|  |  |
| --- | --- |
| EMB00001764328f | **Hanyang University**  **Industrial Engineering Department**  **Graduate School** |

**Regression Analysis**

Prof. Seok Joo Bae

PREDICTING AND EVALUATING THE POPULARITY OF ONLINE NEWS

Evans Sowah **OKPOTI**

Nazli **REIS**

Young Eun **KOO**

Contents

[1.0 Introduction 2](#_Toc454406416)

[2.0 Data Information 3](#_Toc454406417)

[3.0 Theoretical Background 4](#_Toc454406418)

[3.1 Logistic Regression 4](#_Toc454406419)

[3.2 The Logit Transformation 4](#_Toc454406420)

[3.3 The Logistic Regression Model 5](#_Toc454406421)

[3.4 Estimation and Hypothesis Testing 5](#_Toc454406422)

[3.5 ROC (Receiver Operating Characteristic) 6](#_Toc454406423)

[3.6 Area Under the Curve 6](#_Toc454406424)

[3.7 Confusion Matrix 6](#_Toc454406425)

[4.0 Analysis and Results 8](#_Toc454406426)

[4.1 Filter Methods and Model Selection 8](#_Toc454406427)

[4.2 Goodness of Fit, Likelihood Ratio and Odds Ratio 10](#_Toc454406428)

[5.0 Conclusion 12](#_Toc454406429)

[References 12](#_Toc454406430)

[Appendix 13](#_Toc454406431)

[Plot of Odd ratio and 95% C.I. 13](#_Toc454406432)

[Codes for the Fisher criterion feature selection model 14](#_Toc454406433)

[Codes for the MIA feature selection model 16](#_Toc454406434)

# 1.0 Introduction

In this information era, reading and sharing news have become the center of people’s entertainment lives. Therefore, it would be greatly helpful if we could accurately predict the popularity of news prior to its publication, for social media workers (authors, advertisers, etc) [1]. Predicting online news is gaining popularity especially in this social media era [1-4]. The measurement of popularity is based on a number of Web and social network interactions. Tatar et al. [5] mentions two approaches to popularity prediction of web content namely: those that utilize features after publication and those that do not use such features. The first approach is quite easy and is associated with higher prediction performance [4, 6]. The second approach on the other hand is scarce with low prediction performance, however, it offers useful insights about the article for possible content improvement before publication.

For the purpose of this project, we intend to make use of a largely and recently collected dataset with over 39000 articles from Mashable website (figure 1), to first select informative features and then analyze and compare the performance two feature selection approaches. Note that we use the work of Ren and Yang [7] as a baseline. In this project, we intend to find the best model and set of feature to predict the popularity of online news, using machine learning techniques.

|  |  |
| --- | --- |
| C:\Users\hy\Desktop\mashable.png | C:\Users\hy\Desktop\mashable2.png |
| Figure 1. A screenshot of the Mashable website | |

# 2.0 Data Information

We retrieved the data from the UCI Machine Learning respository [3]. The data was collected during a two year period, from January 7 2013 to January 7 2015. A small portion of special occasion articles that did not follow the general HTML structure very recent articles (less than 3 weeks) were discarded. After such preprocessing, we ended with a total of 39,644 articles as shown in table 1. The data contains a selected large list of characteristics that describe different aspects of the article that were considered possibly relevant to influence the number of shares. Some of the features are dependent of particularities of the Mashable service: articles have meta-data, such as keywords, data channel type and total number of shares.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1. Statistical measures of the dataset | | | | | | | |
|  |  |  | | **Articles per day** | | | |
| **Number of article** | **Total days** | **Shares value range** | | **Average** | **Standard deviation** | **Min** | **Max** |
| <1400 | >=1400 |
| 39644 | 709 | 18490 | 21154 | 55.00 | 22.65 | 12 | 105 |

The data contains a total of 61 attributes (features) with 58 predictive, 2 non-predictive and 1 goal field. The attributes of the dataset have been categorized and summarized in table 2 below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 2. List of attributes by category | | | | |
| **Feature** | **Type (#)** |  | **Feature** | **Type (#)** |
| **Words** | |  | **Keywords** | |
| Number of words in the title | number (1) |  | Number of keywords | number (1) |
| Number of words in the article | number (1) |  | Worst keyword (min./avg./max. shares) | number (3) |
| Average word length | number (1) |  | Average keyword (min./avg./max. shares) | number (3) |
| Rate of non-stop words | ratio (1) |  | Best keyword (min./avg./max. shares) | number (3) |
| Rate of unique words | ratio (1) |  | Article category (Mashable data channel) | nominal (7) |
| Rate of unique non-stop words | ratio (1) |  | **Natural Language Processing** | |
| **Links** | |  | Closeness to top 5 LDA topics | ratio (5) |
| Number of links | number (1) |  | Title subjectivity | ratio (1) |
| Number of Mashable article links | number (1) |  | Title sentiment polarity | ratio (1) |
| Minimum, average and maximum number of shares of Mashable links | number (3) |  | Article text subjectivity score and its absolute difference to 0.5 | ratio (2) |
| **Links** | |  | Rate of positive and negative words | ratio (2) |
| Number of images | number (1) |  | Pos. words rate among non-neutral words | ratio (1) |
| Number of videos | number (1) |  | Neg. words rate among non-neutral words | ratio (1) |
| **Digital Media** | |  | Polarity of positive words (min./avg./max.) | ratio (3) |
| Day of the week | nominal (5) |  | Polarity of negative words (min./avg./max.) | ratio (3) |
| Published on a weekend? | bool (1) |  | Article text polarity score and its absolute difference to 0.5 | ratio (2) |
|  |  |  |  |  |
|  |  |  | **Target** | **Type (#)** |
|  |  |  | Number of article shares | number (1) |

# 3.0 Theoretical Background

## 3.1 Logistic Regression

We consider the case where the response is binary, assuming only two values that for convenience we code as one or zero. In our data, we define;

We view as a realization of a random variable that can take the values one and zero with probabilities and , respectively. The distribution of is called a Bernoulli distribution with parameter , and can be written in a compact form as:

Note that if we obtain , and if we obtain . It is fairly easy to verify by direct calculation that the expected value and variance of are:

The mean and variance depend on the underlying probability . Any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggest that a linear model that allows the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data. Let,. We view as a realization of a random variable that takes the values . If the observations in each group are independent, and they all have the same probability πi of having the attribute of interest, then the distribution of Yi is binomial with parameters πi and ni , which we write

The probability distribution function of is given by:

for = .

Here is the probability of obtaining yi successes and failures in some specific order, and the combinatorial coefficient is the number of ways of obtaining yi successes in ni trials. The mean and variance of Yi can be shown to be yes

## 3.2 The Logit Transformation

The next step in defining a model for our data concerns the systematic structure. We would like to have the probabilities πi depend on a vector of observed covariates xi . The simplest idea would be to let πi be a linear function of the covariates, say , where is a vector of regression coefficients.

First, we move from the probability πi to the odds:

defined as the ratio of the probability to its complement, or the ratio of favorable to unfavorable cases. If the probability of an event is a half, the odds are one-to-one or even. If the probability is , the odds are one-to-two. If the probability is very small, the odds are said to be long. In some contexts the language of odds is more natural than the language of probabilities.

Second, we take logarithms, calculating the logit or log-odds

which has the effect of removing the floor restriction. To see this point note that as the probability goes down to zero the odds approach zero and the logit approaches .

## 3.3 The Logistic Regression Model

Suppose that we have independent observations , and that the observation can be treated as a realization of a random variable . We assume that has a binomial distribution

with binomial denominator and probability . With individual data for all . This defines the stochastic structure of the model. Suppose further that the logit of the underlying probability πi is a linear function of the predictors

where is a vector of covariates and is a vector of regression coefficients. Exponentiating we find that the odds for the unit are given by:

This expression defines a multiplicative model for the odds. Solving for the probability in the logit model in , gives the more complicated model

While the left-hand-side is in the familiar probability scale, the right-handside is a non-linear function of the predictors, and there is no simple way to express the effect on the probability of increasing a predictor by one unit while holding the other variables constant. We can obtain an approximate answer by taking derivatives with respect to , which of course makes sense only for continuous predictors.

## 3.4 Estimation and Hypothesis Testing

1. *Maximum Likelihood Estimation*

The likelihood function for n independent binomial observations is a product of densities given by

Taking logs we find that, except for a constant involving the combinatorial terms, the log-likelihood function is:

,

Alternatives to maximum likelihood estimation include weighted least squares, which can be used with grouped data, and a method that minimizes Pearson’s chi-squared statistic, which can be used with both grouped and individual data.

1. *Goodness of Fit Statistics*

Suppose we have just fitted a model and want to assess how well it fits the data.

A measure of discrepancy between observed and fitted values is the deviance statistic, which is given by:

where yi is the observed and is the fitted value for theobservation.

An alternative measure of goodness of fit is Pearson’s chi-squared statistic, which for binomial data can be written as

With grouped data Pearson’s statistic has approximately in large samples a chi-squared distribution with degrees of freedom, and is asymptotically equivalent to the deviance or likelihood-ratio chi-squared statistic. The statistic can not be used as a goodness of fit test with individual data, but provides a basis for calculating residuals.

1. *Tests of Hypothesis*

We can test the hypothesis

concerning the significance of a single coefficient by calculating the ratio of the estimate to its standard error

This statistic has approximately a standard normal distribution in large samples. For more general problems we consider the likelihood ratio test. The deviance plays a role similar to the residual sum of squares.

## 3.5 ROC (Receiver Operating Characteristic)

ROC Curve (figure 2) is graphical plot that illustrates the performance of binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at vaious threshold settings. The true rate is also known as sensitivity or recall in machine learning.

## 3.6 Area Under the Curve

When using normalized units the area under the curve is equalto the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

## 3.7 Confusion Matrix

In thte field of machine abd specially the problem of statistical classification a confusion matrix also known as an error matrix is a specific table layout that allows visualizaion of the performance of an algorithm typically a supervised learning one. Each column of the matrix (figure 3) represents the instances in a predicted class while each row represents the instance in an actual class.[[1]](#footnote-1)

|  |  |
| --- | --- |
|  |  |
| Figure 2. ROC Curve | Figure 3. Confusion Matrix |

# 4.0 Analysis and Results

As previously mentioned, in this project we use the work of Ren and Yang [7] as a baseline of comparison. In their work, they assert the fact that Principal Component Analysis (PCA) usually to reduce dimensionality made the models perform worse and for that matter used filter methods to select features. These methods inlcude Mutual Information (MI) and Fisher Criterion. For the most part of this section, results for each analysis will be reported for both the feature selection method implemented as well as that proposed by [7]. Also we use a 70:30 ratio for the training and testing datasets.

## 4.1 Filter Methods and model selection

1. *Mutual Information*: The MI between features and class labels to be score to rank features is expressed as Kullback-Leibler(KL) divergence as follows:
2. *Fisher Criterion*: Fisher criterion is another effective way in feature ranking. The numerator indicates the discrimination between popular and unpopular news, and the denominator indicates the scatter within the class. The larger the F-score, the more likely this feature is more discsriminative.

Where

1. *Model Independent Approach (MIA)*: Here we adopted a non-parametric method where for each feature, a loca regression (LOESS) smoother is fit between the predictor and the response. Next the statistic is calculated for this model against the intercept only model.

We report the results of the filter methods for the top-20 features in table 3 below. That for Fisher criterion is reported as given in [7] however, that for MI was not given and is therefore not reported.

|  |  |
| --- | --- |
| Table 3. Top-20 features selected using MIA (left) and Fisher criterion (right) | |
|  |  |
| Model Independent Approach (MIA) | Fisher Criterion |

Both methods ranked the “Average keyword (avg.shares)” and “Closeness to LDA topic 1” features as the first and 20th respectively. That is, our full model at start of analysis contains 20 predictor variables.

Next we fit the models for MIA and Fisher criterion using all 20 features and the plots are shown in table 4. For each model, the following plots are given **[1]** a plot of residuals against fitted values, **[2]** a Scale-Location plot of sqrt(|residuals|) against fitted values, **[3]** a Normal Q-Q plot, **[4]** a plot of residuals against leverages. The plots show that the logistic regression model is appropriate for the dataset. The associated summary for each model is also shown in table 5. In the MIA model, the “LDA\_00” feature had null values for the all measured values and for Fisher criterion, the “is\_weekend” feature also had null values.

|  |  |
| --- | --- |
| Table 4. Model plots MIA (left) and Fisher criterion (right) | |
|  |  |
| Model Independent Approach (MIA) | Fisher Criterion |

|  |  |
| --- | --- |
| Table 5. Model summary MIA (left) and Fisher criterion (right) | |
|  |  |
| Model Independent Approach (MIA) | Fisher Criterion |

We further undertake a stepwise logistic regression to reduce the predictor variables. This is done based on the AIC (Akaike Information Criterion) value for each model. The smaller AIC value indicates the best model for our performance tasks. The stages of the stepwise regression and the associated AIC values are shown in table 6. The MIA model was reduced to 16 predictor variables with an AIC value of 36239.39 which is lower than the full model which was 36244.37. Likewise, the Fisher criterion model was also reduced to 15 predictor variables with an AIC value of 36027.17 which is lesser than the prior full model with an AIC value of 36029.2.

|  |  |
| --- | --- |
| Table 5. Stepwise logistic regression with AIC values MIA (left) and Fisher criterion (right) | |
|  |  |
| Model Independent Approach (MIA) | Fisher Criterion |

## 4.2 Goodness of Fit, Likelihood Ratio and Odds Ratio

Suppose two alternative models are under consideration, one model is simpler or more parsimonious than the other more often than not, one of the models is the saturated model. Another common situation is to consider ‘nested’ models, where one model is obtained from the other one by putting some of the parameters to be zero. In this project we seek to test the following hypothesis:

The likelihood-ratio statistic is

The result of this is same as the deviance value obtained from the summary of the model. Next the deviance value is compared to a chi-square value at an significance level with degrees of freedom.

For the MIA model, the deviance value was which less than and so we fail to reject the null hypothesis. That is the 4 predictor variables that were eliminated from the full model have regression parameter . Similary the test for the Fisher criterion model had a deviance value of which was also less than the . We further conduct the odds-ratio test. In this test, we ascertain the odds of “news being popular” (versus “unpopular”), the amount of increase associated with a unit increase in the value of the predictor variable. The odds-ratio with the associated 95% confidence intervals are summarized in table 6 for both MIA and Fisher criterion.

|  |  |
| --- | --- |
| Table 6. Odds-ratio test and confidence intervals MIA (left) and Fisher criterion (right) | |
|  |  |
| Model Independent Approach (MIA) | Fisher Criterion |

**4.3 Model Accuracy, AUC and ROC curve**

The area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example. It measures the classifiers skill in ranking a set of patterns according to the degree to which they belong to the positive class, but without actually assigning patterns to classes.

The overall accuracy also depends on the ability of the classifier to rank patterns, but also on its ability to select a threshold in the ranking used to assign patterns to the positive class if above the threshold and to the negative class if below. The model accuracy, AUC and ROC curves are given in table 7 below.

|  |  |  |
| --- | --- | --- |
| Table 7. AUC and ROC Curves MIA (left) and Fisher criterion (right) | | |
| **Accuracy** | **0.6459** | **0.6485** |
| **AUC** | **0.6944** | **0.6894** |
| **ROC curve** |  |  |
|  | Model Independent Approach (MIA) | Fisher Criterion |

# 5.0 Conclusion

We have analysed the the news popularity dataset of 39644 articles. We have used a different filter method and compared to that of the baseline given in [7]. The accuracy results obtained shows similar values to know state-of-the-art algorithms. In regression analysis and as well prediction requires models with smaller number of predictor variables. However, datasets of like that considered in this project poses dimensionality issues. As studied in this project, a further research could be carried to find alternative methods to extract important features. An approach could be to extract features using a number of methods and finally selecting a subset.

# References

1. Bandari, R., S. Asur, and B.A. Huberman, *The pulse of news in social media: Forecasting popularity.* arXiv preprint arXiv:1202.0332, 2012.

2. Ahmed, M., et al. *A peek into the future: predicting the evolution of popularity in user generated content*. in *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013. ACM.

3. Fernandes, K., P. Vinagre, and P. Cortez. *A proactive intelligent decision support system for predicting the popularity of online news*. in *Portuguese Conference on Artificial Intelligence*. 2015. Springer.

4. Szabo, G. and B.A. Huberman, *Predicting the popularity of online content.* Communications of the ACM, 2010. **53**(8): p. 80-88.

5. Tatar, A., et al., *A survey on predicting the popularity of web content.* Journal of Internet Services and Applications, 2014. **5**(1): p. 1.

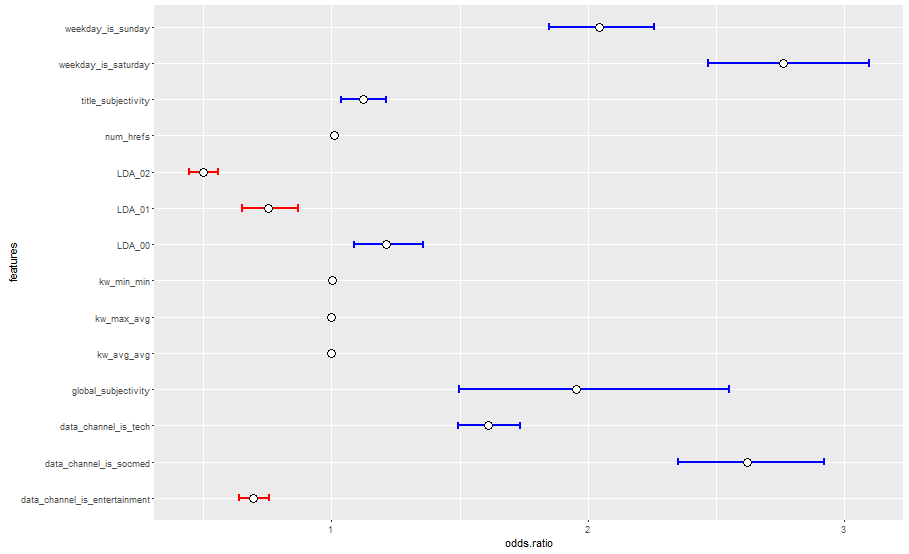
6. Tatar, A., et al., *From popularity prediction to ranking online news.* Social Network Analysis and Mining, 2014. **4**(1): p. 1-12.

7. Ren, H. and Q. Yang, *Predicting and Evaluating the Popularity of Online News*.

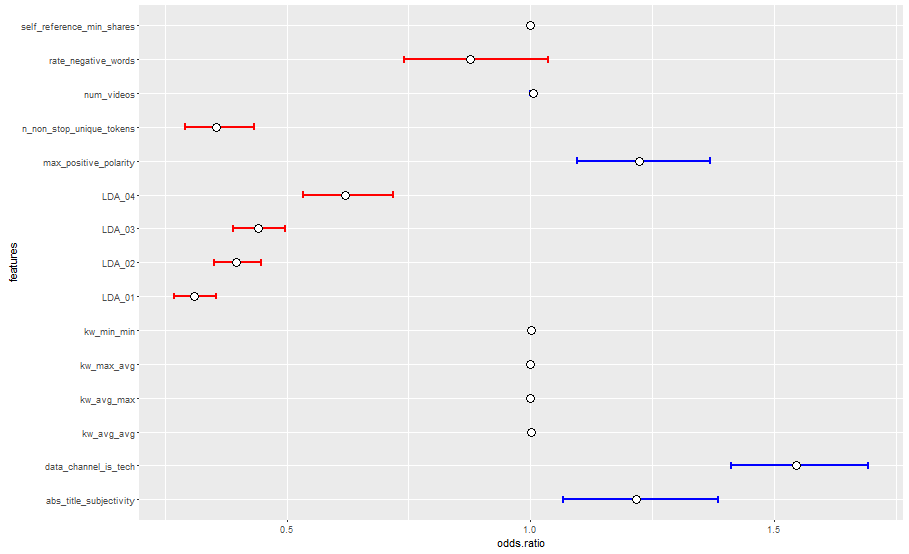
# Appendix

## Plot of Odd ratio and 95% C.I.

1. *Fisher Criterion*



1. *Model Independent Approach (MIA)*



## Codes for the Fisher criterion feature selection model

**#Read input data**

online.data<- read.csv("data/OnlineNewsPopularity.csv")

summary(online.data)

train.percentage<-0.7 #70% of original length

end.index<- ceiling(train.percentage\*length(online.data$shares)) #27751

train.data<-online.data[1:end.index,]

test.data<-online.data[(end.index+1):length(online.data$shares),]

**#Full model fitting**

online.logit.model<-glm(shares ~.,family=binomial(link='logit'),data=train.data)

online.logit.model2<-glm(shares ~.,family=binomial(link='logit'),data=train.data[,-12])

summary(online.logit.model)

anova(online.logit.model,test="Chisq")

summary(online.logit.model)$cov.scaled

layout(matrix(c(1,2,3,4),2,2))

plot((online.logit.model))

**#Stepwise AIC model**

**library(MASS)**

online.model.AIC<-stepAIC(online.logit.model,direction="both")

**#Reduced model fitting**

online.logit.reduced<-glm(formula = shares ~ num\_hrefs + data\_channel\_is\_entertainment +

data\_channel\_is\_socmed + data\_channel\_is\_tech + kw\_min\_min +

kw\_max\_avg + kw\_avg\_avg + weekday\_is\_saturday + weekday\_is\_sunday +

LDA\_00 + LDA\_01 + LDA\_02 + global\_subjectivity +

title\_subjectivity, family = binomial(link = "logit"), data = train.data)

summary(online.logit.reduced)

anova(online.logit.reduced,test="Chisq")

**## odds ratios and 95% CI**

online.model.or<-exp(cbind(OR = coef(online.logit.reduced), confint(online.logit.reduced)))

**library(ggplot2)**

d=data.frame(features=c("num\_hrefs","data\_channel\_is\_entertainment",

"data\_channel\_is\_socmed", "data\_channel\_is\_tech", "kw\_min\_min",

"kw\_max\_avg" , "kw\_avg\_avg" , "weekday\_is\_saturday", "weekday\_is\_sunday" ,

"LDA\_00","LDA\_01" , "LDA\_02" , "global\_subjectivity", "title\_subjectivity"),

odds.ratio=online.model.or[2:15,1], lower=online.model.or[2:15,2], upper=online.model.or[2:15,3])

ggplot() +

geom\_errorbar(data=d, mapping=aes(x=features, ymin=upper, ymax=lower), width=0.2, size=1, color= ifelse(d$odds.ratio < 1,'red','blue')) +

geom\_point(data=d, mapping=aes(x=features, y=odds.ratio), size=4, shape=21, fill="white")+ coord\_flip()

#+ opts(title="geom\_errorbar", plot.title=theme\_text(size=40, vjust=1.5))

**#Perform wald test**

**library(aod)**

online.model.wald<-wald.test(b = coef(online.logit.reduced), Sigma = vcov(online.logit.reduced), Terms = 2:14)

**#model prediction**

fitted.results <- predict(online.logit.reduced,newdata=subset(test.data,select=c(1:20)),type='response')

fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != test.data$shares)

print(paste('Accuracy',1-misClasificError))

**#Likelihood ratio test**

**library(lmtest)**

likelihood.test<-lrtest(online.logit.reduced,online.logit.model)

likelihood.test

**#ROC**

**library(ROCR)**

**library(ggplot2)**

p <- predict(online.logit.reduced,newdata=subset(test.data,select=c(1:20)),type='response')

pr <- prediction(p, test.data$shares)

prf <- performance(pr, measure = "tpr", x.measure = "fpr")

plot(prf,colorize=TRUE, lwd=5)

**#AUC - Area Under Curve**

auc <- performance(pr, measure = "auc")

auc <- auc@y.values[[1]]

print(paste('AUC = ',auc))

## Codes for the MIA feature selection model

**#Read input data**

online.data<- read.csv("data/OnlineNewsPopularity.csv")

online.data<-subset(online.data,select=c(27,42,25,44,41,28,24,18,29,7,53,9,50,20,21,11,59,43,10,40,61))

summary(online.data)

shares.trans<-ifelse(online.data$shares > 1400,1,0)

online.data$shares<-shares.trans

train.percentage<-0.7 #70% of original length

end.index<- ceiling(train.percentage\*length(online.data$shares)) #27751

train.data<-online.data[1:end.index,]

test.data<-online.data[(end.index+1):length(online.data$shares),]

**#Full model fitting**

online.logit.model<-glm(shares ~.,family=binomial(link='logit'),data=train.data)

online.logit.model2<-glm(shares ~.,family=binomial(link='logit'),data=train.data[,-20])

summary(online.logit.model)

anova(online.logit.model,test="Chisq")

summary(online.logit.model)$cov.scaled

layout(matrix(c(1,2,3,4),2,2))

plot((online.logit.model))

**## odds ratios and 95% CI**

online.model.or<-exp(cbind(OR = coef(online.logit.model), confint(online.logit.model)))

**#Stepwise AIC model**

**library(MASS)**

online.model.AIC<-stepAIC(online.logit.model,direction="both")

**#Reduced model fitting**

online.logit.reduced<-glm(formula = shares ~ kw\_max\_avg + LDA\_02 + kw\_avg\_max + LDA\_04 +

LDA\_01 + kw\_avg\_avg + data\_channel\_is\_tech + self\_reference\_min\_shares +

n\_non\_stop\_unique\_tokens + max\_positive\_polarity + rate\_negative\_words +

kw\_min\_min + num\_videos + abs\_title\_subjectivity + LDA\_03,

family = binomial(link = "logit"), data = train.data)

summary(online.logit.reduced)

anova(online.logit.reduced,test="Chisq")

**## odds ratios and 95% CI**

online.model.or<-exp(cbind(OR = coef(online.logit.reduced), confint(online.logit.reduced)))

**library(ggplot2)**

d=data.frame(features=c("kw\_max\_avg" , "LDA\_02", "kw\_avg\_max", "LDA\_04" ,"LDA\_01", "kw\_avg\_avg" , "data\_channel\_is\_tech", "self\_reference\_min\_shares",

"n\_non\_stop\_unique\_tokens" , "max\_positive\_polarity" , "rate\_negative\_words" ,

"kw\_min\_min" , "num\_videos" , "abs\_title\_subjectivity" , "LDA\_03"),

odds.ratio=online.model.or[2:16,1], lower=online.model.or[2:16,2], upper=online.model.or[2:16,3])

ggplot() +

geom\_errorbar(data=d, mapping=aes(x=features, ymin=upper, ymax=lower), width=0.2, size=1, color= ifelse(d$odds.ratio < 1,'red','blue')) +

geom\_point(data=d, mapping=aes(x=features, y=odds.ratio), size=4, shape=21, fill="white")+ coord\_flip()

#+ opts(title="geom\_errorbar", plot.title=theme\_text(size=40, vjust=1.5))

**#Perform wald test**

**library(aod)**

online.model.wald<-wald.test(b = coef(online.logit.model2), Sigma = vcov(online.logit.model2), Terms = 11)

online.model.wald<-wald.test(b = coef(online.logit.reduced), Sigma = vcov(online.logit.reduced), Terms = 1:15)

**#model prediction**

fitted.results <- predict(online.logit.reduced,newdata=subset(test.data,select=c(1:20)),type='response')

fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != test.data$shares)

print(paste('Accuracy',1-misClasificError))

**#Likelihood ratio test**

**library(lmtest)**

likelihood.test<-lrtest(online.logit.reduced,online.logit.model)

likelihood.test

**#ROC**

**library(ROCR)**

**library(ggplot2)**

p <- predict(online.logit.reduced,newdata=subset(test.data,select=c(1:20)),type='response')

pr <- prediction(p, test.data$shares)

prf <- performance(pr, measure = "tpr", x.measure = "fpr")

plot(prf,colorize=TRUE, lwd=5)

**#AUC - Area Under Curve**

auc <- performance(pr, measure = "auc")

auc <- auc@y.values[[1]]

print(paste('AUC = ',auc))

**#Feature Selection**

***install.packages("subselect")***

online.data.original<- read.csv("data/OnlineNewsPopularity.csv")

**library(mlbench)**

**library(caret)**

**# calculate correlation matrix**

correlationMatrix <- cor(online.data.original[,2:60])

# summarize the correlation matrix

print(correlationMatrix)

# find attributes that are highly corrected (ideally >0.75)

highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.5)

# print indexes of highly correlated attributes

print(highlyCorrelated)

**#Rank features by importance**

*# prepare training scheme*

control <- trainControl(method="repeatedcv", number=10, repeats=3)

# train the model

model <- train(shares~., data=online.data.original[,2:61], method="leapSeq", preProcess="scale", trControl=control)

# estimate variable importance

importance <- varImp(model, scale=FALSE,useModel = FALSE)

# summarize importance

print(importance)

# plot importance

plot(importance, top=20)

1. Source:

   https://www.wikipedia.org/ [↑](#footnote-ref-1)