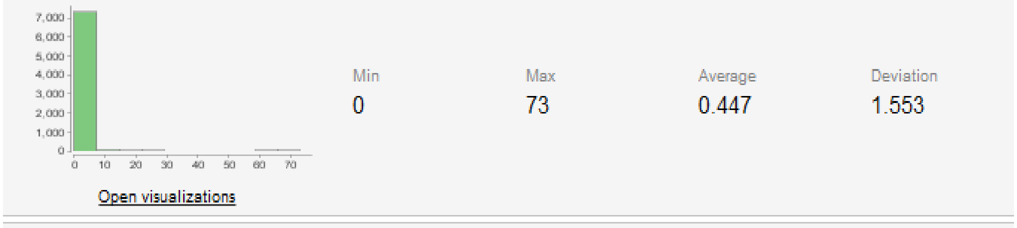
Question 1 As the category leader, before any analysis, you want to understand the data and check for clear patterns. This is a helpful exercise, since it does not make any model assumptions, and the plots can be used in communications and presentations. Report either in plots or in a table the summary statistics (5 points) of mean and standard deviation, and comment on how popular (vs non-popular) news articles differ on the following specific variables are of interest

a. **Number of Videos – this approximates the visual movement in the news**

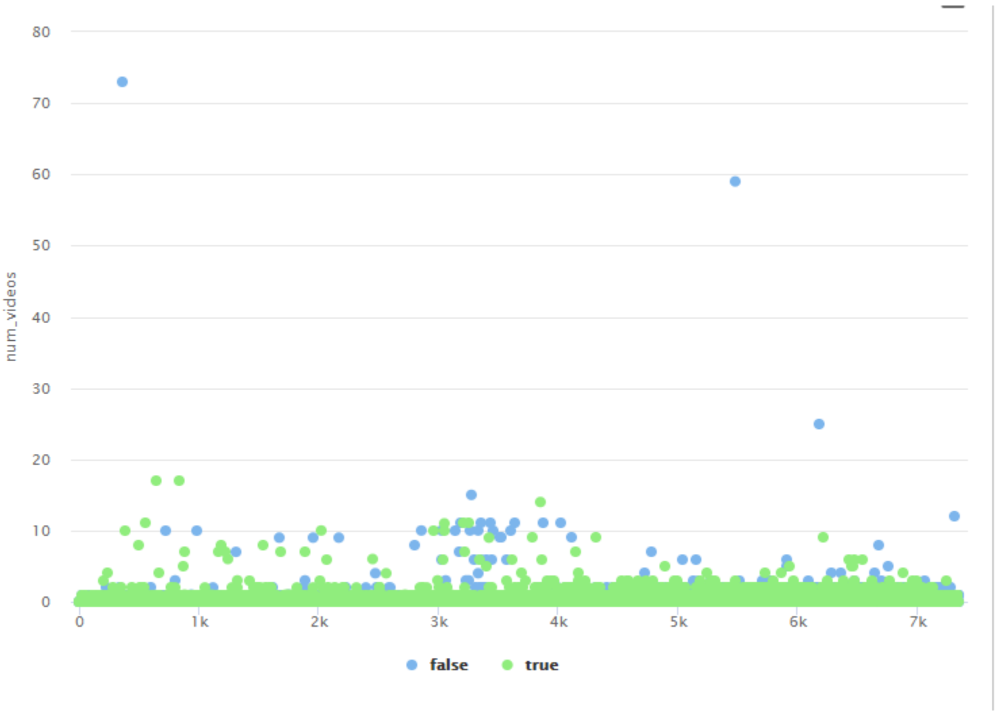
Summary statistics



Min	Max	Average	Deviation
0	73	0.447	1.553

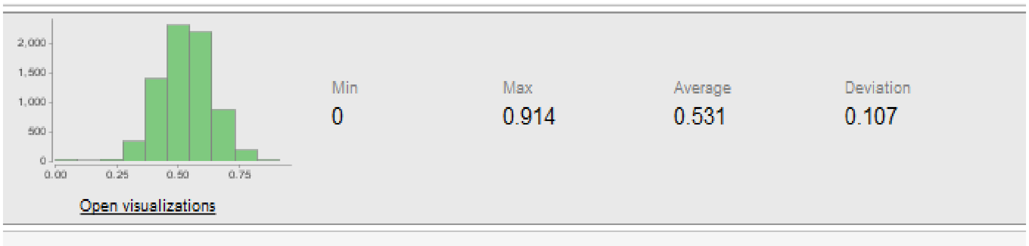
Open visualizations

When there are no videos (num_videos =0) popular field is shown as false, the minimum average number of videos for a channel to be popular is 3 as shown in the plot below



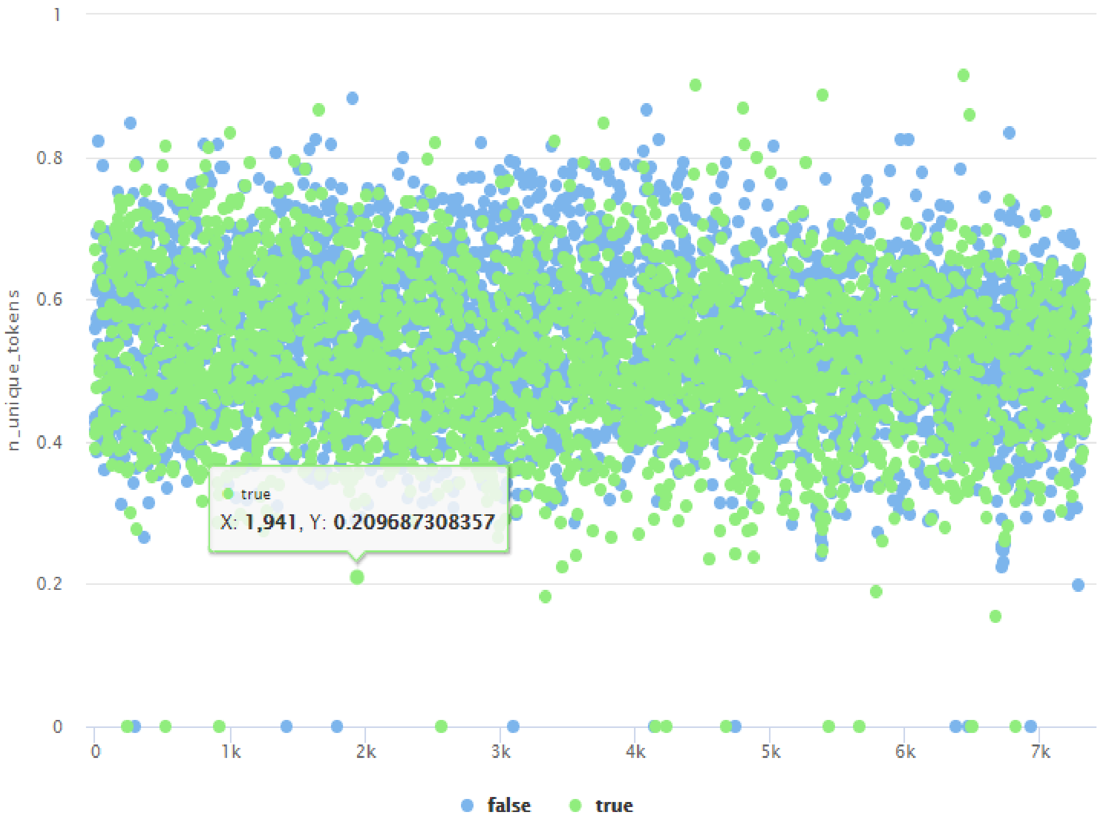
b. **Number of unique tokens – this approximates the complexity of the news**

Summary statistics



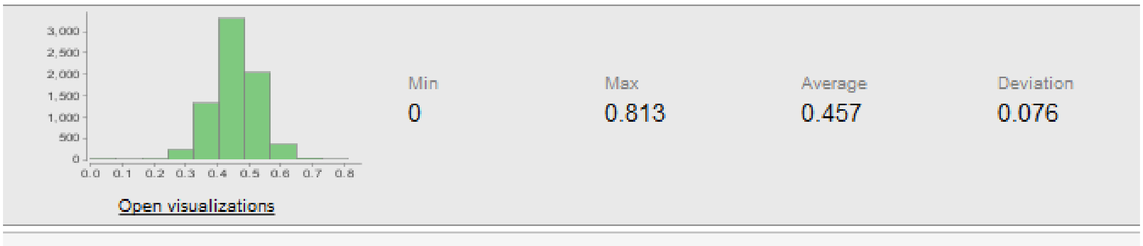
Min	Max	Average	Deviation
0	0.914	0.531	0.107

The number of unique tokens attributes have weak correlation to whether the channel is popular or not as shown on the plot below



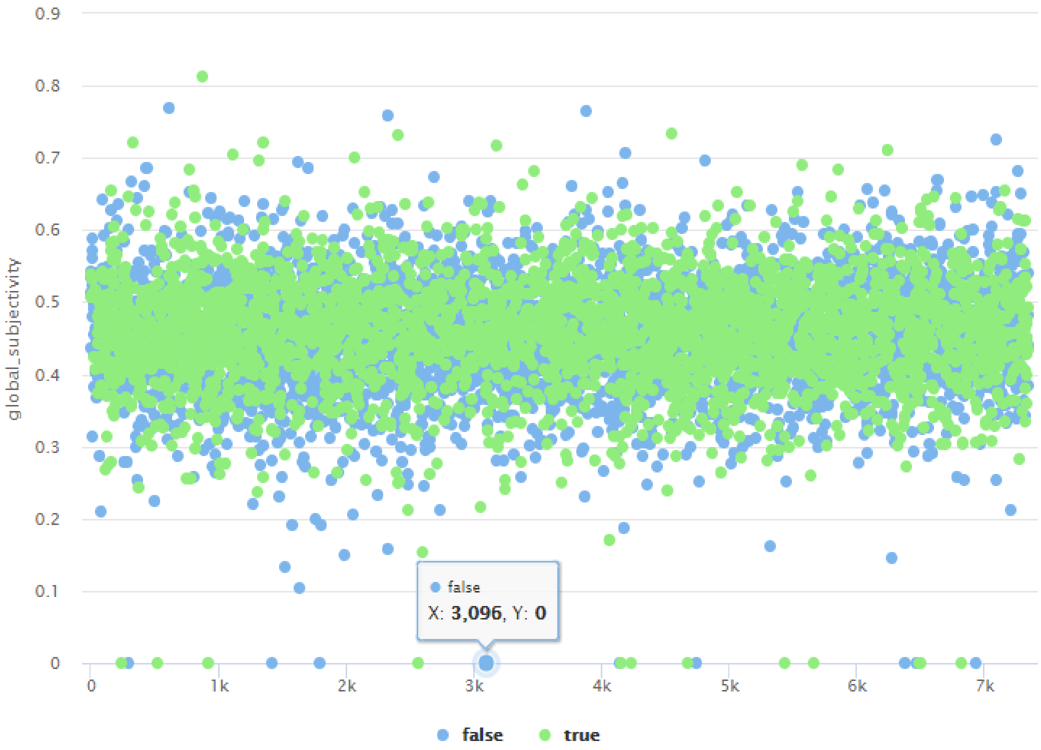
c. Subjectivity – this approximates the factual vs opinion content in news (global_subjectivity)

Summary Statistics



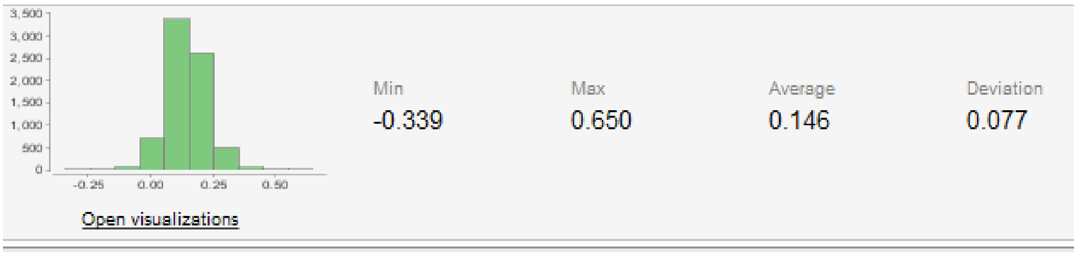
Min	Max	Average	Deviation
0	0.813	0.457	0.076

If the global_subjectivity is between 0.4 and 0.5, the popular attribute is true as shown on the high concentration of green dots on the region.



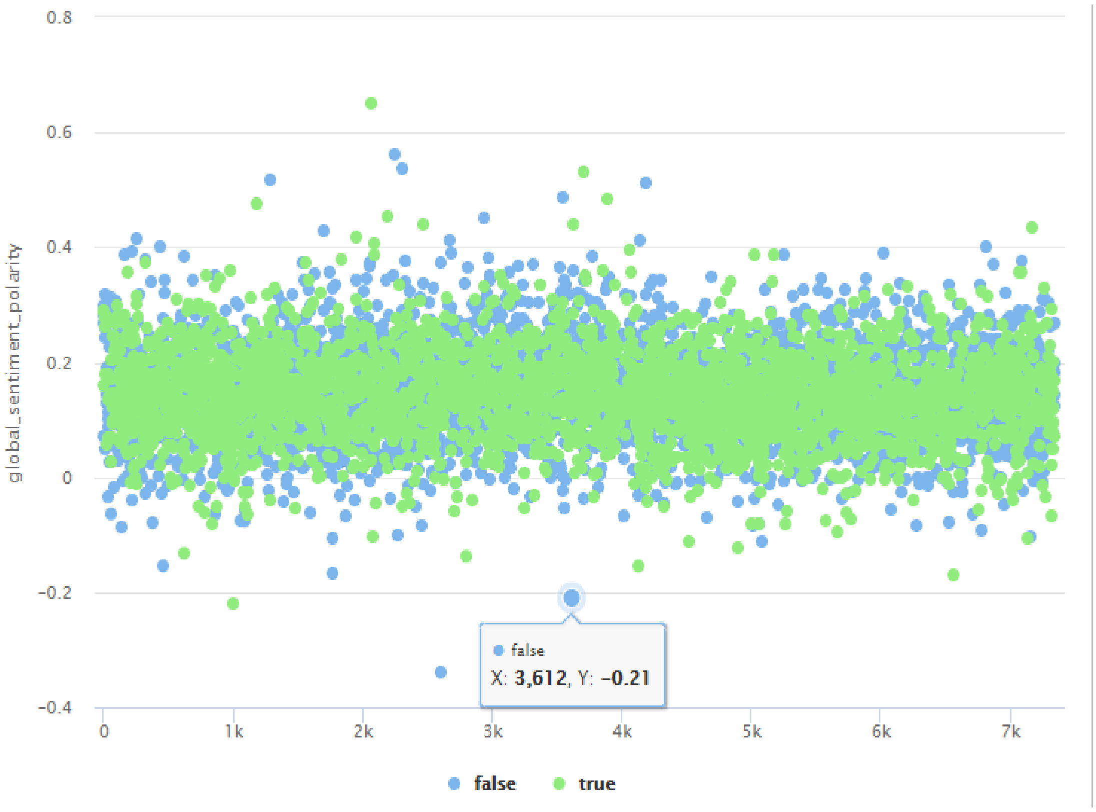
d. **Polarity – this approximates the positive vs negative sentiment in the news (global_polarity)**

Summary statistics

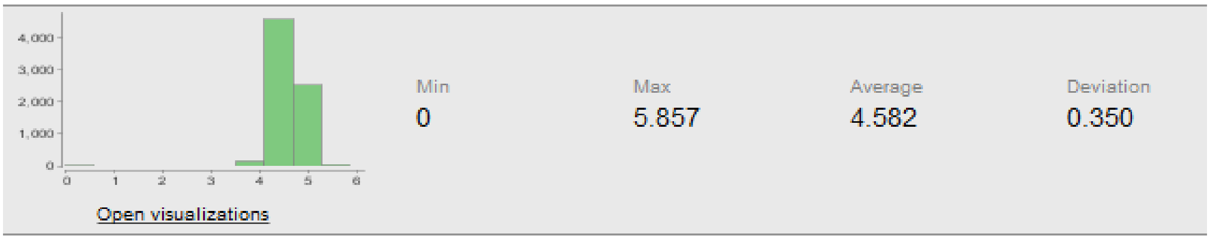


Min	Max	Average	Deviation
-0.339	0.650	0.146	0.077

The average global_polarity value for a channel to be popular is 0.2 and the value goes further up or below 0.2, the popular attribute becomes false



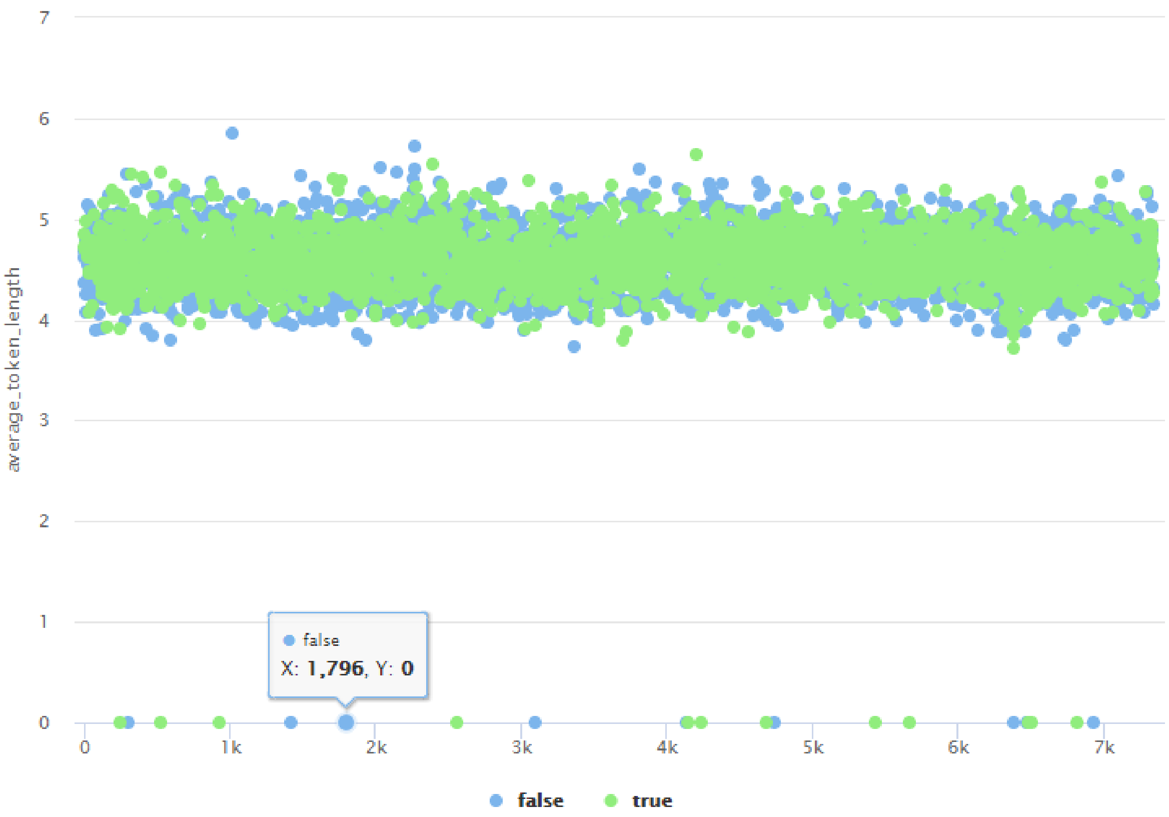
e. Length of title – this approximates the complexity of the news title
Summary statistics



Open visualizations

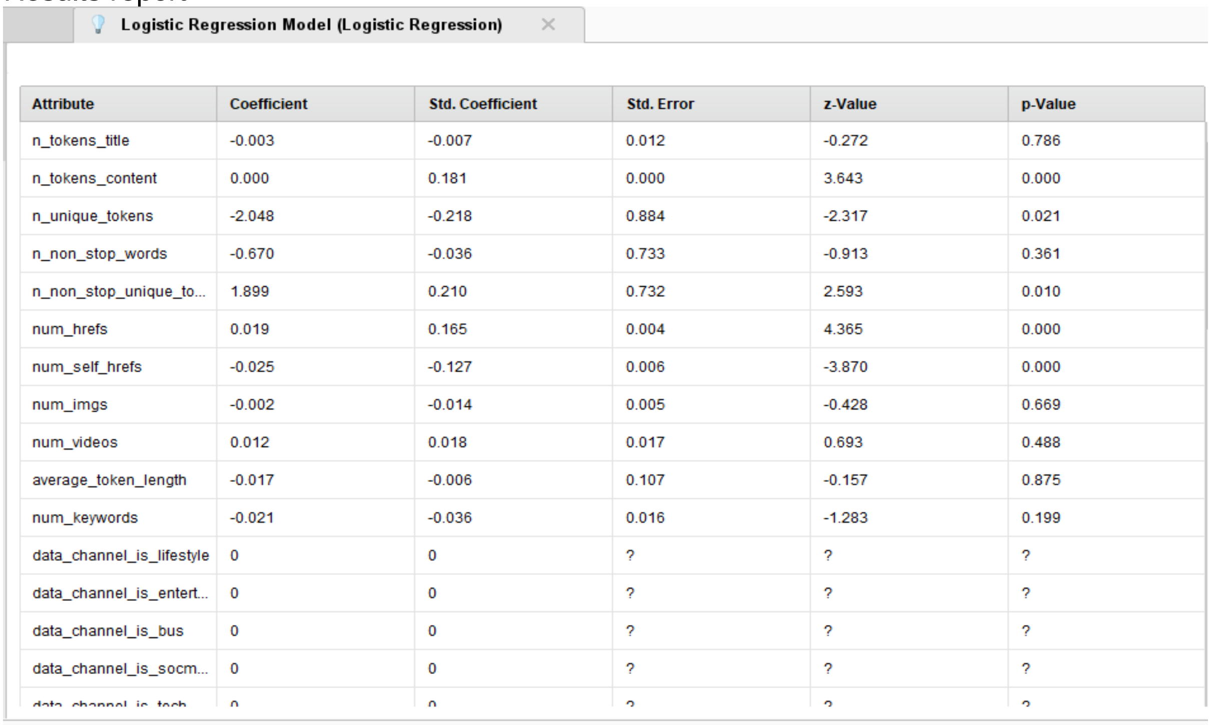
Min	Max	Average	Deviation
0	5.857	4.582	0.350

The average token length for when the channel is popular has to be between 4 and 5.



Question 2 You are also interested in how the different variables are associated with news popularity. For this, you turn to the benchmark method with binomial data – logistic regressions. For this analysis, it's important to refer to the note-setting up the data for the list of variables to exclude. Report the results from logistic regression (5 points), and comment on the following specific variables:

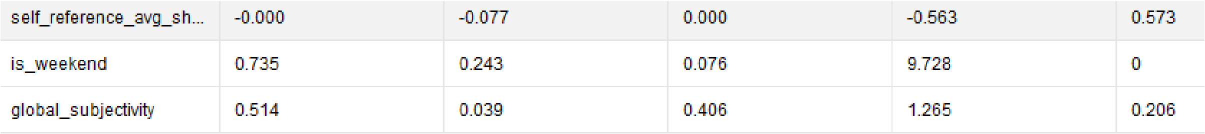
Results report



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
n_tokens_title	-0.003	-0.007	0.012	-0.272	0.786
n_tokens_content	0.000	0.181	0.000	3.643	0.000
n_unique_tokens	-2.048	-0.218	0.884	-2.317	0.021
n_non_stop_words	-0.670	-0.036	0.733	-0.913	0.361
n_non_stop_unique_to...	1.899	0.210	0.732	2.593	0.010
num_hrefs	0.019	0.165	0.004	4.365	0.000
num_self_hrefs	-0.025	-0.127	0.006	-3.870	0.000
num_imgs	-0.002	-0.014	0.005	-0.428	0.669
num_videos	0.012	0.018	0.017	0.693	0.488
average_token_length	-0.017	-0.006	0.107	-0.157	0.875
num_keywords	-0.021	-0.036	0.016	-1.283	0.199
data_channel_is_lifestyle	0	0	?	?	?
data_channel_is_entert...	0	0	?	?	?
data_channel_is_bus	0	0	?	?	?
data_channel_is_socm...	0	0	?	?	?
data_channel_is_tech	0	0	?	?	?

a. What is the weekend effect (2.5 points)? I.e., what is the odds multiplier forthe variable is_weekend?

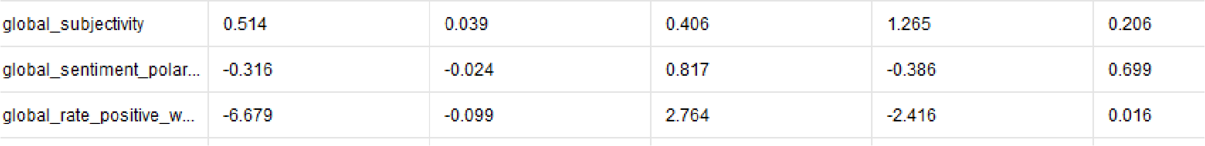
Odds multiplier: 0.735



self_reference_avg_sh...	-0.000	-0.077	0.000	-0.563	0.573
is_weekend	0.735	0.243	0.076	9.728	0
global_subjectivity	0.514	0.039	0.406	1.265	0.206

b. What is the polarity effect (2.5 points)? I.e., what is the odds multiplier for the variable global_sentiment_polarity?

Odds multiplier: -0.316



global_subjectivity	0.514	0.039	0.406	1.265	0.206
global_sentiment_polar...	-0.316	-0.024	0.817	-0.386	0.699
global_rate_positive_w...	-6.679	-0.099	2.764	-2.416	0.016

c. In your analysis, which variables are statistically significant at 5% threshold that you find increase the odds and which variables decrease the odds of a news article being popular? (5 points)

The following variables are statistically significant because their p-values are way less than 0.05

is_weekend (p-value=0.00)

num_hrefs (p-value=0.00)

num_self_hrefs(p-value=0.00)

kw_min_min(p-value=0.00)

kw_min_avg(p-value=0.00)

kw_max_avg(p-value=0.00)

kw_avg_avg(p-value=0.00)

The following variables are statistically significant because their p-values are way above 0.05

title_subjectivity(p-value=0.549)

global_rate_negative_words(p-value=0.712)

kw_avg_min(p-value = 0.954)

kw_min_max(p-value=0.867)

kw_max_max(p-value=0.886)

global_rate_negative_words(p-value=0.712) global_sentiment_polarity(p-value=0.669)

average_token_length(p-value=0.875)

num_imgs(p-value=0.669)

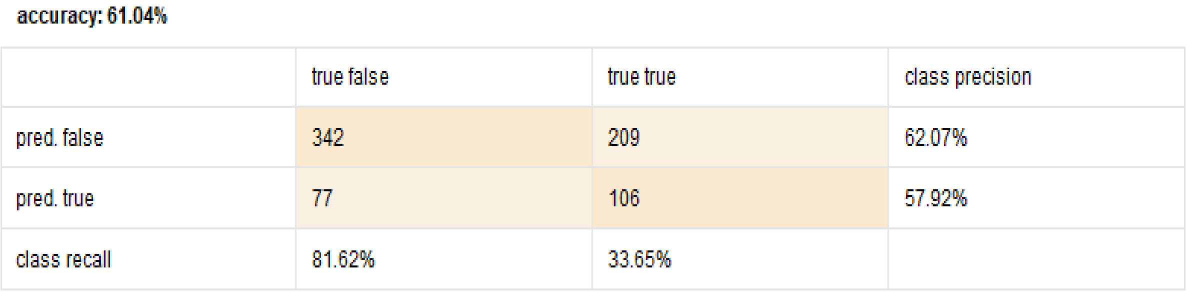
n_tokens_title(p-value=0.786)

Question 3]: You are now interested in developing a prediction model than predicts whether a news article will be popular or not. For this analysis, split the data in 90% train and 10% test, and predict using Logistic Regression, Random Forests (max depth: 10, max trees: 100) , and neural network (use operator “Deep Learning”, max layers = 5, max neurons per layer = 500).

For comparison, we use the Apply Model operator with our chosen Model along with the performance operator to see the confusion matrix.

a. As an overall predictive accuracy comparison, report the f1 scores (class = true) for the three methods. Which method is preferred from an overall predictive accuracy perspective? (5 points)

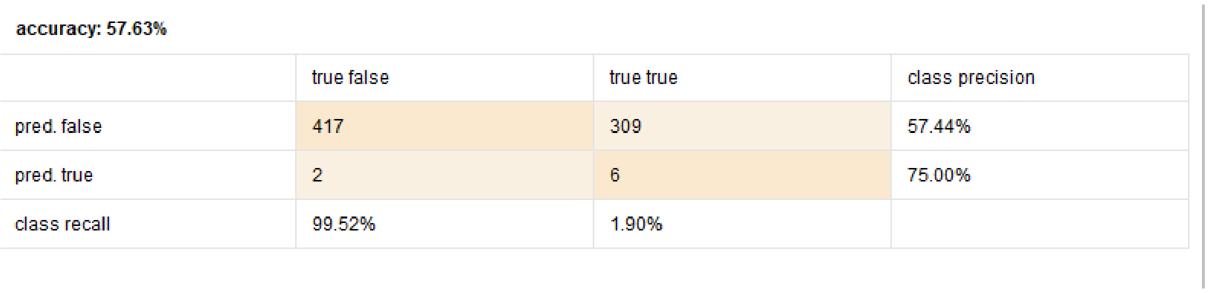
Logistic Regression



accuracy: 61.04%

	true false	true true	class precision
pred. false	342	209	62.07%
pred. true	77	106	57.92%
class recall	81.62%	33.65%	

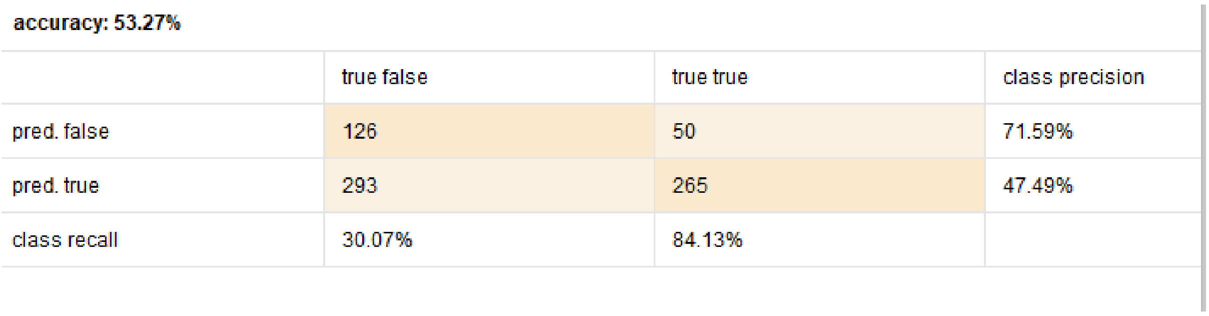
Random Forests



accuracy: 57.63%

	true false	true true	class precision
pred. false	417	309	57.44%
pred. true	2	6	75.00%
class recall	99.52%	1.90%	

Neural Networks



accuracy: 53.27%

	true false	true true	class precision
pred. false	126	50	71.59%
pred. true	293	265	47.49%
class recall	30.07%	84.13%	

Using accuracy Logistic Regression is preferred because it has the highest accuracy of 61.04% compared to Random Forests (57.63%) and Neural Networks(53.27%)

b. Imagine a situation where you worry about missing out on potentially popular news articles. I.e., recall (class=true) is the main metric of interest. Which of the three methods works best in this context? (5 points)

Neural Networks have the highest class=true metric of 84.13% compared to Logistic Regression(33.65%) and Random Forests(1.90%). The best method in this context would therefore be Neural Networks.

c. Imagine a situation where the cost and budgeting team asks you to prioritize marketing spends on promoting news articles. I.e., your main metric of interest is precision (class = true). Which if the three methods works best in this context? (5 points)

Random Forests have the highest precision(class=true) of 75% and this method would work best in this context as compared to Logistic Regression(57.92%) and Neural Networks(47.49%).