TASK GOAL

Use machine learning methods to understand drivers of popular news stories on the news website and develop a predictive model for optimizing marketing spends as well as predicting future popular news articles.

DATA PREPARATION

A csv file containing data attributes that we are interested in examining is first loaded into Rapid Miner.

Row No.	url	timedelta	n_tokens_tit...	n_tokens_c...	n_unique_to...	n_non_stop...	n_non_stop...
1	http://mashab...	731	12	219	0.664	1.000	0.815
2	http://mashab...	731	9	255	0.605	1.000	0.792
3	http://mashab...	731	9	211	0.575	1.000	0.664
4	http://mashab...	731	9	531	0.504	1.000	0.666
5	http://mashab...	731	13	1072	0.416	1.000	0.541
6	http://mashab...	731	10	370	0.560	1.000	0.698
7	http://mashab...	731	8	960	0.418	1.000	0.550
8	http://mashab...	731	12	989	0.434	1.000	0.572
9	http://mashab...	731	11	97	0.670	1.000	0.837
10	http://mashab...	731	10	231	0.636	1.000	0.797
11	http://mashab...	731	9	1248	0.490	1.000	0.732
12	http://mashab...	731	10	187	0.667	1.000	0.800
13	http://mashab...	731	9	274	0.609	1.000	0.708

ExampleSet (39,644 examples, 0 special attributes, 61 regular attributes)

Then we filter the data so that we can only do the analysis on category of technology where the attribute data_channel_is_tech = 1.

average_tok...	num_keywo...	data_chann...	data_chann...	data_chann...	data_chann...	data_chann...	data_chann...	kw_min_min
4.683	7	0	0	0	0	1	0	0
4.359	9	0	0	0	0	1	0	0
4.618	9	0	0	0	0	1	0	0
4.856	7	0	0	0	0	1	0	0
4.717	8	0	0	0	0	1	0	0
4.687	10	0	0	0	0	1	0	0
4.630	7	0	0	0	0	1	0	0
4.259	10	0	0	0	0	1	0	0
4.782	9	0	0	0	0	1	0	0
4.636	9	0	0	0	0	1	0	0
4.986	7	0	0	0	0	1	0	0
4.069	9	0	0	0	0	1	0	0
4.752	10	0	0	0	0	1	0	0
4.728	9	0	0	0	0	1	0	0

ExampleSet (7,346 examples, 0 special attributes, 61 regular attributes)

An outcome variable name popular (added as an attribute to the results table) indicating whether the new channel is popular or not is generated using the function call (“if(shares>=2000,TRUE,FALSE)”.

tiv...	avg_negativ...	min_negativ...	max_negati...	title_subject...	title_sentim...	abs_title_su...	abs_title_se...	shares	popular
	-0.220	-0.500	-0.050	0.455	0.136	0.045	0.136	505	false
	-0.195	-0.400	-0.100	0.643	0.214	0.143	0.214	855	false
	-0.243	-0.500	-0.050	1	0.500	0.500	0.500	891	false
	-0.125	-0.125	-0.125	0.125	0	0.375	0	3600	true
	-0.227	-0.500	-0.050	0.500	0	0	0	17100	true
	-0.207	-0.500	-0.050	0	0	0.500	0	2800	true
	-0.230	-0.500	-0.050	0	0	0.500	0	445	false
	-0.117	-0.200	-0.050	0.900	0.400	0.400	0.400	783	false
	-0.264	-0.500	-0.125	0	0	0.500	0	1500	false
	-0.202	-0.500	-0.050	0.500	0.500	0	0.500	1800	false
	-0.342	-0.800	-0.100	0	0	0.500	0	3900	true
	-0.178	-0.400	-0.008	0	0	0.500	0	480	false
	-0.230	-0.600	-0.050	1	-0.600	0.500	0.600	7700	true
	-0.215	-0.500	-0.050	0	0	0.500	0	1100	false

ExampleSet (7,346 examples, 0 special attributes, 62 regular attributes)

Attributes that are not relevant to the analysis shown below are excluded

a. url
b. timedelta
c. weekday_is_monday: Was the article published on a Monday?
d. weekday_is_tuesday: Was the article published on a Tuesday?
e. weekday_is_wednesday: Was the article published on a Wednesday?
f. weekday_is_thursday: Was the article published on a Thursday?
g. weekday_is_friday: Was the article published on a Friday?
h. weekday_is_saturday: Was the article published on a Saturday?
i. weekday_is_sunday: Was the article published on a Sunday?
j. LDA_00
k. LDA_01
l. LDA_02
m. LDA_03
n. LDA_04
o. rate_positive_words
p. rate_negative_words
q. shares

iv...	max_positiv...	avg_negativ...	min_negativ...	max_negati...	title_subject...	title_sentim...	abs_title_su...	abs_title_se...	popular
	1	-0.220	-0.500	-0.050	0.455	0.136	0.045	0.136	false
	0.600	-0.195	-0.400	-0.100	0.643	0.214	0.143	0.214	false
	1	-0.243	-0.500	-0.050	1	0.500	0.500	0.500	false
	0.800	-0.125	-0.125	-0.125	0.125	0	0.375	0	true
	1	-0.227	-0.500	-0.050	0.500	0	0	0	true
	1	-0.207	-0.500	-0.050	0	0	0.500	0	true
	1	-0.230	-0.500	-0.050	0	0	0.500	0	false
	0.350	-0.117	-0.200	-0.050	0.900	0.400	0.400	0.400	false
	1	-0.264	-0.500	-0.125	0	0	0.500	0	false
	1	-0.202	-0.500	-0.050	0.500	0.500	0	0.500	false
	0.600	-0.342	-0.800	-0.100	0	0	0.500	0	true
	1	-0.178	-0.400	-0.008	0	0	0.500	0	false
	1	-0.230	-0.600	-0.050	1	-0.600	0.500	0.600	true
	1	-0.215	-0.500	-0.050	0	0	0.500	0	false

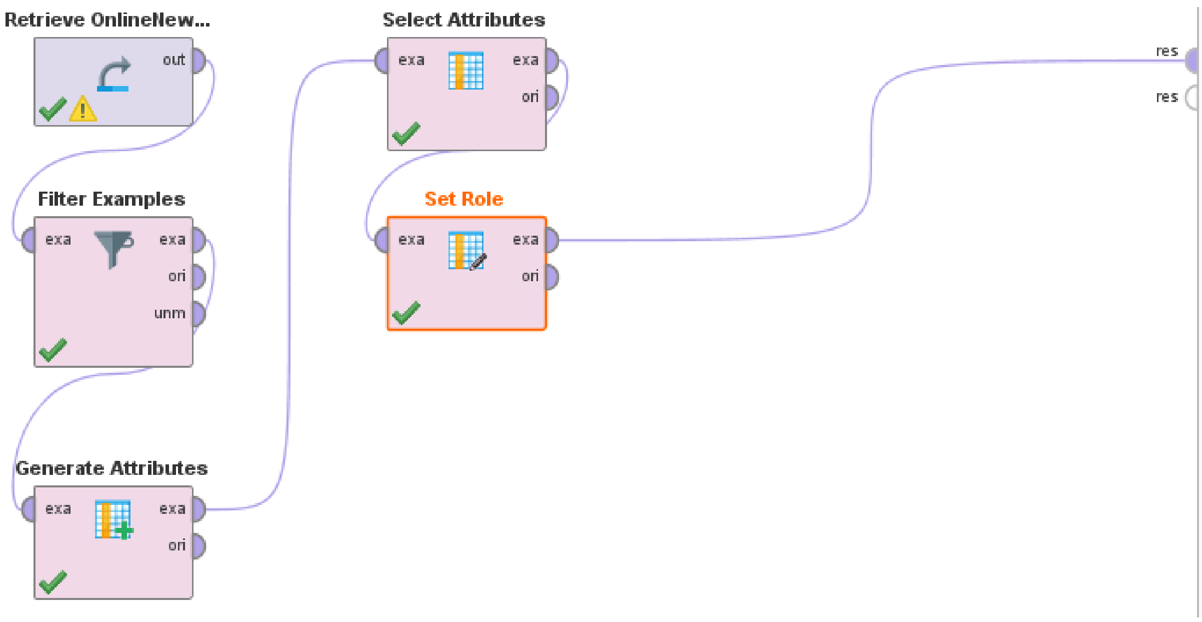
ExampleSet (7,346 examples, 0 special attributes, 45 regular attributes)

Since the predictive analysis is done on the attribute popular, we specify popular variable as the “label” variable.

Row No.	popular	n_tokens_tit...	n_tokens_c...	n_unique_to...	n_non_stop...	n_non_stop...	num_hrefs	num_self_h...	num
1	false	13	1072	0.416	1.000	0.541	19	19	20
2	false	10	370	0.560	1.000	0.698	2	2	0
3	false	12	989	0.434	1.000	0.572	20	20	20
4	true	11	97	0.670	1.000	0.837	2	0	0
5	true	8	1207	0.411	1.000	0.549	24	24	42
6	true	13	1248	0.391	1.000	0.523	21	19	20
7	false	11	1154	0.427	1.000	0.573	20	20	20
8	false	8	266	0.573	1.000	0.721	5	2	1
9	false	8	331	0.563	1.000	0.724	5	3	1
10	false	12	1225	0.385	1.000	0.509	22	22	28
11	true	10	633	0.476	1.000	0.580	2	2	19
12	false	14	290	0.612	1.000	0.762	0	0	14
13	true	10	1244	0.418	1.000	0.563	27	22	20
14	false	10	1036	0.430	1.000	0.560	21	21	20

ExampleSet (7,346 examples, 1 special attribute, 44 regular attributes)

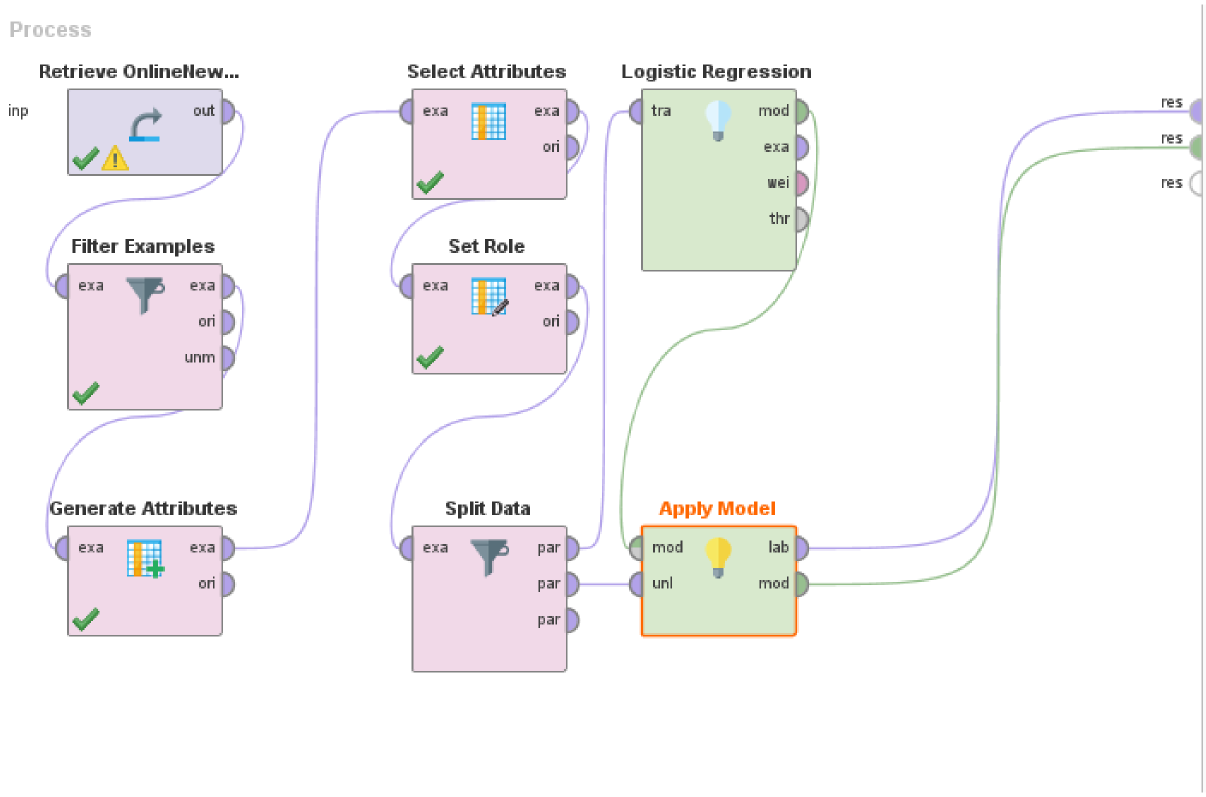
The initial data preparation design is shown below:



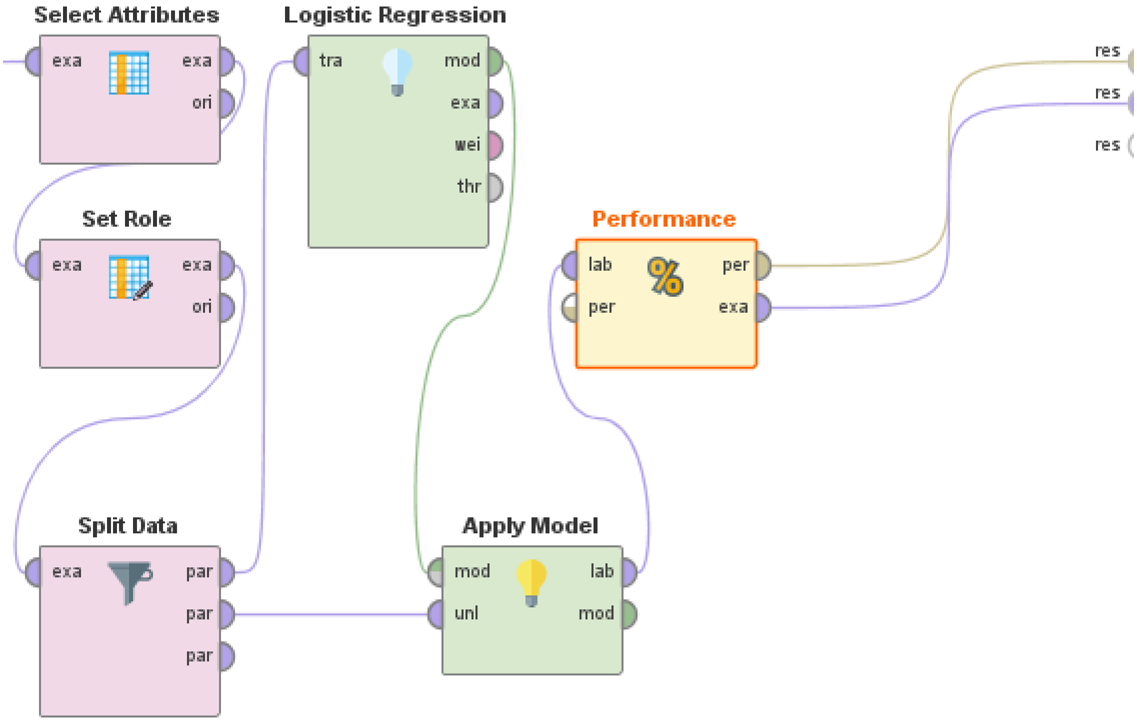
Logistic Regression

Logistic Regression
In designing our logistic regression after the role is set, a split of the data is done into a training set(70%) and a testing set(30%) and use the Logistic regression operator to build the model (with the training partition connected to our model operator). Finally the model is applied to both the testing set to measure the performance.

P1



To check how well the model does, performance(classification) operator is used:
P2



P1 Results

Row No.	popular	prediction(p...	confidence(f...	confidence(t...	n_tokens_tit...	n_tokens_c...	n_unique_to...	n_non_stop...	n_nc
1	false	false	0.889	0.111	6	174	0.692	1.000	0.90
2	false	false	0.924	0.076	13	1024	0.428	1.000	0.55
3	false	false	0.744	0.256	9	268	0.477	1.000	0.58
4	false	false	0.709	0.291	7	925	0.428	1.000	0.54
5	false	false	0.742	0.258	9	965	0.435	1.000	0.56
6	false	false	0.615	0.385	12	981	0.434	1.000	0.57
7	true	false	0.649	0.351	10	951	0.437	1.000	0.57
8	true	false	0.628	0.372	11	1364	0.388	1.000	0.53
9	false	false	0.501	0.499	12	1298	0.413	1.000	0.60
10	false	false	0.528	0.472	9	652	0.532	1.000	0.72
11	false	false	0.647	0.353	8	633	0.464	1.000	0.59
12	false	false	0.715	0.285	13	109	0.663	1.000	0.73
13	false	false	0.601	0.399	10	364	0.519	1.000	0.68
14	true	false	0.575	0.425	12	193	0.591	1.000	0.70