Online News Popularity Data

Julian Murillo
Valeria Park

Data Collection

- UCI Machine Learning Repository
- Features about 39,644 articles published by Mashable in a period of 2 years
- 61 variables
- 58 predictive attributes, 2 non-predictive, 1 Target Variable (Article Shares)

All Potential Variables

- ***Shares (target variable)***
- Number of words in the title
- Number of words in the content
- Rate of unique words in the content
- Rate of non-stop words in the content
- Rate of unique non-stop words in the content
- Number of links
- Number of links to other articles published by Mashable
- Number of images
- Number of videos
- Average length of the words in the content
- Number of keywords in the metadata
- Data channel (dummies)
- Best keyword
- Worst keyword
- Average keyword
- Shares of referenced articles in Mashable
- Article publish day (dummies)
- Closeness to LDA topic 0-4
- Text subjectivity
- Text sentiment polarity
- Rate of positive/negative words
- Polarity of positive/negative words
- Title subjectivity
- Title polarity
- Absolute subjectivity level
- Absolute polarity level

Summary Statistics of Shares Variable

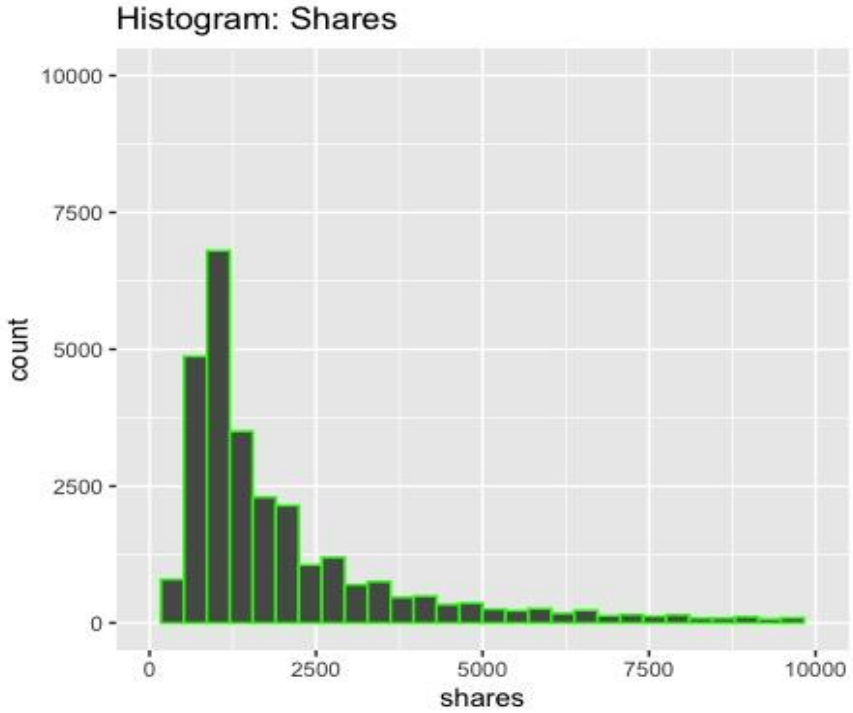
- Tried lasso, ridge, and elastic net models

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	946	1400	3395	2800	843300

- Standard Deviation: 11,626.95!!!
- Variance: 135,185,984!!

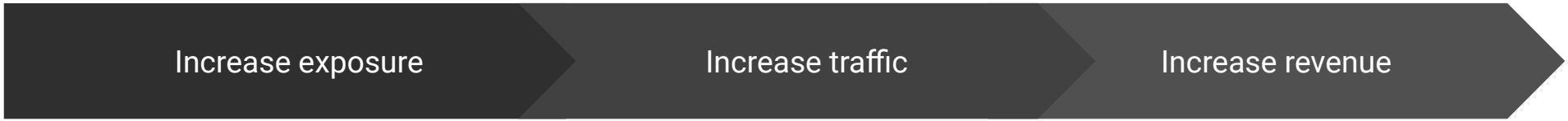
Omit Outliers?

- Difficult to achieve systematically
- Data skewed right and does not follow Normal Distribution



Business Objective

Find which features contribute to a higher number of article shares



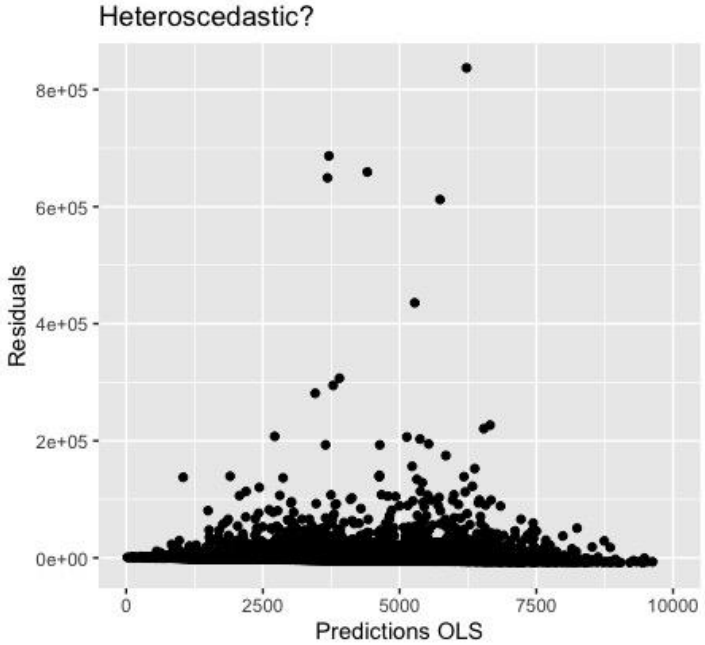
Variable Selection

Classification or OLS?

- Initially: OLS (Shares Target Variable)
- Ridge
  - Super Low Coefficients
- Lasso
  - Selected 0 variables
- Elastic Net
  - Selected 0 variables at optimal alpha of .7
- Created popular_article categorical variable
  - Popular_article if shares > 1400 (Median Shares Value)

Classification or OLS?

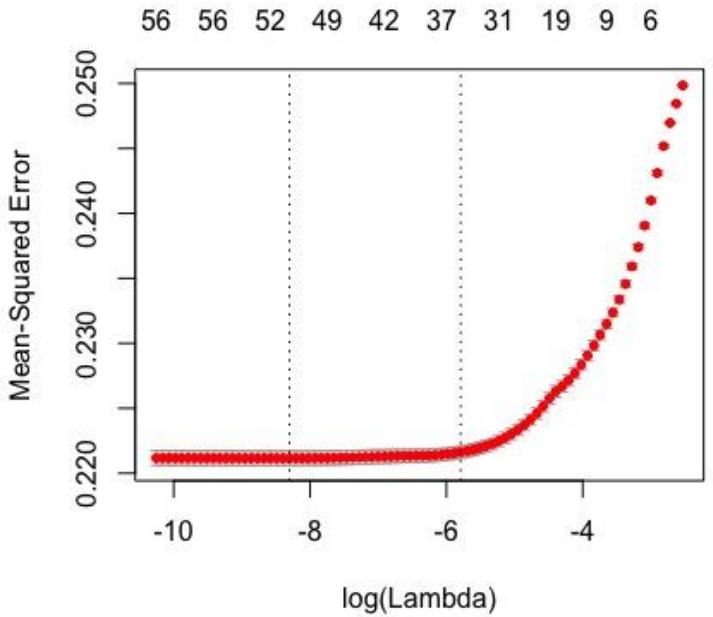
- Ran OLS on Training Set (With All Variables)
  - Output: Adjusted R-squared 2.172%
- Breusch Pagan Test Threshold alpha = .05
  - Output: p-value = 0.1932
  - Null: Homoscedastic
  - Alternative: Heteroscedastic
  - FAIL TO REJECT NULL



Variable Selection: Lasso

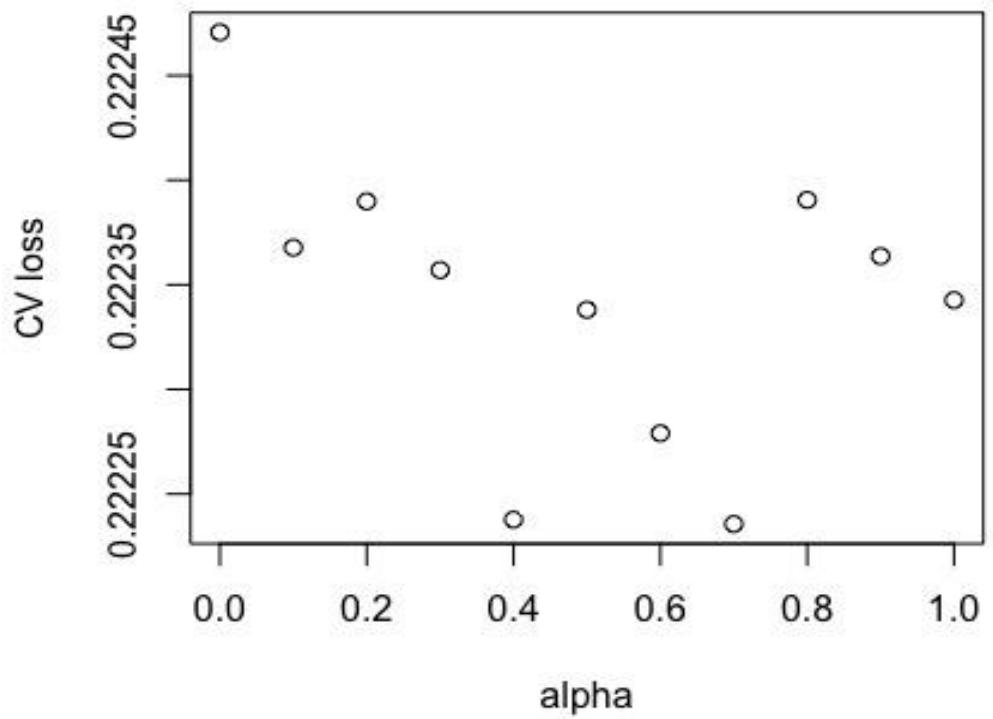
- Greatest Severity on Penalty Term (Lambda)
- Each model had similar misclassification errors

- Variables Selected:
  - 31/58



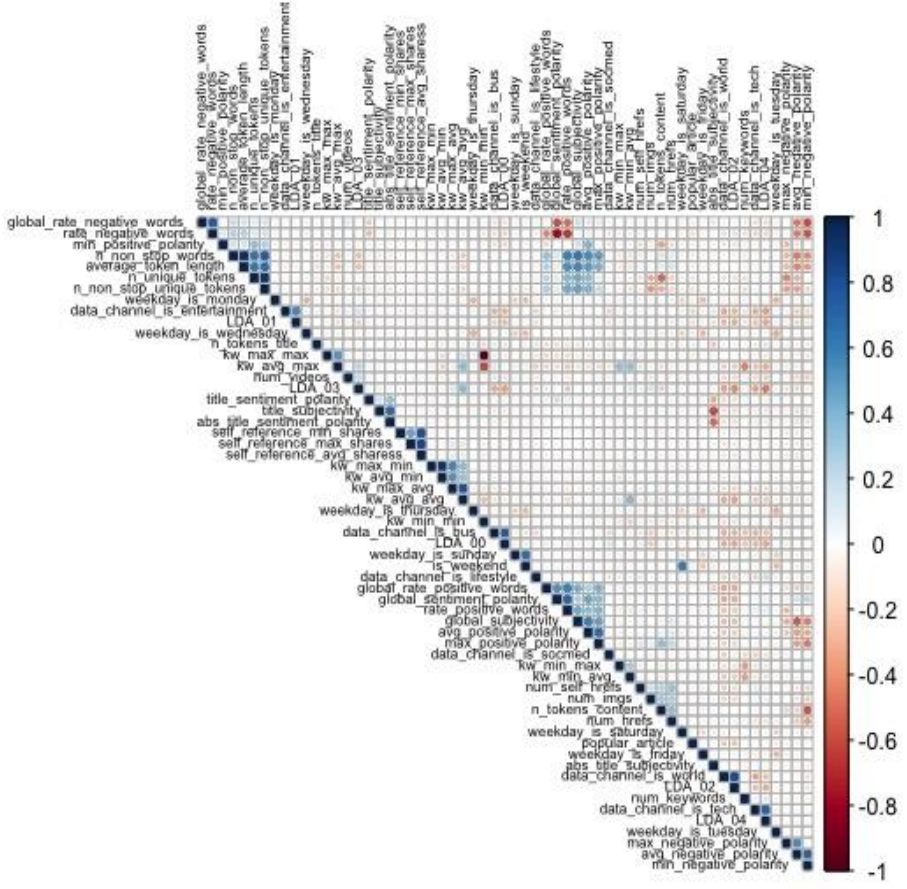
Variable Selection: Elastic Net

- Min C.V. Loss at Alpha = .7
- Variables Selected:
  - 34/58



Testing for Multicollinearity

- Correlation Matrix:
  - Difficult to tell
- VIF
  - Variance Inflation Factor
  - Values above 10 problematic
  - Removed High VIF scores from Variables Selected by Elastic Net



Selecting Variables: Final Variables

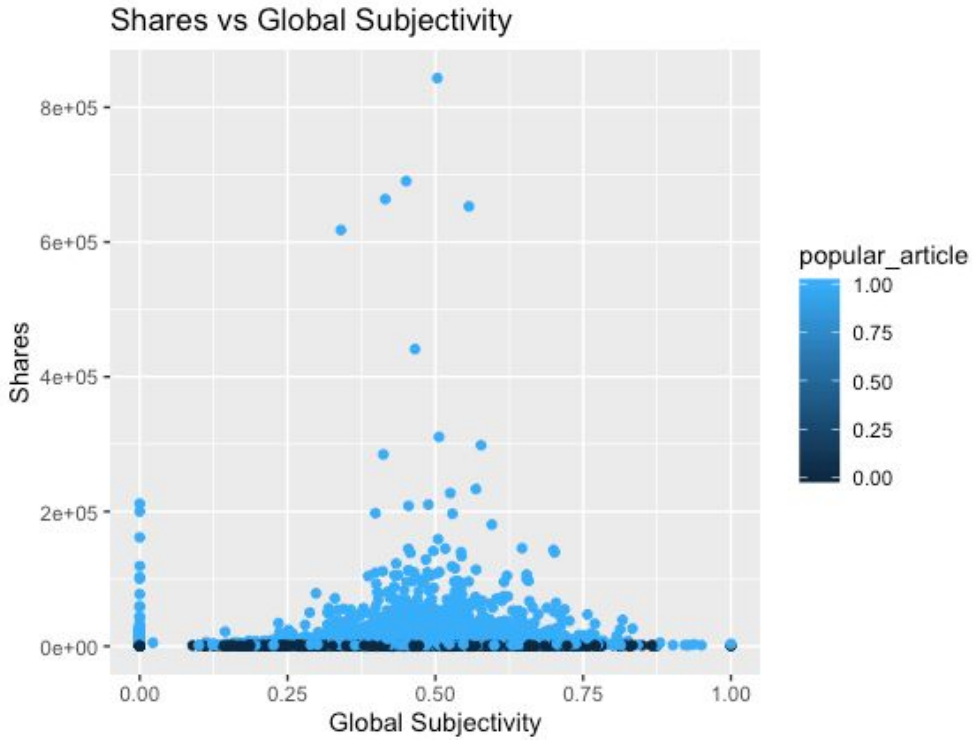
- Why filter so much?
  - We want a model that is as simple and accurate as possible (parsimony)
  - We want to reduce confounding in our model, we don't want to predict on noise or use variables that don't matter much

Selected Variables

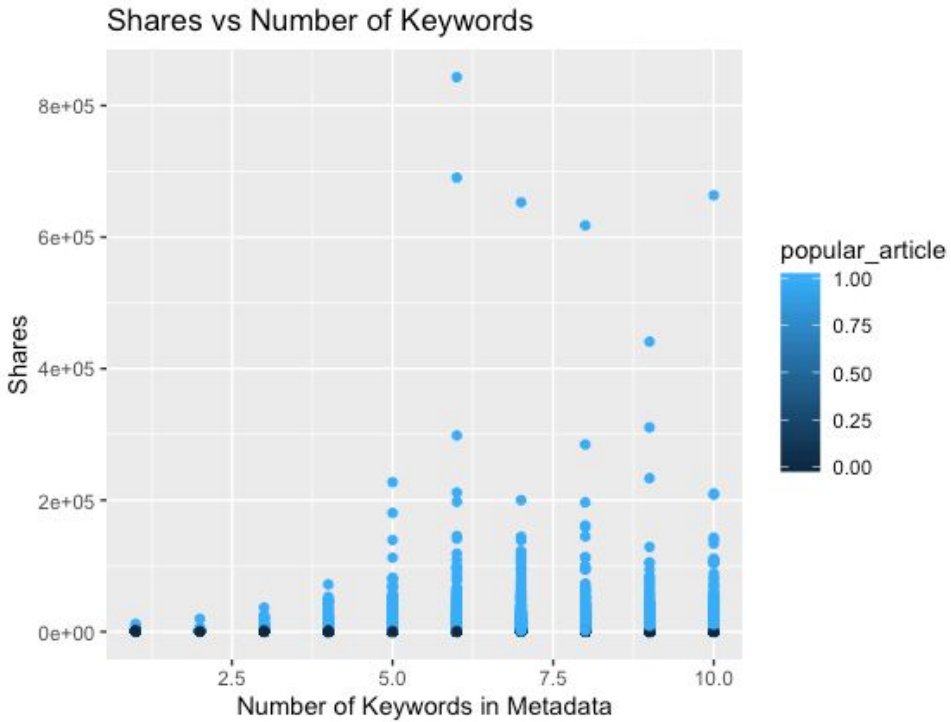
- Popular_article (1 or 0)
- shares
- num_hrefs (# of external links)
- num_self_hrefs (# of internal links)
- num_keywords
- data_channel_is_entertainment
- data_channel_is_bus
- data_channel_is_socmed
- data_channel_is_tech
- weekday_is_tuesday
- weekday_is_wednesday
- weekday_is_friday
- weekday_is_saturday
- is_weekend
- global_subjectivity (total subjectivity)
- min_positive_polarity (min positive words)
- title_subjectivity
- title_sentiment_polarity
- abs_title_subjectivity
- LDA_00
- LDA_01
- LDA_02
- LDA_04

Data Visualization

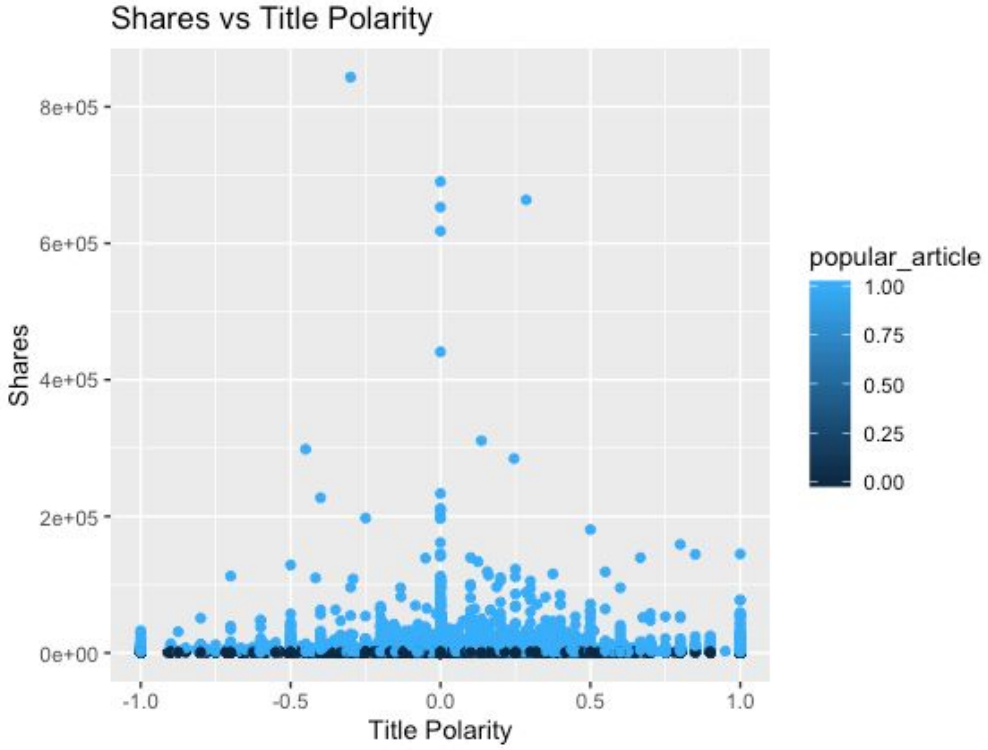
Global Subjectivity



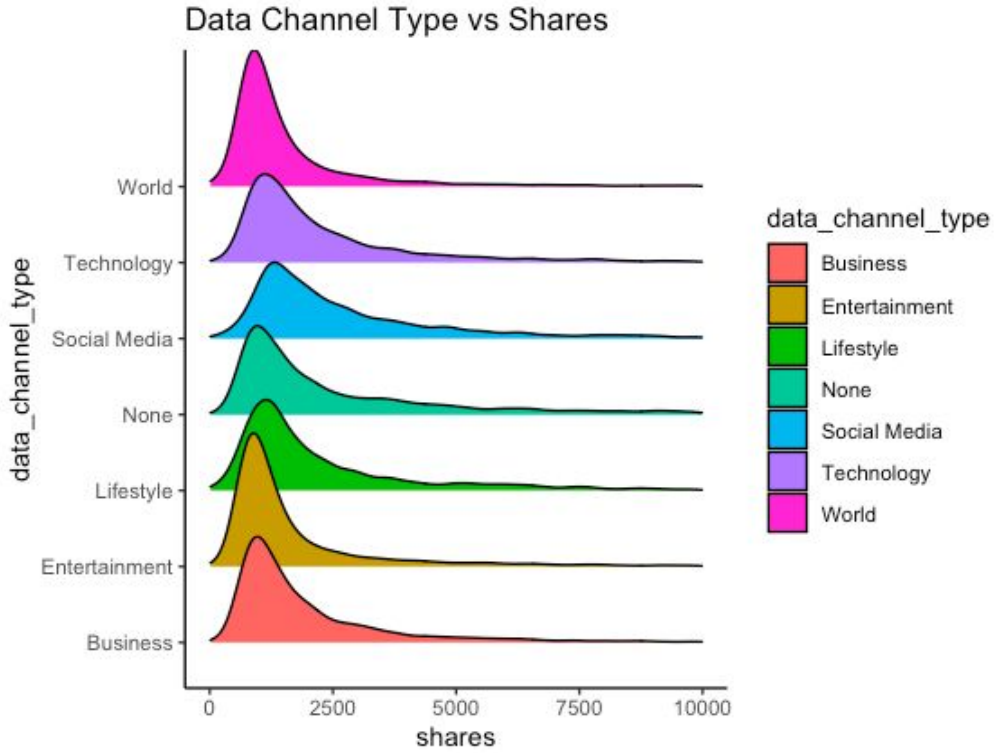
Number of Keywords



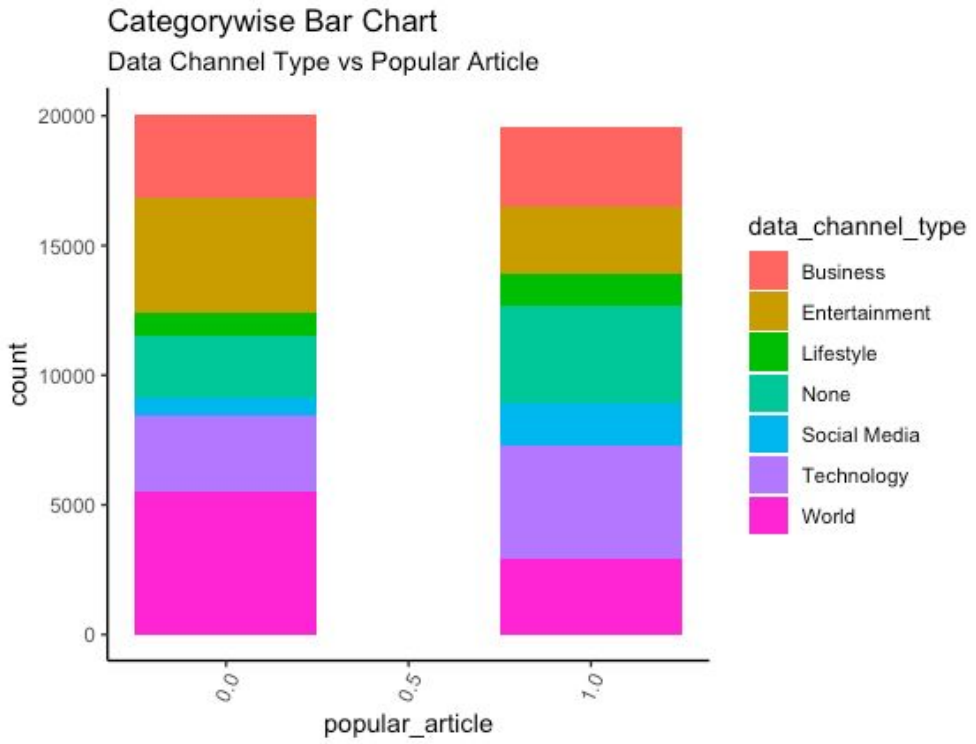
Title Polarity



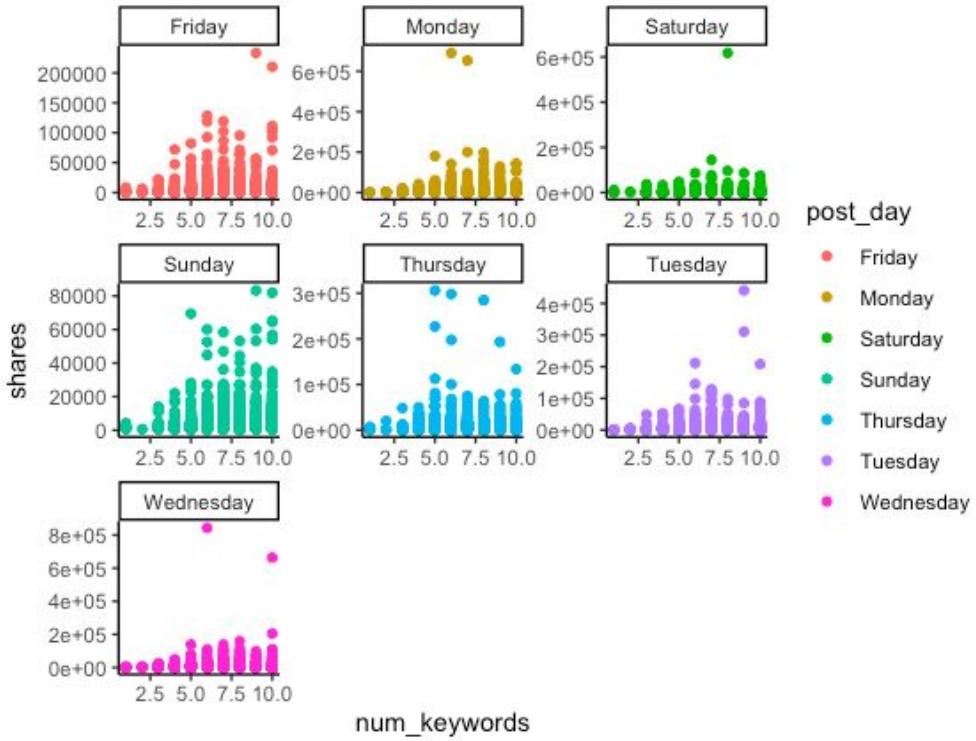
Types of Data Channels



Types of Data Channels



Article Post Days



Summary Statistics

	Overall (n=39644)
Popular Article	
Mean (SD)	0.493 (0.500)
Median [Min, Max]	0.00 [0.00, 1.00]
Shares	
Mean (SD)	3400 (11600)
Median [Min, Max]	1400 [1.00, 843000]
No. of Links	
Mean (SD)	10.9 (11.3)
Median [Min, Max]	8.00 [0.00, 304]
No. of Mashable Links	
Mean (SD)	3.29 (3.86)
Median [Min, Max]	3.00 [0.00, 116]
No. of Keywords	
Mean (SD)	7.22 (1.91)
Median [Min, Max]	7.00 [1.00, 10.0]
Entertainment Data Channel	
Mean (SD)	0.178 (0.383)
Median [Min, Max]	0.00 [0.00, 1.00]
Business Data Channel	
Mean (SD)	0.158 (0.365)
Median [Min, Max]	0.00 [0.00, 1.00]

	Overall (n=39644)
Social Media Data Channel	
Mean (SD)	0.0586 (0.235)
Median [Min, Max]	0.00 [0.00, 1.00]
Technology Data Channel	
Mean (SD)	0.185 (0.389)
Median [Min, Max]	0.00 [0.00, 1.00]
Published on Friday	
Mean (SD)	0.144 (0.351)
Median [Min, Max]	0.00 [0.00, 1.00]
Published on Saturday	
Mean (SD)	0.0619 (0.241)
Median [Min, Max]	0.00 [0.00, 1.00]
Published on Tuesday	
Mean (SD)	0.186 (0.389)
Median [Min, Max]	0.00 [0.00, 1.00]
Published on Wednesday	
Mean (SD)	0.188 (0.390)
Median [Min, Max]	0.00 [0.00, 1.00]
Published on Weekend	
Mean (SD)	0.131 (0.337)
Median [Min, Max]	0.00 [0.00, 1.00]

	Overall (n=39644)
Global Subjectivity	
Mean (SD)	0.443 (0.117)
Median [Min, Max]	0.453 [0.00, 1.00]
Min. of Positive Polarity	
Mean (SD)	0.0954 (0.0713)
Median [Min, Max]	0.100 [0.00, 1.00]
Title Subjectivity	
Mean (SD)	0.282 (0.324)
Median [Min, Max]	0.150 [0.00, 1.00]
Title Sentiment Polarity	
Mean (SD)	0.0714 (0.265)
Median [Min, Max]	0.00 [-1.00, 1.00]
Absolute Title Subjectivity	
Mean (SD)	0.342 (0.189)
Median [Min, Max]	0.500 [0.00, 0.500]
Closeness to LDA Topic 0	
Mean (SD)	0.185 (0.263)
Median [Min, Max]	0.0334 [0.00, 0.927]
Closeness to LDA Topic 1	
Mean (SD)	0.141 (0.220)
Median [Min, Max]	0.0333 [0.00, 0.926]

	Overall (n=39644)
Closeness to LDA Topic 2	
Mean (SD)	0.216 (0.282)
Median [Min, Max]	0.0400 [0.00, 0.920]
Closeness to LDA Topic 4	
Mean (SD)	0.234 (0.289)
Median [Min, Max]	0.0407 [0.00, 0.927]

Predictive Models

Logistic Model

- Interesting Initial Observations: After Exponentiation

data_channel_is_socmed	2.05
data_channel_is_tech	1.44
is_weekend	1.92
global_subjectivity	1.52
num_self_hrefs	0.98

Robustness: Sensitivity, Specificity, Accuracy

Training Set

	0	1
Predicted No	9587	5567
Predicted Yes	5520	9059

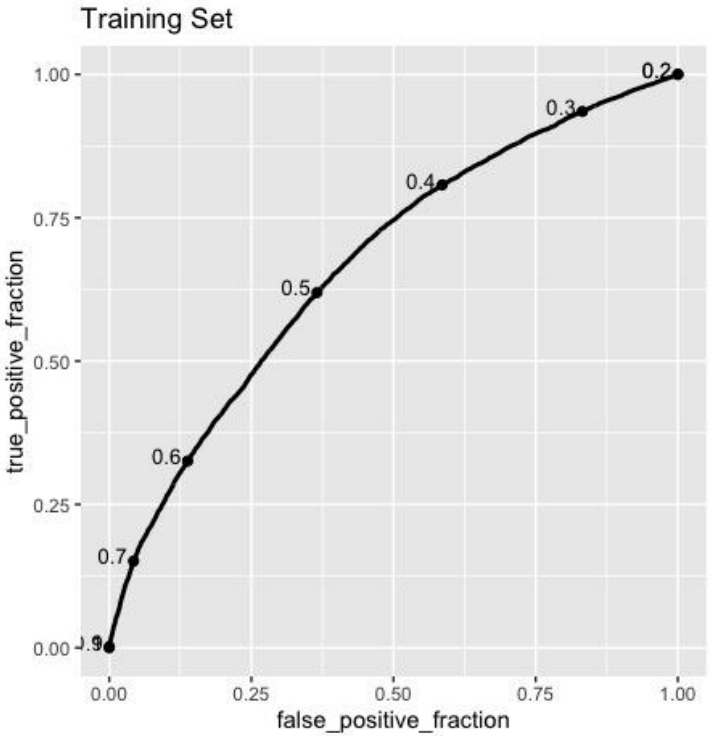
sensi_TPrate	speci_TNrate	FP_rate	accuracy
0.6193765	0.6346065	0.3653935	0.6271147

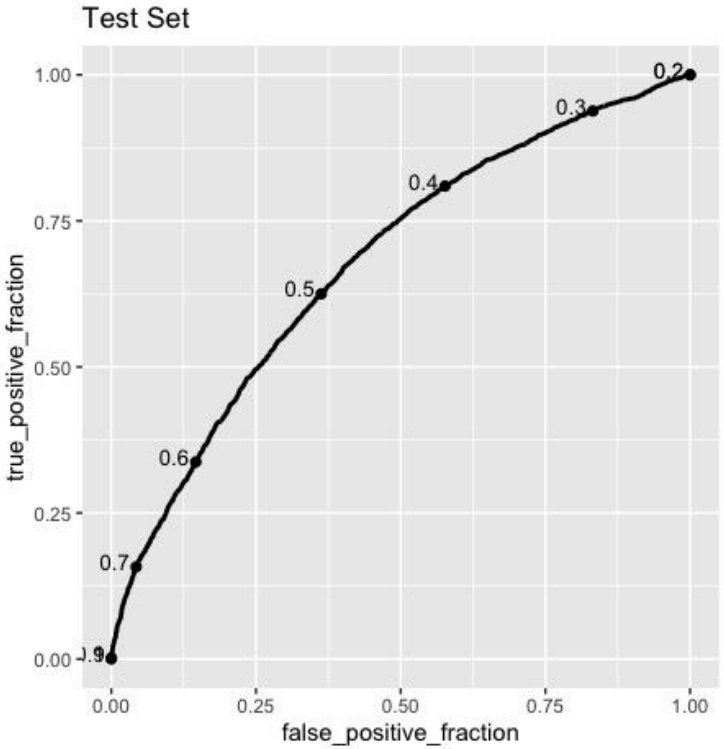
Test set

	0	1
Predicted No	3172	1852
Predicted Yes	1803	3084

sensi_TPrate	speci_TNrate	FP_rate	accuracy
0.6247974	0.6375879	0.3624121	0.6312178

Robustness: ROC Plots





Robustness: AUC

Training Set

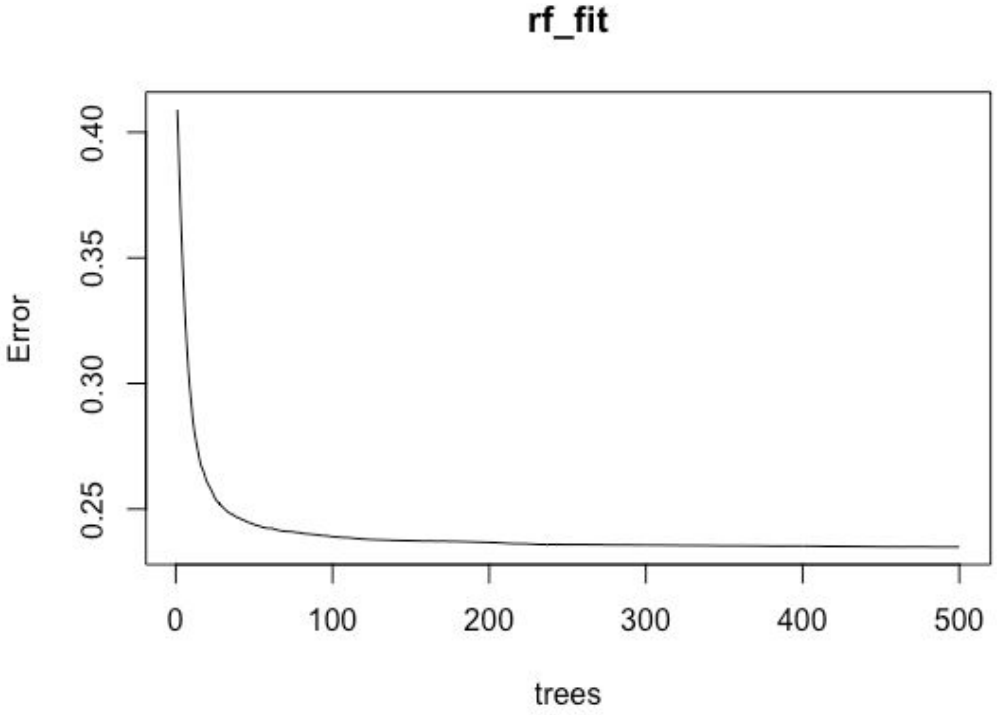
AUC
0.6722459

Test set

AUC
0.6788002

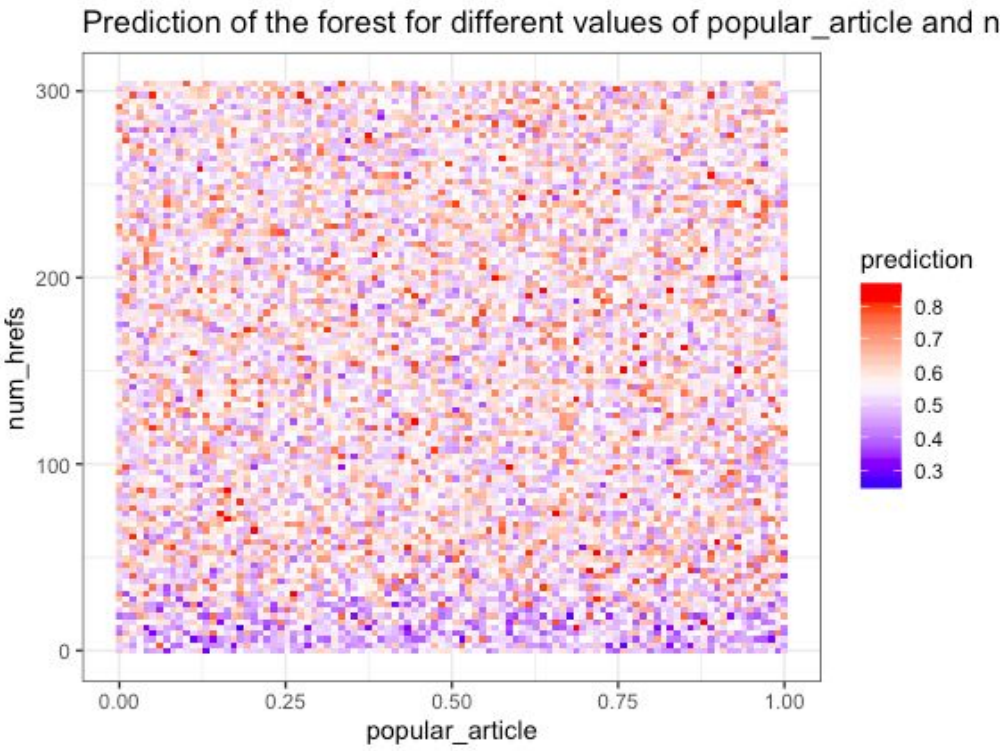
- .7-.8 AUC considered acceptable
- Model very slightly overfit

Random Forest Model



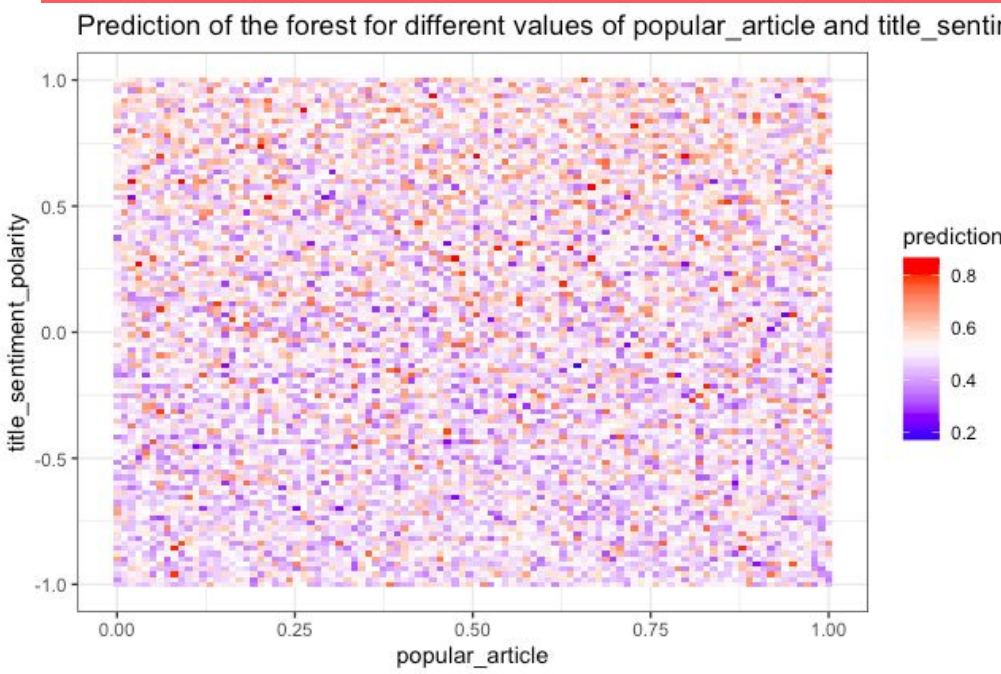
Random Forest Model

Number of external links

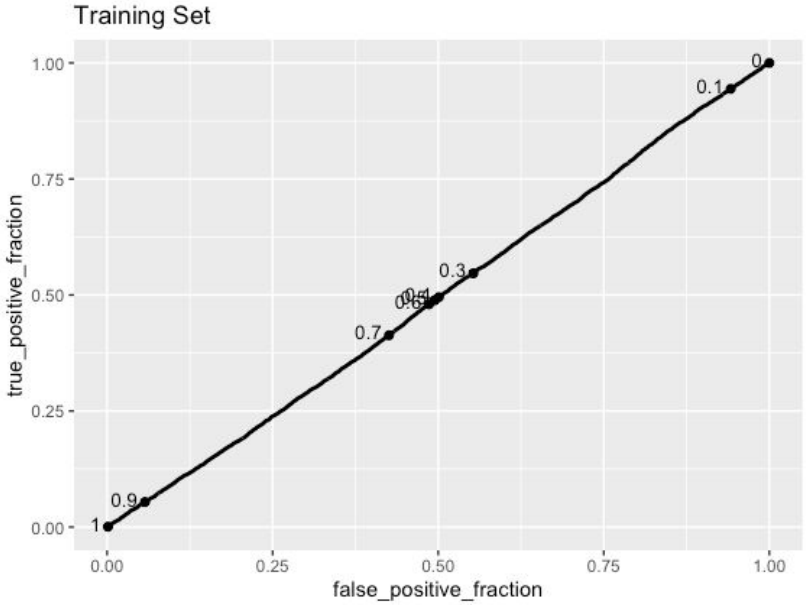


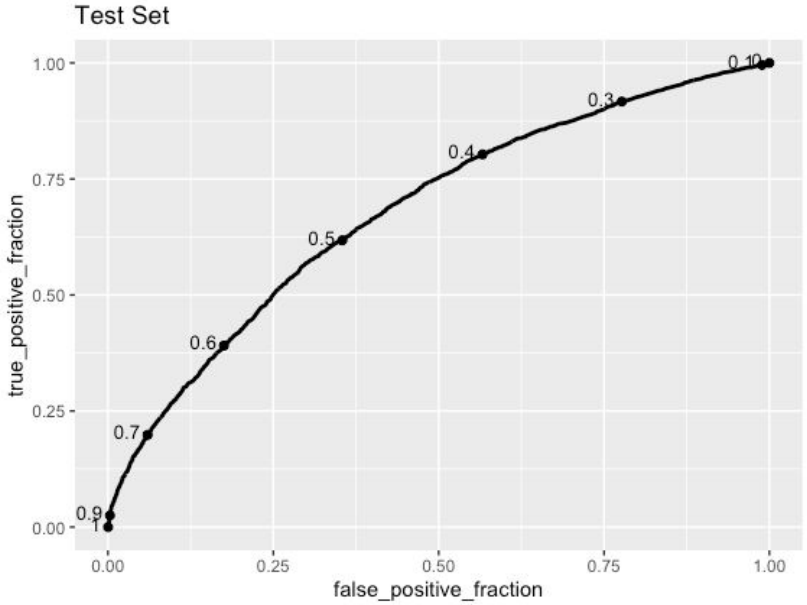
Random Forest Model

Title Sentiment Polarity



Robustness: ROC Plots





Robustness: AUC

Training Set

AUC
0.494185

Test set

AUC
0.6814448

- .7-.8 AUC considered acceptable
- Model very overfit

Robustness: Sensitivity, Specificity, Accuracy

Training Set

	0	1
Predicted No	7538	7569
Predicted Yes	7387	7239

sensi_TPrate	speci_TNrate	FP_rate	accuracy
0.4888574	0.5050586	0.4949414	0.4969899

Test set

	0	1
Predicted No	3246	1870
Predicted Yes	1729	3066

sensi_TPrate	speci_TNrate	FP_rate	accuracy
0.6211507	0.6524623	0.3475377	0.6368681

Conclusion

Limitations

- High variance of target variable (shares)
- Large amount of features
- Useful to know:
  - Demographics of people sharing the articles
  - Demographics of readers
  - Views-to-shares ratio
  - Could use clustering if more data on demographics were given

Business Case

- Hard to predict the nature of the articles and people's preferences
- Recommendations:
  - Include more than 5 keywords in the article
  - Publish on a Friday
  - Post articles about Social Media and Technology

Thank you!