STAT 571 - Miniproject

Vishal Murali

April 7, 2017

Part 1: Executive Summary:

In this section, my goal is to provide background of the study, quick summary about the data and present the methods I used to analyse the data, and my results and findings.

(A) . Background:

Diabetes is a chronic medical condition affecting millions of Americans, but if managed well, with good diet, exercise and medication, patients can lead relatively normal lives. However, if improperly managed, diabetes can lead to patients being continuously admitted and readmitted to hospitals. Readmissions are especially serious - they represent a failure of the health system to provide adequate support to the patient and are extremely costly to the system. As a result, the Centers for Medicare and Medicaid Services announced in 2012 that they would no longer reimburse hospitals for services rendered if a patient was readmitted with complications within 30 days of discharge. Given these policy changes, being able to identify and predict those patients most at risk for costly readmissions has become a pressing priority for hospital administrators. In this project, we shall explore how to use the techniques we have learned in order to help better manage diabetes patients who have been admitted to a hospital. Our goal is to avoid patients being readmitted within 30 days of discharge, which reduces costs for the hospital and improves outcomes for patients. The original data is from the Center for Clinical and Translational Research at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are over 100,000 unique hospital admissions in this dataset, from ~70,000 unique patients. The data includes demographic elements, such as age, gender, and race, as well as clinical attributes such as tests conducted, emergency/inpatient visits, etc

(B). Summary of the data

Our dataset consists of 101766 instances of 31 features.

Description of variables:

The dataset used covers ~50 different variables to describe every hospital diabetes admission. In this section we give an overview and brief description of the variables in this dataset.

a) Patient identi???ers:

a. encounter_id: unique identifier for each admission b. patient_nbr: unique identi???er for each patient

b) Patient Demographics: race, age, gender, weight cover the basic demographic information associated with each patient. Payer_code is an additional variable that identifies which health insurance (Medicare /Medicaid / Commercial) the patient holds.

c) Admission and discharge details:

a. admission_source_id and admission_type_id identify who referred the patient to the hospital (e.g. physician vs. emergency dept.) and what type of admission this was (Emergency vs. Elective vs. Urgent).

b. discharge_disposition_id indicates where the patient was discharged to after treatment.

d) Patient Medical History:

a. num_outpatient: number of outpatient visits by the patient in the year prior to the current encounter

b. num_inpatient: number of inpatient visits by the patient in the year prior to the current encounter

c. num_emergency: number of emergency visits by the patient in the year prior to the current encounter

e) Patient admission details:

a. medical_specialty: the specialty of the physician admitting the patient

b. diag_1, diag_2, diag_3: ICD9 codes for the primary, secondary and tertiary diagnoses of the patient. ICD9 are the universal codes that all physicians use to record diagnoses. There are various easy to use tools to lookup what individual codes mean (Wikipedia is pretty decent on its own)

c. time_in_hospital: the patient's length of stay in the hospital (in days)

d. number_diagnoses: Total no. of diagnosis entered for the patient

e. num_lab_procedures: No. of lab procedures performed in the current encounter

f. num_procedures: No. of non-lab procedures performed in the current encounter g. num_medications: No. of distinct medications prescribed in the current encounter

f) Clinical Results:

a. max_glu_serum: indicates results of the glucose serum test

b. A1Cresult: indicates results of the A1c test

g) Medication Details:

a. diabetesMed: indicates if any diabetes medication was prescribed

b. change: indicates if there was a change in diabetes medication

c. 24 medication variables: indicate whether the dosage of the medicines was changed in any manner during the encounter

h) Readmission indicator: Indicates whether a patient was readmitted after a particular admission. There are 3 levels for this variable: "NO" = no readmission, "< 30" = readmission within 30 days and "> 30" = readmission after more than 30 days. The 30 day distinction is of practical importance to hospitals because federal regulations penalize hospitals for an excessive proportion of such readmissions.

(C). Analysis of the data:

To analyse the data, I first created a 70/30 training/testing split. I then started out with a simple EDA(Exploratory Data Analysis) of the data so see the distribution of various features. It became immediately obvious that certain features such as glimepiride, metformin, diag2_mod, diag3_mod etc. are are unlikely to be predictive of readmission due to low variability. I then performed cross validation to identify features that have non zero coefficients. I used this subset of features and fit a Logistic Regression Model on the training data, and studied the performance . I then fit the models oon the testing data to see how these models generalize to unseen data. The Logistic Regression model performed reasonably well on the testing set, yielding an AUC value of 0.63, and a specificity of 0.47.

(D). Limitations of the Analysis:

The most important limitation in this study is that Diabetic encounters are not all encounters of diabetes patients, but rather only these where diabetes was coded as an existing health condition. Thus we are working with only a fraction of the total number of patients with diabetes.

Part 2: Detailed process of the Analysis:

Step 1: Data Summary

Looking at the data:

```
rm(list=ls()) # Remove all the existing variables
data <- read.csv("readmission.csv")
str(data)
```

```
## 'data.frame':    101766 obs. of  31 variables:
##  $ encounter_id       : int  12522 15738 16680 28236 35754 36900 40926 42570 55842 62256 ...
##  $ patient_nbr        : int  48330783 63555939 42519267 89869032 82637451 77391171 85504905 77586282
##  $ race               : Factor w/ 6 levels "?","AfricanAmerican",..: 4 4 4 2 4 2 4 4 4 2 ...
##  $ gender             : Factor w/ 3 levels "Female","Male",..: 1 1 2 1 2 2 1 2 2 1 ...
##  $ time_in_hospital   : int  13 12 1 9 3 7 7 10 4 1 ...
##  $ num_lab_procedures : int  68 33 51 47 31 62 60 55 70 49 ...
##  $ num_procedures     : int  2 3 0 2 6 0 0 1 1 5 ...
##  $ num_medications    : int  28 18 8 17 16 11 15 31 21 2 ...
##  $ number_outpatient  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ number_emergency   : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ number_inpatient   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ number_diagnoses   : int  8 8 5 9 9 7 8 8 7 8 ...
##  $ max_glu_serum      : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ A1Cresult          : Factor w/ 4 levels ">7",">8","None",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ metformin          : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 3 2 3 2 ...
##  $ glimepiride        : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 3 2 ...
##  $ glipizide          : Factor w/ 4 levels "Down","No","Steady",..: 3 2 3 2 2 2 2 2 2 2 ...
##  $ glyburide          : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 4 2 2 2 2 ...
##  $ pioglitazone       : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ rosiglitazone      : Factor w/ 4 levels "Down","No","Steady",..: 2 3 2 2 2 2 2 2 2 2 ...
##  $ insulin            : Factor w/ 4 levels "Down","No","Steady",..: 3 3 3 3 3 3 1 3 3 3 ...
##  $ change             : Factor w/ 2 levels "Ch","No": 1 1 1 2 2 1 1 2 1 2 ...
##  $ diabetesMed        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ disch_disp_modified: Factor w/ 4 levels "Discharged to home",..: 1 3 1 1 1 1 3 2 1 1 ...
##  $ adm_src_mod        : Factor w/ 4 levels "Emergency Room",..: 2 2 1 1 2 2 1 1 2 2 ...
##  $ adm_typ_mod        : Factor w/ 4 levels "Elective","Emergency",..: 4 1 2 2 4 4 2 2 1 1 ...
##  $ age_mod            : Factor w/ 4 levels "0-19","20-59",..: 4 4 2 2 2 3 2 4 3 3 ...
##  $ diag1_mod          : Factor w/ 24 levels "250.6","250.8",..: 24 9 24 24 6 24 8 8 6 14 ...
##  $ diag2_mod          : Factor w/ 25 levels "250","250.01",..: 13 25 25 7 8 25 25 8 8 25 ...
##  $ diag3_mod          : Factor w/ 21 levels "?","250","250.02",..: 20 20 2 20 2 20 4 13 21 20 ...
##  $ readmitted         : Factor w/ 3 levels "<30",">30","NO": 3 3 3 2 2 1 1 3 3 2 ...
```

Getting a quick data summary:

```
summary(data)
```

```
##   encounter_id        patient_nbr                     race
##  Min.   :    12522   Min.   :      135   ?                : 2273
##  1st Qu.: 84961194   1st Qu.: 23413221   AfricanAmerican:19210
##  Median :152388987   Median : 45505143   Asian           :  641
##  Mean   :165201646   Mean   : 54330401   Caucasian       :76099
##  3rd Qu.:230270888   3rd Qu.: 87545950   Hispanic        : 2037
##  Max.   :443867222   Max.   :189502619   Other           : 1506
##
```

```
##              gender      time_in_hospital num_lab_procedures
##  Female         :54708   Min.   : 1.000   Min.   :  1.0
##  Male           :47055   1st Qu.: 2.000   1st Qu.: 31.0
##  Unknown/Invalid:    3   Median : 4.000   Median : 44.0
##                          Mean   : 4.396   Mean   : 43.1
##                          3rd Qu.: 6.000   3rd Qu.: 57.0
##                          Max.   :14.000   Max.   :132.0
##
##  num_procedures num_medications number_outpatient number_emergency
##  Min.   :0.00   Min.   : 1.00   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :1.00   Median :15.00   Median : 0.0000   Median : 0.0000
##  Mean   :1.34   Mean   :16.02   Mean   : 0.3694   Mean   : 0.1978
##  3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :6.00   Max.   :81.00   Max.   :42.0000   Max.   :76.0000
##
##  number_inpatient  number_diagnoses max_glu_serum A1Cresult
##  Min.   : 0.0000   Min.   : 1.000   >200: 1485    >7  : 3812
##  1st Qu.: 0.0000   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  Median : 0.0000   Median : 8.000   None:96420    None:84748
##  Mean   : 0.6356   Mean   : 7.423   Norm: 2597    Norm: 4990
##  3rd Qu.: 1.0000   3rd Qu.: 9.000
##  Max.   :21.0000   Max.   :16.000
##
##   metformin     glimepiride     glipizide      glyburide
##  Down  :  575   Down  :  194   Down  :  560   Down  :  564
##  No    :81778   No    :96575   No    :89080   No    :91116
##  Steady:18346   Steady: 4670   Steady:11356   Steady: 9274
##  Up    : 1067   Up    :  327   Up    :  770   Up    :  812
##
##
##
##  pioglitazone   rosiglitazone    insulin      change     diabetesMed
##  Down  :  118   Down  :   87   Down  :12218   Ch:47011   No :23403
##  No    :94438   No    :95401   No    :47383   No:54755   Yes:78363
##  Steady: 6976   Steady: 6100   Steady:30849
##  Up    :  234   Up    :  178   Up    :11316
##
##
##
##                                  disch_disp_modified
##  Discharged to home                          :60234
##  Discharged to home with Home Health Service:12902
##  Discharged/Transferred to SNF               :13954
##  Other                                       :14676
##
##
##
##                     adm_src_mod       adm_typ_mod     age_mod
##  Emergency Room           :57494   Elective :18869   0-19 :  852
##  Other                    : 7926   Emergency:53990   20-59:32373
##  Physician Referral       :29565   Other    :10427   60-79:48551
##  Transfer from Home Health: 6781   Urgent   :18480   80+  :19990
##
```

```
##
##
##    diag1_mod       diag2_mod       diag3_mod     readmitted
##  Other  :47056   Other  :37491   Other  :42333   <30:11357
##  428    : 6862   276    : 6752   250    :11555   >30:35545
##  414    : 6581   428    : 6662   401    : 8289   NO :54864
##  786    : 4016   250    : 6071   276    : 5175
##  410    : 3614   427    : 5036   428    : 4577
##  486    : 3508   401    : 3736   427    : 3955
##  (Other):30129   (Other):36018   (Other):25882
```

Refactoring the target variable to have only 0s and 1s.

```
library(ggplot2)
require("car")

data$readmitted <- ifelse(data$readmitted == "<30", "Yes","No")
```

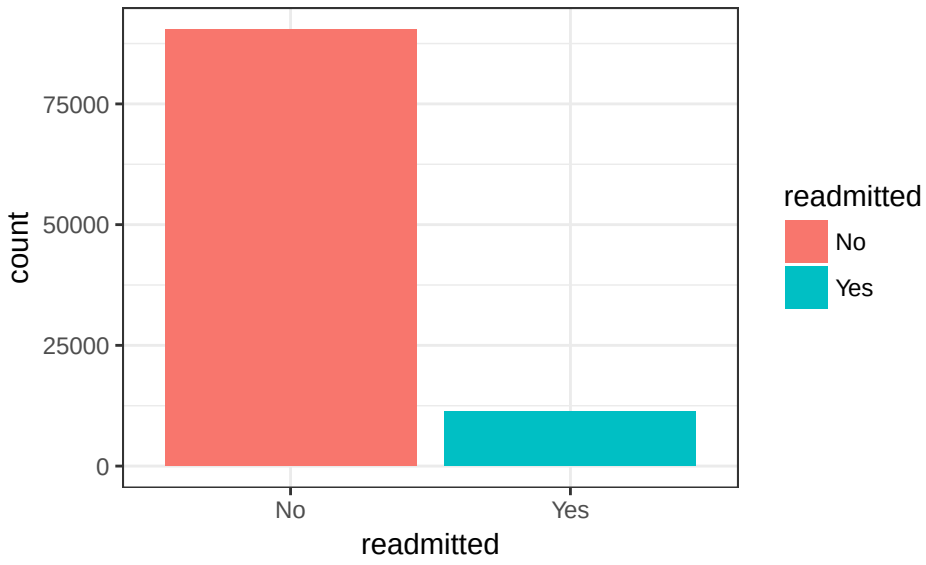
Creating Training and Testing sets.

```
library(caret)
Train <- createDataPartition(data$readmitted, p=0.7, list=FALSE)
training <- data[ Train, ]
testing <- data[ -Train, ]
```

Step 2: Analysis(EDA, Feature Selection, Fitting Models and Evalutaing models on testing data):

Let's do some simple EDA. We'll begin by exploring the distribution by readmission.

```
ggplot(data = data) +
  geom_bar(aes(x = readmitted , fill = readmitted)) +
  theme_bw()
```




```
  labs(list(title="Distribution by Readmission", x = "0 or 1", y = "Count"))
```

```
## $title
```

```
## [1] "Distribution by Readmission"
##
## $x
## [1] "0 or 1"
##
## $y
## [1] "Count"
##
## attr(,"class")
## [1] "labels"
```

It is interesting to note that only ~ 10% of the patients get readmitted.

Next we'll examine some variables that seem to have low variability

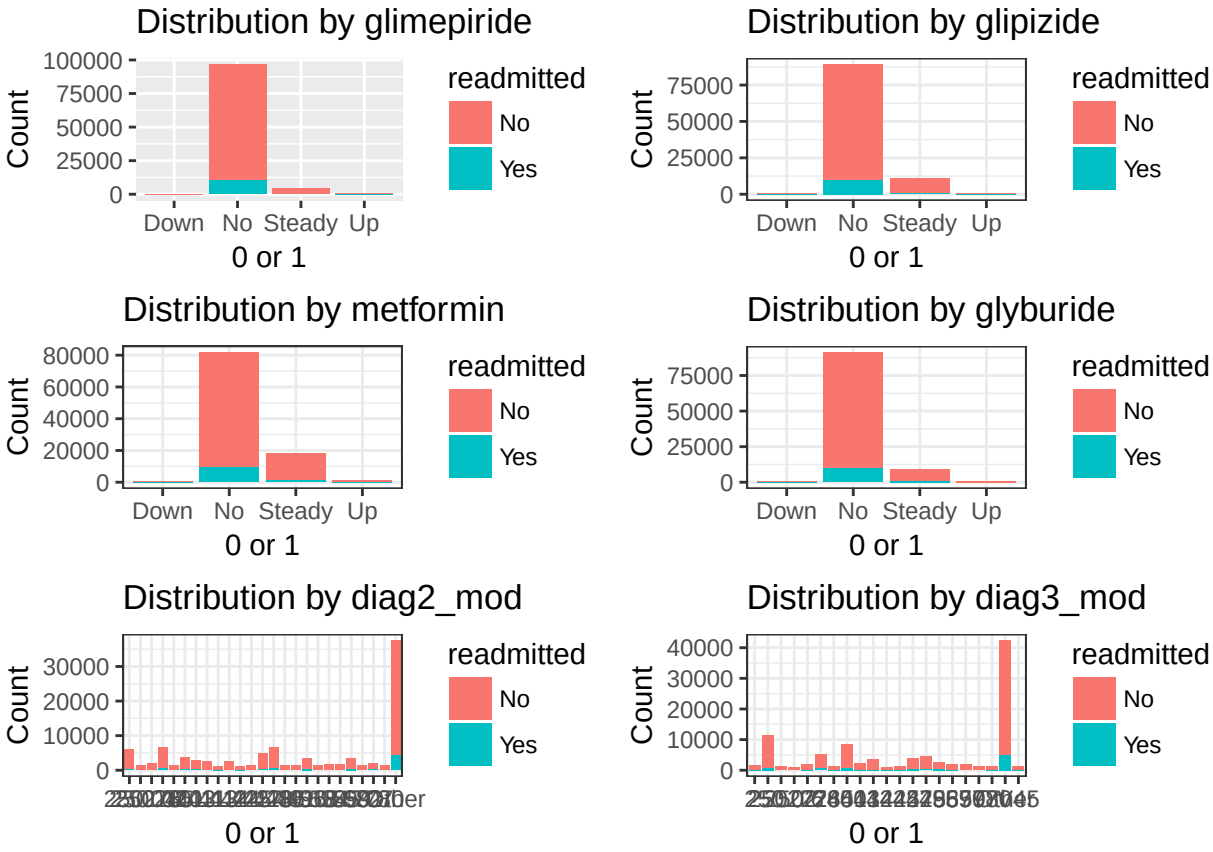
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
## Warning: package 'gridExtra' was built under R version 3.3.3
```

```
plot1 <- ggplot(data = data) +
  geom_bar(aes(x = glimepiride, fill = readmitted)) +
  labs(list(title="Distribution by glimepiride", x = "0 or 1", y = "Count"))
plot2 <- ggplot(data = data) +
  geom_bar(aes(x = glipizide, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by glipizide", x = "0 or 1", y = "Count"))
plot3 <- ggplot(data = data) +
  geom_bar(aes(x = metformin, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by metformin", x = "0 or 1", y = "Count"))
plot4 <- ggplot(data = data) +
  geom_bar(aes(x = glyburide, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by glyburide", x = "0 or 1", y = "Count"))
plot5 <- ggplot(data = data) +
  geom_bar(aes(x = diag2_mod, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by diag2_mod", x = "0 or 1", y = "Count"))
plot6 <- ggplot(data = data) +
  geom_bar(aes(x = diag3_mod, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by diag3_mod", x = "0 or 1", y = "Count"))

grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol=2)
```



As we can see in the plots above, these variables are unlikely to be stastically significant as they have low variability. We can take tese variables out of the data.

Cross Validation and Fitting Models on Data:

I then run cross validation with the cv.glmnet package to identify the most statistically significant variables, and fit a logistic regression model on the data(See Appendix for details). I then made predictions on the testing data.

Step 3: Summary and Conclusions:

My final model contained the following features: race, gender, time_in_hospital, num_lab_procedures, num_procedures, num_medications , number_outpatient , number_emergency , number_inpatient , number_diagnoses, max_glu_serum, disch_disp_modified, adm_src_mod, age_mod and diag1_mod.

The performance statistics of the Logistic Regression Model on the testing set are as follows:

Confusion Matrix and Statistics

```
      Reference
```

Prediction 0 1 0 27068 54 1 3359 48

```
           Accuracy : 0.8882
             95% CI : (0.8846, 0.8917)
No Information Rate : 0.9967
P-Value [Acc > NIR] : 1
```

```
              Kappa : 0.021
```

Mcnemar's Test P-Value : <2e-16

```
        Sensitivity : 0.88960
        Specificity : 0.47059
     Pos Pred Value : 0.99801
     Neg Pred Value : 0.01409
         Prevalence : 0.99666
     Detection Rate : 0.88663
```

Detection Prevalence : 0.88840
Balanced Accuracy : 0.68010 (See full model and summary in Appendix)

For the random forest model, I choose the features that have highest predictive power, and plot it to see the performance. (Check Appendix)

Appendix

In this section I present the full R code for my analysis in the rmd format.

Data Summary

```
rm(list=ls()) # Remove all the existing variables
data <- read.csv("readmission.csv")
str(data)
```

```
## 'data.frame':    101766 obs. of  31 variables:
##  $ encounter_id       : int  12522 15738 16680 28236 35754 36900 40926 42570 55842 62256 ...
##  $ patient_nbr        : int  48330783 63555939 42519267 89869032 82637451 77391171 85504905 77586282
##  $ race               : Factor w/ 6 levels "?","AfricanAmerican",..: 4 4 4 2 4 2 4 4 4 2 ...
##  $ gender             : Factor w/ 3 levels "Female","Male",..: 1 1 2 1 2 2 1 2 2 1 ...
##  $ time_in_hospital   : int  13 12 1 9 3 7 7 10 4 1 ...
##  $ num_lab_procedures : int  68 33 51 47 31 62 60 55 70 49 ...
##  $ num_procedures     : int  2 3 0 2 6 0 0 1 1 5 ...
##  $ num_medications    : int  28 18 8 17 16 11 15 31 21 2 ...
##  $ number_outpatient  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ number_emergency   : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ number_inpatient   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ number_diagnoses   : int  8 8 5 9 9 7 8 8 7 8 ...
##  $ max_glu_serum      : Factor w/ 4 levels ">200",">300",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ A1Cresult          : Factor w/ 4 levels ">7",">8","None",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ metformin          : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 3 2 3 2 ...
##  $ glimepiride        : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 3 2 ...
##  $ glipizide          : Factor w/ 4 levels "Down","No","Steady",..: 3 2 3 2 2 2 2 2 2 2 ...
##  $ glyburide          : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 4 2 2 2 2 ...
##  $ pioglitazone       : Factor w/ 4 levels "Down","No","Steady",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ rosiglitazone      : Factor w/ 4 levels "Down","No","Steady",..: 2 3 2 2 2 2 2 2 2 2 ...
##  $ insulin            : Factor w/ 4 levels "Down","No","Steady",..: 3 3 3 3 3 3 1 3 3 3 ...
##  $ change             : Factor w/ 2 levels "Ch","No": 1 1 1 2 2 1 1 2 1 2 ...
##  $ diabetesMed        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ disch_disp_modified: Factor w/ 4 levels "Discharged to home",..: 1 3 1 1 1 1 3 2 1 1 ...
```

```
##  $ adm_src_mod        : Factor w/ 4 levels "Emergency Room",..: 2 2 1 1 2 2 1 1 2 2 ...
##  $ adm_typ_mod        : Factor w/ 4 levels "Elective","Emergency",..: 4 1 2 2 4 4 2 2 1 1 ...
##  $ age_mod            : Factor w/ 4 levels "0-19","20-59",..: 4 4 2 2 2 3 2 4 3 3 ...
##  $ diag1_mod          : Factor w/ 24 levels "250.6","250.8",..: 24 9 24 24 6 24 8 8 6 14 ...
##  $ diag2_mod          : Factor w/ 25 levels "250","250.01",..: 13 25 25 7 8 25 25 8 8 25 ...
##  $ diag3_mod          : Factor w/ 21 levels "?","250","250.02",..: 20 20 2 20 2 20 4 13 21 20 ...
##  $ readmitted         : Factor w/ 3 levels "<30",">30","NO": 3 3 3 2 2 1 1 3 3 2 ...
```

```
summary(data)
```

```
##   encounter_id        patient_nbr                     race
##  Min.   :    12522   Min.   :      135   ?              : 2273
##  1st Qu.: 84961194   1st Qu.: 23413221   AfricanAmerican:19210
##  Median :152388987   Median : 45505143   Asian          :  641
##  Mean   :165201646   Mean   : 54330401   Caucasian      :76099
##  3rd Qu.:230270888   3rd Qu.: 87545950   Hispanic       : 2037
##  Max.   :443867222   Max.   :189502619   Other          : 1506
##
##              gender      time_in_hospital num_lab_procedures
##  Female         :54708   Min.   : 1.000   Min.   :  1.0
##  Male           :47055   1st Qu.: 2.000   1st Qu.: 31.0
##  Unknown/Invalid:    3   Median : 4.000   Median : 44.0
##                          Mean   : 4.396   Mean   : 43.1
##                          3rd Qu.: 6.000   3rd Qu.: 57.0
##                          Max.   :14.000   Max.   :132.0
##
##  num_procedures num_medications number_outpatient number_emergency
##  Min.   :0.00   Min.   : 1.00   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :1.00   Median :15.00   Median : 0.0000   Median : 0.0000
##  Mean   :1.34   Mean   :16.02   Mean   : 0.3694   Mean   : 0.1978
##  3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :6.00   Max.   :81.00   Max.   :42.0000   Max.   :76.0000
##
##  number_inpatient  number_diagnoses max_glu_serum A1Cresult
##  Min.   : 0.0000   Min.   : 1.000   >200: 1485    >7  : 3812
##  1st Qu.: 0.0000   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  Median : 0.0000   Median : 8.000   None:96420    None:84748
##  Mean   : 0.6356   Mean   : 7.423   Norm: 2597    Norm: 4990
##  3rd Qu.: 1.0000   3rd Qu.: 9.000
##  Max.   :21.0000   Max.   :16.000
##
##   metformin     glimepiride     glipizide      glyburide
##  Down  :  575   Down  :  194   Down  :  560   Down  :  564
##  No    :81778   No    :96575   No    :89080   No    :91116
##  Steady:18346   Steady: 4670   Steady:11356   Steady: 9274
##  Up    : 1067   Up    :  327   Up    :  770   Up    :  812
##
##
##
##  pioglitazone   rosiglitazone    insulin      change     diabetesMed
##  Down  :  118   Down  :   87   Down  :12218   Ch:47011   No :23403
##  No    :94438   No    :95401   No    :47383   No:54755   Yes:78363
##  Steady: 6976   Steady: 6100   Steady:30849
##  Up    :  234   Up    :  178   Up    :11316
```

```
##
##
##
##                                   disch_disp_modified
##  Discharged to home                          :60234
##  Discharged to home with Home Health Service:12902
##  Discharged/Transferred to SNF               :13954
##  Other                                       :14676
##
##
##
##                     adm_src_mod       adm_typ_mod     age_mod
##  Emergency Room           :57494   Elective :18869   0-19 :  852
##  Other                    : 7926   Emergency:53990   20-59:32373
##  Physician Referral       :29565   Other    :10427   60-79:48551
##  Transfer from Home Health: 6781   Urgent   :18480   80+  :19990
##
##
##
##    diag1_mod       diag2_mod       diag3_mod     readmitted
##  Other  :47056   Other  :37491   Other  :42333   <30:11357
##  428    : 6862   276    : 6752   250    :11555   >30:35545
##  414    : 6581   428    : 6662   401    : 8289   NO :54864
##  786    : 4016   250    : 6071   276    : 5175
##  410    : 3614   427    : 5036   428    : 4577
##  486    : 3508   401    : 3736   427    : 3955
##  (Other):30129   (Other):36018   (Other):25882
```

```
head(data)
```

```
##   encounter_id patient_nbr            race gender time_in_hospital
## 1        12522    48330783       Caucasian Female               13
## 2        15738    63555939       Caucasian Female               12
## 3        16680    42519267       Caucasian   Male                1
## 4        28236    89869032 AfricanAmerican Female                9
## 5        35754    82637451       Caucasian   Male                3
## 6        36900    77391171 AfricanAmerican   Male                7
##   num_lab_procedures num_procedures num_medications number_outpatient
## 1                 68              2              28                 0
## 2                 33              3              18                 0
## 3                 51              0               8                 0
## 4                 47              2              17                 0
## 5                 31              6              16                 0
## 6                 62              0              11                 0
##   number_emergency number_inpatient number_diagnoses max_glu_serum
## 1                0                0                8          None
## 2                0                0                8          None
## 3                0                0                5          None
## 4                0                0                9          None
## 5                0                0                9          None
## 6                0                0                7          None
##   A1Cresult metformin glimepiride glipizide glyburide pioglitazone
## 1      None        No          No    Steady        No           No
## 2      None        No          No        No        No           No
## 3      None        No          No    Steady        No           No
```

```
## 4      None        No          No        No        No           No
## 5      None        No          No        No        No           No
## 6      None        No          No        No        Up           No
##   rosiglitazone insulin change diabetesMed           disch_disp_modified
## 1            No  Steady     Ch         Yes            Discharged to home
## 2        Steady  Steady     Ch         Yes Discharged/Transferred to SNF
## 3            No  Steady     Ch         Yes            Discharged to home
## 4            No  Steady     No         Yes            Discharged to home
## 5            No  Steady     No         Yes            Discharged to home
## 6            No  Steady     Ch         Yes            Discharged to home
##      adm_src_mod adm_typ_mod age_mod diag1_mod diag2_mod diag3_mod
## 1          Other      Urgent     80+     Other       427     Other
## 2          Other    Elective     80+       434     Other     Other
## 3 Emergency Room   Emergency   20-59     Other     Other       250
## 4 Emergency Room   Emergency   20-59     Other       403     Other
## 5          Other      Urgent   20-59       414       411       250
## 6          Other      Urgent   60-79     Other     Other     Other
##   readmitted
## 1         NO
## 2         NO
## 3         NO
## 4        >30
## 5        >30
## 6        <30
```

Refactoring the target variable to have only 0s and 1s.

```
library(ggplot2)
require("car")

data$readmitted <- ifelse(data$readmitted == "<30", "Yes","No")
```

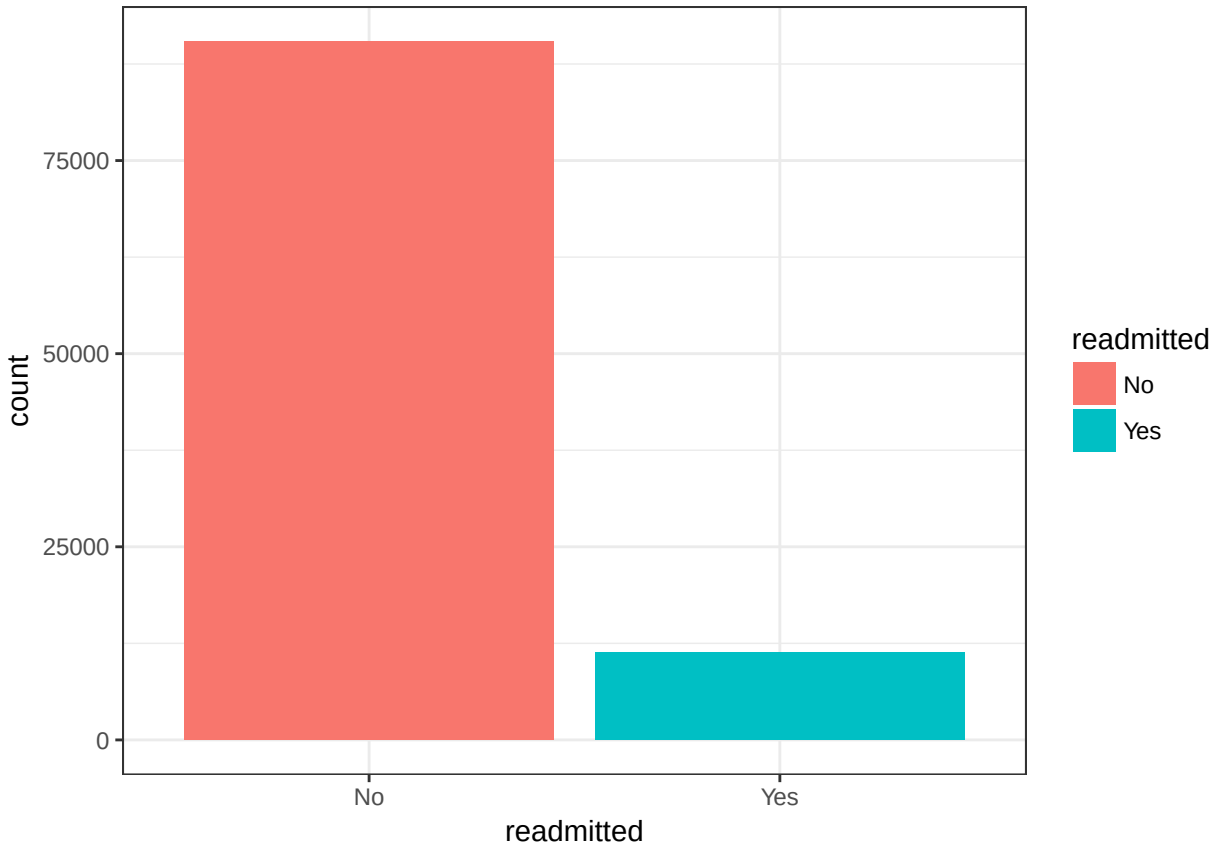
Creating Training and Testing sets.

```
library(caret)
Train <- createDataPartition(data$readmitted, p=0.7, list=FALSE)
training <- data[ Train, ]
testing <- data[ -Train, ]
```

Exploratory Data Analysis

Let's do some simple EDA. We'll begin by exploring the distribution by readmission.

```
ggplot(data = data) +
  geom_bar(aes(x = readmitted , fill = readmitted)) +
  theme_bw()
```



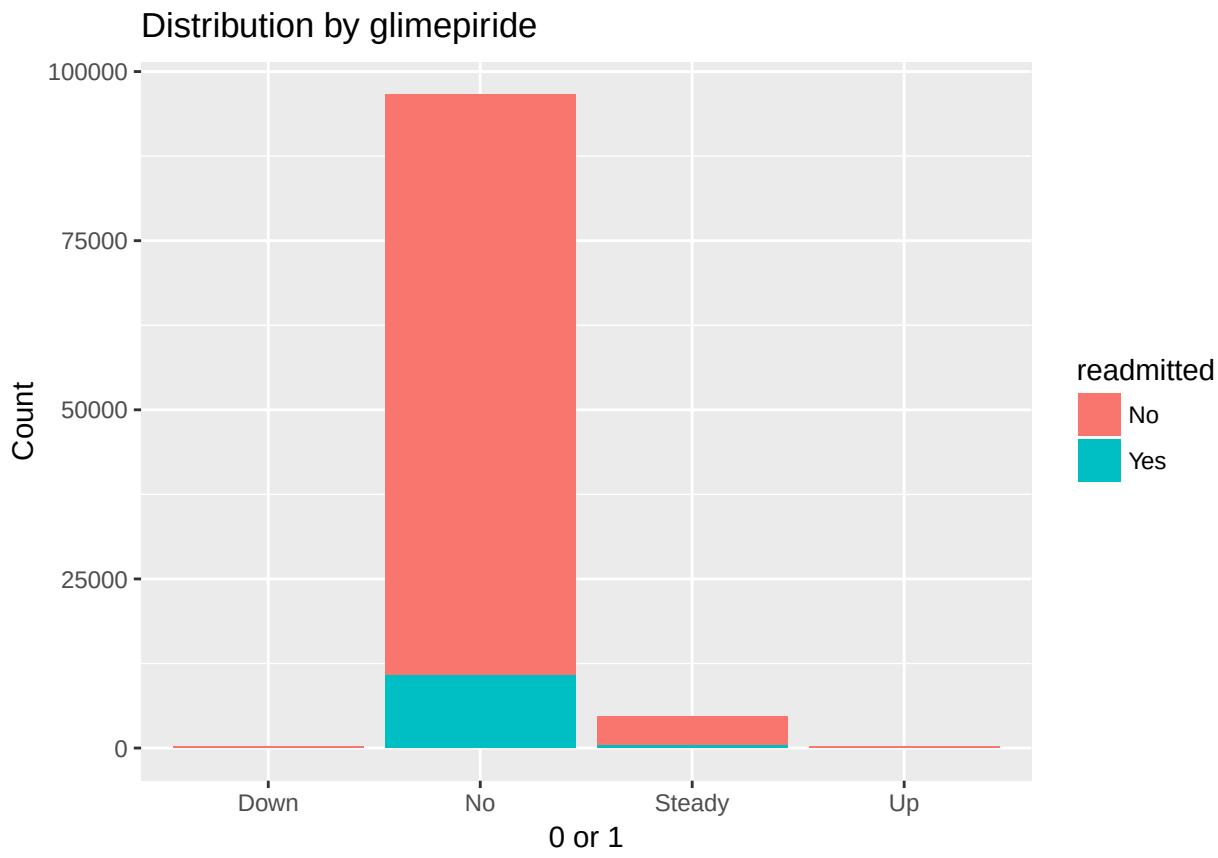

```
  labs(list(title="Distribution by Readmission", x = "0 or 1", y = "Count"))
```

```
## $title
## [1] "Distribution by Readmission"
##
## $x
## [1] "0 or 1"
##
## $y
## [1] "Count"
##
## attr(,"class")
## [1] "labels"
```

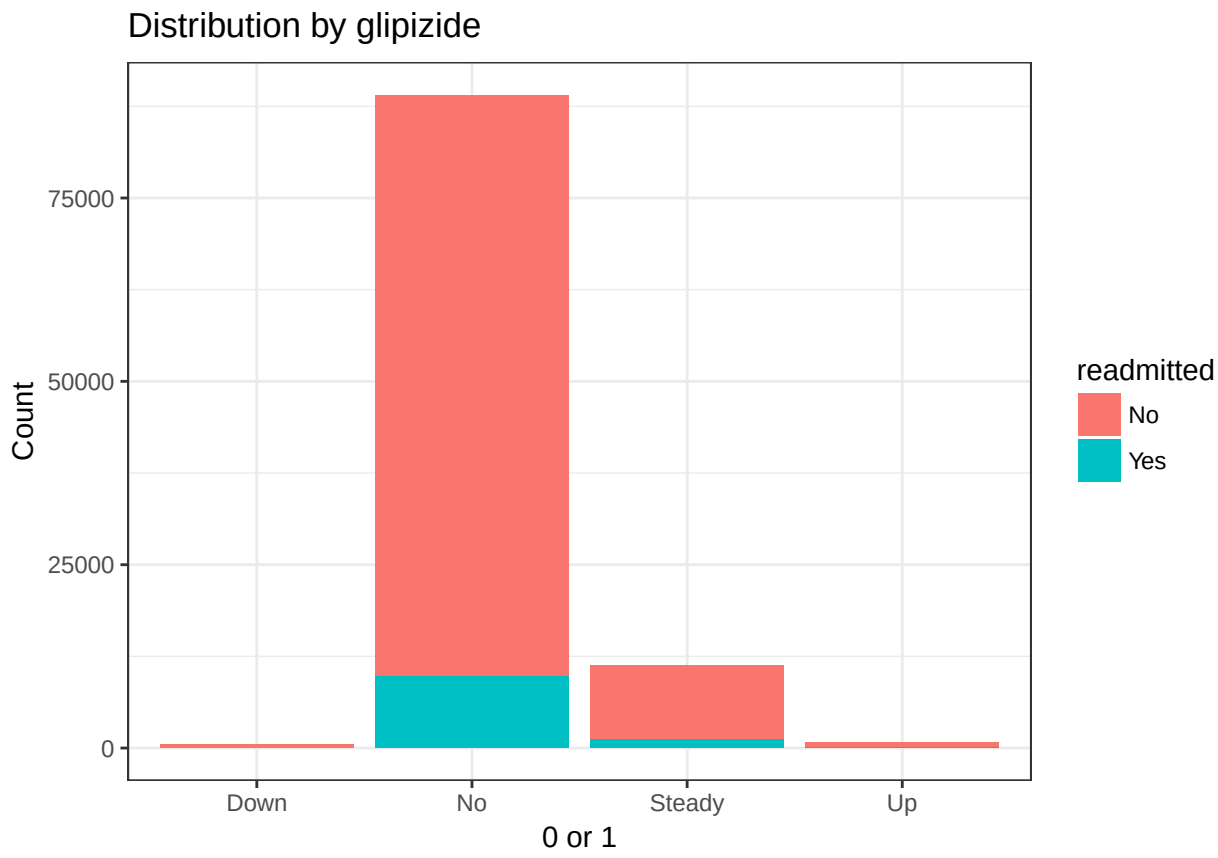
It is interesting to note that only ~ 10% of the patients get readmitted.

Next we'll examine some variables that seem to have low variability

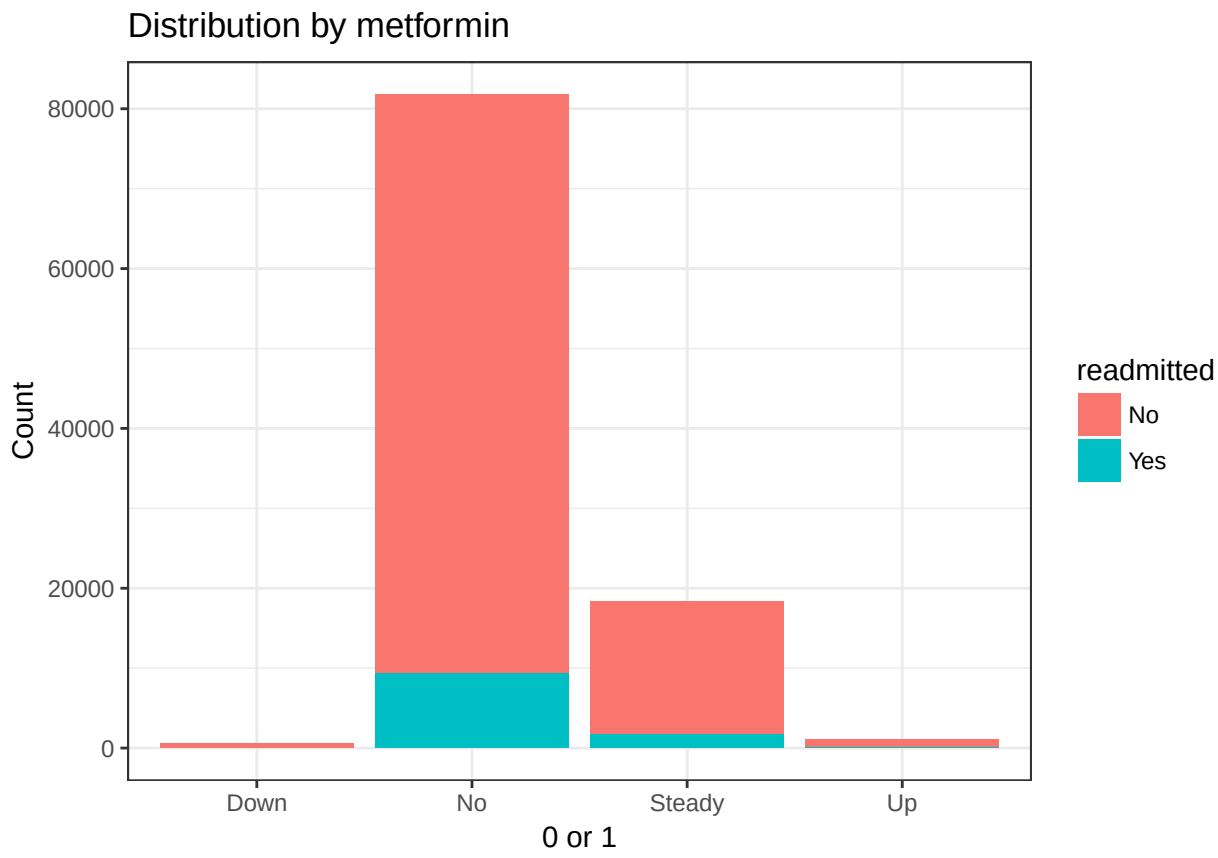
```
library(ggplot2)
ggplot(data = data) +
  geom_bar(aes(x = glimepiride, fill = readmitted)) +
  labs(list(title="Distribution by glimepiride", x = "0 or 1", y = "Count"))
```



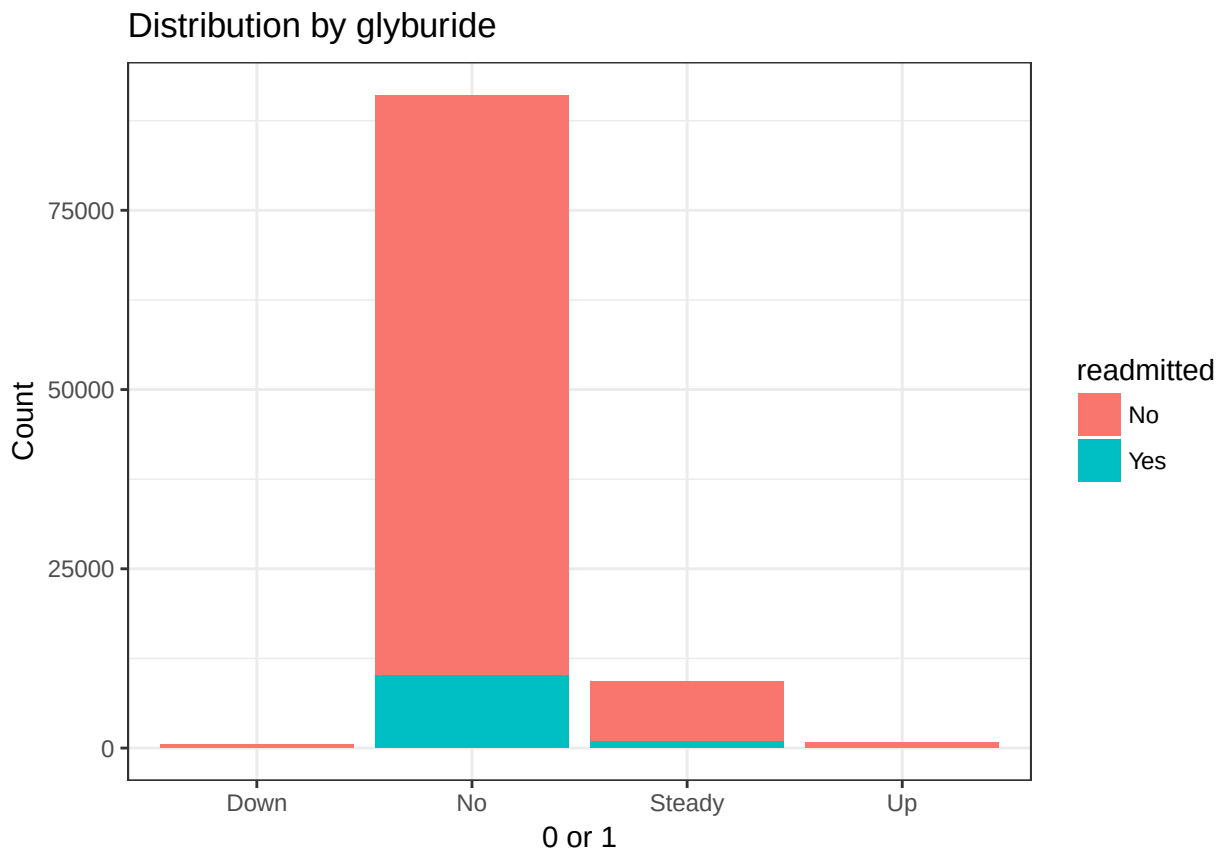

```
ggplot(data = data) +
  geom_bar(aes(x = glipizide, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by glipizide", x = "0 or 1", y = "Count"))
```



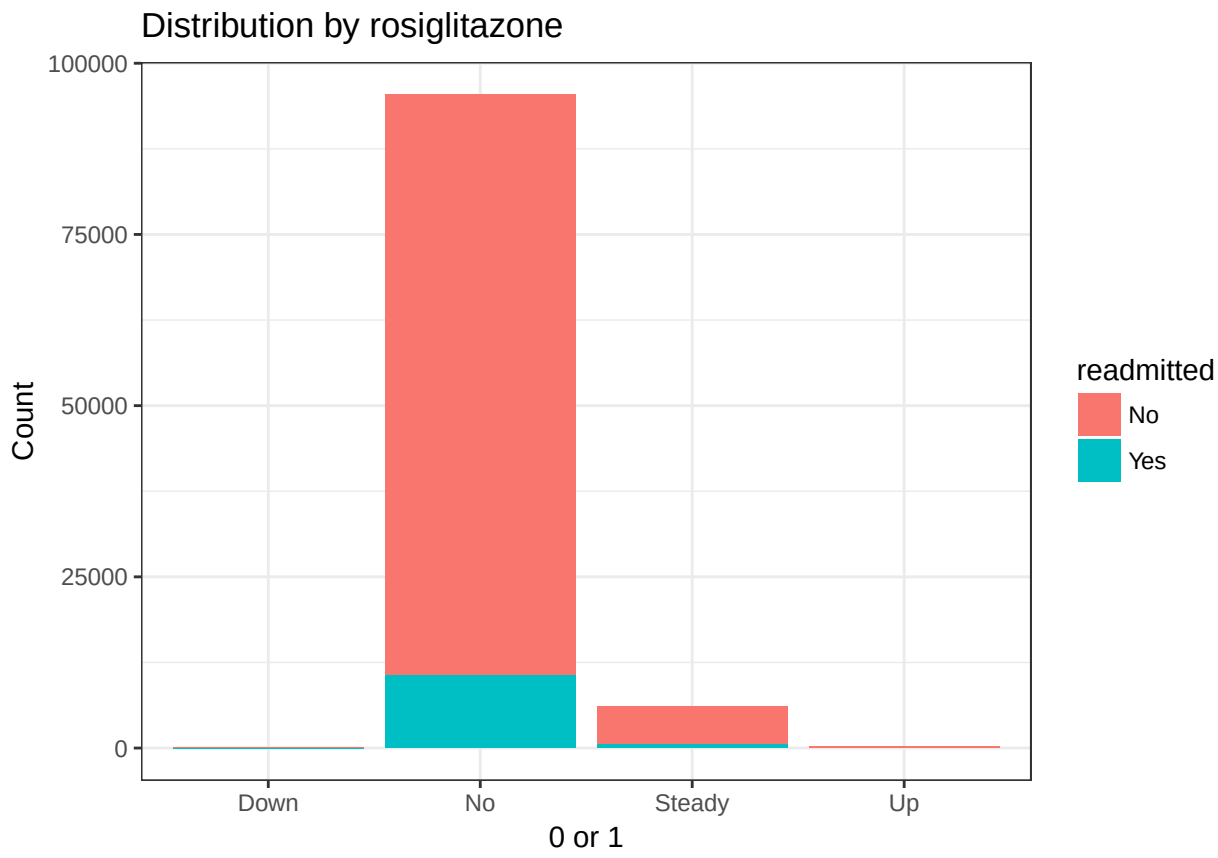

```
ggplot(data = data) +
  geom_bar(aes(x = metformin, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by metformin", x = "0 or 1", y = "Count"))
```



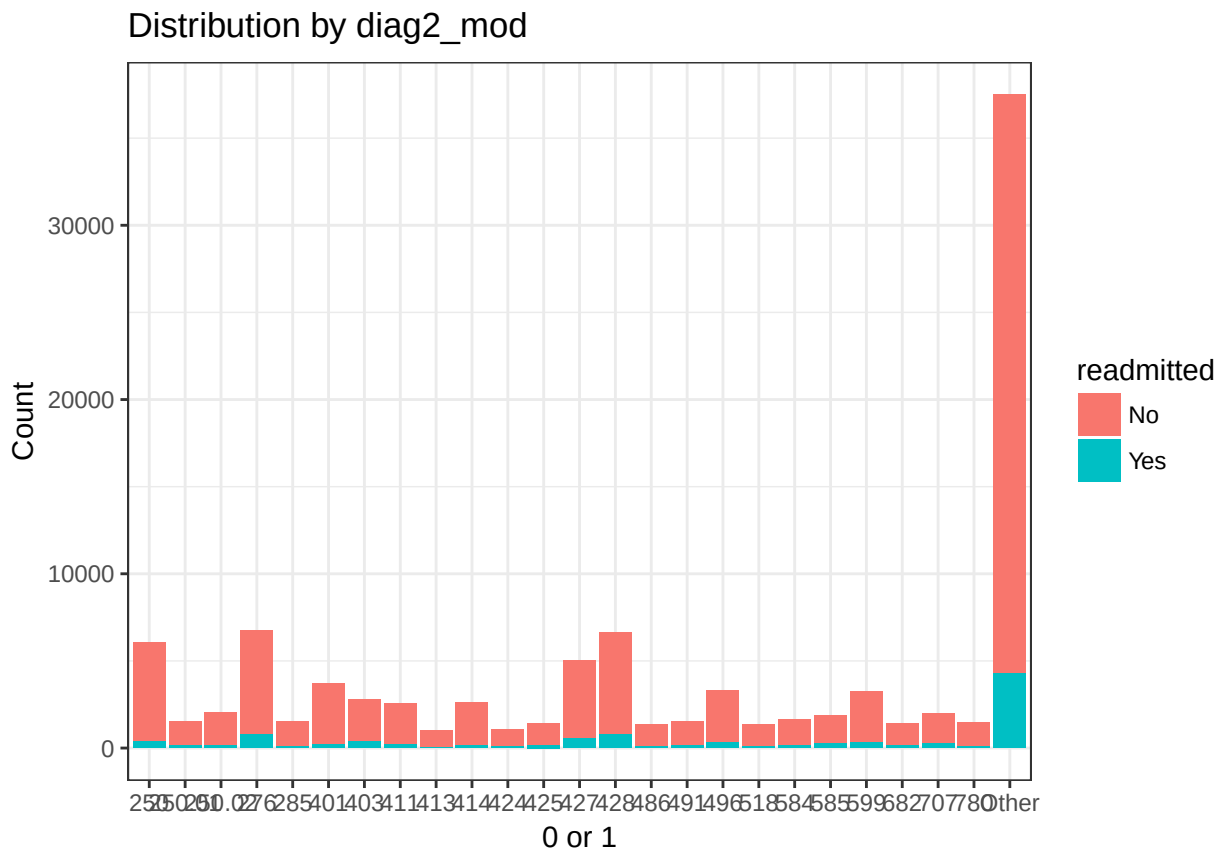

```
ggplot(data = data) +
  geom_bar(aes(x = glyburide, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by glyburide", x = "0 or 1", y = "Count"))
```



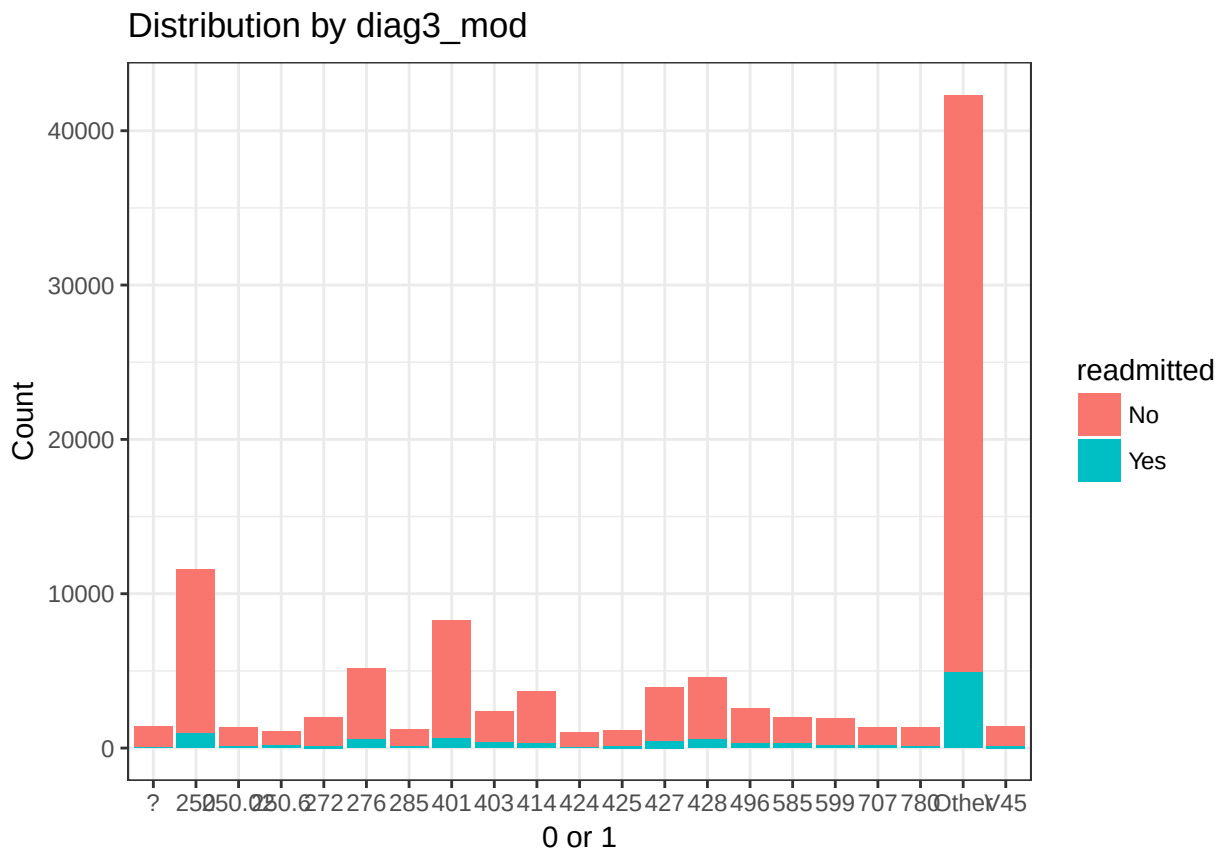

```
ggplot(data = data) +
  geom_bar(aes(x = rosiglitazone, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by rosiglitazone", x = "0 or 1", y = "Count"))
```




```
ggplot(data = data) +
  geom_bar(aes(x = diag2_mod, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by diag2_mod", x = "0 or 1", y = "Count"))
```




```
ggplot(data = data) +
  geom_bar(aes(x = diag3_mod, fill = readmitted)) +
  theme_bw() +
  labs(list(title="Distribution by diag3_mod", x = "0 or 1", y = "Count"))
```



Feature Selection, fitting models on data

Deleting some variables Let's remove the variables that are unlikely to be predictive of readmission

```
training = subset(training, select = -c(patient_nbr, encounter_id,
                            metformin, glimepiride, glipizide, glyburide, rosiglitazone,diag2_mod, diag
testing = subset(testing, select = -c(patient_nbr, encounter_id,
                                metformin, glimepiride, glipizide, glyburide, rosiglitazone, diag2_mod,
```

```
training$readmitted <- ifelse(training$readmitted == "Yes", 1,0)
testing$readmitted <- ifelse(testing$readmitted == "Yes", 1,0)
```

Now lets run cross validation with glmnet to identify important

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.3.3
```

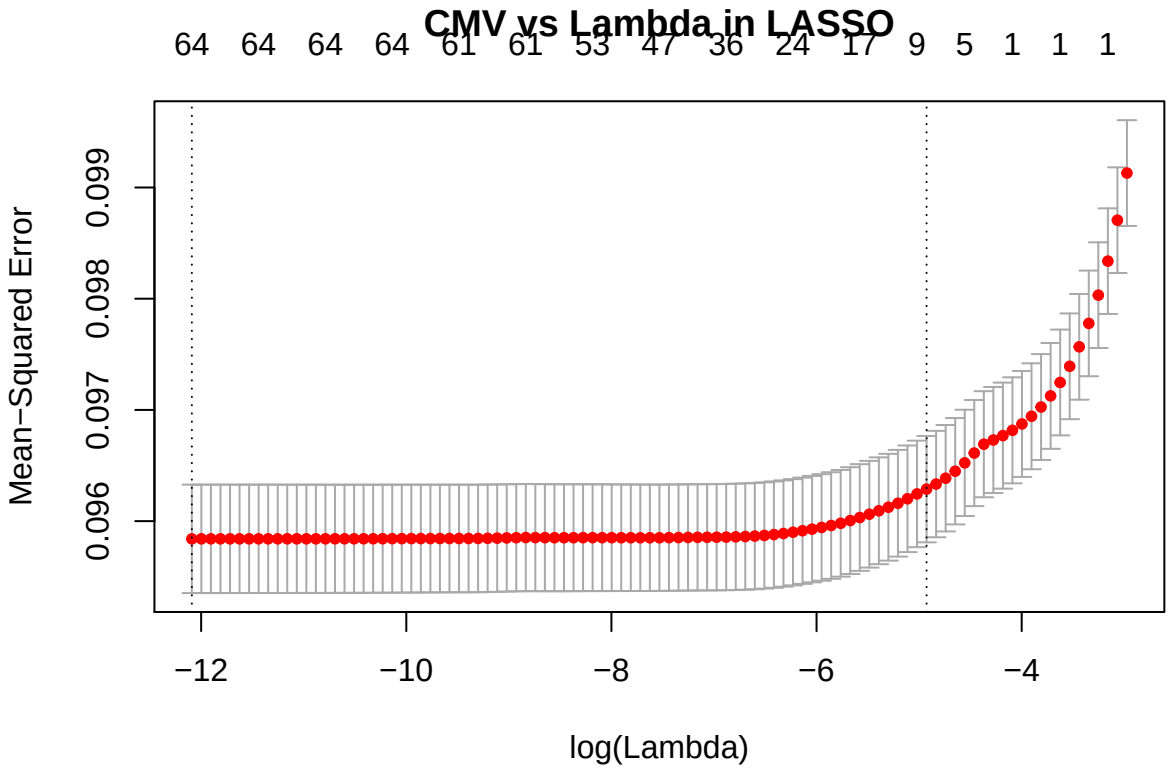
```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-5
```

```
y_col <- training$readmitted
x_col <- model.matrix(readmitted ~. , training)
```

```
fit_glm <- cv.glmnet(x_col, as.numeric(y_col),  alpha = 1)
plot(fit_glm, main = "CMV vs Lambda in LASSO")
```

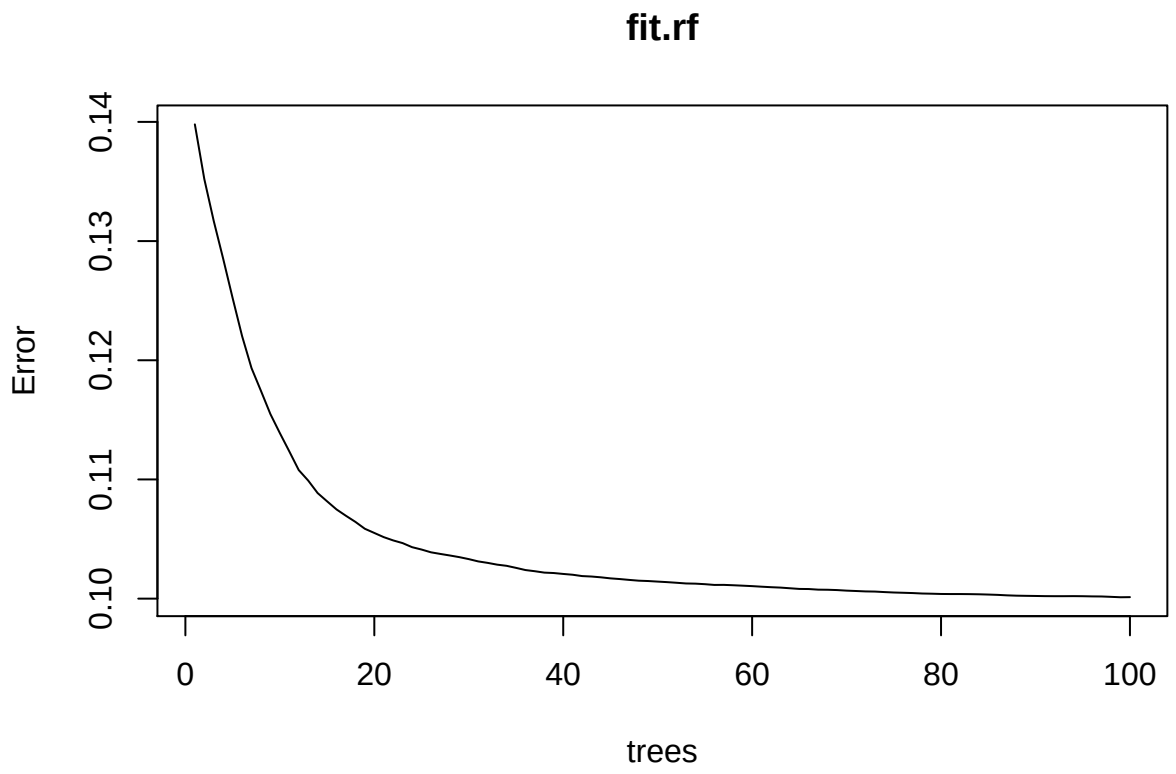



```
coefs_1se = coef(fit_glm, s="lambda.1se")
rownames(coefs_1se)[which((coefs_1se) != 0)]
```

```
## [1] "(Intercept)"
## [2] "time_in_hospital"
## [3] "number_inpatient"
## [4] "number_diagnoses"
## [5] "disch_disp_modifiedDischarged/Transferred to SNF"
## [6] "disch_disp_modifiedOther"
## [7] "diag1_mod434"
```

```
library(randomForest)
rf_formula <- formula(readmitted ~ number_inpatient + number_diagnoses + disch_disp_modified + time_in_h

fit.rf <- randomForest(rf_formula, data = training, ntree = 100)
plot(fit.rf)
```




```
fit2 <- glm(readmitted~., data = training, family = binomial())
summary(fit2)
```

```
##
## Call:
## glm(formula = readmitted ~ ., family = binomial(), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1528  -0.5051  -0.4329  -0.3739   2.6392
##
## Coefficients:
##                                                    Estimate
## (Intercept)                                      -2.9400335
## raceAfricanAmerican                               0.1731020
## raceAsian                                         0.0493442
## raceCaucasian                                     0.1627142
## raceHispanic                                      0.1704561
## raceOther                                         0.0440496
## genderMale                                        0.0260502
## genderUnknown/Invalid                            -8.0458186
## time_in_hospital                                  0.0123335
## num_lab_procedures                                0.0003866
## num_procedures                                   -0.0149152
## num_medications                                   0.0039607
## number_outpatient                                -0.0001969
```

```
## number_emergency                                                    0.0349608
## number_inpatient                                                    0.2526688
## number_diagnoses                                                    0.0414253
## max_glu_serum>300                                                  -0.1324902
## max_glu_serumNone                                                  -0.2050395
## max_glu_serumNorm                                                  -0.0239506
## A1Cresult>8                                                         0.0009436
## A1CresultNone                                                       0.0924990
## A1CresultNorm                                                      -0.0052607
## pioglitazoneNo                                                     -0.2630640
## pioglitazoneSteady                                                 -0.2902726
## pioglitazoneUp                                                     -0.4992847
## insulinNo                                                          -0.2246696
## insulinSteady                                                      -0.2191127
## insulinUp                                                          -0.1028454
## changeNo                                                            0.0837169
## diabetesMedYes                                                      0.1674661
## disch_disp_modifiedDischarged to home with Home Health Service      0.1551814
## disch_disp_modifiedDischarged/Transferred to SNF                    0.3602606
## disch_disp_modifiedOther                                            0.4069118
## adm_src_modOther                                                   -0.0293146
## adm_src_modPhysician Referral                                       0.0409663
## adm_src_modTransfer from Home Health                               -0.1521517
## adm_typ_modEmergency                                                0.0582339
## adm_typ_modOther                                                   -0.0035066
## adm_typ_modUrgent                                                   0.0416309
## age_mod20-59                                                        0.5789864
## age_mod60-79                                                        0.7027100
## age_mod80+                                                          0.6423557
## diag1_mod250.8                                                     -0.4959974
## diag1_mod276                                                       -0.1950224
## diag1_mod38                                                        -0.6195428
## diag1_mod410                                                       -0.3999952
## diag1_mod414                                                       -0.4144789
## diag1_mod427                                                       -0.5055736
## diag1_mod428                                                       -0.2457533
## diag1_mod434                                                       -0.0001159
## diag1_mod435                                                       -0.5831841
## diag1_mod486                                                       -0.7040061
## diag1_mod491                                                       -0.3725894
## diag1_mod493                                                       -0.6939894
## diag1_mod518                                                       -0.9115549
## diag1_mod577                                                       -0.1772259
## diag1_mod584                                                       -0.3568387
## diag1_mod599                                                       -0.4847901
## diag1_mod682                                                       -0.5948915
## diag1_mod715                                                       -0.3659068
## diag1_mod780                                                       -0.5239524
## diag1_mod786                                                       -0.6922672
## diag1_mod820                                                       -0.3490068
## diag1_mod996                                                       -0.3245211
## diag1_modOther                                                     -0.3817989
##                                                                    Std. Error
## (Intercept)                                                         0.4155563
```

```
## raceAfricanAmerican                                                  0.0953968
## raceAsian                                                            0.1921037
## raceCaucasian                                                        0.0922154
## raceHispanic                                                         0.1267525
## raceOther                                                            0.1409831
## genderMale                                                           0.0247418
## genderUnknown/Invalid                                               68.0888890
## time_in_hospital                                                     0.0048799
## num_lab_procedures                                                   0.0007374
## num_procedures                                                       0.0092423
## num_medications                                                      0.0019914
## number_outpatient                                                    0.0088409
## number_emergency                                                     0.0101637
## number_inpatient                                                     0.0078407
## number_diagnoses                                                     0.0074772
## max_glu_serum>300                                                    0.1407309
## max_glu_serumNone                                                    0.1114802
## max_glu_serumNorm                                                    0.1212509
## A1Cresult>8                                                          0.0796072
## A1CresultNone                                                        0.0670816
## A1CresultNorm                                                        0.0871544
## pioglitazoneNo                                                       0.3054241
## pioglitazoneSteady                                                   0.3084446
## pioglitazoneUp                                                       0.4162937
## insulinNo                                                            0.0486323
## insulinSteady                                                        0.0440366
## insulinUp                                                            0.0470193
## changeNo                                                             0.0353994
## diabetesMedYes                                                       0.0393628
## disch_disp_modifiedDischarged to home with Home Health Service       0.0384547
## disch_disp_modifiedDischarged/Transferred to SNF                     0.0379060
## disch_disp_modifiedOther                                             0.0349811
## adm_src_modOther                                                     0.0521924
## adm_src_modPhysician Referral                                        0.0434689
## adm_src_modTransfer from Home Health                                 0.0789745
## adm_typ_modEmergency                                                 0.0504089
## adm_typ_modOther                                                     0.0661500
## adm_typ_modUrgent                                                    0.0441558
## age_mod20-59                                                         0.1988633
## age_mod60-79                                                         0.1991937
## age_mod80+                                                           0.2007264
## diag1_mod250.8                                                       0.1361249
## diag1_mod276                                                         0.1260477
## diag1_mod38                                                          0.1346723
## diag1_mod410                                                         0.1189497
## diag1_mod414                                                         0.1124085
## diag1_mod427                                                         0.1260539
## diag1_mod428                                                         0.1052634
## diag1_mod434                                                         0.1215166
## diag1_mod435                                                         0.1663194
## diag1_mod486                                                         0.1202047
## diag1_mod491                                                         0.1234585
## diag1_mod493                                                         0.1652733
## diag1_mod518                                                         0.1601648
```

```
## diag1_mod577                                                    0.1472232
## diag1_mod584                                                    0.1327914
## diag1_mod599                                                    0.1358130
## diag1_mod682                                                    0.1343954
## diag1_mod715                                                    0.1321289
## diag1_mod780                                                    0.1332946
## diag1_mod786                                                    0.1221289
## diag1_mod820                                                    0.1465613
## diag1_mod996                                                    0.1266154
## diag1_modOther                                                  0.0976994
##                                                                  z value
## (Intercept)                                                     -7.075
## raceAfricanAmerican                                              1.815
## raceAsian                                                        0.257
## raceCaucasian                                                    1.765
## raceHispanic                                                     1.345
## raceOther                                                        0.312
## genderMale                                                       1.053
## genderUnknown/Invalid                                           -0.118
## time_in_hospital                                                 2.527
## num_lab_procedures                                               0.524
## num_procedures                                                  -1.614
## num_medications                                                  1.989
## number_outpatient                                               -0.022
## number_emergency                                                 3.440
## number_inpatient                                                32.225
## number_diagnoses                                                 5.540
## max_glu_serum>300                                               -0.941
## max_glu_serumNone                                               -1.839
## max_glu_serumNorm                                               -0.198
## A1Cresult>8                                                      0.012
## A1CresultNone                                                    1.379
## A1CresultNorm                                                   -0.060
## pioglitazoneNo                                                  -0.861
## pioglitazoneSteady                                              -0.941
## pioglitazoneUp                                                  -1.199
## insulinNo                                                       -4.620
## insulinSteady                                                   -4.976
## insulinUp                                                       -2.187
## changeNo                                                         2.365
## diabetesMedYes                                                   4.254
## disch_disp_modifiedDischarged to home with Home Health Service   4.035
## disch_disp_modifiedDischarged/Transferred to SNF                 9.504
## disch_disp_modifiedOther                                        11.632
## adm_src_modOther                                                -0.562
## adm_src_modPhysician Referral                                    0.942
## adm_src_modTransfer from Home Health                            -1.927
## adm_typ_modEmergency                                             1.155
## adm_typ_modOther                                                -0.053
## adm_typ_modUrgent                                                0.943
## age_mod20-59                                                     2.911
## age_mod60-79                                                     3.528
## age_mod80+                                                       3.200
## diag1_mod250.8                                                  -3.644
```

```
## diag1_mod276                                                          -1.547
## diag1_mod38                                                           -4.600
## diag1_mod410                                                          -3.363
## diag1_mod414                                                          -3.687
## diag1_mod427                                                          -4.011
## diag1_mod428                                                          -2.335
## diag1_mod434                                                          -0.001
## diag1_mod435                                                          -3.506
## diag1_mod486                                                          -5.857
## diag1_mod491                                                          -3.018
## diag1_mod493                                                          -4.199
## diag1_mod518                                                          -5.691
## diag1_mod577                                                          -1.204
## diag1_mod584                                                          -2.687
## diag1_mod599                                                          -3.570
## diag1_mod682                                                          -4.426
## diag1_mod715                                                          -2.769
## diag1_mod780                                                          -3.931
## diag1_mod786                                                          -5.668
## diag1_mod820                                                          -2.381
## diag1_mod996                                                          -2.563
## diag1_modOther                                                        -3.908
##                                                                     Pr(>|z|)
## (Intercept)                                                         1.50e-12
## raceAfricanAmerican                                                 0.069593
## raceAsian                                                           0.797285
## raceCaucasian                                                       0.077648
## raceHispanic                                                        0.178691
## raceOther                                                           0.754702
## genderMale                                                          0.292396
## genderUnknown/Invalid                                               0.905936
## time_in_hospital                                                    0.011491
## num_lab_procedures                                                  0.600097
## num_procedures                                                      0.106570
## num_medications                                                     0.046710
## number_outpatient                                                   0.982235
## number_emergency                                                    0.000582
## number_inpatient                                                     < 2e-16
## number_diagnoses                                                    3.02e-08
## max_glu_serum>300                                                   0.346477
## max_glu_serumNone                                                   0.065879
## max_glu_serumNorm                                                   0.843413
## A1Cresult>8                                                         0.990542
## A1CresultNone                                                       0.167925
## A1CresultNorm                                                       0.951869
## pioglitazoneNo                                                      0.389069
## pioglitazoneSteady                                                  0.346661
## pioglitazoneUp                                                      0.230389
## insulinNo                                                           3.84e-06
## insulinSteady                                                       6.50e-07
## insulinUp                                                           0.028720
## changeNo                                                            0.018034
## diabetesMedYes                                                      2.10e-05
## disch_disp_modifiedDischarged to home with Home Health Service      5.45e-05
```

```
## disch_disp_modifiedDischarged/Transferred to SNF                < 2e-16
## disch_disp_modifiedOther                                         < 2e-16
## adm_src_modOther                                                0.574344
## adm_src_modPhysician Referral                                   0.345974
## adm_src_modTransfer from Home Health                            0.054030
## adm_typ_modEmergency                                            0.247996
## adm_typ_modOther                                                0.957724
## adm_typ_modUrgent                                               0.345773
## age_mod20-59                                                    0.003597
## age_mod60-79                                                    0.000419
## age_mod80+                                                      0.001374
## diag1_mod250.8                                                  0.000269
## diag1_mod276                                                    0.121812
## diag1_mod38                                                     4.22e-06
## diag1_mod410                                                    0.000772
## diag1_mod414                                                    0.000227
## diag1_mod427                                                    6.05e-05
## diag1_mod428                                                    0.019562
## diag1_mod434                                                    0.999239
## diag1_mod435                                                    0.000454
## diag1_mod486                                                    4.72e-09
## diag1_mod491                                                    0.002545
## diag1_mod493                                                    2.68e-05
## diag1_mod518                                                    1.26e-08
## diag1_mod577                                                    0.228671
## diag1_mod584                                                    0.007205
## diag1_mod599                                                    0.000358
## diag1_mod682                                                    9.58e-06
## diag1_mod715                                                    0.005617
## diag1_mod780                                                    8.47e-05
## diag1_mod786                                                    1.44e-08
## diag1_mod820                                                    0.017252
## diag1_mod996                                                    0.010376
## diag1_modOther                                                  9.31e-05
##
## (Intercept)                                                     ***
## raceAfricanAmerican                                             .
## raceAsian
## raceCaucasian                                                   .
## raceHispanic
## raceOther
## genderMale
## genderUnknown/Invalid
## time_in_hospital                                                *
## num_lab_procedures
## num_procedures
## num_medications                                                 *
## number_outpatient
## number_emergency                                                ***
## number_inpatient                                                ***
## number_diagnoses                                                ***
## max_glu_serum>300
## max_glu_serumNone                                               .
## max_glu_serumNorm
```

```
## A1Cresult>8
## A1CresultNone
## A1CresultNorm
## pioglitazoneNo
## pioglitazoneSteady
## pioglitazoneUp
## insulinNo                                                      ***
## insulinSteady                                                  ***
## insulinUp                                                      *
## changeNo                                                       *
## diabetesMedYes                                                 ***
## disch_disp_modifiedDischarged to home with Home Health Service ***
## disch_disp_modifiedDischarged/Transferred to SNF               ***
## disch_disp_modifiedOther                                       ***
## adm_src_modOther
## adm_src_modPhysician Referral
## adm_src_modTransfer from Home Health                           .
## adm_typ_modEmergency
## adm_typ_modOther
## adm_typ_modUrgent
## age_mod20-59                                                   **
## age_mod60-79                                                   ***
## age_mod80+                                                     **
## diag1_mod250.8                                                 ***
## diag1_mod276
## diag1_mod38                                                    ***
## diag1_mod410                                                   ***
## diag1_mod414                                                   ***
## diag1_mod427                                                   ***
## diag1_mod428                                                   *
## diag1_mod434
## diag1_mod435                                                   ***
## diag1_mod486                                                   ***
## diag1_mod491                                                   **
## diag1_mod493                                                   ***
## diag1_mod518                                                   ***
## diag1_mod577
## diag1_mod584                                                   **
## diag1_mod599                                                   ***
## diag1_mod682                                                   ***
## diag1_mod715                                                   **
## diag1_mod780                                                   ***
## diag1_mod786                                                   ***
## diag1_mod820                                                   *
## diag1_mod996                                                   *
## diag1_modOther                                                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49844  on 71236  degrees of freedom
## Residual deviance: 47759  on 71172  degrees of freedom
## AIC: 47889
```

```
##
## Number of Fisher Scoring iterations: 9
```

Evaluating Models on Testing Data

The Random forest package uses bagging, so the plot fucntion gives a good idea of how the model performs on unseen data. However, for logistic regression, we have to make sure that the model performs well on testing data.

```
library(caret)
predictions <- predict(fit2, testing,  type="response")
predictions <- ifelse(predictions>0.5, 1, 0)
confusionMatrix(testing$readmitted, predictions)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 27073    49
##          1  3360    47
##
##                Accuracy : 0.8883
##                  95% CI : (0.8847, 0.8918)
##     No Information Rate : 0.9969
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0208
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8896
##             Specificity : 0.4896
##          Pos Pred Value : 0.9982
##          Neg Pred Value : 0.0138
##              Prevalence : 0.9969
##          Detection Rate : 0.8868
##    Detection Prevalence : 0.8884
##       Balanced Accuracy : 0.6896
##
##        'Positive' Class : 0
##
```