**Problem 4**

4.1 **Building Prediction Models**

Tools: RStudio, Python

Steps:

1. Data Preprocessing:
2. Concatenated all the csv files having test data with the help of Python

Python Code for concatenation:

import glob

import pandas as pd

path ='C:/Users/vanwu/Desktop/INFO 7390 ADS/Midterm/BlogFeedback/test'

allFiles = glob.glob(path + "/\*.csv")

frame = pd.DataFrame()

list\_ = []

for file\_ in allFiles:

df = pd.read\_csv(file\_, header = None)

list\_.append(df)

frame = pd.concat(list\_)

frame.to\_csv('blog\_data\_test.csv', sep=',')

1. Looking into the data and referring the paper, we inferred that the significance of 200 bag-of-words is very weak in predicting the number of comments in next 24 hours. Also, the trend of the words may change which is not predictable. Hence, we deleted the 200 columns containing bag-of-words.

2. Build Models:

1. Prediction has been performed using the following machine learning algorithms: Multiple Linear Regression, CART(Classification and Regression Tree) and Random Forest.

4.2 Performance Evaluation

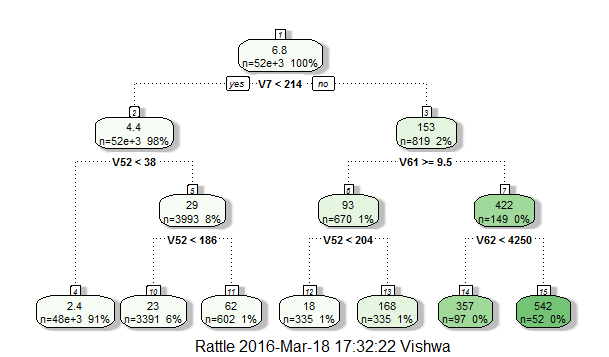
Tool: RStudio

Measurement: RMSE value

1. Linear Regression:
2. Applying the linear regression formula in training data and predicting the rmse value by applying the model to the testing data gives rmse value to be 25.24 and multiple R-squared value to be 0.36

2. CART:

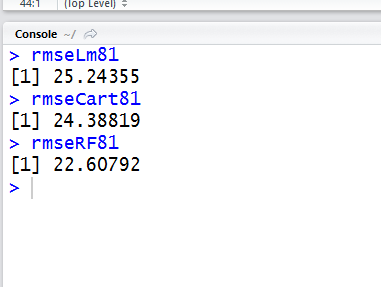
1. The packages used for CART algorithm for calculating RMSE value and visualization of the tree are: rpart, rattle, dynamicGraph, rattle, rpart.plot, cartFit81
2. The RMSE value calculated with CART is 24.39 which is less compared to Regression model.
3. The regression tree that we got from CART algorithm is:



3. Random Forest:

1. The package used for Random Forest in R is ‘randomForest’
2. The RMSE value calculated with Random Forest is 2.20 which is the lowest of all the models.

The list of RMSE values calculated using all the algorithms in R are:



4.3 Model Selection and Recommendation

1. Model Selection:

In order to perform model selection, the following assumptions need to be made:

**Positive**

Assumption 1:

More Total/Average/Max/Min (1,3-5, 6, 8-10, 11, 13-15) comments before basetime (51)/ within the 24 hours before basetime (52)/ within 48-24 hours before basetime, there may be more comments in next 24 hours (53).

Assumption 2:

More Total/Average/Max/Min (26, 28-30, 31, 33-35, 36, 38-40) trackbacks before basetime(56)/ within the 24 hours before basetime (57)/ within 48-24 hours before basetime(58), there may be more comments in next 24 hours.

Assumption 3:

More Average/Max/Min (278 - 280) of comments the parent posts (277) received, there may be more comments in next 24 hours.

**Negative**

Assumption 1:

Longer the hours between publication & basetime (61), lesser comment will be in the next 24 hours.

**Uncertain**

Assumption 1:

Weekday of post publication (270-276) may affect the comments in the next 24 hours.

Assumption 2:

Weekday of basetime (263-269) Sat, Sun More

Assumption 3:

Length of post (62), too short or too long, less comments

Assumption 4:

Bag of words (63-262) not known, trends change.

Assumption 5:

Standard deviation (2,7,12,17,23,27,32,37,42,47) may affect consistency

Assumption 6:

24 hours after publication & before basetime:

1. Published at least 24 hours before basetime. More comments/links (54, 59), more comments in the next 24 hours. Positive
2. Published less than 24 hours before basetime. -**--- to be determined.**

Assumption 7:

Difference of comments/Link (55,60) number between last 24 hours and last last 24 hours.

1. If >=0, attention grows/maintains, positive
2. If <0, attention drops, negative

At first we calculated the RMSE value of all the algorithms using all the parameters of the training and testing dataset. The results we gained using 281 parameters are:

Linear Regression: 21.96

CART: 27.06

Random Forest: 28.39

According to this, regression model would be a good model.

After close analysis, we found out that the significance of the bag-of-words on the model very less. Hence, we removed columns with bag-of-words from the csv files and use those files for all the algorithms. The result we received with 81 parameters are:

Linear Regression: 25.24

CART: 24.39

Random Forest: 22.61

2. Model Recommendation:

If we take into account all the parameters, then Linear Regression is good, but if we remove bag-of-words columns and evaluate the model Random Forest is good

http://www.cs. bme.hu/~buza/pdfs/gfkl\_buza\_social\_media.pdf