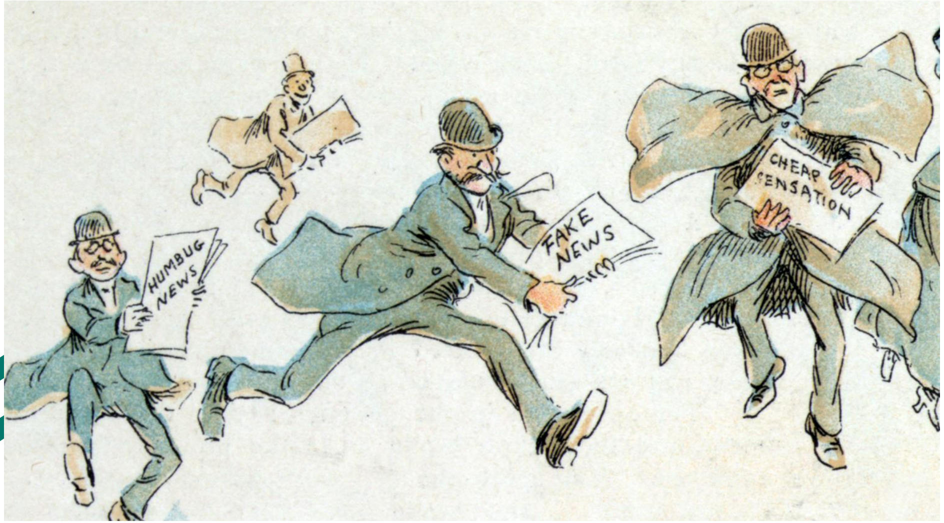
What Makes an Article Worthy to Share?

By Sam Raykhman



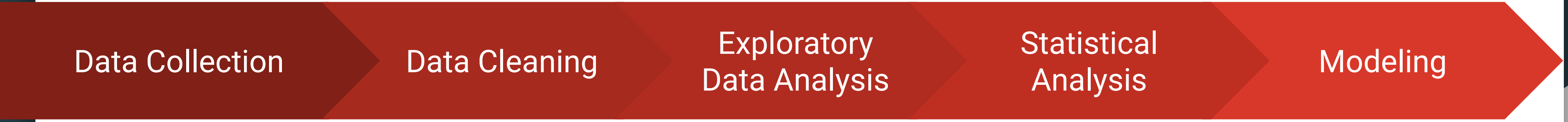
What Could the Data Tell Us?

I looked at the characteristics of 2 years worth of Mashable articles to try and find the factors that represent the amount of shares that the article gets.

I predict that “subjective” variables such as polarity of the text, number of images, or title sentimentality would have the greatest effect on the data.

Mashable
All That’s New on the Web

My Process



Data Collection

Data Cleaning

Exploratory Data Analysis

Statistical Analysis

Modeling

- I used the Online News Popularity dataset from the UCI Machine Learning Repository

- Removed non-predictive factors
- Removed just under 1,000 rows of articles that no longer exist

- Observed the relationships between factors and shares
- Graphed relationships of categorical factors to shares

- Used several analyses to observe correlation between features and target
- Analyzed variance to make sure all data is relevant

- Created polynomial features to create interaction terms since not a lot of my data is represented otherwise

Which to Choose?

There are 58 factors that can be used to predict the number of shares that an article on Mashable would receive. I assumed that variables such as the rate of positive words, the day the article was posted, or the genre of the article would be clear defining factors. But, the data proved otherwise…

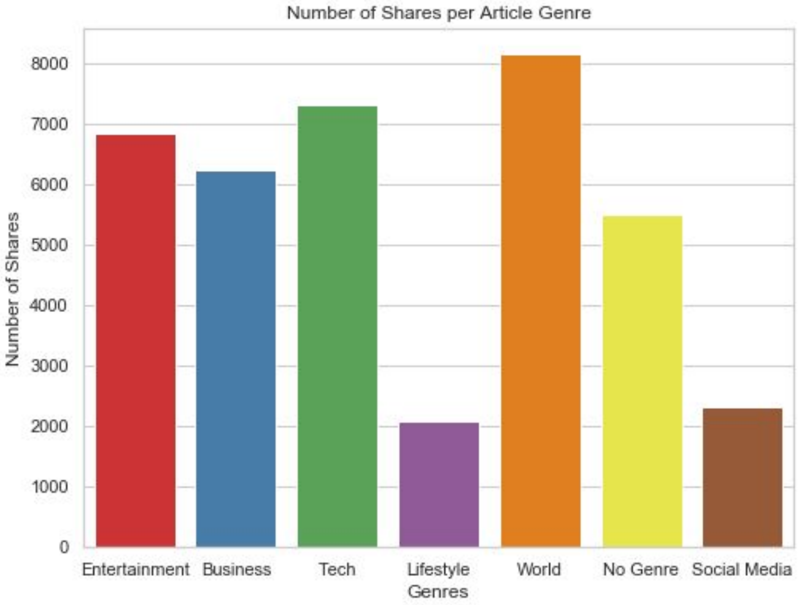
Which Model is Best?

	Training RMSE	Test RMSE	R^2
Standard	0.00202	0.00207	0.158
Lasso	0.15308	0.15587	0.073
Ridge	$7.35276 * 10^{-5}$	$7.57865 * 10^{-5}$	0.152

Examining Factors Closer

Even factors such as genre of the article, which I thought would heavily correlate to the amount of shares, ended up not not represent the data very well.

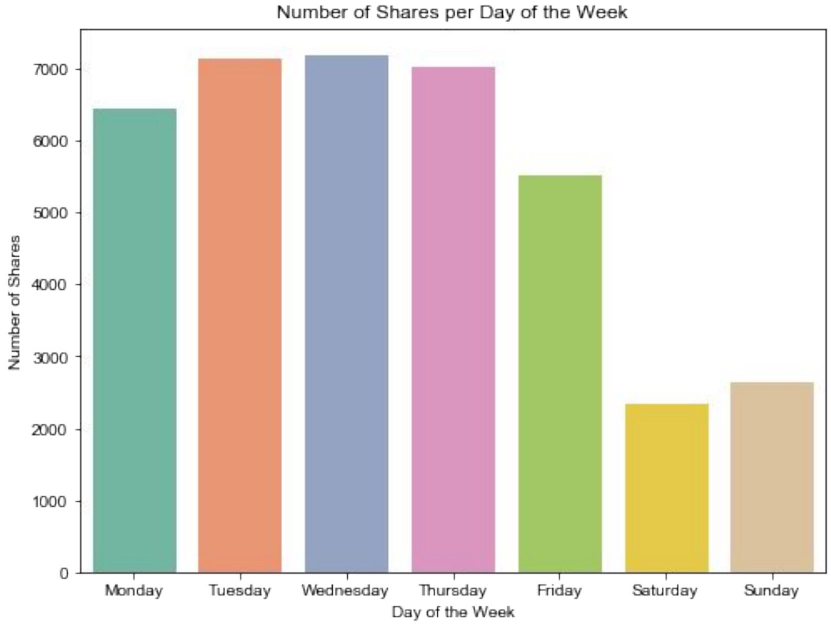
R^2 for OLS model = 0.008



Examining Factors Closer Cont.

Similarly to the genre representation, I assumed the day of the week would correlate heavily but the factors were worse represented than the genre factors.

R^2 for OLS model = 0.000



Takeaways

- I expected certain factors to have an effect on the data, but it turned out that all of my factors (even after creating interaction terms) did not represent the data well.

- Also, there are clear correlations of certain categorical variables between each other, so it's possible to look at the categorical variables compared to each other to determine which alters the target the most.

If I Had More Time

- I would not have dropped the non-predictive columns and the columns that didn’t make sense to me.
  - I would also try to have kept the rows for the articles that do not exist anymore and see what effect they would have on my data
- I would also have attempted to use time to categorize my variables further
  - For example, creating categories for season or month