Mashable Popularity Prediction

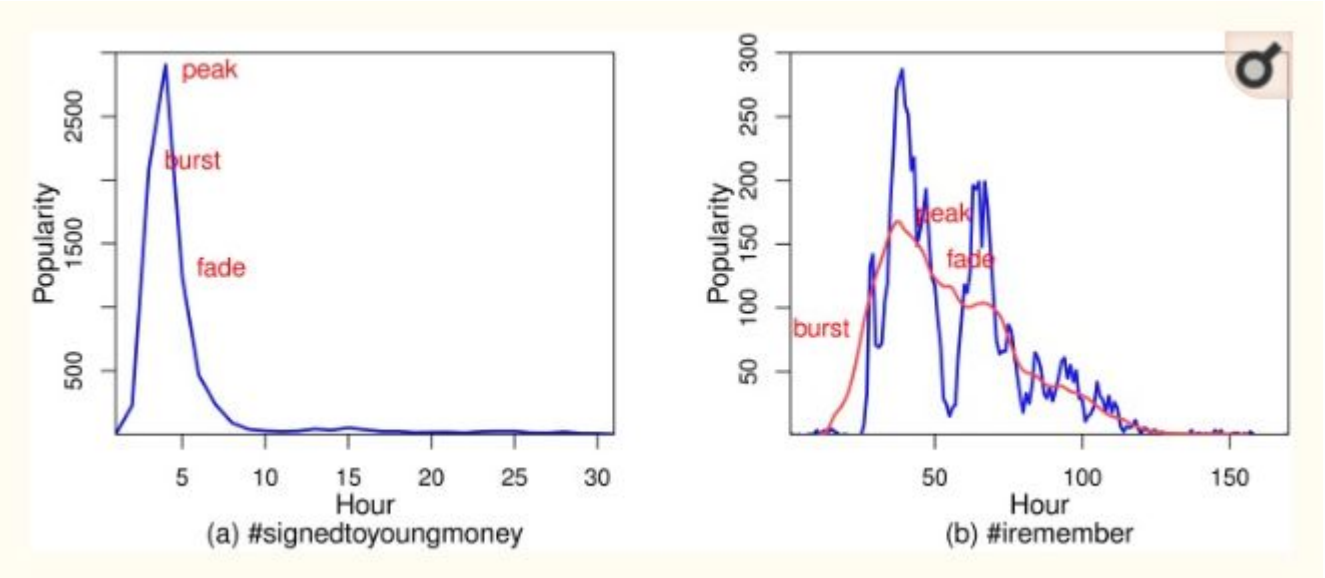
Puneet Auluck
Michele Bradley
Ahsanul Choudhury

Introduction

- Analyzing a dataset that contains 39644 URL links uploaded via Mashable.com
  - Summarizes features pertaining to articles published by Mashable in a two-year period
  - Include:
    - Days between Publication and Dataset Acquisition
    - Type of Article
    - Date of Publication
    - # Positive Words
    - # Negative Words
- Goal is to predict the number of shares in social networks (popularity) of an article

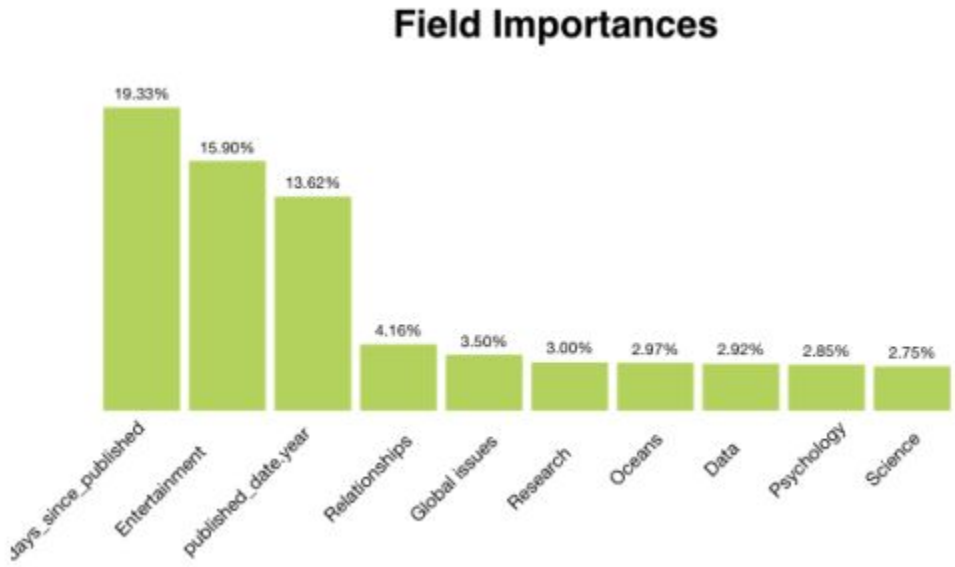
Literature Review - News and Hashtags

- Understanding the dimension of time in News Popularity Articles
- Three periods of evolution for a news article
  - Peak
  - Burst
  - Fade



Literature Review - TED Talks

- Topics have significant influence on popularity
- Positive topics such as entertainment or motivation are more likely to have the greatest number of views



Methodology

- Utilized open source statistics package R
- Generated basic statistics on variables
  - Mean, medians, standard deviations and etc
- Generated boxplots and density plots to visualize distributions
- Analyzed correlation between variables
- Transformed skewed data using logs and square-root functions
- Built models
  - Multiple Linear Regression
  - Tweedie Regression
  - Negative Binomial Regression
- Selected best model based on AIC value

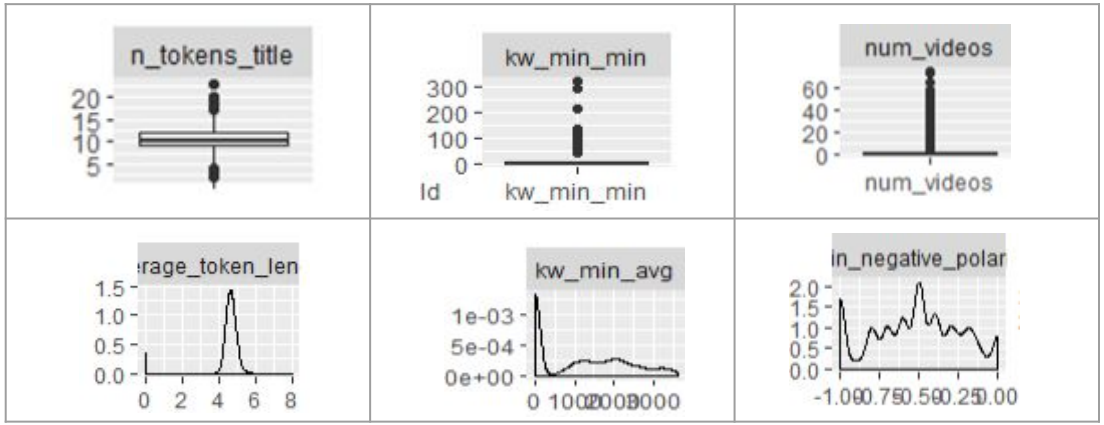
Data Exploration

- Basic statistics

	n	mean	median	sd	se	min	max	range	kurtosis
timedelta	29733	355.63	340.00	213.90	1.24	8.00	731.00	723.00	2
n_tokens_title	29733	10.40	10.00	2.11	0.01	2.00	23.00	21.00	3
n_tokens_content	29733	547.19	410.00	473.20	2.74	0.00	7764.00	7764.00	22
n_unique_tokens	29733	0.55	0.54	4.06	0.02	0.00	701.00	701.00	29663

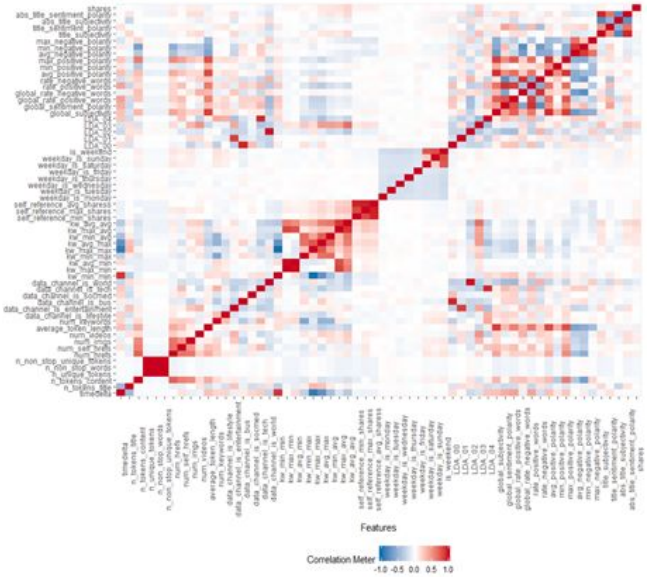
- Box plots
  - Identified outliers

- Density plots
  - Examined distribution



Multicollinearity

Discovered that many variables have correlation with other variables



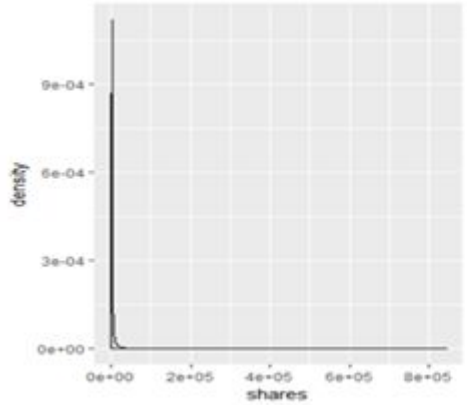
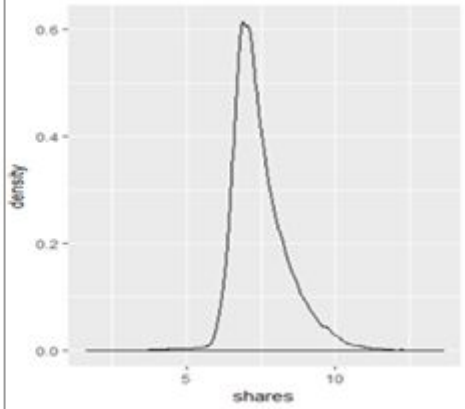
Index	Var1	Var2	value
1	kw_max_max	timedelta	-0.64
2	global_subjectivity	average_token_length	0.60
3	LDA_01	data_channel_is_entertainment	0.60
4	LDA_00	data_channel_is_bus	0.77
5	LDA_04	data_channel_is_tech	0.75
6	LDA_02	data_channel_is_world	0.84
7	kw_max_max	kw_min_min	-0.86
8	kw_avg_min	kw_max_min	0.94
9	kw_max_avg	kw_max_min	0.62
10	kw_avg_avg	kw_max_avg	0.81
11	self_reference_avg_sharess	self_reference_min_shares	0.81
12	self_reference_avg_sharess	self_reference_max_shares	0.85
13	is_weekend	weekday_is_saturday	0.66
14	is_weekend	weekday_is_sunday	0.70
15	avg_positive_polarity	global_subjectivity	0.63
16	rate_positive_words	global_sentiment_polarity	0.73
17	rate_negative_words	global_sentiment_polarity	-0.65
18	rate_positive_words	global_rate_positive_words	0.63
19	rate_negative_words	global_rate_negative_words	0.78
20	max_positive_polarity	avg_positive_polarity	0.70
21	min_negative_polarity	avg_negative_polarity	0.75
22	abs_title_sentiment_polarity	title_subjectivity	0.71

Data Preparation

No imputation required, all variables were complete

Normalized variables with log and square-root transformations

Dropped highly correlated variables

Before Transformation	*After Log Transformation*
	

Building Models

- Multiple Linear Regression
- Tweedie Regression
- Negative Binomial Regression

AIC Metrics

		AIC
Multiple Linear Regression	Forward Selection	75884.89
	Backward Elimination	75859.42
Tweedie Regression	Backward Elimination (.2)	177662.00
	Backward Elimination (.1)	177697.00
Negative Binomial Regression	Full Model	117850.10
	Reduced model	117838.40

Model Selection

•Selected multiple linear regression with backward elimination method

•Lowest AIC - 75859.42

$$\widehat{share\_log} = 1.48E{-}04 * timedelta + 5.18E{-}03 * n\_tokens\_title + 5.87E{-}03 * num\_hrefs - 9.08E{-}03 * num\_self\_hrefs + 3.09E{-}03 * num\_imgs - 9.06E{-}02 * average\_token\_length + 1.03E{-}02 * num\_keywords - 5.56E{-}05 * kw\_avg\_min - 3.55E{-}07 * kw\_avg\_max + 4.12E{-}01 * global\_subjectivity - 1.39E{-}01 * avg\_positive\_polarity - 1.14E{-}01 * avg\_negative\_polarity + 4.63E{-}02 * title\_subjectivity + 7.53E{-}02 * title\_sentiment\_polarity - 2.31E{-}01 * news\_typeEntertainment - 3.05E{-}03 * news\_typeLifestyle + 2.24E{-}01 * news\_typeSocialMedia + 1.24E{-}01 * news\_typeTech - 1.58E{-}01 * news\_typeWorld - 8.43E{-}03 * news\_dayMonday + 1.99E{-}01 * news\_daySaturday + 2.06E{-}01 * news\_daySunday - 7.31E{-}02 * news\_dayThursday - 7.05E{-}02 * news\_dayTuesday - 7.69E{-}02 * news\_dayWednesday + 1.89E{-}02 * kw\_avg\_avg\_sqrt + 1.04E{-}03 * self\_reference\_min\_shares\_sqrt + 1.27E{-}03 * self\_reference\_avg\_sharess\_sqrt$$

Predictions

	shares	Backward Selection
10	1100	492
13	3200	968
18	2400	588
23	1900	809
32	1200	770
33	1700	846
38	3400	857
49	308	812
50	1200	801
57	569	691
58	1100	809
64	1400	762
66	730	834
73	658	517
81	1400	819

Conclusion

- The boxplots and density plots helped identify the skewness and outliers
- From the knowledge of distributions, transformations were applied which helped bring accuracy of the model
- By identifying correlations, we were able to drop variables and thus minimize errors within the model
- Even after transformations and variable selection, the selected model is not fully ideal. The resulted prediction values are not close to the actual values
- Always room for improvements and explore other techniques