Predict the popularity of the Online news

URL: (https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity)

Name: Zhihao Liu

Netid: zliu91

Group: Group 5

Introduction

Due to the Web expansion, the prediction of online news popularity is becoming a trendy research topic. During this analysis, we choose a large and recently collected dataset with over 39,797 articles from the Mashable website with 61 variables. These variables include 58 predictive attributes, 2 non-predictive and 1 goal named **shares**. However, I want to focus on the problem that what kind of features that an article has can make it to be popular. So I transformed the task into a binary task using a decision threshold of 1400. And I create a new response named "popularity", when shares >=1400, the article is defined to be popular and the "popularity" is 1 and when shares < 1400, the article is defined to be popular and the "popularity" is 0.

There are 58 predictive attributes in this dataset, I think it is too many. So I decide to merge some variables. I merge the articles published from Monday to Friday as article published on weekday, and articles published on Saturday and Sunday as articles published on weekend. After merging, there are 51 variables as predictors, and one binary outcome "popularity" as response.

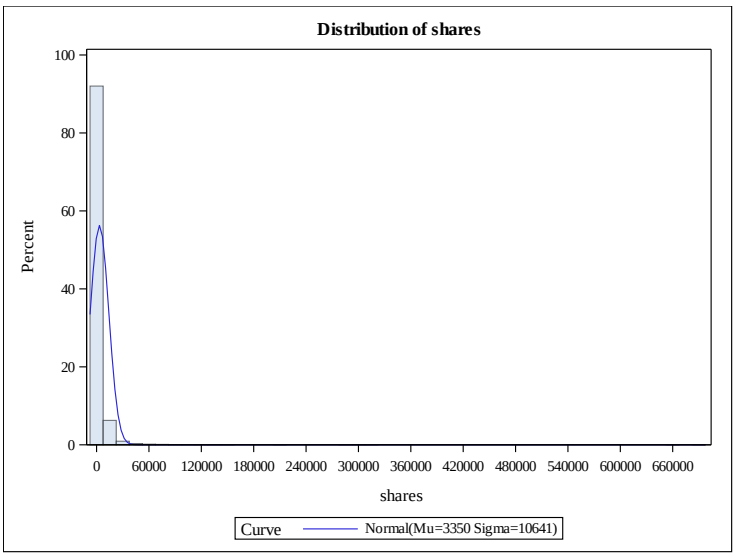
Moments			
N	39797	**Sum Weights**	39797
Mean	3350	**Sum Observations**	66403700
Std Deviation	10641.1124	**Variance**	113233272
Skewness	31.9597406	**Kurtosis**	1710.36259
Uncorrected SS	2.46685E12	**Corrected SS**	2.2444E12
Coeff Variation	317.645145	**Std Error Mean**	75.5811148

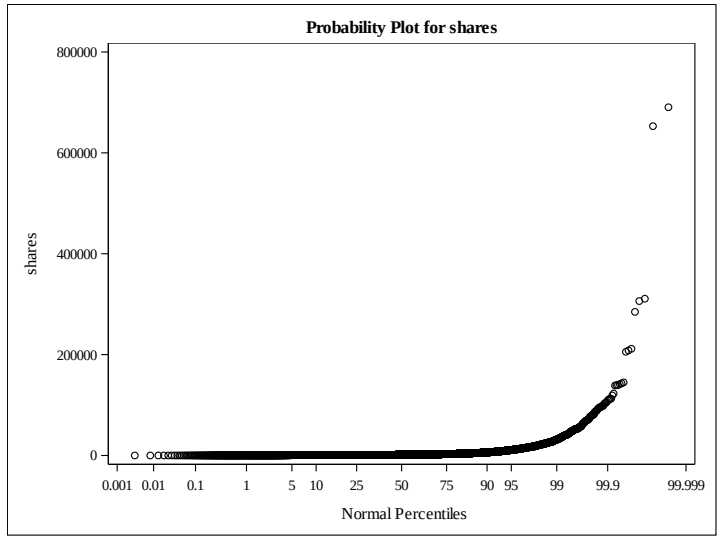
(Table 1)

Basic Statistical Measures			
Location		**Variability**	
Mean	3350.000	**Std Deviation**	10641
Median	1400.000	**Variance**	113233272
Mode	1100.000	**Range**	690399
		Interquartile Range	1855

(Table 2)

Table 1 and Table 2 provide the basic description for observations for **shares**. We can see there is no missing value. And the mean is 3350 and median is 1400. Because mean is larger than median, and the Skewness is 31.9597, so there is positive skew in the dataset. And the range is very large, it is 690399.





The distribution of shares and probability plot for shares are shown above. We can see that the distribution for shares has long right tail and the probability plots tell us that almost articles have shares less than 20000.

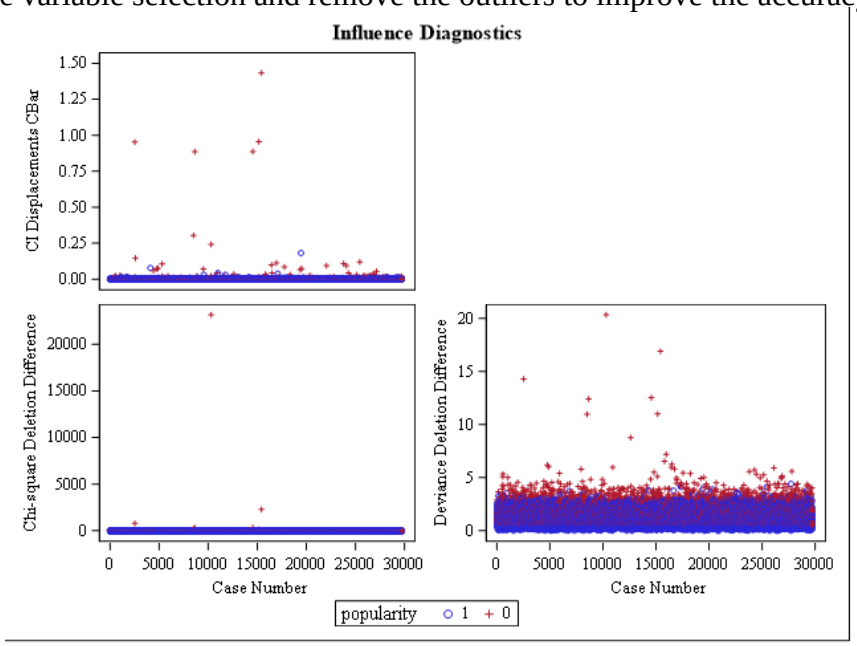
Algorithm and metrics

During this project, because the response is a binary outcome, so I choose Logistic Regression as benchmark algorithm. And I also choose Support Vector Machine as another model to do prediction and make comparison. And in this project, I choose misclassification error between predicted value and actual value as evaluation metrics. And I also split the dataset into train data and test data, 0.75/0.25 split randomly.

Logistic regression

Training model

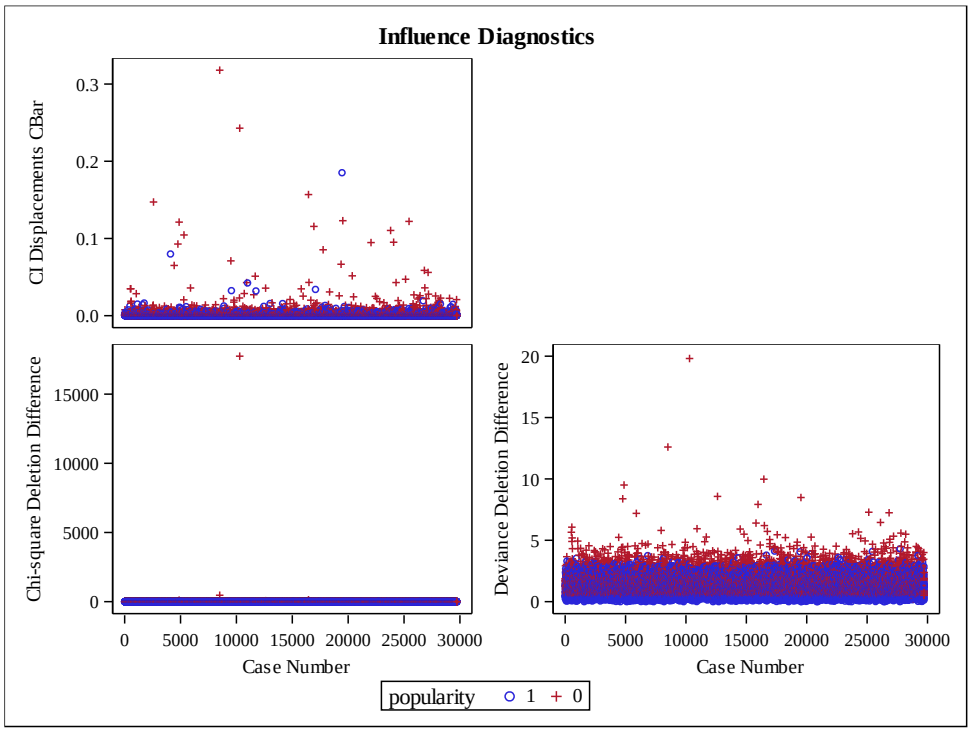
Firstly, I use training data to build model to do prediction. I use stepwise selection and diagnostic plots to make variable selection and remove the outliers to improve the accuracy.



Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	kw_avg_avg		1	1	773.7542		<.0001
2	is_weekend		1	2	557.1502		<.0001
3	data_channel_is_tech		1	3	547.8759		<.0001
4	data_channel_is_socmed		1	4	477.6287		<.0001
5	LDA_00		1	5	278.4454		<.0001
6	kw_min_min		1	6	230.3168		<.0001
7	kw_max_avg		1	7	247.7872		<.0001
8	num_keywords		1	8	192.0168		<.0001
9	num_hrefs		1	9	89.5726		<.0001
10	LDA_04		1	10	74.2705		<.0001
11	kw_min_avg		1	11	58.9695		<.0001
12	self_reference_avg_sharess		1	12	44.9899		<.0001
13	n_tokens_content		1	13	29.9859		<.0001
14	data_channel_is_entertainment		1	14	33.0666		<.0001
15	kw_avg_max		1	15	37.8305		<.0001
16	average_token_length		1	16	21.5450		<.0001
17	global_subjectivity		1	17	36.8294		<.0001
18	num_self_hrefs		1	18	14.5066		0.0001
19	min_positive_polarity		1	19	12.6096		0.0004
20	kw_max_max		1	20	11.2907		0.0008
21	num_imgs		1	21	7.2570		0.0071
22	title_sentiment_polarity		1	22	6.5344		0.0106
23	abs_title_subjectivity		1	23	9.1849		0.0024
24	data_channel_is_bus		1	24	5.3420		0.0208
25	kw_min_max		1	25	4.6271		0.0315
26	self_reference_min_shares		1	26	4.6540		0.0310
27	title_subjectivity		1	27	3.9998		0.0455

Based on the variable selection result, there are 27 variables are kept in the model. And according to the diagnostic plot, we can see that most observations have Cbar less than 0.25, and there are 5 observations have Cbar larger than 0.75. So I decide to remove there 5 observations.

After removing 5 outliers we rebuild the model.



Based on the diagnostics plot above, we can see that there is no outliers anymore.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	41071.451	37056.384
SC	41079.751	37288.779
-2 Log L	41069.451	37000.384

(Table 3)

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	4069.0677	27	<.0001
Score	3697.6366	27	<.0001
Wald	3316.6511	27	<.0001

(Table 4)

The AIC is noticeably lower for the Logistic model (37056.384) than for the intercept only model (41071.451), and the global tests all agree that the parameter is different from 0 at 5% statistical significance level.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.0467	0.1972	0.0560	0.8129
kw_avg_avg		1	0.000733	0.000029	639.6898	<.0001
is_weekend	**0**	1	-0.9196	0.0402	524.0900	<.0001
data_channel_is_tech	**0**	1	-0.5929	0.0507	136.8436	<.0001
data_channel_is_socm	**0**	1	-0.8677	0.0675	165.2038	<.0001
LDA_00		1	1.0871	0.0898	146.6731	<.0001
kw_min_min		1	0.00144	0.000357	16.2042	<.0001
kw_max_avg		1	-0.00009	5.153E-6	319.3341	<.0001
num_keywords		1	0.0412	0.00797	26.6727	<.0001
num_hrefs		1	0.00939	0.00150	39.2661	<.0001
LDA_04		1	0.4993	0.0703	50.3786	<.0001
kw_min_avg		1	-0.00010	0.000016	37.9088	<.0001
self_reference_avg_s		1	4.105E-6	1.134E-6	13.1078	0.0003
n_tokens_content		1	0.000167	0.000034	23.9242	<.0001
data_channel_is_ente	**0**	1	0.2459	0.0371	43.8763	<.0001
kw_avg_max		1	-5.34E-7	1.676E-7	10.1358	0.0015
average_token_length		1	-0.1274	0.0203	39.5415	<.0001
global_subjectivity		1	0.8239	0.1442	32.6486	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
num_self_hrefs		1	-0.0150	0.00393	14.4690	0.0001
min_positive_polarit		1	-0.7402	0.1932	14.6875	0.0001
kw_max_max		1	-4.63E-7	1.237E-7	13.9757	0.0002
num_imgs		1	0.00432	0.00172	6.2770	0.0122
title_sentiment_pola		1	0.1493	0.0498	8.9754	0.0027
abs_title_subjectivi		1	0.2746	0.0774	12.5920	0.0004
data_channel_is_bus	0	1	0.1394	0.0638	4.7713	0.0289
kw_min_max		1	-5.4E-7	2.483E-7	4.7243	0.0297
self_reference_min_s		1	0.000013	2.065E-6	38.3144	<.0001
title_subjectivity		1	0.0910	0.0461	3.8981	0.0483

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
kw_avg_avg	1.001	1.001	1.001
is_weekend 0 vs 1	0.399	0.368	0.431
data_channel_is_tech 0 vs 1	0.553	0.500	0.610
data_channel_is_socm 0 vs 1	0.420	0.368	0.479
LDA_00	2.966	2.487	3.536
kw_min_min	1.001	1.001	1.002
kw_max_avg	1.000	1.000	1.000
num_keywords	1.042	1.026	1.058
num_hrefs	1.009	1.006	1.012
LDA_04	1.648	1.435	1.891
kw_min_avg	1.000	1.000	1.000
self_reference_avg_s	1.000	1.000	1.000
n_tokens_content	1.000	1.000	1.000
data_channel_is_ente 0 vs 1	1.279	1.189	1.375

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
kw_avg_max	1.000	1.000	1.000
average_token_length	0.880	0.846	0.916
global_subjectivity	2.279	1.718	3.024
num_self_hrefs	0.985	0.978	0.993
min_positive_polarit	0.477	0.327	0.697
kw_max_max	1.000	1.000	1.000
num_imgs	1.004	1.001	1.008
title_sentiment_pola	1.161	1.053	1.280
abs_title_subjectivi	1.316	1.131	1.532
data_channel_is_bus 0 vs 1	1.150	1.014	1.303
kw_min_max	1.000	1.000	1.000
self_reference_min_s	1.000	1.000	1.000
title_subjectivity	1.095	1.001	1.199

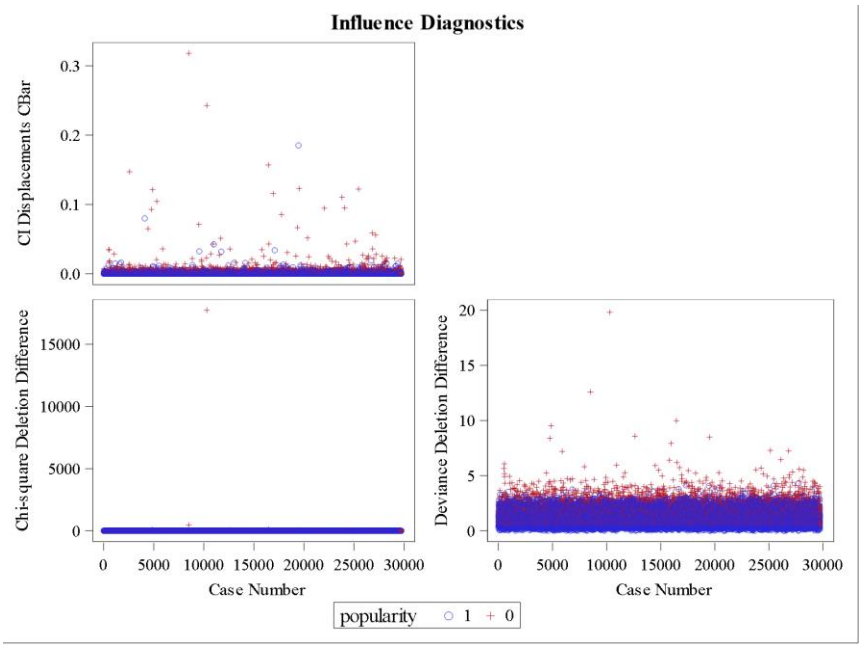
(Table 5)

Based on the result of Odds Ratio Estimate Table, we can see 13 variables have odds ratio greater than 1, indicating that the effect of these 13 variables on the odds ratio is positive and significant. This also implies that the odds of the article being popular are multiplied by more than 1 for every unit increase in these variables. In these variables, **global_subjectivity** has largest positive odds ratio. When **global_subjectivity** increases one unit, the odds of the article being popular are multiplied by 2.279. And 6 variables have odds ratio less than 1, indicating that the effect of these 6 variables on the odds ratio is negative and significant. And the variable **is_weekend** has the largest negative influence on the odds ratio, which means the article is published not on weekend, but on weekdays, the odds of the article being popular are multiplied by 0.399. So the article will be less popular.

However, in Table 5, there are also 8 variables which have odds ratio equal to 1, which means the change of these variables do not have influence on the odds of the article being popular. So I plan to scale there variables, because according to the maximize likelihood estimate table, we can see that all these variables are significant, we should not remove them.

And we also find that all these variables are related to the number of words, so I decided to scale these variables by divided by 1000.

After scaling,



Based on the diagnostics plot, we can see there is no outliers.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
kwavg_avg	2.082	1.967	2.204
is_weekend 0 vs 1	0.399	0.368	0.431
data_channel_is_tech 0 vs 1	0.553	0.500	0.610
data_channel_is_socm 0 vs 1	0.420	0.368	0.479
LDA_00	2.966	2.487	3.536
kwworst_min	4.203	2.089	8.455
kwavg_max	0.912	0.903	0.921
num_keywords	1.042	1.026	1.058
num_hrefs	1.009	1.006	1.012
LDA_04	1.648	1.435	1.891
kwavg_min	0.905	0.877	0.934
self_avg	1.004	1.002	1.006
n_content	1.182	1.105	1.264
data_channel_is_ente 0 vs 1	1.279	1.189	1.375

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
kwbest_avg	0.999	0.999	1.000
average_token_length	0.880	0.846	0.916
global_subjectivity	2.279	1.718	3.024
num_self_hrefs	0.985	0.978	0.993
min_positive_polarit	0.477	0.327	0.697
kwbest_max	1.000	0.999	1.000
num_imgs	1.004	1.001	1.008
title_sentiment_pola	1.161	1.053	1.280
abs_title_subjectivi	1.316	1.131	1.532
data_channel_is_bus 0 vs 1	1.150	1.014	1.303
kwbest_min	0.999	0.999	1.000
self_min	1.013	1.009	1.017
title_subjectivity	1.095	1.001	1.199

(Table 6)

According to the Table 6 result, we can see all the variables have significant influence on the popularity of the article.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
12.2663	8	0.1397

(Table 7)

The P-value of the Goodness-of-Fit test is 0.1397, which is larger than 0.05, so it means that there is no lack of fit problem.

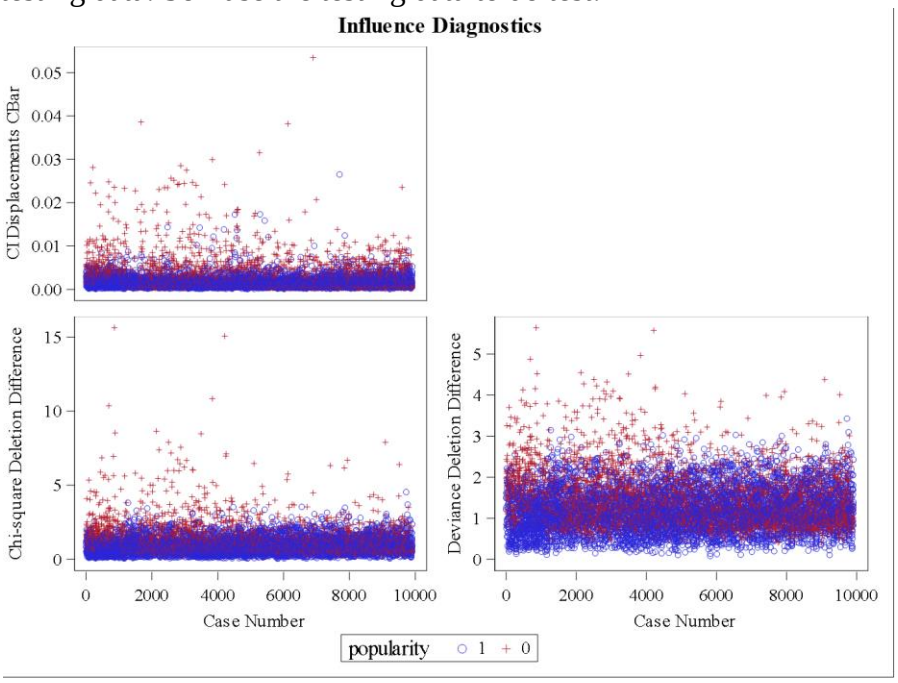
Table of popularity by _INTO_			
popularity	_INTO_(Formatted Value of the Predicted Response)		
Frequency Row Pct	0	1	Total
0	8308 60.05	5528 39.95	13836
1	4701 29.58	11191 70.42	15892
Total	13009	16719	29728

(Table 8)

Table 8 gives us the result of comparison between actual values and predicted values. We can see the result is good in the training data. The misclassification error is 0.34408638. It is an acceptable error.

Testing model

Although the model is fitted well on training data, what we really care is the performance of the model on the testing data. So I use the testing data to do test.



Based on result of diagnostics plot, there is no outlier problem.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	13703.625	12857.912
SC	13710.826	12994.738
-2 Log L	13701.625	12819.912

(Table 9)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	881.7131	18	<.0001
Score	834.9372	18	<.0001
Wald	764.5636	18	<.0001

(Table 10)

The AIC is noticeably lower for the Logistic model (12857.912) than for the intercept only model (13703.625), and the global tests all agree that the parameter is different from 0 at 5% statistical significance level.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.6154	0.3439	3.2016	0.0736
kw_avg_avg		1	0.000630	0.000050	156.1608	<.0001
is_weekend	0	1	-0.7944	0.0684	135.0381	<.0001
data_channel_is_tech	0	1	-0.5729	0.0877	42.6770	<.0001
data_channel_is_socm	0	1	-0.9097	0.1169	60.5469	<.0001
LDA_00		1	0.9438	0.1537	37.7146	<.0001
kw_min_min		1	0.00320	0.000612	27.4367	<.0001
kw_max_avg		1	-0.00008	9.409E-6	79.9489	<.0001
num_keywords		1	0.0584	0.0137	18.1224	<.0001
num_hrefs		1	0.0101	0.00247	16.7250	<.0001
LDA_04		1	0.4146	0.1205	11.8335	0.0006
kw_min_avg		1	-0.00006	0.000028	4.1448	0.0418
self_reference_avg_s		1	3.228E-6	1.96E-6	2.7120	0.0996

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
n_tokens_content		1	0.000185	0.000059	9.9524	0.0016
data_channel_is_ente	0	1	0.2638	0.0639	17.0356	<.0001
kw_avg_max		1	-4.52E-9	2.941E-7	0.0002	0.9877
average_token_length		1	-0.1158	0.0350	10.9747	0.0009
global_subjectivity		1	0.9237	0.2469	13.9920	0.0002
num_self_hrefs		1	-0.0232	0.00679	11.6510	0.0006
min_positive_polarit		1	-0.7532	0.3375	4.9800	0.0256
kw_max_max		1	7.467E-8	2.128E-7	0.1231	0.7257
num_imgs		1	0.00530	0.00314	2.8592	0.0909
title_sentiment_pola		1	0.3631	0.0848	18.3170	<.0001
abs_title_subjectivi		1	0.2755	0.1334	4.2612	0.0390
data_channel_is_bus	0	1	0.1474	0.1099	1.7981	0.1799
kw_min_max		1	-9.28E-7	4.12E-7	5.0767	0.0242
self_reference_min_s		1	0.000013	3.537E-6	14.5072	0.0001
title_subjectivity		1	0.1528	0.0786	3.7821	0.0518

(Table 11)

According to the Table 11, the p-value of 21 variables are less than 0.05, they are significant. And 6 variables are not significant , which are significant in train data.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
is_weekend 0 vs 1	0.440	0.386	0.502
data_channel_is_tech 0 vs 1	0.621	0.525	0.736
data_channel_is_socm 0 vs 1	0.409	0.327	0.512
LDA_00	2.642	1.967	3.549
kw_min_min	1.001	1.001	1.002
num_keywords	1.049	1.025	1.073

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
num_hrefs	1.017	1.012	1.022
LDA_04	1.307	1.041	1.640
data_channel_is_ente 0 vs 1	1.444	1.280	1.630
average_token_length	0.751	0.704	0.800
global_subjectivity	7.145	4.504	11.334
num_self_hrefs	0.974	0.961	0.986
min_positive_polarit	0.667	0.359	1.240
num_imgs	1.016	1.010	1.022
title_sentiment_pola	1.455	1.237	1.711
abs_title_subjectivi	1.333	1.032	1.722
data_channel_is_bus 0 vs 1	1.305	1.060	1.607
title_subjectivity	1.234	1.062	1.435

(Table 12)

Table 12 tells us that 12 variables have positive influence on odds ratio. And **global_subjectivity** has largest influence which is the same as the result from training data. And 7 variables have negative influence on odds ratio. **Data_channel_is_socm** has the largest negative influence, which is different from the variable (**is_weekend**) in training data. This means if an article's channel is not socmed, the odds of this article to be popular will decrease.

Table of popularity by _INTO_			
popularity	_INTO_(Formatted Value of the Predicted Response)		
Frequency Row Pct	0	1	Total
0	2827 60.83	1820 39.17	4647
1	1631 31.00	3631 69.00	5262
Total	4458	5451	9909

(Table 13)

Table 13 gives us the result of comparison between actual values and predicted values on testing data. The misclassification error is 0.37503783. It is just 0.03 higher than the error on training data, so it means the performance of this model is really good.

Support Vector Machine

Support vector machine is a method for classification. The support vector machine is a generalization of a simple and intuitive classifier called the maximal margin classifier which requires that the classes be separable by a linear boundary, so this is a limitation. And support vector classifier, an extension of the maximal margin classifier that can be applied in a broader range of cases. support vector machine, which is a further extension of the support vector classifier in order to accommodate non-linear class boundaries.

```
## Linear SVM
library(ElemStatLearn)
library('e1071')
library(nnet)
### perform cross-validation to find the best cost
svmfit.linear <-svm(y~.,data=train_data,kernel='linear',cost=1,gamma=0.1,decision.values=T)
ypred_linear<- predict(svmfit.linear,test_data)
table(predict=ypred_linear,truth=test_data$y)
### the misclassification error
error<-function(predict,truth){
  sum(predict != truth)/length(truth)
}
error(ypred_linear,test_data$y)

summary(Onlinenews$shares)
```

I choose linear SVM with cost=1 and gamma=0.1 to do analyze.

```
        truth
predict    0    1
      0 2833 1637
      1 1810 3631
```

```
> error(ypred_linear,test_data$y)
[1] 0.3477954
```

(Table 14)

Table 14 is the result of comparison between actual values and predicted values on testing data. The misclassification error is 0.3477954 which is a little better than the error from Logistic regression.

Conclusion

In this project, I plan to find what kind of predictors have significant influence on the popularity of Online News and want to know the accuracy of predictions by these variables. I choose logistic regression and SVM to do analyze. After variable selection and remove the variables do not have influence on the popularity. There are 19 variables in the final model of logistic regression. And the misclassification error on testing data is around 0.37, it is not very good, but acceptable. And the error from SVM is 0.3477954, it is a little better than Logistic regression, but it cost almost half hour to perform SVM.

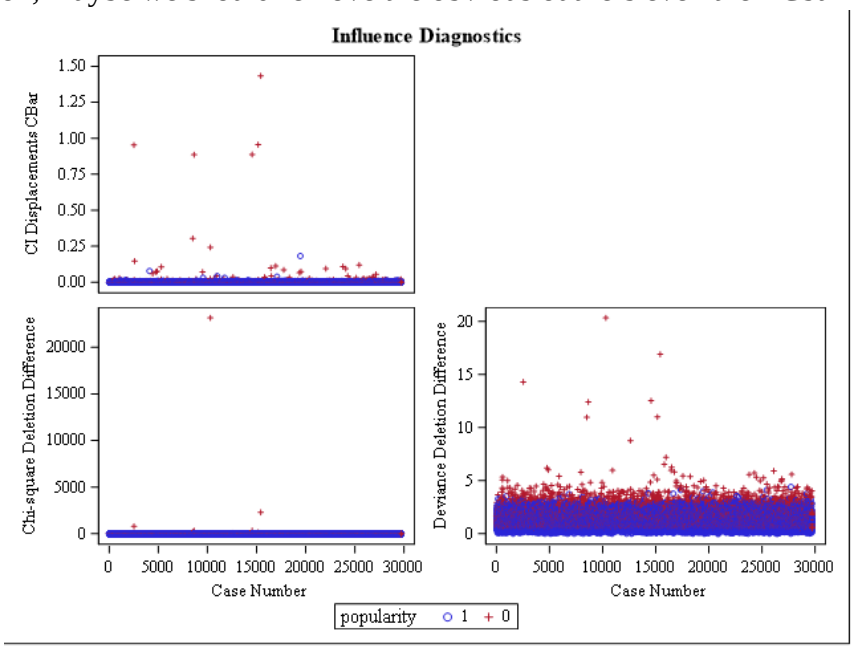
Based on the Logistic Regression result, we can see the articles published on weekend will have more chance to be popular. I think it is maybe people get rest on weekend and have more time to read an article. And also, there are 4 variables that related to channel, we can see that articles' data channel is "tech", "social media", "business" and "entertainment" will make the articles to be popular. I think the reason for this is that all these four channel types are really related to people's daily life, people want to know more information for their lives by reading these articles. And on the other hand, an article has more image will become more popular, because I think images can make illustration more clear.

As to the future work, I think we should do feature engineering on the variables. Polynomial or quadratic terms might increase prediction power if they are construct in the right way. Beyond this, more advanced model like neural network could get even better result than Logistic Regression in a big dataset like this.

Appendices

1. During this project, I used R to split data into train an test randomly. And I also did Support Vector Machine by R. I think R and SAS are both good tools for analysis. R is more flexible and read data easily. And SAS can handle large dataset, such as the dataset in this project. So making a combination of them can make the analysis effective.
2. Odds Ratio table gives us the illustration for the influence of different variables, and when the 95% Wald Confidence Limits does not include 1, means that the variable has significant influence on response. The point estimate > 1, means the influence is positive; point < 1 means the influence is negative. The value in the point estimate means increase 1 unit of that variable will cause the odds ratio of event happen multiple that value.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
is_weekend 0 vs 1	0.440	0.386	0.502
data_channel_is_tech 0 vs 1	0.621	0.525	0.736
data_channel_is_socm 0 vs 1	0.409	0.327	0.512
LDA_00	2.642	1.967	3.549
kw_min_min	1.001	1.001	1.002
num_keywords	1.049	1.025	1.073
num_hrefs	1.017	1.012	1.022
LDA_04	1.307	1.041	1.640
data_channel_is_ente 0 vs 1	1.444	1.280	1.630
average_token_length	0.751	0.704	0.800

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
global_subjectivity	7.145	4.504	11.334
num_self_hrefs	0.974	0.961	0.986
min_positive_polarit	0.667	0.359	1.240
num_imgs	1.016	1.010	1.022
title_sentiment_pola	1.455	1.237	1.711
abs_title_subjectivi	1.333	1.032	1.722
data_channel_is_bus 0 vs 1	1.305	1.060	1.607
title_subjectivity	1.234	1.062	1.435

3. In diagnostic plots, for logistic regression, we just need to see CI Displacements Cbar plots. Generally, we consider the observation with Cbar > 1 as highly influential and we should remove it. However, sometimes whether we should remove some outliers depend on the real situation, maybe we should remove the obvious outliers even their Cbar < 1.