Predicting Popularity of an Online News Article

(or not)

by Joshua Szymanowski

Background on the data

SHARES

vs.

Mashable

- Genre
- Length of article, title, avg. word
- Day of publication
- Sentiment (positivity/negativity)
- Subjectivity (fact/opinion)
- SEO metadata (keywords)

- 39,797 Mashable articles
- Data from January 2013 to December 2014
- 58 predictive variables

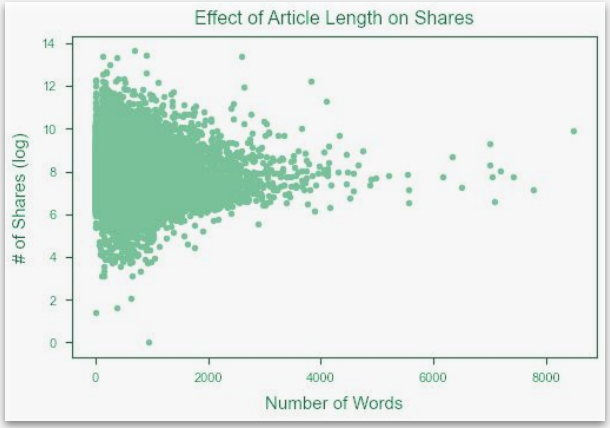
Preview of struggles to come

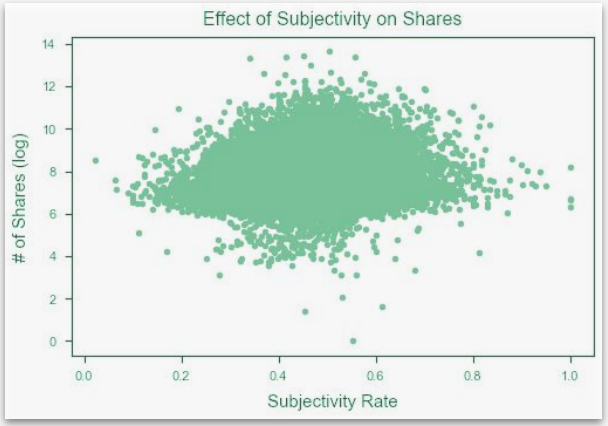
	Untouched	Coefficient scaling	Eliminate extremes	Remove high p-values	All interactions	Top interactions
R^2 value	0.123	0.116	0.134	0.133	0.166	0.152

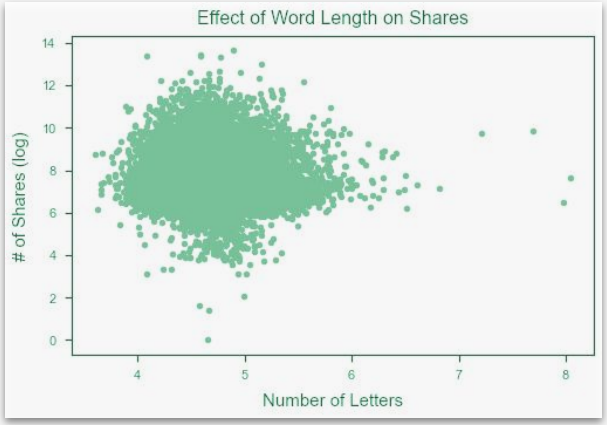
	# of Features	Training RMSE	Testing RMSE
Baseline	49	1,975	1,973
F-Test	8	1,877	1,868
Lasso	8	1,813	1,809

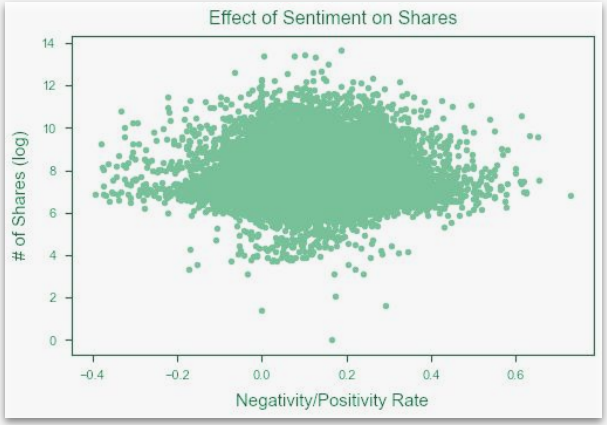
vs. Mean = 3395

Can we tell writers how to write?

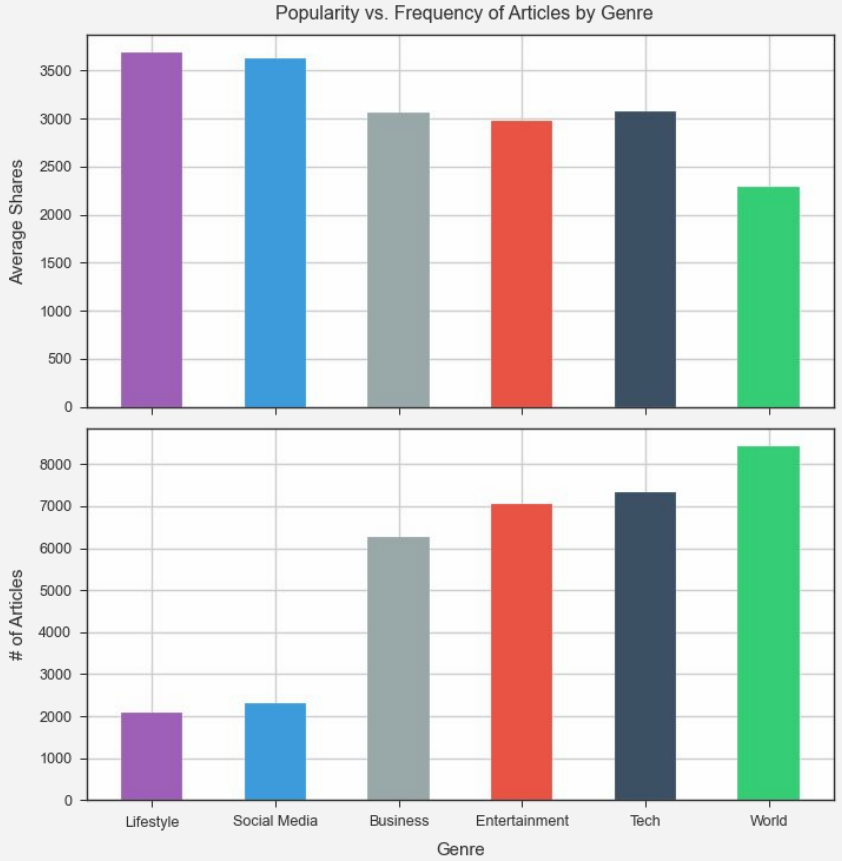






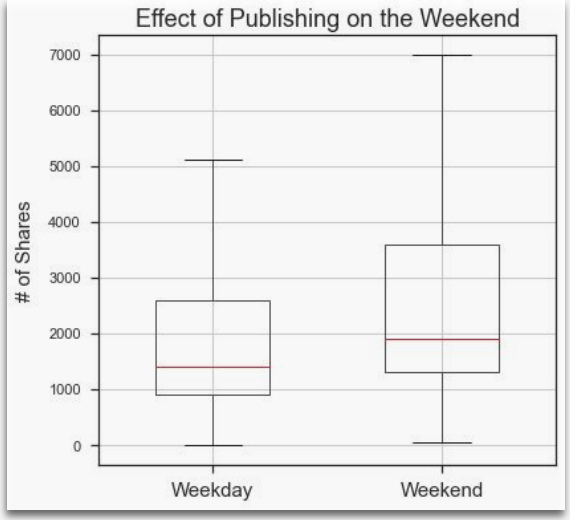


Important discoveries



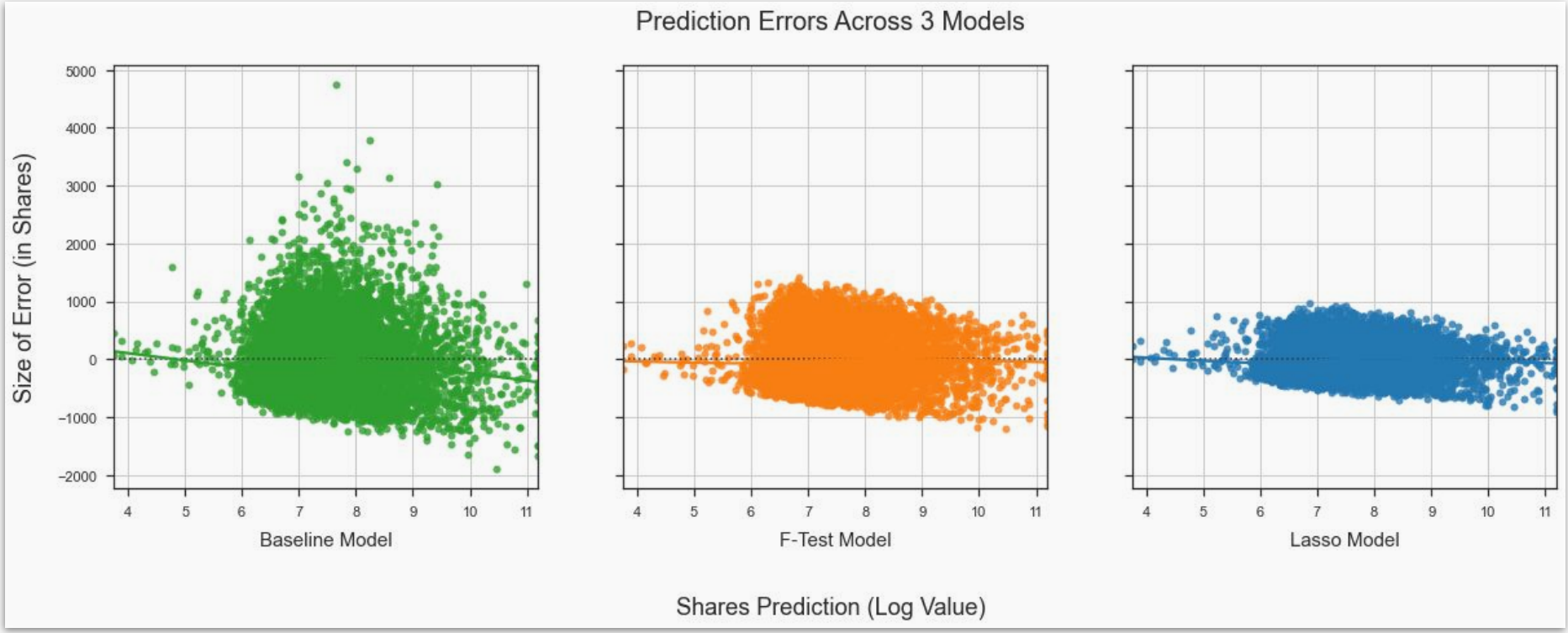
- Niche genres have more active readership.
- World news readers are less active.

- Weekend articles are more popular.
  - No social media at work?
  - More time to read on weekends?
  - Sharing weekend articles during the week?



A closer look at model error

Lasso RMSE = 8 standard deviations



Training RMSE:	1,975	Training RMSE:	1,877	Training RMSE:	1,813
Testing RMSE:	1,973	Testing RMSE:	1,868	Testing RMSE:	1,809

Top variables and conclusions

Variable	Coefficient
Avg. shares of all articles with each keywords (log value)	0.528
Weekend = True	0.102
Genre = Entertainment	-0.085

- New models required
- Significant factors
  - Keywords
  - Genre/channel
  - Publishing on weekday/weekend

- Insignificant factors
  - Article length
  - Title length
  - Average word length
  - Sentiment
  - Subjectivity

Questions?