Online news

Method 1: Tree based classification

step 1: collecting data

```
getwd()
```

```
## [1] "C:/Users/Madhu/Side Projects/Machine_Learning_Course"
```

```
setwd('C:/Users/Madhu/Side Projects/Machine_Learning_Course')

#Read data file

news <- read.csv('OnlineNewsPopularity.csv')

str(news)
```

```
## 'data.frame':    39644 obs. of  61 variables:
##  $ url                           : Factor w/ 39644 levels
"http://mashable.com/2013/01/07/amazon-instant-video-browser/",..: 1 2 3 4 5
6 7 8 9 10 ...
##  $ timedelta                     : num  731 731 731 731 731 731 731 731 731
731 ...
##  $ n_tokens_title                : num  12 9 9 9 13 10 8 12 11 10 ...
##  $ n_tokens_content              : num  219 255 211 531 1072 ...
##  $ n_unique_tokens               : num  0.664 0.605 0.575 0.504 0.416 ...
##  $ n_non_stop_words              : num  1 1 1 1 1 ...
##  $ n_non_stop_unique_tokens      : num  0.815 0.792 0.664 0.666 0.541 ...
##  $ num_hrefs                     : num  4 3 3 9 19 2 21 20 2 4 ...
##  $ num_self_hrefs                : num  2 1 1 0 19 2 20 20 0 1 ...
##  $ num_imgs                      : num  1 1 1 1 20 0 20 20 0 1 ...
##  $ num_videos                    : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ average_token_length          : num  4.68 4.91 4.39 4.4 4.68 ...
##  $ num_keywords                  : num  5 4 6 7 7 9 10 9 7 5 ...
##  $ data_channel_is_lifestyle     : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ data_channel_is_entertainment: num  1 0 0 1 0 0 0 0 0 0 ...
##  $ data_channel_is_bus           : num  0 1 1 0 0 0 0 0 0 0 ...
##  $ data_channel_is_socmed        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ data_channel_is_tech          : num  0 0 0 0 1 1 0 1 1 0 ...
##  $ data_channel_is_world         : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ kw_min_min                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_max_min                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_avg_min                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_min_max                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_max_max                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_avg_max                    : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ kw_min_avg                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_max_avg                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ kw_avg_avg                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ self_reference_min_shares     : num  496 0 918 0 545 8500 545 545 0 0
...
##  $ self_reference_max_shares     : num  496 0 918 0 16000 8500 16000 16000
0 0 ...
##  $ self_reference_avg_sharess    : num  496 0 918 0 3151 ...
##  $ weekday_is_monday             : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday_is_tuesday            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday_is_wednesday          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday_is_thursday           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday_is_friday             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday_is_saturday           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday_is_sunday             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ is_weekend                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LDA_00                        : num  0.5003 0.7998 0.2178 0.0286 0.0286
...
##  $ LDA_01                        : num  0.3783 0.05 0.0333 0.4193 0.0288
...
##  $ LDA_02                        : num  0.04 0.0501 0.0334 0.4947 0.0286
...
##  $ LDA_03                        : num  0.0413 0.0501 0.0333 0.0289 0.0286
...
##  $ LDA_04                        : num  0.0401 0.05 0.6822 0.0286 0.8854
...
##  $ global_subjectivity           : num  0.522 0.341 0.702 0.43 0.514 ...
##  $ global_sentiment_polarity     : num  0.0926 0.1489 0.3233 0.1007 0.281
...
##  $ global_rate_positive_words    : num  0.0457 0.0431 0.0569 0.0414 0.0746
...
##  $ global_rate_negative_words    : num  0.0137 0.01569 0.00948 0.02072
0.01213 ...
##  $ rate_positive_words           : num  0.769 0.733 0.857 0.667 0.86 ...
##  $ rate_negative_words           : num  0.231 0.267 0.143 0.333 0.14 ...
##  $ avg_positive_polarity         : num  0.379 0.287 0.496 0.386 0.411 ...
##  $ min_positive_polarity         : num  0.1 0.0333 0.1 0.1364 0.0333 ...
##  $ max_positive_polarity         : num  0.7 0.7 1 0.8 1 0.6 1 1 0.8 0.5 ...
##  $ avg_negative_polarity         : num  -0.35 -0.119 -0.467 -0.37 -0.22 ...
##  $ min_negative_polarity         : num  -0.6 -0.125 -0.8 -0.6 -0.5 -0.4 -
0.5 -0.5 -0.125 -0.5 ...
##  $ max_negative_polarity         : num  -0.2 -0.1 -0.133 -0.167 -0.05 ...
##  $ title_subjectivity            : num  0.5 0 0 0 0.455 ...
##  $ title_sentiment_polarity      : num  -0.188 0 0 0 0.136 ...
##  $ abs_title_subjectivity        : num  0 0.5 0.5 0.5 0.0455 ...
##  $ abs_title_sentiment_polarity : num  0.188 0 0 0 0.136 ...
##  $ shares                        : int  593 711 1500 1200 505 855 556 891
3600 710 ...
```

Step 2: Exploring the data

```
#Using sumary to check the statistics of amunt column in the dataset

summary(news$shares)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     946    1400    3395    2800  843300

# Adding new variable Popularity based on the market share value. Considering
the median value of 1400 to make a splut in the data as popular and unpopular

news$Popularity <- ifelse( news$shares <= 1400, "Unpopular","Popular")

news$Popularity <- as.factor(news$Popularity)


#Using table to see how many were defaulted and how many were not

table(news$Popularity)

##
##   Popular Unpopular
##     19562     20082
```

Steps to develop tree based classification

```
#Before creating the testing and training data, randomizing the observations

set.seed(12345)

news_rand <- news[order(runif(39644)),]

#checking the summary of the original data with the randomized one to notice
any substantial changes

summary(news$shares)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     946    1400    3395    2800  843300

summary(news_rand$shares)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     946    1400    3395    2800  843300
```

Splitting data to training and testing set

```
# Choosing 90% for training set and remaining

news_train <- news_rand[1:35679,]
news_test <- news_rand[35680:39644,]
```

```
#Looking at percentage split of the testing and training to see if
randomization went well

prop.table(table(news_train$Popularity))

##
##   Popular Unpopular
## 0.4932874 0.5067126

prop.table(table(news_test$Popularity))

##
##   Popular Unpopular
## 0.4948298 0.5051702
```

Training a model on the data

```
#install.packages('C50')
library(C50)

## Warning: package 'C50' was built under R version 3.5.2

# Buiding ecision tree. Since we are predicting defaulted or not, we must
specify the 17th column to represent it as the class or response variable

news_model <- C5.0(x = news_train[29:60], y = news_train$Popularity)
news_model

##
## Call:
## C5.0.default(x = news_train[29:60], y = news_train$Popularity)
##
## Classification Tree
## Number of samples: 35679
## Number of predictors: 32
##
## Tree size: 127
##
## Non-standard options: attempt to group attributes
```

Examining the decision tree

```
summary(news_model)

##
## Call:
## C5.0.default(x = news_train[29:60], y = news_train$Popularity)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Feb 19 13:29:34 2019
## -------------------------------
##
```

```
## Class specified by attribute `outcome'
##
## Read 35679 cases (33 attributes) from undefined.data
##
## Decision tree:
##
## LDA_02 > 0.5926673:
## :...is_weekend <= 0:
## :   :...rate_positive_words <= 0.04411765:
## :   :   :...weekday_is_tuesday <= 0: Unpopular (98/43)
## :   :   :   weekday_is_tuesday > 0: Popular (32/10)
## :   :   rate_positive_words > 0.04411765:
## :   :   :...self_reference_avg_sharess <= 3290.875:
## :   :       :...self_reference_min_shares <= 1100: Unpopular (2641/626)
## :   :       :   self_reference_min_shares > 1100:
## :   :       :   :...title_subjectivity <= 0.45625: Unpopular (748/207)
## :   :       :       title_subjectivity > 0.45625:
## :   :       :       :...max_negative_polarity <= -0.1333333:
## :   :       :           :...min_negative_polarity <= -0.375: Unpopular
(17/7)
## :   :       :           :   min_negative_polarity > -0.375: Popular (15)
## :   :       :           max_negative_polarity > -0.1333333:
## :   :       :           :...weekday_is_friday <= 0: Unpopular (155/55)
## :   :       :               weekday_is_friday > 0: Popular (36/16)
## :   :       self_reference_avg_sharess > 3290.875:
## :   :       :...abs_title_subjectivity <= 0.07222223:
## :   :           :...max_negative_polarity <= -0.05: Unpopular (107/26)
## :   :           :   max_negative_polarity > -0.05: Popular (8/2)
## :   :           abs_title_subjectivity > 0.07222223:
## :   :           :...weekday_is_monday <= 0: Unpopular (768/303)
## :   :               weekday_is_monday > 0:
## :   :               :...title_subjectivity <= 0.3959596: Unpopular
(126/51)
## :   :                   title_subjectivity > 0.3959596:
## :   :                   :...self_reference_max_shares <= 37400: Popular
(34/8)
## :   :                       self_reference_max_shares > 37400: Unpopular
(8/1)
## :   is_weekend > 0:
## :   :...max_negative_polarity <= -0.175:
## :       :...avg_negative_polarity <= -0.5041667: Unpopular (3)
## :       :   avg_negative_polarity > -0.5041667: Popular (37/6)
## :       max_negative_polarity > -0.175:
## :       :...self_reference_min_shares > 1400:
## :           :...global_subjectivity <= 0.4055577: Unpopular (96/43)
## :           :   global_subjectivity > 0.4055577: Popular (104/24)
## :           self_reference_min_shares <= 1400:
## :           :...min_positive_polarity <= 0.03333334: Popular (92/44)
## :               min_positive_polarity > 0.03333334:
## :               :...min_positive_polarity <= 0.1: Unpopular (350/111)
```

```
## :                   min_positive_polarity > 0.1:
## :                   :...min_positive_polarity > 0.2142857: Unpopular (6)
## :                       min_positive_polarity <= 0.2142857:
## :                       :...weekday_is_saturday <= 0: Unpopular (33/14)
## :                           weekday_is_saturday > 0: Popular (26/8)
## LDA_02 <= 0.5926673:
## :...weekday_is_saturday > 0: Popular (1834/469)
##     weekday_is_saturday <= 0:
##     :...self_reference_min_shares <= 1600:
##         :...weekday_is_sunday > 0:
##         :   :...title_subjectivity > 0.7583333: Popular (113/20)
##         :   :   title_subjectivity <= 0.7583333:
##         :   :   :...self_reference_min_shares <= 341:
##         :   :       :...LDA_01 <= 0.04000007: Popular (176/35)
##         :   :       :   LDA_01 > 0.04000007:
##         :   :       :   :...LDA_04 > 0.4476274: Popular (14/2)
##         :   :       :       LDA_04 <= 0.4476274:
##         :   :       :       :...avg_positive_polarity > 0.5148148:
Unpopular (10)
##         :   :       :           avg_positive_polarity <= 0.5148148:
##         :   :       :           :...min_negative_polarity <= -0.4666667:
[S1]
##         :   :       :               min_negative_polarity > -0.4666667:
[S2]
##         :   :       self_reference_min_shares > 341:
##         :   :       :...global_sentiment_polarity > 0.2835377: Unpopular
(36/10)
##         :   :           global_sentiment_polarity <= 0.2835377:
##         :   :           :...self_reference_min_shares <= 1200: Popular
(505/210)
##         :   :               self_reference_min_shares > 1200:
##         :   :               :...self_reference_min_shares <= 1400: Popular
(101/25)
##         :   :                   self_reference_min_shares > 1400:
##         :   :                   :...max_negative_polarity <= -0.1428571:
##         :   :                       :...LDA_00 <= 0.02131728: Popular (5)
##         :   :                       :   LDA_00 > 0.02131728: Unpopular
(17/3)
##         :   :                       max_negative_polarity > -0.1428571:
[S3]
##         :   weekday_is_sunday <= 0:
##         :   :...LDA_01 <= 0.04032645:
##         :       :...min_positive_polarity > 0.03333334:
##         :       :   :...global_subjectivity <= 0.4148676:
##         :       :   :   :...self_reference_max_shares <= 17900: Unpopular
(1759/637)
##         :       :   :   :   self_reference_max_shares > 17900: Popular
(95/40)
##         :       :   :   global_subjectivity > 0.4148676:
##         :       :   :   :...LDA_02 > 0.305243: Unpopular (778/307)
```

```
##         :       :   :       LDA_02 <= 0.305243:
##         :       :   :       :...min_negative_polarity > -0.125: Unpopular
(227/87)
##         :       :   :           min_negative_polarity <= -0.125: [S4]
##         :       :   min_positive_polarity <= 0.03333334:
##         :       :   :...LDA_00 > 0.9199746: Unpopular (26/2)
##         :       :       LDA_00 <= 0.9199746:
##         :       :       :...rate_positive_words > 0.9315069:
##         :       :           :...self_reference_max_shares <= 4700:
Unpopular (78/16)
##         :       :           :   self_reference_max_shares > 4700:
##         :       :           :   :...LDA_02 <= 0.02865748: Unpopular (4)
##         :       :           :       LDA_02 > 0.02865748: Popular (13/1)
##         :       :           rate_positive_words <= 0.9315069:
##         :       :           :...max_negative_polarity <= -0.2:
##         :       :               :...weekday_is_friday <= 0: Unpopular
(77/26)
##         :       :               :   weekday_is_friday > 0: Popular (14/4)
##         :       :               max_negative_polarity > -0.2:
##         :       :               :...max_positive_polarity > 0.8: Popular
(845/312)
##         :       :                   max_positive_polarity <= 0.8: [S5]
##         :       LDA_01 > 0.04032645:
##         :       :...title_sentiment_polarity > 0.85: Popular (112/51)
##         :           title_sentiment_polarity <= 0.85:
##         :           :...min_positive_polarity <= 0.03333334:
##         :               :...LDA_04 > 0.4622981: Popular (106/39)
##         :               :   LDA_04 <= 0.4622981:
##         :               :   :...LDA_00 <= 0.1497432: Unpopular (702/250)
##         :               :       LDA_00 > 0.1497432:
##         :               :       :...weekday_is_tuesday <= 0: Popular
(457/208)
##         :               :           weekday_is_tuesday > 0: Unpopular
(120/51)
##         :               min_positive_polarity > 0.03333334:
##         :               :...self_reference_avg_sharess <= 2800.333: [S6]
##         :                   self_reference_avg_sharess > 2800.333:
##         :                   :...self_reference_max_shares > 61600: Popular
(54/17)
##         :                       self_reference_max_shares <= 61600: [S7]
##         self_reference_min_shares > 1600:
##         :...weekday_is_sunday > 0: Popular (858/210)
##             weekday_is_sunday <= 0:
##             :...LDA_01 > 0.3645431:
##                 :...self_reference_max_shares > 11300: Popular (410/178)
##                 :   self_reference_max_shares <= 11300:
##                 :   :...global_subjectivity <= 0.6339827: Unpopular
(924/371)
##                 :       global_subjectivity > 0.6339827: Popular (30/7)
##                 LDA_01 <= 0.3645431:
```

```
##                 :...min_positive_polarity <= 0.03333334: Popular
(1929/583)
##                     min_positive_polarity > 0.03333334:
##                     :...min_negative_polarity > -0.1428571:
##                         :...title_sentiment_polarity > 0.3875: Popular
(94/25)
##                         :   title_sentiment_polarity <= 0.3875:
##                         :   :...weekday_is_friday <= 0: Unpopular
(484/206)
##                         :       weekday_is_friday > 0: Popular (96/36)
##                         min_negative_polarity <= -0.1428571:
##                         :...self_reference_avg_sharess > 14302.25: Popular
(1209/387)
##                             self_reference_avg_sharess <= 14302.25:
##                             :...LDA_01 <= 0.02009844:
##                                 :...weekday_is_wednesday <= 0: Popular
(399/106)
##                                 :   weekday_is_wednesday > 0:
##                                 :   :...min_positive_polarity <= 0.16:
[S8]
##                                 :       min_positive_polarity > 0.16:
##                                 :       :...LDA_00 <= 0.02001115: Popular
(4)
##                                 :           LDA_00 > 0.02001115: Unpopular
(8)
##                                 LDA_01 > 0.02009844:
##                                 :...LDA_02 <= 0.03339405:
##                                     :...weekday_is_friday <= 0: Popular
(1998/774)
##                                     :   weekday_is_friday > 0: [S9]
##                                     LDA_02 > 0.03339405: [S10]
##
## SubTree [S1]
##
## self_reference_min_shares <= 194: Unpopular (103/38)
## self_reference_min_shares > 194: Popular (6)
##
## SubTree [S2]
##
## title_sentiment_polarity <= 0.4166667: Popular (66/18)
## title_sentiment_polarity > 0.4166667: Unpopular (14/4)
##
## SubTree [S3]
##
## max_negative_polarity <= -0.025: Popular (69/18)
## max_negative_polarity > -0.025: Unpopular (4)
##
## SubTree [S4]
##
## self_reference_min_shares <= 151: Popular (929/410)
```

```
## self_reference_min_shares > 151:
## :...LDA_00 > 0.4795403: Unpopular (442/158)
##     LDA_00 <= 0.4795403:
##     :...weekday_is_friday > 0: Popular (415/179)
##         weekday_is_friday <= 0:
##         :...self_reference_max_shares <= 13800:
##             :...self_reference_min_shares <= 1000: Unpopular (879/380)
##             :   self_reference_min_shares > 1000: Popular (870/410)
##             self_reference_max_shares > 13800:
##             :...min_positive_polarity <= 0.05: Popular (54/12)
##                 min_positive_polarity > 0.05:
##                 :...self_reference_min_shares <= 551: Unpopular (26/6)
##                     self_reference_min_shares > 551: Popular (262/105)
##
## SubTree [S5]
##
## min_positive_polarity <= 0: Popular (532/228)
## min_positive_polarity > 0:
## :...LDA_00 <= 0.1456803: Unpopular (197/69)
##     LDA_00 > 0.1456803:
##     :...weekday_is_wednesday <= 0: Popular (341/137)
##         weekday_is_wednesday > 0:
##         :...self_reference_min_shares > 1000:
##             :...min_negative_polarity <= -0.1666667: Popular (33/9)
##             :   min_negative_polarity > -0.1666667: Unpopular (4)
##             self_reference_min_shares <= 1000:
##             :...global_rate_positive_words > 0.07556675: Popular (4)
##                 global_rate_positive_words <= 0.07556675:
##                 :...avg_negative_polarity <= -0.4351852: Popular (4)
##                     avg_negative_polarity > -0.4351852: Unpopular (63/19)
##
## SubTree [S6]
##
## self_reference_avg_sharess > 331: Unpopular (2910/848)
## self_reference_avg_sharess <= 331:
## :...max_positive_polarity <= 0.375: Popular (26/7)
##     max_positive_polarity > 0.375:
##     :...weekday_is_thursday <= 0: Unpopular (856/340)
##         weekday_is_thursday > 0:
##         :...min_positive_polarity > 0.15:
##             :...abs_title_subjectivity <= 0.3571429: Popular (26/3)
##             :   abs_title_subjectivity > 0.3571429: Unpopular (18/7)
##             min_positive_polarity <= 0.15:
##             :...LDA_02 > 0.311181: Unpopular (31/4)
##                 LDA_02 <= 0.311181:
##                 :...min_positive_polarity > 0.05: Unpopular (121/51)
##                     min_positive_polarity <= 0.05:
##                     :...LDA_03 <= 0.02686957: Unpopular (4)
##                         LDA_03 > 0.02686957: Popular (15/3)
##
```

```
## SubTree [S7]
##
## self_reference_min_shares <= 860: Unpopular (439/149)
## self_reference_min_shares > 860:
## :...weekday_is_friday > 0: Popular (98/43)
##     weekday_is_friday <= 0:
##     :...abs_title_subjectivity <= 0.475: Unpopular (303/112)
##         abs_title_subjectivity > 0.475:
##         :...LDA_01 > 0.4830296:
##             :...LDA_03 <= 0.02397273: Popular (9/1)
##             :   LDA_03 > 0.02397273:
##             :   :...LDA_00 > 0.04000217: Unpopular (23)
##             :       LDA_00 <= 0.04000217:
##             :       :...weekday_is_wednesday <= 0: Unpopular (32/8)
##             :           weekday_is_wednesday > 0:
##             :           :...LDA_01 <= 0.8497025: Popular (9)
##             :               LDA_01 > 0.8497025: Unpopular (3)
##             LDA_01 <= 0.4830296:
##             :...avg_negative_polarity > -0.1461111: Unpopular (21/4)
##                 avg_negative_polarity <= -0.1461111:
##                 :...max_positive_polarity <= 0.55:
##                     :...title_subjectivity <= 0.5: Unpopular (36/14)
##                     :   title_subjectivity > 0.5: Popular (3)
##                     max_positive_polarity > 0.55:
##                     :...weekday_is_tuesday > 0: Popular (43/10)
##                         weekday_is_tuesday <= 0:
##                         :...min_negative_polarity > -0.875: Popular
(95/28)
##                             min_negative_polarity <= -0.875:
##                             :...title_sentiment_polarity <= -0.25: Popular
(3)
##                                 title_sentiment_polarity > -0.25:
Unpopular (25/8)
##
## SubTree [S8]
##
## min_negative_polarity <= -1: Unpopular (10/3)
## min_negative_polarity > -1: Popular (84/30)
##
## SubTree [S9]
##
## min_negative_polarity > -0.8: Popular (337/107)
## min_negative_polarity <= -0.8:
## :...min_positive_polarity <= 0.0625: Unpopular (12/3)
##     min_positive_polarity > 0.0625:
##     :...max_positive_polarity > 0.75: Popular (60/20)
##         max_positive_polarity <= 0.75:
##         :...LDA_03 <= 0.8297771: Unpopular (19/4)
##             LDA_03 > 0.8297771: Popular (9/1)
##
```

```
## SubTree [S10]
##
## global_rate_positive_words <= 0.01834862: Unpopular (198/73)
## global_rate_positive_words > 0.01834862:
## :...title_sentiment_polarity > 0.205: Popular (445/162)
##     title_sentiment_polarity <= 0.205:
##     :...max_negative_polarity <= -0.375: Unpopular (79/28)
##         max_negative_polarity > -0.375:
##         :...weekday_is_friday <= 0: Popular (1413/667)
##             weekday_is_friday > 0:
##             :...min_negative_polarity > -0.1785714: Popular (17/2)
##                 min_negative_polarity <= -0.1785714:
##                 :...abs_title_subjectivity > 0.2583333: Popular (199/74)
##                     abs_title_subjectivity <= 0.2583333:
##                     :...global_subjectivity <= 0.3878394: Unpopular (12)
##                         global_subjectivity > 0.3878394:
##                         :...min_positive_polarity > 0.15: Popular (4)
##                             min_positive_polarity <= 0.15:
##                             :...avg_negative_polarity <= -0.3045549:
Unpopular (17/2)
##                                 avg_negative_polarity > -0.3045549:
##                                 :...LDA_01 <= 0.04559081: Unpopular
(26/10)
##                                     LDA_01 > 0.04559081: Popular (11)
##
##
## Evaluation on training data (35679 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##     127 12336(34.6%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##   11805  5795    (a): class Popular
##    6541 11538    (b): class Unpopular
##
##
##  Attribute usage:
##
##  100.00% LDA_02
##   91.44% self_reference_min_shares
##   84.64% weekday_is_saturday
##   79.33% weekday_is_sunday
##   74.55% LDA_01
##   70.74% min_positive_polarity
##   46.07% self_reference_avg_sharess
```

```
##   32.87% min_negative_polarity
##   27.12% title_sentiment_polarity
##   23.00% weekday_is_friday
##   22.31% global_subjectivity
##   19.62% rate_positive_words
##   18.62% self_reference_max_shares
##   18.39% LDA_00
##   15.53% is_weekend
##   14.22% max_negative_polarity
##    9.57% max_positive_polarity
##    6.98% global_rate_positive_words
##    6.77% title_subjectivity
##    5.52% abs_title_subjectivity
##    4.48% LDA_04
##    3.00% weekday_is_thursday
##    2.80% weekday_is_wednesday
##    2.62% weekday_is_monday
##    2.45% weekday_is_tuesday
##    2.07% global_sentiment_polarity
##    1.08% avg_negative_polarity
##    0.56% avg_positive_polarity
##    0.34% LDA_03
##
##
## Time: 2.6 secs
```

As we can see, 11805 were classified as class a - Popular, 11538 were classified as class b - Unpopular .

step 4: Evaluating Model performance

```
news_pred <- predict(news_model, news_test)
# using gmodels ro create confusion matrix

#install.packages('gmodels')

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.5.2

CrossTable(news_test$Popularity, news_pred, prop.chisq = FALSE, prop.c =
FALSE, prop.r = FALSE, dnn = c('actual Popularity', 'predicted Popularity'))

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
```

```
##
##
## Total Observations in Table:  3965
##
##
##                   | predicted Popularity
## actual Popularity |   Popular | Unpopular | Row Total |
## ------------------|-----------|-----------|-----------|
##           Popular |      1228 |       734 |      1962 |
##                   |     0.310 |     0.185 |           |
## ------------------|-----------|-----------|-----------|
##         Unpopular |       810 |      1193 |      2003 |
##                   |     0.204 |     0.301 |           |
## ------------------|-----------|-----------|-----------|
##      Column Total |      2038 |      1927 |      3965 |
## ------------------|-----------|-----------|-----------|
##
##
```

As seen from the above confusion matrix, 810 were misclassified as a Type 2 error and 734 were type 1 that is it was Popular but classifed as Unpopular.

Method#2: Adding Regression to trees

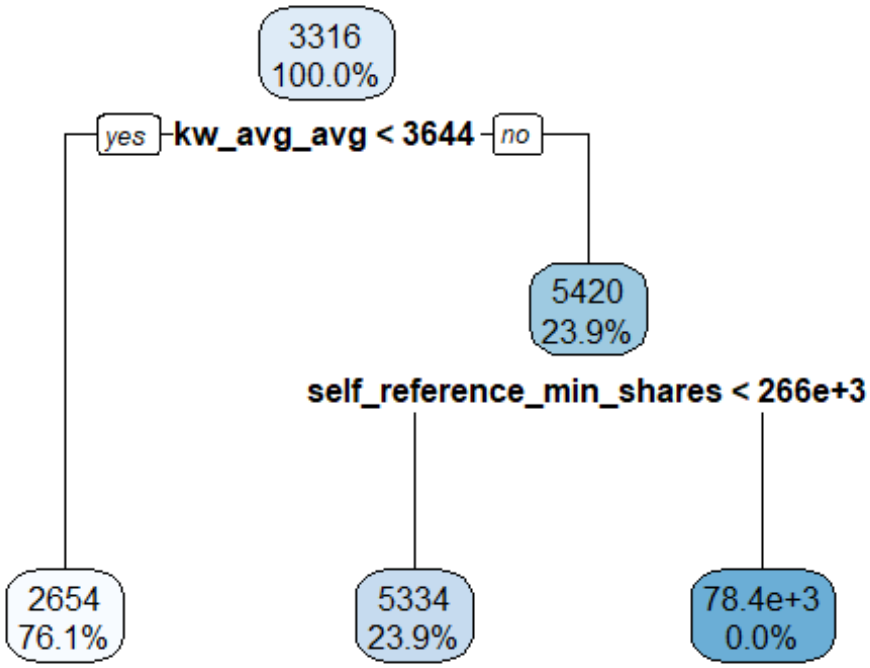
```
# Installing rpart - recursive partitioning

#install.packages('rpart')
library(rpart)

## Warning: package 'rpart' was built under R version 3.5.2

#reg_data_train <- news_train[c(29:60,62)]

m.rpart <- rpart(shares ~ ., data=news_train[c(2:61)])
m.rpart

## n= 35679
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
## 1) root 35679 3.655753e+12  3316.064
##   2) kw_avg_avg< 3643.616 27138 1.469896e+12  2654.039 *
##   3) kw_avg_avg>=3643.616 8541 2.136172e+12  5419.566
##     6) self_reference_min_shares< 265900 8531 1.706600e+12  5334.007 *
##     7) self_reference_min_shares>=265900 10 3.762334e+11 78410.000 *

#install.packages('rpart.plot')

library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.5.2
```

```
#Visualizing the tree
```

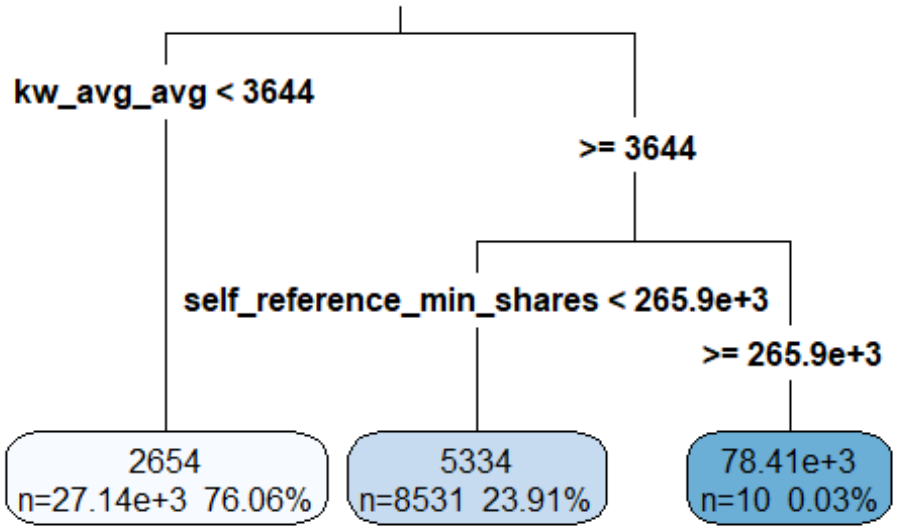
```
rpart.plot(m.rpart, digits=3)
```

```
## Warning: Bad 'data' field in model 'call' (expected a data.frame or a
matrix).
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```




```
rpart.plot(m.rpart, digits=4, fallen.leaves = TRUE, type = 3, extra = 101)
```

```
## Warning: Bad 'data' field in model 'call' (expected a data.frame or a
matrix).
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```



The last and final step 4 of evaluating model performance

```
p.rpart <- predict(m.rpart, news_test)
summary(p.rpart)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2654    2654    2654    3331    2654   78410

summary(news_test$Popularity)

##   Popular Unpopular
##      1962      2003

cor(p.rpart, news_test$shares)

## [1] 0.0615543
```

A 61% correlation is seen.