> library(devtools)

> library(sentiment)

> library(twitteR)

> library(plyr)

> library(ggplot2)

> library(wordcloud)

> library(RColorBrewer)

> library(devtools)

> library(stringr)

> getwd()

[1] "C:/Users/Alok Satpathy/Desktop/Fall 2016/EDA/Final Project"

> setwd("C:/Users/Alok Satpathy/Desktop/Fall 2016/EDA/Final Project")

> Q3csv=read.csv("Question3Dataset.csv")

> attach(Q3csv)

> str(Q3csv)

'data.frame': 50000 obs. of 32 variables:

$ CommentId : int 68362950 64796201 64796476 63094214 63092337 63093585 66041688 66041726 66286341 62891724 ...

$ CommentScore : int 0 0 0 0 0 0 1 0 0 0 ...

$ CommentText : Factor w/ 49819 levels "'$(this).prop('checked', 'checked');' try to add this. it will work perfectly",..: 38483 40932 19118 37842 32232 41297 41340 2442 23023 11801 ...

$ CommentCrDays : int 28 131 131 179 179 179 95 95 87 185 ...

$ CommentLength : int 182 45 90 31 69 66 176 115 16 526 ...

$ CommentCrDt : Factor w/ 49907 levels "2016-01-25 18:53:34",..: 47069 31873 31878 24652 24646 24650 37492 37493 38477 23842 ...

$ CommentUserId : int 5922757 4769409 4769409 26095 2684539 3807729 845568 1317944 4363659 256196 ...

$ comment\_owner\_reputation : int 1070 110 110 35552 66919 21688 15210 3649 36 221242 ...

$ comment\_owner\_profile\_summary : int 2 0 0 2 3 0 2 0 0 3 ...

$ comment\_owner\_views : int 95 14 14 1661 2560 1051 4567 731 19 26642 ...

$ comment\_owner\_upvotes : int 346 1 1 2811 780 3943 3752 1093 5 6857 ...

$ comment\_owner\_downvotes : int 30 0 0 66 574 877 6245 21 0 1760 ...

$ comment\_owner\_lastactivity\_days: int 7 7 7 7 7 6 7 7 7 6 ...

$ PostId : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ editDurationAfterCreation : int NA 0 0 0 0 0 0 0 NA 0 ...

$ activityDurationAfterCreation : int 0 0 0 0 0 0 0 0 0 0 ...

$ title\_length : int NA NA NA NA NA NA NA NA NA NA ...

$ num\_tags : int NA NA NA NA NA NA NA NA NA NA ...

$ PostAnswerCount : int NA NA NA NA NA NA NA NA NA NA ...

$ num\_favorite : int NA NA NA NA NA NA NA NA NA NA ...

$ hascode : int 1 1 1 1 1 1 1 1 1 1 ...

$ post\_views : int NA NA NA NA NA NA NA NA NA NA ...

$ postTypeId : int 2 2 2 2 2 2 2 2 2 2 ...

$ Id : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ AcceptedAnswerId : int NA NA NA NA NA NA NA NA NA NA ...

$ IsAcceptedAnswer : int 0 0 0 0 0 0 0 0 0 1 ...

$ postScore : int 1 1 1 1 1 1 1 1 3 2 ...

$ post\_length : int 1333 2274 2274 846 846 846 147 147 853 1242 ...

$ PostCommentCount : int 1 2 2 3 3 3 2 2 1 5 ...

$ PostCrDt : Factor w/ 15208 levels "2016-01-25 18:51:52",..: 14377 10095 10095 7861 7861 7861 11703 11703 11993 7639 ...

$ PostUpVotes : int 1 1 1 1 1 1 1 1 3 2 ...

$ PostDownVotes : int NA NA NA NA NA NA NA NA NA NA ...

> Q3csv[is.na(Q3csv)]=0

> str(Q3csv)

'data.frame': 50000 obs. of 32 variables:

$ CommentId : int 68362950 64796201 64796476 63094214 63092337 63093585 66041688 66041726 66286341 62891724 ...

$ CommentScore : int 0 0 0 0 0 0 1 0 0 0 ...

$ CommentText : Factor w/ 49819 levels "'$(this).prop('checked', 'checked');' try to add this. it will work perfectly",..: 38483 40932 19118 37842 32232 41297 41340 2442 23023 11801 ...

$ CommentCrDays : int 28 131 131 179 179 179 95 95 87 185 ...

$ CommentLength : int 182 45 90 31 69 66 176 115 16 526 ...

$ CommentCrDt : Factor w/ 49907 levels "2016-01-25 18:53:34",..: 47069 31873 31878 24652 24646 24650 37492 37493 38477 23842 ...

$ CommentUserId : int 5922757 4769409 4769409 26095 2684539 3807729 845568 1317944 4363659 256196 ...

$ comment\_owner\_reputation : int 1070 110 110 35552 66919 21688 15210 3649 36 221242 ...

$ comment\_owner\_profile\_summary : int 2 0 0 2 3 0 2 0 0 3 ...

$ comment\_owner\_views : int 95 14 14 1661 2560 1051 4567 731 19 26642 ...

$ comment\_owner\_upvotes : int 346 1 1 2811 780 3943 3752 1093 5 6857 ...

$ comment\_owner\_downvotes : int 30 0 0 66 574 877 6245 21 0 1760 ...

$ comment\_owner\_lastactivity\_days: int 7 7 7 7 7 6 7 7 7 6 ...

$ PostId : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ editDurationAfterCreation : num 0 0 0 0 0 0 0 0 0 0 ...

$ activityDurationAfterCreation : int 0 0 0 0 0 0 0 0 0 0 ...

$ title\_length : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_tags : num 0 0 0 0 0 0 0 0 0 0 ...

$ PostAnswerCount : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_favorite : num 0 0 0 0 0 0 0 0 0 0 ...

$ hascode : int 1 1 1 1 1 1 1 1 1 1 ...

$ post\_views : num 0 0 0 0 0 0 0 0 0 0 ...

$ postTypeId : int 2 2 2 2 2 2 2 2 2 2 ...

$ Id : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ AcceptedAnswerId : num 0 0 0 0 0 0 0 0 0 0 ...

$ IsAcceptedAnswer : int 0 0 0 0 0 0 0 0 0 1 ...

$ postScore : int 1 1 1 1 1 1 1 1 3 2 ...

$ post\_length : int 1333 2274 2274 846 846 846 147 147 853 1242 ...

$ PostCommentCount : int 1 2 2 3 3 3 2 2 1 5 ...

$ PostCrDt : Factor w/ 15208 levels "2016-01-25 18:51:52",..: 14377 10095 10095 7861 7861 7861 11703 11703 11993 7639 ...

$ PostUpVotes : int 1 1 1 1 1 1 1 1 3 2 ...

$ PostDownVotes : num 0 0 0 0 0 0 0 0 0 0 ...

> Q3csv["SentimentScore"]=NA

> str(Q3csv)

'data.frame': 50000 obs. of 33 variables:

$ CommentId : int 68362950 64796201 64796476 63094214 63092337 63093585 66041688 66041726 66286341 62891724 ...

$ CommentScore : int 0 0 0 0 0 0 1 0 0 0 ...

$ CommentText : Factor w/ 49819 levels "'$(this).prop('checked', 'checked');' try to add this. it will work perfectly",..: 38483 40932 19118 37842 32232 41297 41340 2442 23023 11801 ...

$ CommentCrDays : int 28 131 131 179 179 179 95 95 87 185 ...

$ CommentLength : int 182 45 90 31 69 66 176 115 16 526 ...

$ CommentCrDt : Factor w/ 49907 levels "2016-01-25 18:53:34",..: 47069 31873 31878 24652 24646 24650 37492 37493 38477 23842 ...

$ CommentUserId : int 5922757 4769409 4769409 26095 2684539 3807729 845568 1317944 4363659 256196 ...

$ comment\_owner\_reputation : int 1070 110 110 35552 66919 21688 15210 3649 36 221242 ...

$ comment\_owner\_profile\_summary : int 2 0 0 2 3 0 2 0 0 3 ...

$ comment\_owner\_views : int 95 14 14 1661 2560 1051 4567 731 19 26642 ...

$ comment\_owner\_upvotes : int 346 1 1 2811 780 3943 3752 1093 5 6857 ...

$ comment\_owner\_downvotes : int 30 0 0 66 574 877 6245 21 0 1760 ...

$ comment\_owner\_lastactivity\_days: int 7 7 7 7 7 6 7 7 7 6 ...

$ PostId : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ editDurationAfterCreation : num 0 0 0 0 0 0 0 0 0 0 ...

$ activityDurationAfterCreation : int 0 0 0 0 0 0 0 0 0 0 ...

$ title\_length : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_tags : num 0 0 0 0 0 0 0 0 0 0 ...

$ PostAnswerCount : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_favorite : num 0 0 0 0 0 0 0 0 0 0 ...

$ hascode : int 1 1 1 1 1 1 1 1 1 1 ...

$ post\_views : num 0 0 0 0 0 0 0 0 0 0 ...

$ postTypeId : int 2 2 2 2 2 2 2 2 2 2 ...

$ Id : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ AcceptedAnswerId : num 0 0 0 0 0 0 0 0 0 0 ...

$ IsAcceptedAnswer : int 0 0 0 0 0 0 0 0 0 1 ...

$ postScore : int 1 1 1 1 1 1 1 1 3 2 ...

$ post\_length : int 1333 2274 2274 846 846 846 147 147 853 1242 ...

$ PostCommentCount : int 1 2 2 3 3 3 2 2 1 5 ...

$ PostCrDt : Factor w/ 15208 levels "2016-01-25 18:51:52",..: 14377 10095 10095 7861 7861 7861 11703 11703 11993 7639 ...

$ PostUpVotes : int 1 1 1 1 1 1 1 1 3 2 ...

$ PostDownVotes : num 0 0 0 0 0 0 0 0 0 0 ...

$ SentimentScore : logi NA NA NA NA NA NA ...

#function to clean data

> cleancomments = function(commentstest)

+ {

+ commentstest\_cl = gsub("(RT|via)((?:\\b\\W\*@\\w+)+)","",commentstest)

+ commentstest\_cl = gsub("http[^[:blank:]]+", "", commentstest\_cl)

+ commentstest\_cl = gsub("@\\w+", "", commentstest\_cl)

+ commentstest\_cl = gsub("[ \t]{2,}", "", commentstest\_cl)

+ commentstest\_cl = gsub("^\\s+|\\s+$", "", commentstest\_cl)

+ commentstest\_cl = gsub("[[:punct:]]", " ", commentstest\_cl)

+ commentstest\_cl = gsub("[^[:alnum:]]", " ", commentstest\_cl)

+ commentstest\_cl <- gsub('\\d+', '', commentstest\_cl)

+ return(commentstest\_cl)

+ }

#function to calculate number of words in each category within a sentence

> sentimentScore <- function(sentences, vNegTerms, negTerms, posTerms, vPosTerms){

+ final\_scores <- matrix('', 0, 5)

+ scores <- lapply(sentences, function(sentence, vNegTerms, negTerms, posTerms, vPosTerms){

+ initial\_sentence <- sentence

+ #remove unnecessary characters and split up by word

+ sentence = cleancomments(sentence)

+ sentence <- tolower(sentence)

+ wordList <- strsplit(sentence, '\\s+')

+ words <- unlist(wordList)

+ #build vector with matches between sentence and each category

+ vPosMatches <- match(words, vPosTerms)

+ posMatches <- match(words, posTerms)

+ vNegMatches <- match(words, vNegTerms)

+ negMatches <- match(words, negTerms)

+ #sum up number of words in each category

+ vPosMatches <- sum(!is.na(vPosMatches))

+ posMatches <- sum(!is.na(posMatches))

+ vNegMatches <- sum(!is.na(vNegMatches))

+ negMatches <- sum(!is.na(negMatches))

+ score <- c(vNegMatches, negMatches, posMatches, vPosMatches)

+ #add row to scores table

+ newrow <- c(initial\_sentence, score)

+ final\_scores <- rbind(final\_scores, newrow)

+ return(final\_scores)

+ }, vNegTerms, negTerms, posTerms, vPosTerms)

+ return(scores)

+ }

#load pos,neg statements

> afinn\_list <- read.delim(file='AFINN-111.txt', header=FALSE, stringsAsFactors=FALSE)

> names(afinn\_list) <- c('word', 'score')

> summary(afinn\_list$score)

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

-5.0000 -2.0000 -2.0000 -0.5894 2.0000 5.0000

> afinn\_list$word <- tolower(afinn\_list$word)

#categorize words as very negative to very positive and add some movie-specific words

> vNegTerms <- afinn\_list$word[afinn\_list$score==-5 | afinn\_list$score==-4]

> negTerms <- afinn\_list$word[afinn\_list$score==-2 | afinn\_list$score==-1 | afinn\_list$score==-3]

> posTerms <- afinn\_list$word[afinn\_list$score==3 | afinn\_list$score==2 | afinn\_list$score==1]

> vPosTerms <- afinn\_list$word[afinn\_list$score==5 | afinn\_list$score==4]

> str(Q3csv)

'data.frame': 50000 obs. of 33 variables:

$ CommentId : int 68362950 64796201 64796476 63094214 63092337 63093585 66041688 66041726 66286341 62891724 ...

$ CommentScore : int 0 0 0 0 0 0 1 0 0 0 ...

$ CommentText : Factor w/ 49819 levels "'$(this).prop('checked', 'checked');' try to add this. it will work perfectly",..: 38483 40932 19118 37842 32232 41297 41340 2442 23023 11801 ...

$ CommentCrDays : int 28 131 131 179 179 179 95 95 87 185 ...

$ CommentLength : int 182 45 90 31 69 66 176 115 16 526 ...

$ CommentCrDt : Factor w/ 49907 levels "2016-01-25 18:53:34",..: 47069 31873 31878 24652 24646 24650 37492 37493 38477 23842 ...

$ CommentUserId : int 5922757 4769409 4769409 26095 2684539 3807729 845568 1317944 4363659 256196 ...

$ comment\_owner\_reputation : int 1070 110 110 35552 66919 21688 15210 3649 36 221242 ...

$ comment\_owner\_profile\_summary : int 2 0 0 2 3 0 2 0 0 3 ...

$ comment\_owner\_views : int 95 14 14 1661 2560 1051 4567 731 19 26642 ...

$ comment\_owner\_upvotes : int 346 1 1 2811 780 3943 3752 1093 5 6857 ...

$ comment\_owner\_downvotes : int 30 0 0 66 574 877 6245 21 0 1760 ...

$ comment\_owner\_lastactivity\_days: int 7 7 7 7 7 6 7 7 7 6 ...

$ PostId : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ editDurationAfterCreation : num 0 0 0 0 0 0 0 0 0 0 ...

$ activityDurationAfterCreation : int 0 0 0 0 0 0 0 0 0 0 ...

$ title\_length : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_tags : num 0 0 0 0 0 0 0 0 0 0 ...

$ PostAnswerCount : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_favorite : num 0 0 0 0 0 0 0 0 0 0 ...

$ hascode : int 1 1 1 1 1 1 1 1 1 1 ...

$ post\_views : num 0 0 0 0 0 0 0 0 0 0 ...

$ postTypeId : int 2 2 2 2 2 2 2 2 2 2 ...

$ Id : int 40559915 38709311 38709311 37813455 37813455 37813455 39355415 39355415 39483899 37708274 ...

$ AcceptedAnswerId : num 0 0 0 0 0 0 0 0 0 0 ...

$ IsAcceptedAnswer : int 0 0 0 0 0 0 0 0 0 1 ...

$ postScore : int 1 1 1 1 1 1 1 1 3 2 ...

$ post\_length : int 1333 2274 2274 846 846 846 147 147 853 1242 ...

$ PostCommentCount : int 1 2 2 3 3 3 2 2 1 5 ...

$ PostCrDt : Factor w/ 15208 levels "2016-01-25 18:51:52",..: 14377 10095 10095 7861 7861 7861 11703 11703 11993 7639 ...

$ PostUpVotes : int 1 1 1 1 1 1 1 1 3 2 ...

$ PostDownVotes : num 0 0 0 0 0 0 0 0 0 0 ...

$ SentimentScore : logi NA NA NA NA NA NA ...

> for(i in 1:nrow(Q3csv))

+ {

+ rowSenti=as.data.frame(sentimentScore(Q3csv[i,"CommentText"], vNegTerms, negTerms, posTerms, vPosTerms))

+ #print(rowSenti)

+ totalSenti=((as.numeric(as.character(rowSenti[,2])))\*(-2)) + ((as.numeric(as.character(rowSenti[,3])))\*(-1)) + ((as.numeric(as.character(rowSenti[,4])))\*(1)) + ((as.numeric(as.character(rowSenti[,5])))\*(2))

+ #print(totalSenti)

+

+ if(totalSenti>0)

+ {

+ totalSenti=1

+ }

+ else if(totalSenti<0)

+ {

+ totalSenti=-1

+ }

+

+ Q3csv[i,"SentimentScore"]=totalSenti

+

+ #Uncomment to calculate polarity -- very slow process

+ #polar=sentiment(Q3csv[i,"CommentText"])

+ #Q3csv[i,"SentimentPolarity"]=polar[1,"polarity"]

+

+ if(i%%1000 == 0)

+ {

+ cat(i, " Sentiments Analyzed\n")

+ }

+ }

1000 Sentiments Analyzed

…..

50000 Sentiments Analyzed

> attach(Q3csv)

> Q3df=data.frame(CommentScore, CommentCrDays, CommentLength, comment\_owner\_reputation, comment\_owner\_profile\_summary, comment\_owner\_views, comment\_owner\_upvotes, comment\_owner\_downvotes, comment\_owner\_lastactivity\_days, editDurationAfterCreation, activityDurationAfterCreation, title\_length, num\_tags, PostAnswerCount, num\_favorite, hascode, post\_views, postTypeId, IsAcceptedAnswer, postScore, post\_length, PostCommentCount, PostUpVotes, PostDownVotes, SentimentScore)

> attach(Q3csv)

> Q3df=data.frame(CommentScore, CommentCrDays, CommentLength, comment\_owner\_reputation, comment\_owner\_profile\_summary, comment\_owner\_views, comment\_owner\_upvotes, comment\_owner\_downvotes, comment\_owner\_lastactivity\_days, editDurationAfterCreation, activityDurationAfterCreation, title\_length, num\_tags, PostAnswerCount, num\_favorite, hascode, post\_views, postTypeId, IsAcceptedAnswer, postScore, post\_length, PostCommentCount, PostUpVotes, PostDownVotes, SentimentScore)

> Q3df$SentimentScore <- as.factor(Q3df$SentimentScore)

> str(Q3df)

'data.frame': 50000 obs. of 25 variables:

$ CommentScore : int 0 0 0 0 0 0 1 0 0 0 ...

$ CommentCrDays : int 28 131 131 179 179 179 95 95 87 185 ...

$ CommentLength : int 182 45 90 31 69 66 176 115 16 526 ...

$ comment\_owner\_reputation : int 1070 110 110 35552 66919 21688 15210 3649 36 221242 ...

$ comment\_owner\_profile\_summary : int 2 0 0 2 3 0 2 0 0 3 ...

$ comment\_owner\_views : int 95 14 14 1661 2560 1051 4567 731 19 26642 ...

$ comment\_owner\_upvotes : int 346 1 1 2811 780 3943 3752 1093 5 6857 ...

$ comment\_owner\_downvotes : int 30 0 0 66 574 877 6245 21 0 1760 ...

$ comment\_owner\_lastactivity\_days: int 7 7 7 7 7 6 7 7 7 6 ...

$ editDurationAfterCreation : num 0 0 0 0 0 0 0 0 0 0 ...

$ activityDurationAfterCreation : int 0 0 0 0 0 0 0 0 0 0 ...

$ title\_length : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_tags : num 0 0 0 0 0 0 0 0 0 0 ...

$ PostAnswerCount : num 0 0 0 0 0 0 0 0 0 0 ...

$ num\_favorite : num 0 0 0 0 0 0 0 0 0 0 ...

$ hascode : int 1 1 1 1 1 1 1 1 1 1 ...

$ post\_views : num 0 0 0 0 0 0 0 0 0 0 ...

$ postTypeId : int 2 2 2 2 2 2 2 2 2 2 ...

$ IsAcceptedAnswer : int 0 0 0 0 0 0 0 0 0 1 ...

$ postScore : int 1 1 1 1 1 1 1 1 3 2 ...

$ post\_length : int 1333 2274 2274 846 846 846 147 147 853 1242 ...

$ PostCommentCount : int 1 2 2 3 3 3 2 2 1 5 ...

$ PostUpVotes : int 1 1 1 1 1 1 1 1 3 2 ...

$ PostDownVotes : num 0 0 0 0 0 0 0 0 0 0 ...

$ SentimentScore : Factor w/ 3 levels "-1","0","1": 3 3 2 3 2 2 3 3 3 2 ...

> counts <- ddply(Q3df, .(Q3df$SentimentScore), nrow)

> counts

Q3df$SentimentScore V1

1 -1 7470

2 0 21688

3 1 20842

# Subset Selection Methods

# Best Subset Selection

#lapply(Q3df["SentimentScore"], unique)

#install.packages("leaps")

> library(leaps)

> regfit.full=regsubsets(SentimentScore~.,Q3df)

> summary(regfit.full)

Subset selection object

Call: regsubsets.formula(SentimentScore ~ ., Q3df)

24 Variables (and intercept)

Forced in Forced out

CommentScore FALSE FALSE

CommentCrDays FALSE FALSE

CommentLength FALSE FALSE

comment\_owner\_reputation FALSE FALSE

comment\_owner\_profile\_summary FALSE FALSE

comment\_owner\_views FALSE FALSE

comment\_owner\_upvotes FALSE FALSE

comment\_owner\_downvotes FALSE FALSE

comment\_owner\_lastactivity\_days FALSE FALSE

editDurationAfterCreation FALSE FALSE

activityDurationAfterCreation FALSE FALSE

title\_length FALSE FALSE

num\_tags FALSE FALSE

PostAnswerCount FALSE FALSE

num\_favorite FALSE FALSE

hascode FALSE FALSE

post\_views FALSE FALSE

postTypeId FALSE FALSE

IsAcceptedAnswer FALSE FALSE

postScore FALSE FALSE

post\_length FALSE FALSE

PostCommentCount FALSE FALSE

PostUpVotes FALSE FALSE

PostDownVotes FALSE FALSE

1 subsets of each size up to 8

Selection Algorithm: exhaustive

CommentScore CommentCrDays CommentLength comment\_owner\_reputation

1 ( 1 ) " " " " " " " "

2 ( 1 ) " " " " " " " "

3 ( 1 ) " " " " " " " "

4 ( 1 ) " " " " " " " "

5 ( 1 ) " " " " " " " "

6 ( 1 ) " " " " "\*" " "

7 ( 1 ) " " " " "\*" " "

8 ( 1 ) " " " " "\*" " "

comment\_owner\_profile\_summary comment\_owner\_views comment\_owner\_upvotes

1 ( 1 ) " " " " " "

2 ( 1 ) " " " " " "

3 ( 1 ) " " " " " "

4 ( 1 ) " " " " " "

5 ( 1 ) " " " " "\*"

6 ( 1 ) " " " " "\*"

7 ( 1 ) " " " " "\*"

8 ( 1 ) " " " " "\*"

comment\_owner\_downvotes comment\_owner\_lastactivity\_days editDurationAfterCreation

1 ( 1 ) " " " " " "

2 ( 1 ) " " " " " "

3 ( 1 ) " " " " " "

4 ( 1 ) " " "\*" " "

5 ( 1 ) " " "\*" " "

6 ( 1 ) " " "\*" " "

7 ( 1 ) " " "\*" " "

8 ( 1 ) "\*" "\*" " "

activityDurationAfterCreation title\_length num\_tags PostAnswerCount num\_favorite

1 ( 1 ) " " " " " " " " " "

2 ( 1 ) " " " " " " " " " "

3 ( 1 ) " " " " " " " " " "

4 ( 1 ) " " " " " " " " " "

5 ( 1 ) " " " " " " " " " "

6 ( 1 ) " " " " " " " " " "

7 ( 1 ) " " " " " " " " "\*"

8 ( 1 ) " " " " " " " " "\*"

hascode post\_views postTypeId IsAcceptedAnswer postScore post\_length

1 ( 1 ) " " " " "\*" " " " " " "

2 ( 1 ) " " " " "\*" " " " " " "

3 ( 1 ) " " " " "\*" "\*" " " " "

4 ( 1 ) " " " " "\*" "\*" " " " "

5 ( 1 ) " " " " "\*" "\*" " " " "

6 ( 1 ) " " " " "\*" "\*" " " " "

7 ( 1 ) " " " " "\*" "\*" " " " "

8 ( 1 ) " " " " "\*" "\*" " " " "

PostCommentCount PostUpVotes PostDownVotes

1 ( 1 ) " " " " " "

2 ( 1 ) "\*" " " " "

3 ( 1 ) "\*" " " " "

4 ( 1 ) "\*" " " " "

5 ( 1 ) "\*" " " " "

6 ( 1 ) "\*" " " " "

7 ( 1 ) "\*" " " " "

8 ( 1 ) "\*" " " " "

> regfit.full=regsubsets(SentimentScore~.,data=Q3df,nvmax=25)

> reg.summary=summary(regfit.full)

> names(reg.summary)

[1] "which" "rsq" "rss" "adjr2" "cp" "bic" "outmat" "obj"

> reg.summary$rsq

[1] 0.009506542 0.013498615 0.016337592 0.017578879 0.018515803 0.019007624 0.019412887

[8] 0.019733968 0.019924789 0.020082617 0.020221754 0.020311359 0.020398851 0.020476067

[15] 0.020548119 0.020594568 0.020640106 0.020652275 0.020667584 0.020679801 0.020689862

[22] 0.020697619 0.020703239 0.020704819

> par(mfrow=c(2,2))

> plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")

> plot(reg.summary$adjr2,xlab="Number of Variables",ylab="Adjusted

+ RSq",type="l")

> which.max(reg.summary$adjr2)

[1] 17

> points(25,reg.summary$adjr2[25], col="red",cex=2,pch=20)

> plot(reg.summary$cp,xlab="Number of Variables",ylab="Cp",type='l')

> which.min(reg.summary$cp)

[1] 17

> points(24,reg.summary$cp[24],col="red",cex=2,pch=20)

> which.min(reg.summary$bic)

[1] 8

> plot(reg.summary$bic,xlab="Number of Variables",ylab="BIC",type='l')

> points(6,reg.summary$bic[6],col="red",cex=2,pch=20)

> coef(regfit.full,6)

(Intercept) CommentLength

2.131011e+00 1.445234e-04

comment\_owner\_upvotes comment\_owner\_lastactivity\_days

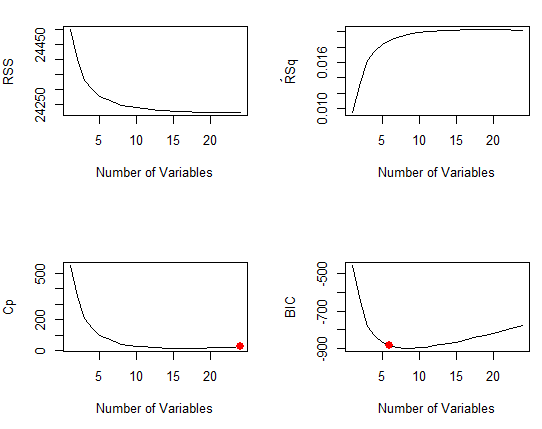
-9.117726e-06 5.661424e-04

postTypeId IsAcceptedAnswer

9.301572e-02 9.779436e-02

PostCommentCount

-9.108750e-03



# Forward and Backward Stepwise Selection

> regfit.fwd=regsubsets(SentimentScore~.,data=Q3df,nvmax=25,method="forward")

> summary(regfit.fwd)

Subset selection object

Call: regsubsets.formula(SentimentScore ~ ., data = Q3df, nvmax = 25,

method = "forward")

24 Variables (and intercept)

Forced in Forced out

CommentScore FALSE FALSE

CommentCrDays FALSE FALSE

CommentLength FALSE FALSE

comment\_owner\_reputation FALSE FALSE

comment\_owner\_profile\_summary FALSE FALSE

comment\_owner\_views FALSE FALSE

comment\_owner\_upvotes FALSE FALSE

comment\_owner\_downvotes FALSE FALSE

comment\_owner\_lastactivity\_days FALSE FALSE

editDurationAfterCreation FALSE FALSE

activityDurationAfterCreation FALSE FALSE

title\_length FALSE FALSE

num\_tags FALSE FALSE

PostAnswerCount FALSE FALSE

num\_favorite FALSE FALSE

hascode FALSE FALSE

post\_views FALSE FALSE

postTypeId FALSE FALSE

IsAcceptedAnswer FALSE FALSE

postScore FALSE FALSE

post\_length FALSE FALSE

PostCommentCount FALSE FALSE

PostUpVotes FALSE FALSE

PostDownVotes FALSE FALSE

1 subsets of each size up to 24

Selection Algorithm: forward

CommentScore CommentCrDays CommentLength comment\_owner\_reputation

1 ( 1 ) " " " " " " " "

2 ( 1 ) " " " " " " " "

3 ( 1 ) " " " " " " " "

4 ( 1 ) " " " " " " " "

5 ( 1 ) " " " " " " " "

6 ( 1 ) " " " " "\*" " "

7 ( 1 ) " " " " "\*" " "

8 ( 1 ) " " " " "\*" " "

9 ( 1 ) " " " " "\*" " "

10 ( 1 ) "\*" " " "\*" " "

11 ( 1 ) "\*" " " "\*" " "

12 ( 1 ) "\*" " " "\*" " "

13 ( 1 ) "\*" "\*" "\*" " "

14 ( 1 ) "\*" "\*" "\*" " "

15 ( 1 ) "\*" "\*" "\*" "\*"

16 ( 1 ) "\*" "\*" "\*" "\*"

17 ( 1 ) "\*" "\*" "\*" "\*"

18 ( 1 ) "\*" "\*" "\*" "\*"

19 ( 1 ) "\*" "\*" "\*" "\*"

20 ( 1 ) "\*" "\*" "\*" "\*"

21 ( 1 ) "\*" "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*"

comment\_owner\_profile\_summary comment\_owner\_views comment\_owner\_upvotes

1 ( 1 ) " " " " " "

2 ( 1 ) " " " " " "

3 ( 1 ) " " " " " "

4 ( 1 ) " " " " " "

5 ( 1 ) " " " " "\*"

6 ( 1 ) " " " " "\*"

7 ( 1 ) " " " " "\*"

8 ( 1 ) " " " " "\*"

9 ( 1 ) " " " " "\*"

10 ( 1 ) " " " " "\*"

11 ( 1 ) "\*" " " "\*"

12 ( 1 ) "\*" " " "\*"

13 ( 1 ) "\*" " " "\*"

14 ( 1 ) "\*" "\*" "\*"

15 ( 1 ) "\*" "\*" "\*"

16 ( 1 ) "\*" "\*" "\*"

17 ( 1 ) "\*" "\*" "\*"

18 ( 1 ) "\*" "\*" "\*"

19 ( 1 ) "\*" "\*" "\*"

20 ( 1 ) "\*" "\*" "\*"

21 ( 1 ) "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*"

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

1 ( 1 ) " " " "

2 ( 1 ) " " " "

3 ( 1 ) " " " "

4 ( 1 ) " " "\*"

5 ( 1 ) " " "\*"

6 ( 1 ) " " "\*"

7 ( 1 ) " " "\*"

8 ( 1 ) "\*" "\*"

9 ( 1 ) "\*" "\*"

10 ( 1 ) "\*" "\*"

11 ( 1 ) "\*" "\*"

12 ( 1 ) "\*" "\*"

13 ( 1 ) "\*" "\*"

14 ( 1 ) "\*" "\*"

15 ( 1 ) "\*" "\*"

16 ( 1 ) "\*" "\*"

17 ( 1 ) "\*" "\*"

18 ( 1 ) "\*" "\*"

19 ( 1 ) "\*" "\*"

20 ( 1 ) "\*" "\*"

21 ( 1 ) "\*" "\*"

22 ( 1 ) "\*" "\*"

23 ( 1 ) "\*" "\*"

24 ( 1 ) "\*" "\*"

editDurationAfterCreation activityDurationAfterCreation title\_length num\_tags

1 ( 1 ) " " " " " " " "

2 ( 1 ) " " " " " " " "

3 ( 1 ) " " " " " " " "

4 ( 1 ) " " " " " " " "

5 ( 1 ) " " " " " " " "

6 ( 1 ) " " " " " " " "

7 ( 1 ) " " " " " " " "

8 ( 1 ) " " " " " " " "

9 ( 1 ) " " " " " " " "

10 ( 1 ) " " " " " " " "

11 ( 1 ) " " " " " " " "

12 ( 1 ) " " " " " " " "

13 ( 1 ) " " " " " " " "

14 ( 1 ) " " " " " " " "

15 ( 1 ) " " " " " " " "

16 ( 1 ) " " "\*" " " " "

17 ( 1 ) " " "\*" " " " "

18 ( 1 ) " " "\*" "\*" " "

19 ( 1 ) " " "\*" "\*" "\*"

20 ( 1 ) "\*" "\*" "\*" "\*"

21 ( 1 ) "\*" "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*"

PostAnswerCount num\_favorite hascode post\_views postTypeId IsAcceptedAnswer

1 ( 1 ) " " " " " " " " "\*" " "

2 ( 1 ) " " " " " " " " "\*" " "

3 ( 1 ) " " " " " " " " "\*" "\*"

4 ( 1 ) " " " " " " " " "\*" "\*"

5 ( 1 ) " " " " " " " " "\*" "\*"

6 ( 1 ) " " " " " " " " "\*" "\*"

7 ( 1 ) " " "\*" " " " " "\*" "\*"

8 ( 1 ) " " "\*" " " " " "\*" "\*"

9 ( 1 ) " " "\*" " " " " "\*" "\*"

10 ( 1 ) " " "\*" " " " " "\*" "\*"

11 ( 1 ) " " "\*" " " " " "\*" "\*"

12 ( 1 ) "\*" "\*" " " " " "\*" "\*"

13 ( 1 ) "\*" "\*" " " " " "\*" "\*"

14 ( 1 ) "\*" "\*" " " " " "\*" "\*"

15 ( 1 ) "\*" "\*" " " " " "\*" "\*"

16 ( 1 ) "\*" "\*" " " " " "\*" "\*"

17 ( 1 ) "\*" "\*" " " " " "\*" "\*"

18 ( 1 ) "\*" "\*" " " " " "\*" "\*"

19 ( 1 ) "\*" "\*" " " " " "\*" "\*"

20 ( 1 ) "\*" "\*" " " " " "\*" "\*"

21 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"

postScore post\_length PostCommentCount PostUpVotes PostDownVotes

1 ( 1 ) " " " " " " " " " "

2 ( 1 ) " " " " "\*" " " " "

3 ( 1 ) " " " " "\*" " " " "

4 ( 1 ) " " " " "\*" " " " "

5 ( 1 ) " " " " "\*" " " " "

6 ( 1 ) " " " " "\*" " " " "

7 ( 1 ) " " " " "\*" " " " "

8 ( 1 ) " " " " "\*" " " " "

9 ( 1 ) " " "\*" "\*" " " " "

10 ( 1 ) " " "\*" "\*" " " " "

11 ( 1 ) " " "\*" "\*" " " " "

12 ( 1 ) " " "\*" "\*" " " " "

13 ( 1 ) " " "\*" "\*" " " " "

14 ( 1 ) " " "\*" "\*" " " " "

15 ( 1 ) " " "\*" "\*" " " " "

16 ( 1 ) " " "\*" "\*" " " " "

17 ( 1 ) " " "\*" "\*" " " "\*"

18 ( 1 ) " " "\*" "\*" " " "\*"

19 ( 1 ) " " "\*" "\*" " " "\*"

20 ( 1 ) " " "\*" "\*" " " "\*"

21 ( 1 ) " " "\*" "\*" " " "\*"

22 ( 1 ) " " "\*" "\*" " " "\*"

23 ( 1 ) " " "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*" "\*"

> regfit.bwd=regsubsets(SentimentScore~.,data=Q3df,nvmax=25,method="backward"

+ )

> summary(regfit.bwd)

Subset selection object

Call: regsubsets.formula(SentimentScore ~ ., data = Q3df, nvmax = 25,

method = "backward")

24 Variables (and intercept)

Forced in Forced out

CommentScore FALSE FALSE

CommentCrDays FALSE FALSE

CommentLength FALSE FALSE

comment\_owner\_reputation FALSE FALSE

comment\_owner\_profile\_summary FALSE FALSE

comment\_owner\_views FALSE FALSE

comment\_owner\_upvotes FALSE FALSE

comment\_owner\_downvotes FALSE FALSE

comment\_owner\_lastactivity\_days FALSE FALSE

editDurationAfterCreation FALSE FALSE

activityDurationAfterCreation FALSE FALSE

title\_length FALSE FALSE

num\_tags FALSE FALSE

PostAnswerCount FALSE FALSE

num\_favorite FALSE FALSE

hascode FALSE FALSE

post\_views FALSE FALSE

postTypeId FALSE FALSE

IsAcceptedAnswer FALSE FALSE

postScore FALSE FALSE

post\_length FALSE FALSE

PostCommentCount FALSE FALSE

PostUpVotes FALSE FALSE

PostDownVotes FALSE FALSE

1 subsets of each size up to 24

Selection Algorithm: backward

CommentScore CommentCrDays CommentLength comment\_owner\_reputation

1 ( 1 ) " " " " " " " "

2 ( 1 ) " " " " " " " "

3 ( 1 ) " " " " " " " "

4 ( 1 ) " " " " " " " "

5 ( 1 ) " " " " " " " "

6 ( 1 ) " " " " "\*" " "

7 ( 1 ) " " " " "\*" " "

8 ( 1 ) " " " " "\*" "\*"

9 ( 1 ) " " " " "\*" "\*"

10 ( 1 ) " " " " "\*" "\*"

11 ( 1 ) "\*" " " "\*" "\*"

12 ( 1 ) "\*" " " "\*" "\*"

13 ( 1 ) "\*" " " "\*" "\*"

14 ( 1 ) "\*" "\*" "\*" "\*"

15 ( 1 ) "\*" "\*" "\*" "\*"

16 ( 1 ) "\*" "\*" "\*" "\*"

17 ( 1 ) "\*" "\*" "\*" "\*"

18 ( 1 ) "\*" "\*" "\*" "\*"

19 ( 1 ) "\*" "\*" "\*" "\*"

20 ( 1 ) "\*" "\*" "\*" "\*"

21 ( 1 ) "\*" "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*"

comment\_owner\_profile\_summary comment\_owner\_views comment\_owner\_upvotes

1 ( 1 ) " " " " " "

2 ( 1 ) " " " " " "

3 ( 1 ) " " " " " "

4 ( 1 ) " " " " " "

5 ( 1 ) " " " " " "

6 ( 1 ) " " " " " "

7 ( 1 ) " " " " " "

8 ( 1 ) " " " " " "

9 ( 1 ) " " " " " "

10 ( 1 ) " " "\*" " "

11 ( 1 ) " " "\*" " "

12 ( 1 ) "\*" "\*" " "

13 ( 1 ) "\*" "\*" " "

14 ( 1 ) "\*" "\*" " "

15 ( 1 ) "\*" "\*" "\*"

16 ( 1 ) "\*" "\*" "\*"

17 ( 1 ) "\*" "\*" "\*"

18 ( 1 ) "\*" "\*" "\*"

19 ( 1 ) "\*" "\*" "\*"

20 ( 1 ) "\*" "\*" "\*"

21 ( 1 ) "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*"

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

1 ( 1 ) " " " "

2 ( 1 ) " " " "

3 ( 1 ) " " " "

4 ( 1 ) " " "\*"

5 ( 1 ) "\*" "\*"

6 ( 1 ) "\*" "\*"

7 ( 1 ) "\*" "\*"

8 ( 1 ) "\*" "\*"

9 ( 1 ) "\*" "\*"

10 ( 1 ) "\*" "\*"

11 ( 1 ) "\*" "\*"

12 ( 1 ) "\*" "\*"

13 ( 1 ) "\*" "\*"

14 ( 1 ) "\*" "\*"

15 ( 1 ) "\*" "\*"

16 ( 1 ) "\*" "\*"

17 ( 1 ) "\*" "\*"

18 ( 1 ) "\*" "\*"

19 ( 1 ) "\*" "\*"

20 ( 1 ) "\*" "\*"

21 ( 1 ) "\*" "\*"

22 ( 1 ) "\*" "\*"

23 ( 1 ) "\*" "\*"

24 ( 1 ) "\*" "\*"

editDurationAfterCreation activityDurationAfterCreation title\_length num\_tags

1 ( 1 ) " " " " " " " "

2 ( 1 ) " " " " " " " "

3 ( 1 ) " " " " " " " "

4 ( 1 ) " " " " " " " "

5 ( 1 ) " " " " " " " "

6 ( 1 ) " " " " " " " "

7 ( 1 ) " " " " " " " "

8 ( 1 ) " " " " " " " "

9 ( 1 ) " " " " " " " "

10 ( 1 ) " " " " " " " "

11 ( 1 ) " " " " " " " "

12 ( 1 ) " " " " " " " "

13 ( 1 ) " " " " " " " "

14 ( 1 ) " " " " " " " "

15 ( 1 ) " " " " " " " "

16 ( 1 ) " " "\*" " " " "

17 ( 1 ) " " "\*" " " " "

18 ( 1 ) " " "\*" " " " "

19 ( 1 ) " " "\*" " " " "

20 ( 1 ) " " "\*" "\*" " "

21 ( 1 ) " " "\*" "\*" "\*"

22 ( 1 ) " " "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*"

PostAnswerCount num\_favorite hascode post\_views postTypeId IsAcceptedAnswer

1 ( 1 ) " " " " " " " " "\*" " "

2 ( 1 ) " " " " " " " " "\*" " "

3 ( 1 ) " " " " " " " " "\*" "\*"

4 ( 1 ) " " " " " " " " "\*" "\*"

5 ( 1 ) " " " " " " " " "\*" "\*"

6 ( 1 ) " " " " " " " " "\*" "\*"

7 ( 1 ) " " "\*" " " " " "\*" "\*"

8 ( 1 ) " " "\*" " " " " "\*" "\*"

9 ( 1 ) " " "\*" " " " " "\*" "\*"

10 ( 1 ) " " "\*" " " " " "\*" "\*"

11 ( 1 ) " " "\*" " " " " "\*" "\*"

12 ( 1 ) " " "\*" " " " " "\*" "\*"

13 ( 1 ) "\*" "\*" " " " " "\*" "\*"

14 ( 1 ) "\*" "\*" " " " " "\*" "\*"

15 ( 1 ) "\*" "\*" " " " " "\*" "\*"

16 ( 1 ) "\*" "\*" " " " " "\*" "\*"

17 ( 1 ) "\*" "\*" " " " " "\*" "\*"

18 ( 1 ) "\*" "\*" " " " " "\*" "\*"

19 ( 1 ) "\*" "\*" " " " " "\*" "\*"

20 ( 1 ) "\*" "\*" " " " " "\*" "\*"

21 ( 1 ) "\*" "\*" " " " " "\*" "\*"

22 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" " " "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"

postScore post\_length PostCommentCount PostUpVotes PostDownVotes

1 ( 1 ) " " " " " " " " " "

2 ( 1 ) " " " " "\*" " " " "

3 ( 1 ) " " " " "\*" " " " "

4 ( 1 ) " " " " "\*" " " " "

5 ( 1 ) " " " " "\*" " " " "

6 ( 1 ) " " " " "\*" " " " "

7 ( 1 ) " " " " "\*" " " " "

8 ( 1 ) " " " " "\*" " " " "

9 ( 1 ) " " "\*" "\*" " " " "

10 ( 1 ) " " "\*" "\*" " " " "

11 ( 1 ) " " "\*" "\*" " " " "

12 ( 1 ) " " "\*" "\*" " " " "

13 ( 1 ) " " "\*" "\*" " " " "

14 ( 1 ) " " "\*" "\*" " " " "

15 ( 1 ) " " "\*" "\*" " " " "

16 ( 1 ) " " "\*" "\*" " " " "

17 ( 1 ) " " "\*" "\*" " " "\*"

18 ( 1 ) " " "\*" "\*" "\*" "\*"

19 ( 1 ) "\*" "\*" "\*" "\*" "\*"

20 ( 1 ) "\*" "\*" "\*" "\*" "\*"

21 ( 1 ) "\*" "\*" "\*" "\*" "\*"

22 ( 1 ) "\*" "\*" "\*" "\*" "\*"

23 ( 1 ) "\*" "\*" "\*" "\*" "\*"

24 ( 1 ) "\*" "\*" "\*" "\*" "\*"

> coef(regfit.full,7)

(Intercept) CommentLength

2.121948e+00 1.391808e-04

comment\_owner\_upvotes comment\_owner\_lastactivity\_days

-9.200970e-06 5.703298e-04

num\_favorite postTypeId

9.054091e-03 9.883489e-02

IsAcceptedAnswer PostCommentCount

9.811977e-02 -9.464200e-03

> coef(regfit.fwd,7)

(Intercept) CommentLength

2.121948e+00 1.391808e-04

comment\_owner\_upvotes comment\_owner\_lastactivity\_days

-9.200970e-06 5.703298e-04

num\_favorite postTypeId

9.054091e-03 9.883489e-02

IsAcceptedAnswer PostCommentCount

9.811977e-02 -9.464200e-03

> coef(regfit.bwd,7)

(Intercept) CommentLength

2.119595e+00 1.284322e-04

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-6.656976e-06 6.057056e-04

num\_favorite postTypeId

8.785009e-03 9.752432e-02

IsAcceptedAnswer PostCommentCount

9.879619e-02 -9.483886e-03

> set.seed(1)

> train=sample(c(TRUE,FALSE), nrow(Q3df),rep=TRUE)

> test=(!train)

> regfit.best=regsubsets(SentimentScore~.,data=Q3df[train,],nvmax=25)

> test.mat=model.matrix(SentimentScore~.,data=Q3df[test,])

> val.errors=rep(NA,25)

> for(i in 1:24){

+ coefi=coef(regfit.best,id=i)

+ pred=test.mat[,names(coefi)]%\*%coefi

+ val.errors[i]=mean((Q2df$accepted\_answer\_flag[test]-pred)^2)

+ }

> val.errors

[1] 4.943091 4.948854 4.949073 4.949220 4.948741 4.950904 4.951578 4.951579 4.951543

[10] 4.950707 4.950675 4.951003 4.950558 4.950523 4.950217 4.950279 4.950223 4.949929

[19] 4.950079 4.950081 4.950154 4.950113 4.950113 4.950134 NA

> which.min(val.errors)

[1] 1

> coef(regfit.best,24)

(Intercept) CommentScore

2.050387e+00 -2.527981e-03

CommentCrDays CommentLength

-1.113496e-04 1.381652e-04

comment\_owner\_reputation comment\_owner\_profile\_summary

-2.040678e-07 -5.646625e-03

comment\_owner\_views comment\_owner\_upvotes

3.086106e-07 -3.316175e-06

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-2.657463e-06 4.328995e-04

editDurationAfterCreation activityDurationAfterCreation

9.331199e-05 2.269298e-04

title\_length num\_tags

3.245885e-04 1.975285e-03

PostAnswerCount num\_favorite

6.764146e-03 9.980540e-03

hascode post\_views

-5.754213e-03 -4.360660e-06

postTypeId IsAcceptedAnswer

1.477826e-01 1.037886e-01

postScore post\_length

-3.318392e-03 6.331601e-06

PostCommentCount PostUpVotes

-1.051133e-02 3.217659e-03

PostDownVotes

-6.953369e-03

> predict.regsubsets=function(object,newdata,id,...){

+ form=as.formula(object$call[[2]])

+ mat=model.matrix(form,newdata)

+ coefi=coef(object,id=id)

+ xvars=names(coefi)

+ mat[,xvars]%\*%coefi

+ }

> regfit.best=regsubsets(SentimentScore~.,data=Q3df,nvmax=26)

> coef(regfit.best,24)

(Intercept) CommentScore

4.653228e-02 -5.486034e-03

CommentCrDays CommentLength

-7.392533e-05 1.285232e-04

comment\_owner\_reputation comment\_owner\_profile\_summary

-1.863706e-07 -7.202571e-03

comment\_owner\_views comment\_owner\_upvotes

2.790236e-07 -3.499492e-06

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-3.429080e-06 5.395105e-04

editDurationAfterCreation activityDurationAfterCreation

1.226762e-04 1.258002e-04

title\_length num\_tags

2.273232e-04 -3.243711e-03

PostAnswerCount num\_favorite

1.014764e-02 9.449765e-03

hascode post\_views

3.714281e-03 -2.830536e-06

postTypeId IsAcceptedAnswer

1.449792e-01 9.685918e-02

postScore post\_length

-2.889953e-02 6.184775e-06

PostCommentCount PostUpVotes

-9.523447e-03 2.895800e-02

PostDownVotes

-3.299746e-02

> k=24

> set.seed(1)

> folds=sample(1:k,nrow(Q3df),replace=TRUE)

> cv.errors=matrix(NA,k,25, dimnames=list(NULL, paste(1:25)))

> for(j in 1:k){

+ best.fit=regsubsets(SentimentScore~.,data=Q3df[folds!=j,],nvmax=24)

+ for(i in 1:24){

+ pred=predict(best.fit,Q3df[folds==j,],id=i)

+ cv.errors[j,i]=mean( (Q3df$SentimentScore[folds==j]-pred)^2)

+ }

+ }

> mean.cv.errors=apply(cv.errors,2,mean)

> mean.cv.errors

1 2 3 4 5 6 7 8 9

0.4901022 0.4881375 0.4867505 0.4865834 0.4859859 0.4858126 0.4854753 0.4851784 0.4852343

10 11 12 13 14 15 16 17 18

0.4852620 0.4850431 0.4852058 0.4851952 0.4851196 0.4850597 0.4850249 0.4849512 0.4850309

19 20 21 22 23 24 25

0.4850688 0.4850668 0.4850721 0.4850371 0.4850080 0.4850067 NA

> par(mfrow=c(1,1))

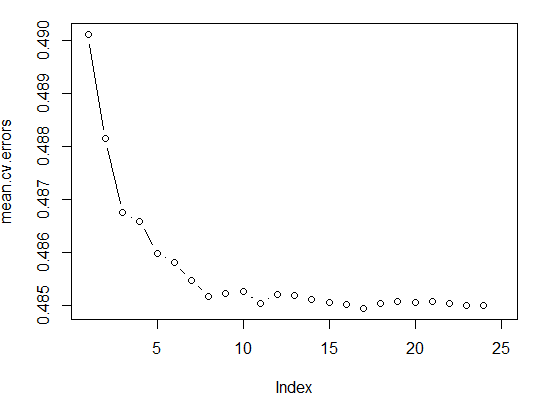
> plot(mean.cv.errors,type='b')

> reg.best=regsubsets(SentimentScore~.,data=Q3df, nvmax=24)

> coef(reg.best,3)

(Intercept) postTypeId IsAcceptedAnswer PostCommentCount

0.141909889 0.097384891 0.101440036 -0.009116773



> fix(Q3df)

> names(Q3df)

[1] "CommentScore" "CommentCrDays"

[3] "CommentLength" "comment\_owner\_reputation"

[5] "comment\_owner\_profile\_summary" "comment\_owner\_views"

[7] "comment\_owner\_upvotes" "comment\_owner\_downvotes"

[9] "comment\_owner\_lastactivity\_days" "editDurationAfterCreation"

[11] "activityDurationAfterCreation" "title\_length"

[13] "num\_tags" "PostAnswerCount"

[15] "num\_favorite" "hascode"

[17] "post\_views" "postTypeId"

[19] "IsAcceptedAnswer" "postScore"

[21] "post\_length" "PostCommentCount"

[23] "PostUpVotes" "PostDownVotes"

[25] "SentimentScore"

> dim(Q3df)

[1] 50000 25

> sum(is.na(Q3df$accepted\_answer\_flag))

[1] 0

> Q2df=na.omit(Q3df)

> dim(Q3df)

[1] 50000 25

> sum(is.na(Q3df))

[1] 0

> library(glmnet)

> x=model.matrix(SentimentScore~.,Q3df)[,-1]

> y=Q3df$SentimentScore

> grid=10^seq(10,-2,length=100)

> ridge.mod=glmnet(x,y,alpha=0,lambda=grid)

> dim(coef(ridge.mod))

[1] 25 100

> ridge.mod$lambda[50]

[1] 11497.57

> coef(ridge.mod)[,50]

(Intercept) CommentScore

2.674400e-01 -7.545200e-07

CommentCrDays CommentLength

-1.504906e-09 4.643934e-09

comment\_owner\_reputation comment\_owner\_profile\_summary

-2.566712e-11 -1.387635e-06

comment\_owner\_views comment\_owner\_upvotes

-1.233101e-11 -7.920546e-10

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-5.901519e-10 4.785541e-08

editDurationAfterCreation activityDurationAfterCreation

1.654108e-08 -7.134305e-09

title\_length num\_tags

-1.366800e-07 -2.376027e-06

PostAnswerCount num\_favorite

-1.625277e-06 -2.733356e-07

hascode post\_views

2.802591e-08 -7.188613e-10

postTypeId IsAcceptedAnswer

8.787701e-06 8.716925e-06

postScore post\_length

-3.057351e-09 -9.243461e-11

PostCommentCount PostUpVotes

-6.355428e-07 -1.579046e-08

PostDownVotes

-1.810618e-06

> sqrt(sum(coef(ridge.mod)[-1,50]^2))

[1] 1.295259e-05

> coef(ridge.mod)[,60]

(Intercept) CommentScore

2.674395e-01 -1.225817e-05

CommentCrDays CommentLength

-2.451127e-08 7.569856e-08

comment\_owner\_reputation comment\_owner\_profile\_summary

-4.171656e-10 -2.256107e-05

comment\_owner\_views comment\_owner\_upvotes

-1.996953e-10 -1.287272e-08

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-9.592233e-09 7.784721e-07

editDurationAfterCreation activityDurationAfterCreation

2.695786e-07 -1.150359e-07

title\_length num\_tags

-2.219797e-06 -3.858298e-05

PostAnswerCount num\_favorite

-2.638460e-05 -4.400206e-06

hascode post\_views

4.625604e-07 -1.161979e-08

postTypeId IsAcceptedAnswer

1.426654e-04 1.417003e-04

postScore post\_length

-4.692492e-08 -1.483142e-09

PostCommentCount PostUpVotes

-1.033788e-05 -2.536661e-07

PostDownVotes

-2.940462e-05

> sqrt(sum(coef(ridge.mod)[-1,60]^2))

[1] 0.000210412

> predict(ridge.mod,s=500,type="coefficients")[1:25,]

(Intercept) CommentScore

2.674393e-01 -1.752933e-05

CommentCrDays CommentLength

-3.509461e-08 1.084246e-07

comment\_owner\_reputation comment\_owner\_profile\_summary

-5.966685e-10 -3.227444e-05

comment\_owner\_views comment\_owner\_upvotes

-2.851228e-10 -1.841141e-08

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-1.372007e-08 1.113916e-06

editDurationAfterCreation activityDurationAfterCreation

3.860876e-07 -1.638989e-07

title\_length num\_tags

-3.173814e-06 -5.516110e-05

PostAnswerCount num\_favorite

-3.771628e-05 -6.264268e-06

hascode post\_views

6.665143e-07 -1.657552e-08

postTypeId IsAcceptedAnswer

2.039421e-04 2.026904e-04

postScore post\_length

-6.520343e-08 -2.108022e-09

PostCommentCount PostUpVotes

-1.479208e-05 -3.607602e-07

PostDownVotes

-4.204118e-05

> set.seed(1)

> train=sample(1:nrow(x), nrow(x)/2)

> test=(-train)

> y.test=y[test]

> ridge.mod=glmnet(x[train,],y[train],alpha=0,lambda=grid, thresh=1e-12)

> ridge.pred=predict(ridge.mod,s=4,newx=x[test,])

> mean((ridge.pred-y.test)^2)

[1] 0.4869801

> mean((mean(y[train])-y.test)^2)

[1] 0.4920119

> ridge.pred=predict(ridge.mod,s=1e10,newx=x[test,])

> mean((ridge.pred-y.test)^2)

[1] 0.4920119

> ridge.pred=predict(ridge.mod,s=0,newx=x[test,],exact=T)

> mean((ridge.pred-y.test)^2)

[1] 0.4822871

> lm(y~x, subset=train)

Call:

lm(formula = y ~ x, subset = train)

Coefficients:

(Intercept) xCommentScore

3.034e-02 -8.285e-03

xCommentCrDays xCommentLength

-1.218e-04 1.336e-04

xcomment\_owner\_reputation xcomment\_owner\_profile\_summary

-2.228e-07 -4.584e-03

xcomment\_owner\_views xcomment\_owner\_upvotes

4.941e-07 -1.288e-06

xcomment\_owner\_downvotes xcomment\_owner\_lastactivity\_days

-3.821e-06 5.516e-04

xeditDurationAfterCreation xactivityDurationAfterCreation

8.688e-05 1.410e-04

xtitle\_length xnum\_tags

1.445e-04 -5.078e-03

xPostAnswerCount xnum\_favorite

1.539e-02 5.591e-03

xhascode xpost\_views

-2.208e-04 1.811e-06

xpostTypeId xIsAcceptedAnswer

1.520e-01 1.127e-01

xpostScore xpost\_length

-1.139e-02 4.312e-06

xPostCommentCount xPostUpVotes

-9.086e-03 1.122e-02

xPostDownVotes

-1.561e-02

> predict(ridge.mod,s=0,exact=T,type="coefficients")[1:25,]

(Intercept) CommentScore

3.030178e-02 -8.284729e-03

CommentCrDays CommentLength

-1.217552e-04 1.336419e-04

comment\_owner\_reputation comment\_owner\_profile\_summary

-2.228017e-07 -4.584323e-03

comment\_owner\_views comment\_owner\_upvotes

4.941145e-07 -1.287735e-06

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-3.821179e-06 5.515542e-04

editDurationAfterCreation activityDurationAfterCreation

8.691441e-05 1.409845e-04

title\_length num\_tags

1.445228e-04 -5.081293e-03

PostAnswerCount num\_favorite

1.539464e-02 5.592095e-03

hascode post\_views

-2.207117e-04 1.808175e-06

postTypeId IsAcceptedAnswer

1.519684e-01 1.126509e-01

postScore post\_length

-1.066573e-02 4.313745e-06

PostCommentCount PostUpVotes

-9.086300e-03 1.048876e-02

PostDownVotes

-1.487913e-02

> set.seed(1)

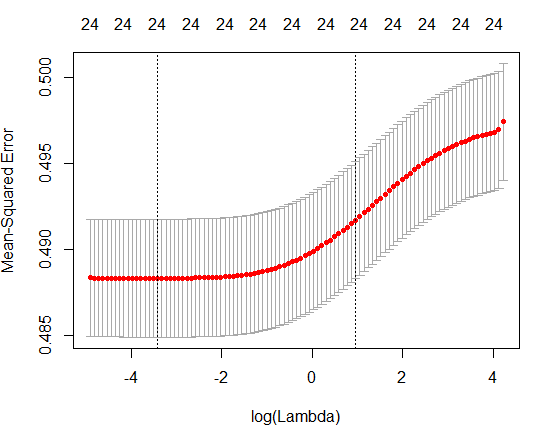
> cv.out=cv.glmnet(x[train,],y[train],alpha=0)

> plot(cv.out)

> bestlam=cv.out$lambda.min

> bestlam

[1] 0.03287641



> ridge.pred=predict(ridge.mod,s=bestlam,newx=x[test,])

> mean((ridge.pred-y.test)^2)

[1] 0.482261

> out=glmnet(x,y,alpha=0)

> predict(out,type="coefficients",s=bestlam)[1:25,]

(Intercept) CommentScore

1.803943e-01 -5.261442e-03

CommentCrDays CommentLength

-6.501377e-05 1.216346e-04

comment\_owner\_reputation comment\_owner\_profile\_summary

-1.687244e-07 -7.269208e-03

comment\_owner\_views comment\_owner\_upvotes

2.470619e-07 -3.653344e-06

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-3.466172e-06 5.169837e-04

editDurationAfterCreation activityDurationAfterCreation

1.349340e-04 1.151007e-04

title\_length num\_tags

-7.302651e-05 -6.221418e-03

PostAnswerCount num\_favorite

1.101935e-03 8.984145e-03

hascode post\_views

3.824967e-03 -7.478733e-07

postTypeId IsAcceptedAnswer

7.657585e-02 9.278628e-02

postScore post\_length

4.596677e-05 5.831033e-06

PostCommentCount PostUpVotes

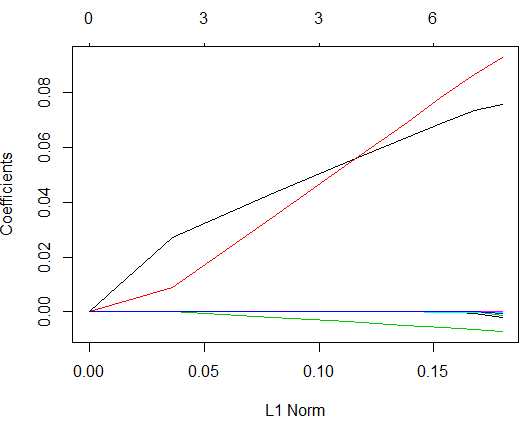
-8.992794e-03 6.428105e-05

PostDownVotes

-4.927510e-03

> lasso.mod=glmnet(x[train,],y[train],alpha=1,lambda=grid)

> plot(lasso.mod)

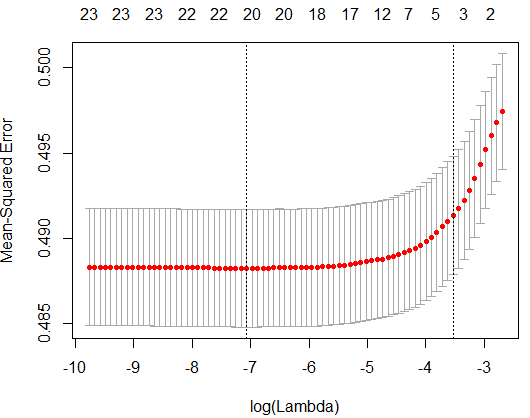


> set.seed(1)

> cv.out=cv.glmnet(x[train,],y[train],alpha=1)

> plot(cv.out)

> bestlam=cv.out$lambda.min



> lasso.pred=predict(lasso.mod,s=bestlam,newx=x[test,])

> mean((lasso.pred-y.test)^2)

[1] 0.4831451

> out=glmnet(x,y,alpha=1,lambda=grid)

> lasso.coef=predict(out,type="coefficients",s=bestlam)[1:25,]

> lasso.coef

(Intercept) CommentScore

1.616890e-01 0.000000e+00

CommentCrDays CommentLength

0.000000e+00 3.132565e-05

comment\_owner\_reputation comment\_owner\_profile\_summary

-2.980265e-08 -3.489407e-03

comment\_owner\_views comment\_owner\_upvotes

0.000000e+00 -3.290342e-06

comment\_owner\_downvotes comment\_owner\_lastactivity\_days

-2.532104e-06 3.298435e-04

editDurationAfterCreation activityDurationAfterCreation

0.000000e+00 0.000000e+00

title\_length num\_tags

0.000000e+00 0.000000e+00

PostAnswerCount num\_favorite

0.000000e+00 3.343166e-04

hascode post\_views

0.000000e+00 0.000000e+00

postTypeId IsAcceptedAnswer

8.235835e-02 7.838857e-02

postScore post\_length

0.000000e+00 0.000000e+00

PostCommentCount PostUpVotes

-7.186783e-03 0.000000e+00

PostDownVotes

-1.788575e-04

> lasso.coef[lasso.coef!=0]

(Intercept) CommentLength

1.616890e-01 3.132565e-05

comment\_owner\_reputation comment\_owner\_profile\_summary

-2.980265e-08 -3.489407e-03

comment\_owner\_upvotes comment\_owner\_downvotes

-3.290342e-06 -2.532104e-06

comment\_owner\_lastactivity\_days num\_favorite

3.298435e-04 3.343166e-04

postTypeId IsAcceptedAnswer

8.235835e-02 7.838857e-02

PostCommentCount PostDownVotes

-7.186783e-03 -1.788575e-04

> library(pls)

> set.seed(2)

> pcr.fit=pcr(SentimentScore~., data=Q3df,scale=TRUE,validation="CV")

> summary(pcr.fit)

Data: X dimension: 50000 24

Y dimension: 50000 1

Fit method: svdpc

Number of components considered: 24

VALIDATION: RMSEP

Cross-validated using 10 random segments.

(Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

CV 0.7034 0.7002 0.6994 0.6989 0.6988 0.6988 0.6986 0.6985 0.6976

adjCV 0.7034 0.7002 0.6994 0.6989 0.6988 0.6988 0.6985 0.6985 0.6976

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps 16 comps

CV 0.6976 0.6976 0.6974 0.6974 0.6974 0.6973 0.6968 0.6965

adjCV 0.6976 0.6976 0.6974 0.6974 0.6974 0.6973 0.6968 0.6965

17 comps 18 comps 19 comps 20 comps 21 comps 22 comps 23 comps 24 comps

CV 0.6966 0.6966 0.6965 0.6965 0.6966 0.6966 0.6965 0.6965

adjCV 0.6966 0.6965 0.6965 0.6965 0.6965 0.6965 0.6964 0.6964

TRAINING: % variance explained

1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

X 18.9523 31.158 41.682 48.53 53.879 59.071 63.469 67.612

SentimentScore 0.8968 1.151 1.296 1.31 1.337 1.403 1.426 1.672

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps

X 71.578 75.236 78.645 81.987 85.267 88.036 90.623

SentimentScore 1.674 1.694 1.733 1.753 1.753 1.798 1.928

16 comps 17 comps 18 comps 19 comps 20 comps 21 comps 22 comps

X 93.154 95.308 96.612 97.634 98.600 99.329 99.852

SentimentScore 2.014 2.014 2.024 2.028 2.032 2.033 2.034

23 comps 24 comps

X 100.000 100.00

SentimentScore 2.068 2.07

> validationplot(pcr.fit,val.type="MSEP")

> set.seed(1)

> pcr.fit=pcr(SentimentScore~., data=Q3df,subset=train,scale=TRUE, validation="CV")

> validationplot(pcr.fit,val.type="MSEP")

> library(pls)

> set.seed(2)

> pcr.fit=pcr(SentimentScore~., data=Q3df,scale=TRUE,validation="CV")

> summary(pcr.fit)

Data: X dimension: 50000 24

Y dimension: 50000 1

Fit method: svdpc

Number of components considered: 24

VALIDATION: RMSEP

Cross-validated using 10 random segments.

(Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

CV 0.7034 0.7002 0.6994 0.6989 0.6988 0.6988 0.6986 0.6985 0.6976

adjCV 0.7034 0.7002 0.6994 0.6989 0.6988 0.6988 0.6985 0.6985 0.6976

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps 16 comps

CV 0.6976 0.6976 0.6974 0.6974 0.6974 0.6973 0.6968 0.6965

adjCV 0.6976 0.6976 0.6974 0.6974 0.6974 0.6973 0.6968 0.6965

17 comps 18 comps 19 comps 20 comps 21 comps 22 comps 23 comps 24 comps

CV 0.6966 0.6966 0.6965 0.6965 0.6966 0.6966 0.6965 0.6965

adjCV 0.6966 0.6965 0.6965 0.6965 0.6965 0.6965 0.6964 0.6964

TRAINING: % variance explained

1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

X 18.9523 31.158 41.682 48.53 53.879 59.071 63.469 67.612

SentimentScore 0.8968 1.151 1.296 1.31 1.337 1.403 1.426 1.672

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps

X 71.578 75.236 78.645 81.987 85.267 88.036 90.623

SentimentScore 1.674 1.694 1.733 1.753 1.753 1.798 1.928

16 comps 17 comps 18 comps 19 comps 20 comps 21 comps 22 comps

X 93.154 95.308 96.612 97.634 98.600 99.329 99.852

SentimentScore 2.014 2.014 2.024 2.028 2.032 2.033 2.034

23 comps 24 comps

X 100.000 100.00

SentimentScore 2.068 2.07

> validationplot(pcr.fit,val.type="MSEP")

> set.seed(1)

> pcr.fit=pcr(SentimentScore~., data=Q3df,subset=train,scale=TRUE, validation="CV")

> validationplot(pcr.fit,val.type="MSEP")

> pcr.pred=predict(pcr.fit,x[test,],ncomp=7)

> mean((pcr.pred-y.test)^2)

[1] 0.4846027

> pcr.fit=pcr(y~x,scale=TRUE,ncomp=7)

> summary(pcr.fit)

Data: X dimension: 50000 24

Y dimension: 50000 1

Fit method: svdpc

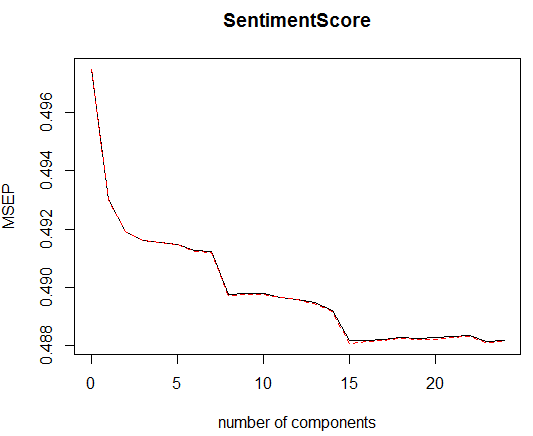
Number of components considered: 7

TRAINING: % variance explained

1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps

X 18.9523 31.158 41.682 48.53 53.879 59.071 63.469

y 0.8968 1.151 1.296 1.31 1.337 1.403 1.426



> set.seed(1)

> pls.fit=plsr(SentimentScore~., data=Q3df,subset=train,scale=TRUE, validation="CV")

> summary(pls.fit)

Data: X dimension: 25000 24

Y dimension: 25000 1

Fit method: kernelpls

Number of components considered: 24

VALIDATION: RMSEP

Cross-validated using 10 random segments.

(Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

CV 0.7053 0.7003 0.6991 0.6988 0.6987 0.6987 0.6988 0.6988 0.6988

adjCV 0.7053 0.7003 0.6991 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps 16 comps

CV 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987

adjCV 0.6987 0.6987 0.6986 0.6986 0.6986 0.6986 0.6986 0.6986

17 comps 18 comps 19 comps 20 comps 21 comps 22 comps 23 comps 24 comps

CV 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987

adjCV 0.6986 0.6986 0.6986 0.6986 0.6986 0.6987 0.6987 0.6987

TRAINING: % variance explained

1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

X 17.84 26.523 34.755 41.117 48.739 53.318 56.670 60.986

SentimentScore 1.48 1.884 2.012 2.032 2.039 2.047 2.057 2.062

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps

X 62.782 65.444 69.018 71.575 74.610 77.534 79.608

SentimentScore 2.071 2.074 2.075 2.075 2.075 2.075 2.075

16 comps 17 comps 18 comps 19 comps 20 comps 21 comps 22 comps

X 82.187 83.898 87.045 89.855 91.142 94.229 94.728

SentimentScore 2.075 2.075 2.075 2.075 2.075 2.075 2.075

23 comps 24 comps

X 96.757 100.000

SentimentScore 2.075 2.075

> set.seed(1)

> pls.fit=plsr(SentimentScore~., data=Q3df,subset=train,scale=TRUE, validation="CV")

> summary(pls.fit)

Data: X dimension: 25000 24

Y dimension: 25000 1

Fit method: kernelpls

Number of components considered: 24

VALIDATION: RMSEP

Cross-validated using 10 random segments.

(Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

CV 0.7053 0.7003 0.6991 0.6988 0.6987 0.6987 0.6988 0.6988 0.6988

adjCV 0.7053 0.7003 0.6991 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps 16 comps

CV 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987

adjCV 0.6987 0.6987 0.6986 0.6986 0.6986 0.6986 0.6986 0.6986

17 comps 18 comps 19 comps 20 comps 21 comps 22 comps 23 comps 24 comps

CV 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987 0.6987

adjCV 0.6986 0.6986 0.6986 0.6986 0.6986 0.6987 0.6987 0.6987

TRAINING: % variance explained

1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps

X 17.84 26.523 34.755 41.117 48.739 53.318 56.670 60.986

SentimentScore 1.48 1.884 2.012 2.032 2.039 2.047 2.057 2.062

9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps

X 62.782 65.444 69.018 71.575 74.610 77.534 79.608

SentimentScore 2.071 2.074 2.075 2.075 2.075 2.075 2.075

16 comps 17 comps 18 comps 19 comps 20 comps 21 comps 22 comps

X 82.187 83.898 87.045 89.855 91.142 94.229 94.728

SentimentScore 2.075 2.075 2.075 2.075 2.075 2.075 2.075

23 comps 24 comps

X 96.757 100.000

SentimentScore 2.075 2.075

> validationplot(pls.fit,val.type="MSEP")

> pls.pred=predict(pls.fit,x[test,],ncomp=2)

> mean((pls.pred-y.test)^2)

[1] 0.4822001

> pls.fit=plsr(SentimentScore~., data=Q3df,scale=TRUE,ncomp=2)

> summary(pls.fit)

Data: X dimension: 50000 24

Y dimension: 50000 1

Fit method: kernelpls

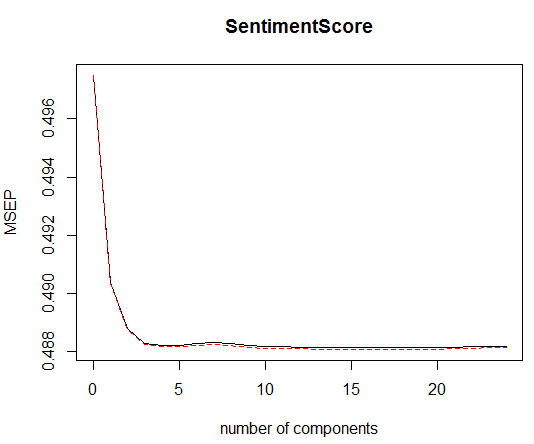
Number of components considered: 2

TRAINING: % variance explained

1 comps 2 comps

X 17.537 26.943

SentimentScore 1.529 1.888



#Dividing dataset 75% for training and 25% for testing

> Q3smp\_size=floor(0.75\*nrow(Q3df))

> set.seed(123)

> Q3train\_ind=sample(seq\_len(nrow(Q3df)), size = Q3smp\_size)

> Q3Train=Q3df[Q3train\_ind,]

> Q3Test=Q3df[-Q3train\_ind,]

> nrow(Q3Train)

[1] 37500

> nrow(Q3Test)

[1] 12500

#Initial logistic regression model with all parameters

> attach(Q3Train)

> Q3lm=glm(as.factor(SentimentScore)~.,data=Q3Train,family="binomial")

> summary(Q3lm)

Call:

glm(formula = as.factor(SentimentScore) ~ ., family = "binomial",

data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5474 0.4756 0.5244 0.5818 1.3509

Coefficients:

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

(Intercept) 2.121e+00 2.763e-01 7.676 1.64e-14 \*\*\*

CommentScore -3.614e-03 9.630e-03 -0.375 0.70743

CommentCrDays -1.988e-04 1.661e-04 -1.197 0.23140

CommentLength -2.490e-03 1.210e-04 -20.580 < 2e-16 \*\*\*

comment\_owner\_reputation -1.956e-07 3.463e-07 -0.565 0.57212

comment\_owner\_profile\_summary 3.013e-02 1.397e-02 2.157 0.03100 \*

comment\_owner\_views 8.864e-07 6.556e-07 1.352 0.17641

comment\_owner\_upvotes -1.002e-05 7.976e-06 -1.256 0.20920

comment\_owner\_downvotes -1.309e-05 4.714e-06 -2.776 0.00551 \*\*

comment\_owner\_lastactivity\_days 6.155e-04 4.127e-04 1.492 0.13582

editDurationAfterCreation 4.020e-04 1.085e-03 0.370 0.71114

activityDurationAfterCreation 2.744e-04 7.631e-04 0.360 0.71919

title\_length -8.666e-04 1.266e-03 -0.684 0.49380

num\_tags -3.500e-02 2.058e-02 -1.701 0.08893 .

PostAnswerCount 5.031e-02 2.184e-02 2.304 0.02124 \*

num\_favorite 2.683e-02 1.479e-02 1.814 0.06964 .

hascode -1.241e-02 6.087e-02 -0.204 0.83848

post\_views -3.564e-05 2.509e-05 -1.421 0.15538

postTypeId 3.721e-02 1.356e-01 0.274 0.78376

IsAcceptedAnswer 2.831e-01 4.191e-02 6.755 1.43e-11 \*\*\*

postScore 4.567e-02 1.187e-01 0.385 0.70036

post\_length -4.597e-06 1.009e-05 -0.456 0.64854

PostCommentCount -2.387e-02 2.712e-03 -8.802 < 2e-16 \*\*\*

PostUpVotes -4.520e-02 1.187e-01 -0.381 0.70335

PostDownVotes 5.253e-02 1.193e-01 0.440 0.65975

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31004 on 37475 degrees of freedom

AIC: 31054

Number of Fisher Scoring iterations: 4

#Checking for significant parameters with forward step

> Q3modelstepforward=step(Q3lm, direction="forward")

Start: AIC=31054.04

as.factor(SentimentScore) ~ CommentScore + CommentCrDays + CommentLength +

comment\_owner\_reputation + comment\_owner\_profile\_summary +

comment\_owner\_views + comment\_owner\_upvotes + comment\_owner\_downvotes +

comment\_owner\_lastactivity\_days + editDurationAfterCreation +

activityDurationAfterCreation + title\_length + num\_tags +

PostAnswerCount + num\_favorite + hascode + post\_views + postTypeId +

IsAcceptedAnswer + postScore + post\_length + PostCommentCount +

PostUpVotes + PostDownVotes

> summary(Q3modelstepforward)

Call:

glm(formula = as.factor(SentimentScore) ~ CommentScore + CommentCrDays +

CommentLength + comment\_owner\_reputation + comment\_owner\_profile\_summary +

comment\_owner\_views + comment\_owner\_upvotes + comment\_owner\_downvotes +

comment\_owner\_lastactivity\_days + editDurationAfterCreation +

activityDurationAfterCreation + title\_length + num\_tags +

PostAnswerCount + num\_favorite + hascode + post\_views + postTypeId +

IsAcceptedAnswer + postScore + post\_length + PostCommentCount +

PostUpVotes + PostDownVotes, family = "binomial", data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5474 0.4756 0.5244 0.5818 1.3509

Coefficients:

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

(Intercept) 2.121e+00 2.763e-01 7.676 1.64e-14 \*\*\*

CommentScore -3.614e-03 9.630e-03 -0.375 0.70743

CommentCrDays -1.988e-04 1.661e-04 -1.197 0.23140

CommentLength -2.490e-03 1.210e-04 -20.580 < 2e-16 \*\*\*

comment\_owner\_reputation -1.956e-07 3.463e-07 -0.565 0.57212

comment\_owner\_profile\_summary 3.013e-02 1.397e-02 2.157 0.03100 \*

comment\_owner\_views 8.864e-07 6.556e-07 1.352 0.17641

comment\_owner\_upvotes -1.002e-05 7.976e-06 -1.256 0.20920

comment\_owner\_downvotes -1.309e-05 4.714e-06 -2.776 0.00551 \*\*

comment\_owner\_lastactivity\_days 6.155e-04 4.127e-04 1.492 0.13582

editDurationAfterCreation 4.020e-04 1.085e-03 0.370 0.71114

activityDurationAfterCreation 2.744e-04 7.631e-04 0.360 0.71919

title\_length -8.666e-04 1.266e-03 -0.684 0.49380

num\_tags -3.500e-02 2.058e-02 -1.701 0.08893 .

PostAnswerCount 5.031e-02 2.184e-02 2.304 0.02124 \*

num\_favorite 2.683e-02 1.479e-02 1.814 0.06964 .

hascode -1.241e-02 6.087e-02 -0.204 0.83848

post\_views -3.564e-05 2.509e-05 -1.421 0.15538

postTypeId 3.721e-02 1.356e-01 0.274 0.78376

IsAcceptedAnswer 2.831e-01 4.191e-02 6.755 1.43e-11 \*\*\*

postScore 4.567e-02 1.187e-01 0.385 0.70036

post\_length -4.597e-06 1.009e-05 -0.456 0.64854

PostCommentCount -2.387e-02 2.712e-03 -8.802 < 2e-16 \*\*\*

PostUpVotes -4.520e-02 1.187e-01 -0.381 0.70335

PostDownVotes 5.253e-02 1.193e-01 0.440 0.65975

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31004 on 37475 degrees of freedom

AIC: 31054

Number of Fisher Scoring iterations: 4

#Checking for significant parameters with backward step

> Q3modelstepbackward=step(Q3lm, direction="backward")

> summary(Q3modelstepbackward)

#Removing insignificant parameters and remodelling

> Q3lm2=glm(as.factor(SentimentScore)~CommentLength+comment\_owner\_profile\_summary+comment\_owner\_downvotes+num\_tags+ +PostAnswerCount+num\_favorite+IsAcceptedAnswer+PostCommentCount,data=Q3Train,family="binomial")

> summary(Q3lm2)

Call:

glm(formula = as.factor(SentimentScore) ~ CommentLength + comment\_owner\_profile\_summary +

comment\_owner\_downvotes + num\_tags + +PostAnswerCount + num\_favorite +

IsAcceptedAnswer + PostCommentCount, family = "binomial",

data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4205 0.4782 0.5245 0.5807 1.3445

Coefficients:

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

(Intercept) 2.167e+00 3.651e-02 59.346 < 2e-16 \*\*\*

CommentLength -2.511e-03 1.192e-04 -21.062 < 2e-16 \*\*\*

comment\_owner\_profile\_summary 1.974e-02 1.316e-02 1.499 0.13379

comment\_owner\_downvotes -1.640e-05 4.117e-06 -3.984 6.78e-05 \*\*\*

num\_tags -4.091e-02 1.770e-02 -2.311 0.02084 \*

PostAnswerCount 3.658e-02 1.228e-02 2.980 0.00288 \*\*

num\_favorite 1.632e-02 1.014e-02 1.609 0.10760

IsAcceptedAnswer 2.849e-01 4.127e-02 6.905 5.02e-12 \*\*\*

PostCommentCount -2.385e-02 2.519e-03 -9.468 < 2e-16 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31016 on 37491 degrees of freedom

AIC: 31034

Number of Fisher Scoring iterations: 4

#Again removing insignificant parameters and remodelling

> Q3lm3=glm(as.factor(SentimentScore)~CommentLength+comment\_owner\_downvotes+num\_tags+PostAnswerCount+IsAcceptedAnswer+PostCommentCount,data=Q3Train,family="binomial")

> summary(Q3lm3)

Call:

glm(formula = as.factor(SentimentScore) ~ CommentLength + comment\_owner\_downvotes +

num\_tags + PostAnswerCount + IsAcceptedAnswer + PostCommentCount,

family = "binomial", data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2508 0.4788 0.5246 0.5804 1.3227

Coefficients:

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

(Intercept) 2.190e+00 3.055e-02 71.709 < 2e-16 \*\*\*

CommentLength -2.506e-03 1.190e-04 -21.061 < 2e-16 \*\*\*

comment\_owner\_downvotes -1.526e-05 4.049e-06 -3.769 0.000164 \*\*\*

num\_tags -4.288e-02 1.762e-02 -2.433 0.014954 \*

PostAnswerCount 4.116e-02 1.199e-02 3.433 0.000596 \*\*\*

IsAcceptedAnswer 2.838e-01 4.123e-02 6.882 5.88e-12 \*\*\*

PostCommentCount -2.334e-02 2.504e-03 -9.322 < 2e-16 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31021 on 37493 degrees of freedom

AIC: 31035

Number of Fisher Scoring iterations: 4

> BIC(Q3lm3)

[1] 31095.11

#Predicting based on the logistic model built

> logistic\_probs=predict(Q3lm3, Q3Train, type="response")

> head(logistic\_probs)

14379 39415 20449 44149 47020 2278

0.8985759 0.8931714 0.6793667 0.8320429 0.8284823 0.8087790

> testing\_y=Q3Test$SentimentScore

> logistic\_pred\_y=rep(-1,length(testing\_y))

> logistic\_pred\_y[logistic\_probs>0.8]=0

> logistic\_pred\_y[logistic\_probs>0.9]=1

> training\_y=Q3Train$SentimentScore

> table(logistic\_pred\_y,training\_y)

training\_y

logistic\_pred\_y -1 0 1

-1 349 371 724

0 4419 14239 12164

1 144 822 1339

> mean(logistic\_pred\_y!=training\_y,na.rm=TRUE)

[1] 0.5392959

> logistic\_probs=predict(Q3lm3, Q3Test, type="response")

> head(logistic\_probs)

2 8 10 15 20 23

0.8840112 0.8647405 0.7335037 0.6166388 0.7108554 0.6406810

> logistic\_pred\_y=rep(-1,length(testing\_y))

> logistic\_pred\_y[logistic\_probs>0.8]=0

> logistic\_pred\_y[logistic\_probs>0.9]=1

> table(logistic\_pred\_y,testing\_y)

testing\_y

logistic\_pred\_y -1 0 1

-1 333 396 750

0 1482 4774 4055

1 47 281 382

> mean(logistic\_pred\_y!=testing\_y,na.rm=TRUE)

[1] 0.56088

#Cross Validation for logistic regression

#K-fold

> library(boot)

> MSE\_10\_Fold\_CV=cv.glm(Q3Train,Q3lm3,K=10)$delta[1]

> MSE\_10\_Fold\_CV

[1] 0.1251676

> MSE\_10\_Fold\_CV=NULL

> MSE\_10\_Fold\_CV=NULL

> for(i in 1:10){

+ model=glm(SentimentScore~CommentLength+comment\_owner\_downvotes+num\_tags+PostAnswerCount+IsAcceptedAnswer+PostCommentCount,data=Q3Train)

+ MSE\_10\_Fold\_CV[i]=cv.glm(Q3Train,model,K=10)$delta[1]

+ }

> MSE\_10\_Fold\_CV

[1] 0.4865397 0.4865629 0.4865638 0.4865315 0.4865661 0.4865629 0.4865668 0.4865968

[9] 0.4866280 0.4865788

> library(ROCR)

> Q3smp\_size=floor(0.75\*nrow(Q3df))

> set.seed(123)

> Q3train\_ind=sample(seq\_len(nrow(Q3df)), size = Q3smp\_size)

> Q3Train=Q3df[Q3train\_ind, ]

> Q3Test=Q3df[-Q3train\_ind, ]

> nrow(Q3Train)

[1] 37500

> nrow(Q3Test)

[1] 12500

> Q3lm=glm(SentimentScore~.,data=Q3Train,family="binomial")

> summary(Q3lm)

Call:

glm(formula = SentimentScore ~ ., family = "binomial", data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5474 0.4756 0.5244 0.5818 1.3509

Coefficients:

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

(Intercept) 2.121e+00 2.763e-01 7.676 1.64e-14 \*\*\*

CommentScore -3.614e-03 9.630e-03 -0.375 0.70743

CommentCrDays -1.988e-04 1.661e-04 -1.197 0.23140

CommentLength -2.490e-03 1.210e-04 -20.580 < 2e-16 \*\*\*

comment\_owner\_reputation -1.956e-07 3.463e-07 -0.565 0.57212

comment\_owner\_profile\_summary 3.013e-02 1.397e-02 2.157 0.03100 \*

comment\_owner\_views 8.864e-07 6.556e-07 1.352 0.17641

comment\_owner\_upvotes -1.002e-05 7.976e-06 -1.256 0.20920

comment\_owner\_downvotes -1.309e-05 4.714e-06 -2.776 0.00551 \*\*

comment\_owner\_lastactivity\_days 6.155e-04 4.127e-04 1.492 0.13582

editDurationAfterCreation 4.020e-04 1.085e-03 0.370 0.71114

activityDurationAfterCreation 2.744e-04 7.631e-04 0.360 0.71919

title\_length -8.666e-04 1.266e-03 -0.684 0.49380

num\_tags -3.500e-02 2.058e-02 -1.701 0.08893 .

PostAnswerCount 5.031e-02 2.184e-02 2.304 0.02124 \*

num\_favorite 2.683e-02 1.479e-02 1.814 0.06964 .

hascode -1.241e-02 6.087e-02 -0.204 0.83848

post\_views -3.564e-05 2.509e-05 -1.421 0.15538

postTypeId 3.721e-02 1.356e-01 0.274 0.78376

IsAcceptedAnswer 2.831e-01 4.191e-02 6.755 1.43e-11 \*\*\*

postScore 4.567e-02 1.187e-01 0.385 0.70036

post\_length -4.597e-06 1.009e-05 -0.456 0.64854

PostCommentCount -2.387e-02 2.712e-03 -8.802 < 2e-16 \*\*\*

PostUpVotes -4.520e-02 1.187e-01 -0.381 0.70335

PostDownVotes 5.253e-02 1.193e-01 0.440 0.65975

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31004 on 37475 degrees of freedom

AIC: 31054

Number of Fisher Scoring iterations: 4

> Q3modelstepforward=step(Q3lm, direction="forward")

Start: AIC=31054.04

SentimentScore ~ CommentScore + CommentCrDays + CommentLength +

comment\_owner\_reputation + comment\_owner\_profile\_summary +

comment\_owner\_views + comment\_owner\_upvotes + comment\_owner\_downvotes +

comment\_owner\_lastactivity\_days + editDurationAfterCreation +

activityDurationAfterCreation + title\_length + num\_tags +

PostAnswerCount + num\_favorite + hascode + post\_views + postTypeId +

IsAcceptedAnswer + postScore + post\_length + PostCommentCount +

PostUpVotes + PostDownVotes

> summary(Q3modelstepforward)

Call:

glm(formula = SentimentScore ~ CommentScore + CommentCrDays +

CommentLength + comment\_owner\_reputation + comment\_owner\_profile\_summary +

comment\_owner\_views + comment\_owner\_upvotes + comment\_owner\_downvotes +

comment\_owner\_lastactivity\_days + editDurationAfterCreation +

activityDurationAfterCreation + title\_length + num\_tags +

PostAnswerCount + num\_favorite + hascode + post\_views + postTypeId +

IsAcceptedAnswer + postScore + post\_length + PostCommentCount +

PostUpVotes + PostDownVotes, family = "binomial", data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5474 0.4756 0.5244 0.5818 1.3509

Coefficients:

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

(Intercept) 2.121e+00 2.763e-01 7.676 1.64e-14 \*\*\*

CommentScore -3.614e-03 9.630e-03 -0.375 0.70743

CommentCrDays -1.988e-04 1.661e-04 -1.197 0.23140

CommentLength -2.490e-03 1.210e-04 -20.580 < 2e-16 \*\*\*

comment\_owner\_reputation -1.956e-07 3.463e-07 -0.565 0.57212

comment\_owner\_profile\_summary 3.013e-02 1.397e-02 2.157 0.03100 \*

comment\_owner\_views 8.864e-07 6.556e-07 1.352 0.17641

comment\_owner\_upvotes -1.002e-05 7.976e-06 -1.256 0.20920

comment\_owner\_downvotes -1.309e-05 4.714e-06 -2.776 0.00551 \*\*

comment\_owner\_lastactivity\_days 6.155e-04 4.127e-04 1.492 0.13582

editDurationAfterCreation 4.020e-04 1.085e-03 0.370 0.71114

activityDurationAfterCreation 2.744e-04 7.631e-04 0.360 0.71919

title\_length -8.666e-04 1.266e-03 -0.684 0.49380

num\_tags -3.500e-02 2.058e-02 -1.701 0.08893 .

PostAnswerCount 5.031e-02 2.184e-02 2.304 0.02124 \*

num\_favorite 2.683e-02 1.479e-02 1.814 0.06964 .

hascode -1.241e-02 6.087e-02 -0.204 0.83848

post\_views -3.564e-05 2.509e-05 -1.421 0.15538

postTypeId 3.721e-02 1.356e-01 0.274 0.78376

IsAcceptedAnswer 2.831e-01 4.191e-02 6.755 1.43e-11 \*\*\*

postScore 4.567e-02 1.187e-01 0.385 0.70036

post\_length -4.597e-06 1.009e-05 -0.456 0.64854

PostCommentCount -2.387e-02 2.712e-03 -8.802 < 2e-16 \*\*\*

PostUpVotes -4.520e-02 1.187e-01 -0.381 0.70335

PostDownVotes 5.253e-02 1.193e-01 0.440 0.65975

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31004 on 37475 degrees of freedom

AIC: 31054

Number of Fisher Scoring iterations: 4

Q3modelstepbackward=step(Q3lm, direction="backward")

summary(Q3modelstepbackward)

> Q3lm2=glm(SentimentScore~CommentLength+comment\_owner\_profile\_summary+comment\_owner\_downvotes+num\_tags+ +PostAnswerCount+num\_favorite+IsAcceptedAnswer+PostCommentCount,data=Q3Train,family="binomial")

> summary(Q3lm2)

Call:

glm(formula = SentimentScore ~ CommentLength + comment\_owner\_profile\_summary +

comment\_owner\_downvotes + num\_tags + +PostAnswerCount + num\_favorite +

IsAcceptedAnswer + PostCommentCount, family = "binomial",

data = Q3Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4205 0.4782 0.5245 0.5807 1.3445

Coefficients:

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

(Intercept) 2.167e+00 3.651e-02 59.346 < 2e-16 \*\*\*

CommentLength -2.511e-03 1.192e-04 -21.062 < 2e-16 \*\*\*

comment\_owner\_profile\_summary 1.974e-02 1.316e-02 1.499 0.13379

comment\_owner\_downvotes -1.640e-05 4.117e-06 -3.984 6.78e-05 \*\*\*

num\_tags -4.091e-02 1.770e-02 -2.311 0.02084 \*

PostAnswerCount 3.658e-02 1.228e-02 2.980 0.00288 \*\*

num\_favorite 1.632e-02 1.014e-02 1.609 0.10760

IsAcceptedAnswer 2.849e-01 4.127e-02 6.905 5.02e-12 \*\*\*

PostCommentCount -2.385e-02 2.519e-03 -9.468 < 2e-16 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 31644 on 37499 degrees of freedom

Residual deviance: 31016 on 37491 degrees of freedom

AIC: 31034

Number of Fisher Scoring iterations: 4