MidtermPDF

Michael Guel

10/28/2022

```
library(readxl)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.1.3
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
library(lubridate)
```

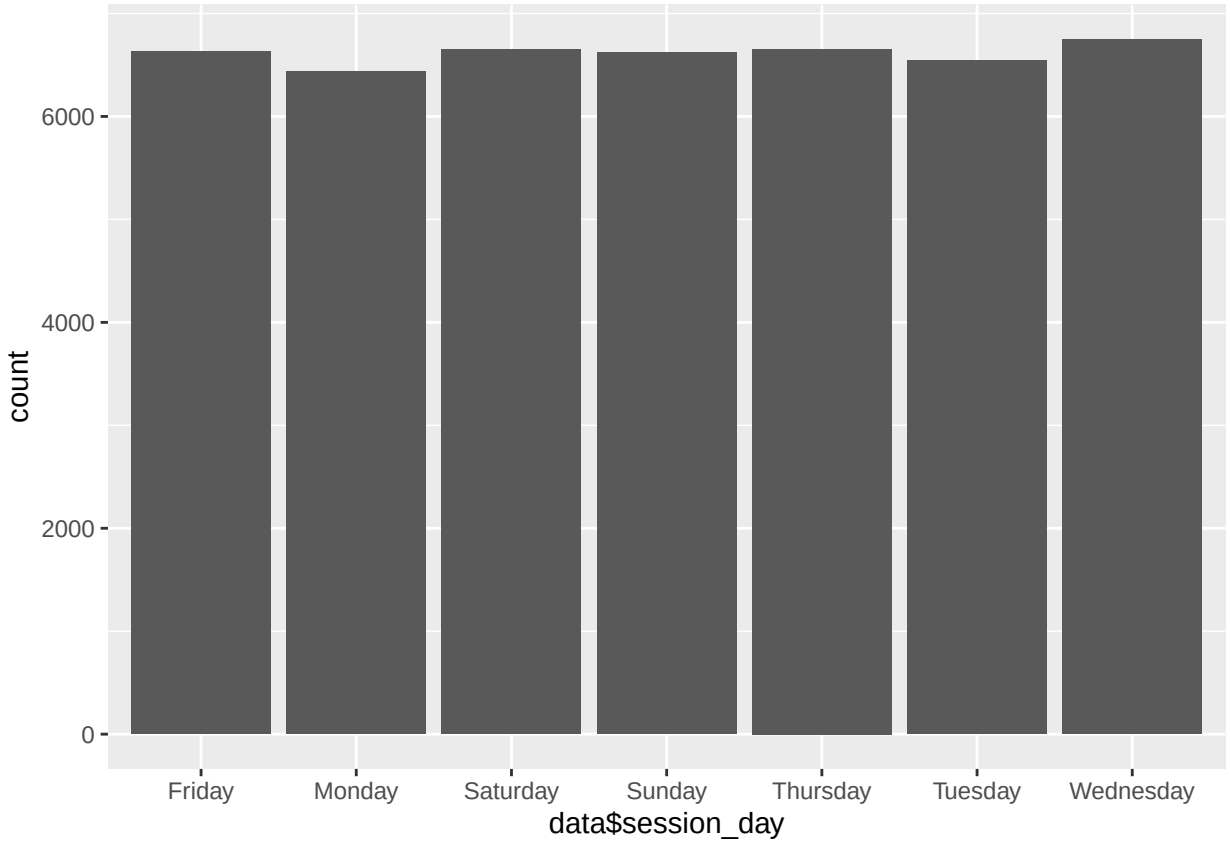
```
## Warning: package 'lubridate' was built under R version 4.1.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

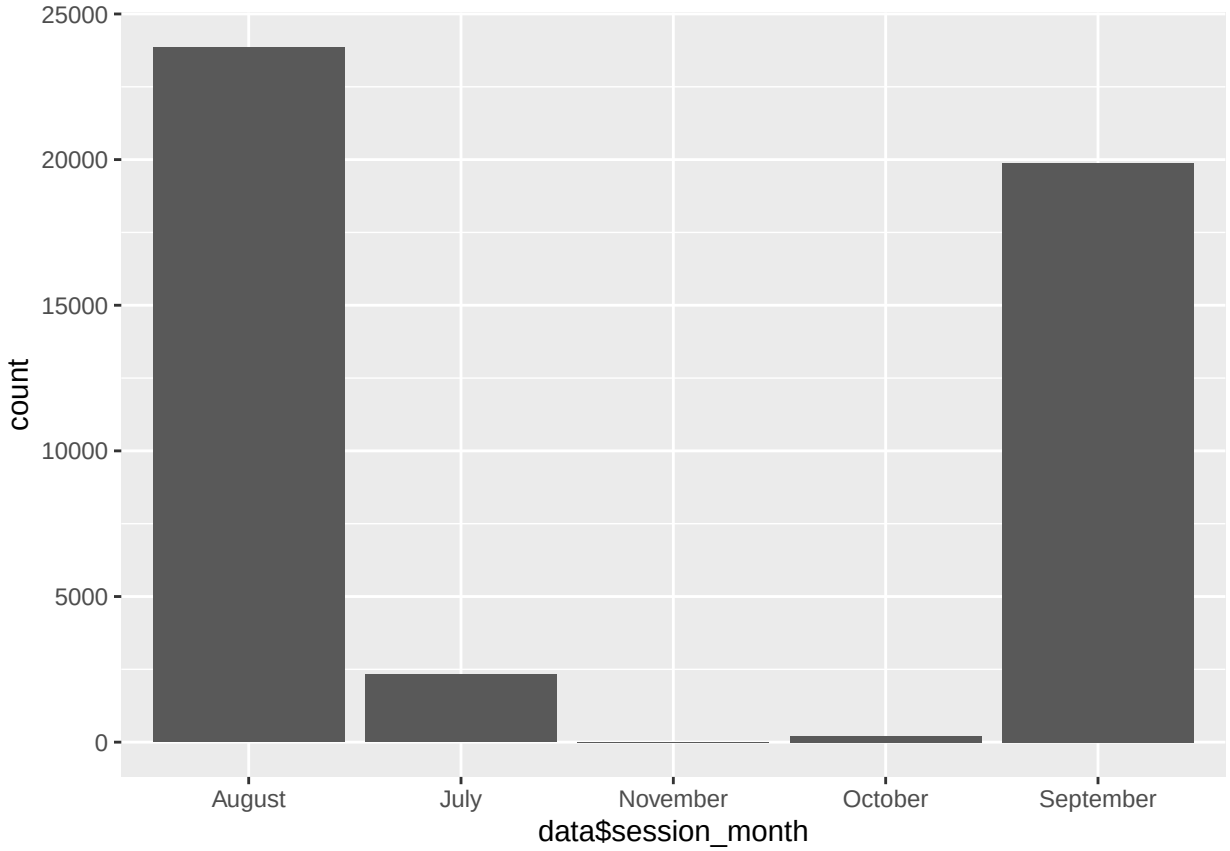
```
library(scatterplot3d)


### LOAD THE DATA IN AND MERGE

session = read.csv('sessions.csv', sep = ',')

transactions = read.csv('transactions.csv', sep = ',')

data= merge(session,transactions,by=c("session_id"))

### USE YMD TO CONVERT DATE TO DATE TYPE

data$session_dt = ymd(data$session_dt)


### ADD TWO FEATURE FOR DATE WHICH TELLS WHAT DAY AND WHAT MONTH

data = data %>% mutate(session_day = weekdays(as.Date(data$session_dt)),session_month = months(as.Date(

data$session_day = as.factor(data$session_day)

data$session_month = as.factor(data$session_month)

### EDA

ggplot(data = data, aes(data$session_day)) +
  geom_bar()
```

```
## Warning: Use of ‘data$session_day‘ is discouraged. Use ‘session_day‘ instead.
```



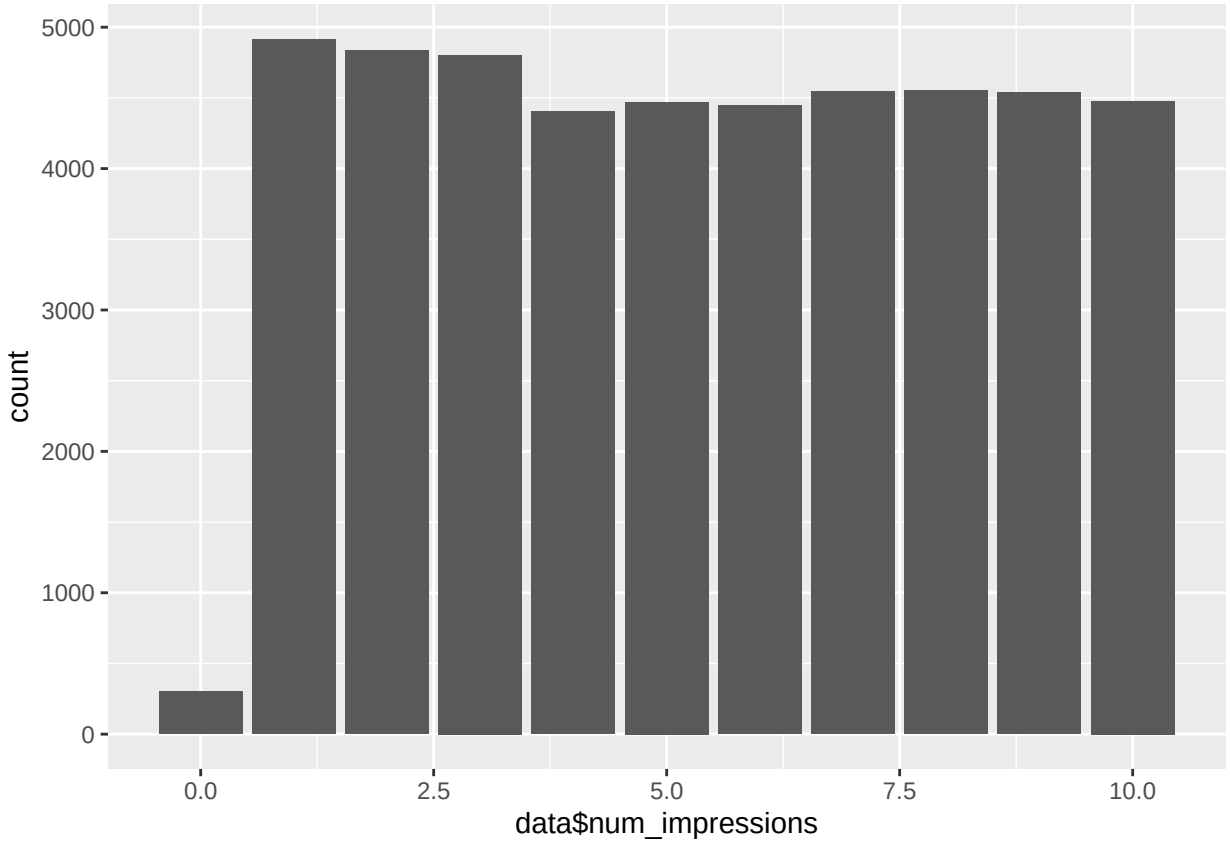

```
ggplot(data = data, aes(data$session_month)) +
  geom_bar()
```

```
## Warning: Use of ‘data$session_month‘ is discouraged. Use ‘session_month‘
## instead.
```



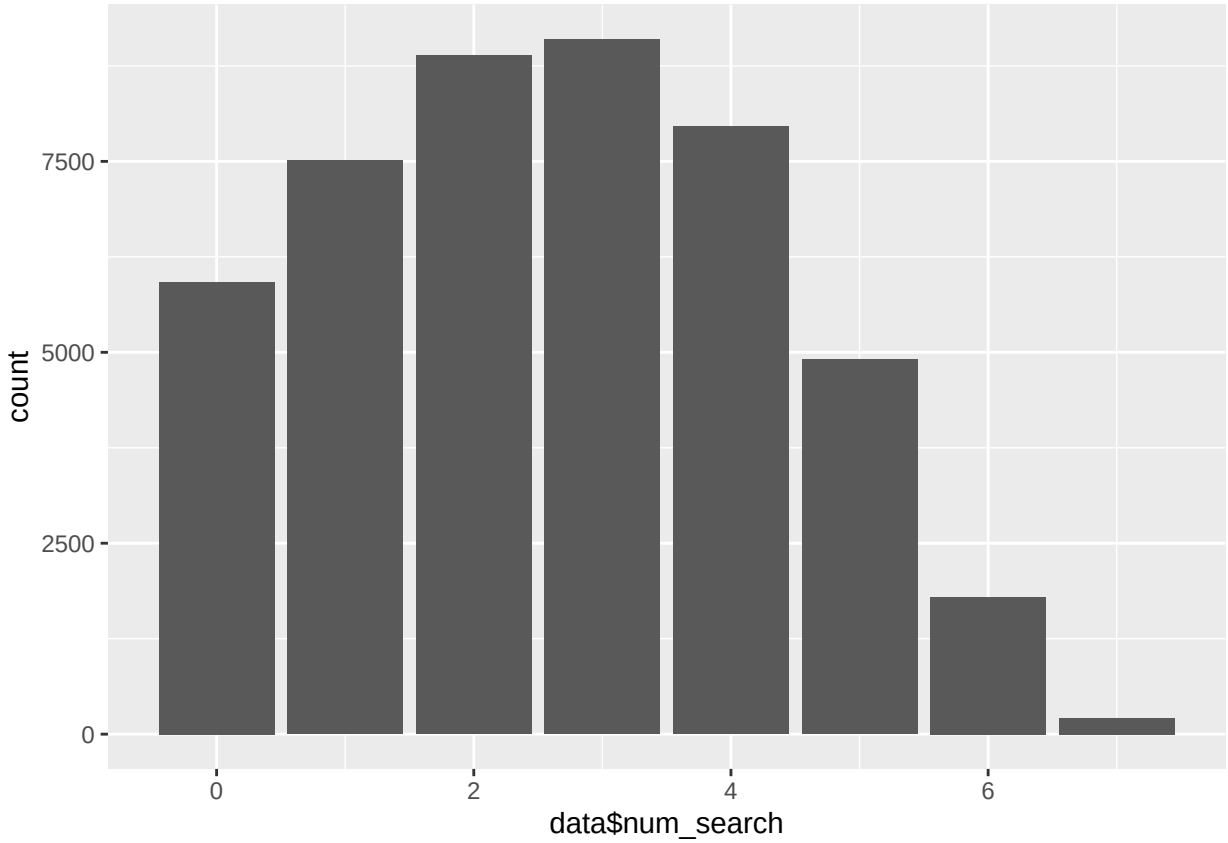

```
ggplot(data = data, aes(data$num_impressions)) +
  geom_bar()
```

```
## Warning: Use of ‘data$num_impressions‘ is discouraged. Use ‘num_impressions‘
## instead.
```




```
ggplot(data = data, aes(data$num_search)) +
  geom_bar()
```

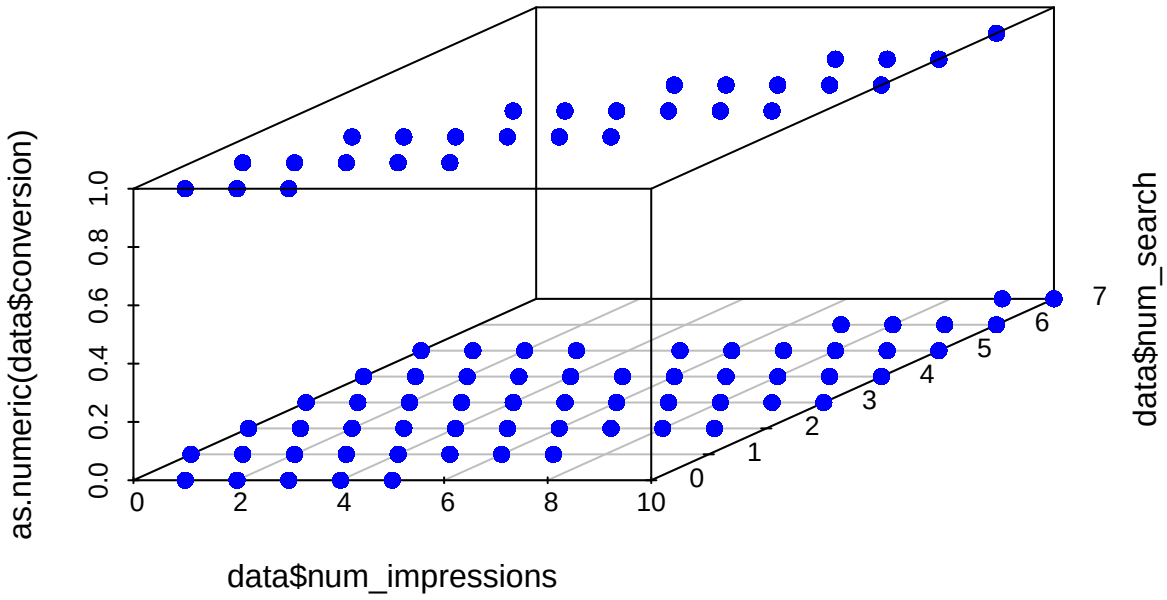
```
## Warning: Use of ‘data$num_search‘ is discouraged. Use ‘num_search‘ instead.
```




```
totconversion = data %>% group_by(conversion) %>% summarise(count = n())

howmanysessions = data %>% group_by(user_id) %>% summarise(count = n())

unique(howmanysessions$count)
```

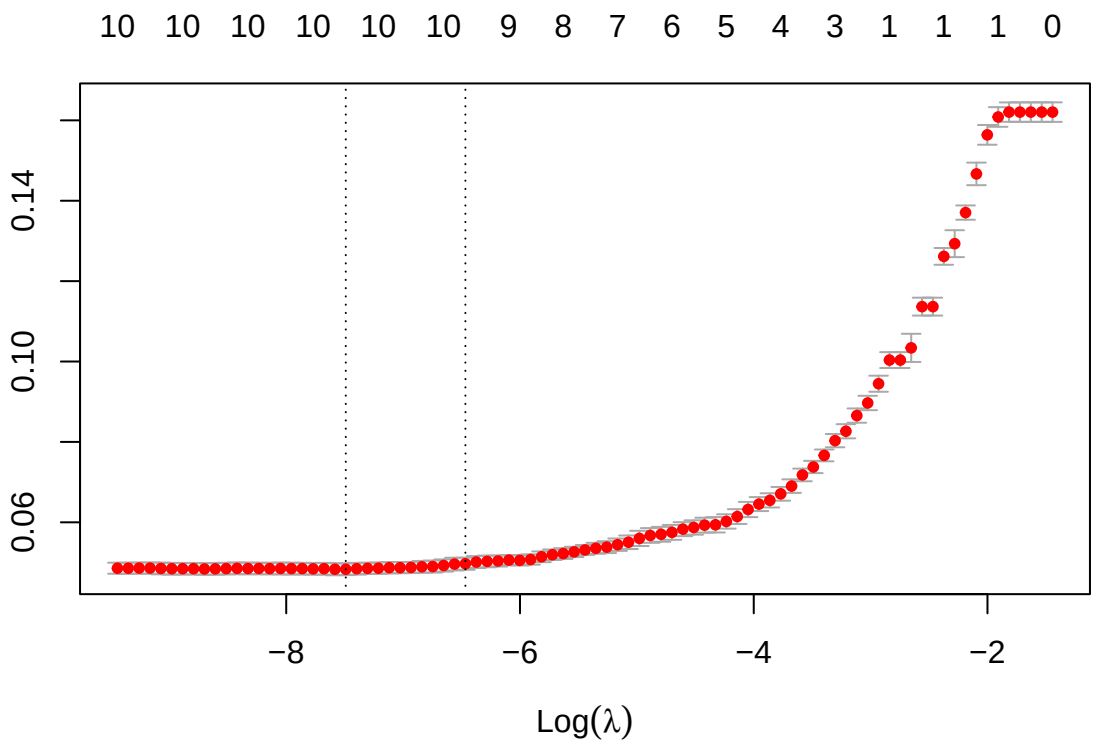
```
## [1] 3 4 5 6
```

```
scatterplot3d(data$num_impressions, data$num_search, as.numeric(data$conversion), pch = 19, color = "blu
```




```
### ADD RANKING FEATURE FOR SEVERAL FEATURES

nexttr = data %>% group_by(user_id) %>% arrange(session_dt) %>% mutate(rank = rank(session_dt,ties.metho

nexttr = nexttr %>% group_by(user_id) %>% mutate(prevrel = ifelse(rank == 1,0,ifelse(rank == 2, avg_relev

nexttr = nexttr %>% group_by(user_id) %>% mutate(prevsearch = ifelse(rank == 1,0,ifelse(rank == 2, num_s

nexttr = nexttr %>% group_by(user_id) %>% mutate(previmp = ifelse(rank == 1,0,ifelse(rank == 2, num_impr

nexttr = nexttr %>% mutate(b4thisses = (imp2sess - num_impressions))

nexttr = nexttr %>% mutate(searchb4 = (num2ses - num_search))

nexttr$prevrel = as.numeric(nexttr$prevrel)

nexttr$prevsearch = as.numeric(nexttr$prevsearch)

nexttr$previmp = as.numeric(nexttr$previmp)

### DROP FEATURES THAT ARE OF NO VALUE AND NOT BEING USED
```

```
use = nexttr[,!(colnames(nexttr)%in% c("session_id","session_dt","user_id","avgrelacross","totsearch","a

### Split into train, validation and test

train = use[use$train == TRUE,]

train = train[,!(colnames(train)%in% c("train","score","test"))]

validation = use[use$score == TRUE,]

validation = validation[,!(colnames(validation)%in% c("train","score","test"))]

test = use[use$test == TRUE,]

test = test[,!(colnames(test)%in% c("train","score","test"))]

### CREATE MATRIX FOR FEATURES AND PULL TARGET VARIABLE INTO ITS OWN DATAFRAME

x = model.matrix(conversion ~ ., train)[,-1]

y = as.factor(train$conversion)


### FIT MODEL USING CV AND L1 PENALTY

fit_ridge_cv = cv.glmnet(x, y, alpha = 1,family = "binomial",type.measure = "class")
plot(fit_ridge_cv)
```




```
coef(fit_ridge_cv,s = "lambda.min")
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                 s1
## (Intercept)            -23.87421763
## num_impressions          0.29675549
## avg_relevance            6.47144354
## num_search              -0.21340286
## session_monthJuly       -2.14653932
## session_monthNovember   -7.42853052
## session_monthOctober   -10.23620813
## session_monthSeptember  -0.09768519
## rank                     2.57566389
## imp2sess                 0.79607007
## num2ses                 -0.34527211
```

```
coef(fit_ridge_cv,s = "lambda.1se")
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                 s1
## (Intercept)            -21.58971152
## num_impressions          0.25822492
## avg_relevance            5.83868398
## num_search              -0.19027101
```

```
## session_monthJuly       -0.88948354
## session_monthNovember   -6.01963371
## session_monthOctober    -8.09907818
## session_monthSeptember  -0.03401194
## rank                     2.25039755
## imp2sess                 0.69466985
## num2ses                 -0.23339529
```

```
### PREDICT AND CHECK FOR ACCURACY TRAIN DATA

xtrain = model.matrix(conversion ~ ., train)[,-1]

ne = data.frame(predict(fit_ridge_cv,xtrain,type = "class", s="lambda.min"))

see = data.frame(train$conversion)

see$pred = ne$lambda.min

table(see$pred,see$train.conversion)
```

```
##
##         FALSE  TRUE
##   FALSE 25571   867
##   TRUE    646  4203
```

```
### CALCULATE PREDICTION STATISTICS

cm <- table(see$pred, see$train.conversion)

accuracy <- sum(cm[1], cm[4]) / sum(cm[1:4])
precision <- cm[4] / sum(cm[4], cm[2])
sensitivity <- cm[4] / sum(cm[4], cm[3])
fscore <- (2 * (sensitivity * precision))/(sensitivity + precision)
specificity <- cm[1] / sum(cm[1], cm[2])

accuracy
```

```
## [1] 0.9516413
```

```
precision
```

```
## [1] 0.8667767
```

```
sensitivity
```

```
## [1] 0.8289941
```

```
fscore
```

```
## [1] 0.8474645
```

```
specificity
```

```
## [1] 0.9753595
```

```
### PREDICT AND CHECK ACCURACY

xval = model.matrix(conversion ~ ., validation)[,-1]

ne = data.frame(predict(fit_ridge_cv,xval,type = "class", s="lambda.min"))

see = data.frame(validation$conversion)

see$pred = ne$lambda.min

table(see$pred,see$validation.conversion)
```

```
##
##         FALSE  TRUE
##   FALSE 12384   424
##   TRUE    186  2006
```

```
### CALCULATE PREDICTION STATISTICS

cm <- table(see$pred, see$validation.conversion)

accuracy <- sum(cm[1], cm[4]) / sum(cm[1:4])
precision <- cm[4] / sum(cm[4], cm[2])
sensitivity <- cm[4] / sum(cm[4], cm[3])
fscore <- (2 * (sensitivity * precision))/(sensitivity + precision)
specificity <- cm[1] / sum(cm[1], cm[2])

accuracy
```

```
## [1] 0.9593333
```

```
precision
```

```
## [1] 0.915146
```

```
sensitivity
```

```
## [1] 0.8255144
```

```
fscore
```

```
## [1] 0.8680225
```

```
specificity
```

```
## [1] 0.9852029
```

```
### PREDICT AND CHECK ACCURACY ON TEST DATA

xtest = model.matrix(conversion ~ ., test)[,-1]

ne = data.frame(predict(fit_ridge_cv,xtest,type = "class", s="lambda.min"))

see = data.frame(test$conversion)

see$pred = ne$lambda.min

table(see$pred,see$test.conversion)
```

```
##
##         FALSE TRUE
##   FALSE  2470  424
##   TRUE    100 2006
```

```
### CALCULATE PREDICTION STATISTICS

cm <- table(see$pred, see$test.conversion)

accuracy <- sum(cm[1], cm[4]) / sum(cm[1:4])
precision <- cm[4] / sum(cm[4], cm[2])
sensitivity <- cm[4] / sum(cm[4], cm[3])
fscore <- (2 * (sensitivity * precision))/(sensitivity + precision)
specificity <- cm[1] / sum(cm[1], cm[2])

accuracy
```

```
## [1] 0.8952
```

```
precision
```

```
## [1] 0.9525166
```

```
sensitivity
```

```
## [1] 0.8255144
```

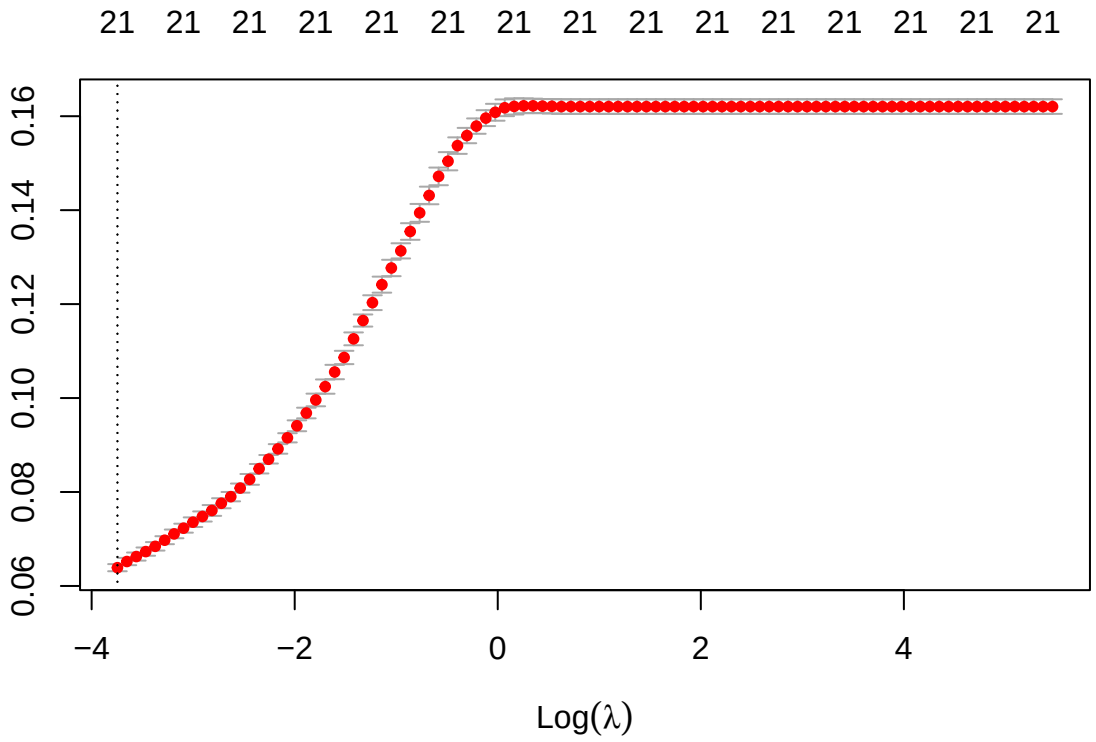
```
fscore
```

```
## [1] 0.8844797
```

```
specificity
```

```
## [1] 0.9610895
```

```
### COMPARE TO RIDGE REGRESSION
### FIT MODEL USING CV AND L2 PENALTY

use = nexttr[,!(colnames(nexttr)%in% c("session_id","session_dt","user_id"))]


### Split into train, validation and test

train = use[use$train == TRUE,]

train = train[,!(colnames(train)%in% c("train","score","test"))]

validation = use[use$score == TRUE,]

validation = validation[,!(colnames(validation)%in% c("train","score","test"))]

test = use[use$test == TRUE,]

test = test[,!(colnames(test)%in% c("train","score","test"))]


### CREATE MATRIX FOR FEATURES AND PULL TARGET VARIABLE INTO ITS OWN DATAFRAME

x = model.matrix(conversion ~ ., train)[,-1]

y = as.factor(train$conversion)

fit_ridge_cv = cv.glmnet(x, y, alpha = 0,family = "binomial",type.measure = "class")
plot(fit_ridge_cv)
```




```
coef(fit_ridge_cv,s = "lambda.min")
```

```
## 22 x 1 sparse Matrix of class "dgCMatrix"
##                                      s1
## (Intercept)             -10.711373572
## num_impressions           0.214240304
## avg_relevance             2.710532109
## num_search               -0.026915621
## session_dayMonday        -0.001526865
## session_daySaturday       0.035605375
## session_daySunday        -0.047022567
## session_dayThursday       0.059461677
## session_dayTuesday       -0.015620058
## session_dayWednesday     -0.008202798
## session_monthJuly        -0.674642977
## session_monthNovember    -4.776626929
## session_monthOctober     -3.670439849
## session_monthSeptember   -0.040613281
## rank                      0.727445438
## imp2sess                  0.146908669
## num2ses                   0.081202507
## prevrel                   0.281116712
## prevsearch               -0.053205885
## previmp                  -0.004763806
## b4thisses                 0.124708855
## searchb4                  0.113196483
```

```
coef(fit_ridge_cv,s = "lambda.1se")
```

```
## 22 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)             -10.711373572
## num_impressions           0.214240304
## avg_relevance             2.710532109
## num_search               -0.026915621
## session_dayMonday        -0.001526865
## session_daySaturday       0.035605375
## session_daySunday        -0.047022567
## session_dayThursday       0.059461677
## session_dayTuesday       -0.015620058
## session_dayWednesday     -0.008202798
## session_monthJuly        -0.674642977
## session_monthNovember    -4.776626929
## session_monthOctober     -3.670439849
## session_monthSeptember   -0.040613281
## rank                      0.727445438
## imp2sess                  0.146908669
## num2ses                   0.081202507
## prevrel                   0.281116712
## prevsearch               -0.053205885
## previmp                  -0.004763806
## b4thisses                 0.124708855
## searchb4                  0.113196483
```

```
### PREDICT AND CHECK FOR ACCURACY TEST DATA

xtest = model.matrix(conversion ~ ., test)[,-1]

ne = data.frame(predict(fit_ridge_cv,xtest,type = "class", s="lambda.min"))

see = data.frame(test$conversion)

see$pred = ne$lambda.min

table(see$pred,see$test.conversion)
```

```
##
##         FALSE TRUE
##   FALSE  2515  722
##   TRUE     55 1708
```

```
### CALCULATE PREDICTION STATISTICS

cm <- table(see$pred, see$test.conversion)

accuracy <- sum(cm[1], cm[4]) / sum(cm[1:4])
precision <- cm[4] / sum(cm[4], cm[2])
sensitivity <- cm[4] / sum(cm[4], cm[3])
fscore <- (2 * (sensitivity * precision))/(sensitivity + precision)
```

```r
specificity <- cm[1] / sum(cm[1], cm[2])

accuracy
```

```
## [1] 0.8446
```

```r
precision
```

```
## [1] 0.9688032
```

```r
sensitivity
```

```
## [1] 0.7028807
```

```r
fscore
```

```
## [1] 0.8146912
```

```r
specificity
```

```
## [1] 0.9785992
```