**Assignment No. 15.1**

**Facebook Comment Volume Dataset Data Set**

1 Page Popularity/likes Defines the popularity or support for the source of the document.   
  
2 Page Checkinsâ€™s Describes how many individuals so far visited this place. This feature is only associated with the places eg:some institution, place, theater etc.   
  
3 Page talking about Defines the daily interest of individuals towards source of the document/ Post. The people who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares, etc by visitors to the page.   
  
4 Page Category Defines the category of the source of the document eg: place, institution, brand etc.   
  
5 - 29 Derived These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features.   
  
30 CC1 The total number of comments before selected base date/time.   
  
31 CC2 The number of comments in last 24 hours, relative to base date/time.   
  
32 CC3 The number of comments in last 48 to last 24 hours relative to base date/time.   
  
33 CC4 The number of comments in the first 24 hours after the publication of post but before base date/time.   
  
34 CC5 The difference between CC2 and CC3.   
  
35 Base time Selected time in order to simulate the scenario.   
  
36 Post length Character count in the post.   
  
37 Post Share Count This features counts the no of shares of the post, that how many peoples had shared this post on to their timeline.   
  
38 Post Promotion Status To reach more people with posts in News Feed, individual promote their post and this features tells that whether the post is promoted(1) or not(0).   
  
39 H Local This describes the H hrs, for which we have the target variable/ comments received.   
  
40-46 Post published weekday This represents the day(Sunday...Saturday) on which the post was published.   
  
47-53 Base DateTime weekday This represents the day(Sunday...Saturday) on selected base Date/Time.   
  
54 Target Variable The no of comments in next H hrs(H is given in Feature no 39).

a. Predict the no of comments in next H hrs

b. Use regression technique

c. Report the training accuracy and test accuracy

Features\_Variant\_1 <- read.csv("E:/kamagyana/Computing/DARET/Assignments/Dataset/Training/Features\_Variant\_1.csv", header=FALSE, stringsAsFactors=FALSE)

View(Features\_Variant\_1)

FV1 <- Features\_Variant\_1

nrow(FV1)

ncol(FV1)

sum(is.na(FV1))

Features\_Variant\_2 <- read.csv("E:/kamagyana/Computing/DARET/Assignments/Dataset/Training/Features\_Variant\_2.csv", header=FALSE, stringsAsFactors=FALSE)

View(Features\_Variant\_2)

FV2 <- Features\_Variant\_2

nrow(FV2)

ncol(FV2)

sum(is.na(FV2))

Features\_TestSet <- read.csv("E:/kamagyana/Computing/DARET/Assignments/Dataset/Testing/Features\_TestSet.csv", header=FALSE, stringsAsFactors=FALSE)

View(Features\_TestSet)

test <- Features\_TestSet

nrow(test)

ncol(test)

sum(is.na(test))

colnames(FV1)[c(1:4)] <- c("PPop","PChek","PTalk","PCat")

colnames(FV1)

colnames(FV1)[c(30:54)] <- c("tcbbt","cl24bbt","c4824bbt","cf24ap","diff4824","bt","Polen","PoShare","PoPromo","H","Psun","Pmon","Ptue","Pwed","Pthu","Pfri","Psat","Bsun","Bmon","Btue","Bwed","Bthu","Bfri","Bsat","ncH")

colnames(FV1)

basemodel <- lm(ncH~.,data = FV1)

summary(basemodel)

Call:

lm(formula = ncH ~ ., data = FV1)

Residuals:

Min 1Q Median 3Q Max

-346.17 -5.34 -1.13 2.94 1271.58

Coefficients: (5 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.728e+00 2.025e+00 -1.841 0.065618 .

PPop 2.323e-08 2.866e-08 0.810 0.417776

PChek -1.268e-05 7.571e-06 -1.675 0.093959 .

PTalk -1.930e-05 2.271e-06 -8.499 < 2e-16 \*\*\*

PCat -1.013e-02 7.691e-03 -1.318 0.187622

V5 -5.684e-01 1.219e-01 -4.661 3.16e-06 \*\*\*

V6 1.439e-02 6.206e-03 2.319 0.020399 \*

V7 9.640e-01 1.225e-01 7.866 3.74e-15 \*\*\*

V8 1.328e-01 5.869e-02 2.263 0.023624 \*

V9 -1.608e-01 8.272e-02 -1.944 0.051925 .

V10 -1.170e-01 4.164e-02 -2.810 0.004955 \*\*

V11 -3.753e-03 5.901e-03 -0.636 0.524860

V12 1.971e-01 5.551e-02 3.551 0.000384 \*\*\*

V13 1.307e-01 3.322e-02 3.933 8.39e-05 \*\*\*

V14 -3.238e-02 3.992e-02 -0.811 0.417301

V15 1.116e-01 1.140e-01 0.979 0.327554

V16 1.202e-02 4.022e-03 2.990 0.002796 \*\*

V17 1.245e-01 4.895e-02 2.543 0.010986 \*

V18 -2.496e-02 4.024e-02 -0.620 0.535138

V19 -1.406e-01 3.519e-02 -3.995 6.47e-05 \*\*\*

V20 5.264e-01 1.342e-01 3.924 8.74e-05 \*\*\*

V21 -1.644e-02 6.521e-03 -2.522 0.011682 \*

V22 -1.046e+00 1.262e-01 -8.293 < 2e-16 \*\*\*

V23 -1.495e-01 5.882e-02 -2.541 0.011062 \*

V24 2.144e-01 8.734e-02 2.455 0.014099 \*

V25 9.299e-04 3.719e-03 0.250 0.802557

V26 2.191e-03 5.359e-03 0.409 0.682658

V27 NA NA NA NA

V28 -3.228e-02 2.684e-02 -1.203 0.229069

V29 9.731e-03 2.394e-02 0.406 0.684394

tcbbt 4.422e-02 1.640e-02 2.697 0.006992 \*\*

cl24bbt 2.197e-01 3.726e-03 58.965 < 2e-16 \*\*\*

c4824bbt -2.262e-02 3.511e-03 -6.444 1.18e-10 \*\*\*

cf24ap -6.825e-02 1.826e-02 -3.739 0.000185 \*\*\*

diff4824 NA NA NA NA

bt -1.924e-01 7.406e-03 -25.972 < 2e-16 \*\*\*

Polen -3.277e-05 3.850e-04 -0.085 0.932168

PoShare 2.910e-03 1.768e-04 16.456 < 2e-16 \*\*\*

PoPromo NA NA NA NA

H 4.130e-01 7.877e-02 5.243 1.59e-07 \*\*\*

Psun -1.070e-02 5.926e-01 -0.018 0.985596

Pmon -1.522e-01 6.238e-01 -0.244 0.807230

Ptue -4.862e-02 6.546e-01 -0.074 0.940795

Pwed 1.180e+00 6.513e-01 1.812 0.070070 .

Pthu 8.018e-01 6.184e-01 1.297 0.194795

Pfri 1.064e+00 5.673e-01 1.876 0.060623 .

Psat NA NA NA NA

Bsun 1.114e-01 5.632e-01 0.198 0.843281

Bmon 1.163e+00 6.199e-01 1.877 0.060565 .

Btue 5.220e-01 6.553e-01 0.797 0.425705

Bwed 1.491e+00 6.466e-01 2.305 0.021146 \*

Bthu 4.801e-01 6.091e-01 0.788 0.430632

Bfri 5.957e-01 5.659e-01 1.053 0.292563

Bsat NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.17 on 40900 degrees of freedom

Multiple R-squared: 0.3252, Adjusted R-squared: 0.3244

F-statistic: 410.7 on 48 and 40900 DF, p-value: < 2.2e-16

fitmodel <- step(basemodel)

summary(fitmodel)

Call:

lm(formula = ncH ~ PChek + PTalk + V5 + V6 + V7 + V8 + V9 + V10 +

V11 + V12 + V13 + V16 + V17 + V19 + V20 + V21 + V22 + V23 +

V24 + tcbbt + cl24bbt + c4824bbt + cf24ap + bt + PoShare +

H + Pwed + Pthu + Pfri + Bmon + Bwed, data = FV1)

Residuals:

Min 1Q Median 3Q Max

-347.06 -5.35 -1.15 2.92 1271.60

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.571e+00 1.897e+00 -1.882 0.059808 .

PChek -1.300e-05 7.388e-06 -1.760 0.078432 .

PTalk -1.814e-05 1.825e-06 -9.942 < 2e-16 \*\*\*

V5 -5.223e-01 1.134e-01 -4.607 4.11e-06 \*\*\*

V6 1.445e-02 6.033e-03 2.396 0.016596 \*

V7 9.406e-01 1.181e-01 7.965 1.69e-15 \*\*\*

V8 1.441e-01 5.371e-02 2.683 0.007290 \*\*

V9 -1.553e-01 7.913e-02 -1.963 0.049663 \*

V10 -1.137e-01 2.961e-02 -3.839 0.000124 \*\*\*

V11 -2.530e-03 1.040e-03 -2.432 0.015028 \*

V12 1.527e-01 2.888e-02 5.286 1.26e-07 \*\*\*

V13 1.311e-01 2.310e-02 5.676 1.39e-08 \*\*\*

V16 1.107e-02 1.892e-03 5.850 4.95e-09 \*\*\*

V17 1.279e-01 3.907e-02 3.273 0.001065 \*\*

V19 -1.307e-01 2.524e-02 -5.178 2.25e-07 \*\*\*

V20 4.777e-01 1.269e-01 3.764 0.000168 \*\*\*

V21 -1.596e-02 6.234e-03 -2.560 0.010473 \*

V22 -1.008e+00 1.208e-01 -8.344 < 2e-16 \*\*\*

V23 -1.637e-01 5.505e-02 -2.974 0.002939 \*\*

V24 1.957e-01 8.085e-02 2.421 0.015494 \*

tcbbt 4.440e-02 1.639e-02 2.708 0.006764 \*\*

cl24bbt 2.197e-01 3.725e-03 58.986 < 2e-16 \*\*\*

c4824bbt -2.260e-02 3.510e-03 -6.439 1.21e-10 \*\*\*

cf24ap -6.845e-02 1.825e-02 -3.750 0.000177 \*\*\*

bt -1.927e-01 7.398e-03 -26.041 < 2e-16 \*\*\*

PoShare 2.907e-03 1.764e-04 16.482 < 2e-16 \*\*\*

H 4.081e-01 7.773e-02 5.251 1.52e-07 \*\*\*

Pwed 1.242e+00 4.282e-01 2.900 0.003738 \*\*

Pthu 7.461e-01 4.500e-01 1.658 0.097313 .

Pfri 9.110e-01 4.353e-01 2.093 0.036357 \*

Bmon 8.403e-01 4.504e-01 1.866 0.062073 .

Bwed 1.070e+00 4.326e-01 2.474 0.013366 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.17 on 40917 degrees of freedom

Multiple R-squared: 0.3251, Adjusted R-squared: 0.3246

F-statistic: 635.9 on 31 and 40917 DF, p-value: < 2.2e-16

AIC(basemodel)

[1] 392523.1

AIC(fitmodel)

[1] 392496

predFV1 <- predict(fitmodel,FV1)

msemodel1 <- mean((FV1$ncH - predFV1)^2)

msemodel1

[1] 850.231

colnames(test)[c(1:4)] <- c("PPop","PChek","PTalk","PCat")

colnames(test)[c(30:54)] <- c("tcbbt","cl24bbt","c4824bbt","cf24ap","diff4824","bt","Polen","PoShare","PoPromo","H","Psun","Pmon","Ptue","Pwed","Pthu","Pfri","Psat","Bsun","Bmon","Btue","Bwed","Bthu","Bfri","Bsat","ncH")

colnames(test)

testpred <- predict(fitmodel, test)

msetest <- mean((test$ncH - testpred)^2)

msetest

[1] 11344.59

basemodel

ls(basemodel)

cor(FV1$ncH,predFV1)

[1] 0.5701962

cor(test$ncH,testpred)

[1] 0.3901208

minmaxFV1 <- min\_max\_accuracy(FV1$ncH,predFV1)

minmaxFV1 <- mean(min(FV1$ncH,predFV1)/max(FV1$ncH,predFV1))

minmaxFV1

[1] -0.04702975

minmaxFV1 <- mean(min(test$ncH,testpred)/max(test$ncH,testpred))

minmaxFV1 <- mean(min(FV1$ncH,predFV1)/max(FV1$ncH,predFV1))

minmaxtest <- mean(min(test$ncH,testpred)/max(test$ncH,testpred))

minmaxtest

[1] -0.03619449

mapeFV1 <- mean(abs(predFV1-FV1$ncH)/(FV1$ncH))

mapeFV1

[1] Inf

mapetest <- mean(abs(testpred - test$ncH)/(test$ncH))

mapetest

[1] Inf

FV1$dev <- (predFV1-FV1$ncH)

summary(FV1$dev)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-1271.596 -2.924 1.147 0.000 5.349 347.057

test$dev <- (testpred - test$ncH)

summary(test$dev)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-2387.3224 -5.5675 0.8095 -9.3566 7.0571 604.9548