NYC MTA Subway Data - Feature Selection (Regression)

Shraddha Somani

```
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(caTools)
```

Load the data

```
df_train = read.csv("C:/Users/Shraddha Somani/nyc_mta_train.csv")
head(df_train)
```

```
##    C.A UNIT      SCP         STATION LINENAME DIVISION       DATE     TIME
## 1 R203 R043 00-05-03         WALL ST       45      IRT 02/21/2016 07:13:52
## 2 R644 R135 01-00-01      NEWKIRK AV       25      IRT 09/08/2016 17:00:00
## 3 N319 R298 01-06-01   NORTHERN BLVD       MR      IND 01/26/2016 08:00:00
## 4 N102 R127 01-06-02 JAY ST-METROTEC      ACF      IND 11/11/2016 11:00:00
## 5 R409 R449 01-00-01        E 149 ST        6      IRT 09/28/2016 17:00:00
## 6 R334 R367 00-00-02          233 ST       25      IRT 11/24/2016 04:00:00
##      DESC  ENTRIES   EXITS            DATETIME YEAR MONTH DAY   WEEKDAY
## 1 REGULAR 11849652 1770727 2016-02-21 07:13:52 2016     2  21    Sunday
## 2 REGULAR   217900 1742518 2016-09-08 17:00:00 2016     9   8  Thursday
## 3 REGULAR  1913887  909905 2016-01-26 08:00:00 2016     1  26   Tuesday
## 4 REGULAR  5199304   96659 2016-11-11 11:00:00 2016    11  11    Friday
## 5 REGULAR  2078806 2720360 2016-09-28 17:00:00 2016     9  28 Wednesday
## 6 REGULAR 84611779  151407 2016-11-24 04:00:00 2016    11  24  Thursday
##   HOUR TOTAL_COUNT
## 1    7    13620379
## 2   17     1960418
## 3    8     2823792
## 4   11     5295963
## 5   17     4799166
## 6    4    84763186
```

```
df_train_pca = df_train[c(-7,-8,-11,-12,-13,-18)]
head(df_train_pca)
```

```
##    C.A UNIT      SCP         STATION LINENAME DIVISION    DESC  ENTRIES
## 1 R203 R043 00-05-03         WALL ST       45      IRT REGULAR 11849652
## 2 R644 R135 01-00-01      NEWKIRK AV       25      IRT REGULAR   217900
## 3 N319 R298 01-06-01   NORTHERN BLVD       MR      IND REGULAR  1913887
## 4 N102 R127 01-06-02 JAY ST-METROTEC      ACF      IND REGULAR  5199304
## 5 R409 R449 01-00-01        E 149 ST        6      IRT REGULAR  2078806
## 6 R334 R367 00-00-02          233 ST       25      IRT REGULAR 84611779
##   MONTH DAY   WEEKDAY HOUR
## 1     2  21    Sunday    7
## 2     9   8  Thursday   17
## 3     1  26   Tuesday    8
## 4    11  11    Friday   11
## 5     9  28 Wednesday   17
## 6    11  24  Thursday    4
```

```
sample = sample.split(df_train_pca$ENTRIES, 2/3)
train = df_train_pca[sample,]
test = df_train_pca[!sample,]
str(train)
```

```
## 'data.frame':    4459982 obs. of  12 variables:
##  $ C.A     : Factor w/ 699 levels "A002","A006",..: 504 687 232 527 490 559 423 516 29 381
...
##  $ UNIT    : Factor w/ 448 levels "R001","R003",..: 37 119 111 157 252 222 445 141 74 268 ...
##  $ SCP     : Factor w/ 210 levels "00-00-00","00-00-01",..: 50 67 103 67 3 2 88 3 30 1 ...
##  $ STATION : Factor w/ 357 levels "1 AV","103 ST",..: 346 288 241 55 15 24 284 115 182 214
...
##  $ LINENAME: Factor w/ 120 levels "1","123","1237ACENQRS",..: 23 19 50 32 1 22 1 32 114 82
...
##  $ DIVISION: Factor w/ 6 levels "BMT","IND","IRT",..: 3 3 2 3 3 3 4 3 1 2 ...
##  $ DESC    : Factor w/ 2 levels "RECOVR AUD","REGULAR": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ENTRIES : int  11849652 217900 5199304 4540089 9056387 2804576 747017 15126457 6324268 380
2964 ...
##  $ MONTH   : int  2 9 11 9 4 12 1 6 10 10 ...
##  $ DAY     : int  21 8 11 10 21 11 24 17 1 10 ...
##  $ WEEKDAY : Factor w/ 7 levels "Friday","Monday",..: 4 5 1 3 5 4 4 1 3 2 ...
##  $ HOUR    : int  7 17 11 13 16 19 0 5 5 1 ...
```

Convert entire dataset to numeric value

```
train$C.A = as.factor(train$C.A) %>% as.numeric()
train$UNIT = as.factor(train$UNIT) %>% as.numeric()
train$SCP = as.factor(train$SCP) %>% as.numeric()
train$STATION = as.factor(train$STATION) %>% as.numeric()
train$LINENAME = as.factor(train$LINENAME) %>% as.numeric()
train$DIVISION = as.factor(train$DIVISION) %>% as.numeric()
train$DESC = as.factor(train$DESC) %>% as.numeric()
train$ENTRIES = as.integer(train$ENTRIES) %>% as.numeric()
train$MONTH = as.integer(train$MONTH) %>% as.numeric()
train$WEEKDAY = as.factor(train$WEEKDAY) %>% as.numeric()
train$HOUR = as.integer(train$HOUR) %>% as.numeric()
head(train)
```

```
##   C.A UNIT SCP STATION LINENAME DIVISION DESC  ENTRIES MONTH DAY WEEKDAY
## 1 504   37  50     346       23        3    2 11849652     2  21       4
## 2 687  119  67     288       19        3    2   217900     9   8       5
## 4 232  111 103     241       50        2    2  5199304    11  11       1
## 7 527  157  67      55       32        3    2  4540089     9  10       3
## 8 490  252   3      15        1        3    2  9056387     4  21       5
## 9 559  222   2      24       22        3    2  2804576    12  11       4
##   HOUR
## 1    7
## 2   17
## 4   11
## 7   13
## 8   16
## 9   19
```

Multi Linear Regression

```
df_fit = lm(ENTRIES ~ ., data = train)
summary(df_fit)
```

```
## 
## Call:
## lm(formula = ENTRIES ~ ., data = train)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
##  -69828530  -44022410  -28996736  -13961855 2115074501 
## 
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)    
## (Intercept) 89839331.0  2918686.6   30.781  < 2e-16 ***
## C.A           -25925.0      824.4  -31.449  < 2e-16 ***
## UNIT          -41973.5      730.8  -57.431  < 2e-16 ***
## SCP            41034.8     1995.2   20.567  < 2e-16 ***
## STATION      -137578.3      934.8 -147.176  < 2e-16 ***
## LINENAME     -167485.2     4646.0  -36.049  < 2e-16 ***
## DIVISION    -3389363.6   233078.4  -14.542  < 2e-16 ***
## DESC        -1174000.4  1422189.4   -0.825    0.409    
## MONTH         120726.8    26434.4    4.567 4.95e-06 ***
## DAY             8499.4    10372.1    0.819    0.413    
## WEEKDAY       -27410.6    45389.0   -0.604    0.546    
## HOUR          -10690.6    13220.6   -0.809    0.419    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 192300000 on 4459970 degrees of freedom
## Multiple R-squared:  0.008477,   Adjusted R-squared:  0.008475 
## F-statistic:  3467 on 11 and 4459970 DF,  p-value: < 2.2e-16
```

Step Fit

```
fit_step = step(df_fit)
```

```
## Start:  AIC=170146648
## ENTRIES ~ C.A + UNIT + SCP + STATION + LINENAME + DIVISION +
##     DESC + MONTH + DAY + WEEKDAY + HOUR
##
##            Df  Sum of Sq        RSS       AIC
## - WEEKDAY   1 1.3493e+16 1.6501e+23 170146647
## - HOUR      1 2.4192e+16 1.6501e+23 170146647
## - DAY       1 2.4843e+16 1.6501e+23 170146647
## - DESC      1 2.5211e+16 1.6501e+23 170146647
## <none>                   1.6501e+23 170146648
## - MONTH     1 7.7169e+17 1.6501e+23 170146667
## - DIVISION  1 7.8235e+18 1.6502e+23 170146858
## - SCP       1 1.5650e+19 1.6502e+23 170147069
## - C.A       1 3.6592e+19 1.6504e+23 170147635
## - LINENAME  1 4.8079e+19 1.6506e+23 170147946
## - UNIT      1 1.2203e+20 1.6513e+23 170149943
## - STATION   1 8.0139e+20 1.6581e+23 170168255
##
## Step:  AIC=170146647
## ENTRIES ~ C.A + UNIT + SCP + STATION + LINENAME + DIVISION +
##     DESC + MONTH + DAY + HOUR
##
##            Df  Sum of Sq        RSS       AIC
## - HOUR      1 2.4150e+16 1.6501e+23 170146645
## - DAY       1 2.4320e+16 1.6501e+23 170146645
## - DESC      1 2.5256e+16 1.6501e+23 170146645
## <none>                   1.6501e+23 170146647
## - MONTH     1 7.7018e+17 1.6501e+23 170146665
## - DIVISION  1 7.8248e+18 1.6502e+23 170146856
## - SCP       1 1.5650e+19 1.6502e+23 170147068
## - C.A       1 3.6590e+19 1.6504e+23 170147634
## - LINENAME  1 4.8081e+19 1.6506e+23 170147944
## - UNIT      1 1.2203e+20 1.6513e+23 170149942
## - STATION   1 8.0139e+20 1.6581e+23 170168253
##
## Step:  AIC=170146645
## ENTRIES ~ C.A + UNIT + SCP + STATION + LINENAME + DIVISION +
##     DESC + MONTH + DAY
##
##            Df  Sum of Sq        RSS       AIC
## - DAY       1 2.4506e+16 1.6501e+23 170146644
## - DESC      1 2.5849e+16 1.6501e+23 170146644
## <none>                   1.6501e+23 170146645
## - MONTH     1 7.7261e+17 1.6501e+23 170146664
## - DIVISION  1 7.8357e+18 1.6502e+23 170146855
## - SCP       1 1.5649e+19 1.6502e+23 170147066
## - C.A       1 3.6571e+19 1.6504e+23 170147632
## - LINENAME  1 4.8079e+19 1.6506e+23 170147943
## - UNIT      1 1.2201e+20 1.6513e+23 170149940
## - STATION   1 8.0142e+20 1.6581e+23 170168253
##
## Step:  AIC=170146644
## ENTRIES ~ C.A + UNIT + SCP + STATION + LINENAME + DIVISION +
##     DESC + MONTH
## 
##            Df  Sum of Sq        RSS       AIC
## - DESC      1 2.5921e+16 1.6501e+23 170146643
## <none>                   1.6501e+23 170146644
## - MONTH     1 7.7194e+17 1.6501e+23 170146663
## - DIVISION  1 7.8362e+18 1.6502e+23 170146854
## - SCP       1 1.5649e+19 1.6502e+23 170147065
## - C.A       1 3.6571e+19 1.6504e+23 170147630
## - LINENAME  1 4.8078e+19 1.6506e+23 170147941
## - UNIT      1 1.2201e+20 1.6513e+23 170149939
## - STATION   1 8.0142e+20 1.6581e+23 170168251
## 
## Step:  AIC=170146643
## ENTRIES ~ C.A + UNIT + SCP + STATION + LINENAME + DIVISION + 
##     MONTH
## 
##            Df  Sum of Sq        RSS       AIC
## <none>                   1.6501e+23 170146643
## - MONTH     1 7.6872e+17 1.6501e+23 170146661
## - DIVISION  1 7.8373e+18 1.6502e+23 170146853
## - SCP       1 1.5649e+19 1.6502e+23 170147064
## - C.A       1 3.6571e+19 1.6504e+23 170147629
## - LINENAME  1 4.8078e+19 1.6506e+23 170147940
## - UNIT      1 1.2201e+20 1.6513e+23 170149937
## - STATION   1 8.0143e+20 1.6581e+23 170168250
```