# Exploratory data analysis for Predicting Hospital Readmissions

*Youliang Yu*

*Nov. 11, 2016*

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(data.table)
cat("load data")
```

```
## load data
```

```
train <- fread('data/Challenge_1_Training.csv',header =TRUE,stringsAsFactors = FALSE,na.strings=c("?","
```

```
## readimitted
```

```
## [1] ">30" NA    "<30"
```

```
## There are NA labels involved, remove them first.
```

```
## check fraction of patients get admitted >30 days
```

```
## [1] 0.756036
```

```
## About 3 quarters, no problem.
```

```
## Check levels for each variables since most variables are discrete.
```

```
##                     race                   gender                      age
##                        6                        2                       10
##                   weight       admission_type_id discharge_disposition_id
##                       10                        7                       22
##      admission_source_id        time_in_hospital               payer_code
##                       12                       14                       17
##        medical_specialty      num_lab_procedures           num_procedures
##                       61                      109                        7
##          num_medications       number_outpatient         number_emergency
##                       70                       31                       25
##         number_inpatient                  diag_1                   diag_2
##                       21                      553                      540
##                   diag_3        number_diagnoses            max_glu_serum
##                      570                       16                        4
```

```
##                A1Cresult                metformin              repaglinide
##                        4                        4                        4
##             nateglinide           chlorpropamide             glimepiride
##                        4                        3                        4
##            acetohexamide                glipizide                glyburide
##                        2                        4                        4
##             tolbutamide            pioglitazone            rosiglitazone
##                        2                        4                        4
##                acarbose                miglitol            troglitazone
##                        3                        4                        2
##             tolazamide                  examide             citoglipton
##                        2                        1                        1
##                 insulin      glyburide.metformin      glipizide.metformin
##                        4                        2                        2
## glimepiride.pioglitazone  metformin.rosiglitazone   metformin.pioglitazone
##                        1                        1                        1
##                  change              diabetesMed              readmitted
##                        2                        2                        2
```

```
## Diagnosis 1/2/3 in train intersect with test set since only 500 out of 800/900 appears in train...
```

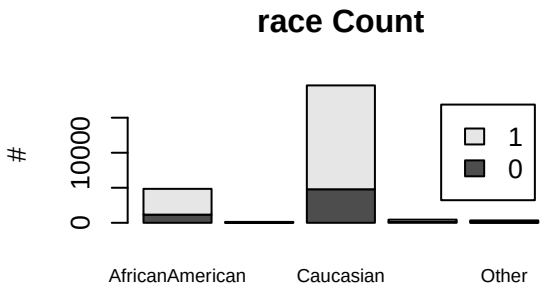
```
## check NA fraction in each var
```

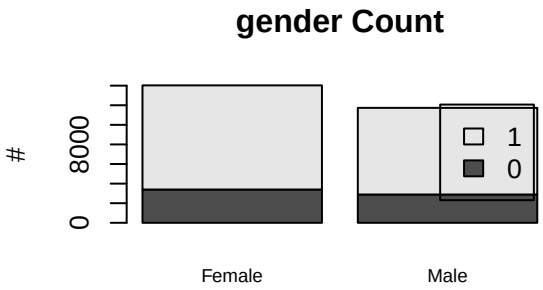
```
##                     race                   gender                      age
##               1.51385762               0.00000000               0.00000000
##                   weight        admission_type_id discharge_disposition_id
##              96.09890536               0.00000000               0.00000000
##      admission_source_id         time_in_hospital               payer_code
##               0.00000000               0.00000000              39.33312631
##        medical_specialty       num_lab_procedures           num_procedures
##              51.59149134               0.00000000               0.00000000
##          num_medications        number_outpatient         number_emergency
##               0.00000000               0.00000000               0.00000000
##         number_inpatient                   diag_1                   diag_2
##               0.00000000               0.01552674               0.17855757
##                   diag_3         number_diagnoses            max_glu_serum
##               0.85008928               0.00000000              94.46859716
##                A1Cresult                metformin              repaglinide
##              83.85994876              81.65126931              98.30370313
##              nateglinide           chlorpropamide             glimepiride
##              99.21978107              99.91460290              94.72090676
##             acetohexamide                glipizide                glyburide
##              99.99611831              86.66252620              89.70188650
##              tolbutamide            pioglitazone            rosiglitazone
##              99.99223663              92.21333747              93.39337008
##                 acarbose                miglitol            troglitazone
##              99.62347644              99.94177471              99.99223663
##              tolazamide                  examide             citoglipton
##              99.97670988             100.00000000             100.00000000
##                  insulin      glyburide.metformin      glipizide.metformin
##              44.06102011              99.32846829              99.98059157
## glimepiride.pioglitazone  metformin.rosiglitazone   metformin.pioglitazone
##             100.00000000             100.00000000             100.00000000
```

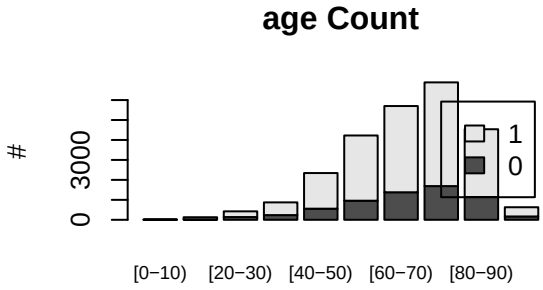
```
##                     change             diabetesMed              readmitted
##              50.90443289             20.08772611              0.00000000
```

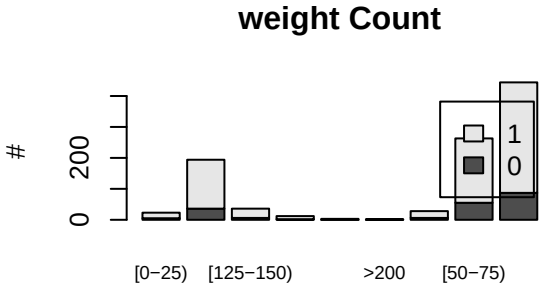
```
## For almost all 24 features for medications, over 90% of them are NAs, 6 of them gives 100% NAs, each
```

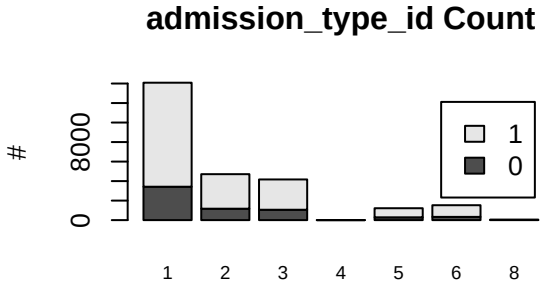
```
## Visualize other features
```

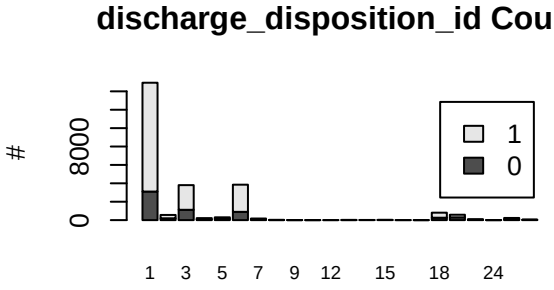


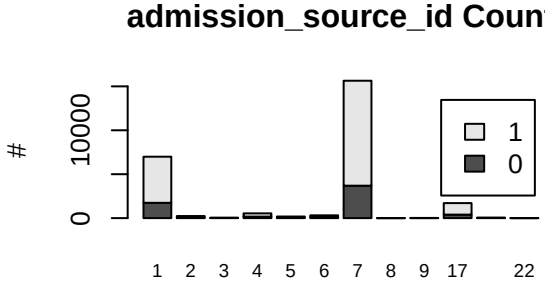



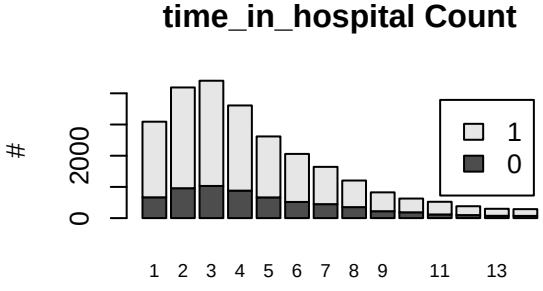



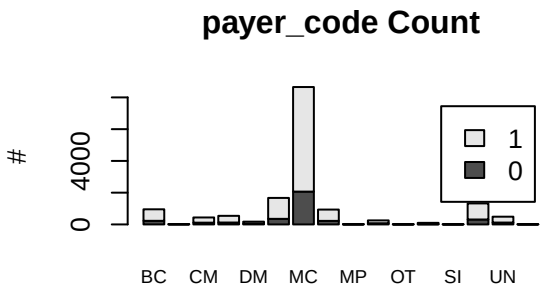



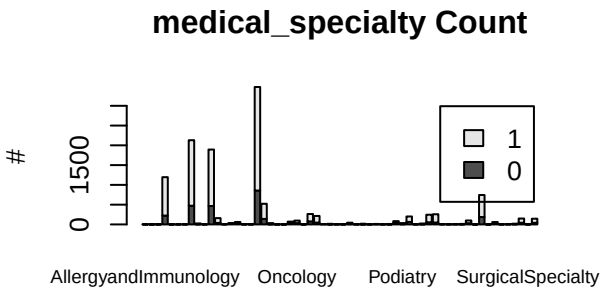



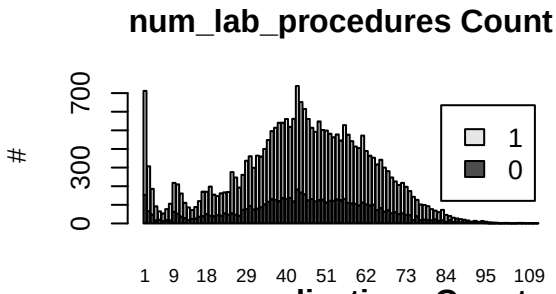



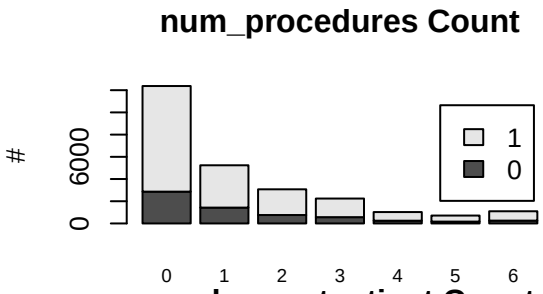



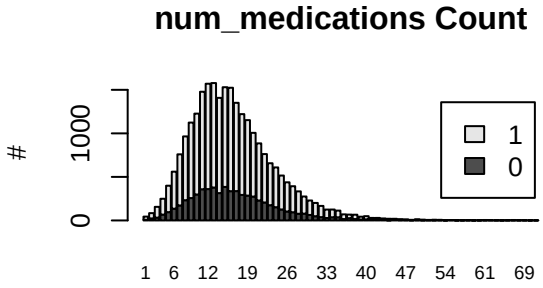



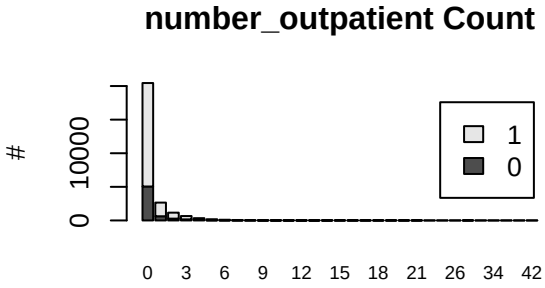

payer_code Count
#
0
4000
1
0
BC
CM
DM
MC
MP
OT
SI
UN

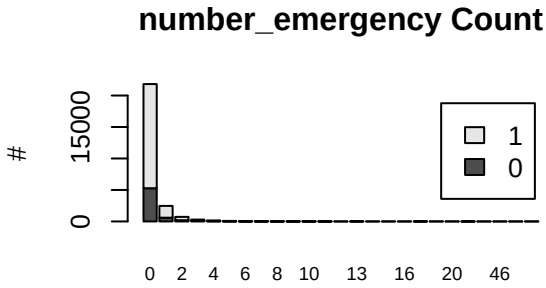

medical_specialty Count
#
0
1500
1
0
AllergyandImmunology
Oncology
Podiatry
SurgicalSpecialty

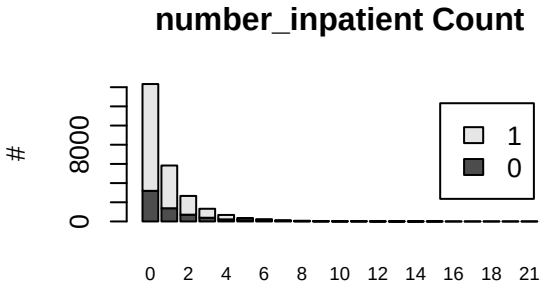

num_lab_procedures Count
#
0
300
700
1
0
1
9
18
29
40
51
62
73
84
95
109

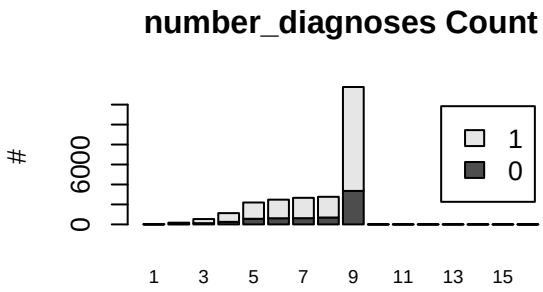

num_procedures Count
#
0
6000
1
0
0
1
2
3
4
5
6

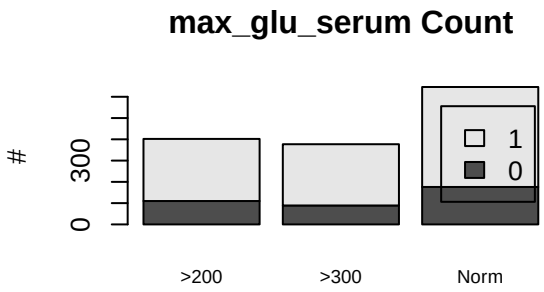

num_medications Count
#
0
1000
1
0
1
6
12
19
26
33
40
47
54
61
69

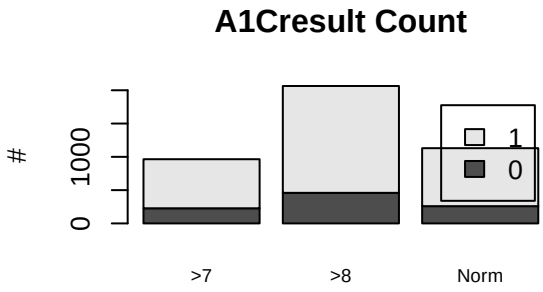

number_outpatient Count
#
0
10000
1
0
0
3
6
9
12
15
18
21
26
34
42

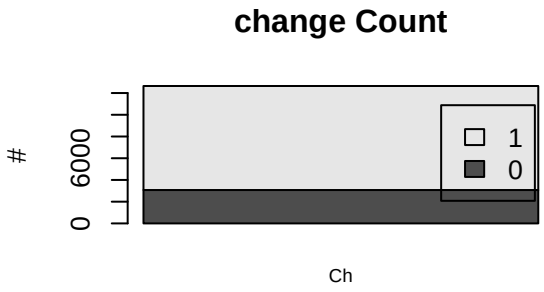

number_emergency Count
#
0
15000
1
0
0
2
4
6
8
10
13
16
20
46


number_inpatient Count
#
0
8000
1
0
0
2
4
6
8
10
12
14
16
18
21



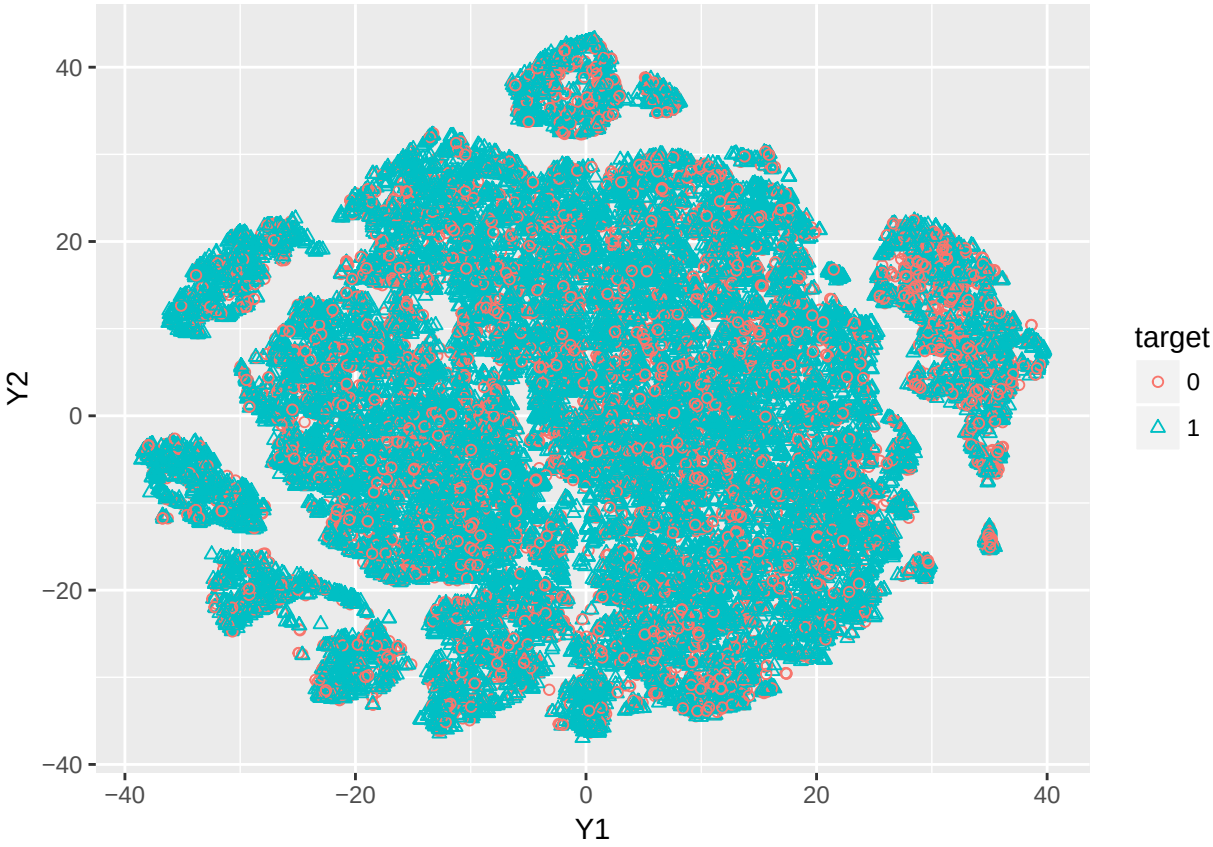




```
## Notice 'change' is a contant feature, there are several other features are constant, shouldn't be hel

## Also, no noticable feature surpress target-mean significantly

## try visualization using tSNE, replace variables with target-mean

## t-Distributed Stochastic Neighbor Embedding(tsne) is a dimensionality reduction technique that maps
## low dimensional space(usually 2 or 3) such that distribution of distances between training samples s

## Read the 25762 x 42 data matrix successfully!
## Using no_dims = 2, perplexity = 30.000000, and theta = 0.500000
## Computing input similarities...
## Normalizing input...
## Building tree...
##  - point 0 of 25762
##  - point 10000 of 25762
##  - point 20000 of 25762
## Done in 16.00 seconds (sparsity = 0.004785)!
## Learning embedding...
## Iteration 50: error is 108.993345 (50 iterations in 25.58 seconds)
## Iteration 100: error is 108.790515 (50 iterations in 33.45 seconds)
## Iteration 150: error is 92.150367 (50 iterations in 23.05 seconds)
## Iteration 200: error is 88.327358 (50 iterations in 21.74 seconds)
## Iteration 250: error is 87.125481 (50 iterations in 21.83 seconds)
## Iteration 300: error is 3.733740 (50 iterations in 21.56 seconds)
## Iteration 350: error is 3.342363 (50 iterations in 21.48 seconds)
## Iteration 400: error is 3.103327 (50 iterations in 22.47 seconds)
## Iteration 450: error is 2.933514 (50 iterations in 22.75 seconds)
## Iteration 500: error is 2.802670 (50 iterations in 23.02 seconds)
```

```
## Iteration 550: error is 2.697601 (50 iterations in 22.03 seconds)
## Iteration 600: error is 2.610497 (50 iterations in 22.86 seconds)
## Iteration 650: error is 2.536751 (50 iterations in 21.81 seconds)
## Iteration 700: error is 2.473012 (50 iterations in 22.01 seconds)
## Iteration 750: error is 2.417086 (50 iterations in 21.99 seconds)
## Iteration 800: error is 2.368043 (50 iterations in 23.02 seconds)
## Fitting performed in 370.65 seconds.
```




```
## On this 2D clustering pic, one could hardly see much of the 2 classes separable, indicating the targ

## Will dig more on the relation between target and features, probably non-linear
```