Machine Learning - Week 8 Final Project

Amol Gote

10/23/2020

set.seed(123)  
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(ROCR)  
library(gbm)

## Loaded gbm 2.1.8

library(ROSE)

## Loaded ROSE 0.0-3

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(parallel)

medical\_ds <- read.csv("./data/data.csv")

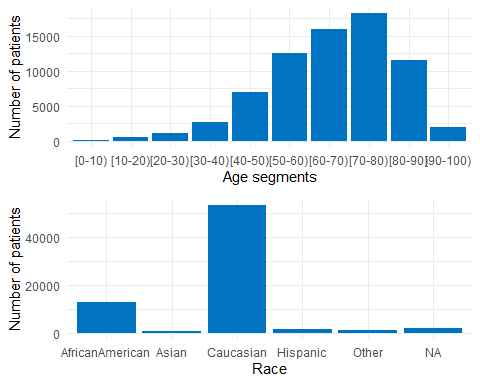
summary(medical\_ds)

## encounter\_id patient\_nbr race   
## Min. : 12522 Min. : 135 AfricanAmerican:12887   
## 1st Qu.: 81458068 1st Qu.: 23375768 Asian : 497   
## Median :144057399 Median : 48166798 Caucasian :53491   
## Mean :156815588 Mean : 55051604 Hispanic : 1517   
## 3rd Qu.:215551324 3rd Qu.: 87522217 Other : 1178   
## Max. :443867222 Max. :189502619 NA's : 1948   
##   
## gender age weight admission\_type\_id  
## Female :38025 [70-80):18210 [75-100) : 1195 Min. :1.0   
## Male :33490 [60-70):15960 [50-75) : 781 1st Qu.:1.0   
## Unknown/Invalid: 3 [50-60):12466 [100-125): 566 Median :1.0   
## [80-90):11589 [125-150): 131 Mean :2.1   
## [40-50): 6878 [25-50) : 89 3rd Qu.:3.0   
## [30-40): 2699 (Other) : 91 Max. :8.0   
## (Other): 3716 NA's :68665   
## discharge\_disposition\_id admission\_source\_id time\_in\_hospital payer\_code   
## Min. : 1.000 Min. : 1.000 Min. : 1.000 MC :20414   
## 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 2.000 HM : 4022   
## Median : 1.000 Median : 7.000 Median : 3.000 BC : 3432   
## Mean : 3.593 Mean : 5.657 Mean : 4.289 SP : 3347   
## 3rd Qu.: 3.000 3rd Qu.: 7.000 3rd Qu.: 6.000 MD : 2197   
## Max. :28.000 Max. :25.000 Max. :14.000 (Other): 7063   
## NA's :31043   
## medical\_specialty num\_lab\_procedures num\_procedures   
## InternalMedicine :10919 Min. : 1.00 Min. :0.000   
## Family/GeneralPractice: 5118 1st Qu.: 31.00 1st Qu.:0.000   
## Emergency/Trauma : 4465 Median : 44.00 Median :1.000   
## Cardiology : 4266 Mean : 43.08 Mean :1.431   
## Surgery-General : 2221 3rd Qu.: 57.00 3rd Qu.:2.000   
## (Other) :10052 Max. :132.00 Max. :6.000   
## NA's :34477   
## num\_medications number\_outpatient number\_emergency number\_inpatient   
## Min. : 1.00 Min. : 0.0000 Min. : 0.0000 Min. : 0.0000   
## 1st Qu.:10.00 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median :14.00 Median : 0.0000 Median : 0.0000 Median : 0.0000   
## Mean :15.71 Mean : 0.2801 Mean : 0.1035 Mean : 0.1778   
## 3rd Qu.:20.00 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0000   
## Max. :81.00 Max. :42.0000 Max. :42.0000 Max. :12.0000   
##   
## diag\_1 diag\_2 diag\_3 number\_diagnoses max\_glu\_serum  
## 414 : 5233 250 : 5009 250 : 9039 Min. : 1.000 >200: 969   
## 428 : 3980 276 : 4604 401 : 6588 1st Qu.: 6.000 >300: 756   
## 786 : 3040 428 : 4335 276 : 3448 Median : 8.000 None:68062   
## 410 : 2902 427 : 3539 428 : 2847 Mean : 7.246 Norm: 1731   
## 486 : 2439 401 : 3088 427 : 2707 3rd Qu.: 9.000   
## (Other):53913 (Other):50649 (Other):45664 Max. :16.000   
## NA's : 11 NA's : 294 NA's : 1225   
## A1Cresult metformin repaglinide nateglinide chlorpropamide  
## >7 : 2891 Down : 435 Down : 29 Down : 8 Down : 1   
## >8 : 6304 No :56527 No :70592 No :71020 No :71446   
## None:58532 Steady:13718 Steady: 824 Steady: 474 Steady: 67   
## Norm: 3791 Up : 838 Up : 73 Up : 16 Up : 4   
##   
##   
##   
## glimepiride acetohexamide glipizide glyburide tolbutamide   
## Down : 138 No :71517 Down : 378 Down : 421 No :71499   
## No :67773 Steady: 1 No :62412 No :63664 Steady: 19   
## Steady: 3372 Steady: 8150 Steady: 6812   
## Up : 235 Up : 578 Up : 621   
##   
##   
##   
## pioglitazone rosiglitazone acarbose miglitol troglitazone   
## Down : 81 Down : 75 No :71316 Down : 1 No :71515   
## No :66210 No :66817 Steady: 192 No :71498 Steady: 3   
## Steady: 5047 Steady: 4490 Up : 10 Steady: 18   
## Up : 180 Up : 136 Up : 1   
##   
##   
##   
## tolazamide examide citoglipton insulin glyburide.metformin  
## No :71488 No:71518 No:71518 Down : 7505 Down : 4   
## Steady: 30 No :34921 No :71016   
## Steady:22129 Steady: 491   
## Up : 6963 Up : 7   
##   
##   
##   
## glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone  
## No :71511 No:71518 No :71516   
## Steady: 7 Steady: 2   
##   
##   
##   
##   
##   
## metformin.pioglitazone change diabetesMed readmitted   
## No :71517 Ch:32024 No :17199 <30: 6293   
## Steady: 1 No:39494 Yes:54319 >30:22240   
## NO :42985   
##   
##   
##   
##

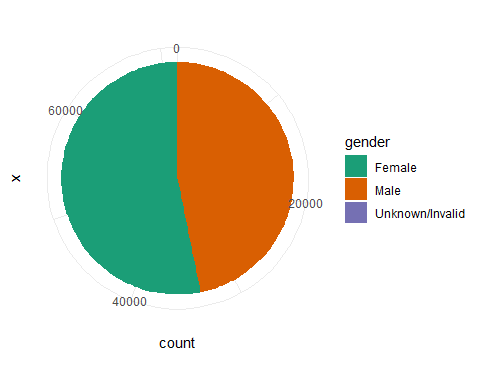
nrow(medical\_ds)

## [1] 71518

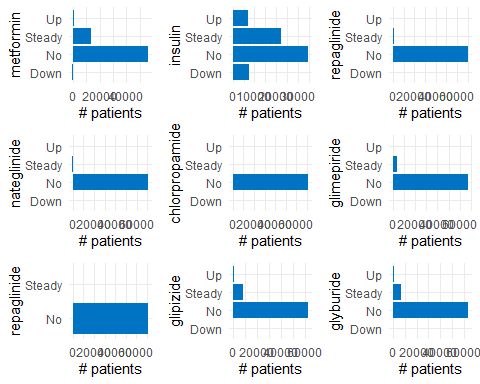
age\_plot <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = age), fill = "#0073C2FF") +  
 xlab("Age segments") +   
 ylab("Number of patients") +   
 theme\_minimal()   
  
race\_plot <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = race), fill = "#0073C2FF") +  
 xlab("Race") +   
 ylab("Number of patients") +   
 theme\_minimal()   
  
  
gender\_plot <- ggplot(data = medical\_ds, aes(x="", fill=gender))+  
 geom\_bar(width = 1) +  
 coord\_polar("y") +  
 scale\_fill\_brewer(palette="Dark2") +  
 theme\_minimal()  
  
grid.arrange(age\_plot, race\_plot)



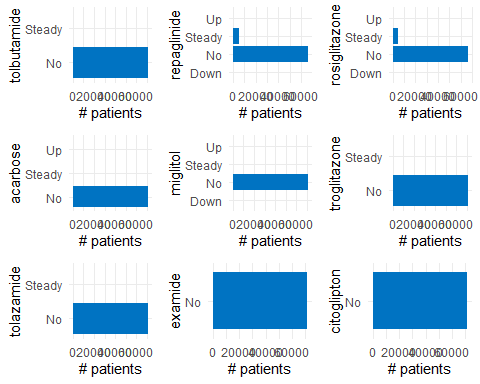
gender\_plot



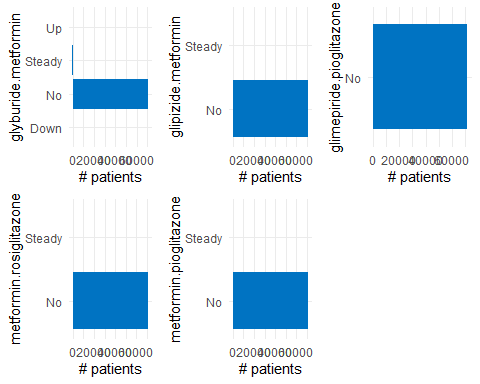
metformin <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = metformin), fill = "#0073C2FF") +  
 xlab("metformin") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
insulin <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = insulin), fill = "#0073C2FF") +  
 xlab("insulin") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
repaglinide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = repaglinide), fill = "#0073C2FF") +  
 xlab("repaglinide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
nateglinide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = nateglinide), fill = "#0073C2FF") +  
 xlab("nateglinide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
chlorpropamide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = chlorpropamide), fill = "#0073C2FF") +  
 xlab("chlorpropamide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
glimepiride <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = glimepiride), fill = "#0073C2FF") +  
 xlab("glimepiride") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
acetohexamide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = acetohexamide), fill = "#0073C2FF") +  
 xlab("repaglinide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
glipizide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = glipizide), fill = "#0073C2FF") +  
 xlab("glipizide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
glyburide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = glyburide), fill = "#0073C2FF") +  
 xlab("glyburide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
grid.arrange(metformin, insulin,repaglinide,nateglinide,chlorpropamide,glimepiride,acetohexamide,glipizide,glyburide, ncol=3)



tolbutamide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = tolbutamide), fill = "#0073C2FF") +  
 xlab("tolbutamide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
pioglitazone <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = pioglitazone), fill = "#0073C2FF") +  
 xlab("repaglinide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
rosiglitazone <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = rosiglitazone), fill = "#0073C2FF") +  
 xlab("rosiglitazone") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
acarbose <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = acarbose), fill = "#0073C2FF") +  
 xlab("acarbose") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
miglitol <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = miglitol), fill = "#0073C2FF") +  
 xlab("miglitol") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
troglitazone <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = troglitazone), fill = "#0073C2FF") +  
 xlab("troglitazone") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
tolazamide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = tolazamide), fill = "#0073C2FF") +  
 xlab("tolazamide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
examide <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = examide), fill = "#0073C2FF") +  
 xlab("examide") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
citoglipton <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = citoglipton), fill = "#0073C2FF") +  
 xlab("citoglipton") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
grid.arrange(tolbutamide,pioglitazone,rosiglitazone,acarbose,miglitol,troglitazone,tolazamide,examide,citoglipton, ncol=3)



glyburide.metformin <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = glyburide.metformin), fill = "#0073C2FF") +  
 xlab("glyburide.metformin") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
glipizide.metformin <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = glipizide.metformin), fill = "#0073C2FF") +  
 xlab("glipizide.metformin") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
glimepiride.pioglitazone <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = glimepiride.pioglitazone), fill = "#0073C2FF") +  
 xlab("glimepiride.pioglitazone") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
metformin.rosiglitazone <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = metformin.rosiglitazone), fill = "#0073C2FF") +  
 xlab("metformin.rosiglitazone") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
metformin.pioglitazone <- ggplot(data = medical\_ds)+  
 geom\_bar(aes(x = metformin.pioglitazone), fill = "#0073C2FF") +  
 xlab("metformin.pioglitazone") +   
 ylab("# patients") +   
 coord\_flip() +  
 theme\_minimal()   
  
grid.arrange(glyburide.metformin, glipizide.metformin,glimepiride.pioglitazone,metformin.rosiglitazone,metformin.pioglitazone, ncol=3)



medical\_ds\_1 <- medical\_ds[,c(  
 "age", "race", "gender", "admission\_type\_id","discharge\_disposition\_id", "admission\_source\_id", "time\_in\_hospital","num\_lab\_procedures",  
 "num\_procedures","num\_medications","number\_outpatient","number\_emergency","number\_inpatient",  
 "diag\_1","diag\_2","diag\_3","number\_diagnoses","change","max\_glu\_serum","A1Cresult","metformin",  
 "glimepiride","glipizide","glyburide","pioglitazone","rosiglitazone",  
 "insulin","readmitted"  
 )]  
summary(medical\_ds\_1)

## age race gender   
## [70-80):18210 AfricanAmerican:12887 Female :38025   
## [60-70):15960 Asian : 497 Male :33490   
## [50-60):12466 Caucasian :53491 Unknown/Invalid: 3   
## [80-90):11589 Hispanic : 1517   
## [40-50): 6878 Other : 1178   
## [30-40): 2699 NA's : 1948   
## (Other): 3716   
## admission\_type\_id discharge\_disposition\_id admission\_source\_id  
## Min. :1.0 Min. : 1.000 Min. : 1.000   
## 1st Qu.:1.0 1st Qu.: 1.000 1st Qu.: 1.000   
## Median :1.0 Median : 1.000 Median : 7.000   
## Mean :2.1 Mean : 3.593 Mean : 5.657   
## 3rd Qu.:3.0 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :8.0 Max. :28.000 Max. :25.000   
##   
## time\_in\_hospital num\_lab\_procedures num\_procedures num\_medications  
## Min. : 1.000 Min. : 1.00 Min. :0.000 Min. : 1.00   
## 1st Qu.: 2.000 1st Qu.: 31.00 1st Qu.:0.000 1st Qu.:10.00   
## Median : 3.000 Median : 44.00 Median :1.000 Median :14.00   
## Mean : 4.289 Mean : 43.08 Mean :1.431 Mean :15.71   
## 3rd Qu.: 6.000 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:20.00   
## Max. :14.000 Max. :132.00 Max. :6.000 Max. :81.00   
##   
## number\_outpatient number\_emergency number\_inpatient diag\_1   
## Min. : 0.0000 Min. : 0.0000 Min. : 0.0000 414 : 5233   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000 428 : 3980   
## Median : 0.0000 Median : 0.0000 Median : 0.0000 786 : 3040   
## Mean : 0.2801 Mean : 0.1035 Mean : 0.1778 410 : 2902   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0000 486 : 2439   
## Max. :42.0000 Max. :42.0000 Max. :12.0000 (Other):53913   
## NA's : 11   
## diag\_2 diag\_3 number\_diagnoses change max\_glu\_serum  
## 250 : 5009 250 : 9039 Min. : 1.000 Ch:32024 >200: 969   
## 276 : 4604 401 : 6588 1st Qu.: 6.000 No:39494 >300: 756   
## 428 : 4335 276 : 3448 Median : 8.000 None:68062   
## 427 : 3539 428 : 2847 Mean : 7.246 Norm: 1731   
## 401 : 3088 427 : 2707 3rd Qu.: 9.000   
## (Other):50649 (Other):45664 Max. :16.000   
## NA's : 294 NA's : 1225   
## A1Cresult metformin glimepiride glipizide glyburide   
## >7 : 2891 Down : 435 Down : 138 Down : 378 Down : 421   
## >8 : 6304 No :56527 No :67773 No :62412 No :63664   
## None:58532 Steady:13718 Steady: 3372 Steady: 8150 Steady: 6812   
## Norm: 3791 Up : 838 Up : 235 Up : 578 Up : 621   
##   
##   
##   
## pioglitazone rosiglitazone insulin readmitted   
## Down : 81 Down : 75 Down : 7505 <30: 6293   
## No :66210 No :66817 No :34921 >30:22240   
## Steady: 5047 Steady: 4490 Steady:22129 NO :42985   
## Up : 180 Up : 136 Up : 6963   
##   
##   
##

medical\_ds\_2 <- na.omit(medical\_ds\_1, cols = c("race","gender", "diag\_1", "diag\_2", "diag\_3"))  
nrow(medical\_ds\_2)

## [1] 68358

medical\_ds\_3 <- medical\_ds\_2 %>%  
 filter(!(discharge\_disposition\_id %in% c(11,19,20,21)))  
nrow(medical\_ds\_3)

## [1] 67301

medical\_ds\_3 <- medical\_ds\_3 %>%  
 filter(gender != 'Unknown/Invalid')  
nrow(medical\_ds\_3)

## [1] 67300

summary(medical\_ds\_3)

## age race gender   
## [70-80):17275 AfricanAmerican:12416 Female :35896   
## [60-70):15182 Asian : 477 Male :31404   
## [50-60):11864 Caucasian :51810 Unknown/Invalid: 0   
## [80-90):10886 Hispanic : 1458   
## [40-50): 6462 Other : 1139   
## [30-40): 2498   
## (Other): 3133   
## admission\_type\_id discharge\_disposition\_id admission\_source\_id  
## Min. :1.000 Min. : 1.00 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.: 1.00 1st Qu.: 1.000   
## Median :1.000 Median : 1.00 Median : 7.000   
## Mean :2.109 Mean : 3.52 Mean : 5.665   
## 3rd Qu.:3.000 3rd Qu.: 3.00 3rd Qu.: 7.000   
## Max. :8.000 Max. :28.00 Max. :25.000   
##   
## time\_in\_hospital num\_lab\_procedures num\_procedures num\_medications  
## Min. : 1.000 Min. : 1.00 Min. :0.000 Min. : 1.00   
## 1st Qu.: 2.000 1st Qu.: 31.00 1st Qu.:0.000 1st Qu.:10.00   
## Median : 4.000 Median : 44.00 Median :1.000 Median :14.00   
## Mean : 4.313 Mean : 42.98 Mean :1.437 Mean :15.78   
## 3rd Qu.: 6.000 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:20.00   
## Max. :14.000 Max. :132.00 Max. :6.000 Max. :81.00   
##   
## number\_outpatient number\_emergency number\_inpatient diag\_1   
## Min. : 0.0000 Min. : 0.0000 Min. : 0.0000 414 : 5027   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000 428 : 3805   
## Median : 0.0000 Median : 0.0000 Median : 0.0000 786 : 2929   
## Mean : 0.2862 Mean : 0.1063 Mean : 0.1804 410 : 2704   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0000 486 : 2307   
## Max. :42.0000 Max. :42.0000 Max. :12.0000 427 : 1967   
## (Other):48561   
## diag\_2 diag\_3 number\_diagnoses change max\_glu\_serum  
## 250 : 4396 250 : 8700 Min. : 3.000 Ch:30225 >200: 913   
## 276 : 4329 401 : 6377 1st Qu.: 6.000 No:37075 >300: 691   
## 428 : 4141 276 : 3281 Median : 8.000 None:64038   
## 427 : 3370 428 : 2718 Mean : 7.336 Norm: 1658   
## 401 : 2964 414 : 2569 3rd Qu.: 9.000   
## 599 : 2175 427 : 2551 Max. :16.000   
## (Other):45925 (Other):41104   
## A1Cresult metformin glimepiride glipizide glyburide   
## >7 : 2779 Down : 413 Down : 131 Down : 360 Down : 396   
## >8 : 5712 No :53029 No :63754 No :58588 No :59828   
## None:55175 Steady:13065 Steady: 3192 Steady: 7789 Steady: 6479   
## Norm: 3634 Up : 793 Up : 223 Up : 563 Up : 597   
##   
##   
##   
## pioglitazone rosiglitazone insulin readmitted   
## Down : 79 Down : 72 Down : 7083 <30: 6090   
## No :62223 No :62797 No :33183 >30:21539   
## Steady: 4827 Steady: 4302 Steady:20547 NO :39671   
## Up : 171 Up : 129 Up : 6487   
##   
##   
##

medical\_ds\_3$readmitted\_enc[medical\_ds\_3$readmitted == '<30'] <- 1  
medical\_ds\_3$readmitted\_enc[medical\_ds\_3$readmitted == '>30'] <- 1  
medical\_ds\_3$readmitted\_enc[medical\_ds\_3$readmitted == 'NO'] <- 0  
medical\_ds\_3$readmitted\_enc <- factor(medical\_ds\_3$readmitted\_enc)

medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[0-10)'] <- 'Young'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[10-20)'] <- 'Young'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[20-30)'] <- 'Young'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[30-40)'] <- 'Young'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[40-50)'] <- 'Middle-aged'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[50-60)'] <- 'Middle-aged'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[60-70)'] <- 'Old'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[70-80)'] <- 'Old'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[80-90)'] <- 'Older'  
medical\_ds\_3$age\_enc[medical\_ds\_3$age == '[90-100)'] <- 'Older'  
medical\_ds\_3$age\_enc <- factor(medical\_ds\_3$age\_enc)  
summary(medical\_ds\_3)

## age race gender   
## [70-80):17275 AfricanAmerican:12416 Female :35896   
## [60-70):15182 Asian : 477 Male :31404   
## [50-60):11864 Caucasian :51810 Unknown/Invalid: 0   
## [80-90):10886 Hispanic : 1458   
## [40-50): 6462 Other : 1139   
## [30-40): 2498   
## (Other): 3133   
## admission\_type\_id discharge\_disposition\_id admission\_source\_id  
## Min. :1.000 Min. : 1.00 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.: 1.00 1st Qu.: 1.000   
## Median :1.000 Median : 1.00 Median : 7.000   
## Mean :2.109 Mean : 3.52 Mean : 5.665   
## 3rd Qu.:3.000 3rd Qu.: 3.00 3rd Qu.: 7.000   
## Max. :8.000 Max. :28.00 Max. :25.000   
##   
## time\_in\_hospital num\_lab\_procedures num\_procedures num\_medications  
## Min. : 1.000 Min. : 1.00 Min. :0.000 Min. : 1.00   
## 1st Qu.: 2.000 1st Qu.: 31.00 1st Qu.:0.000 1st Qu.:10.00   
## Median : 4.000 Median : 44.00 Median :1.000 Median :14.00   
## Mean : 4.313 Mean : 42.98 Mean :1.437 Mean :15.78   
## 3rd Qu.: 6.000 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:20.00   
## Max. :14.000 Max. :132.00 Max. :6.000 Max. :81.00   
##   
## number\_outpatient number\_emergency number\_inpatient diag\_1   
## Min. : 0.0000 Min. : 0.0000 Min. : 0.0000 414 : 5027   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000 428 : 3805   
## Median : 0.0000 Median : 0.0000 Median : 0.0000 786 : 2929   
## Mean : 0.2862 Mean : 0.1063 Mean : 0.1804 410 : 2704   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0000 486 : 2307   
## Max. :42.0000 Max. :42.0000 Max. :12.0000 427 : 1967   
## (Other):48561   
## diag\_2 diag\_3 number\_diagnoses change max\_glu\_serum  
## 250 : 4396 250 : 8700 Min. : 3.000 Ch:30225 >200: 913   
## 276 : 4329 401 : 6377 1st Qu.: 6.000 No:37075 >300: 691   
## 428 : 4141 276 : 3281 Median : 8.000 None:64038   
## 427 : 3370 428 : 2718 Mean : 7.336 Norm: 1658   
## 401 : 2964 414 : 2569 3rd Qu.: 9.000   
## 599 : 2175 427 : 2551 Max. :16.000   
## (Other):45925 (Other):41104   
## A1Cresult metformin glimepiride glipizide glyburide   
## >7 : 2779 Down : 413 Down : 131 Down : 360 Down : 396   
## >8 : 5712 No :53029 No :63754 No :58588 No :59828   
## None:55175 Steady:13065 Steady: 3192 Steady: 7789 Steady: 6479   
## Norm: 3634 Up : 793 Up : 223 Up : 563 Up : 597   
##   
##   
##   
## pioglitazone rosiglitazone insulin readmitted readmitted\_enc  
## Down : 79 Down : 72 Down : 7083 <30: 6090 0:39671   
## No :62223 No :62797 No :33183 >30:21539 1:27629   
## Steady: 4827 Steady: 4302 Steady:20547 NO :39671   
## Up : 171 Up : 129 Up : 6487   
##   
##   
##   
## age\_enc   
## Middle-aged:18326   
## Old :32457   
## Older :12629   
## Young : 3888   
##   
##   
##

medical\_ds\_3$gender <- factor(medical\_ds\_3$gender)

medical\_ds\_3$race <- factor(medical\_ds\_3$race)

medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id <= 3] <- "Referral"  
  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 4] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 5] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 6] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 10] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 18] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 22] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 25] <- "Transfer"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 26] <- "Transfer"  
  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 11] <- "Birth"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 12] <- "Birth"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 13] <- "Birth"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 14] <- "Birth"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 23] <- "Birth"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 24] <- "Birth"  
  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 7] <- "Misc"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 8] <- "Misc"  
  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 9] <- "Unknown"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 15] <- "Unknown"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 17] <- "Unknown"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 19] <- "Unknown"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 20] <- "Unknown"  
medical\_ds\_3$admission\_source\_id\_enc[medical\_ds\_3$admission\_source\_id == 21] <- "Unknown"  
medical\_ds\_3$admission\_source\_id\_enc <- factor(medical\_ds\_3$admission\_source\_id\_enc)

medical\_ds\_3$admission\_type\_id\_enc[medical\_ds\_3$admission\_type\_id == 1] <- "Emergency"  
medical\_ds\_3$admission\_type\_id\_enc[medical\_ds\_3$admission\_type\_id == 2] <- "Urgent"  
medical\_ds\_3$admission\_type\_id\_enc[medical\_ds\_3$admission\_type\_id == 3] <- "Elective"  
medical\_ds\_3$admission\_type\_id\_enc[medical\_ds\_3$admission\_type\_id >= 4] <- "Other"  
medical\_ds\_3$admission\_type\_id\_enc <- factor(medical\_ds\_3$admission\_type\_id\_enc)

medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 6] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 8] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 15] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 16] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 17] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 22] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 23] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 24] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 27] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 28] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 29] <- "Discharged"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 30] <- "Discharged"  
  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 13] <- "Hospice"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 14] <- "Hospice"  
  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 7] <- "Misc"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 9] <- "Misc"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 10] <- "Misc"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 12] <- "Misc"  
  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 18] <- "Unknown"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 25] <- "Unknown"  
medical\_ds\_3$discharge\_disposition\_id\_enc[medical\_ds\_3$discharge\_disposition\_id <= 26] <- "Unknown"  
medical\_ds\_3$discharge\_disposition\_id\_enc <- factor(medical\_ds\_3$discharge\_disposition\_id\_enc)

medical\_ds\_3$max\_glu\_serum <- factor(medical\_ds\_3$max\_glu\_serum)  
medical\_ds\_3$A1Cresult <- factor(medical\_ds\_3$A1Cresult)  
medical\_ds\_3$change <- factor(medical\_ds\_3$change)  
medical\_ds\_3$metformin <- factor(medical\_ds\_3$metformin)  
medical\_ds\_3$insulin <- factor(medical\_ds\_3$insulin)

medical\_ds\_3$diag\_1 <- as.numeric(medical\_ds\_3$diag\_1)  
#medical\_ds\_3$diag\_1\_enc <- 0  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 390 & medical\_ds\_3$diag\_1 <= 459] <- "Circulatory"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 == 785] <- "Circulatory"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 460 & medical\_ds\_3$diag\_1 <= 519] <- "Respiratory"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 == 786] <- "Respiratory"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 520 & medical\_ds\_3$diag\_1 <= 579] <- "Digestive"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 == 787] <- "Digestive"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 == 250] <- "Diabetes"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 250.00 & medical\_ds\_3$diag\_1 <= 250.99] <- "Diabetes"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 800 & medical\_ds\_3$diag\_1 <= 999] <- "Injury"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 710 & medical\_ds\_3$diag\_1 <= 739] <- "Musculoskeletal"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 580 & medical\_ds\_3$diag\_1 <= 629] <- "Genitourinary"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 == 788] <- "Genitourinary"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 140 & medical\_ds\_3$diag\_1 <= 239] <- "Neoplasms"  
  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 == 780 | medical\_ds\_3$diag\_1 == 782 | medical\_ds\_3$diag\_1 == 781 | medical\_ds\_3$diag\_1 == 784] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 790 & medical\_ds\_3$diag\_1 <= 799] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 240 & medical\_ds\_3$diag\_1 <= 249] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 251 & medical\_ds\_3$diag\_1 <= 279] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 680 & medical\_ds\_3$diag\_1 <= 709] <- "Other"  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 1 & medical\_ds\_3$diag\_1 <= 139] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 290 & medical\_ds\_3$diag\_1 <= 319] <- "Other"  
  
  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 280 & medical\_ds\_3$diag\_1 <= 289] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 320 & medical\_ds\_3$diag\_1 <= 359] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 630 & medical\_ds\_3$diag\_1 <= 679] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 360 & medical\_ds\_3$diag\_1 <= 389] <- "Other"  
medical\_ds\_3$diag\_1\_enc[medical\_ds\_3$diag\_1 >= 740 & medical\_ds\_3$diag\_1 <= 759] <- "Other"  
medical\_ds\_3$diag\_1\_enc <- factor(medical\_ds\_3$diag\_1\_enc)  
levels(medical\_ds\_3$diag\_1\_enc)

## [1] "Circulatory" "Diabetes" "Digestive" "Genitourinary"  
## [5] "Neoplasms" "Other" "Respiratory"

summary(medical\_ds\_3$diag\_1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 254.0 293.0 331.4 453.0 696.0

medical\_ds\_3$diag\_2 <- as.numeric(medical\_ds\_3$diag\_2)  
#medical\_ds\_3$diag\_2\_enc <- 0  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 390 & medical\_ds\_3$diag\_2 <= 459] <- "Circulatory"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 == 785] <- "Circulatory"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 460 & medical\_ds\_3$diag\_2 <= 519] <- "Respiratory"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 == 786] <- "Respiratory"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 520 & medical\_ds\_3$diag\_2 <= 579] <- "Digestive"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 == 787] <- "Digestive"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 == 250] <- "Diabetes"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 250.00 & medical\_ds\_3$diag\_2 <= 250.99] <- "Diabetes"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 800 & medical\_ds\_3$diag\_2 <= 999] <- "Injury"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 710 & medical\_ds\_3$diag\_2 <= 739] <- "Musculoskeletal"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 580 & medical\_ds\_3$diag\_2 <= 629] <- "Genitourinary"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 == 788] <- "Genitourinary"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 140 & medical\_ds\_3$diag\_2 <= 239] <- "Neoplasms"  
  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 == 780 | medical\_ds\_3$diag\_2 == 782 | medical\_ds\_3$diag\_2 == 781 | medical\_ds\_3$diag\_2 == 784] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 790 & medical\_ds\_3$diag\_2 <= 799] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 240 & medical\_ds\_3$diag\_2 <= 249] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 251 & medical\_ds\_3$diag\_2 <= 279] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 680 & medical\_ds\_3$diag\_2 <= 709] <- "Other"  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 1 & medical\_ds\_3$diag\_2 <= 139] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 290 & medical\_ds\_3$diag\_2 <= 319] <- "Other"  
  
  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 280 & medical\_ds\_3$diag\_2 <= 289] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 320 & medical\_ds\_3$diag\_2 <= 359] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 630 & medical\_ds\_3$diag\_2 <= 679] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 360 & medical\_ds\_3$diag\_2 <= 389] <- "Other"  
medical\_ds\_3$diag\_2\_enc[medical\_ds\_3$diag\_2 >= 740 & medical\_ds\_3$diag\_2 <= 759] <- "Other"  
medical\_ds\_3$diag\_2\_enc <- factor(medical\_ds\_3$diag\_2\_enc)  
levels(medical\_ds\_3$diag\_2\_enc)

## [1] "Circulatory" "Diabetes" "Digestive" "Genitourinary"   
## [5] "Musculoskeletal" "Neoplasms" "Other" "Respiratory"

summary(medical\_ds\_3$diag\_2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 138.0 251.0 267.7 364.0 725.0

medical\_ds\_3$diag\_3 <- as.numeric(medical\_ds\_3$diag\_3)  
#medical\_ds\_3$diag\_3\_enc <- 0  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 390 & medical\_ds\_3$diag\_3 <= 459] <- "Circulatory"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 == 785] <- "Circulatory"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 460 & medical\_ds\_3$diag\_3 <= 519] <- "Respiratory"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 == 786] <- "Respiratory"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 520 & medical\_ds\_3$diag\_3 <= 579] <- "Digestive"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 == 787] <- "Digestive"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 == 250] <- "Diabetes"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 250.00 & medical\_ds\_3$diag\_3 <= 250.99] <- "Diabetes"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 800 & medical\_ds\_3$diag\_3 <= 999] <- "Injury"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 710 & medical\_ds\_3$diag\_3 <= 739] <- "Musculoskeletal"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 580 & medical\_ds\_3$diag\_3 <= 629] <- "Genitourinary"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 == 788] <- "Genitourinary"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 140 & medical\_ds\_3$diag\_3 <= 239] <- "Neoplasms"  
  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 == 780 | medical\_ds\_3$diag\_3 == 782 | medical\_ds\_3$diag\_3 == 781 | medical\_ds\_3$diag\_3 == 784] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 790 & medical\_ds\_3$diag\_3 <= 799] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 240 & medical\_ds\_3$diag\_3 <= 249] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 251 & medical\_ds\_3$diag\_3 <= 279] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 680 & medical\_ds\_3$diag\_3 <= 709] <- "Other"  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 1 & medical\_ds\_3$diag\_3 <= 139] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 290 & medical\_ds\_3$diag\_3 <= 319] <- "Other"  
  
  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 280 & medical\_ds\_3$diag\_3 <= 289] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 320 & medical\_ds\_3$diag\_3 <= 359] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 630 & medical\_ds\_3$diag\_3 <= 679] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 360 & medical\_ds\_3$diag\_3 <= 389] <- "Other"  
medical\_ds\_3$diag\_3\_enc[medical\_ds\_3$diag\_3 >= 740 & medical\_ds\_3$diag\_3 <= 759] <- "Other"  
medical\_ds\_3$diag\_3\_enc <- factor(medical\_ds\_3$diag\_3\_enc)  
levels(medical\_ds\_3$diag\_3\_enc)

## [1] "Circulatory" "Diabetes" "Digestive" "Genitourinary"   
## [5] "Musculoskeletal" "Neoplasms" "Other" "Respiratory"

summary(medical\_ds\_3$diag\_3)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 131.0 242.0 266.4 344.0 758.0

final\_medical\_ds\_1 <- medical\_ds\_3  
summary(final\_medical\_ds\_1)

## age race gender admission\_type\_id  
## [70-80):17275 AfricanAmerican:12416 Female:35896 Min. :1.000   
## [60-70):15182 Asian : 477 Male :31404 1st Qu.:1.000   
## [50-60):11864 Caucasian :51810 Median :1.000   
## [80-90):10886 Hispanic : 1458 Mean :2.109   
## [40-50): 6462 Other : 1139 3rd Qu.:3.000   
## [30-40): 2498 Max. :8.000   
## (Other): 3133   
## discharge\_disposition\_id admission\_source\_id time\_in\_hospital  
## Min. : 1.00 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 1.00 1st Qu.: 1.000 1st Qu.: 2.000   
## Median : 1.00 Median : 7.000 Median : 4.000   
## Mean : 3.52 Mean : 5.665 Mean : 4.313   
## 3rd Qu.: 3.00 3rd Qu.: 7.000 3rd Qu.: 6.000   
## Max. :28.00 Max. :25.000 Max. :14.000   
##   
## num\_lab\_procedures num\_procedures num\_medications number\_outpatient  
## Min. : 1.00 Min. :0.000 Min. : 1.00 Min. : 0.0000   
## 1st Qu.: 31.00 1st Qu.:0.000 1st Qu.:10.00 1st Qu.: 0.0000   
## Median : 44.00 Median :1.000 Median :14.00 Median : 0.0000   
## Mean : 42.98 Mean :1.437 Mean :15.78 Mean : 0.2862   
## 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:20.00 3rd Qu.: 0.0000   
## Max. :132.00 Max. :6.000 Max. :81.00 Max. :42.0000   
##   
## number\_emergency number\_inpatient diag\_1 diag\_2   
## Min. : 0.0000 Min. : 0.0000 Min. : 1.0 Min. : 1.0   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.:254.0 1st Qu.:138.0   
## Median : 0.0000 Median : 0.0000 Median :293.0 Median :251.0   
## Mean : 0.1063 Mean : 0.1804 Mean :331.4 Mean :267.7   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.:453.0 3rd Qu.:364.0   
## Max. :42.0000 Max. :12.0000 Max. :696.0 Max. :725.0   
##   
## diag\_3 number\_diagnoses change max\_glu\_serum A1Cresult   
## Min. : 1.0 Min. : 3.000 Ch:30225 >200: 913 >7 : 2779   
## 1st Qu.:131.0 1st Qu.: 6.000 No:37075 >300: 691 >8 : 5712   
## Median :242.0 Median : 8.000 None:64038 None:55175   
## Mean :266.4 Mean : 7.336 Norm: 1658 Norm: 3634   
## 3rd Qu.:344.0 3rd Qu.: 9.000   
## Max. :758.0 Max. :16.000   
##   
## metformin glimepiride glipizide glyburide pioglitazone   
## Down : 413 Down : 131 Down : 360 Down : 396 Down : 79   
## No :53029 No :63754 No :58588 No :59828 No :62223   
## Steady:13065 Steady: 3192 Steady: 7789 Steady: 6479 Steady: 4827   
## Up : 793 Up : 223 Up : 563 Up : 597 Up : 171   
##   
##   
##   
## rosiglitazone insulin readmitted readmitted\_enc age\_enc   
## Down : 72 Down : 7083 <30: 6090 0:39671 Middle-aged:18326   
## No :62797 No :33183 >30:21539 1:27629 Old :32457   
## Steady: 4302 Steady:20547 NO :39671 Older :12629   
## Up : 129 Up : 6487 Young : 3888   
##   
##   
##   
## admission\_source\_id\_enc admission\_type\_id\_enc discharge\_disposition\_id\_enc  
## Birth : 4 Elective :13259 Discharged: 92   
## Misc :36320 Emergency:34293 Unknown :67208   
## Referral:21764 Other : 7697   
## Transfer: 4328 Urgent :12051   
## Unknown : 4884   
##   
##   
## diag\_1\_enc diag\_2\_enc diag\_3\_enc   
## Circulatory : 4293 Other :46260 Other :44733   
## Diabetes : 304 Neoplasms :10999 Neoplasms :12840   
## Digestive : 7591 Respiratory : 4359 Respiratory : 3698   
## Genitourinary: 409 Circulatory : 3101 Circulatory : 2771   
## Neoplasms : 5511 Diabetes : 859 Digestive : 1903   
## Other :43567 Genitourinary: 820 Musculoskeletal: 1076   
## Respiratory : 5625 (Other) : 902 (Other) : 279

# split the data into training and (held-out) test sets  
training\_ind <- createDataPartition(final\_medical\_ds\_1$readmitted\_enc,  
p = 0.75,  
list = FALSE,  
times = 1)  
training\_set <- final\_medical\_ds\_1[training\_ind, ]  
test\_set <- final\_medical\_ds\_1[-training\_ind, ]

summary(training\_set)

## age race gender admission\_type\_id  
## [70-80):12940 AfricanAmerican: 9254 Female:26947 Min. :1.000   
## [60-70):11351 Asian : 362 Male :23529 1st Qu.:1.000   
## [50-60): 8928 Caucasian :38876 Median :1.000   
## [80-90): 8157 Hispanic : 1103 Mean :2.112   
## [40-50): 4867 Other : 881 3rd Qu.:3.000   
## [30-40): 1886 Max. :8.000   
## (Other): 2347   
## discharge\_disposition\_id admission\_source\_id time\_in\_hospital  
## Min. : 1.000 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 2.000   
## Median : 1.000 Median : 7.000 Median : 4.000   
## Mean : 3.536 Mean : 5.658 Mean : 4.322   
## 3rd Qu.: 3.000 3rd Qu.: 7.000 3rd Qu.: 6.000   
## Max. :28.000 Max. :25.000 Max. :14.000   
##   
## num\_lab\_procedures num\_procedures num\_medications number\_outpatient  
## Min. : 1.00 Min. :0.000 Min. : 1.00 Min. : 0.0000   
## 1st Qu.: 31.00 1st Qu.:0.000 1st Qu.:10.00 1st Qu.: 0.0000   
## Median : 44.00 Median :1.000 Median :14.00 Median : 0.0000   
## Mean : 43.03 Mean :1.442 Mean :15.79 Mean : 0.2851   
## 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:20.00 3rd Qu.: 0.0000   
## Max. :132.00 Max. :6.000 Max. :75.00 Max. :36.0000   
##   
## number\_emergency number\_inpatient diag\_1 diag\_2   
## Min. : 0.0000 Min. : 0.0000 Min. : 1.0 Min. : 1.0   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.:254.0 1st Qu.:138.0   
## Median : 0.0000 Median : 0.0000 Median :293.0 Median :251.0   
## Mean : 0.1062 Mean : 0.1826 Mean :331.3 Mean :267.2   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.:451.0 3rd Qu.:362.0   
## Max. :42.0000 Max. :12.0000 Max. :696.0 Max. :725.0   
##   
## diag\_3 number\_diagnoses change max\_glu\_serum A1Cresult   
## Min. : 1.0 Min. : 3.000 Ch:22664 >200: 672 >7 : 2081   
## 1st Qu.:131.0 1st Qu.: 6.000 No:27812 >300: 518 >8 : 4311   
## Median :242.0 Median : 8.000 None:48078 None:41348   
## Mean :266.7 Mean : 7.337 Norm: 1208 Norm: 2736   
## 3rd Qu.:344.0 3rd Qu.: 9.000   
## Max. :758.0 Max. :16.000   
##   
## metformin glimepiride glipizide glyburide pioglitazone   
## Down : 301 Down : 98 Down : 268 Down : 291 Down : 53   
## No :39801 No :47836 No :43889 No :44880 No :46701   
## Steady: 9773 Steady: 2373 Steady: 5887 Steady: 4851 Steady: 3593   
## Up : 601 Up : 169 Up : 432 Up : 454 Up : 129   
##   
##   
##   
## rosiglitazone insulin readmitted readmitted\_enc age\_enc   
## Down : 53 Down : 5296 <30: 4567 0:29754 Middle-aged:13795   
## No :47114 No :24801 >30:16155 1:20722 Old :24291   
## Steady: 3213 Steady:15488 NO :29754 Older : 9464   
## Up : 96 Up : 4891 Young : 2926   
##   
##   
##   
## admission\_source\_id\_enc admission\_type\_id\_enc discharge\_disposition\_id\_enc  
## Birth : 3 Elective : 9952 Discharged: 70   
## Misc :27164 Emergency:25649 Unknown :50406   
## Referral:16360 Other : 5779   
## Transfer: 3292 Urgent : 9096   
## Unknown : 3657   
##   
##   
## diag\_1\_enc diag\_2\_enc diag\_3\_enc   
## Circulatory : 3254 Other :34731 Other :33585   
## Diabetes : 227 Neoplasms : 8249 Neoplasms : 9593   
## Digestive : 5691 Respiratory : 3243 Respiratory : 2768   
## Genitourinary: 309 Circulatory : 2314 Circulatory : 2082   
## Neoplasms : 4173 Diabetes : 649 Digestive : 1421   
## Other :32627 Genitourinary: 613 Musculoskeletal: 813   
## Respiratory : 4195 (Other) : 677 (Other) : 214

threshold <- 250  
target\_enc\_train <- function(variable, level) {  
 training\_set$readmitted\_enc <- as.numeric(as.vector(training\_set$readmitted\_enc))  
 train\_avg\_target <- mean(training\_set[, "readmitted\_enc"])  
 if (nrow(training\_set[training\_set[, variable]==level, ])==0) {  
 return(train\_avg\_target)  
 } else {  
 level\_num\_obs <- nrow(training\_set[training\_set[, variable]==level,])  
 level\_avg\_target <- mean(training\_set[training\_set[, variable]==level,  
 "readmitted\_enc"])  
 return((level\_num\_obs\*level\_avg\_target+threshold\*train\_avg\_target)/(level\_num\_obs+threshold))  
 }  
}  
age\_target <- mapply(target\_enc\_train, variable = "age\_enc", level = levels(training\_set$age\_enc), USE.NAMES = FALSE)  
names(age\_target) <- levels(training\_set$age\_enc)  
training\_set$age\_target <- 0  
for (level in levels(training\_set$age\_enc)) {  
 training\_set[training\_set[, "age\_enc"]==level, "age\_target"] <- age\_target[level]  
}  
  
test\_set$age\_target <- 0  
for (level in levels(training\_set$age\_enc)) {  
 test\_set[test\_set[, "age\_enc"]==level, "age\_target"] <- age\_target[level]  
}  
  
  
#gender\_target <- mapply(target\_enc\_train, variable = "gender", level = levels(training\_set$gender), USE.NAMES = FALSE)  
#names(gender\_target) <- levels(training\_set$gender)  
#training\_set$gender\_target <- 0  
#for (level in levels(training\_set$gender)) {  
# training\_set[training\_set[, "gender"]==level, "gender\_target"] <- gender\_target[level]  
#}  
  
#test\_set$gender\_target <- 0  
#for (level in levels(training\_set$gender)) {  
# test\_set[test\_set[, "gender"]==level, "gender\_target"] <- gender\_target[level]  
#}  
  
  
race\_target <- mapply(target\_enc\_train, variable = "race", level = levels(training\_set$race), USE.NAMES = FALSE)  
names(race\_target) <- levels(training\_set$race)  
training\_set$race\_target <- 0  
for (level in levels(training\_set$race)) {  
 training\_set[training\_set[, "race"]==level, "race\_target"] <- race\_target[level]  
}  
  
test\_set$race\_target <- 0  
for (level in levels(training\_set$race)) {  
 test\_set[test\_set[, "race"]==level, "race\_target"] <- race\_target[level]  
}  
  
  
admission\_source\_id\_target <- mapply(target\_enc\_train, variable = "admission\_source\_id\_enc", level = levels(training\_set$admission\_source\_id\_enc), USE.NAMES = FALSE)  
names(admission\_source\_id\_target) <- levels(training\_set$admission\_source\_id\_enc)  
training\_set$admission\_source\_id\_target <- 0  
for (level in levels(training\_set$admission\_source\_id\_enc)) {  
 training\_set[training\_set[, "admission\_source\_id\_enc"]==level, "admission\_source\_id\_target"] <- admission\_source\_id\_target[level]  
}  
  
test\_set$admission\_source\_id\_target <- 0  
for (level in levels(training\_set$admission\_source\_id\_enc)) {  
 test\_set[test\_set[, "admission\_source\_id\_enc"]==level, "admission\_source\_id\_target"] <- admission\_source\_id\_target[level]  
}  
  
admission\_type\_id\_target <- mapply(target\_enc\_train, variable = "admission\_type\_id\_enc", level = levels(training\_set$admission\_type\_id\_enc), USE.NAMES = FALSE)  
names(admission\_type\_id\_target) <- levels(training\_set$admission\_type\_id\_enc)  
training\_set$admission\_type\_id\_target <- 0  
for (level in levels(training\_set$admission\_type\_id\_enc)) {  
 training\_set[training\_set[, "admission\_type\_id\_enc"]==level, "admission\_type\_id\_target"] <- admission\_type\_id\_target[level]  
}  
  
test\_set$admission\_type\_id\_target <- 0  
for (level in levels(training\_set$admission\_type\_id\_enc)) {  
 test\_set[test\_set[, "admission\_type\_id\_enc"]==level, "admission\_type\_id\_target"] <- admission\_type\_id\_target[level]  
}  
  
discharge\_disposition\_id\_target <- mapply(target\_enc\_train,   
 variable = "discharge\_disposition\_id\_enc", level = levels(training\_set$discharge\_disposition\_id\_enc), USE.NAMES = FALSE)  
names(discharge\_disposition\_id\_target) <- levels(training\_set$discharge\_disposition\_id\_enc)  
training\_set$discharge\_disposition\_id\_target <- 0  
for (level in levels(training\_set$discharge\_disposition\_id\_enc)) {  
 training\_set[training\_set[, "discharge\_disposition\_id\_enc"]==level, "discharge\_disposition\_id\_target"] <- discharge\_disposition\_id\_target[level]  
}  
  
test\_set$discharge\_disposition\_id\_target <- 0  
for (level in levels(training\_set$discharge\_disposition\_id\_enc)) {  
 test\_set[test\_set[, "discharge\_disposition\_id\_enc"]==level, "discharge\_disposition\_id\_target"] <- discharge\_disposition\_id\_target[level]  
}  
  
  
max\_glu\_serum\_target <- mapply(target\_enc\_train, variable = "max\_glu\_serum", level = levels(training\_set$max\_glu\_serum), USE.NAMES = FALSE)  
names(max\_glu\_serum\_target) <- levels(training\_set$max\_glu\_serum)  
training\_set$max\_glu\_serum\_target <- 0  
for (level in levels(training\_set$max\_glu\_serum)) {  
 training\_set[training\_set[, "max\_glu\_serum"]==level, "max\_glu\_serum\_target"] <- max\_glu\_serum\_target[level]  
}  
  
test\_set$max\_glu\_serum\_target <- 0  
for (level in levels(training\_set$max\_glu\_serum)) {  
 test\_set[test\_set[, "max\_glu\_serum"]==level, "max\_glu\_serum\_target"] <- max\_glu\_serum\_target[level]  
}  
  
A1Cresult\_target <- mapply(target\_enc\_train, variable = "A1Cresult", level = levels(training\_set$A1Cresult), USE.NAMES = FALSE)  
names(A1Cresult\_target) <- levels(training\_set$A1Cresult)  
training\_set$A1Cresult\_target <- 0  
for (level in levels(training\_set$A1Cresult)) {  
 training\_set[training\_set[, "A1Cresult"]==level, "A1Cresult\_target"] <- A1Cresult\_target[level]  
}  
  
test\_set$A1Cresult\_target <- 0  
for (level in levels(training\_set$A1Cresult)) {  
 test\_set[test\_set[, "A1Cresult"]==level, "A1Cresult\_target"] <- A1Cresult\_target[level]  
}  
  
#change\_target <- mapply(target\_enc\_train, variable = "change", level = levels(training\_set$change), USE.NAMES = FALSE)  
#names(change\_target) <- levels(training\_set$change)  
#training\_set$change\_target <- 0  
#for (level in levels(training\_set$change)) {  
# training\_set[training\_set[, "change"]==level, "change\_target"] <- change\_target[level]  
#}  
  
#test\_set$change\_target <- 0  
#for (level in levels(training\_set$change)) {  
# test\_set[test\_set[, "change"]==level, "change\_target"] <- change\_target[level]  
#}  
  
metformin\_target <- mapply(target\_enc\_train, variable = "metformin", level = levels(training\_set$metformin), USE.NAMES = FALSE)  
names(metformin\_target) <- levels(training\_set$metformin)  
training\_set$metformin\_target <- 0  
for (level in levels(training\_set$metformin)) {  
 training\_set[training\_set[, "metformin"]==level, "metformin\_target"] <- metformin\_target[level]  
}  
  
test\_set$metformin\_target <- 0  
for (level in levels(training\_set$metformin)) {  
 test\_set[test\_set[, "metformin"]==level, "metformin\_target"] <- metformin\_target[level]  
}  
  
insulin\_target <- mapply(target\_enc\_train, variable = "insulin", level = levels(training\_set$insulin), USE.NAMES = FALSE)  
names(insulin\_target) <- levels(training\_set$insulin)  
training\_set$insulin\_target <- 0  
for (level in levels(training\_set$insulin)) {  
 training\_set[training\_set[, "insulin"]==level, "insulin\_target"] <- insulin\_target[level]  
}  
  
test\_set$insulin\_target <- 0  
for (level in levels(training\_set$insulin)) {  
 test\_set[test\_set[, "insulin"]==level, "insulin\_target"] <- insulin\_target[level]  
}  
  
  
diag\_1\_enc\_target <- mapply(target\_enc\_train, variable = "diag\_1\_enc", level = levels(training\_set$diag\_1\_enc), USE.NAMES = FALSE)  
names(diag\_1\_enc\_target) <- levels(training\_set$diag\_1\_enc)  
training\_set$diag\_1\_enc\_target <- 0  
for (level in levels(training\_set$diag\_1\_enc)) {  
 training\_set[training\_set[, "diag\_1\_enc"]==level, "diag\_1\_enc\_target"] <- diag\_1\_enc\_target[level]  
}  
  
test\_set$diag\_1\_enc\_target <- 0  
for (level in levels(training\_set$diag\_1\_enc)) {  
 test\_set[test\_set[, "diag\_1\_enc"]==level, "diag\_1\_enc\_target"] <- diag\_1\_enc\_target[level]  
}  
  
diag\_2\_enc\_target <- mapply(target\_enc\_train, variable = "diag\_2\_enc", level = levels(training\_set$diag\_2\_enc), USE.NAMES = FALSE)  
names(diag\_2\_enc\_target) <- levels(training\_set$diag\_2\_enc)  
training\_set$diag\_2\_enc\_target <- 0  
for (level in levels(training\_set$diag\_2\_enc)) {  
 training\_set[training\_set[, "diag\_2\_enc"]==level, "diag\_2\_enc\_target"] <- diag\_2\_enc\_target[level]  
}  
  
test\_set$diag\_2\_enc\_target <- 0  
for (level in levels(training\_set$diag\_2\_enc)) {  
 test\_set[test\_set[, "diag\_2\_enc"]==level, "diag\_2\_enc\_target"] <- diag\_2\_enc\_target[level]  
}  
  
diag\_3\_enc\_target <- mapply(target\_enc\_train, variable = "diag\_3\_enc", level = levels(training\_set$diag\_3\_enc), USE.NAMES = FALSE)  
names(diag\_3\_enc\_target) <- levels(training\_set$diag\_3\_enc)  
training\_set$diag\_3\_enc\_target <- 0  
for (level in levels(training\_set$diag\_3\_enc)) {  
 training\_set[training\_set[, "diag\_3\_enc"]==level, "diag\_3\_enc\_target"] <- diag\_3\_enc\_target[level]  
}  
  
test\_set$diag\_3\_enc\_target <- 0  
for (level in levels(training\_set$diag\_3\_enc)) {  
 test\_set[test\_set[, "diag\_3\_enc"]==level, "diag\_3\_enc\_target"] <- diag\_3\_enc\_target[level]  
}

# onehot\_encoder <- dummyVars(~ gender + change + diag\_1\_enc + diag\_2\_enc + diag\_3\_enc,   
# final\_medical\_ds\_1[, c("gender", "change","diag\_1\_enc", "diag\_2\_enc", "diag\_3\_enc")])  
# onehot\_enc\_training <- predict(onehot\_encoder, training\_set[, c("gender", "change","diag\_1\_enc", "diag\_2\_enc", "diag\_3\_enc")])  
# training\_set <- cbind(training\_set, onehot\_enc\_training)  
# onehot\_enc\_test <- predict(onehot\_encoder, test\_set[, c("gender", "change","diag\_1\_enc", "diag\_2\_enc", "diag\_3\_enc")])  
# test\_set <- cbind(test\_set, onehot\_enc\_test)

onehot\_encoder <- dummyVars(~ gender + change,  
 final\_medical\_ds\_1[, c("gender", "change")])  
onehot\_enc\_training <- predict(onehot\_encoder, training\_set[, c("gender", "change")])  
training\_set <- cbind(training\_set, onehot\_enc\_training)  
onehot\_enc\_test <- predict(onehot\_encoder, test\_set[, c("gender", "change")])  
test\_set <- cbind(test\_set, onehot\_enc\_test)

test\_sample\_size <- 15000  
train\_sample\_size <- test\_sample\_size \* 3  
training\_subset\_ds <- training\_set[sample(nrow(training\_set), train\_sample\_size),]  
nrow(training\_subset\_ds)

## [1] 45000

test\_subset\_ds <- test\_set[sample(nrow(test\_set), test\_sample\_size),]  
nrow(test\_subset\_ds)

## [1] 15000

head(training\_subset\_ds)

## age race gender admission\_type\_id discharge\_disposition\_id  
## 6056 [80-90) AfricanAmerican Female 1 6  
## 58809 [40-50) Caucasian Female 8 1  
## 46865 [80-90) Caucasian Female 3 1  
## 27798 [80-90) Caucasian Female 1 1  
## 18628 [70-80) AfricanAmerican Female 1 1  
## 32447 [40-50) Caucasian Male 2 1  
## admission\_source\_id time\_in\_hospital num\_lab\_procedures num\_procedures  
## 6056 7 4 42 0  
## 58809 7 1 68 2  
## 46865 1 3 18 1  
## 27798 7 2 49 0  
## 18628 7 4 47 1  
## 32447 7 4 72 1  
## num\_medications number\_outpatient number\_emergency number\_inpatient  
## 6056 6 0 0 0  
## 58809 11 0 0 1  
## 46865 6 2 0 0  
## 27798 11 0 0 0  
## 18628 20 0 2 0  
## 32447 12 0 0 0  
## diag\_1 diag\_2 diag\_3 number\_diagnoses change max\_glu\_serum A1Cresult  
## 6056 372 44 725 5 Ch None None  
## 58809 529 241 237 6 Ch None None  
## 46865 424 630 666 4 No None None  
## 27798 320 77 740 4 No None None  
## 18628 105 129 131 9 Ch None None  
## 32447 342 325 381 5 No None Norm  
## metformin glimepiride glipizide glyburide pioglitazone rosiglitazone  
## 6056 Steady No No Up No No  
## 58809 No Steady No No No No  
## 46865 No No No No No No  
## 27798 No No No No No Steady  
## 18628 No No No No No No  
## 32447 No No No No No No  
## insulin readmitted readmitted\_enc age\_enc admission\_source\_id\_enc  
## 6056 No NO 0 Older Misc  
## 58809 No NO 0 Middle-aged Misc  
## 46865 No >30 1 Older Referral  
## 27798 No NO 0 Older Misc  
## 18628 Up >30 1 Old Misc  
## 32447 No NO 0 Middle-aged Misc  
## admission\_type\_id\_enc discharge\_disposition\_id\_enc diag\_1\_enc diag\_2\_enc  
## 6056 Emergency Unknown Other Other  
## 58809 Other Unknown Digestive Other  
## 46865 Elective Unknown Circulatory Other  
## 27798 Emergency Unknown Other Other  
## 18628 Emergency Unknown Other Other  
## 32447 Urgent Unknown Other Other  
## diag\_3\_enc age\_target race\_target admission\_source\_id\_target  
## 6056 Musculoskeletal 0.4382986 0.3906390 0.4349468  
## 58809 Neoplasms 0.3739860 0.4193793 0.4349468  
## 46865 Other 0.4382986 0.4193793 0.3850471  
## 27798 Other 0.4382986 0.4193793 0.4349468  
## 18628 Other 0.4295926 0.3906390 0.4349468  
## 32447 Other 0.3739860 0.4193793 0.4349468  
## admission\_type\_id\_target discharge\_disposition\_id\_target  
## 6056 0.4155617 0.4103489  
## 58809 0.4690053 0.4103489  
## 46865 0.3551885 0.4103489  
## 27798 0.4155617 0.4103489  
## 18628 0.4155617 0.4103489  
## 32447 0.4192845 0.4103489  
## max\_glu\_serum\_target A1Cresult\_target metformin\_target insulin\_target  
## 6056 0.4085547 0.4146025 0.3869733 0.3947001  
## 58809 0.4085547 0.4146025 0.4170341 0.3947001  
## 46865 0.4085547 0.4146025 0.4170341 0.3947001  
## 27798 0.4085547 0.4146025 0.4170341 0.3947001  
## 18628 0.4085547 0.4146025 0.4170341 0.4442001  
## 32447 0.4085547 0.3652488 0.4170341 0.3947001  
## diag\_1\_enc\_target diag\_2\_enc\_target diag\_3\_enc\_target gender.Female  
## 6056 0.4231114 0.4149862 0.3966443 1  
## 58809 0.3877517 0.4149862 0.3876494 1  
## 46865 0.3711852 0.4149862 0.4184316 1  
## 27798 0.4231114 0.4149862 0.4184316 1  
## 18628 0.4231114 0.4149862 0.4184316 1  
## 32447 0.4231114 0.4149862 0.4184316 0  
## gender.Male change.Ch change.No  
## 6056 0 1 0  
## 58809 0 1 0  
## 46865 0 0 1  
## 27798 0 0 1  
## 18628 0 1 0  
## 32447 1 0 1

head(test\_subset\_ds)

## age race gender admission\_type\_id discharge\_disposition\_id  
## 20919 [50-60) AfricanAmerican Female 1 2  
## 32203 [60-70) Caucasian Female 1 1  
## 53503 [50-60) Caucasian Male 1 1  
## 41538 [70-80) Caucasian Male 3 1  
## 64115 [60-70) Hispanic Female 6 1  
## 12700 [80-90) Caucasian Male 2 18  
## admission\_source\_id time\_in\_hospital num\_lab\_procedures num\_procedures  
## 20919 7 4 40 0  
## 32203 7 3 64 1  
## 53503 7 2 51 0  
## 41538 1 4 3 0  
## 64115 7 1 41 0  
## 12700 6 3 29 0  
## num\_medications number\_outpatient number\_emergency number\_inpatient  
## 20919 18 1 0 0  
## 32203 16 0 0 0  
## 53503 9 0 0 0  
## 41538 9 0 0 0  
## 64115 4 0 0 0  
## 12700 7 0 0 1  
## diag\_1 diag\_2 diag\_3 number\_diagnoses change max\_glu\_serum A1Cresult  
## 20919 166 156 368 9 No None None  
## 32203 554 591 673 9 Ch None None  
## 53503 528 388 78 9 No None >8  
## 41538 293 79 127 7 No None None  
## 64115 529 706 78 5 No None None  
## 12700 270 249 742 8 Ch None None  
## metformin glimepiride glipizide glyburide pioglitazone rosiglitazone  
## 20919 No No No No No No  
## 32203 Steady Steady No No Steady No  
## 53503 No No Steady No No No  
## 41538 No No No No No No  
## 64115 No No No No No No  
## 12700 No No No Down No Steady  
## insulin readmitted readmitted\_enc age\_enc admission\_source\_id\_enc  
## 20919 Steady >30 1 Middle-aged Misc  
## 32203 No NO 0 Old Misc  
## 53503 No NO 0 Middle-aged Misc  
## 41538 Steady NO 0 Old Referral  
## 64115 No NO 0 Old Misc  
## 12700 No <30 1 Older Transfer  
## admission\_type\_id\_enc discharge\_disposition\_id\_enc diag\_1\_enc  
## 20919 Emergency Unknown Neoplasms  
## 32203 Emergency Unknown Digestive  
## 53503 Emergency Unknown Digestive  
## 41538 Elective Unknown Other  
## 64115 Other Unknown Digestive  
## 12700 Urgent Unknown Other  
## diag\_2\_enc diag\_3\_enc age\_target race\_target  
## 20919 Neoplasms Other 0.3739860 0.3906390  
## 32203 Genitourinary Other 0.4295926 0.4193793  
## 53503 Other Other 0.3739860 0.4193793  
## 41538 Other Other 0.4295926 0.4193793  
## 64115 Other Other 0.4295926 0.3714951  
## 12700 Other Other 0.4382986 0.4193793  
## admission\_source\_id\_target admission\_type\_id\_target  
## 20919 0.4349468 0.4155617  
## 32203 0.4349468 0.4155617  
## 53503 0.4349468 0.4155617  
## 41538 0.3850471 0.3551885  
## 64115 0.4349468 0.4690053  
## 12700 0.3000093 0.4192845  
## discharge\_disposition\_id\_target max\_glu\_serum\_target A1Cresult\_target  
## 20919 0.4103489 0.4085547 0.4146025  
## 32203 0.4103489 0.4085547 0.4146025  
## 53503 0.4103489 0.4085547 0.4127676  
## 41538 0.4103489 0.4085547 0.4146025  
## 64115 0.4103489 0.4085547 0.4146025  
## 12700 0.4103489 0.4085547 0.4146025  
## metformin\_target insulin\_target diag\_1\_enc\_target diag\_2\_enc\_target  
## 20919 0.4170341 0.4104481 0.4091415 0.4023571  
## 32203 0.3869733 0.3947001 0.3877517 0.4213591  
## 53503 0.4170341 0.3947001 0.3877517 0.4149862  
## 41538 0.4170341 0.4104481 0.4231114 0.4149862  
## 64115 0.4170341 0.3947001 0.3877517 0.4149862  
## 12700 0.4170341 0.3947001 0.4231114 0.4149862  
## diag\_3\_enc\_target gender.Female gender.Male change.Ch change.No  
## 20919 0.4184316 1 0 0 1  
## 32203 0.4184316 1 0 1 0  
## 53503 0.4184316 0 1 0 1  
## 41538 0.4184316 0 1 0 1  
## 64115 0.4184316 1 0 0 1  
## 12700 0.4184316 0 1 1 0

training\_subset\_ds\_final <- training\_subset\_ds[, -which(names(training\_subset\_ds) %in%   
 c("age", "age\_enc", "race",   
 "admission\_type\_id",  
 "discharge\_disposition\_id",  
 "admission\_source\_id",  
 "max\_glu\_serum",  
 "A1Cresult",  
 "metformin","insulin","glimepiride","glipizide" ,"glyburide","pioglitazone","rosiglitazone",  
 "gender", "change",  
 "diag\_1","diag\_2","diag\_3",  
 "diag\_1\_enc","diag\_2\_enc","diag\_3\_enc", "readmitted"))]  
  
test\_subset\_ds\_final <- test\_subset\_ds[, -which(names(training\_subset\_ds) %in%   
 c("age", "age\_enc", "race",   
 "admission\_type\_id",  
 "discharge\_disposition\_id",  
 "admission\_source\_id",  
 "max\_glu\_serum",  
 "A1Cresult",  
 "metformin","insulin","glimepiride","glipizide" ,"glyburide","pioglitazone","rosiglitazone",  
 "gender", "change",  
 "diag\_1","diag\_2","diag\_3",  
 "diag\_1\_enc","diag\_2\_enc","diag\_3\_enc", "readmitted"))]

colnames(training\_subset\_ds\_final)

## [1] "time\_in\_hospital" "num\_lab\_procedures"   
## [3] "num\_procedures" "num\_medications"   
## [5] "number\_outpatient" "number\_emergency"   
## [7] "number\_inpatient" "number\_diagnoses"   
## [9] "readmitted\_enc" "admission\_source\_id\_enc"   
## [11] "admission\_type\_id\_enc" "discharge\_disposition\_id\_enc"   
## [13] "age\_target" "race\_target"   
## [15] "admission\_source\_id\_target" "admission\_type\_id\_target"   
## [17] "discharge\_disposition\_id\_target" "max\_glu\_serum\_target"   
## [19] "A1Cresult\_target" "metformin\_target"   
## [21] "insulin\_target" "diag\_1\_enc\_target"   
## [23] "diag\_2\_enc\_target" "diag\_3\_enc\_target"   
## [25] "gender.Female" "gender.Male"   
## [27] "change.Ch" "change.No"

num\_var\_start\_index <- 1  
num\_var\_end\_index <- 8  
target\_var\_index <- 9  
test\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index] <- scale(test\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index],   
 center = apply(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index], 2, mean),   
 scale = apply(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index], 2, sd))  
training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index] <- scale(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index])  
levels(training\_subset\_ds\_final$readmitted\_enc)[levels(training\_subset\_ds\_final$readmitted\_enc) == 1] <- "Yes"  
levels(training\_subset\_ds\_final$readmitted\_enc)[levels(training\_subset\_ds\_final$readmitted\_enc) == 0] <- "No"

head(training\_subset\_ds\_final)

## time\_in\_hospital num\_lab\_procedures num\_procedures num\_medications  
## 6056 -0.1079482 -0.05310131 -0.8186237 -1.1809420  
## 58809 -1.1204553 1.24777464 0.3163045 -0.5776750  
## 46865 -0.4454506 -1.25390988 -0.2511596 -1.1809420  
## 27798 -0.7829529 0.29713452 -0.8186237 -0.5776750  
## 18628 -0.1079482 0.19706714 -0.2511596 0.5082055  
## 32447 -0.1079482 1.44790940 -0.2511596 -0.4570216  
## number\_outpatient number\_emergency number\_inpatient number\_diagnoses  
## 6056 -0.2619965 -0.2008944 -0.2994909 -1.2376194  
## 58809 -0.2619965 -0.2008944 1.3366741 -0.7091939  
## 46865 1.5827604 -0.2008944 -0.2994909 -1.7660449  
## 27798 -0.2619965 -0.2008944 -0.2994909 -1.7660449  
## 18628 -0.2619965 3.5887721 -0.2994909 0.8760824  
## 32447 -0.2619965 -0.2008944 -0.2994909 -1.2376194  
## readmitted\_enc admission\_source\_id\_enc admission\_type\_id\_enc  
## 6056 No Misc Emergency  
## 58809 No Misc Other  
## 46865 Yes Referral Elective  
## 27798 No Misc Emergency  
## 18628 Yes Misc Emergency  
## 32447 No Misc Urgent  
## discharge\_disposition\_id\_enc age\_target race\_target  
## 6056 Unknown 0.4382986 0.3906390  
## 58809 Unknown 0.3739860 0.4193793  
## 46865 Unknown 0.4382986 0.4193793  
## 27798 Unknown 0.4382986 0.4193793  
## 18628 Unknown 0.4295926 0.3906390  
## 32447 Unknown 0.3739860 0.4193793  
## admission\_source\_id\_target admission\_type\_id\_target  
## 6056 0.4349468 0.4155617  
## 58809 0.4349468 0.4690053  
## 46865 0.3850471 0.3551885  
## 27798 0.4349468 0.4155617  
## 18628 0.4349468 0.4155617  
## 32447 0.4349468 0.4192845  
## discharge\_disposition\_id\_target max\_glu\_serum\_target A1Cresult\_target  
## 6056 0.4103489 0.4085547 0.4146025  
## 58809 0.4103489 0.4085547 0.4146025  
## 46865 0.4103489 0.4085547 0.4146025  
## 27798 0.4103489 0.4085547 0.4146025  
## 18628 0.4103489 0.4085547 0.4146025  
## 32447 0.4103489 0.4085547 0.3652488  
## metformin\_target insulin\_target diag\_1\_enc\_target diag\_2\_enc\_target  
## 6056 0.3869733 0.3947001 0.4231114 0.4149862  
## 58809 0.4170341 0.3947001 0.3877517 0.4149862  
## 46865 0.4170341 0.3947001 0.3711852 0.4149862  
## 27798 0.4170341 0.3947001 0.4231114 0.4149862  
## 18628 0.4170341 0.4442001 0.4231114 0.4149862  
## 32447 0.4170341 0.3947001 0.4231114 0.4149862  
## diag\_3\_enc\_target gender.Female gender.Male change.Ch change.No  
## 6056 0.3966443 1 0 1 0  
## 58809 0.3876494 1 0 1 0  
## 46865 0.4184316 1 0 0 1  
## 27798 0.4184316 1 0 0 1  
## 18628 0.4184316 1 0 1 0  
## 32447 0.4184316 0 1 0 1

head(test\_subset\_ds\_final)

## time\_in\_hospital num\_lab\_procedures num\_procedures num\_medications  
## 20919 -0.1079482 -0.1531687 -0.8186237 0.26689871  
## 32203 -0.4454506 1.0476399 -0.2511596 0.02559193  
## 53503 -0.7829529 0.3972019 -0.8186237 -0.81898182  
## 41538 -0.1079482 -2.0044152 -0.8186237 -0.81898182  
## 64115 -1.1204553 -0.1031350 -0.8186237 -1.42224878  
## 12700 -0.4454506 -0.7035393 -0.8186237 -1.06028861  
## number\_outpatient number\_emergency number\_inpatient number\_diagnoses  
## 20919 0.6603819 -0.2008944 -0.2994909 0.8760824  
## 32203 -0.2619965 -0.2008944 -0.2994909 0.8760824  
## 53503 -0.2619965 -0.2008944 -0.2994909 0.8760824  
## 41538 -0.2619965 -0.2008944 -0.2994909 -0.1807685  
## 64115 -0.2619965 -0.2008944 -0.2994909 -1.2376194  
## 12700 -0.2619965 -0.2008944 1.3366741 0.3476570  
## readmitted\_enc admission\_source\_id\_enc admission\_type\_id\_enc  
## 20919 1 Misc Emergency  
## 32203 0 Misc Emergency  
## 53503 0 Misc Emergency  
## 41538 0 Referral Elective  
## 64115 0 Misc Other  
## 12700 1 Transfer Urgent  
## discharge\_disposition\_id\_enc age\_target race\_target  
## 20919 Unknown 0.3739860 0.3906390  
## 32203 Unknown 0.4295926 0.4193793  
## 53503 Unknown 0.3739860 0.4193793  
## 41538 Unknown 0.4295926 0.4193793  
## 64115 Unknown 0.4295926 0.3714951  
## 12700 Unknown 0.4382986 0.4193793  
## admission\_source\_id\_target admission\_type\_id\_target  
## 20919 0.4349468 0.4155617  
## 32203 0.4349468 0.4155617  
## 53503 0.4349468 0.4155617  
## 41538 0.3850471 0.3551885  
## 64115 0.4349468 0.4690053  
## 12700 0.3000093 0.4192845  
## discharge\_disposition\_id\_target max\_glu\_serum\_target A1Cresult\_target  
## 20919 0.4103489 0.4085547 0.4146025  
## 32203 0.4103489 0.4085547 0.4146025  
## 53503 0.4103489 0.4085547 0.4127676  
## 41538 0.4103489 0.4085547 0.4146025  
## 64115 0.4103489 0.4085547 0.4146025  
## 12700 0.4103489 0.4085547 0.4146025  
## metformin\_target insulin\_target diag\_1\_enc\_target diag\_2\_enc\_target  
## 20919 0.4170341 0.4104481 0.4091415 0.4023571  
## 32203 0.3869733 0.3947001 0.3877517 0.4213591  
## 53503 0.4170341 0.3947001 0.3877517 0.4149862  
## 41538 0.4170341 0.4104481 0.4231114 0.4149862  
## 64115 0.4170341 0.3947001 0.3877517 0.4149862  
## 12700 0.4170341 0.3947001 0.4231114 0.4149862  
## diag\_3\_enc\_target gender.Female gender.Male change.Ch change.No  
## 20919 0.4184316 1 0 0 1  
## 32203 0.4184316 1 0 1 0  
## 53503 0.4184316 0 1 0 1  
## 41538 0.4184316 0 1 0 1  
## 64115 0.4184316 1 0 0 1  
## 12700 0.4184316 0 1 1 0

# Calculate the number of cores  
no\_cores <- detectCores() - 1  
  
library(doParallel)

## Warning: package 'doParallel' was built under R version 3.6.3

## Loading required package: foreach

## Warning: package 'foreach' was built under R version 3.6.3

##   
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':  
##   
## accumulate, when

## Loading required package: iterators

## Warning: package 'iterators' was built under R version 3.6.3

# create the cluster for caret to use  
cl <- makePSOCKcluster(no\_cores)  
registerDoParallel(cl)

modelLookup("rf")

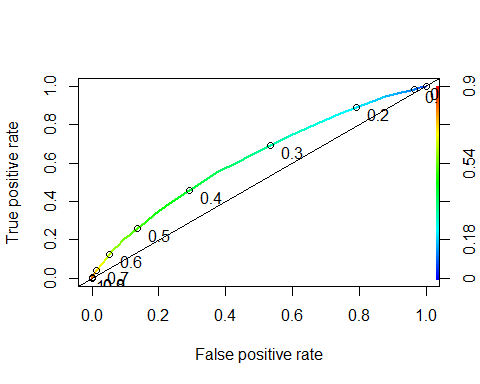
## model parameter label forReg forClass probModel  
## 1 rf mtry #Randomly Selected Predictors TRUE TRUE TRUE

recommended\_mtry <- floor(sqrt(ncol(training\_subset\_ds\_final[,-1\*c(target\_var\_index:target\_var\_index)])))  
rfGrid <- expand.grid(mtry = c(recommended\_mtry-2, recommended\_mtry, recommended\_mtry+2))

rfControl <- trainControl(method = "oob", classProbs = TRUE)  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)

test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes

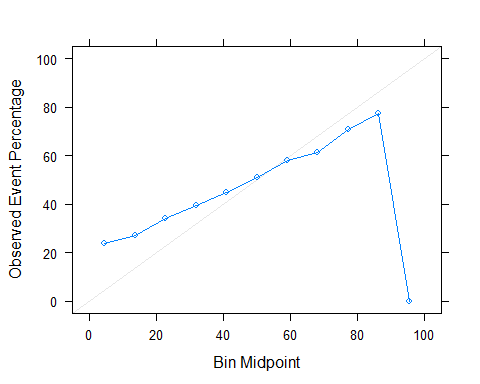
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



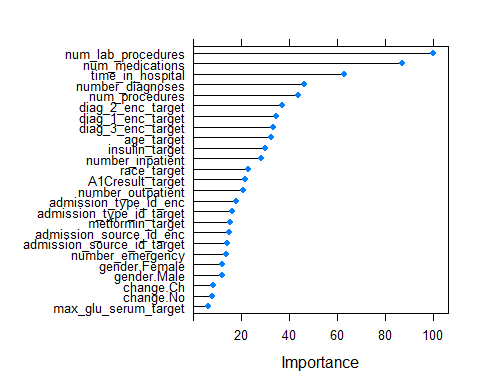
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf <- rocr\_auc@y.values[[1]]  
auc\_rf

## [1] 0.6215247

calibration\_curve <- calibration(readmitted\_enc ~ class\_probabilities\_onehot,  
data = test\_subset\_ds\_final,  
class = 1)  
plot(calibration\_curve)



rf\_varImp <- varImp(rf\_onehot\_model, type = 2)  
plot(rf\_varImp, top = 25)



rf\_varImp

## rf variable importance  
##   
## only 20 most important variables shown (out of 27)  
##   
## Overall  
## num\_lab\_procedures 100.00  
## num\_medications 87.11  
## time\_in\_hospital 63.19  
## number\_diagnoses 46.41  
## num\_procedures 43.80  
## diag\_2\_enc\_target 37.16  
## diag\_1\_enc\_target 34.87  
## diag\_3\_enc\_target 33.54  
## age\_target 32.70  
## insulin\_target 30.02  
## number\_inpatient 28.38  
## race\_target 22.88  
## A1Cresult\_target 21.71  
## number\_outpatient 21.07  
## admission\_type\_id\_enc 17.98  
## admission\_type\_id\_target 16.38  
## metformin\_target 15.42  
## admission\_source\_id\_enc 14.93  
## admission\_source\_id\_target 14.09  
## number\_emergency 13.88

modelLookup("nnet")

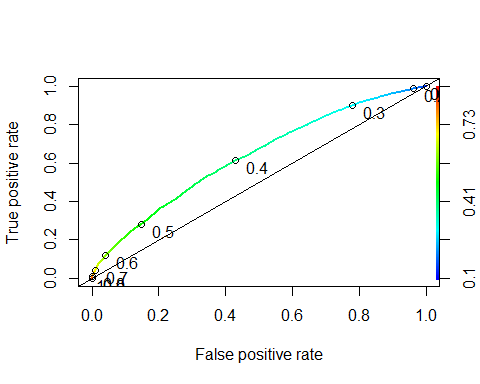
## model parameter label forReg forClass probModel  
## 1 nnet size #Hidden Units TRUE TRUE TRUE  
## 2 nnet decay Weight Decay TRUE TRUE TRUE

nnGrid <- expand.grid(size = 8:10, decay = 0.2)  
nnControl <- trainControl(method = "repeatedcv",  
 repeats = 5,  
 number=10,  
 classProbs = TRUE)

nn\_target\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 800)

test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes

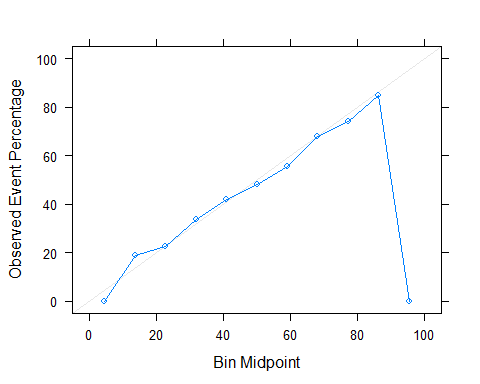
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_nn <- rocr\_auc@y.values[[1]]  
auc\_nn

## [1] 0.6321121

calibration\_curve <- calibration(readmitted\_enc ~ class\_probabilities\_target,  
data = test\_subset\_ds\_final,  
class = 1)  
plot(calibration\_curve)



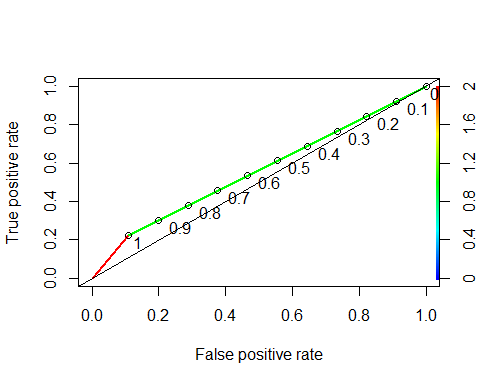
levels(test\_subset\_ds\_final$readmitted\_enc)[levels(test\_subset\_ds\_final$readmitted\_enc) == 1] <- "Yes"  
 levels(test\_subset\_ds\_final$readmitted\_enc)[levels(test\_subset\_ds\_final$readmitted\_enc) == 0] <- "No"  
 test\_subset\_ds\_final$readmitted\_enc <- factor(test\_subset\_ds\_final$readmitted\_enc)  
 rf\_confusion\_matrix <- confusionMatrix(test\_subset\_ds\_final$prediction\_target, test\_subset\_ds\_final$readmitted\_enc, positive = "Yes")  
 rf\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 7512 4429  
## Yes 1310 1749  
##   
## Accuracy : 0.6174   
## 95% CI : (0.6096, 0.6252)  
## No Information Rate : 0.5881   
## P-Value [Acc > NIR] : 1.415e-13   
##   
## Kappa : 0.1456   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.2831   
## Specificity : 0.8515   
## Pos Pred Value : 0.5718   
## Neg Pred Value : 0.6291   
## Prevalence : 0.4119   
## Detection Rate : 0.1166   
## Detection Prevalence : 0.2039   
## Balanced Accuracy : 0.5673   
##   
## 'Positive' Class : Yes   
##

## Logistic Regression

logistic\_regression\_model <- glm(readmitted\_enc ~   
 num\_medications +   
 time\_in\_hospital +  
 number\_diagnoses +  
 num\_procedures +   
 age\_target +  
 number\_inpatient +  
 insulin\_target +  
 race\_target +   
 A1Cresult\_target +  
 number\_outpatient +  
 admission\_type\_id\_enc +  
 admission\_source\_id\_enc +  
 metformin\_target +   
 number\_emergency +   
 diag\_3\_enc\_target +   
 diag\_2\_enc\_target +   
 diag\_1\_enc\_target   
 ,data = training\_subset\_ds\_final, family = binomial)  
  
test\_subset\_ds\_final$prediction\_target <- predict(logistic\_regression\_model,newdata = test\_subset\_ds\_final[,-1\*c(target\_var\_index:target\_var\_index)],type="response")  
test\_subset\_ds\_final$prediction\_target <- ifelse(test\_subset\_ds\_final$prediction\_target > 0.5, 1, 0)

rocr\_pred <- prediction(test\_subset\_ds\_final$prediction\_target, test\_subset\_ds\_final$readmitted\_enc)  
 rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
 plot(rocr\_roc,  
 colorize = TRUE,  
 print.cutoffs.at = seq(0, 1, by = 0.1),  
 text.adj = c(-0.5, 1),  
 lwd = 2)  
 abline(a = 0, b = 1)



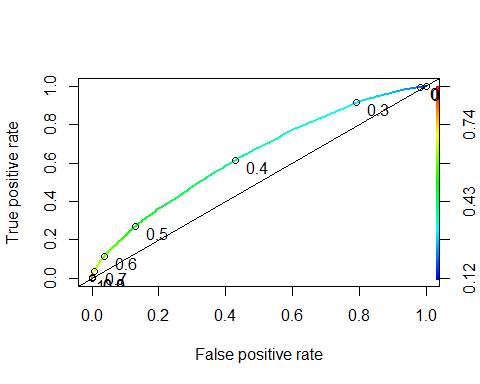
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
 auc\_glm <- rocr\_auc@y.values[[1]]  
 auc\_glm

## [1] 0.5570181

gbmControl <- trainControl(method = "repeatedcv", number = 10, repeats = 10)  
gbm\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "gbm",  
 trControl = gbmControl,  
 verbose = FALSE)

test\_subset\_ds\_final$prediction\_onehot <- predict(gbm\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(gbm\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes

rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_gbm <- rocr\_auc@y.values[[1]]  
auc\_gbm

## [1] 0.6357168

table(training\_subset\_ds\_final$readmitted\_enc)

##   
## No Yes   
## 26506 18494

training\_subset\_ds\_final\_no <- training\_subset\_ds\_final %>%  
 filter(readmitted\_enc == 'No')  
no\_count <- nrow(training\_subset\_ds\_final\_no)  
  
training\_subset\_ds\_final\_yes <- training\_subset\_ds\_final %>%  
 filter(readmitted\_enc == 'Yes')  
yes\_count <- nrow(training\_subset\_ds\_final\_yes)  
no\_count

## [1] 26506

yes\_count

## [1] 18494

training\_subset\_ds\_final\_over <- ovun.sample(readmitted\_enc ~., data = training\_subset\_ds\_final, method = "over", N = no\_count\*2)$data  
table(training\_subset\_ds\_final\_over$readmitted\_enc)

##   
## No Yes   
## 26506 26506

rfControl <- trainControl(method = "oob", classProbs = TRUE)  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final\_over[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final\_over[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)  
  
test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf\_over <- rocr\_auc@y.values[[1]]  
auc\_rf\_over

## [1] 0.6090756

nn\_target\_model <- train(x = training\_subset\_ds\_final\_over[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final\_over[, target\_var\_index],  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 800)  
  
test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes  
  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_nn\_over <- rocr\_auc@y.values[[1]]  
auc\_nn\_over

## [1] 0.6317249

training\_subset\_ds\_final\_under <- ovun.sample(readmitted\_enc ~., data = training\_subset\_ds\_final, method = "under", N = yes\_count\*2)$data  
table(training\_subset\_ds\_final\_under$readmitted\_enc)

##   
## No Yes   
## 18494 18494

rfControl <- trainControl(method = "oob", classProbs = TRUE)  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final\_under[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final\_under[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)  
test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf\_under <- rocr\_auc@y.values[[1]]  
auc\_rf\_under

## [1] 0.6186688

nn\_target\_model <- train(x = training\_subset\_ds\_final\_under[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final\_under[, target\_var\_index],  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 800)  
  
test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes  
  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_nn\_under <- rocr\_auc@y.values[[1]]  
auc\_nn\_under

## [1] 0.6266792

training\_subset\_ds\_final\_both <- ovun.sample(readmitted\_enc ~., data = training\_subset\_ds\_final, method = "both", N = no\_count + yes\_count,  
 p = 0.5, seed = 222)$data  
table(training\_subset\_ds\_final\_both$readmitted\_enc)

##   
## No Yes   
## 22533 22467

rfControl <- trainControl(method = "oob", classProbs = TRUE)  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final\_both[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final\_both[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)  
  
test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf\_both <- rocr\_auc@y.values[[1]]  
auc\_rf\_both

## [1] 0.6064393

nn\_target\_model <- train(x = training\_subset\_ds\_final\_both[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final\_both[, target\_var\_index],  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 800)  
  
test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes  
  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_nn\_both <- rocr\_auc@y.values[[1]]  
auc\_nn\_both

## [1] 0.6209691

rfControl <- trainControl(method = "oob", classProbs = TRUE, sampling = "smote")  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)

## Warning: package 'DMwR' was built under R version 3.6.3

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf\_smote <- rocr\_auc@y.values[[1]]  
auc\_rf\_smote

## [1] 0.6063941

rfControl <- trainControl(method = "oob", classProbs = TRUE, sampling = "up")  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)  
  
test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf\_up\_caret <- rocr\_auc@y.values[[1]]  
auc\_rf\_up\_caret

## [1] 0.6068535

rfControl <- trainControl(method = "oob", classProbs = TRUE, sampling = "down")  
rf\_onehot\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "rf",  
 tuneGrid = rfGrid,  
 trControl = rfControl,  
 ntree = 50,  
 importance = TRUE,  
 trace = FALSE)  
  
test\_subset\_ds\_final$prediction\_onehot <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(rf\_onehot\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_rf\_down\_caret <- rocr\_auc@y.values[[1]]  
auc\_rf\_down\_caret

## [1] 0.6193377

nnControl$sampling <- "up"  
nn\_target\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 800)  
  
test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes  
  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_nn\_up\_caret <- rocr\_auc@y.values[[1]]  
auc\_nn\_up\_caret

## [1] 0.6335705

nnControl$sampling <- "down"  
nn\_target\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 800)  
  
test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes  
  
rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
rocr\_auc <- performance(rocr\_pred, measure = "auc")  
auc\_nn\_down\_caret <- rocr\_auc@y.values[[1]]  
auc\_nn\_down\_caret

## [1] 0.6307908

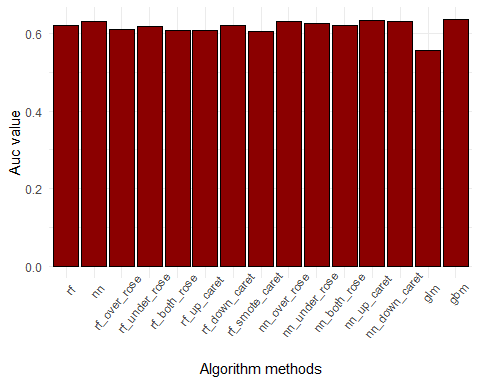
# nnControl$sampling <- "smote"  
# nn\_target\_model <- train(x = training\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], y = training\_subset\_ds\_final[, target\_var\_index],  
# method = "nnet",  
# tuneGrid = nnGrid,  
# trControl = nnControl,  
# importance = TRUE,  
# trace = FALSE,  
# MaxNWts = 800)  
#   
# test\_subset\_ds\_final$prediction\_target <- predict(nn\_target\_model,newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)])  
# class\_probabilities <- predict(nn\_target\_model, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
# test\_subset\_ds\_final$class\_probabilities\_target <- class\_probabilities$Yes  
#   
# rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_target, test\_subset\_ds\_final$readmitted\_enc)  
# rocr\_roc <- performance(rocr\_pred, measure = "tpr", x.measure = "fpr")  
# rocr\_auc <- performance(rocr\_pred, measure = "auc")  
# auc\_nn\_smote\_caret <- rocr\_auc@y.values[[1]]  
# auc\_nn\_smote\_caret

stopCluster(cl)  
registerDoSEQ()

auc\_df <- data.frame("algorithm\_method" = c("rf"), "auc" = c(auc\_rf))  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn"), "auc" = c(auc\_nn))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_over\_rose"), "auc" = c(auc\_rf\_over))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_under\_rose"), "auc" = c(auc\_rf\_under))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_both\_rose"), "auc" = c(auc\_rf\_both))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_up\_caret"), "auc" = c(auc\_rf\_up\_caret))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_down\_caret"), "auc" = c(auc\_rf\_down\_caret))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_smote\_caret"), "auc" = c(auc\_rf\_smote))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn\_over\_rose"), "auc" = c(auc\_nn\_over))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn\_under\_rose"), "auc" = c(auc\_nn\_under))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn\_both\_rose"), "auc" = c(auc\_nn\_both))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn\_up\_caret"), "auc" = c(auc\_nn\_up\_caret))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn\_down\_caret"), "auc" = c(auc\_nn\_down\_caret))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glm"), "auc" = c(auc\_glm))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("gbm"), "auc" = c(auc\_gbm))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
  
# auc\_df\_temp <- data.frame("algorithm\_method" = c("nn\_smote\_caret"), "auc" = c(auc\_nn\_smote\_caret))  
# auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
  
auc\_df

## algorithm\_method auc  
## 1 rf 0.6215247  
## 2 nn 0.6321121  
## 3 rf\_over\_rose 0.6090756  
## 4 rf\_under\_rose 0.6186688  
## 5 rf\_both\_rose 0.6064393  
## 6 rf\_up\_caret 0.6068535  
## 7 rf\_down\_caret 0.6193377  
## 8 rf\_smote\_caret 0.6063941  
## 9 nn\_over\_rose 0.6317249  
## 10 nn\_under\_rose 0.6266792  
## 11 nn\_both\_rose 0.6209691  
## 12 nn\_up\_caret 0.6335705  
## 13 nn\_down\_caret 0.6307908  
## 14 glm 0.5570181  
## 15 gbm 0.6357168

ggplot(data=auc\_df, aes(x=algorithm\_method, y=auc)) +  
 geom\_bar(colour="black", stat="identity", fill = "darkred") +  
 ylab("Auc value") +  
 xlab("Algorithm methods") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))



#detach(package:plyr)

library(dplyr)  
post\_model\_ds <- medical\_ds\_3  
post\_model\_ds$readmitted\_flag[post\_model\_ds$readmitted == '<30'] <- "Yes"  
post\_model\_ds$readmitted\_flag[post\_model\_ds$readmitted == '>30'] <- "Yes"  
post\_model\_ds$readmitted\_flag[post\_model\_ds$readmitted == 'NO'] <- "No"  
  
grouped\_data <- post\_model\_ds %>%  
 group\_by(readmitted\_flag) %>%  
 summarise(  
 mean\_num\_lab\_procedures = mean(num\_lab\_procedures),  
 median\_num\_lab\_procedures = median(num\_lab\_procedures),  
 mean\_num\_medications = mean(num\_medications),  
 median\_num\_medications = median(num\_medications),  
 mean\_time\_in\_hospital = mean(time\_in\_hospital),  
 median\_time\_in\_hospital = median(time\_in\_hospital),  
 mean\_number\_diagnoses = mean(number\_diagnoses),  
 median\_number\_diagnoses = median(number\_diagnoses)  
 )

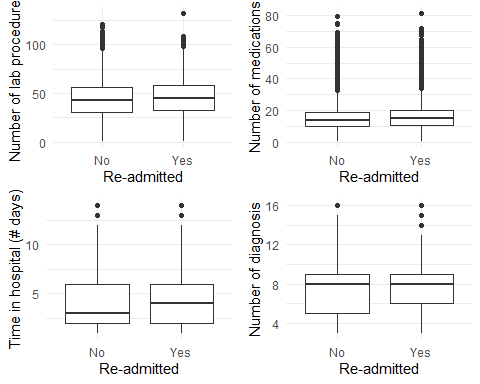
## `summarise()` ungrouping output (override with `.groups` argument)

grouped\_data

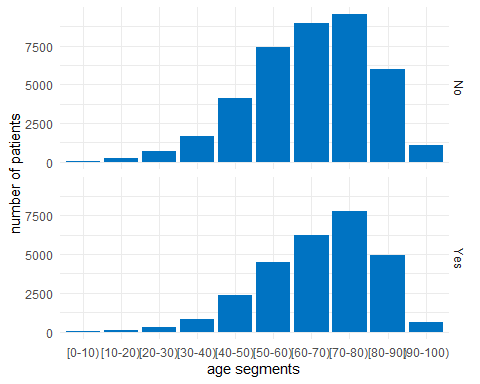
## Warning: `...` is not empty.  
##   
## We detected these problematic arguments:  
## \* `needs\_dots`  
##   
## These dots only exist to allow future extensions and should be empty.  
## Did you misspecify an argument?

## # A tibble: 2 x 9  
## readmitted\_flag mean\_num\_lab\_pr~ median\_num\_lab\_~ mean\_num\_medica~  
## <chr> <dbl> <int> <dbl>  
## 1 No 42.1 43 15.5  
## 2 Yes 44.2 45 16.1  
## # ... with 5 more variables: median\_num\_medications <int>,  
## # mean\_time\_in\_hospital <dbl>, median\_time\_in\_hospital <int>,  
## # mean\_number\_diagnoses <dbl>, median\_number\_diagnoses <int>

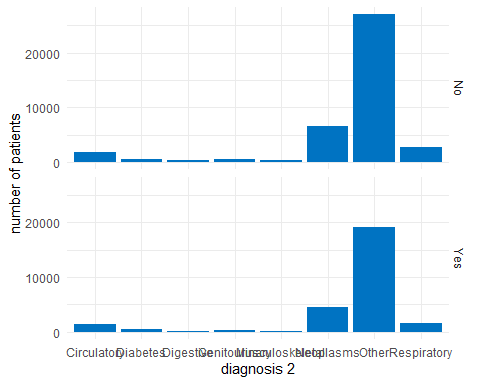
num\_lab\_procedures <- ggplot(data=post\_model\_ds, aes(x=readmitted\_flag, y=num\_lab\_procedures)) +   
 geom\_boxplot() +  
 xlab("Re-admitted") +  
 ylab("Number of lab procedures") +  
 theme\_minimal()  
  
num\_medications <- ggplot(data=post\_model\_ds, aes(x=readmitted\_flag, y=num\_medications)) +   
 geom\_boxplot() +  
 xlab("Re-admitted") +  
 ylab("Number of medications") +  
 theme\_minimal()  
  
time\_in\_hospital <- ggplot(data=post\_model\_ds, aes(x=readmitted\_flag, y=time\_in\_hospital)) +   
 geom\_boxplot() +  
 xlab("Re-admitted") +  
 ylab("Time in hospital (# days)") +  
 theme\_minimal()  
  
number\_diagnoses <- ggplot(data=post\_model\_ds, aes(x=readmitted\_flag, y=number\_diagnoses)) +   
 geom\_boxplot() +  
 xlab("Re-admitted") +  
 ylab("Number of diagnosis") +  
 theme\_minimal()  
  
grid.arrange(num\_lab\_procedures, num\_medications,time\_in\_hospital,number\_diagnoses, ncol=2)



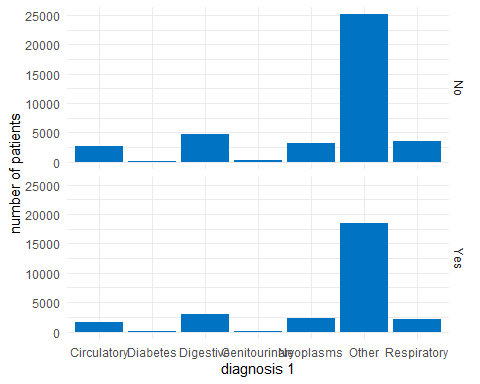
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = age), fill = "#0073c2ff") +  
 xlab("age segments") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



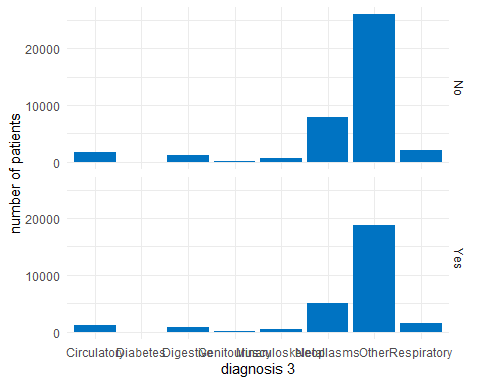
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = diag\_2\_enc), fill = "#0073c2ff") +  
 xlab("diagnosis 2") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



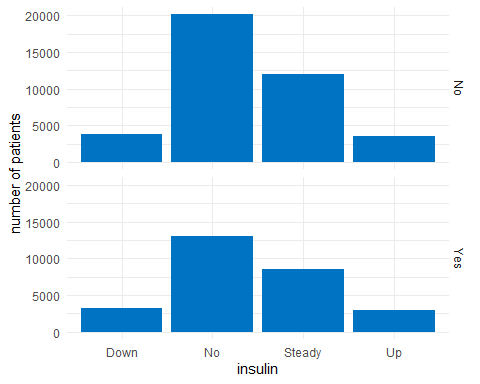
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = diag\_1\_enc), fill = "#0073c2ff") +  
 xlab("diagnosis 1") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



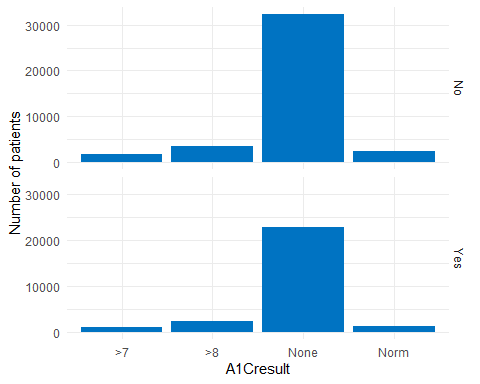
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = diag\_3\_enc), fill = "#0073c2ff") +  
 xlab("diagnosis 3") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



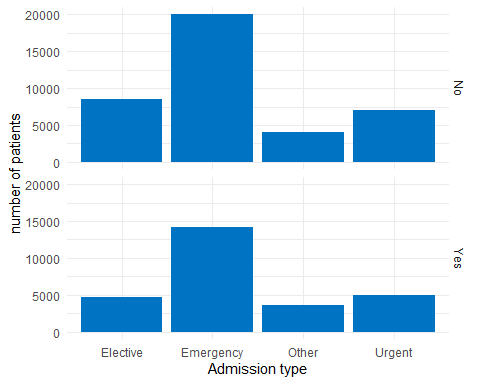
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = insulin), fill = "#0073c2ff") +  
 xlab("insulin") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



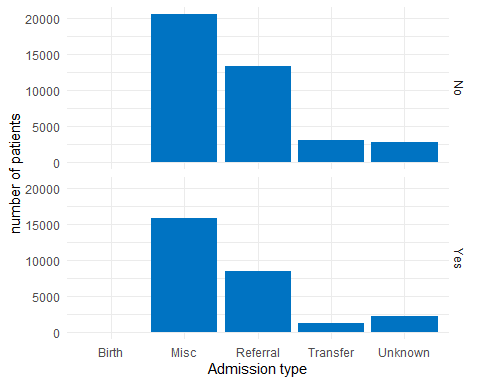
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = A1Cresult), fill = "#0073C2FF") +  
 xlab("A1Cresult") +  
 ylab("Number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



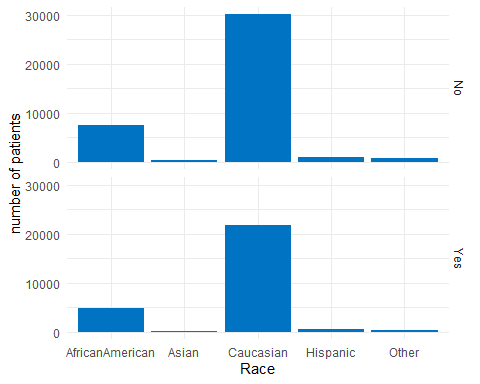
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = admission\_type\_id\_enc), fill = "#0073c2ff") +  
 xlab("Admission type") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



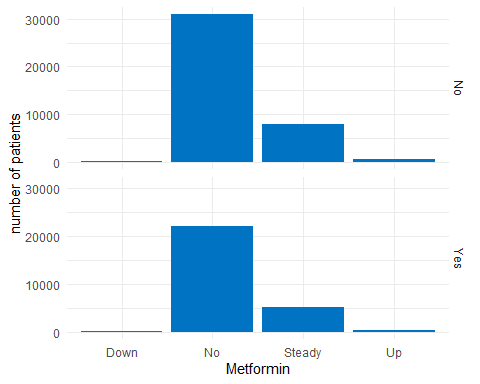
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = admission\_source\_id\_enc), fill = "#0073c2ff") +  
 xlab("Admission type") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



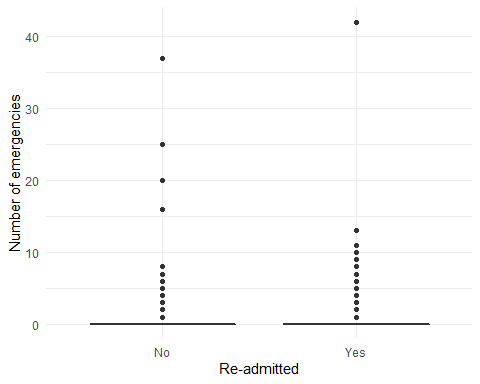
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = race), fill = "#0073c2ff") +  
 xlab("Race") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



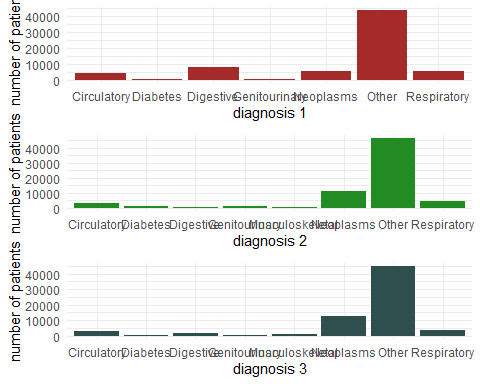
ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = metformin), fill = "#0073c2ff") +  
 xlab("Metformin") +  
 ylab("number of patients") +  
 facet\_grid(readmitted\_flag ~ .) +  
 theme\_minimal()



ggplot(data=post\_model\_ds, aes(x=readmitted\_flag, y=number\_emergency)) +   
 geom\_boxplot() +  
 xlab("Re-admitted") +  
 ylab("Number of emergencies") +  
 theme\_minimal()



diag\_1\_dist\_plot <- ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = diag\_1\_enc), fill = "brown") +  
 xlab("diagnosis 1") +  
 ylab("number of patients") +  
 theme\_minimal()  
  
diag\_2\_dist\_plot <- ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = diag\_2\_enc), fill = "forestgreen") +  
 xlab("diagnosis 2") +  
 ylab("number of patients") +  
 theme\_minimal()  
  
  
diag\_3\_dist\_plot <- ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = diag\_3\_enc), fill = "darkslategrey") +  
 xlab("diagnosis 3") +  
 ylab("number of patients") +  
 theme\_minimal()  
  
grid.arrange(diag\_1\_dist\_plot, diag\_2\_dist\_plot,diag\_3\_dist\_plot, ncol=1)



admission\_source\_id\_dist\_plot <- ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = admission\_source\_id\_enc), fill = "#0073c2ff") +  
 xlab("Admission source distribution") +  
 ylab("Number of patients") +  
 theme\_minimal()  
  
  
admission\_type\_id\_dist\_plot <- ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = admission\_type\_id\_enc), fill = "darkmagenta") +  
 xlab("Admission type distribution") +  
 ylab("Number of patients") +  
 theme\_minimal()  
  
age\_distribution\_plot <- ggplot(data = post\_model\_ds)+  
 geom\_bar(aes(x = age\_enc), fill = "deepskyblue4") +  
 xlab("Age distribution") +  
 ylab("Number of patients") +  
 theme\_minimal()  
  
  
grid.arrange(admission\_source\_id\_dist\_plot, admission\_type\_id\_dist\_plot,age\_distribution\_plot,ncol=2)

