Predicting Popularity in News Media

Jess Beering, Molly Carmody, Bella Hutchins, Ryan Nicholson, Catalina Sanchez-Carrion

The Problem

News and media outlets aim for their posts to reach the largest number of people as possible, therefore have a high popularity among social media users. We wanted to determine how the different factors that determine popularity impact the total number of shares an article gets.

Main Goals for the Project

- Explore attributes that best predict popularity/number of shares for an article
- Explore accuracy of three predictive models
  - Naïve Bayes
  - Random Forest
  - Regression
- Importance/relevance of number of Google Trending Topics in article and its correlation with number of shares

Data Sets Used

- UCI Online News Popularity Data Set including 39,000 Mashable Articles
- Google Trending Topics from 10 Categories per month (2013 and 2014)

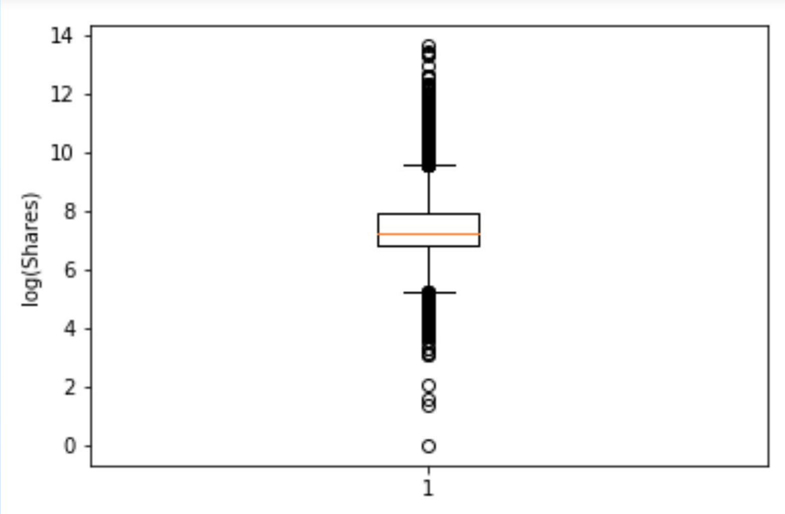
Factors Used in Our Classifier

- Week day posted
- Number of visual aides
- Category
  - Lifestyle
  - Entertainment
  - Business
  - Social Media
  - Technology
  - World
- Trending Word Count
- Automated Readability Index (ARI)

Data Cleaning

1. Cleaning original csv file: concise csv file that only contained the features of interest
2. Cleaning to get trending word counts: upload the content of each article into the csv file
3. Cleaning to evaluate the readability index: created a feature for the automated readability index (ARI) for the content of each article

Naïve Bayes



Unpopular Article: less than 957 shares [0]

Normal Article: between 957 and 2800 shares [1]

Popular Article: greater than 2801 shares [2]

Naïve Bayes

Target Popularity Values	Predicted Popularity Values
[0 1 1 0 1 0 2 2 2 1 1 1 2 1 2 1 2 1 0 2 1 1 2 2 0 1 1 0 1 1 2 1 0 1 2 2 2 1 2 0 2 1 1 2 2 0 2 1 1 2]	[0 0 0 0 2 0 1 0 1 0 0 2 2 2 0 1 2 2 0 0 1 0 0 0 0 1 0 0 0 0 2 0 0 1 1 2 0 0 2 0 0 0 0 1 0 0 0 0 2 0]

Above is a chart comparing target popularity values with the prediction popularity values from the classifier.

Naïve Bayes

We performed Naïve Bayes on each separate feature category (visual aides, article category, and day of week)

Features	Accuracy
Visual Aides	49.42%
Article Category	56.27%
Day of the week	52.43%

Random Forest

Training: As the model takes in a group of features of an article along with the accompanying number of shares.

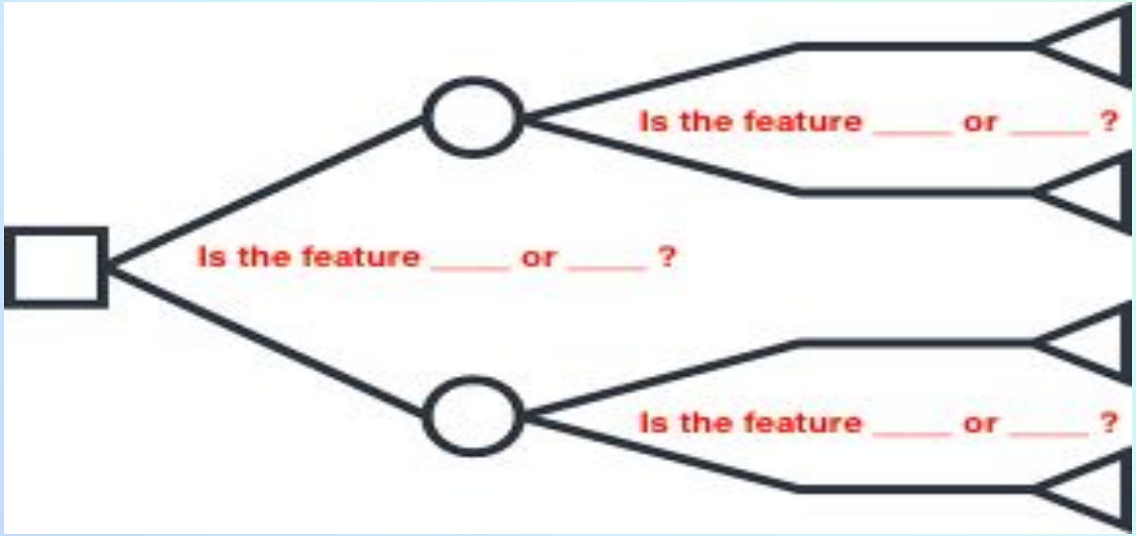
It creates decision trees that reflect how the features impact the shares in different ways.

Random Forest

These decision trees are essentially structures made up of many “questions” about an article’s features.

It uses the “answers” to those questions to arrive at the total number of shares of that article.

Random Forest


Is the feature ____ or ____ ?
Is the feature ____ or ____ ?
Is the feature ____ or ____ ?

Random Forest

The code is still running on our dataset! Will include results in the final powerpoint.

Regression

After