**ASSIGNMENT NO. 14.1**

**Data Set Information:**

This data originates from blog posts. The raw HTML-documents   
of the blog posts were crawled and processed.   
The prediction task associated with the data is the prediction   
of the number of comments in the upcoming 24 hours. In order   
to simulate this situation, we choose a basetime (in the past)   
and select the blog posts that were published at most   
72 hours before the selected base date/time. Then, we calculate   
all the features of the selected blog posts from the information   
that was available at the basetime, therefore each instance   
corresponds to a blog post. The target is the number of   
comments that the blog post received in the next 24 hours   
relative to the basetime.   
  
In the train data, the basetimes were in the years   
2010 and 2011. In the test data the basetimes were   
in February and March 2012. This simulates the real-world   
situtation in which training data from the past is available   
to predict events in the future.   
  
The train data was generated from different basetimes that may   
temporally overlap. Therefore, if you simply split the train   
into disjoint partitions, the underlying time intervals may   
overlap. Therefore, the you should use the provided, temporally   
disjoint train and test splits in order to ensure that the   
evaluation is fair.

**Attribute Information:**

1...50:   
Average, standard deviation, min, max and median of the   
Attributes 51...60 for the source of the current blog post   
With source we mean the blog on which the post appeared.   
For example, myblog.blog.org would be the source of   
the post myblog.blog.org/post\_2010\_09\_10   
51: Total number of comments before basetime   
52: Number of comments in the last 24 hours before the   
basetime   
53: Let T1 denote the datetime 48 hours before basetime,   
Let T2 denote the datetime 24 hours before basetime.   
This attribute is the number of comments in the time period   
between T1 and T2   
54: Number of comments in the first 24 hours after the   
publication of the blog post, but before basetime   
55: The difference of Attribute 52 and Attribute 53   
56...60:   
The same features as the attributes 51...55, but   
features 56...60 refer to the number of links (trackbacks),   
while features 51...55 refer to the number of comments.   
61: The length of time between the publication of the blog post   
and basetime   
62: The length of the blog post   
63...262:   
The 200 bag of words features for 200 frequent words of the   
text of the blog post   
263...269: binary indicator features (0 or 1) for the weekday   
(Monday...Sunday) of the basetime   
270...276: binary indicator features (0 or 1) for the weekday   
(Monday...Sunday) of the date of publication of the blog   
post   
277: Number of parent pages: we consider a blog post P as a   
parent of blog post B, if B is a reply (trackback) to   
blog post P.   
278...280:   
Minimum, maximum, average number of comments that the   
parents received   
281: The target: the number of comments in the next 24 hours   
(relative to basetime)

1. a. Read the dataset and identify the right features

> blogData\_train <- read.csv("E:/kamagyana/Computing/DARET/Assignments/blogData\_train.csv", header=FALSE, stringsAsFactors=FALSE)

> View(blogData\_train)

head(blogData\_train)

# Each of the column is mapped with the respective description of the column to enter into further analysis as a feature.

colnames(blogData\_train)[c(51,52,53,54,55,56,57,58,59,60)] <- c("tcbbt","l24cbbt","48-24cbbt","cf24apbbt","cdif4824","tlbbt","l24lbbt","48-24lbbt","lf24apbbt","ldif4824")

colnames(blogData\_train)

colnames(blogData\_train)[c(61,62)] <- c("bt-pt","lpt")

colnames(blogData\_train)[c(277,278,279,280,281)] <- c("npp","mincnpp","maxcnpp","avcnpp","cnextbbt")

colnames(blogData\_train)

name <- c("mean\_","sd\_","min\_","max\_","med\_")

colnames(blogData\_train)[c(1,2,3,4,5)] <- paste(name,colnames(blogData\_train)[51])

colnames(blogData\_train[c(1,2,3,4,5)])

colnames(blogData\_train)[c(6,7,8,9,10)] <- paste(name,colnames(blogData\_train)[52])

colnames(blogData\_train)[c(11,12,13,14,15)] <- paste(name,colnames(blogData\_train)[53])

colnames(blogData\_train)[c(16,17,18,19,20)] <- paste(name,colnames(blogData\_train)[54])

colnames(blogData\_train)[c(21,22,23,24,25)] <- paste(name,colnames(blogData\_train)[55])

colnames(blogData\_train)[c(26,27,28,29,30)] <- paste(name,colnames(blogData\_train)[56])

colnames(blogData\_train)[c(31,32,33,34,35)] <- paste(name,colnames(blogData\_train)[57])

colnames(blogData\_train)[c(36,37,38,39,40)] <- paste(name,colnames(blogData\_train)[58])

colnames(blogData\_train)[c(41,42,43,44,45)] <- paste(name,colnames(blogData\_train)[59])

colnames(blogData\_train)[c(46,47,48,49,50)] <- paste(name,colnames(blogData\_train)[60])

colnames(blogData\_train)

head(blogData\_train)sum(is.na(blogData\_train))

colnames(blogData\_train)[c(263,264,265,266,267,268,269)] <- c("Monbt","Tuebt","Wedbt","Thubt","Fribt",'Satbt',"Sunbt")

colnames(blogData\_train)[c(270,271,272,273,274,275,276)] <- c("Monpt","Tuept","Wedpt","Thupt","Fript","Satpt","Sunpt")

colnames(blogData\_train)

colnames(blogData\_train)[c(63:262)] <- c(paste0("Word",c(seq(1:200))))

colnames(blogData\_train)

regdata <- cbind(blogData\_train[c(1,6,11,16,21,26,31,36,41,46,c(51:62),277,280,281)])

regword <- cbind(blogData\_train[c(63:262,281)])

regday <- cbind(blogData\_train[c(263:276,281)])

cor(regdata)

1. b. Clean dataset, impute missing values and perform exploratory data analysis.

> sum(is.na(blogData\_train))

[1] 0

NOTE:This shows that there is no missing data in the training dataset

> summary(regdata)

mean\_ tcbbt mean\_ l24cbbt mean\_ 48-24cbbt mean\_ cf24apbbt

Min. : 0.000 Min. : 0.0000 Min. : 0.000 Min. : 0.000

1st Qu.: 2.286 1st Qu.: 0.8916 1st Qu.: 0.775 1st Qu.: 1.825

Median : 10.631 Median : 4.1507 Median : 3.817 Median : 9.777

Mean : 39.444 Mean : 15.2146 Mean : 14.053 Mean : 34.898

3rd Qu.: 40.305 3rd Qu.: 15.9986 3rd Qu.: 14.641 3rd Qu.: 35.831

Max. :1122.667 Max. :442.6667 Max. :438.000 Max. :1102.000

mean\_ cdif4824 mean\_ tlbbt mean\_ l24lbbt mean\_ 48-24lbbt

Min. :-0.66667 Min. :0.0000 Min. :0.00000 Min. :0.00000

1st Qu.: 0.05797 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000

Median : 0.22381 Median :0.1087 Median :0.04317 Median :0.03763

Mean : 1.16150 Mean :0.5013 Mean :0.19161 Mean :0.17990

3rd Qu.: 0.94595 3rd Qu.:0.4985 3rd Qu.:0.20382 3rd Qu.:0.19385

Max. :30.68948 Max. :8.9527 Max. :3.15385 Max. :3.00000

mean\_ lf24apbbt mean\_ ldif4824 tcbbt l24cbbt

Min. :0.00000 Min. :-0.05556 Min. : 0.00 Min. : 0.00

1st Qu.:0.00000 1st Qu.: 0.00000 1st Qu.: 0.00 1st Qu.: 0.00

Median :0.09264 Median : 0.00000 Median : 3.00 Median : 0.00

Mean :0.45718 Mean : 0.01172 Mean : 39.44 Mean : 15.21

3rd Qu.:0.43590 3rd Qu.: 0.01429 3rd Qu.: 25.00 3rd Qu.: 4.00

Max. :8.49112 Max. : 0.66667 Max. :2044.00 Max. :1424.00

48-24cbbt cf24apbbt cdif4824 tlbbt

Min. : 0.00 Min. : 0.0 Min. :-1256.000 Min. : 0.0000

1st Qu.: 0.00 1st Qu.: 0.0 1st Qu.: -1.000 1st Qu.: 0.0000

Median : 0.00 Median : 2.0 Median : 0.000 Median : 0.0000

Mean : 14.05 Mean : 34.9 Mean : 1.161 Mean : 0.5013

3rd Qu.: 3.00 3rd Qu.: 21.0 3rd Qu.: 1.000 3rd Qu.: 0.0000

Max. :1424.00 Max. :1932.0 Max. : 1422.000 Max. :30.0000

l24lbbt 48-24lbbt lf24apbbt ldif4824

Min. : 0.0000 Min. : 0.0000 Min. : 0.0000 Min. :-20.00000

1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.00000

Median : 0.0000 Median : 0.0000 Median : 0.0000 Median : 0.00000

Mean : 0.1916 Mean : 0.1799 Mean : 0.4572 Mean : 0.01172

3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.00000

Max. :24.0000 Max. :24.0000 Max. :26.0000 Max. : 23.00000

bt-pt lpt npp avcnpp

Min. : 0.00 Min. : 0 Min. : 0.0000 Min. : 0.0000

1st Qu.:14.00 1st Qu.: 93 1st Qu.: 0.0000 1st Qu.: 0.0000

Median :35.00 Median : 1859 Median : 0.0000 Median : 0.0000

Mean :34.78 Mean : 2850 Mean : 0.1192 Mean : 0.7695

3rd Qu.:55.00 3rd Qu.: 3959 3rd Qu.: 0.0000 3rd Qu.: 0.0000

Max. :72.00 Max. :57894 Max. :136.0000 Max. :1778.0000

cnextbbt

Min. : 0.000

1st Qu.: 0.000

Median : 0.000

Mean : 6.765

3rd Qu.: 1.000

Max. :1424.000

1. c. Visualize the dataset and make inferences from that
2. > plot(regdata$tcbbt,regdata$cnextbbt)
3. > plot(regdata$cnextbbt,regdata$tcbbt)
4. > plot(regdata$cnextbbt,regdata$l24cbbt)
5. > plot(regdata$cnextbbt,regdata$`48-24cbbt`)
6. > plot(x=regdata$`48-24cbbt`,y=regdata$cnextbbt)
7. > plot(x=regdata$tcbbt,y=regdata$cnextbbt)
8. > plot(x=regdata$l24cbbt,y=regdata$cnextbbt)
9. > plot(x=regdata$cf24apbbt,y=regdata$cnextbbt)
10. > plot(x=regdata$tlbbt,y=regdata$cnextbbt)
11. > plot(x=regdata$l24lbbt,y=regdata$cnextbbt)
12. > plot(x=regdata$`48-24lbbt`,y=regdata$cnextbbt)
13. > plot(x=regdata$lf24apbbt,y=regdata$cnextbbt)
14. > plot(x=regdata$ldif4824,y=regdata$cnextbbt)
15. > plot(x=regdata$`bt-pt`,y=regdata$cnextbbt)
16. > plot(x=regdata$lpt,y=regdata$cnextbbt)
17. > plot(x=regdata$npp,y=regdata$cnextbbt)
18. > plot(x=regdata$avcnpp,y=regdata$cnextbbt)
19. d. Perform any 3 hypothesis tests using columns of your choice, make conclusions

NOTE: To test the hypothesis that the number of comments in the upcoming next 24 hours relative to the base time are dependent on the weekday of the base chosen, we use the anova test treating the number of comments as the dependent variable and the weekday as the independent variable. Since there is no column which captures the week day directly, we create separate columns for weekday through mutate command and then undertake the anova test.

> regday <- mutate(regday, daybt = ifelse(Monbt==1,"Monday",ifelse(Tuebt==1,"Tuesday",ifelse(Wedbt==1,"Wednesday",ifelse(Thubt==1,"Thursday",ifelse(Fribt==1,"Friday",ifelse(Satbt==1,"Saturday",ifelse(Sunbt==1,"Sunday","NA"))))))))

> regday <- mutate(regday, daypt = ifelse(Monpt==1,"Monday",ifelse(Tuept==1,"Tuesday",ifelse(Wedpt==1,"Wednesday",ifelse(Thupt==1,"Thursday",ifelse(Fript==1,"Friday",ifelse(Satpt==1,"Saturday",ifelse(Sunpt==1,"Sunday","NA"))))))))

> afit1 <-aov(regday$cnextbbt~regday$daybt)

> summary(afit1)

1. Df Sum Sq Mean Sq F value Pr(>F)
2. regday$daybt 6 44586 7431 5.229 2.16e-05 \*\*\*
3. Residuals 52390 74451262 1421
4. ---
5. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

NOTE: The test clearly shows that the comments in the ensuing 24 hours after the base time are statistically significantly dependent on the weekday of the base time chosen. The Tukey’s HSD results show the specific days which are statistically significant over another chosen day is presented below. It shows clearly that if the base time is Tuesday or Wednesday, then the comments on the next 24 hours would be higher than any other day chosen, and more so when the day is the weekend, i.e. Friday, Saturday, or Sunday.

> TukeyHSD(afit1)

1. Tukey multiple comparisons of means
2. 95% family-wise confidence level
3. Fit: aov(formula = regday$cnextbbt ~ regday$daybt)
4. $`regday$daybt`
5. diff lwr upr p adj
6. Tuesday-Friday 2.66023055 0.80520706 4.515254 0.0004705
7. Wednesday-Friday 1.88405894 0.11982850 3.648289 0.0273829
8. Tuesday-Monday 2.35751233 0.33136037 4.383664 0.0107867
9. Tuesday-Saturday 2.58344290 0.71416316 4.452723 0.0009075
10. Wednesday-Saturday 1.80727130 0.02805695 3.586486 0.0436745
11. Tuesday-Sunday 2.54329790 0.60661378 4.479982 0.0020810
12. Wednesday-Sunday 1.76712629 -0.08277702 3.617030 0.0722825

> afit2 <- aov(regday$cnextbbt~regday$daypt)

> summary(afit2)

1. Df Sum Sq Mean Sq F value Pr(>F)
2. regday$daypt 6 10345 1724 1.213 0.296
3. Residuals 52390 74485503 1422

NOTE: When the same anova test is performed on the comments on the next 24 hours of the base time, the day of the week of the post, did not show any significant influence. Hence the weekday of the posting in the blog would not influence the number of comments in the upcoming 24 hours relative to base time.

NOTE: Since the weekday of the post and the comments in the ensuing 24 hours of the base time did not show any kind of relationship, we tested for whether the comments in the first 24 hours of the post, but before the base time show any significant relationship, in the sense the number of comments is significantly influenced by the weekday of the post.

> afit3 <- aov(regdata$`mean\_ cf24apbbt`~regday$daypt)

> summary(afit3)

Df Sum Sq Mean Sq F value Pr(>F)

regday$daypt 6 561836 93639 20.95 <2e-16 \*\*\*

Residuals 52390 234109242 4469

---

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

NOTE: The results clearly show a strong influence of the weekday on the number of comments in the first 24 hours after the blog is posted and before the base time. The TukeyHSD results again show that blogs that are posted on Tuesdays, Wednesdays, and Thursdays always receive more comments than the Fridays, Saturdays, and Sundays.

> TukeyHSD(afit3)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = regdata$`mean\_ cf24apbbt` ~ regday$daypt)

$`regday$daypt`

diff lwr upr p adj

Saturday-Friday -7.16328722 -10.700203 -3.626371 0.0000000

Sunday-Friday -7.54150917 -11.166006 -3.917012 0.0000000

Saturday-Monday -8.82985287 -12.346379 -5.313327 0.0000000

Sunday-Monday -9.20807482 -12.812677 -5.603473 0.0000000

Thursday-Saturday 7.57510264 4.070902 11.079303 0.0000000

Tuesday-Saturday 9.11707842 5.634091 12.600066 0.0000000

Wednesday-Saturday 7.54254271 4.073085 11.012000 0.0000000

Thursday-Sunday 7.95332459 4.360746 11.545903 0.0000000

Tuesday-Sunday 9.49530038 5.923410 13.067191 0.0000000

Wednesday-Sunday 7.92076467 4.362066 11.479463 0.0000000

NOTE: To test the hypothesis whether the mean of total comments before base time in the given training data is greater than 40, we use the following functions. The results and the p value clearly shows that the mean total comments before the base time are not greater than 40 at 5% significance level, rather only they are significant at 12.6% level.

> z <- (40-(mean(regdata$tcbbt)))/((sd(regdata$tcbbt))/(sqrt(length(regdata$tcbbt))))

> pnorm(z,lower.tail=FALSE)

1. [1] 0.1260308

e. Create a linear regression model to predict the number of comments in the next 24 hours (relative to basetime)

> model1 <- lm(cnextbbt~.,data=regdata)

> summary(model1)

Call:

lm(formula = cnextbbt ~ ., data = regdata)

Residuals:

Min 1Q Median 3Q Max

-275.39 -4.70 -1.04 2.68 1341.18

Coefficients: (2 not defined because of singularities)

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

(Intercept) 4.429e+00 3.063e-01 14.460 < 2e-16 \*\*\*

`mean\_ tcbbt` -2.239e-01 8.428e-02 -2.656 0.00791 \*\*

`mean\_ l24cbbt` -9.402e+05 1.404e+05 -6.695 2.17e-11 \*\*\*

`mean\_ 48-24cbbt` 9.402e+05 1.404e+05 6.695 2.17e-11 \*\*\*

`mean\_ cf24apbbt` -3.184e-01 6.804e-02 -4.679 2.88e-06 \*\*\*

`mean\_ cdif4824` 9.402e+05 1.404e+05 6.695 2.17e-11 \*\*\*

`mean\_ tlbbt` -1.297e+01 6.083e+00 -2.132 0.03305 \*

`mean\_ l24lbbt` -3.355e+07 1.049e+07 -3.198 0.00139 \*\*

`mean\_ 48-24lbbt` 3.355e+07 1.049e+07 3.198 0.00139 \*\*

`mean\_ lf24apbbt` 5.010e+00 4.612e+00 1.086 0.27735

`mean\_ ldif4824` 3.355e+07 1.049e+07 3.198 0.00139 \*\*

tcbbt -3.375e-02 6.820e-03 -4.948 7.50e-07 \*\*\*

l24cbbt 2.168e-01 4.333e-03 50.044 < 2e-16 \*\*\*

`48-24cbbt` -4.203e-02 4.820e-03 -8.722 < 2e-16 \*\*\*

cf24apbbt -3.546e-02 6.989e-03 -5.073 3.93e-07 \*\*\*

cdif4824 NA NA NA NA

tlbbt -1.010e+00 5.535e-01 -1.826 0.06793 .

l24lbbt 4.984e-01 2.763e-01 1.804 0.07123 .

`48-24lbbt` 3.508e-01 2.897e-01 1.211 0.22583

lf24apbbt 9.531e-01 5.422e-01 1.758 0.07879 .

ldif4824 NA NA NA NA

`bt-pt` -1.494e-01 6.803e-03 -21.953 < 2e-16 \*\*\*

lpt 2.135e-04 3.520e-05 6.064 1.34e-09 \*\*\*

npp 4.378e-02 9.201e-02 0.476 0.63416

avcnpp 2.721e-03 6.499e-03 0.419 0.67541

---

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

Residual standard error: 30.19 on 52374 degrees of freedom

Multiple R-squared: 0.3591, Adjusted R-squared: 0.3589

F-statistic: 1334 on 22 and 52374 DF, p-value: < 2.2e-16

> AIC(model1)

[1] 505815.3