independent variable

OLS Regression Results			
Omnibus:	303.051	**Durbin-Watson:**	2.032
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1067.922
Skew:	0.660	**Prob(JB):**	1.27e-232
Kurtosis:	6.139	**Cond. No.**	2.99

	VIF Factor	features
0	1.000	intercept
1	1.121	duration
2	1.054	age
3	2.535	log_prev_views
4	2.518	log_prev10_views
5	1.026	tags
6	1.008	family_friendly
7	1.045	title_length
8	1.010	title_compound_score