E6690_Fall_2019_Project_dso2119_diabetes

Dwiref Oza

12/18/2019

Introduction

As per the 2014 National DM Statistics Report by the Centers for Disease Control and Prevention, an estimated 9.3% of the US population is affected by diabetes mellitus, 28% of which remains undiagnosed. The average rate of readmission for a hospital patient hovers between 8.5% to 13.5%, while for diabetes patients this figure is worryingly much higher. Thirty-day readmission for diabetes patients has been charted to lie between 14.4% and 22.7%. A study by Strack et. al investigated the impact of HbA1c measurement on readmission rates by analyzing a databse of 70,000 patient records.

Dataset and Paper

The data used in the study was submitted to the UC Irvine Machine Learning Repository by the authors on behalf of the Center for Clinical and Translational Research at Virginia Commonwealth University. The dataset has records of 10 years worth of in-atient, out-patient and emergency patient data from 1999 - 2008. Each entry has 50 features, ranging from hormone levels to biological indicators relevant to diabetes mellitus, along with descriptors such as patient age, race, gender, age, duration of hospital care, specialty of the attending physician, etc. All features are relevant to predicting the rate of readmission, however the raw data has gaps in some of these fields, which reduces the possible tenable predictors available. One of the key columns in the data which was the primary thrust of the study, is the testing of the HbA1c blood sugar levels. For a readmission prediction task, there are 3 possible outcomes:

1. No readmission
2. Readmission in less than 30 days
3. Readmission post a 30 day period

```
library(dplyr)
```

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

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

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

```
library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

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

library(ggplot2)
library(corrplot)

## corrplot 0.84 loaded

library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(caret)

## Loading required package: lattice

library(rpart)
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:psych':
##
##     outlier

## The following object is masked from 'package:ggplot2':
##
##     margin

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

```r
library(nnet)
library(e1071)
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(CORElearn)
library(lasso2)
```

```
## R Package to solve regression problems while imposing
##   an L1 constraint on the parameters. Based on S-plus Release 2.1
## Copyright (C) 1998, 1999
## Justin Lokhorst   <jlokhors@stats.adelaide.edu.au>
## Berwin A. Turlach <bturlach@stats.adelaide.edu.au>
## Bill Venables     <wvenable@stats.adelaide.edu.au>
##
## Copyright (C) 2002
## Martin Maechler <maechler@stat.math.ethz.ch>
```

```
##
## Attaching package: 'lasso2'
```

```
## The following object is masked from 'package:psych':
##
##     tr
```

```r
# read in data
filename <- 'diabetic_data.csv'
data <- read.table(filename, sep = ",", header = T, na.strings = "?")
head(data)
```

```
##   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
readmitted
## 1                      No                    No     No          No
NO
## 2                      No                    No     Ch         Yes
>30
## 3                      No                    No     No         Yes
NO
## 4                      No                    No     Ch         Yes
NO
## 5                      No                    No     Ch         Yes
NO
## 6                      No                    No     No         Yes
>30

#load(file = "data2.rdata")
```

Data Cleanup

No prediction task is complete without pruning the data so that it can be orderd and meaningful. Numerical data must be uniform, and if any columns are categorical, any and all values equivalent to missing, other or NA must be dealt with. Some rows of the dataset have NA values. Columns where a majority of such entries exist are best ignored since they cannot be representative of every patient. Columns 25 to 41 and 43 - 47 are thus discarded. Further, the encouter ID and payer code are discarded as well.

```
# PREPROCESSING, CLEANING

data <- select(data, -encounter_id, -patient_nbr, -weight,-(25:41),-
(43:47))

# exploratory analysis and plots
summary(data)

##               race                     gender           age
##  AfricanAmerican:19210   Female         :54708   [70-80):26068
##  Asian          :  641   Male           :47055   [60-70):22483
##  Caucasian      :76099   Unknown/Invalid:    3   [50-60):17256
##  Hispanic       : 2037                            [80-90):17197
##  Other          : 1506                            [40-50): 9685
##  NA's           : 2273                            [30-40): 3775
##                                                   (Other): 5302
##  admission_type_id discharge_disposition_id admission_source_id
##  Min.   :1.000     Min.   : 1.000           Min.   : 1.000
##  1st Qu.:1.000     1st Qu.: 1.000           1st Qu.: 1.000
##  Median :1.000     Median : 1.000           Median : 7.000
##  Mean   :2.024     Mean   : 3.716           Mean   : 5.754
##  3rd Qu.:3.000     3rd Qu.: 4.000           3rd Qu.: 7.000
##  Max.   :8.000     Max.   :28.000           Max.   :25.000
##
##  time_in_hospital   payer_code                medical_specialty
##  Min.   : 1.000   MC     :32439   InternalMedicine      :14635
##  1st Qu.: 2.000   HM     : 6274   Emergency/Trauma      : 7565
##  Median : 4.000   SP     : 5007   Family/GeneralPractice: 7440
##  Mean   : 4.396   BC     : 4655   Cardiology            : 5352
##  3rd Qu.: 6.000   MD     : 3532   Surgery-General       : 3099
##  Max.   :14.000   (Other): 9603   (Other)               :13726
##                   NA's   :40256   NA's                  :49949
##  num_lab_procedures num_procedures num_medications
number_outpatient
##  Min.   :  1.0      Min.   :0.00   Min.   : 1.00   Min.   : 0.0000

##  1st Qu.: 31.0      1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000

##  Median : 44.0      Median :1.00   Median :15.00   Median : 0.0000

##  Mean   : 43.1      Mean   :1.34   Mean   :16.02   Mean   : 0.3694
```

```
##  3rd Qu.: 57.0      3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000

##  Max.   :132.0      Max.   :6.00   Max.   :81.00   Max.   :42.0000

##

##  number_emergency  number_inpatient      diag_1          diag_2

##  Min.   : 0.0000   Min.   : 0.0000   428    : 6862   276    : 6752

##  1st Qu.: 0.0000   1st Qu.: 0.0000   414    : 6581   428    : 6662

##  Median : 0.0000   Median : 0.0000   786    : 4016   250    : 6071

##  Mean   : 0.1978   Mean   : 0.6356   410    : 3614   427    : 5036

##  3rd Qu.: 0.0000   3rd Qu.: 1.0000   486    : 3508   401    : 3736

##  Max.   :76.0000   Max.   :21.0000   (Other):77164   (Other):73151

##                                      NA's   :   21   NA's   :  358

##      diag_3      number_diagnoses max_glu_serum A1Cresult
insulin
##  250    :11555   Min.   : 1.000   >200: 1485    >7  : 3812
Down  :12218
##  401    : 8289   1st Qu.: 6.000   >300: 1264    >8  : 8216
No    :47383
##  276    : 5175   Median : 8.000   None:96420    None:84748
Steady:30849
##  428    : 4577   Mean   : 7.423   Norm: 2597    Norm: 4990
Up    :11316
##  427    : 3955   3rd Qu.: 9.000

##  (Other):66792   Max.   :16.000

##  NA's   : 1423

##  change     diabetesMed readmitted
##  Ch:47011   No :23403   <30:11357
##  No:54755   Yes:78363   >30:35545
##                         NO :54864
##
##
##
##
```

```
# time-in-hospital is positively correlated with number of lab
procedures,
# number of non-lab procedures, number of medications and number of
diagnoses
# number of emergency visits correlates with number of inpatient
visits

# fix some missing values
data$race[is.na(data$race)] <- "Other"
any(is.na(data$race)) # false
```

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

Categorizing ICD-9 codes

In the UCI dataset, columns 20, 21 and 22 signify diagnoses for patient visits. The values of these columns are the International Classification of Diseases (ICD-9) medical codes. The range of these values is from 001 to 999, which are too numermous and only serve to thin out the density of data. It would be much more useful to condense these codes into categorical variables that define the broad area of the diagnosis instead of the actual diagnosis itself. Thus, these can be reduced to the labels: 1. Circulatory 2. Respiratory 3. Digestive 4. Diabetes 5. Injury 6. Muscoskeletal 7. Genitourinary 8. Neoplasms 9. Other

Below is the code to achieve this.

```
# FEATURE EXTRACTION

data2 <- data

data2$diag_1 <- as.numeric(levels(data2$diag_1)[data2$diag_1])

## Warning: NAs introduced by coercion

data2$diag_2 <- as.numeric(levels(data2$diag_2)[data2$diag_2])

## Warning: NAs introduced by coercion

data2$diag_3 <- as.numeric(levels(data2$diag_3)[data2$diag_3])

## Warning: NAs introduced by coercion

# diagnosis1
data2$diagnosis_group <- factor( rep("other",nrow(data2)),ordered = F,

                                 levels =
c("circulatory","respiratory","Digestive","Diabetes","Injury",

"Musculoskeletal","Genitourinary","Neoplasms","other"))
data2$diagnosis_group[data2$diag_1>=390 & data2$diag_1 <= 459 |
data2$diag_1==785] <- "circulatory"
data2$diagnosis_group[data2$diag_1>=460 & data2$diag_1 <= 519 |
```

```
data2$diag_1==786] <- "respiratory"
data2$diagnosis_group[data2$diag_1>=520 & data2$diag_1 <= 579 |
data2$diag_1==787] <- "Digestive"
data2$diagnosis_group[data2$diag_1>=250 & data2$diag_1 < 251] <-
"Diabetes"
data2$diagnosis_group[data2$diag_1>800 & data2$diag_1 <= 999] <-
"Injury"
data2$diagnosis_group[data2$diag_1>=710 & data2$diag_1 <= 739] <-
"Musculoskeletal"
data2$diagnosis_group[data2$diag_1>=580 & data2$diag_1 <= 629 |
data2$diag_1==788] <- "Genitourinary"
data2$diagnosis_group[data2$diag_1>=140 & data2$diag_1 <= 239 |
data2$diag_1>=790 &
                        data2$diag_1 <= 799 | data2$diag_1==780 |
data2$diag_1>=240 & data2$diag_1 < 250 |
                        data2$diag_1>=251 & data2$diag_1 <= 279 |
data2$diag_1>=680 & data2$diag_1 <= 709 |
                        data2$diag_1>=001 & data2$diag_1 <= 139 |
data2$diag_1==781 |
                      data2$diag_1==782 | data2$diag_1==784] <-
"Neoplasms"

# diagnosis_2
data2$diagnosis_2 <- factor( rep("other",nrow(data2)),ordered = F,
                                levels =
c("circulatory","respiratory","Digestive","Diabetes","Injury",

"Musculoskeletal","Genitourinary","Neoplasms","other"))
data2$diagnosis_2[data2$diag_2>=390 & data2$diag_2 <= 459 |
data2$diag_2==785] <- "circulatory"
data2$diagnosis_2[data2$diag_2>=460 & data2$diag_2 <= 519 |
data2$diag_2==786] <- "respiratory"
data2$diagnosis_2[data2$diag_2>=520 & data2$diag_2 <= 579 |
data2$diag_2==787] <- "Digestive"
data2$diagnosis_2[data2$diag_2>=250 & data2$diag_2 < 251] <-
"Diabetes"
data2$diagnosis_2[data2$diag_2>800 & data2$diag_2 <= 999] <- "Injury"
data2$diagnosis_2[data2$diag_2>=710 & data2$diag_2 <= 739] <-
"Musculoskeletal"
data2$diagnosis_2[data2$diag_2>=580 & data2$diag_2 <= 629 |
data2$diag_2==788] <- "Genitourinary"
data2$diagnosis_2[data2$diag_2>=140 & data2$diag_2 <= 239 |
data2$diag_2>=790 &
                        data2$diag_2 <= 799 | data2$diag_2==780 |
data2$diag_2>=240 & data2$diag_2 < 250 |
                        data2$diag_2>=251 & data2$diag_2 <= 279 |
data2$diag_2>=680 & data2$diag_2 <= 709 |
                        data2$diag_2>=001 & data2$diag_2 <= 139 |
data2$diag_2==781 |
                        data2$diag_2==782 | data2$diag_2==784] <-
```

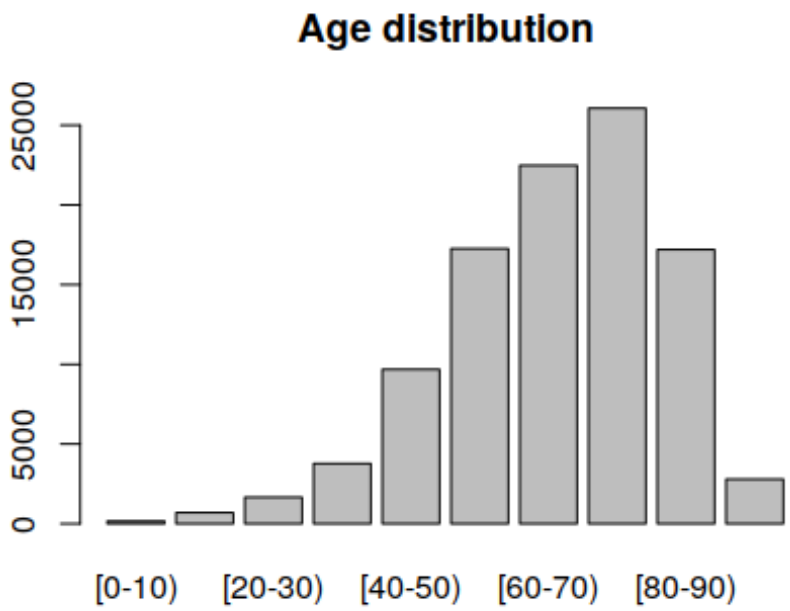
```
"Neoplasms"

# diagnosis_3
data2$diagnosis_3 <- factor( rep("other",nrow(data2)),ordered = F,
                                  levels =
c("circulatory","respiratory","Digestive","Diabetes","Injury",

"Musculoskeletal","Genitourinary","Neoplasms","other"))
data2$diagnosis_3[data2$diag_3>=390 & data2$diag_3 <= 459 |
data2$diag_3==785] <- "circulatory"
data2$diagnosis_3[data2$diag_3>=460 & data2$diag_3 <= 519 |
data2$diag_3==786] <- "respiratory"
data2$diagnosis_3[data2$diag_3>=520 & data2$diag_3 <= 579 |
data2$diag_3==787] <- "Digestive"
data2$diagnosis_3[data2$diag_3>=250 & data2$diag_3 < 251] <-
"Diabetes"
data2$diagnosis_3[data2$diag_3>800 & data2$diag_3 <= 999] <- "Injury"
data2$diagnosis_3[data2$diag_3>=710 & data2$diag_3 <= 739] <-
"Musculoskeletal"
data2$diagnosis_3[data2$diag_3>=580 & data2$diag_3 <= 629 |
data2$diag_3==788] <- "Genitourinary"
data2$diagnosis_3[data2$diag_3>=140 & data2$diag_3 <= 239 |
data2$diag_3>=790 &
                        data2$diag_3 <= 799 | data2$diag_3==780 |
data2$diag_3>=240 & data2$diag_3 < 250 |
                        data2$diag_3>=251 & data2$diag_3 <= 279 |
data2$diag_3>=680 & data2$diag_3 <= 709 |
                        data2$diag_3>=001 & data2$diag_3 <= 139 |
data2$diag_3==781 |
                        data2$diag_3==782 | data2$diag_3==784] <-
"Neoplasms"
# admission_source
data2$admission_source <- factor( rep("other",nrow(data2)),ordered =
F,
                             levels = c("clinic_referral",
"emergency","other"))
data2$admission_source[data2$admission_source_id==c(1,2,3)]<-
"clinic_referral"
data2$admission_source[data2$admission_source_id==7]<- "emergency"

# discharged_to
data2$discharged_to <- factor( rep("transferred",nrow(data2)),ordered
= F,
                                  levels = c("home",
"transferred","left_AMA"))
data2$discharged_to[data2$discharge_disposition_id==c(1,6,8)]<- "home"
data2$discharged_to[data2$discharge_disposition_id==7]<- "left_AMA"

data2 <- select(data2, -diag_1, -diag_2, -diag_3, -admission_type_id,
-discharge_disposition_id)
```

```
data2 <- select(data2, -medical_specialty)
data2 <- rename(data2, diag1 = diagnosis_group, diag2=diagnosis_2,
diag3 = diagnosis_3)

# payer_code
data2$payer_code2 <- factor( rep("other",nrow(data2)),ordered = F,
                             levels = c("other", "self_pay"))
data2$payer_code2[data2$payer_code=="SP"]<- "self_pay"
data2 <- select(data2, -payer_code)
data2 <- select(data2, -admission_source_id)
data2 <- rename(data2, payer_code=payer_code2)
```

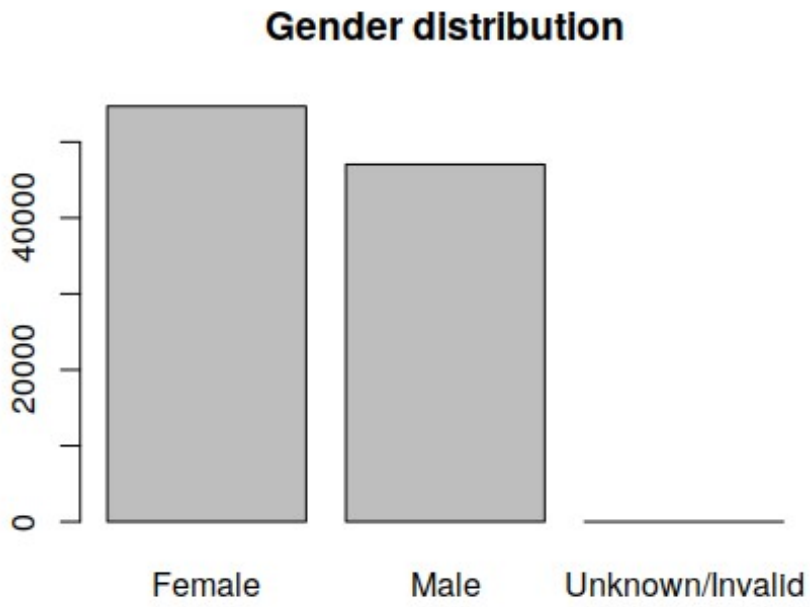
Data Visualization

To start with, here are the patient distributions by race, age, gender and their readmissions (or lack thereof).

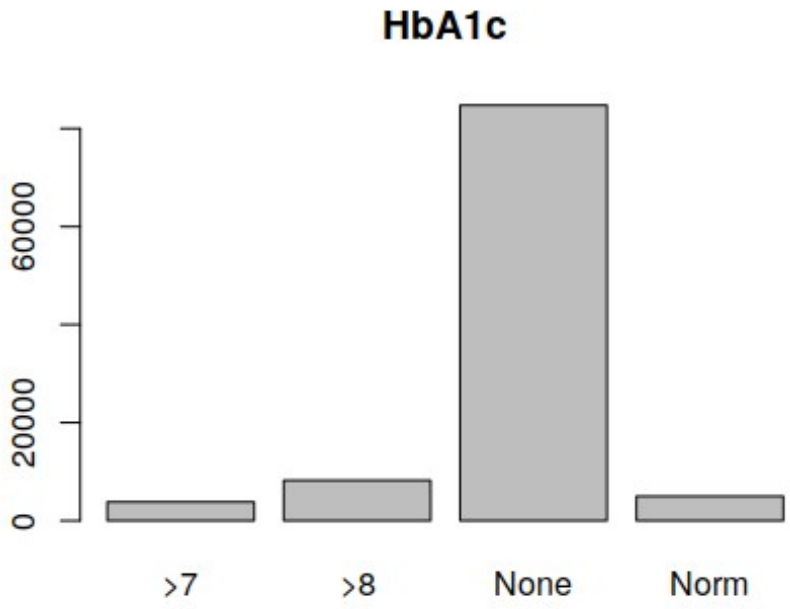
```
# variable distributions
plot(data$age, main = "Age distribution")
```



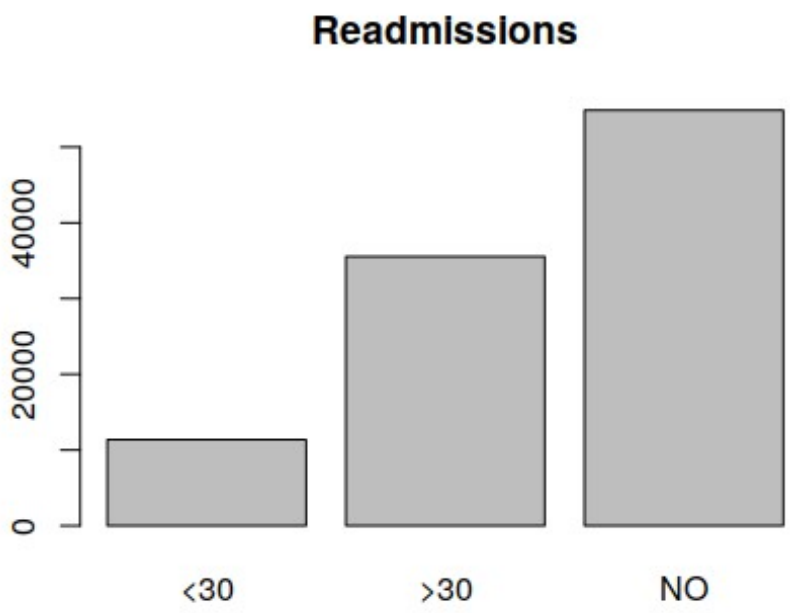

```
plot(data$gender, main = "Gender distribution")
```



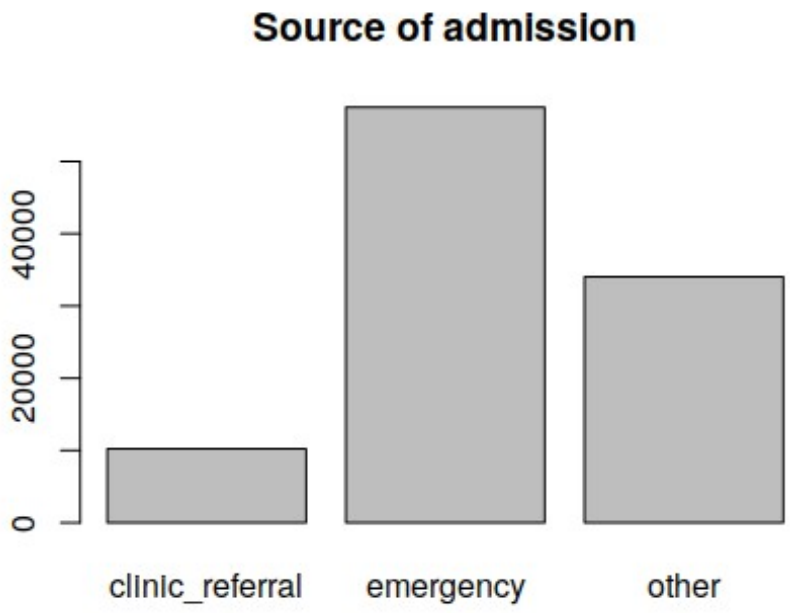

```
plot(data$A1Cresult, main = "HbA1c")
```



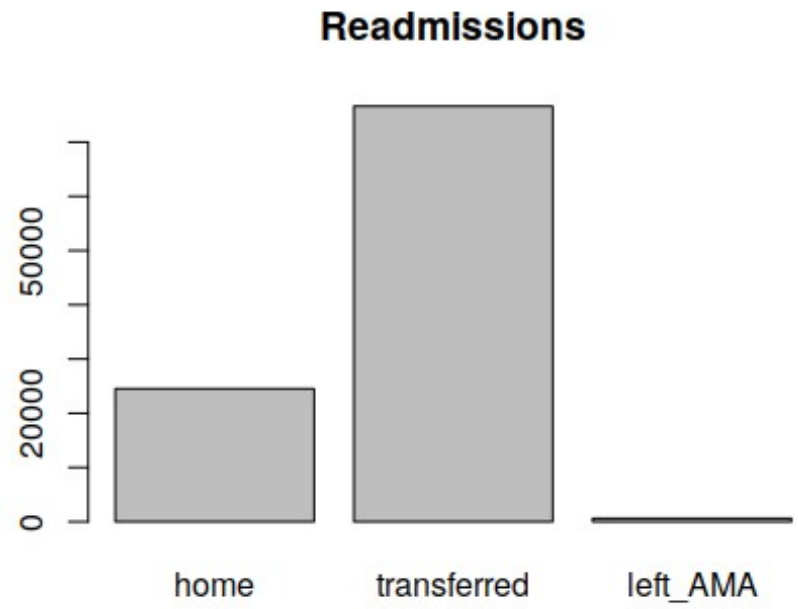

```
plot(data$readmitted, main = "Readmissions")
```



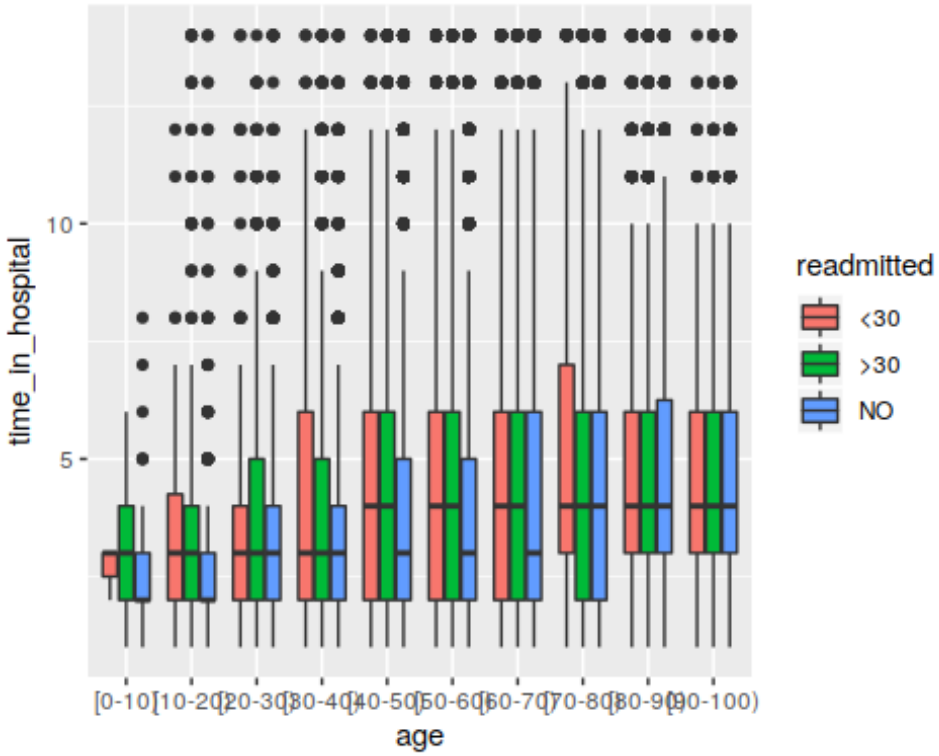

```
plot(data2$admission_source, main = "Source of admission")
```



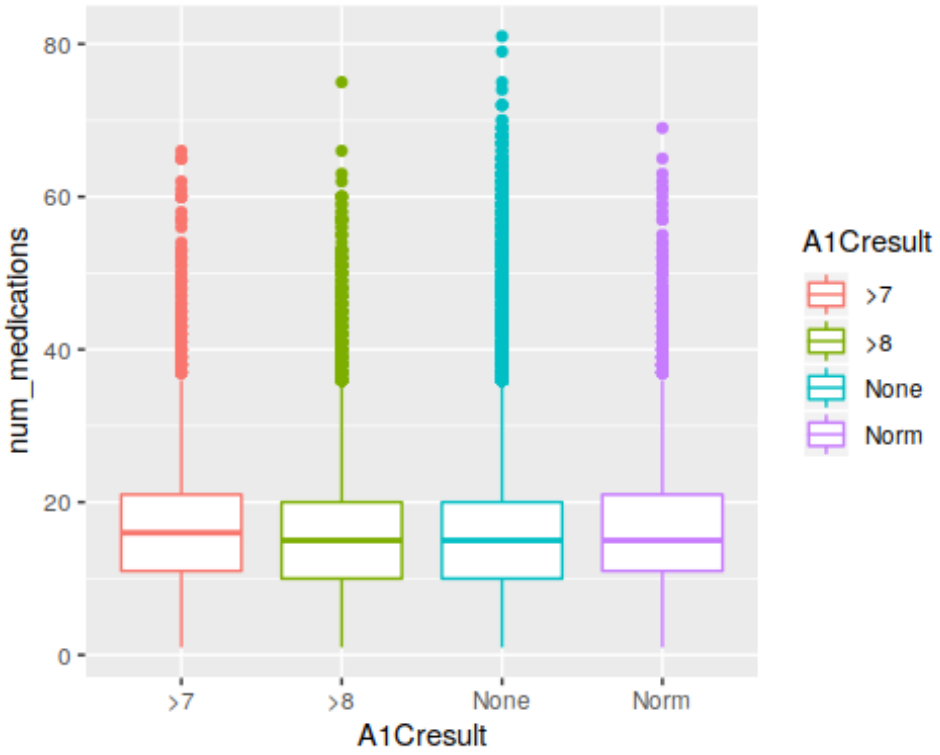

```
plot(data2$discharged_to, main = "Readmissions")
```



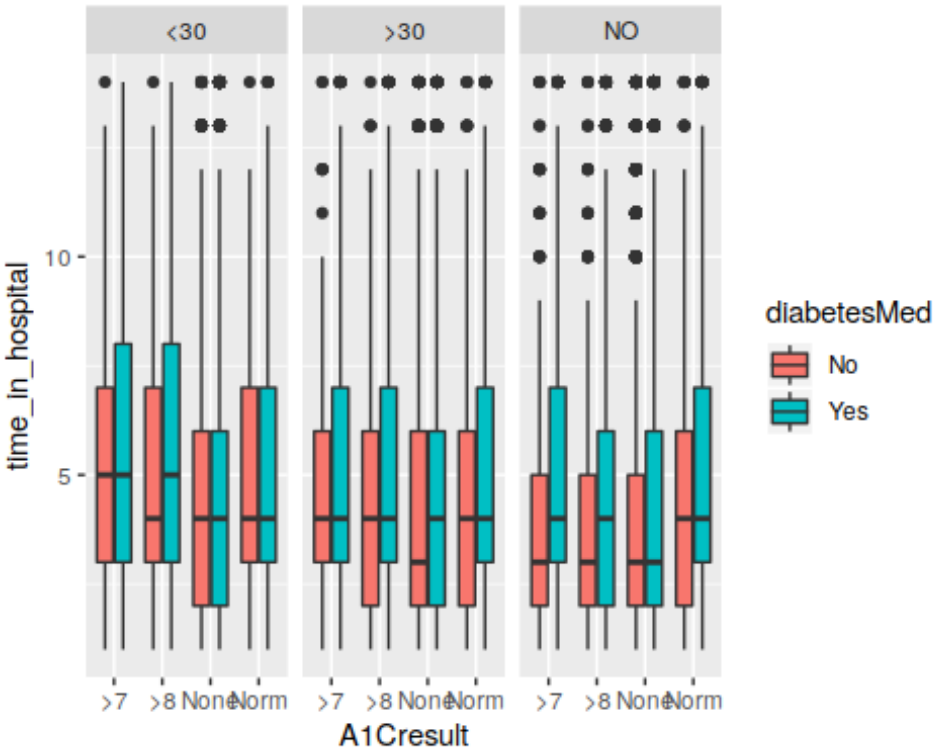

```
g <- ggplot(data2, aes(x=age, y=time_in_hospital))
g + geom_boxplot(aes(fill=readmitted))
```



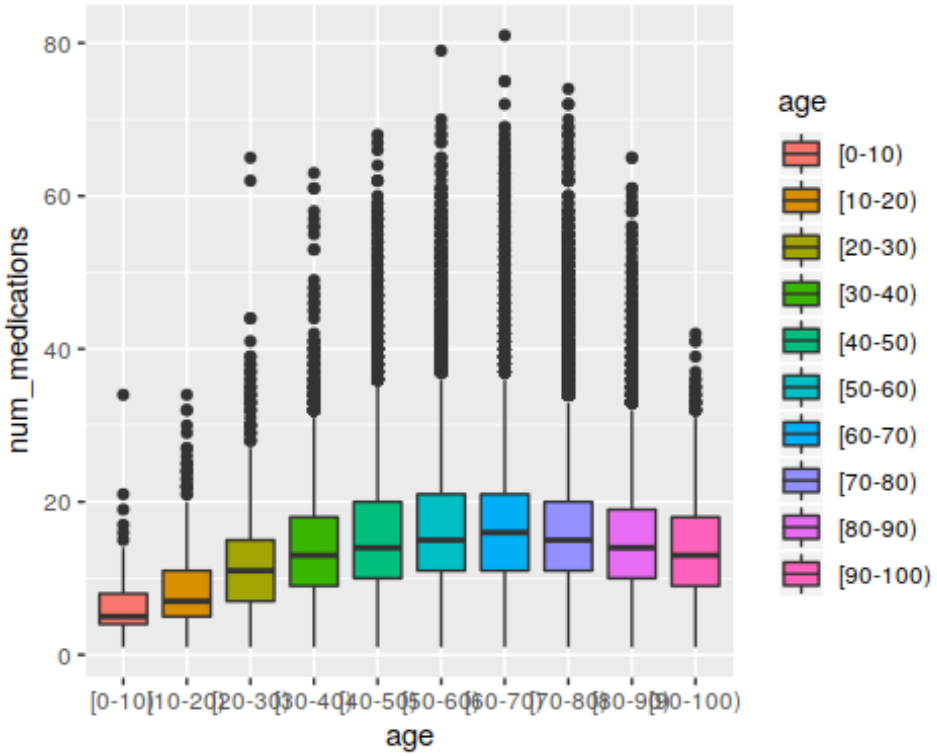
```
g <- ggplot(data2,aes(x=A1Cresult, y=num_medications))
g + geom_boxplot(aes(color=A1Cresult))
```



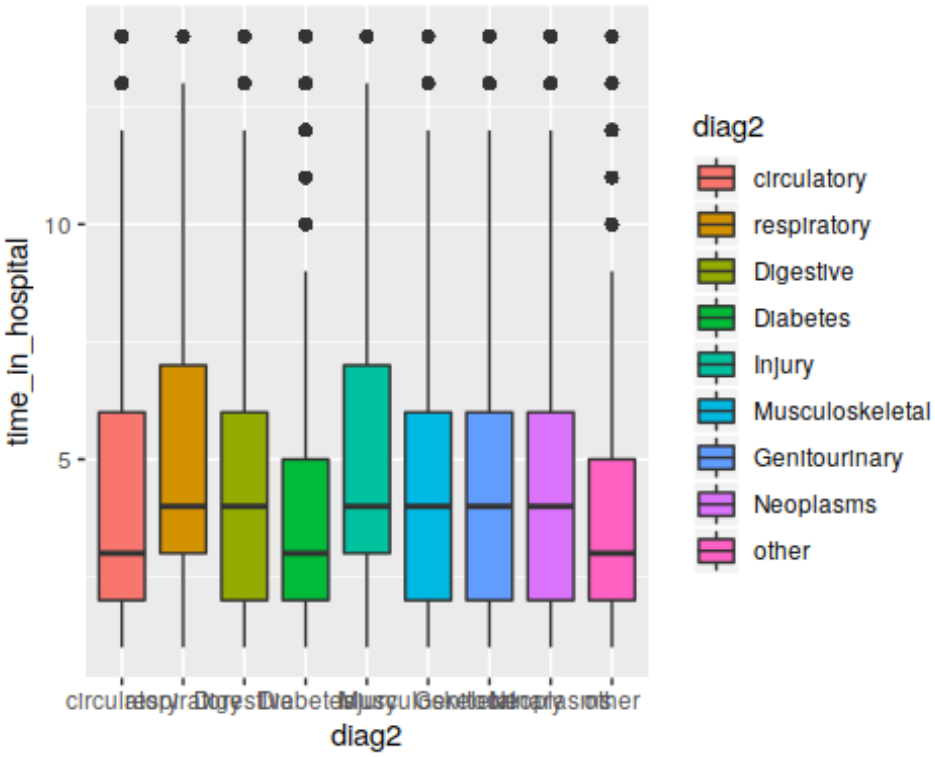

```
g <- ggplot(data2,aes(x=A1Cresult, y=time_in_hospital))
g + geom_boxplot(aes(fill=diabetesMed)) + facet_grid(. ~ readmitted)
```




```
g <- ggplot(data2,aes(x=age, y=num_medications))
g + geom_boxplot(aes(fill=age))
```



```
g <- ggplot(data2,aes(x=diag2, y=time_in_hospital))
g + geom_boxplot(aes(fill=diag2))
```



Mode is 70-80yrs normal distribution, right skewed.

53% of the patients were female, while 47% were males.

84% of the patients had no A1c results.

More than 50% patients werer not readmitted.

Emergency 60%.

70% of patients were transferred to another facility.

75% of patients were Caucasian, while the mode of stay in hospital was 3 days. Patients with readmission inside of 30 days were in their 70s-80s and had longer stints at the hospital. Patients in their 30s-40s readmitted within 30 days spent longer time in the hospital as well. Patients with no readmission had generally spent less time in hospital, which is self-explanatory. The number of medications being taken by patients was highest in 60-70yr olds. Finally, patients with either respiratory and/or injury diagnoses stayed for longer in the hospital.

Principal Component Analysis for Potential Predictors

Even after data pre-processing, there remain over 20 columns of data for what is can be modeled through multivariate logistic regression, as is the case in the study by Strack et. al, or through a support vector machine (SVM) or R-part decision tree, or even Random

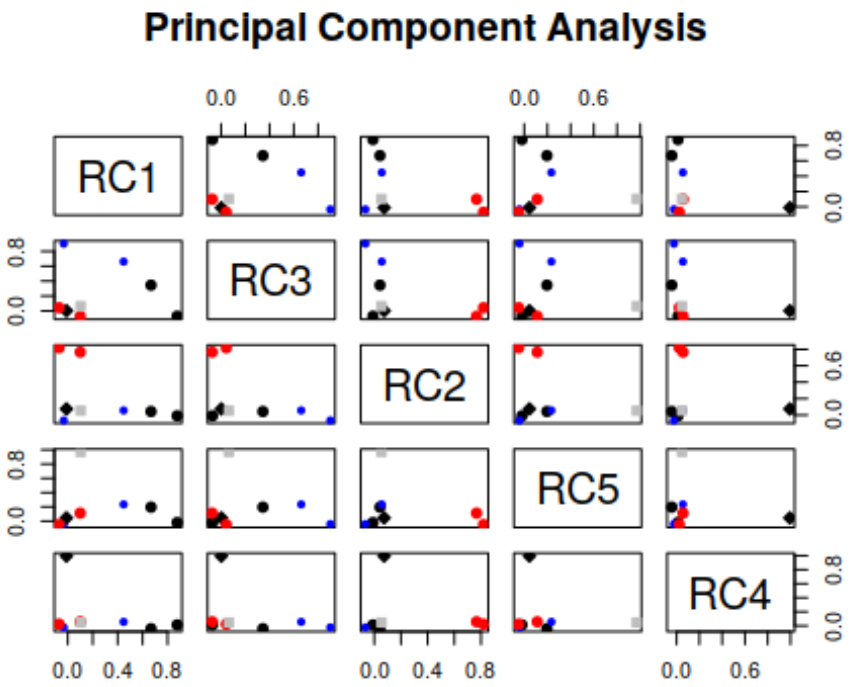
Forests. Prior to deploying these models to predict patient readmission, it would be beneficial to identify which features contribute to the pricipal components.

```
# QUICK PCA with numeric variables

y <- select(data2, readmitted)
X <- select(data2, time_in_hospital, num_lab_procedures,
num_procedures, num_medications,
            number_outpatient, number_emergency, number_inpatient,
number_diagnoses)

pca_noRot <- principal(X, nfactors = 5, rotate = "none")
rotation2_noRot <- data.frame(cbind(pca_noRot$score, y))
head(rotation2_noRot)

##           PC1         PC2         PC3         PC4        PC5
readmitted
## 1 -2.25316154 -0.62717313 -0.55923530 -0.73955236  2.4753866
NO
## 2  0.09670017 -0.15201901 -0.91473185  0.73852281 -0.5917657
>30
## 3 -0.47502214 -0.06655793  2.91742195 -0.37807450  0.4665627
NO
## 4 -0.44440274 -0.41069854 -0.08463501  0.04481760 -0.0449856
NO
## 5 -1.25997530 -0.35914236 -0.73236998  0.02918209  0.9124258
NO
## 6  0.46559588 -0.95044487  1.96758288 -0.74207955 -1.0661839
>30

pca_noRot$loadings

##
## Loadings:
##                     PC1    PC2    PC3    PC4    PC5
## time_in_hospital     0.742         -0.198          0.107
## num_lab_procedures   0.546         -0.559  0.160  0.371
## num_procedures       0.490 -0.331  0.593 -0.342
## num_medications      0.804          0.181
## number_outpatient           0.448  0.483  0.648  0.369
## number_emergency            0.692        -0.436
## number_inpatient     0.168  0.716 -0.101 -0.260
## number_diagnoses     0.506  0.227         0.351 -0.727
##
##                   PC1   PC2   PC3   PC4   PC5
## SS loadings     2.031 1.371 0.984 0.952 0.819
## Proportion Var 0.254 0.171 0.123 0.119 0.102
## Cumulative Var 0.254 0.425 0.548 0.667 0.770

# linear model of class as a function of PCs
linModel_noRot <- glm(readmitted ~ PC1 + PC2 + PC3 + PC4 + PC5, data =
```

```
rotation2_noRot, family = binomial)
summary(linModel_noRot)

##
## Call:
## glm(formula = readmitted ~ PC1 + PC2 + PC3 + PC4 + PC5, family =
binomial,
##     data = rotation2_noRot)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4314   0.4052   0.4549   0.4989   4.9863
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.127068   0.010367 205.182  < 2e-16 ***
## PC1         -0.190832   0.010017 -19.051  < 2e-16 ***
## PC2         -0.306258   0.008669 -35.329  < 2e-16 ***
## PC3          0.084091   0.010088   8.335  < 2e-16 ***
## PC4          0.052632   0.009552   5.510 3.59e-08 ***
## PC5          0.067664   0.010380   6.519 7.08e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 71205  on 101765  degrees of freedom
## Residual deviance: 69353  on 101760  degrees of freedom
## AIC: 69365
##
## Number of Fisher Scoring iterations: 5

# all PCs are significant ***

# PCA with varimax rotation
pca2 <- principal(X, nfactors = 5, rotate = "varimax")
rotation2 <- data.frame(cbind(pca2$score, y))
pca2$loadings

##
## Loadings:
##                     RC1    RC3    RC2    RC5    RC4
## time_in_hospital     0.668  0.345         0.196
## num_lab_procedures   0.877
## num_procedures              0.901
## num_medications      0.448  0.661         0.234
## number_outpatient                                0.995
## number_emergency                    0.819
## number_inpatient                    0.767  0.113
## number_diagnoses     0.103                0.969
##
```

```
##                    RC1   RC3   RC2   RC5   RC4
## SS loadings    1.443 1.385 1.276 1.052 1.001
## Proportion Var 0.180 0.173 0.159 0.132 0.125
## Cumulative Var 0.180 0.353 0.513 0.644 0.770
```

```
plot(pca2)
```




```
summary(rotation2)
```

```
##       RC1                  RC3                RC2                  RC5
##  Min.   :-7.574275   Min.   :-1.9592   Min.   :-1.99465   Min.   :-
8.4042
##  1st Qu.:-0.660605   1st Qu.:-0.7510   1st Qu.:-0.45102   1st Qu.:-
0.7400
##  Median :-0.000514   Median :-0.2624   Median :-0.37591   Median :
0.3911
##  Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.00000   Mean   :
0.0000
##  3rd Qu.: 0.659315   3rd Qu.: 0.5415   3rd Qu.: 0.08278   3rd Qu.:
0.7883
##  Max.   : 4.449304   Max.   : 6.6907   Max.   :54.79327   Max.   :
4.9339
##       RC4           readmitted
##  Min.   :-3.5602   <30:11357
##  1st Qu.:-0.3211   >30:35545
##  Median :-0.2614   NO :54864
```

```
##  Mean   : 0.0000
##  3rd Qu.:-0.1847
##  Max.   :33.0071

# plot(rotation2)
# commented plot of rotation as it created a gigantic pdf when
knitting.
```

With standard principal component analysis, the 5 principal components come out to be the following groups. The first principal component refers to number of medications and time in hospital. PC2 is number of in-patient visits and emergency. PC3 is the number of procedures, PC4 refers to the number of out-patient visits and PC5 signifies the number of diagnoses.

The varimax roation is applied to the top 5 principal components to maximize the sum of variance. In the above visualizatios, the Rotatoted Componenets are explained as follows. RC1 refers to lab procedures and time in hospital, RC2 signifies emergency visits and status as an in-patient, RC3 refers to number of procedures and medications, while RC4 is outpatient information and finally RC5 is the number of diagnoses.

Splitting the Data

The processed dataset is split 66 to 37% for training and testing respectively.

```
# SPLIT DATA INTO TRAINING AND TESTING SET

set.seed(123)
inTrain <- createDataPartition(y = data2$readmitted, p = .66,list =
FALSE)
train <- data2[ inTrain,]
test <- data2[-inTrain,]
nrow(train) # 67167

## [1] 67167

nrow(test) # 3459

## [1] 34599
```

Logistic Regression

By fitting two linear models with and without the HbA1c test results, a conclusion on the importance of this parameter can be made. For the first trial, a mutlivariate logistic regression has been attempted while excluding the HbA1c test results. The model converges in 8 Fischer iterations.

```
# LOGISTIC REGRESSION

fit_all <- glm(readmitted ~., data=train, family=binomial)
summary(fit_all)
```

```
##
## Call:
## glm(formula = readmitted ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8990   0.3846   0.4395   0.4999   2.2835
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)

## (Intercept)                4.3320752  0.7324285   5.915 3.33e-09
***
## raceAsian                 -0.0018707  0.1642194  -0.011 0.990911

## raceCaucasian             -0.0234787  0.0334207  -0.703 0.482355

## raceHispanic              -0.0469144  0.0944243  -0.497 0.619297

## raceOther                  0.1336233  0.0774369   1.726 0.084424 .

## genderMale                -0.0203878  0.0254299  -0.802 0.422712

## genderUnknown/Invalid      7.4871401 72.4629962   0.103 0.917706

## age[10-20)                -0.8869117  0.7409504  -1.197 0.231310

## age[20-30)                -1.4180359  0.7223842  -1.963 0.049647 *

## age[30-40)                -1.3241135  0.7195225  -1.840 0.065729 .

## age[40-50)                -1.3021107  0.7179984  -1.814 0.069750 .

## age[50-60)                -1.3049306  0.7177178  -1.818 0.069039 .

## age[60-70)                -1.4176661  0.7176487  -1.975 0.048219 *

## age[70-80)                -1.5154410  0.7176014  -2.112 0.034702 *

## age[80-90)                -1.5231336  0.7178422  -2.122 0.033853 *

## age[90-100)               -1.4535488  0.7211296  -2.016 0.043836 *

## time_in_hospital          -0.0209143  0.0048761  -4.289 1.79e-05
***
## num_lab_procedures        -0.0007656  0.0007449  -1.028 0.304010

## num_procedures             0.0363290  0.0088114   4.123 3.74e-05
```

```
***
## num_medications           -0.0059848  0.0020344  -2.942 0.003263 **
## number_outpatient         -0.0005861  0.0093355  -0.063 0.949941
## number_emergency          -0.0410848  0.0110600  -3.715 0.000203
***
## number_inpatient          -0.2579019  0.0081774 -31.538  < 2e-16
***
## number_diagnoses          -0.0434177  0.0079870  -5.436 5.45e-08
***
## max_glu_serum>300         -0.0258251  0.1397134  -0.185 0.853352
## max_glu_serumNone          0.1130685  0.0994652   1.137 0.255637
## max_glu_serumNorm          0.0380231  0.1233544   0.308 0.757898
## A1Cresult>8                0.0097075  0.0817711   0.119 0.905500
## A1CresultNone             -0.0983488  0.0685650  -1.434 0.151462
## A1CresultNorm              0.0385024  0.0892388   0.431 0.666139
## insulinNo                  0.1778067  0.0498493   3.567 0.000361
***
## insulinSteady              0.1236622  0.0451809   2.737 0.006199 **
## insulinUp                  0.0553715  0.0485476   1.141 0.254053
## changeNo                  -0.0773335  0.0355372  -2.176 0.029546 *
## diabetesMedYes            -0.1254690  0.0404473  -3.102 0.001922 **
## diag1respiratory           0.2677983  0.0432513   6.192 5.95e-10
***
## diag1Digestive             0.1642629  0.0511756   3.210 0.001328 **
## diag1Diabetes             -0.0955557  0.0520027  -1.838 0.066134 .
## diag1Injury               -0.0206477  0.0527935  -0.391 0.695721
## diag1Musculoskeletal       0.0862266  0.0669121   1.289 0.197518
## diag1Genitourinary         0.1406363  0.0623199   2.257 0.024028 *
## diag1Neoplasms             0.1341967  0.0426854   3.144 0.001667 **
## diag1other                 0.0063333  0.0539477   0.117 0.906546
```

```
## diag2respiratory           0.1237510  0.0461673   2.680 0.007351 **
## diag2Digestive            -0.0219824  0.0691118  -0.318 0.750432
## diag2Diabetes             -0.1181940  0.0461688  -2.560 0.010466 *
## diag2Injury               -0.0342049  0.0860925  -0.397 0.691144
## diag2Musculoskeletal       0.1226307  0.1053851   1.164 0.244568
## diag2Genitourinary         0.0418495  0.0492916   0.849 0.395871
## diag2Neoplasms            -0.0801934  0.0378499  -2.119 0.034114 *
## diag2other                 0.0470526  0.0481365   0.977 0.328330
## diag3respiratory          -0.0290795  0.0510372  -0.570 0.568833
## diag3Digestive            -0.0737785  0.0689288  -1.070 0.284458
## diag3Diabetes             -0.0882469  0.0405602  -2.176 0.029577 *
## diag3Injury               -0.0538811  0.0930862  -0.579 0.562704
## diag3Musculoskeletal      -0.0591301  0.0954469  -0.620 0.535582
## diag3Genitourinary        -0.1243312  0.0512776  -2.425 0.015322 *
## diag3Neoplasms            -0.0359685  0.0383199  -0.939 0.347916
## diag3other                 0.0208179  0.0429224   0.485 0.627667
## admission_sourceemergency  0.0550461  0.0467263   1.178 0.238775
## admission_sourceother      0.0936103  0.0487297   1.921 0.054731 .
## discharged_totransferred  -0.2285422  0.0325089  -7.030 2.06e-12
***
## discharged_toleft_AMA     -0.4253694  0.1511161  -2.815 0.004880 **
## payer_codeself_pay        -0.0139187  0.0597244  -0.233 0.815723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 46997  on 67166  degrees of freedom
## Residual deviance: 45211  on 67103  degrees of freedom
## AIC: 45339
##
## Number of Fisher Scoring iterations: 8

# pseudo R-squared for logistic regression model
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <-  length(LogModel$fitted.values)
  R.l <-  1 -  dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}

logisticPseudoR2s(fit_all)

## Pseudo R^2 for logistic regression
## Hosmer and Lemeshow R^2   0.038
## Cox and Snell R^2         0.026
## Nagelkerke R^2            0.052
```

Clearly, this model performs rather poorly.

For the second logistic model, the HbA1c results are included as a predictor. The adjusted R-squared and Chi-squared test reveals that this model performs very similarly to the standard model in the previous code chunk.

```
library(pscl)

## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis

# main effects, with A1C result
fit_a1c <- glm(readmitted ~ race+age+discharged_to+time_in_hospital+

num_lab_procedures+num_procedures+num_medications+number_outpatient+
             number_emergency+number_inpatient+number_diagnoses+
             insulin+change+diabetesMed+diag1+diag2+diag3+A1Cresult,
           data=train, family = binomial)
summary(fit_a1c)
```

```
##
## Call:
## glm(formula = readmitted ~ race + age + discharged_to +
time_in_hospital +
##     num_lab_procedures + num_procedures + num_medications +
number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     insulin + change + diabetesMed + diag1 + diag2 + diag3 +
##     A1Cresult, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9010   0.3849   0.4400   0.4996   2.2900
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                4.4836276  0.7249506   6.185 6.22e-10 ***
## raceAsian                 -0.0100058  0.1641053  -0.061 0.951382
## raceCaucasian             -0.0287415  0.0331437  -0.867 0.385845
## raceHispanic              -0.0571373  0.0942881  -0.606 0.544524
## raceOther                  0.1341869  0.0771480   1.739 0.081974 .
## age[10-20)                -0.8912004  0.7408852  -1.203 0.229020
## age[20-30)                -1.4224222  0.7222953  -1.969 0.048918 *
## age[30-40)                -1.3340896  0.7194365  -1.854 0.063689 .
## age[40-50)                -1.3141904  0.7179028  -1.831 0.067161 .
## age[50-60)                -1.3159163  0.7176292  -1.834 0.066699 .
## age[60-70)                -1.4284584  0.7175630  -1.991 0.046513 *
## age[70-80)                -1.5265744  0.7175163  -2.128 0.033372 *
## age[80-90)                -1.5341060  0.7177548  -2.137 0.032568 *
## age[90-100)               -1.4633775  0.7210283  -2.030 0.042400 *
## discharged_totransferred -0.2117733  0.0311287  -6.803 1.02e-11 ***
## discharged_toleft_AMA    -0.4228555  0.1510000  -2.800 0.005104 **
## time_in_hospital         -0.0215273  0.0048580  -4.431 9.37e-06 ***
## num_lab_procedures       -0.0006507  0.0007131  -0.913 0.361492
## num_procedures            0.0377935  0.0086710   4.359 1.31e-05 ***
## num_medications          -0.0059246  0.0020141  -2.942 0.003266 **
## number_outpatient        -0.0010944  0.0092989  -0.118 0.906314
## number_emergency         -0.0415350  0.0110498  -3.759 0.000171 ***
## number_inpatient         -0.2583964  0.0081598 -31.667  < 2e-16 ***
## number_diagnoses         -0.0431737  0.0078984  -5.466 4.60e-08 ***
## insulinNo                 0.1839036  0.0496615   3.703 0.000213 ***
## insulinSteady             0.1295031  0.0449713   2.880 0.003981 **
## insulinUp                 0.0566980  0.0485172   1.169 0.242558
## changeNo                 -0.0779371  0.0355159  -2.194 0.028204 *
## diabetesMedYes           -0.1238852  0.0404245  -3.065 0.002180 **
## diag1respiratory          0.2661413  0.0431844   6.163 7.14e-10 ***
## diag1Digestive            0.1627616  0.0511441   3.182 0.001461 **
## diag1Diabetes            -0.0978257  0.0519358  -1.884 0.059621 .
## diag1Injury              -0.0190415  0.0527559  -0.361 0.718147
## diag1Musculoskeletal      0.0906135  0.0664085   1.364 0.172415
```

```
## diag1Genitourinary        0.1431810  0.0622731   2.299 0.021491 *
## diag1Neoplasms            0.1348785  0.0426365   3.163 0.001559 **
## diag1other                0.0147847  0.0534718   0.276 0.782168
## diag2respiratory          0.1209144  0.0461398   2.621 0.008777 **
## diag2Digestive           -0.0232019  0.0690705  -0.336 0.736935
## diag2Diabetes            -0.1203610  0.0461177  -2.610 0.009058 **
## diag2Injury              -0.0326212  0.0860660  -0.379 0.704669
## diag2Musculoskeletal      0.1241485  0.1053734   1.178 0.238726
## diag2Genitourinary        0.0402716  0.0492546   0.818 0.413573
## diag2Neoplasms           -0.0822905  0.0378277  -2.175 0.029600 *
## diag2other                0.0451978  0.0481163   0.939 0.347554
## diag3respiratory         -0.0299450  0.0510278  -0.587 0.557314
## diag3Digestive           -0.0746104  0.0689025  -1.083 0.278879
## diag3Diabetes            -0.0888472  0.0405388  -2.192 0.028404 *
## diag3Injury              -0.0547867  0.0930757  -0.589 0.556113
## diag3Musculoskeletal     -0.0579821  0.0954050  -0.608 0.543355
## diag3Genitourinary       -0.1258919  0.0512524  -2.456 0.014037 *
## diag3Neoplasms           -0.0376742  0.0382944  -0.984 0.325212
## diag3other                0.0207894  0.0429193   0.484 0.628114
## A1Cresult>8               0.0100658  0.0817676   0.123 0.902026
## A1CresultNone            -0.0988712  0.0685423  -1.442 0.149165
## A1CresultNorm             0.0394730  0.0892272   0.442 0.658209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46997  on 67166  degrees of freedom
## Residual deviance: 45219  on 67111  degrees of freedom
## AIC: 45331
##
## Number of Fisher Scoring iterations: 6

# results not very different from fit_all
logisticPseudoR2s(fit_a1c)

## Pseudo R^2 for logistic regression
## Hosmer and Lemeshow R^2   0.038
## Cox and Snell R^2         0.026
## Nagelkerke R^2            0.052

pR2(fit_a1c)

##           llh       llhNull            G2      McFadden
r2ML
## -2.260935e+04 -2.349854e+04  1.778383e+03  3.784029e-02  2.612960e-
02
##          r2CU
##  5.191983e-02
```

```
# adjusted R-squared mostly same as fit_all
anova(fit_a1c, test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: readmitted
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                67166      46997
## race                 4    17.74     67162      46979  0.001389 **
## age                  9    83.67     67153      46896 3.002e-14 ***
## discharged_to        2    68.11     67151      46828 1.618e-15 ***
## time_in_hospital     1   102.01     67150      46726 < 2.2e-16 ***
## num_lab_procedures   1     2.96     67149      46723  0.085590 .
## num_procedures       1    28.97     67148      46694 7.350e-08 ***
## num_medications      1    46.21     67147      46647 1.063e-11 ***
## number_outpatient    1    19.78     67146      46628 8.675e-06 ***
## number_emergency     1   181.35     67145      46446 < 2.2e-16 ***
## number_inpatient     1  1041.01     67144      45405 < 2.2e-16 ***
## number_diagnoses     1    26.87     67143      45378 2.181e-07 ***
## insulin              3    30.82     67140      45348 9.262e-07 ***
## change               1     1.80     67139      45346  0.179957
## diabetesMed          1     9.63     67138      45336  0.001914 **
## diag1                8    66.76     67130      45269 2.160e-11 ***
## diag2                8    28.44     67122      45241  0.000398 ***
## diag3                8    11.92     67114      45229  0.154731
## A1Cresult            3    10.33     67111      45219  0.015978 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Decision Tree

For the third test, a decision tree is attempted. The breakdown of variable importance belies the skewness of the dataset, which is bound to cripple the performance of the decision tree. As a result, while the tree predicts no readmission with almost 92% accuracy, the model's sensitivity to predicting readmission within or after 30 days is abysmal.

```
# RPART DECISION TREES

rpart_tree <- rpart(formula = readmitted ~
age+discharged_to+time_in_hospital+

num_lab_procedures+num_procedures+num_medications+number_outpatient+

number_emergency+number_inpatient+number_diagnoses+
```

```
insulin+change+diabetesMed+diag1+diag2+diag3+A1Cresult,
                    data=train, method = 'class')
summary(rpart_tree)
```

```
## Call:
## rpart(formula = readmitted ~ age + discharged_to + time_in_hospital
+
##     num_lab_procedures + num_procedures + num_medications +
number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     insulin + change + diabetesMed + diag1 + diag2 + diag3 +
##     A1Cresult, data = train, method = "class")
##   n= 67167
##
##           CP nsplit rel error    xerror        xstd
## 1 0.03973382      0 1.0000000 1.0000000 0.004173206
## 2 0.01408451      1 0.9602662 0.9602662 0.004158332
## 3 0.01000000      2 0.9461817 0.9461817 0.004151688
##
## Variable importance
## number_inpatient number_emergency
##               93                6
##
## Node number 1: 67167 observations,    complexity param=0.03973382
##   predicted class=NO   expected loss=0.4608811  P(node) =1
##     class counts:  7496 23460 36211
##    probabilities: 0.112 0.349 0.539
##   left son=2 (22559 obs) right son=3 (44608 obs)
##   Primary splits:
##       number_inpatient  < 0.5  to the right, improve=1167.9460, (0
missing)
##       number_emergency  < 0.5  to the right, improve= 407.1558, (0
missing)
##       number_outpatient < 0.5  to the right, improve= 322.8942, (0
missing)
##       number_diagnoses  < 5.5  to the right, improve= 299.4891, (0
missing)
##       num_medications   < 10.5 to the right, improve= 147.5778, (0
missing)
##   Surrogate splits:
##       number_emergency  < 0.5  to the right, agree=0.685,
adj=0.062, (0 split)
##       number_outpatient < 4.5  to the right, agree=0.666,
adj=0.005, (0 split)
##
## Node number 2: 22559 observations,    complexity param=0.01408451
##   predicted class=>30  expected loss=0.5560087  P(node) =0.3358643
##     class counts:  3757 10016  8786
##    probabilities: 0.167 0.444 0.389
```

```
##   left son=4 (9552 obs) right son=5 (13007 obs)
##   Primary splits:
##       number_inpatient  < 1.5  to the right,   improve=168.30290,
(0 missing)
##       number_emergency  < 0.5  to the right,   improve= 80.80963,
(0 missing)
##       number_outpatient < 0.5  to the right,   improve= 66.26141,
(0 missing)
##       age               splits as  RLLLLLLRRR, improve= 41.40109,
(0 missing)
##       diag1             splits as  LLLLRRRRR,  improve= 31.46437,
(0 missing)
##   Surrogate splits:
##       number_emergency  < 0.5  to the right,   agree=0.604,
adj=0.065, (0 split)
##       age               splits as  RRLLRRRRRR, agree=0.581,
adj=0.011, (0 split)
##       diag1             splits as  RRRLRRRRR,  agree=0.580,
adj=0.009, (0 split)
##       number_outpatient < 0.5  to the right,   agree=0.579,
adj=0.006, (0 split)
##       discharged_to     splits as  RRL,        agree=0.578,
adj=0.003, (0 split)
##
## Node number 3: 44608 observations
##   predicted class=NO   expected loss=0.3852  P(node) =0.6641357
##     class counts:  3739 13444 27425
##    probabilities: 0.084 0.301 0.615
##
## Node number 4: 9552 observations
##   predicted class=>30  expected loss=0.5183208  P(node) =0.1422127
##     class counts:  2016  4601  2935
##    probabilities: 0.211 0.482 0.307
##
## Node number 5: 13007 observations
##   predicted class=NO   expected loss=0.5501653  P(node) =0.1936516
##     class counts:  1741  5415  5851
##    probabilities: 0.134 0.416 0.450
```

```
test$pred_readmit <- predict(rpart_tree, test, type="class")
table(predict(rpart_tree, test, type="class"), test$readmitted)
```

```
##
##         <30   >30    NO
##   <30     0     0     0
##   >30  1112  2449  1502
##   NO   2749  9636 17151
```

```
prop.table(table(test$readmitted, test$pred_readmit),1)
```

```
##
##              <30        >30         NO
##   <30 0.00000000 0.28800829 0.71199171
##   >30 0.00000000 0.20264791 0.79735209
##   NO  0.00000000 0.08052324 0.91947676
```

```
confusionMatrix(test$pred_readmit, test$readmitted)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   <30   >30    NO
##        <30     0     0     0
##        >30  1112  2449  1502
##        NO   2749  9636 17151
##
## Overall Statistics
##
##                Accuracy : 0.5665
##                  95% CI : (0.5612, 0.5717)
##     No Information Rate : 0.5391
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1129
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity              0.0000    0.20265    0.9195
## Specificity              1.0000    0.88389    0.2233
## Pos Pred Value              NaN    0.48371    0.5807
## Neg Pred Value           0.8884    0.67375    0.7033
## Prevalence               0.1116    0.34929    0.5391
## Detection Rate           0.0000    0.07078    0.4957
## Detection Prevalence     0.0000    0.14633    0.8537
## Balanced Accuracy        0.5000    0.54327    0.5714
```

Random Forest

For the fourth test, a random forest approach is tested for the readmission dataset. This model does not predict readmission as well as the decision tree in the third test. Here, the prediction of readmission is at an accuracy of 84%, as shown in the confusion matrix.

```
# RANDOM FOREST
Rf_fit<-randomForest(formula=readmitted ~
age+discharged_to+time_in_hospital+

num_lab_procedures+num_procedures+num_medications+number_outpatient+
```

```
number_emergency+number_inpatient+number_diagnoses+

insulin+change+diabetesMed+diag1+diag2+diag3+A1Cresult,
                     data=train)
print(Rf_fit)

##
## Call:
##  randomForest(formula = readmitted ~ age + discharged_to +
time_in_hospital +      num_lab_procedures + num_procedures +
num_medications + number_outpatient +      number_emergency +
number_inpatient + number_diagnoses +      insulin + change +
diabetesMed + diag1 + diag2 + diag3 +      A1Cresult, data = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 43.1%
## Confusion matrix:
##     <30  >30    NO class.error
## <30 134 2760  4602   0.9821238
## >30 118 7982 15360   0.6597613
## NO   51 6061 30099   0.1687885

test$pred_readmit <- predict(Rf_fit, test, type = "response")
table(test$readmitted, test$pred_readmit)

##
##         <30   >30    NO
##   <30    74  1388  2399
##   >30    65  4065  7955
##   NO     25  2912 15716

prop.table(table(test$readmitted, test$pred_readmit),1)

##
##                <30         >30          NO
##   <30 0.019166019 0.359492359 0.621341621
##   >30 0.005378568 0.336367398 0.658254034
##   NO  0.001340267 0.156114298 0.842545435

importance(Rf_fit)

##                    MeanDecreaseGini
## age                       2776.7713
## discharged_to              831.4018
## time_in_hospital          3211.9642
## num_lab_procedures        5317.8260
## num_procedures            1942.6983
## num_medications           4473.9832
```

```
## number_outpatient          815.3654
## number_emergency           619.0576
## number_inpatient          1714.0723
## number_diagnoses          1993.8531
## insulin                   1467.4188
## change                     623.9546
## diabetesMed                369.3588
## diag1                     2972.4926
## diag2                     3149.5865
## diag3                     3271.0728
## A1Cresult                 1021.2010
```

Support Vector Machine

A Support Vector Machine approach is the fourth test in this project. Due to the volume of the dataset, a parallelSVM library function is used instead of the standard SVM function call in R. Any loss or gain in model performance by using the parallel implementation was not tested. The SVM performs particularly poorly compared to the Random Forest and Decision Trees with just 56% accuracy, and takes longer to train, making it the least favorable approach tested. Note that the warnings in the code output for the SVM code chunk are inherent to the parallelSVM library and cannot be avoided.

```
# SUPPORT VECTOR MACHINES
library(parallelSVM)
SVMmodel <- parallelSVM(readmitted ~
age+discharged_to+time_in_hospital+

num_lab_procedures+num_procedures+num_medications+number_outpatient+
                  number_emergency+number_inpatient+number_diagnoses+

insulin+change+diabetesMed+diag1+diag2+diag3+A1Cresult,
                data=train, kernel = "linear")
                  #kernel = "rbf", gamma = 0.1, cost = 1)
print(SVMmodel)

##
## Call:
## parallelSVM(formula = readmitted ~ age + discharged_to +
time_in_hospital + num_lab_procedures + num_procedures +
num_medications + number_outpatient + number_emergency +
number_inpatient + number_diagnoses + insulin + change +  diabetesMed
+ diag1 + diag2 + diag3 + A1Cresult, data = train,  kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classficiation
##  SVM-Kernel:  linear
##        cost:  1
##       gamma:  0.01923077
##
```

```
## Average Number of Support Vectors: 11509
##
```

```
summary(SVMmodel)
```

```
##
## Call:
## parallelSVM(formula = readmitted ~ age + discharged_to +
time_in_hospital + num_lab_procedures + num_procedures +
num_medications + number_outpatient + number_emergency +
number_inpatient + number_diagnoses + insulin + change +  diabetesMed
+ diag1 + diag2 + diag3 + A1Cresult, data = train,  kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classficiation
##  SVM-Kernel:  linear
##        cost:  1
##       gamma:  0.01923077
##
## Average Number of Support Vectors: 11509
##
##
##  ( 3836 7673 )
##
##
## Number of classes: 3
##
## Levels:
## <30 >30 NO
##
##
##
```

```
x <- select(test, -readmitted)
y <- select(test, readmitted)
pred <- predict(SVMmodel, x)
test$pred_readmit <- pred
prop.table(table(test$readmitted, test$pred_readmit),1)
```

```
##
##              >30         NO
##   <30 0.22325822 0.77674178
##   >30 0.15821266 0.84178734
##   NO  0.05275291 0.94724709
```

```
confusionMatrix(test$pred_readmit, test$readmitted)
```

```
## Warning in levels(reference) != levels(data): longer object length
is not a
## multiple of shorter object length
```

```
## Warning in confusionMatrix.default(test$pred_readmit,
test$readmitted): Levels
## are not in the same order for reference and data. Refactoring data
to match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   <30   >30    NO
##        <30     0     0     0
##        >30   862  1912   984
##        NO   2999 10173 17669
##
## Overall Statistics
##
##                Accuracy : 0.5659
##                  95% CI : (0.5607, 0.5712)
##     No Information Rate : 0.5391
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.0985
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity              0.0000    0.15821    0.9472
## Specificity              1.0000    0.91801    0.1740
## Pos Pred Value              NaN    0.50878    0.5729
## Neg Pred Value           0.8884    0.67015    0.7382
## Prevalence               0.1116    0.34929    0.5391
## Detection Rate           0.0000    0.05526    0.5107
## Detection Prevalence     0.0000    0.10862    0.8914
## Balanced Accuracy        0.5000    0.53811    0.5606
```

Naive Bayes

As per a study by Caruana and Niculescu-Mizil, Bayes classification has been shown to be outperformed by classifiers such as boosted trees and random forests. The random forest approach has already been attempted, so it would be an interesting experiment to see by what margin the Naive Bayes classifier falls short of the results from the Random Forest classifier.

```
# NAIVE BAYES
# e1071 implementation

nbayesmodel <- naiveBayes(readmitted ~
age+discharged_to+time_in_hospital+
```

```
num_lab_procedures+num_procedures+num_medications+number_outpatient+

number_emergency+number_inpatient+number_diagnoses+

insulin+change+diabetesMed+diag1+diag2+diag3+A1Cresult,
                          data = train)

pred <- predict(nbayesmodel, test, type = "class")
test$pred_readmit <- pred
prop.table(table(test$readmitted, test$pred_readmit),1)

##
##              <30        >30         NO
##   <30 0.10075110 0.16938617 0.72986273
##   >30 0.04973107 0.16764584 0.78262309
##   NO  0.01318823 0.07596633 0.91084544

confusionMatrix(test$pred_readmit, test$readmitted)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   <30   >30    NO
##        <30   389   601   246
##        >30   654  2026  1417
##        NO   2818  9458 16990
##
## Overall Statistics
##
##                Accuracy : 0.5609
##                  95% CI : (0.5556, 0.5661)
##     No Information Rate : 0.5391
##     P-Value [Acc > NIR] : 2.418e-16
##
##                   Kappa : 0.1193
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity             0.10075    0.16765    0.9108
## Specificity             0.97244    0.90801    0.2302
## Pos Pred Value          0.31472    0.49451    0.5805
## Neg Pred Value          0.89593    0.67022    0.6882
## Prevalence              0.11159    0.34929    0.5391
## Detection Rate          0.01124    0.05856    0.4911
## Detection Prevalence    0.03572    0.11841    0.8459
## Balanced Accuracy       0.53660    0.53783    0.5705
```

```
write.csv(data2, file = "processed_data_diabetes.csv", sep=",",
na="?", row.names = F)

## Warning in write.csv(data2, file = "processed_data_diabetes.csv",
sep = ",", :
## attempt to set 'sep' ignored
```

It turns out that for the given data, the Bayes Classifier performs measurably better than the Random Forest, although objectively all models tested in this project perform far from satisfactorily. Presumably a neaural network might perform better, but since this project was coded in R, testing this approach was not immediately feasible.

Summary and Conclusions

Based on hypothesis testing for the multivariate logstic models, the results of the study by Strack et. al can be corroborated, as the addition of HbA1c test results as a descriptor demostrably improves the result, although not by a large margin. This underwhelming delta is explained by the fact that only 84% of the patients in the entire dataset were tested for A1c and thus making meaningful predictions using it as a descriptor is moot. In the real world, the inclusion of this test result may very well improve readmission prediction, but due to the quality of the dataset, this cannot be verified beyond a doubt. Of the tested models, the Naive Bayes classifier achieved a sensitivity of 94% for predicting no readmission, with an overall accuracy of 56.09%. The Random Forest model managed a sensitivity of 84.25% for no readmission, giving an overall accuracy of 56.9%. The decision tree yielded a 91% sensitivity for no readmission prediction, but on overall accuracy of 56.65%. The SVM had an overall accuracy of 56.61%, managing a no readmission sensitivity of 94.48%. Overall, the Bayes classifier and the SVM perform better than the rest, but declaring that either performed the best is misleading, since the overall performance of all models was extremely poor. In particular, the per-class specificities for all models were below 20%, which for medical prediction tasks is extremely dangerous to rely on.