Predicting the Popularity of Online News

Nehemiah Solis

Project Overview

Online News

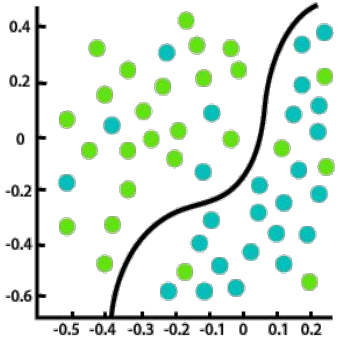
- Go-to source for news and entertainment
- Revenue from “Cost-Per-Click”
- Shares indicate popularity
- More shares = More revenue

Mashable Inc.

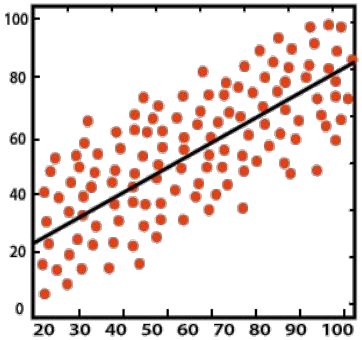
- Digital media website founded in 2005
- 9.7 million Twitter followers
- 7.5 million Facebook fans

Problem Statement

- Raw data indicates a regression problem
- Supervised Learning techniques
  - Regression task
    - Predict the “number of shares”?
  - Classification task
    - Predict whether an article will become popular or not?
- Optimize features to predict “number of shares” or “popularity”



Classification



Regression

Dataset Description

- Originally compiled by K. Fernandes et al.
  - January 2013 - January 2015
  - Pre-Processed
- Number of Instances: 39,643
- Number of Attributes: 61
  - 1 target (‘shares’)
  - 2 non-predictive features (‘URL’ and ‘Days between article publication and dataset acquisition’)
  - 58 predictive features
- Attribute Characteristics: Integer, Real

Dataset Description

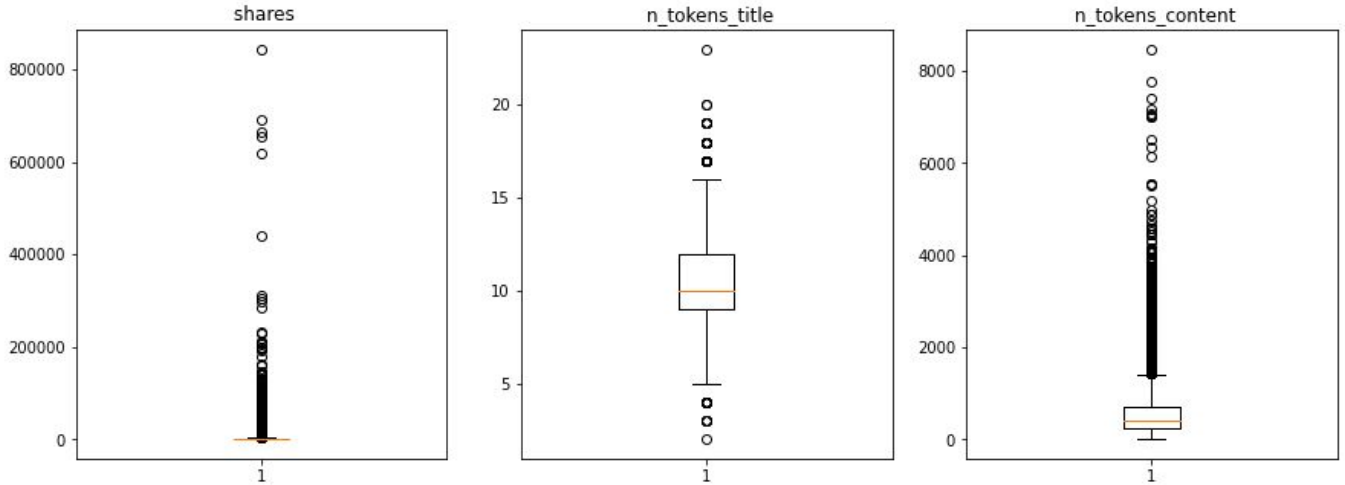
Feature	Type
Words	
Number of words in the title	number
Number of words in the article	number
Average word length	number
Rate of non-stop words	ratio
Rate of unique words	ratio
Rate of unique non-stop words	ratio
Links	
Number of links	number
Number of Mashable article links	number
Minimum, average, and maximum number of shares of Mashable links	number
Digital Media	
Number of images	number
Number of videos	number
Time	
Day of the week	nominal
Published on a weekend?	bool

Target	Type
Number of article Mashable shares	number

Feature	Type
Keywords	
Number of keywords	number
Worst keyword (min/avg/max. shares)	number
Average keyword (min/avg/max. shares)	number
Best keyword (min/avg.max. shares)	number
Article category (Mashable data channel)	nominal
Natural Language Processing	
Closeness to top 5 LDA topics	ratio
Title subjectivity	ratio
Article text subjectivity score and its absolute difference to 0.5	ratio
Title sentiment polarity	ratio
Rate of positive and negative words	ratio
Pos. word rate among non-neutral words	ratio
Neg. word rate among non-neautral words	ratio
Polarity of positive words (min/avg/max)	ratio
Polarity of negative words (min/avg/max)	ratio
Article text polarity score and its absolute difference to 0.5	ratio

Data Cleaning

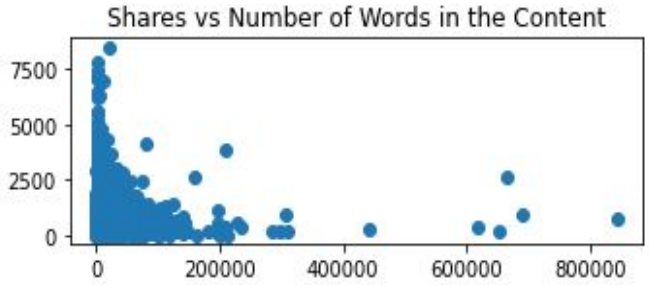
- Omit unnecessary features
- Variable Type and Missing values
- Outlier Detection

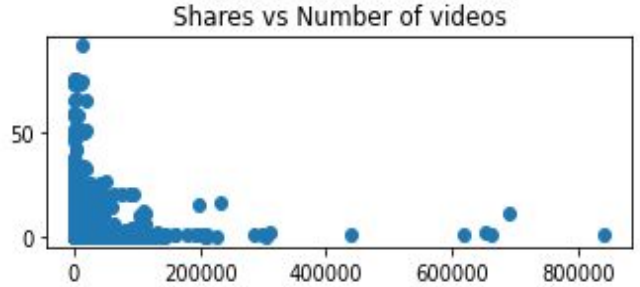


Exploratory Analysis

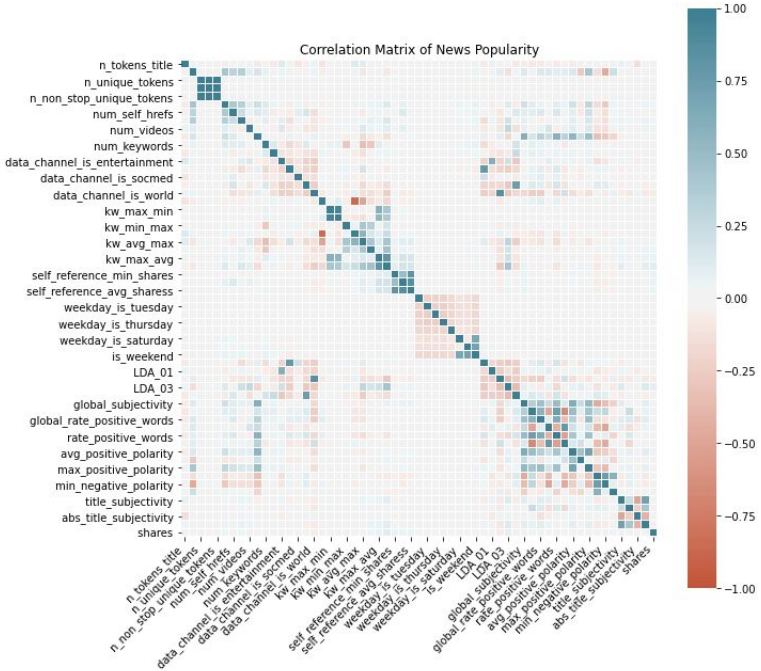
Continuous Variables

- Quick observations about relationships from scatter plots





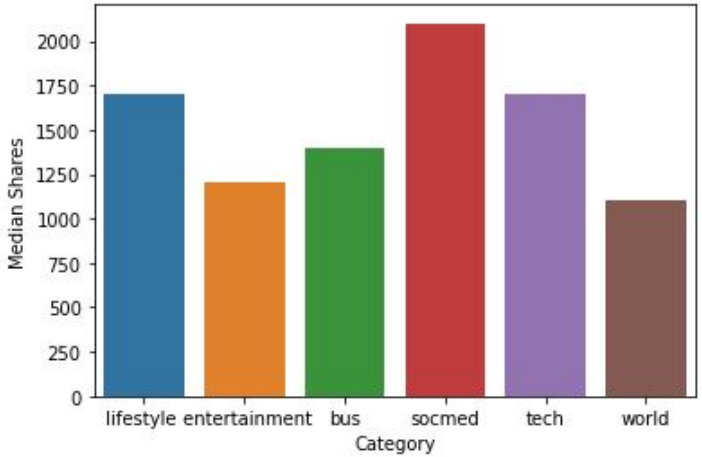
Exploratory Analysis

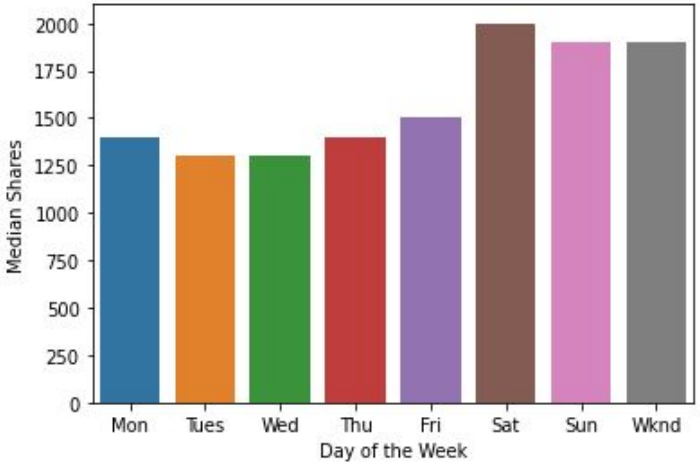


Exploratory Analysis

Categorical Variables

- Day of the week
- Category of article publication





Feature Engineering

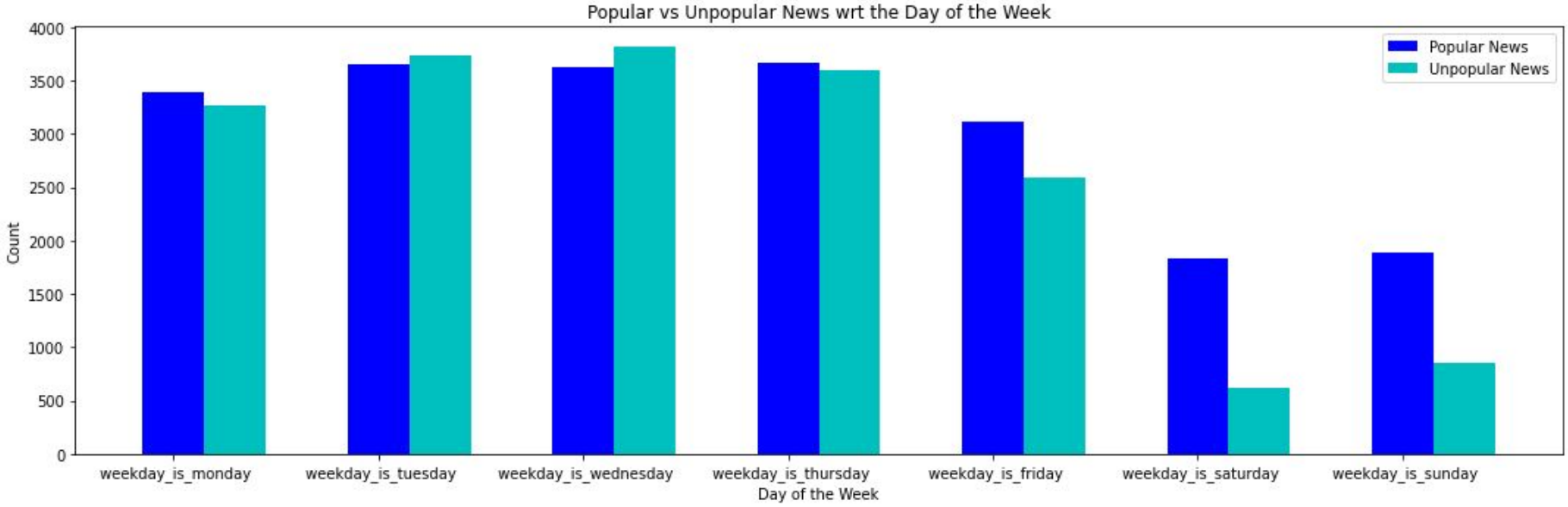
Normalization

- Many features, many ranges of values
- Normalize non-log transformed features

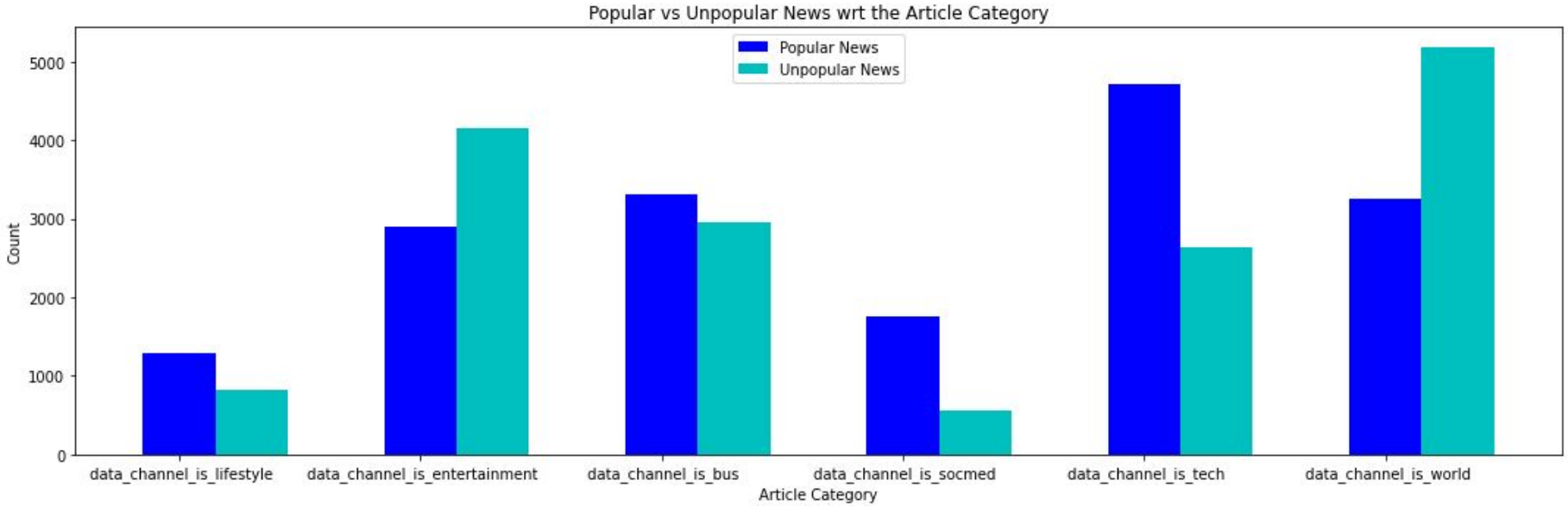
Target Transformation

- Transform ‘shares’ into popular and unpopular categories
- Used for classification tasks
- Unbounded target values for regression tasks

Feature Engineering



Feature Engineering



Feature Engineering

Principal Component Analysis

- One component explains 76% of total variance

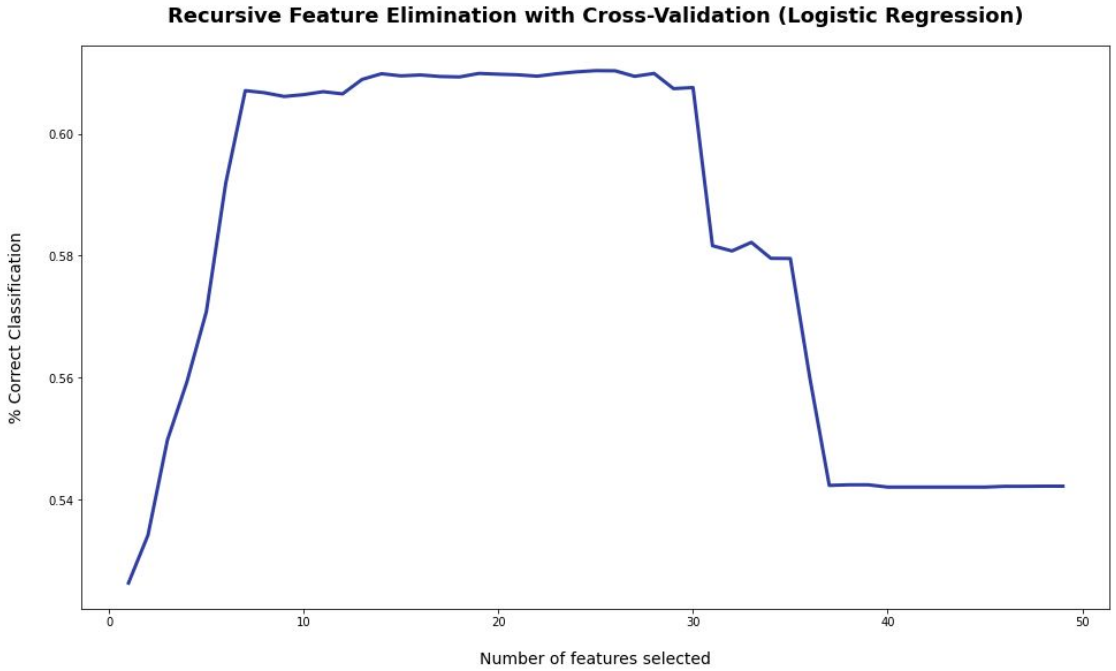
Correlation Analysis

- Are any features correlated with each other?

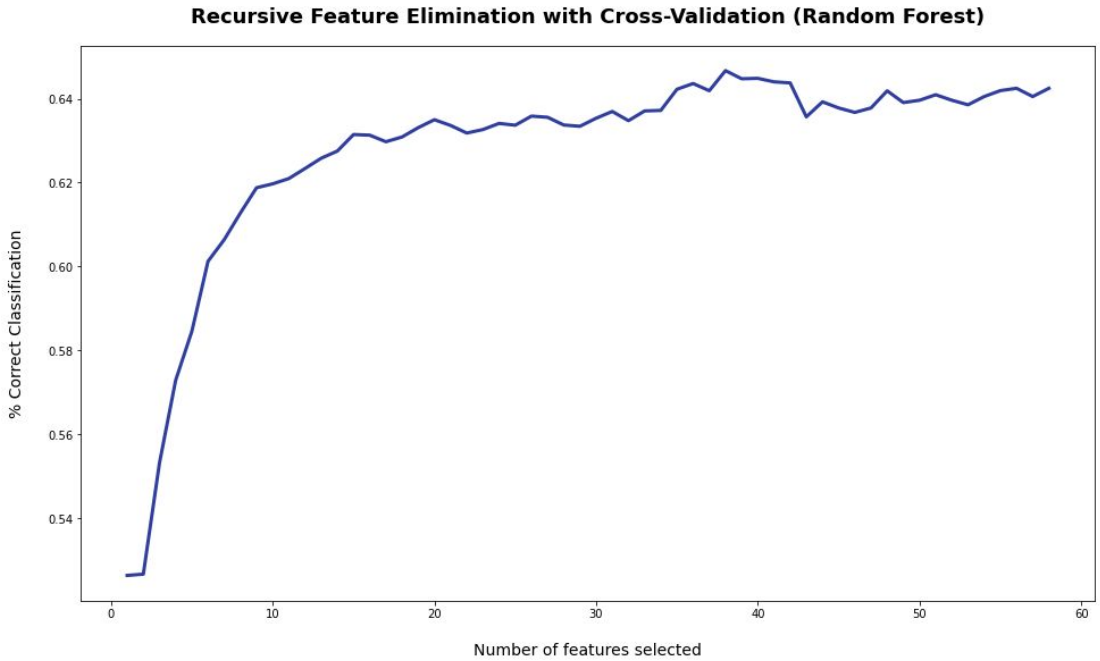
Recursive Feature Elimination Cross-Validation

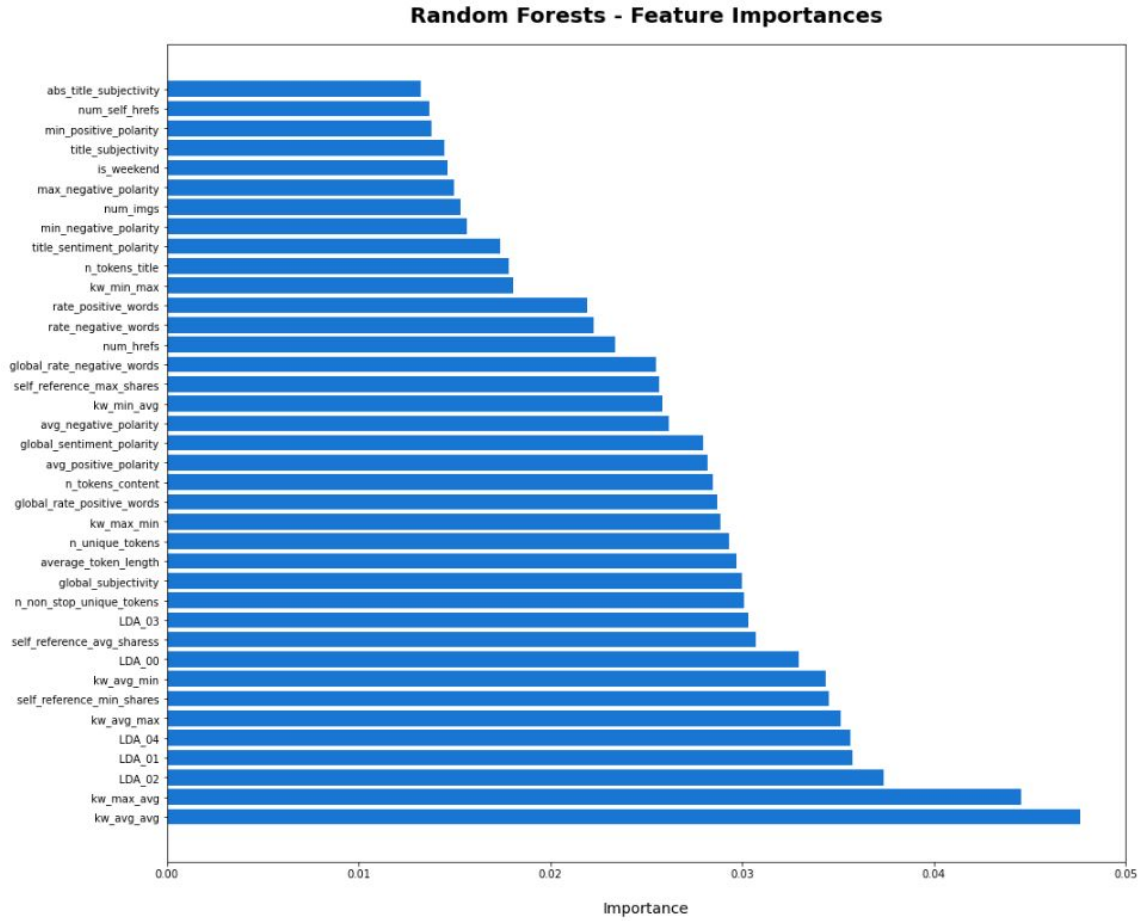
- Logistic Regression
  - Truncated feature set from correlation analysis
- Random Forest Classifier
  - Full feature set

Feature Engineering



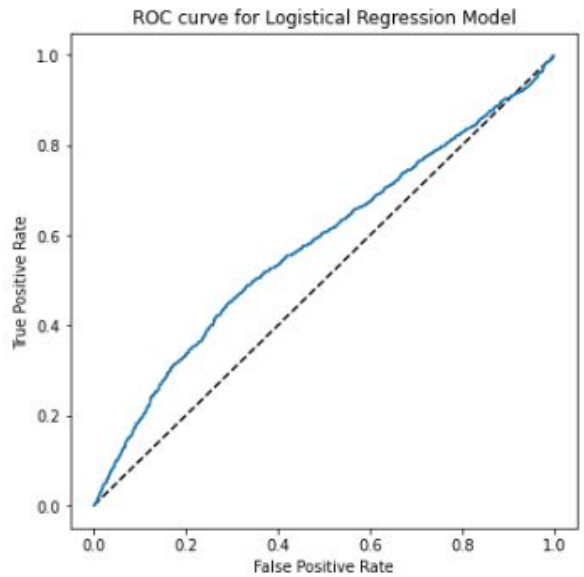
Feature Engineering




Random Forests - Feature Importances
abs_title_subjectivity
num_self_hrefs
min_positive_polarity
title_subjectivity
is_weekend
max_negative_polarity
num_imgs
min_negative_polarity
title_sentiment_polarity
n_tokens_title
kw_min_max
rate_positive_words
rate_negative_words
num_hrefs
global_rate_negative_words
self_reference_max_shares
kw_min_avg
avg_negative_polarity
global_sentiment_polarity
avg_positive_polarity
n_tokens_content
global_rate_positive_words
kw_max_min
n_unique_tokens
average_token_length
global_subjectivity
n_non_stop_unique_tokens
LDA_03
self_reference_avg_sharess
LDA_00
kw_avg_min
self_reference_min_shares
kw_avg_max
LDA_04
LDA_01
LDA_02
kw_max_avg
kw_avg_avg
0.00
0.01
0.02
0.03
0.04
0.05
Importance

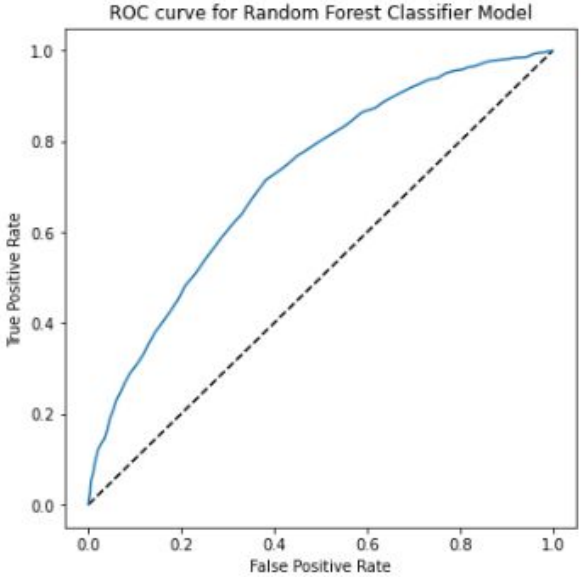
Machine Learning Approaches

Logistic Regression



Area under the ROC curve: 0.579

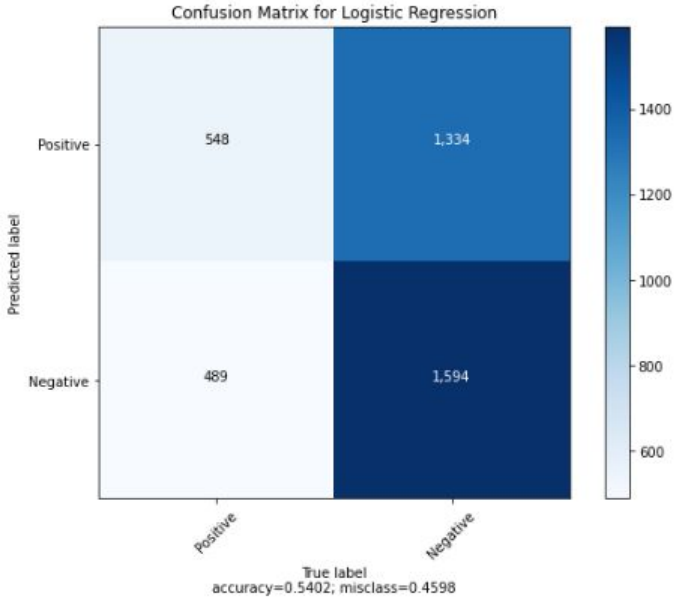
Random Forests



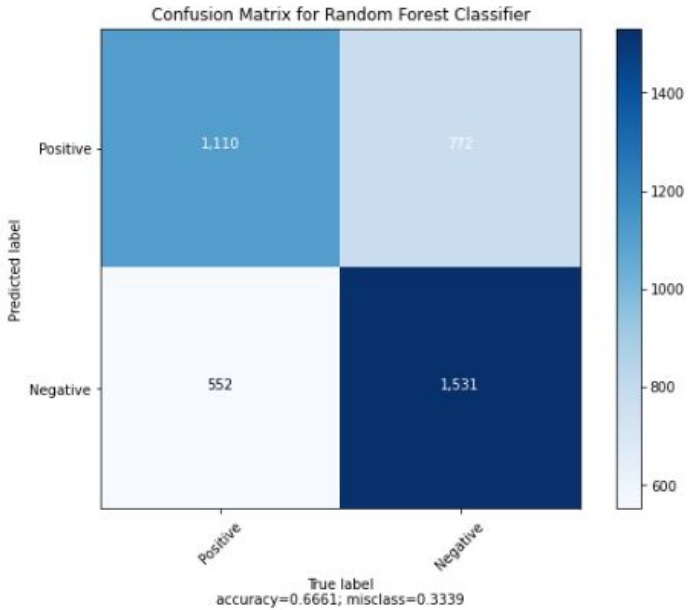
Area under the ROC curve: 0.718

Machine Learning Approaches

Logistic Regression



Random Forests



Machine Learning Approaches

Logistic Regression

	precision	recall	f1-score
0	0.53	0.29	0.38
1	0.54	0.77	0.64
accuracy			0.54
macro avg	0.54	0.53	0.51
weighted avg	0.54	0.54	0.51

Accuracy of Logistic Regression: 0.5402

Random Forests

	precision	recall	f1-score
0	0.67	0.59	0.63
1	0.66	0.73	0.70
accuracy			0.67
macro avg	0.67	0.66	0.66
weighted avg	0.67	0.67	0.66

Accuracy of Random Forest Classifier: 0.6661

Machine Learning Approaches

	OLS	Lasso
R-squared	0.013	0.013
Mean Absolute Error	3040.58	3105.64
Mean Squared Error	71399393.89	66194782.83
Root Mean Squared Error	8449.81	8136.02
Mean Absolute Percentage Error	229.22	305.4

Machine Learning Approaches

	Regression	Classification
Models Used	• Logistic Regression • Ordinary Least Squares • Lasso Regression	• Random Forest • Support Vector Machines
Best Model	Logistic Regression	Random Forest

Conclusion and Recommendations

- Regression Models that intend to predict “number of shares” performed poorly.
- Classification Models that predict “popular articles” performed well.
  - Random Forest performed the best
    - Precision = 0.67
    - Recall = 0.67
- Features to focus on to improve popularity
  - Increase
  - Decrease
  - Day of the Week
  - Article Category
- Future Work

Questions?