```r
#Loading the training Data Set
sampledata <- read.csv("C:\\Users\\Satya Praveen\\Desktop\\coding\\data-challenge\\data-challenge\\training_data.csv",sep=',')
head(sampledata)
```

```
##   encounter_id patient_nbr            race gender     age weight
## 1      2278392     8222157       Caucasian Female  [0-10)      ?
## 2       149190    55629189       Caucasian Female [10-20)      ?
## 3        64410    86047875 AfricanAmerican Female [20-30)      ?
## 4       500364    82442376       Caucasian   Male [30-40)      ?
## 5        16680    42519267       Caucasian   Male [40-50)      ?
## 6        35754    82637451       Caucasian   Male [50-60)      ?
##   admission_type_id discharge_disposition_id admission_source_id
## 1                 6                       25                   1
## 2                 1                        1                   7
## 3                 1                        1                   7
## 4                 1                        1                   7
## 5                 1                        1                   7
## 6                 2                        1                   2
##   time_in_hospital payer_code        medical_specialty num_lab_procedures
## 1                1          ? Pediatrics-Endocrinology                 41
## 2                3          ?                         ?                 59
## 3                2          ?                         ?                 11
## 4                2          ?                         ?                 44
## 5                1          ?                         ?                 51
## 6                3          ?                         ?                 31
##   num_procedures num_medications number_outpatient number_emergency
## 1              0               1                 0                0
## 2              0              18                 0                0
## 3              5              13                 2                0
## 4              1              16                 0                0
## 5              0               8                 0                0
## 6              6              16                 0                0
##   number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
## 1                0 250.83      ?      ?                1          None
## 2                0    276 250.01    255                9          None
## 3                1    648    250    V27                6          None
## 4                0      8 250.43    403                7          None
## 5                0    197    157    250                5          None
## 6                0    414    411    250                9          None
##   A1Cresult metformin repaglinide nateglinide chlorpropamide glimepiride
## 1      None        No          No          No             No          No
## 2      None        No          No          No             No          No
## 3      None        No          No          No             No          No
## 4      None        No          No          No             No          No
## 5      None        No          No          No             No          No
## 6      None        No          No          No             No          No
##   acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone
## 1            No        No        No          No           No            No
## 2            No        No        No          No           No            No
```

```
## 3            No    Steady        No          No           No            No
## 4            No        No        No          No           No            No
## 5            No    Steady        No          No           No            No
## 6            No        No        No          No           No            No
##   acarbose miglitol troglitazone tolazamide examide citoglipton insulin
## 1       No       No           No         No      No          No      No
## 2       No       No           No         No      No          No      Up
## 3       No       No           No         No      No          No      No
## 4       No       No           No         No      No          No      Up
## 5       No       No           No         No      No          No  Steady
## 6       No       No           No         No      No          No  Steady
##   glyburide.metformin glipizide.metformin glimepiride.pioglitazone
## 1                  No                  No                       No
## 2                  No                  No                       No
## 3                  No                  No                       No
## 4                  No                  No                       No
## 5                  No                  No                       No
## 6                  No                  No                       No
##   metformin.rosiglitazone metformin.pioglitazone change diabetesMed
## 1                      No                     No     No          No
## 2                      No                     No     Ch         Yes
## 3                      No                     No     No         Yes
## 4                      No                     No     Ch         Yes
## 5                      No                     No     Ch         Yes
## 6                      No                     No     No         Yes
##   readmitted
## 1          N
## 2          N
## 3          N
## 4          N
## 5          N
## 6          N
```

Replacing the missing values with NA

```
sampledata[sampledata=="?"]<-NA
head(sampledata)
```

```
##   encounter_id patient_nbr            race gender     age weight
## 1      2278392     8222157       Caucasian Female  [0-10)   <NA>
## 2       149190    55629189       Caucasian Female [10-20)   <NA>
## 3        64410    86047875 AfricanAmerican Female [20-30)   <NA>
## 4       500364    82442376       Caucasian   Male [30-40)   <NA>
## 5        16680    42519267       Caucasian   Male [40-50)   <NA>
## 6        35754    82637451       Caucasian   Male [50-60)   <NA>
##   admission_type_id discharge_disposition_id admission_source_id
## 1                 6                      25                   1
## 2                 1                       1                   7
## 3                 1                       1                   7
## 4                 1                       1                   7
```

```
## 5                 1                       1                   7
## 6                 2                       1                   2
##   time_in_hospital payer_code        medical_specialty num_lab_procedures
## 1                1       <NA> Pediatrics-Endocrinology                 41
## 2                3       <NA>                      <NA>                59
## 3                2       <NA>                      <NA>                11
## 4                2       <NA>                      <NA>                44
## 5                1       <NA>                      <NA>                51
## 6                3       <NA>                      <NA>                31
##   num_procedures num_medications number_outpatient number_emergency
## 1              0               1                 0                0
## 2              0              18                 0                0
## 3              5              13                 2                0
## 4              1              16                 0                0
## 5              0               8                 0                0
## 6              6              16                 0                0
##   number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
## 1                0 250.83   <NA>   <NA>                1          None
## 2                0    276 250.01    255                9          None
## 3                1    648    250    V27                6          None
## 4                0      8 250.43    403                7          None
## 5                0    197    157    250                5          None
## 6                0    414    411    250                9          None
##   A1Cresult metformin repaglinide nateglinide chlorpropamide glimepiride
## 1      None        No          No          No             No          No
## 2      None        No          No          No             No          No
## 3      None        No          No          No             No          No
## 4      None        No          No          No             No          No
## 5      None        No          No          No             No          No
## 6      None        No          No          No             No          No
##   acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone
## 1            No        No        No          No           No            No
## 2            No        No        No          No           No            No
## 3            No    Steady        No          No           No            No
## 4            No        No        No          No           No            No
## 5            No    Steady        No          No           No            No
## 6            No        No        No          No           No            No
##   acarbose miglitol troglitazone tolazamide examide citoglipton insulin
## 1       No       No           No         No      No          No      No
## 2       No       No           No         No      No          No      Up
## 3       No       No           No         No      No          No      No
## 4       No       No           No         No      No          No      Up
## 5       No       No           No         No      No          No  Steady
## 6       No       No           No         No      No          No  Steady
##   glyburide.metformin glipizide.metformin glimepiride.pioglitazone
## 1                  No                  No                       No
## 2                  No                  No                       No
## 3                  No                  No                       No
## 4                  No                  No                       No
## 5                  No                  No                       No
```

```
## 6                   No                   No                         No
##   metformin.rosiglitazone metformin.pioglitazone change diabetesMed
## 1                      No                     No     No          No
## 2                      No                     No     Ch         Yes
## 3                      No                     No     No         Yes
## 4                      No                     No     Ch         Yes
## 5                      No                     No     Ch         Yes
## 6                      No                     No     No         Yes
##   readmitted
## 1          N
## 2          N
## 3          N
## 4          N
## 5          N
## 6          N
```

Dimensions of the training data

```
dim(sampledata)
```

```
## [1] 81414    50
```

To see how many NA values are there in each column in the given data.

```
library(dplyr)
```

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

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

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

```
sampledata %>%
select(everything()) %>%  # replace to your needs
summarise_all(funs(sum(is.na(.))))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
##   encounter_id patient_nbr race gender age weight admission_type_id
## 1            0           0 1813      0   0  78844                 0
##   discharge_disposition_id admission_source_id time_in_hospital payer_code
## 1                        0                   0                0      32231
##   medical_specialty num_lab_procedures num_procedures num_medications
## 1             39935                  0              0               0
##   number_outpatient number_emergency number_inpatient diag_1 diag_2 diag_3
## 1                 0               0                0     18    288   1125
```

```
##   number_diagnoses max_glu_serum A1Cresult metformin repaglinide
## 1                0             0         0         0           0
##   nateglinide chlorpropamide glimepiride acetohexamide glipizide glyburide
## 1           0              0           0             0         0         0
##   tolbutamide pioglitazone rosiglitazone acarbose miglitol troglitazone
## 1           0            0             0        0        0            0
##   tolazamide examide citoglipton insulin glyburide.metformin
## 1          0       0           0       0                   0
##   glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone
## 1                   0                         0                       0
##   metformin.pioglitazone change diabetesMed readmitted
## 1                      0      0           0          0
```

So we have 7 variables having NA values which are: Race, Payer_code,medical_speciality, weight, diag-1,diag-2,diag-3 Removing the Weight variable which has more than 78844 missing values.

```
sampledata$weight <- NULL
```

Replacing the odd values like V45, E932 with NA in all the three variables(diag-1,diag-2,diag-3)

```
sampledata$diag_3 <- gsub("V.*",NA, sampledata$diag_3)
sampledata$diag_2 <- gsub("V.*",NA, sampledata$diag_2)
sampledata$diag_1 <- gsub("V.*",NA, sampledata$diag_1)

sampledata$diag_3 <- gsub("E.*",NA, sampledata$diag_3)
sampledata$diag_2 <- gsub("E.*",NA, sampledata$diag_2)
sampledata$diag_1 <- gsub("E.*",NA, sampledata$diag_1)
```

FOr diag-3 variable: checking whether the variable is numeric or not

```
is.numeric(sampledata$diag_3)

## [1] FALSE

#converting to numeric
sampledata$diag_3 <- as.numeric(sampledata$diag_3)
is.numeric(sampledata$diag_3)

## [1] TRUE

# impute (replace NA values with mean)
sampledata$diag_3[is.na(sampledata$diag_3)] <- mean(sampledata$diag_3, na.rm
= T)
```

For diag-2

```
sampledata$diag_2 <- as.numeric(sampledata$diag_2) #converting to numeric
is.numeric(sampledata$diag_2)

## [1] TRUE
```

```
#impute
sampledata$diag_2[is.na(sampledata$diag_2)] <- mean(sampledata$diag_2, na.rm
= T)
```

For diag-1

```
sampledata$diag_1 <- as.numeric(sampledata$diag_1) #converting to numeric
is.numeric(sampledata$diag_1)

## [1] TRUE

#impute
sampledata$diag_1[is.na(sampledata$diag_1)] <- mean(sampledata$diag_1, na.rm
= T)
```

Replacing all NA values in the categorical data with None #### For variable - Race

```
set.seed(1234)
# Get levels and add "None" level
levels <- levels(sampledata$race)
levels[length(levels) + 1] <- "None"

# refactor Race to include "None" as a factor level and replace NA with
"None"
sampledata$race <- factor(sampledata$race, levels = levels)
sampledata$race[is.na(sampledata$race)] <- "None"
```

For variable - payer_code

```
set.seed(1235)

# Get levels and add "None" level
levels <- levels(sampledata$payer_code)
levels[length(levels) + 1] <- "None"

# refactor payer_code to include "None" as a factor level and replace NA with
"None"
sampledata$payer_code <- factor(sampledata$payer_code, levels = levels)
sampledata$payer_code[is.na(sampledata$payer_code)] <- "None"
```

For variable - medical_speciality

```
set.seed(1236)
# Get levels and add "None"
levels <- levels(sampledata$medical_specialty)
levels[length(levels) + 1] <- "None"

# refactor medical_speciality to include "None" as a factor level and replace
NA with "None"
sampledata$medical_specialty <- factor(sampledata$medical_specialty, levels =
levels)
sampledata$medical_specialty[is.na(sampledata$medical_specialty)] <- "None"
```

dealing with categorical variables

```
convert <- c(3,4,5,10,11,22:48)
sampledata[,convert] <- data.frame(apply(sampledata[convert], 2, as.factor))

# checking whether the columns are converted to factor type
is.factor(sampledata[,23])

## [1] TRUE

is.factor(sampledata[,20])

## [1] FALSE
```

Now let us consider the clening of test data.

```
################################ TESTING DATA
#################################
testdata = read.csv("C:\\Users\\Satya Praveen\\Desktop\\coding\\data-
challenge\\data-challenge\\test_data.csv",sep = ',')
testdata[testdata=="?"]<-NA
head(testdata)

##   encounter_id patient_nbr            race gender      age weight
## 1        15738    63555939       Caucasian Female [90-100)   <NA>
## 2        62256    49726791 AfricanAmerican Female  [60-70)   <NA>
## 3       150006    22864131            <NA> Female  [50-60)   <NA>
## 4       183930   107400762       Caucasian Female  [80-90)   <NA>
## 5       248916   115196778       Caucasian Female  [50-60)   <NA>
## 6       260166    80845353       Caucasian Female  [70-80)   <NA>
##   admission_type_id discharge_disposition_id admission_source_id
## 1                 3                        3                   4
## 2                 3                        1                   2
## 3                 2                        1                   4
## 4                 2                        6                   1
## 5                 1                        1                   1
## 6                 1                        1                   7
##   time_in_hospital payer_code      medical_specialty num_lab_procedures
## 1               12       <NA>       InternalMedicine                33
## 2                1       <NA>                   <NA>                49
## 3                2       <NA>                   <NA>                66
## 4               11       <NA>                   <NA>                42
## 5                2       <NA>        Surgery-General                25
## 6                6       <NA> Family/GeneralPractice                27
##   num_procedures num_medications number_outpatient number_emergency
## 1              3              18                 0                0
## 2              5               2                 0                0
## 3              1              19                 0                0
## 4              2              19                 0                0
## 5              2              11                 0                0
## 6              0              16                 0                0
##   number_inpatient diag_1 diag_2 diag_3 number_diagnoses max_glu_serum
```

```
## 1               0    434    198    486                8          None
## 2               0    518    998    627                8          None
## 3               0    410    427    428                7          None
## 4               0    V57    715    V43                8          None
## 5               0    996    585 250.01                3          None
## 6               0    996    999 250.01                8          None
##   A1Cresult metformin repaglinide nateglinide chlorpropamide glimepiride
## 1      None        No          No          No             No          No
## 2      None        No          No          No             No          No
## 3      None        No          No          No             No          No
## 4      None        No          No          No             No          No
## 5      None        No          No          No             No          No
## 6      None        No          No          No             No          No
##   acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone
## 1            No        No        No          No           No        Steady
## 2            No        No        No          No           No            No
## 3            No        No        No          No           No            No
## 4            No        No        No          No           No            No
## 5            No        No        No          No           No            No
## 6            No        No        No          No           No            No
##   acarbose miglitol troglitazone tolazamide examide citoglipton insulin
## 1       No       No           No         No      No          No  Steady
## 2       No       No           No         No      No          No  Steady
## 3       No       No           No         No      No          No    Down
## 4       No       No           No         No      No          No      No
## 5       No       No           No         No      No          No  Steady
## 6       No       No           No         No      No          No  Steady
##   glyburide.metformin glipizide.metformin glimepiride.pioglitazone
## 1                  No                  No                       No
## 2                  No                  No                       No
## 3                  No                  No                       No
## 4                  No                  No                       No
## 5                  No                  No                       No
## 6                  No                  No                       No
##   metformin.rosiglitazone metformin.pioglitazone change diabetesMed
## 1                      No                     No     Ch         Yes
## 2                      No                     No     No         Yes
## 3                      No                     No     Ch         Yes
## 4                      No                     No     No          No
## 5                      No                     No     No         Yes
## 6                      No                     No     No         Yes
```

```r
# dimensions of the test data
dim(testdata)
```

```
## [1] 20352    49
```

To see how many NA values are there in each column in the given data.

```
library(dplyr)
testdata %>%
  select(everything()) %>%  # replace to your needs
  summarise_all(funs(sum(is.na(.))))
```

```
##   encounter_id patient_nbr race gender age weight admission_type_id
## 1            0           0  460      0   0  19725                 0
##   discharge_disposition_id admission_source_id time_in_hospital payer_code
## 1                        0                   0                0       8025
##   medical_specialty num_lab_procedures num_procedures num_medications
## 1             10014                  0              0               0
##   number_outpatient number_emergency number_inpatient diag_1 diag_2 diag_3
## 1                 0                0                0      3     70    298
##   number_diagnoses max_glu_serum A1Cresult metformin repaglinide
## 1                0             0         0         0           0
##   nateglinide chlorpropamide glimepiride acetohexamide glipizide glyburide
## 1           0              0           0             0         0         0
##   tolbutamide pioglitazone rosiglitazone acarbose miglitol troglitazone
## 1           0            0             0        0        0            0
##   tolazamide examide citoglipton insulin glyburide.metformin
## 1          0       0           0       0                   0
##   glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone
## 1                   0                        0                       0
##   metformin.pioglitazone change diabetesMed
## 1                      0      0           0
```

Removing the Weight variable which has more than 19725 missing values.

```
testdata$weight <- NULL
```

Replacing the odd values like V45, E932 and manymore irrelevant numbers with NA in all the three variables

```
testdata$diag_3 <- gsub("V.*",NA, testdata$diag_3)
testdata$diag_2 <- gsub("V.*",NA, testdata$diag_2)
testdata$diag_1 <- gsub("V.*",NA, testdata$diag_1)

testdata$diag_3 <- gsub("E.*",NA, testdata$diag_3)
testdata$diag_2 <- gsub("E.*",NA, testdata$diag_2)
testdata$diag_1 <- gsub("E.*",NA, testdata$diag_1)
```

Imputing the diag-1,2,3 columns missing values with their mean values

```r
#FOr diag-3 variable
#Checking whether the variable is numeric or not
is.numeric(testdata$diag_3)
```

```
## [1] FALSE
```

```r
testdata$diag_3 <- as.numeric(testdata$diag_3)
#converting to numeric
is.numeric(testdata$diag_3)
```

```
## [1] TRUE
```

```r
#impute (replace with mean)
testdata$diag_3[is.na(testdata$diag_3)] <- mean(testdata$diag_3, na.rm = T)

#For diag-2
testdata$diag_2 <- as.numeric(testdata$diag_2) #converting to numeric
is.numeric(testdata$diag_2)
```

```
## [1] TRUE
```

```r
#impute
testdata$diag_2[is.na(testdata$diag_2)] <- mean(testdata$diag_2, na.rm = T)

#For diag-1
testdata$diag_1 <- as.numeric(testdata$diag_1) #converting to numeric
is.numeric(testdata$diag_1)
```

```
## [1] TRUE
```

```r
#impute
testdata$diag_1[is.na(testdata$diag_1)] <- mean(testdata$diag_1, na.rm = T)
```

replacing all NA values in the categorical data with None

For variable - Race

```r
set.seed(45)
# Get levels and add "None"
levels <- levels(testdata$race)
levels[length(levels) + 1] <- "None"
```

Refactor Race to include “None” as a factor level and replace NA with “None”

```r
testdata$race <- factor(testdata$race, levels = levels)
testdata$race[is.na(testdata$race)] <- "None"
```

Variable levels modification

```
#### For variable - payer_code
set.seed(451)

# Get levels and add "None"
levels <- levels(testdata$payer_code)
levels[length(levels) + 1] <- "None"

# refactor Payer_code to include "None" as a factor level and replace NA with
"None"
testdata$payer_code <- factor(testdata$payer_code, levels = levels)
testdata$payer_code[is.na(testdata$payer_code)] <- "None"


#### For variable - medical_speciality
set.seed(452)

# Get levels and add "None"
levels <- levels(testdata$medical_specialty)
levels[length(levels) + 1] <- "None"

# refactor medical_speicality to include "None" as a factor level and replace
NA with "None"
testdata$medical_specialty <- factor(testdata$medical_specialty, levels =
levels)
testdata$medical_specialty[is.na(testdata$medical_specialty)] <- "None"
```

dealing with categorical variables

```
convert <- c(3,4,5,10,11,22:48)
testdata[,convert] <- data.frame(apply(testdata[convert], 2, as.factor))

is.factor(testdata[,23]) # checking whether the columns are converted to
factor type

## [1] TRUE

is.factor(testdata[,20])

## [1] FALSE
```

We need to handle missing values and categorical features before feeding the data into a machine learning algorithm, because the mathematics underlying most machine learning models assumes that the data is numerical and contains no missing values.

```
############################ MODELLING ######################
#install.packages("h2o")
library(h2o)

## Warning: package 'h2o' was built under R version 3.4.4

##
```

```
## ----------------------------------------------------------------------
##
## Your next step is to start H2O:
##     > h2o.init()
##
## For H2O package documentation, ask for help:
##     > ??h2o
##
## After starting H2O, you can use the Web UI at http://localhost:54321
## For more information visit http://docs.h2o.ai
##
## ----------------------------------------------------------------------

##
## Attaching package: 'h2o'

## The following objects are masked from 'package:stats':
##
##     cor, sd, var

## The following objects are masked from 'package:base':
##
##     %*%, %in%, &&, ||, apply, as.factor, as.numeric, colnames,
##     colnames<-, ifelse, is.character, is.factor, is.numeric, log,
##     log10, log1p, log2, round, signif, trunc
```

```
#To launch the H2O cluster
localH2O <- h2o.init(nthreads = -1)
```

```
##
## H2O is not running yet, starting it now...
##
## Note:  In case of errors look at the following log files:
##
C:\Users\SATYAP~1\AppData\Local\Temp\Rtmp2PrCKc/h2o_Satya_Praveen_started_fro
m_r.out
##
C:\Users\SATYAP~1\AppData\Local\Temp\Rtmp2PrCKc/h2o_Satya_Praveen_started_fro
m_r.err
##
##
## Starting H2O JVM and connecting: .. Connection successful!
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         8 seconds 319 milliseconds
##     H2O cluster timezone:       America/Chicago
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.18.0.11
##     H2O cluster version age:    1 month and 11 days
##     H2O cluster name:           H2O_started_from_R_Satya_Praveen_bwl991
```

```
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   0.84 GB
##     H2O cluster total cores:    4
##     H2O cluster allowed cores:  4
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     H2O API Extensions:         Algos, AutoML, Core V3, Core V4
##     R Version:                  R version 3.4.3 (2017-11-30)
```

```
# loading the training data into H2o environment from R
train.h2o <- as.h2o(sampledata)
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

```
test.h2o <- as.h2o(testdata)
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

```
colnames(train.h2o)
```

```
##  [1] "encounter_id"             "patient_nbr"
##  [3] "race"                     "gender"
##  [5] "age"                      "admission_type_id"
##  [7] "discharge_disposition_id" "admission_source_id"
##  [9] "time_in_hospital"         "payer_code"
## [11] "medical_specialty"        "num_lab_procedures"
## [13] "num_procedures"           "num_medications"
## [15] "number_outpatient"        "number_emergency"
## [17] "number_inpatient"         "diag_1"
## [19] "diag_2"                   "diag_3"
## [21] "number_diagnoses"         "max_glu_serum"
## [23] "A1Cresult"                "metformin"
## [25] "repaglinide"              "nateglinide"
## [27] "chlorpropamide"           "glimepiride"
## [29] "acetohexamide"            "glipizide"
## [31] "glyburide"                "tolbutamide"
## [33] "pioglitazone"             "rosiglitazone"
## [35] "acarbose"                 "miglitol"
## [37] "troglitazone"             "tolazamide"
## [39] "examide"                  "citoglipton"
```

```
## [41] "insulin"                   "glyburide.metformin"
## [43] "glipizide.metformin"       "glimepiride.pioglitazone"
## [45] "metformin.rosiglitazone"   "metformin.pioglitazone"
## [47] "change"                    "diabetesMed"
## [49] "readmitted"
```

GBM(Gradient Boosting) Model

Removing the encounter id and patient number from the dependent variables as they don't add any extra information for training the model.

```
y.dep <- 49 #deppendent variable

x.indep <- c(3:48) #independent variables
```

GBM model on the training data set in H2o environment

```
gbm.model <- h2o.gbm(y=y.dep, x=x.indep, training_frame = train.h2o, ntrees =
1000, max_depth = 4, learn_rate = 0.01, seed = 1122)

## Warning in .h2o.startModelJob(algo, params, h2oRestApiVersion): Dropping
bad and constant columns: [citoglipton, examide].

##
  |
  |                                                                      |   0%
  |
  |                                                                      |   1%
  |
  |=                                                                     |   1%
  |
  |=                                                                     |   2%
  |
  |==                                                                    |   3%
  |
  |==                                                                    |   4%
  |
  |===                                                                   |   4%
  |
  |===                                                                   |   5%
  |
  |====                                                                  |   6%
  |
  |====                                                                  |   7%
  |
  |=====                                                                 |   8%
  |
  |======                                                                |   9%
  |
  |======                                                                |  10%
  |
```

```
|=======                                                 |  10%
|
|=======                                                 |  11%
|
|========                                                |  12%
|
|========                                                |  13%
|
|=========                                               |  13%
|
|=========                                               |  14%
|
|=========                                               |  15%
|
|==========                                              |  15%
|
|==========                                              |  16%
|
|===========                                             |  17%
|
|============                                            |  18%
|
|============                                            |  19%
|
|=============                                           |  20%
|
|==============                                          |  21%
|
|==============                                          |  22%
|
|===============                                         |  23%
|
|================                                        |  24%
|
|================                                        |  25%
|
|=================                                       |  25%
|
|=================                                       |  26%
|
|==================                                      |  27%
|
|==================                                      |  28%
|
|===================                                     |  28%
|
|===================                                     |  29%
|
|====================                                    |  30%
|
```

```
|==================                                                    |  31%
|
|==================                                                    |  32%
|
|====================                                                  |  32%
|
|====================                                                  |  33%
|
|======================                                                |  34%
|
|======================                                                |  35%
|
|======================                                                |  35%
|
|========================                                              |  36%
|
|========================                                              |  37%
|
|=========================                                             |  38%
|
|=========================                                             |  39%
|
|=========================                                             |  39%
|
|==========================                                            |  40%
|
|==========================                                            |  41%
|
|==========================                                            |  42%
|
|===========================                                           |  42%
|
|===========================                                           |  43%
|
|============================                                          |  44%
|
|============================                                          |  45%
|
|============================                                          |  46%
|
|=============================                                         |  47%
|
|=============================                                         |  48%
|
|=============================                                         |  49%
|
|==============================                                        |  50%
|
|==============================                                        |  50%
|
```

```
|================================                  |  51%
|
|=================================                 |  52%
|
|==================================                |  53%
|
|===================================               |  54%
|
|====================================              |  55%
|
|=====================================             |  56%
|
|======================================            |  57%
|
|=======================================           |  58%
|
|========================================          |  58%
|
|=========================================         |  59%
|
|==========================================        |  60%
|
|===========================================       |  61%
|
|============================================      |  62%
|
|============================================      |  62%
|
|=============================================     |  63%
|
|==============================================    |  64%
|
|===============================================   |  65%
|
|================================================  |  66%
|
|================================================= |  67%
|
|================================================= |  68%
|
|================================================== |  69%
|
|=================================================  |  70%
|
|=================================================  |  70%
|
|================================================== |  71%
|
|=================================================== |  72%
|
```

```
|=============================================                  |  73%
|
|==============================================                 |  74%
|
|===============================================                |  75%
|
|================================================               |  76%
|
|=================================================              |  77%
|
|==================================================             |  78%
|
|===================================================            |  79%
|
|====================================================           |  80%
|
|=====================================================          |  81%
|
|======================================================         |  81%
|
|=======================================================        |  82%
|
|========================================================       |  83%
|
|=========================================================      |  84%
|
|==========================================================     |  85%
|
|===========================================================    |  86%
|
|============================================================   |  87%
|
|=============================================================  |  88%
|
|==============================================================  |  89%
|
|===============================================================  |  90%
|
|================================================================ |  91%
|
|================================================================= |  92%
|
|================================================================== |  93%
|
|================================================================== |  93%
|
|================================================================== |  94%
|
|================================================================== |  95%
|
```

```
  |=============================================================   |  96%
  |
  |==============================================================  |  97%
  |
  |==============================================================  |  98%
  |
  |============================================================== |  98%
  |
  |============================================================== |  99%
  |
  |================================================================| 100%
```

```
# predicting the readmission results for the test data set
predict.gbm <- as.data.frame(h2o.predict(gbm.model, test.h2o))
```

```
##
  |
  |                                                                 |   0%
  |
  |================================================================| 100%
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): Test/
## Validation dataset column 'medical_specialty' has levels not trained on:
## [Proctology, Surgery-PlasticwithinHeadandNeck]

## Warning in doTryCatch(return(expr), name, parentenv, handler): Test/
## Validation dataset column 'tolazamide' has levels not trained on: [Up]
```

```
summary(predict.gbm)
```

```
##  predict         N                Y
##  N:15776   Min.   :0.2373   Min.   :0.01191
##  Y: 4576   1st Qu.:0.8671   1st Qu.:0.06825
##            Median :0.9060   Median :0.09399
##            Mean   :0.8890   Mean   :0.11105
##            3rd Qu.:0.9318   3rd Qu.:0.13289
##            Max.   :0.9881   Max.   :0.76271
```

Shutting down the h2o environment

```
h2o.shutdown(prompt = FALSE)
```

```
## [1] TRUE
```

Random Forests and othe models are also used but they are taking longer execution times because of large dataset. For reference we can see the code below.

The traditional models in the normal R environment and methods are taking too much time to predict. Hence I have used H2o Environment linked with R for dealing with large datasets.

Final Predicted values of readmission of diabetes patients for the test dataset. Since the final outcomes of the testdata file are not given here, the confusion matrix and the accuracy scores of the model on the testdata couldn't be calculated.

```
# Final Output predictions dataframe containing the row
identifiers(encounter_id and patient number)
Final_Output <- data.frame(encounter_id = testdata$encounter_id, readmitted =
predict.gbm$predict)

#  Exporting the csv file and saving it.
write.csv(Final_Output, file = "Burra_Praveen.csv", row.names = F)
```