PREDICT THE POPULARITY OF ONLINE NEWS

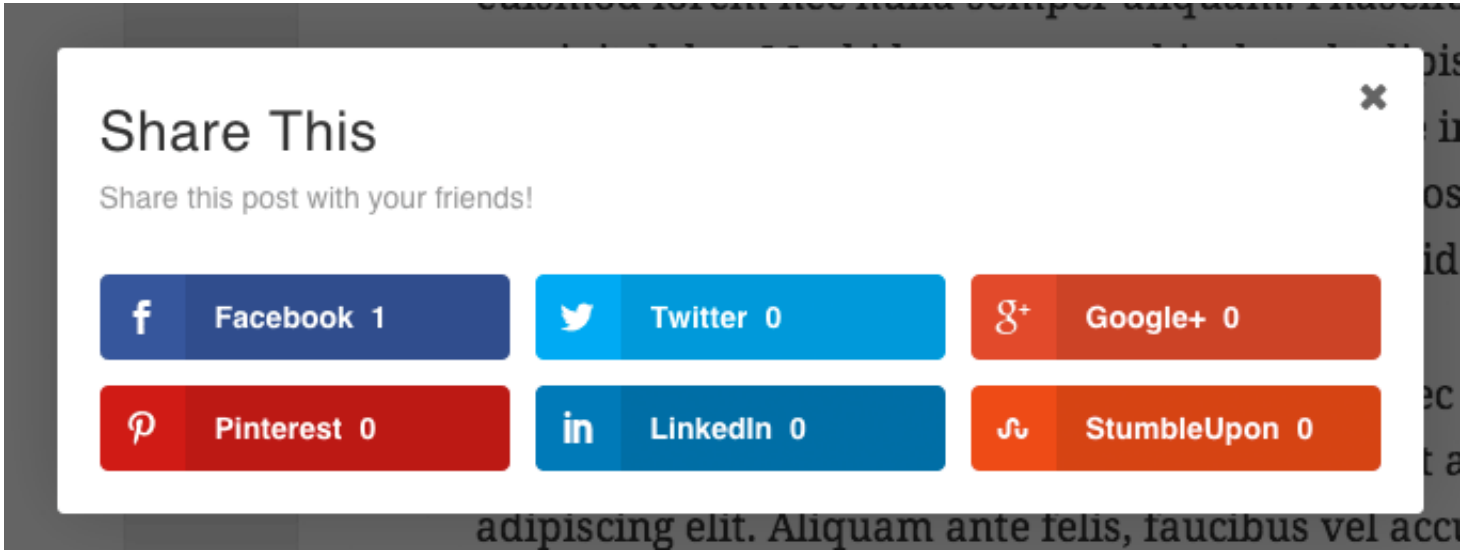
Agenda

- Background
- Methodology
- Findings
- Recommendations

BACKGROUND

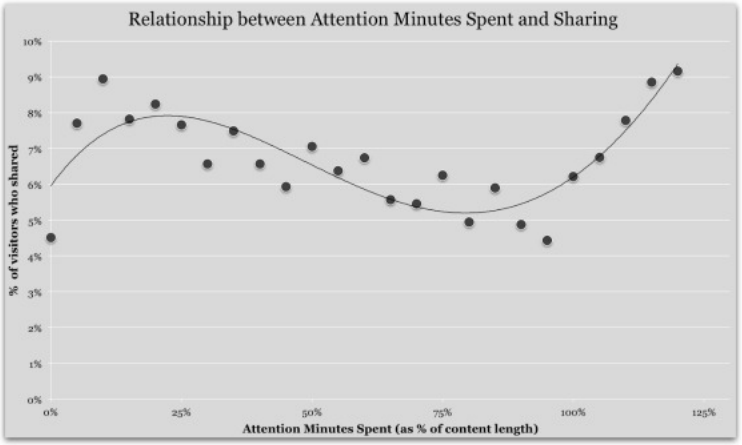
Background

The number of shares under a news article indicates how popular the news is.



UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#

Background



Over 50% of their shares come from people who have read less than 50% of an article.

**Upworthy*

Motivation to share content ?	%
To entertain others	94
To define themselves	68
To network/grow relationships	73
To feel a sense of purpose	69
To support a brand/cause	84

**New York Times Consumer Insight Group*

Objective

The objective of this project is to predict the popularity of an online article

METHODOLOGY

Data Set Information

Aspects	Features	Total Variables	Type
Words	Number of words of the title/content; Average word length; Rate of unique/non-stop words of contents	6	Numerical(6)
Links	Number of links; Number of links to other articles in Mashable	5	Numerical(5)
Digital Media	Number of images/videos	2	Numerical(2)
Publication Time	Day of the week/weekend	7	Categorical(7)
Keywords	Number of keywords; Worst/best/average keywords (#shares); Article category	11	Numerical(10) Categorical(1)
NLP	Closeness to five LDA topics; Title/Text polarity/subjectivity; Rate and polarity of positive/negative words; Absolute subjectivity/polarity level	21	Numerical(21)
Subjects	Tech/Business/Social Media/Life Style/World/Entertainment	6	Categorical(6)
Target	Number of shares at Mashable	1	Numerical(1)

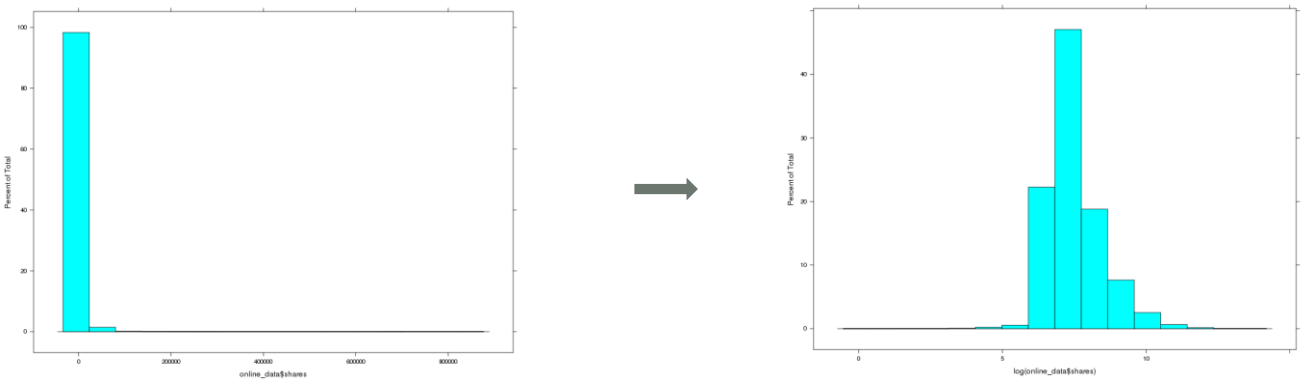
Independent Variables	
Categorical	14
Numerical	44
Dependent Variable	
Numerical	1

Data Exploration and Cleansing

Total Observations in Dataset: 39644
Training(70%): **27750**
Test(30%): **11894**

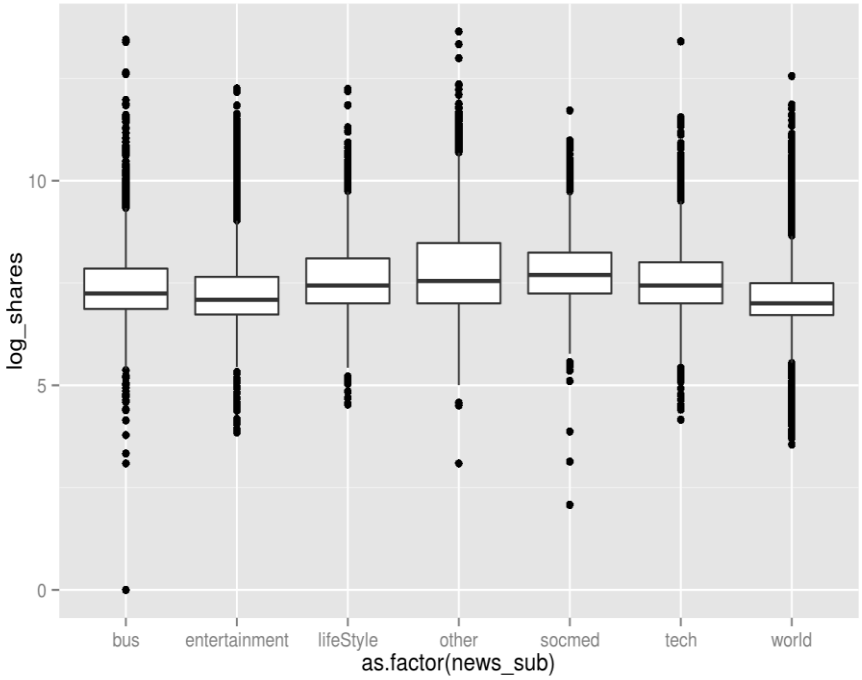
Data Cleansing:
Missing Values: By checking the distributions, **3000** observations with missing values in 9 numerical variables were found.

Dependent Variable Transformation: Shares was heavily skewed, log transformation was used to reduce the skewness.

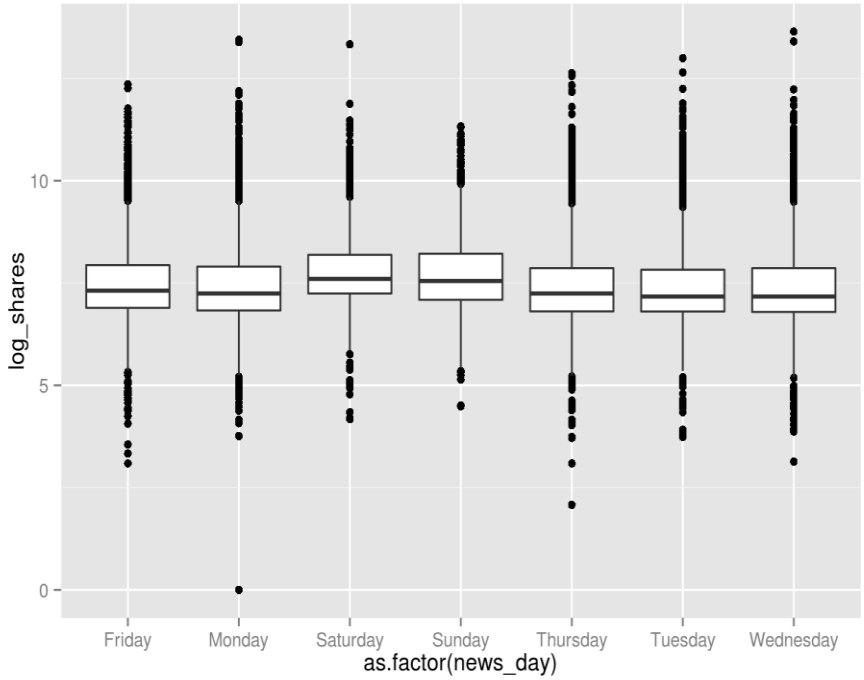


METHODOLOGY – REGRESSION MODELS

Relationship: Categorical with Dependent Variable

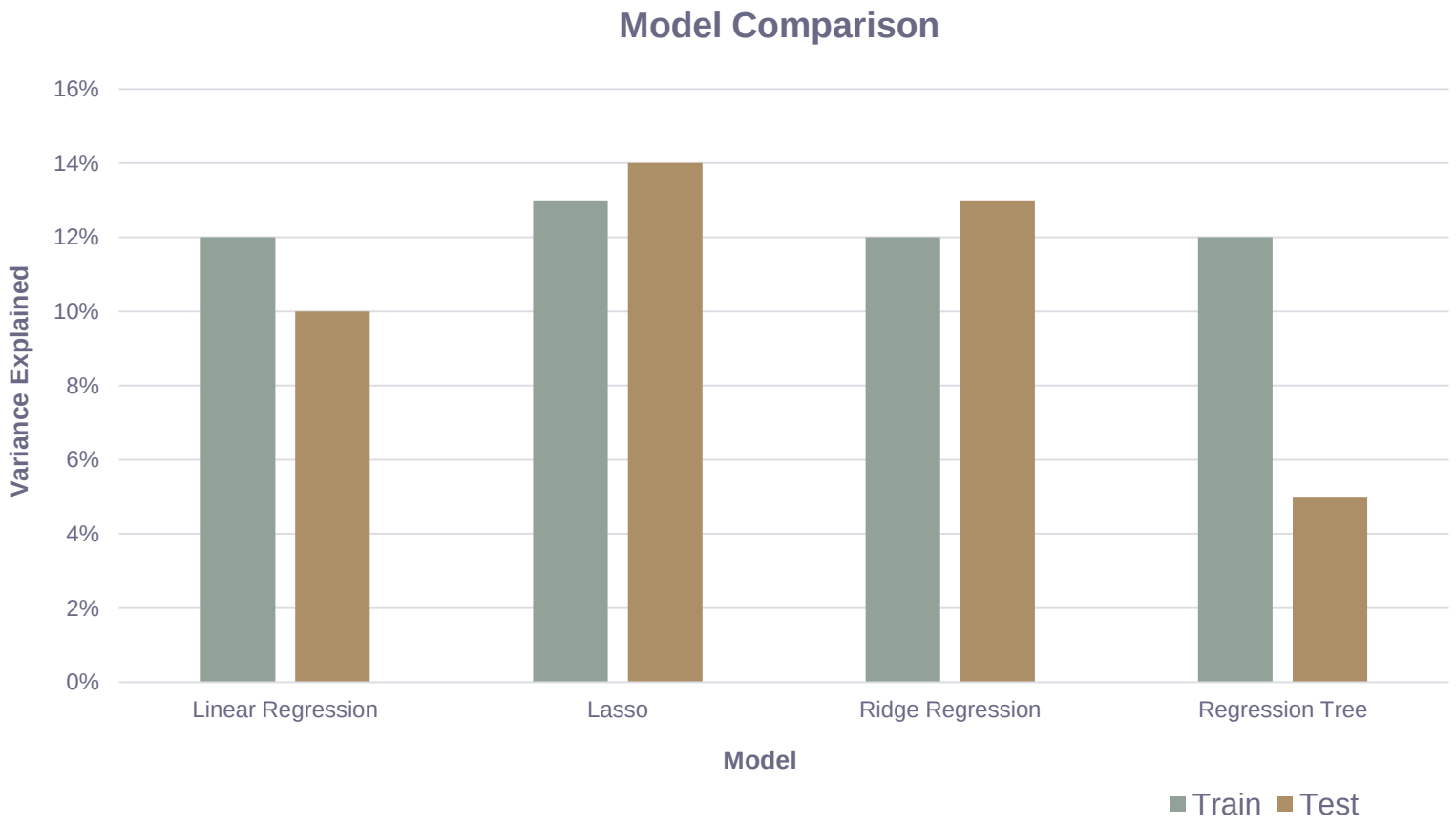


All subjects look similar regarding share numbers



Publishing day did not show much influence on shares

Regression Model Comparison



Based on the lower variance explained percentage, the four regression models did not work well for predicting the output (numbers of shares)

Clusterwise Regression Model

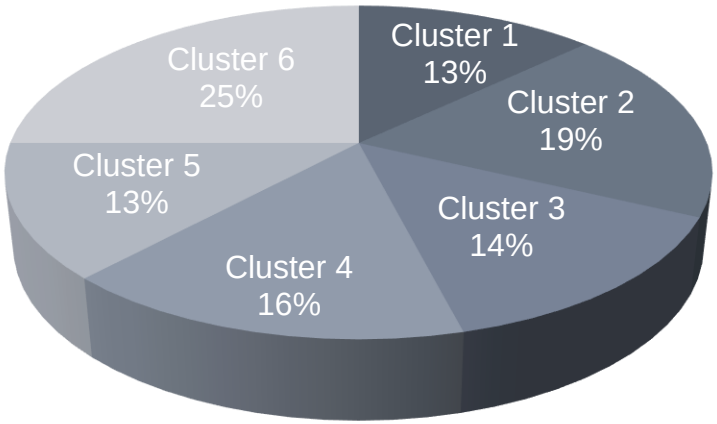
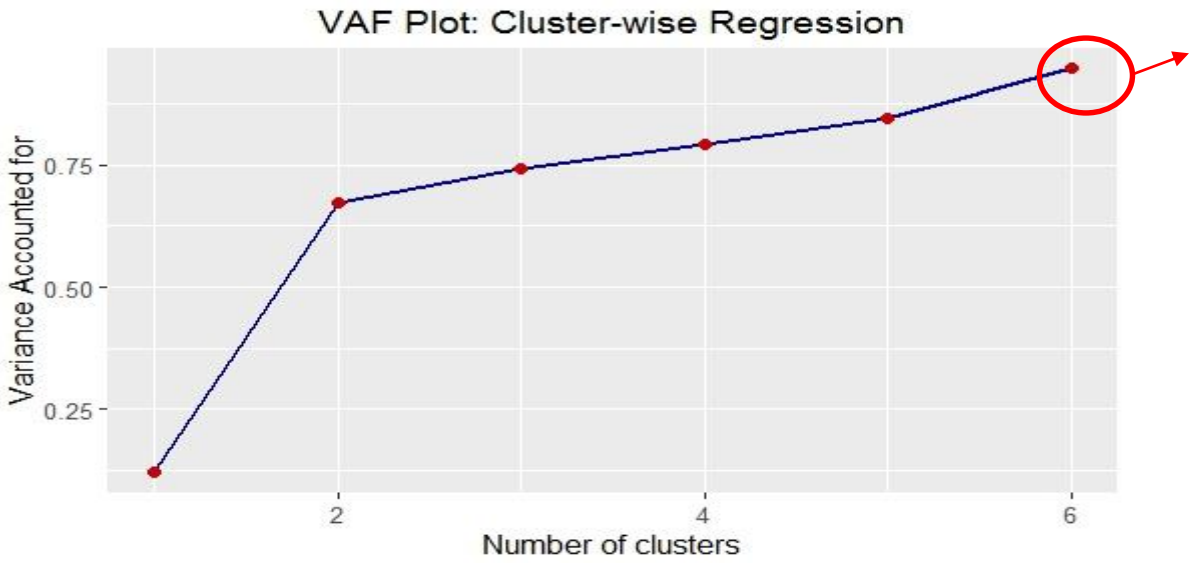


Figure : **Size of each cluster selected in cluster-wise regression**

Cluster	Size	Adjusted R-squared
1	3529	86.93%
2	5272	84.58%
3	3904	89.11%
4	4565	82.97%
5	3637	86.28%
6	6843	81.1%

Clusterwise Regression Model

VAF scree-plot



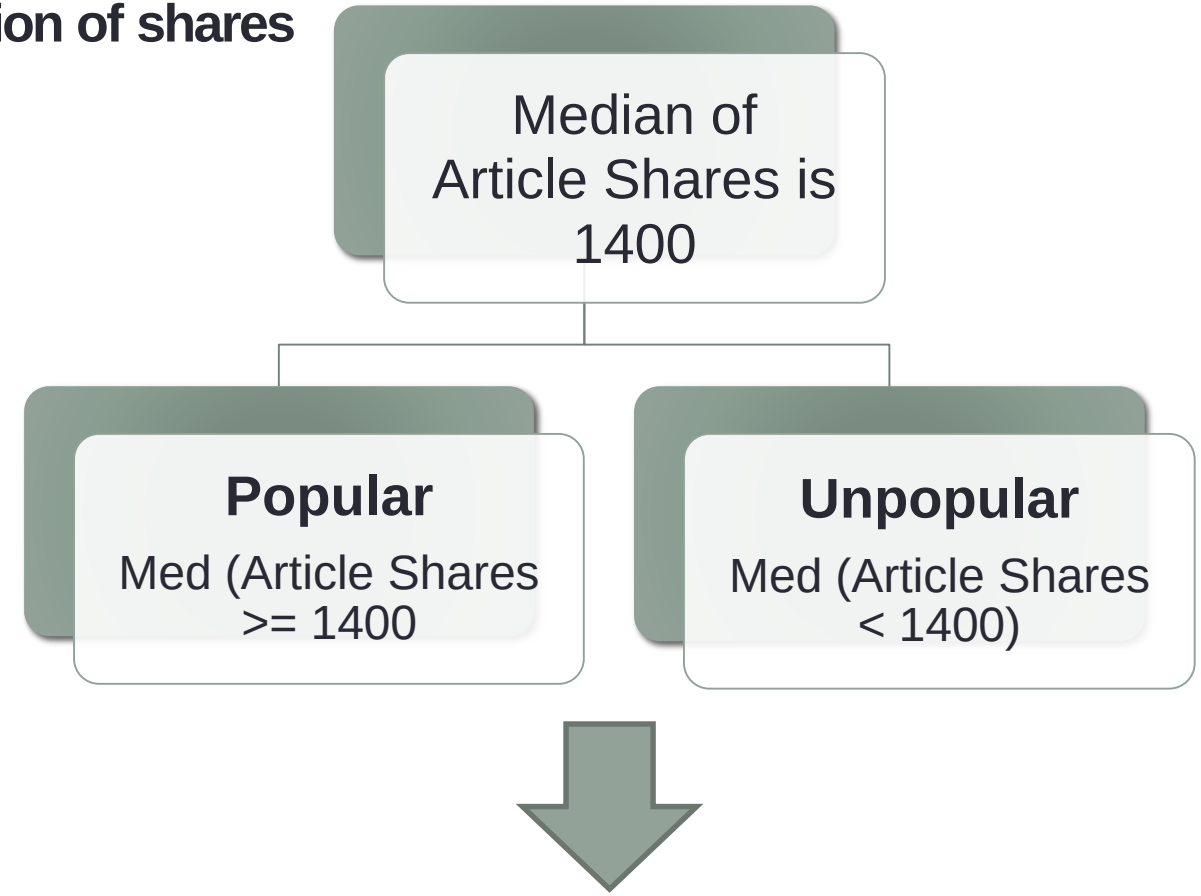
84.3% Variance accounted

6-cluster solution gives the best r-squared value

METHODOLOGY – CLASSIFICATION MODELS

Classification Model Approach

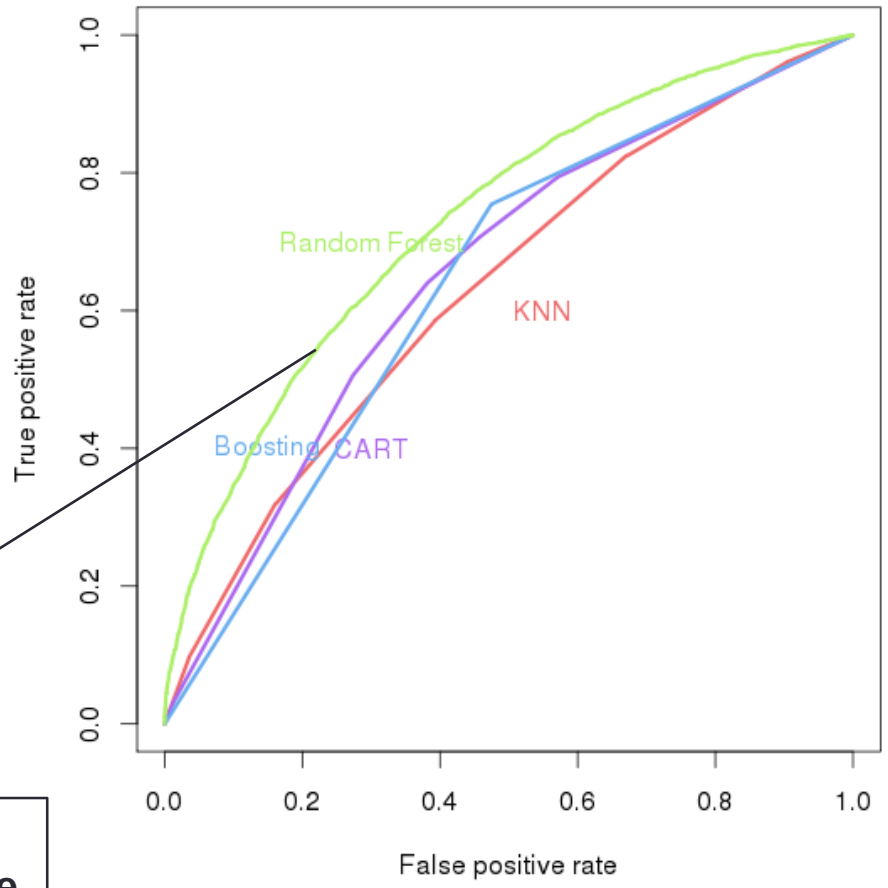
Binary categorization of shares



- Classification Tree
- Random Forest Classification
- KNN
- Boosting

Classification Model Comparison

Model	Accuracy
KNN	59.53%
Classification Tree	63.08%
Boosting	64.85%
Random Forest	66.94%



Random Forest is the best classification model to predict popularity of online article

FINDINGS

Key Findings

- Without transformation, linear regression with all variables gave **R^2 of .02**, indicating irrelevant information been used.
- After transformation and variable selection, linear model gives **R^2 of 0.12**
- **Ridge and lasso doesn't improve** on linear model (evaluated on R^2)
- Based on misclassification rate and area under curve, **Random forest is the best classification model** to predict whether the news was popular or not

RECOMMENDATIONS

Recommendations

- **From Regression Models**
  - *Dataset is not appropriate for building regression models*
  - *Without more information measuring different media environment we cannot establish that general rule that one kind of article will draw more attention than others*
- **From Classification Models**
  - *To make an article popular (shares >=1400) , publishers can-*
    - *Look at increasing:*
      Amount of key words.
      Number of links embedded.
      Number of images
    - *Have a more subjective and positive title*