NewsWorthy

Understanding popularity, *before* publication

Outline

0 Problem

1 Motivation

2 Solution

3 Methodology

4 Business Value

Problem

Tesla's Roadster will go almost 400 miles on a single charge, according to Elon Musk

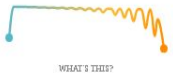
1.9k SHARES | Share on Facebook | Share on Twitter | +

So you wanna be a data scientist? A guide to 2015's hottest profession

6.9k SHARES | Share on Facebook | Share on Twitter | +

Amazon's Streaming Video Library Now a Little Easier to Navigate

593 SHARES | Share on Facebook | Share on Twitter | +



Sci-fi dreaming to desk-side vacations: The evolution of virtual tourism

4.1k SHARES | Share on Facebook | Share on Twitter | +

Facebook adds free group calling to Messenger

1.7k SHARES | Share on Facebook | Share on Twitter | +

4 Minutes of Dad Jokes Is Surprisingly Hysterical

10.0k SHARES | Share on Facebook | Share on Twitter | +

How to Land a Job at Spotify

3.2k SHARES | Share on Facebook | Share on Twitter | +

Facebook apologizes after 'Year in Review' stirs up bad memories for some users

34.7k SHARES | Share on Facebook | Share on Twitter | +

Turkish teen accused of insulting Erdogan freed, but could still get 4 years

1.3k SHARES | Share on Facebook | Share on Twitter | +

This pig savoring his treats is living his best life

1.2k SHARES | Share on Facebook | Share on Twitter | +

Are negative-sentiment articles shared more globally?

Are longer articles read on Mondays?

Do technology articles with short titles perform well?

Does number of images affect sharing?

“

Predict *likely* unpopular content—before publication

Motivation

Market Size

$22.9B

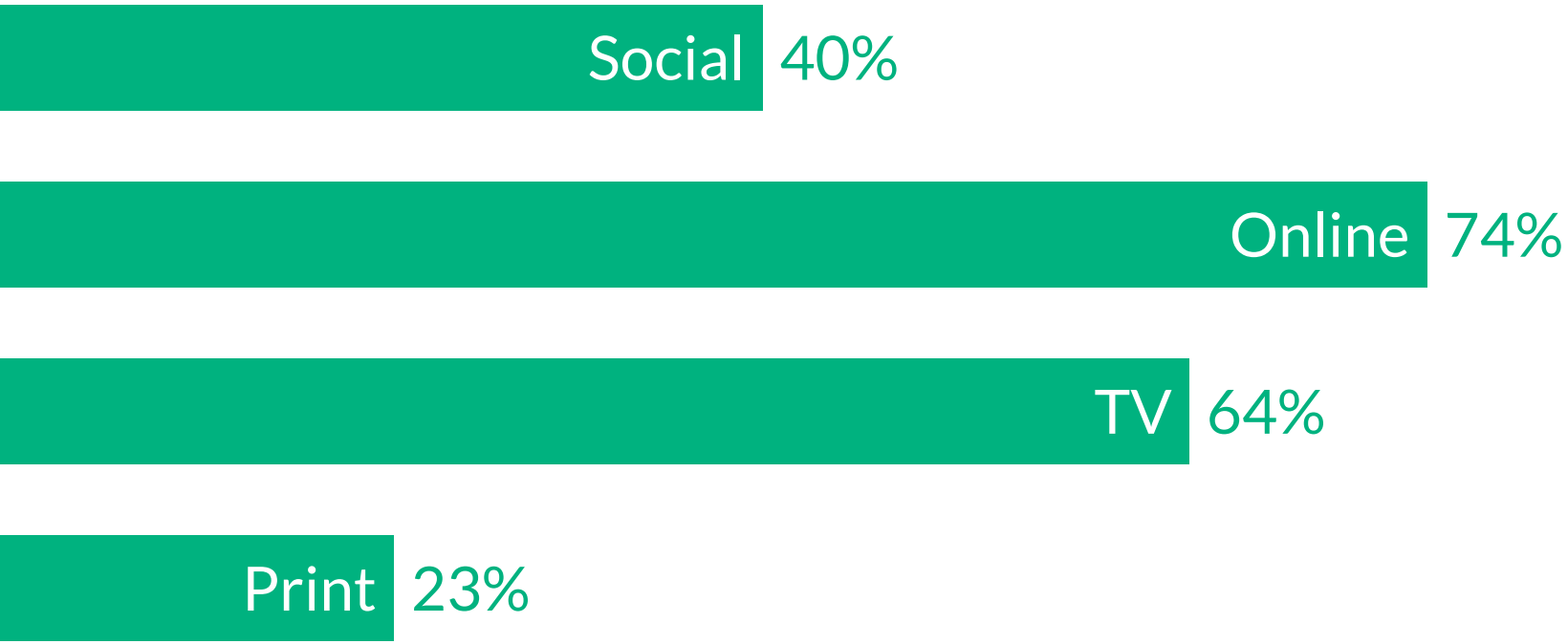
online publishing

January 2015 *

* According to Nielsen

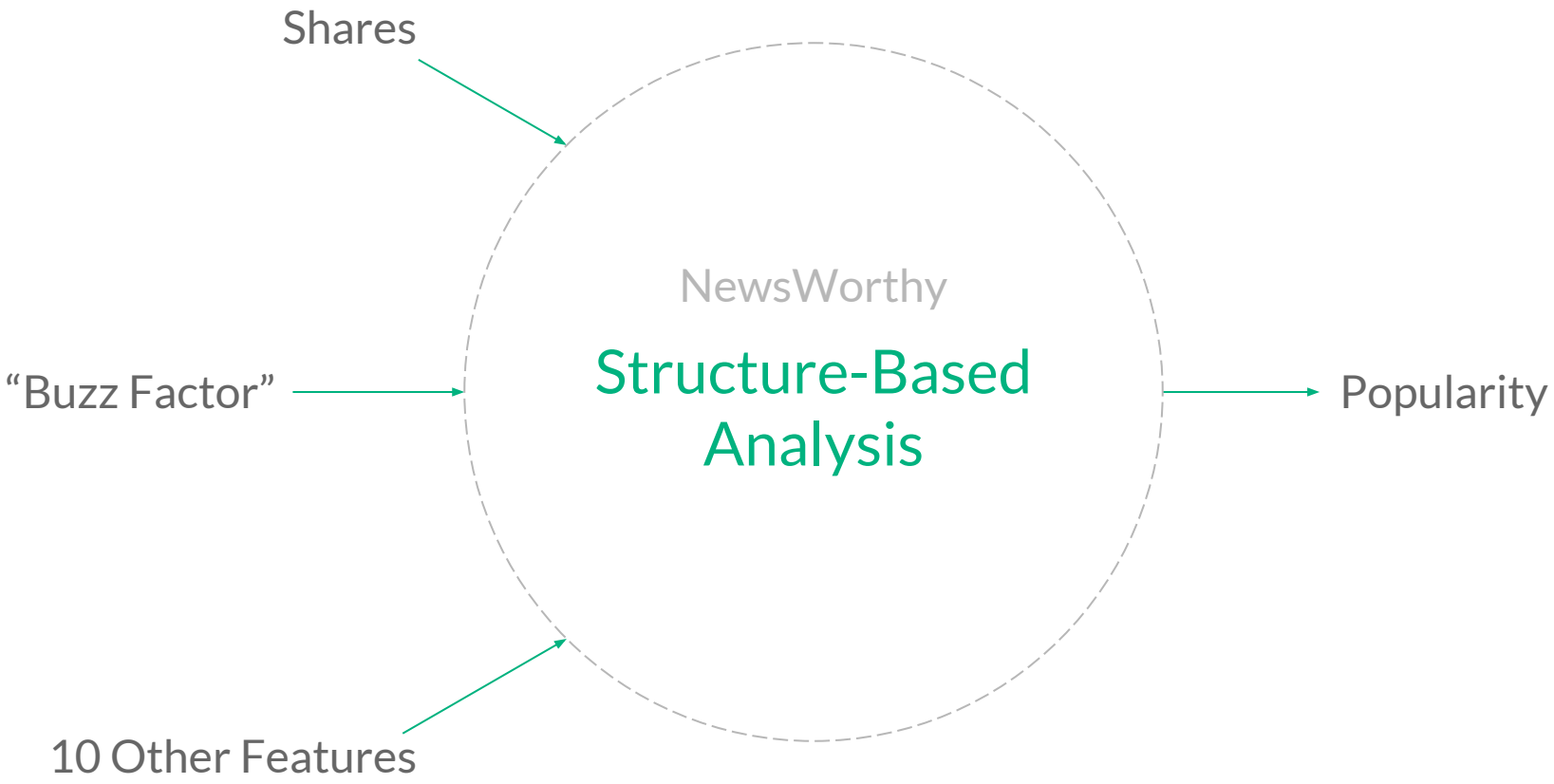
Sources of News

"Which, if any, of the following have you used in the last week as a source of news?" —Reuters Digital News Report



Solution

High-Level Process



Overview

- Structure-based analytics **applicable to all online content**
  - Pre-release content cannot be leaked
  - Large amounts of *training* content already available
- Most individuals *and* large publishing organizations **lack tools (and skills)** to perform data analysis
  - Bloggers depend on article popularity and social media buzz to attract potential advertisers
  - Large organizations *still* compete over pageviews

Methodology

0 Collect raw material

UCI Machine Learning
Online News Popularity (1/8/2015)

39.7k

Mashable articles

with 61 features

shares	title_sentiment_polarity	n_non_stop_words
is_thursday	self_reference_max_shares	num_keywords
n_tokens_title	is_world	is_tech
num_hrefs	global_subjectivity	n_tokens _content

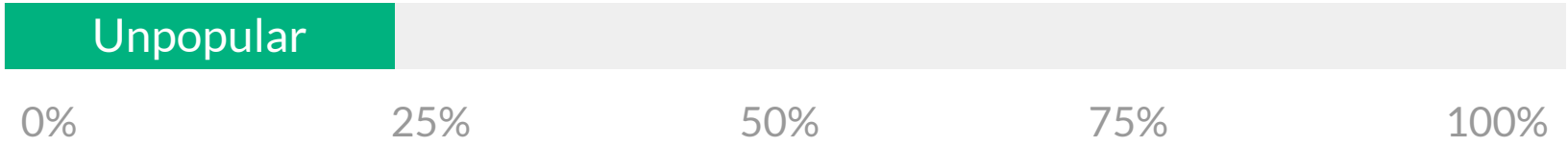
1 Transform articles into data

- Compute attributes based on article *structure*, as opposed to textual content
  - Computed for any text
  - Comparable regardless of the availability of similar documents
  - For small sample sizes, little overlap in text tokens
  - *Software as a Service* (SaaS) enables data analysis without needing access to the original text

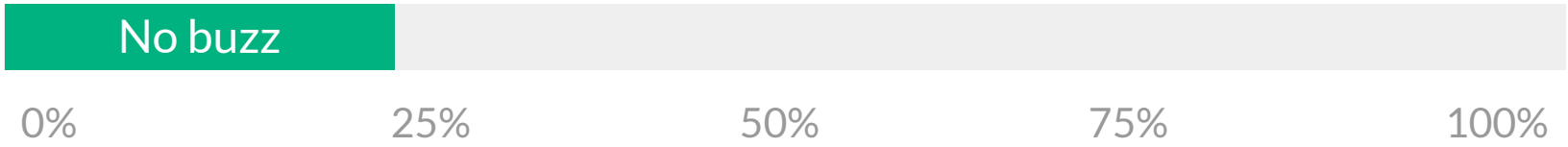
2 Calculate metrics

Given an online news article collection, predict which will rank in the bottom 25% in **popularity** or **buzz factor**.

Popularity raw number of shares an article receives

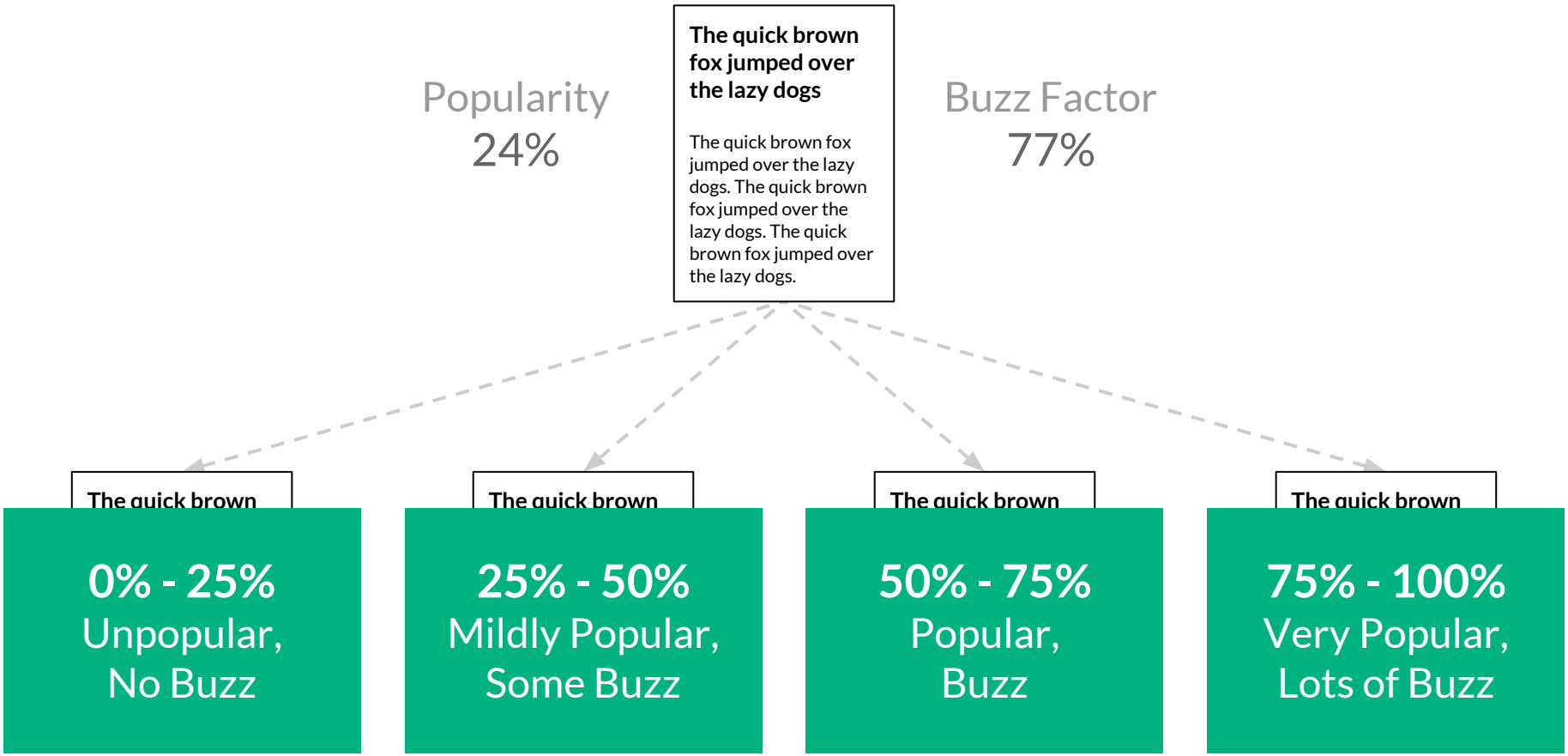


Buzz Factor number of shares received per day



3 Sort into popularity & buzz factor bins

Training Article

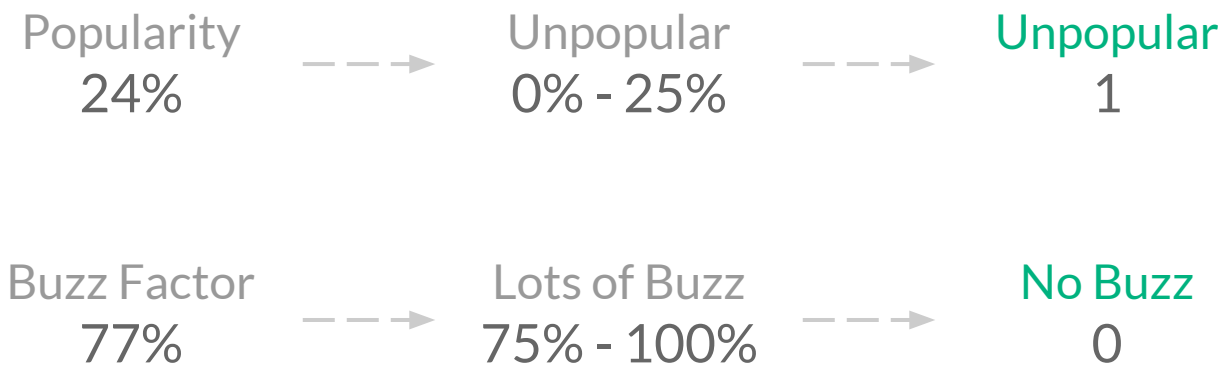


4 Derive (boolean) target variables

Training Article

The quick brown fox jumped over the lazy dogs

The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs.



Predict whether or not an article will be unpopular or generate no buzz based on its popularity and buzz factor bins.

5 Select features

50 Features

↓

“Feature-Importance Metric”

ExtraTreesClassifier

↓

10 Features

Most relevant to target variables

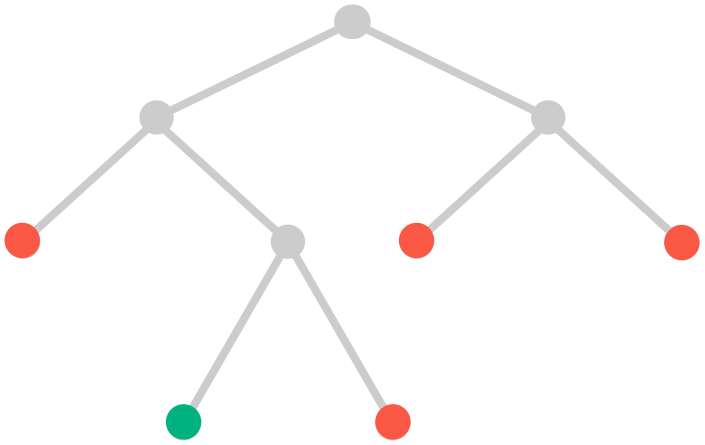
variable	importance	variable	importance
timedelta	0.042349	n_tokens_title	0.034482
num_keywords	0.034670	average_token_length	0.031795
...	...	...	...

6 Train machine learning algorithms

Considering top 10 features from feature selection:

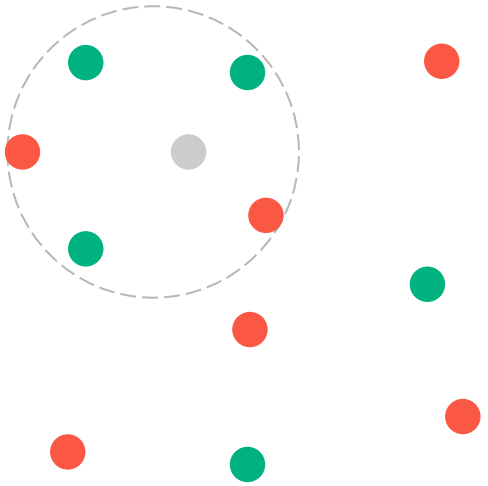
Random Forest Classifier

Estimators = 100



K-Nearest Neighbors

K chosen by cross-validation



7 Make predictions with *ensemble* learning

Test Article

The quick brown fox jumped over the lazy dogs

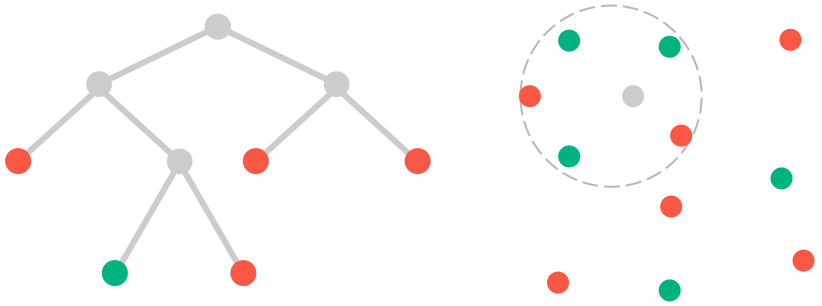
The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs.

Similar Articles

The quick brown fox jumped over the lazy dogs

The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs. The quick brown fox jumped over the lazy dogs.

Train on Algorithms

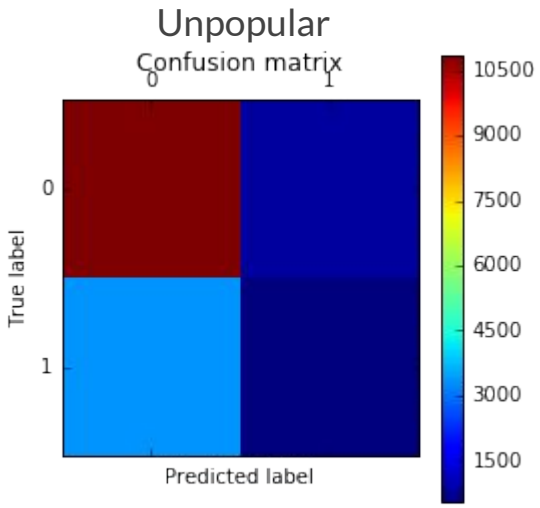


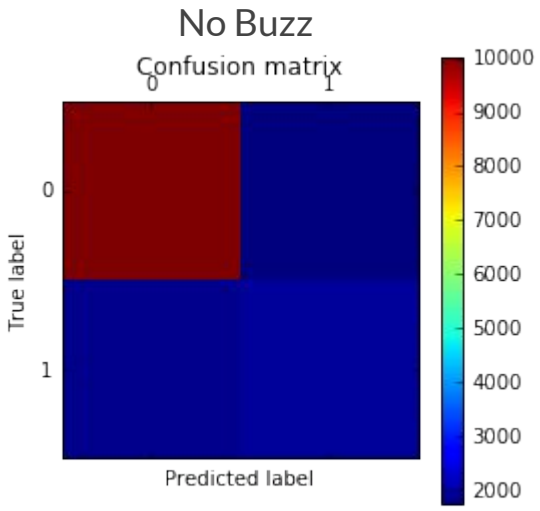
Random Forest (Unpopular) & **K-NN** (Popular) = **Result** (Unpopular)

- Unpopular if **either** of the trained model predictions is unpopular
  - Not the *most* accurate,
  - But minimizes *expensive* false negatives

Prototype Results

	Unpopular	No Buzz
Articles	9,798	9,784
Random Forest Accuracy %	74.62	79.36
K-NN Accuracy % (K = 8, 7)	74.68	77.86
Ensemble Learning %	73.03	77.61





Business Value

NewsWorthy Competitive Advantage

1. Transform content into data
2. Understand popularity *before* publication
3. Apply across all content past, present, future
4. Secure pre-release content cannot be leaked

Future Opportunity

- Structure-based prediction not limited to *just* NewsWorthy
  - BlogWorthy, for professional bloggers
  - AdWorthy, for advertising firms
  - TweetWorthy, for social media gurus
- Online & social content rapidly increasing; others in decline
- Investment opportunity with high ceiling for future returns

Questions

Haley, Helen, Tom, Tyler

“

Tyler is one smart dude.

—Dr. Randy Paffenroth