Name: Saunak Chandra (saunakc@uw.edu)

Class: DATASCI 350 Methods for Data Analysis

File: Project submission Online News Popularity.docx

Purpose: Produce logistic model for predicting online news popularity.

Data Source: <http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>

Created on: 2015-08-22

**Purpose:**

Run statistical linear regression on a dataset of online news popularity to produce a predictive model. The dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to identify features with more influence in predicting the popularity.

**Data Set Information:**

* The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided urls.
* Acquisition date: January 8, 2015

**Attribute Information:**

Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)

Attribute Information:

1. url: URL of the article (non-predictive)
2. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
3. n\_tokens\_title: Number of words in the title
4. n\_tokens\_content: Number of words in the content
5. n\_unique\_tokens: Rate of unique words in the content
6. n\_non\_stop\_words: Rate of non-stop words in the content
7. n\_non\_stop\_unique\_tokens: Rate of unique non-stop words in the content
8. num\_hrefs: Number of links
9. num\_self\_hrefs: Number of links to other articles published by Mashable
10. num\_imgs: Number of images
11. num\_videos: Number of videos
12. average\_token\_length: Average length of the words in the content
13. num\_keywords: Number of keywords in the metadata
14. data\_channel\_is\_lifestyle: Is data channel 'Lifestyle'?
15. data\_channel\_is\_entertainment: Is data channel 'Entertainment'?
16. data\_channel\_is\_bus: Is data channel 'Business'?
17. data\_channel\_is\_socmed: Is data channel 'Social Media'?
18. data\_channel\_is\_tech: Is data channel 'Tech'?
19. data\_channel\_is\_world: Is data channel 'World'?
20. kw\_min\_min: Worst keyword (min. shares)
21. kw\_max\_min: Worst keyword (max. shares)
22. kw\_avg\_min: Worst keyword (avg. shares)
23. kw\_min\_max: Best keyword (min. shares)
24. kw\_max\_max: Best keyword (max. shares)
25. kw\_avg\_max: Best keyword (avg. shares)
26. kw\_min\_avg: Avg. keyword (min. shares)
27. kw\_max\_avg: Avg. keyword (max. shares)
28. kw\_avg\_avg: Avg. keyword (avg. shares)
29. self\_reference\_min\_shares: Min. shares of referenced articles in Mashable
30. self\_reference\_max\_shares: Max. shares of referenced articles in Mashable
31. self\_reference\_avg\_sharess: Avg. shares of referenced articles in Mashable
32. weekday\_is\_monday: Was the article published on a Monday?
33. weekday\_is\_tuesday: Was the article published on a Tuesday?
34. weekday\_is\_wednesday: Was the article published on a Wednesday?
35. weekday\_is\_thursday: Was the article published on a Thursday?
36. weekday\_is\_friday: Was the article published on a Friday?
37. weekday\_is\_saturday: Was the article published on a Saturday?
38. weekday\_is\_sunday: Was the article published on a Sunday?
39. is\_weekend: Was the article published on the weekend?
40. LDA\_00: Closeness to LDA topic 0
41. LDA\_01: Closeness to LDA topic 1
42. LDA\_02: Closeness to LDA topic 2
43. LDA\_03: Closeness to LDA topic 3
44. LDA\_04: Closeness to LDA topic 4
45. global\_subjectivity: Text subjectivity
46. global\_sentiment\_polarity: Text sentiment polarity
47. global\_rate\_positive\_words: Rate of positive words in the content
48. global\_rate\_negative\_words: Rate of negative words in the content
49. rate\_positive\_words: Rate of positive words among non-neutral tokens
50. rate\_negative\_words: Rate of negative words among non-neutral tokens
51. avg\_positive\_polarity: Avg. polarity of positive words
52. min\_positive\_polarity: Min. polarity of positive words
53. max\_positive\_polarity: Max. polarity of positive words
54. avg\_negative\_polarity: Avg. polarity of negative words
55. min\_negative\_polarity: Min. polarity of negative words
56. max\_negative\_polarity: Max. polarity of negative words
57. title\_subjectivity: Title subjectivity
58. title\_sentiment\_polarity: Title polarity
59. abs\_title\_subjectivity: Absolute subjectivity level
60. abs\_title\_sentiment\_polarity: Absolute polarity level
61. shares: Number of shares (target)

**Summary**

I am presenting here a model to predict whether an online news article published in Mashable is going to be popular or not. Out of 61 variables captured in the data set few are more important than others- a list of such variables are a good target for advertising agencies. We will be using Lasso Cross Validation technique in R to narrow down the factors significantly important in predicting the popularity.

I have used Logistics Linear Regression model for my prediction with an Accuracy of **65%** and AUC of **0.70**. The model takes about less about a minute to run in a 2.7G Hz Intel 8 GB RAM Windows environment.

**Data Exploration:**

[1] "Number of cases: 39644"

[1] "Number of attributes: 61"

[1] "Number of attributes with character col: 1"

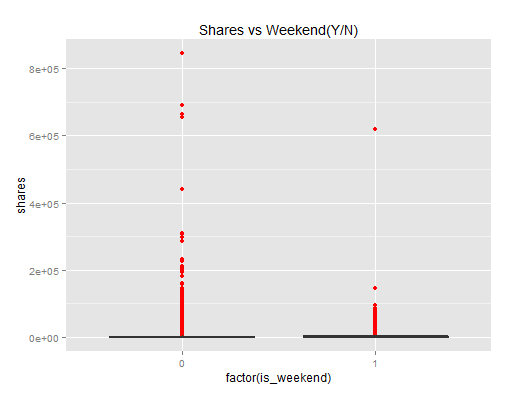
[1] "Number of attributes with numeric col: 59"

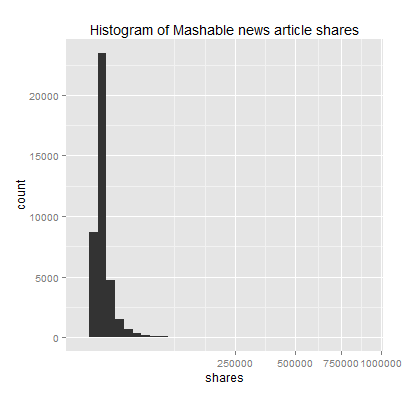
[1] "Number of attributes with integer col: 1"

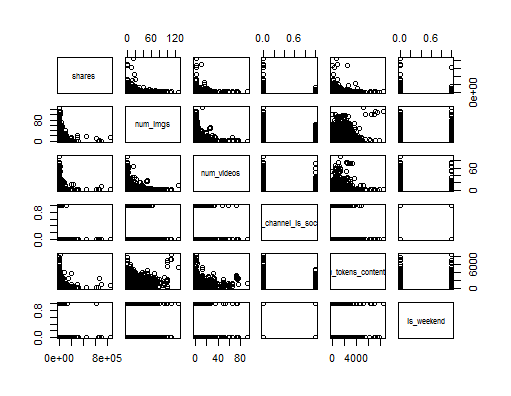
[1] "------Output variable \"shares\" distribution------"

[1] "Mean Median IQR Range"

[1] "3395 1400 1854 1 - 843300"







**Prediction:**

The linear regression model is applied to 25 factors instead of the originally perceived 61 factors. Lasso Regression is applied to the dataset to find out most effective co-efficients. The factors are-

n\_tokens\_title

n\_tokens\_content

n\_non\_stop\_words

num\_hrefs

num\_imgs

num\_videos

num\_keywords

data\_channel\_is\_socmed

data\_channel\_is\_tech

kw\_min\_min

kw\_avg\_avg

self\_reference\_min\_shares

self\_reference\_max\_shares

self\_reference\_avg\_sharess

weekday\_is\_monday

weekday\_is\_friday

weekday\_is\_saturday

is\_weekend

LDA\_00

LDA\_04

global\_subjectivity

max\_positive\_polarity

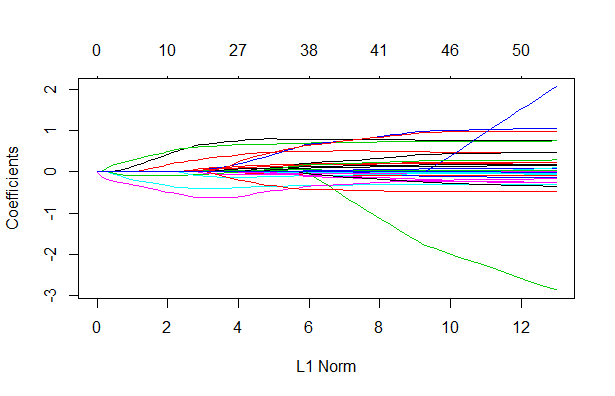
title\_subjectivity

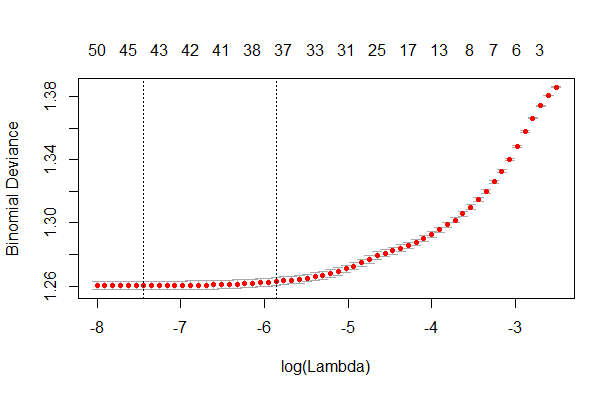
title\_sentiment\_polarity

abs\_title\_subjectivity

Precisely an article’s token title and content, references, number of videos and images in the article, keywords, self references of the article, channel related social and technology, day of week, and subjectivity of the content are significantly important in measurement of popularity. Weekends especially Friday and Saturday articles are more active. Article title subject and sentiment polarity has some influence in the popularity.

**Metrics for Lasso Regression:**





**Performance:**

The model predicts with an accuracy of 64% and the area under curve is 0.7.

[1] "-----Prediction results---------"

[1] "Accuracy = 0.64"

[1] "Sensitivity = 0.68"

[1] "Specificity = 0.6"

[1] "Area under curve (AUC) 0.7"

[1] "--------------------------------"

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 1377 773

1 638 1176

Accuracy : 0.644

95% CI : (0.6289, 0.659)

No Information Rate : 0.5083

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.2871

Mcnemar's Test P-Value : 0.0003607

Sensitivity : 0.6834

Specificity : 0.6034

Pos Pred Value : 0.6405

Neg Pred Value : 0.6483

Prevalence : 0.5083

Detection Rate : 0.3474

Detection Prevalence : 0.5424

Balanced Accuracy : 0.6434

'Positive' Class : 0

Time to run the model:

user system elapsed

42.64 0.25 43.28