Diabetes Related U.S. Hospital Readmission Prediction: A Generalized Model Revisited

Brian Kang *[1,2]

[1]Department of Economics, University of Washington
[2]Departments of Applied Mathematics, Computer Science and Engineering,
Mathematics, and Statistics, University of Washington

June 8, 2019

[*] Group project with Matt Kelly and Tatsuya Okuda for ECON 484: Machine Learning and Data Science for Econometricians

Contents

1 Introduction

Hyperglycemia, or high blood sugar levels, can lead to life threatening outcomes for both type 1 and type 2 diabetes. The lack of insulin, the component that breaks down carbohydrates into energy, burdens the human body with overwhelming levels of sugar remaining in the bloodstream and eventually shuts down normal bodily functions. There is currently little to no cure for this extremely common and slowly spreading condition other than constant care, maintenance, and insulin injections. Awareness for diabetes have increased dramatically the past century and a traditional protocol to treat patients in U.S. hospitals' intensive care unit (ICU) have been established. However, these protocols' necessity and effectiveness are questioned when it comes to non-emergency inpatient admissions. Often times, these patients are often not treated or extremely variable management strategies are administered, possibly leading to additional ailments (Strack et al. 2014b). I believe that current admission protocols should be implemented to all diabetes related hospital admissions because of it safety and accessibility.

In this paper, we will inspect in depth upon this phenomenon from a historical perspective. Patterns of diabetes care and medication prescribed, along with each patient's tendency to be readmitted to a U.S. hospital may inform of certain characteristics of the hospital protocols more effective and safe to be continuously administered. Through the use of statistical methodology and modeling, a large dataset will be researched to extract and explain potential implications in medical treatment for diabetic patients at U.S. hospitals.

The remainder of the paper is organized as follows. Section two will involve introducing the precedent academic literature that has already been written on our related topic of interest using the same dataset. Section three will present our data that was analyzed and the statistical models that were implemented to dive deeper into the information provided. Section four demonstrates our strategy; how we go about answering the relationship between hospital care and medication and patient's chance to be readmitted, hence indirectly evaluate whether the medical protocol serves effectively. An alternative method, the multinomial logistic regression, is implemented to answer the same question but more specifically. Section five describes the conclusion we reach through our data and models through comparison.

2 Literature Review

The precedent of this writing analyzes the same historical dataset including information of diabetic inpatient admissions to U.S. hospitals to "examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety." In particular the authors indicated the HbA1c test results as an indicator of attention to "diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus." The HbA1c tests hemoglobin (a blood pigment that carries oxygen in our blood) that bounds to glucose through a glycation process when exposed to sugar, which accurately reflects of how well diabetes is being controlled. Then, through a multivariable logistic regression, the relationship between the test results and patient readmission was studied to conclude that they response variable and primary feature in interest are endogenously dependent on the type of diagnosis of the patient (Strack et al. 2014b).

Our primary interest is to study the potential relationship between patient readmission and other factors of the overall hospital protocol such as demographics, treatment types, and medications that may explain which combination of treatments lead to the least readmissions, assuming that patients are not readmitted due to proper maintenance and sugar level control through prescriptions.

3 Data

The empirical data we found is collected from the UCI Machine Learning Repository (Dua and Graff 2017). The database contains data collected from 130 U.S. hospitals across a 10 year period (1999 - 2008). The original dataset included patient encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics. The refined data set includes 50 features and 101,766 observations describing diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information. (Strack et al. 2014a).

3.1 Statistical Methods

Since we are interested in factors that lead to no readmission, we will analyze the data in terms of a redefined binary variable of the ”readmitted” variable (is or is not readmitted) instead of the factors: not readmitted, look less than 30 days to be readmitted, and took longer than 30 days to next readmission. Feature variables of our models will initially consist of majority of the variables available in the dataset, including demographics that may affect treatment, measurements of diagnosis, and changes in medications. (Special note: 97% of the weight data is missing from the dataset, but this variable turns out to be insignificant to our sample (Strack et al. 2014b).)

The initial process of preparing our dataset for analysis involved redefining variables, assigning variables to correct values and data types, and deleting certain values. Observing the summary of all variables within the dataset, we were able to select a number of variables that particularly had irregular data values, a heavy skew, or an abnormal spread and center:

number of lab procedures, number of medication, number of outpatients, number of emergency visits, and the number of inpatients. These data were cleaned using the following manner:

- All of these selected variables had outliers exceeding the maximum of their five number summary. With respect to the minimum of the outliers, severe outliers were selected and counted.
- By observing each variables' boxplot and histogram, the distribution was inspected. Also observing the above tail with outliers, the severity and worth of the outliers were evaluated.
- Extreme outliers were deleted. For less severe ones, selected outliers were deleted so that we would not lose too much data but also they would not greatly disturb the remaining data and future modeling.
- Entire rows of data for each outlier were deleted, not assigned NA or null values. This is because modifying individual observations would be considered as if they were missing data points since the beginning, which is false. Rows as a whole were taken out because if they were assigned NA or null values, they will cause errors throughout the modeling, prediction, and analysis process. Also, this method ensures that "remaining information" of the outliers are not included throughout the process.

Overall, 7825 rows were deleted, retaining 93941 observations while eliminating troublesome information. The graphs observed can be found in Appendix 6.1 figure 1 and Appendix 6.2 figure 2a.

One of the selected variables, number of inpatients, seemed like it could fit either an exponential or log-normal distribution. If that was the case, I would have been able to standardize the values in terms of the fitted distribution and better explain the effect of this variable within future models. However, based on the outputs from fitting both distributions, no one distribution actually fit well to the variable. I believe that this is because the data values were discrete, if they were continuous data types, one of the two distributions may have fit. The referenced plots can be found in Appendix 6.2 figure 2. R code to generate the plots and clean data can be found in the R Script attached in Appendix 6.10.

4 Strategy: Fitted Models

In this paper we want to predict the likelihood of readmission using historical data of the components of U.S. hospital protocols for treating diabetes. That is, we plan to explain why parts of the actual treatment and change in medication are effective or not to lead diabetic patients to be either not readmitted or readmitted to hospitals. Ultimately, we aim to illustrate that certain characteristics of the protocol or maybe even some combinations are directly related to the prevention of hospital readmission, hence implying proper maintenance or suppression of diabetes caused symptoms.

To study this, we used various statistical methods and machine learning techniques to train our models and predict a potential patient's readmission. Methods include the LASSO, logistic regression, boosting, and random forest. The LASSO has been separated in two parts: post-LASSO and OLS regression and LASSO with the tuning parameter chosen by doing a K-fold cross validation. In addition to estimating if one will be readmitted or not, we implemented a multinomial logistic neural net with and without variable selection to predict from the original responses: not readmitted, look less than 30 days to be readmitted, and took longer than 30 days to next readmission.

To elaborate more in detail, we will compare all models from one another in terms of the flexibility-interpretability trade off and the bias-variance trade off. Generally, with higher flexibility comes lower errors and worse interpretability, so fitting a support vector machine or random forest will return outstanding predictions, but it is near impossible to know what those numeric values mean. Also, with high flexibility comes with lower bias but higher variance and overall MSE, the error value of a model ($MSE = Bias^2 + Variance$). In addition to these trade-offs, generally we see that the model's accuracy for predicting sum its MSE equals to 1 ($MSE + Accuracy = 1$). These characteristics are demonstrated in our models as well. Therefore, by choosing one specific model over all others, we are weighing our values between flexibility or interpretability, accuracy or error, and machine run time efficiency. All R code for creating the models, predicting, and analyzing can be found in Appendix 6.10.

4.1 Solution: LASSO

I decided to choose the LASSO with default settings to be the best predictor out of all models. As a shrinkage model, all coefficients will be biased towards zero; however, the perk of this model is that with high probability it will select the features that are significant for our prediction. This property is especially useful for our problem. Initially we have hundreds of variables we can possibly insert into our model but we do not know how to choose those that we really need. The LASSO selects them for us. With the variables that are not pushed to zero, we can easily run an OLS regression to get our predictions. By its characteristic, the LASSO is considered not very flexible but very interpretable. We may think this lack of flexibility may return very high errors and low accuracy, but for our model we see observed a MSE value of about 0.22, implying an accuracy rate of 78% which is great with respect to the lack of flexibility, more inflexible than a simple linear regression. Overall, we have the best prediction out of all the models we fitted despite having the lowest flexibility. The LASSO also is the most interpretable. Using the returned model we got we can conclude that the chance of not being readmitted is simply the sum of the products of factor variables and their coefficients and the respective products of numeric variables and their coefficients, as what the post-LASSO OLS returns. However, a bane from using this model is that we lose a large portion of our data while predicting using the test set, about half of the test observations was lost, a proportion incomparably larger than all other models implemented. I believe that this comes from the inflexibility of the LASSO.

The one reason I chose the LASSO over the LASSO with the tuning parameter chosen by cross validation is the increase in MSE (figure 3). Although mathematically speaking, the CV should return the minimal turning parameter and so maximize the accuracy, the default setting for the LASSO preformed better. I predict that this is due to either round off error by separately assigning the regularization parameter or due to LASSO's tendency to work well for lower dimensional models. The default setting may have picked up better variables or dropped more unnecessary variables to ultimately return a MSE lower that its counterpart combined with cross validation. The output from the LASSO and LASSO with CV can be found in Appendices 6.3 and 6.4.

4.2 Alternative: Multinomial Logistic Neural Network

I would like to briefly discuss the results from the multinomial logistic neural net as well because it serves the save goal but in a more accurate method. The multinomial logistic regression is simply a logistic regression but we estimate more than two response variables, unlike the logit that predicts binary values. The neural net is a model that takes advantage of its flexibility to return great predictions, but the final result is very difficult to interpret. From our output (Appendix 6.6) we see that the accuracy rate is about 57% with MSE 43%, summing to 100%. Note that the p-value is extremely low (4.1e-14). P-values for every individual variable is listed in figure 4. We can see that many of these variables are individually insignificant, totaling 75 features within our model. The advantage of using a neural net is the capability of drastically reducing this count while maintaining a reasonable accuracy. After extracting only the variables with very high significance and reasonably importance (just like how we calculate them for our random forest in figure 6a), we reduce the number of features to 25. The MSE only increased by 0.1%. The statistics for this model can be found in Appendix 6.7 and figure 5. The significance of this model is the prediction of patient readmission not in a binary sense but in more specific factors.

5 Result

We now reach the verdict for whether certain characteristics of the U.S. hospital diabetes treatment protocol illustrates a relationship to the prevention of readmission. According to the LASSO and improved neural net, demographic ages are significant along with the expected variables like number of diagnoses, outpatient, emergency, but not really the time in hospital. Mostly steady dosage of medication shows greater significance over increasing the dosage for a few medications. There was almost no decreases in dosage that were significant. Both models strongly agreed that the number of inpatients were strongly significant. According to the LASSO, several medical specialties of the admitting physician shows some effect and a couple discharge dispositions showed great significance to prediction, especially expired and hospice/home.

From these results I conclude that there are some clear indicators that are correlated to the prevention of admission, implying good maintenance or treatment during the hospital visit. However, we cannot say that these components of the treatment protocol is necessarily causal to readmission or not. For example, take the three most significant features from our models: the number of inpatient visits, discharge dispositions, and steady dosage of medication. The number of inpatient visit is the number of overnight stays at a hospital which means significantly prolonged monitoring and care from the hospital. Because blood sugar levels change even overnight, the constant care is directly related to better care, logically speaking. Then, the discharge dispositions may not tell us much information. Expired decisions may imply a wide variety of decisions; it is too vague. Discharge to home is what we could assume to be the primal decision by the doctors after treatment and prescribing medication, this may lead the patient to self-care in the longer term, but that is not necessarily true. This leads us to the steady dosages. Logically speaking, a steady dosage of the regular medication for patients may imply less abrupt fluctuation in blood sugar levels and so describe the patient's responsible habit of good self-care, which a large proportion of diabetes sufferers practice.

In conclusion, our study may be improved through looking deeper into other models we have put down because the LASSO gave us the best accuracy despite its inflexibility. It also is the easiest to interpret out of all models we implemented, including the logit. Other statistical methods have their own boons and banes outside of considering the flexibility-interpretability and bias-variance tradeoff. For example, the logit is quicker than both LASSO and neural net in terms of machine run time efficiency. If we did a tree, we would have superior visual information of how the model behaves. If we succeeded in implementing a support vector machine, its high flexibility may have displayed an accuracy higher than the LASSO's. Not only in terms of trying other models, but also in terms of the formation of our model formula, if could have narrowed our sight of view for our variables or adjust/standardize some data values our final result may have been a step closer to discovering a relationship closer to causality instead of correlation.

References

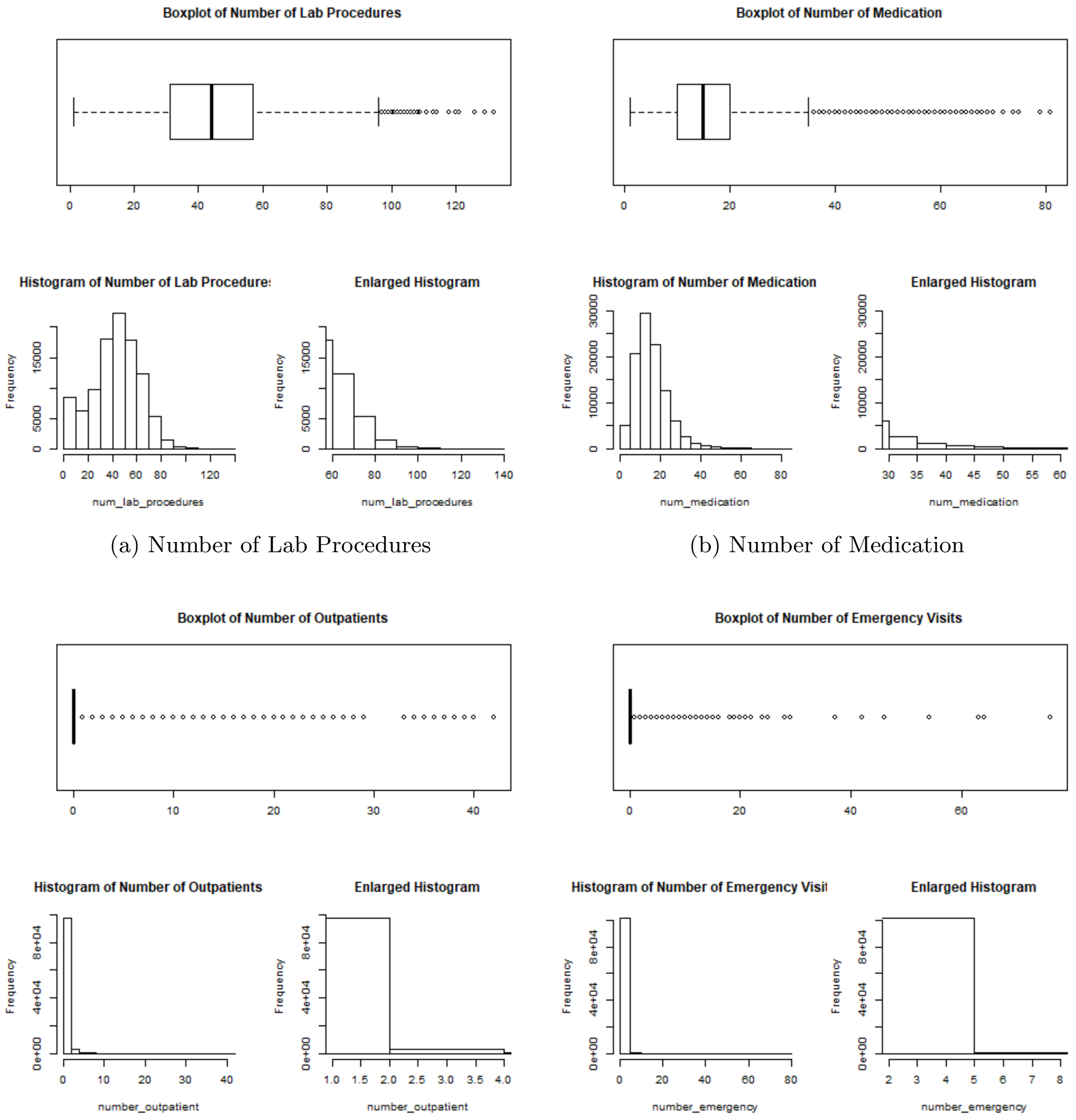
Dua, Dheeru, and Casey Graff. 2017. *UCI Machine Learning Repository.* University of California, Irvine, School of Information and Computer Sciences. `http://archive.ics.uci.edu/ml`.

Strack, Beata, et al. 2014a. "List of features and their descriptions in the initial dataset". `https://www.hindawi.com/journals/bmri/2014/781670/`.

— . 2014b. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records". *BioMed Research International* 2014 (): 11. doi:`http://dx.doi.org/10.1155/2014/781670`.

6 Appendices

6.1 Original Data of Selected Variables



(a) Number of Lab Procedures

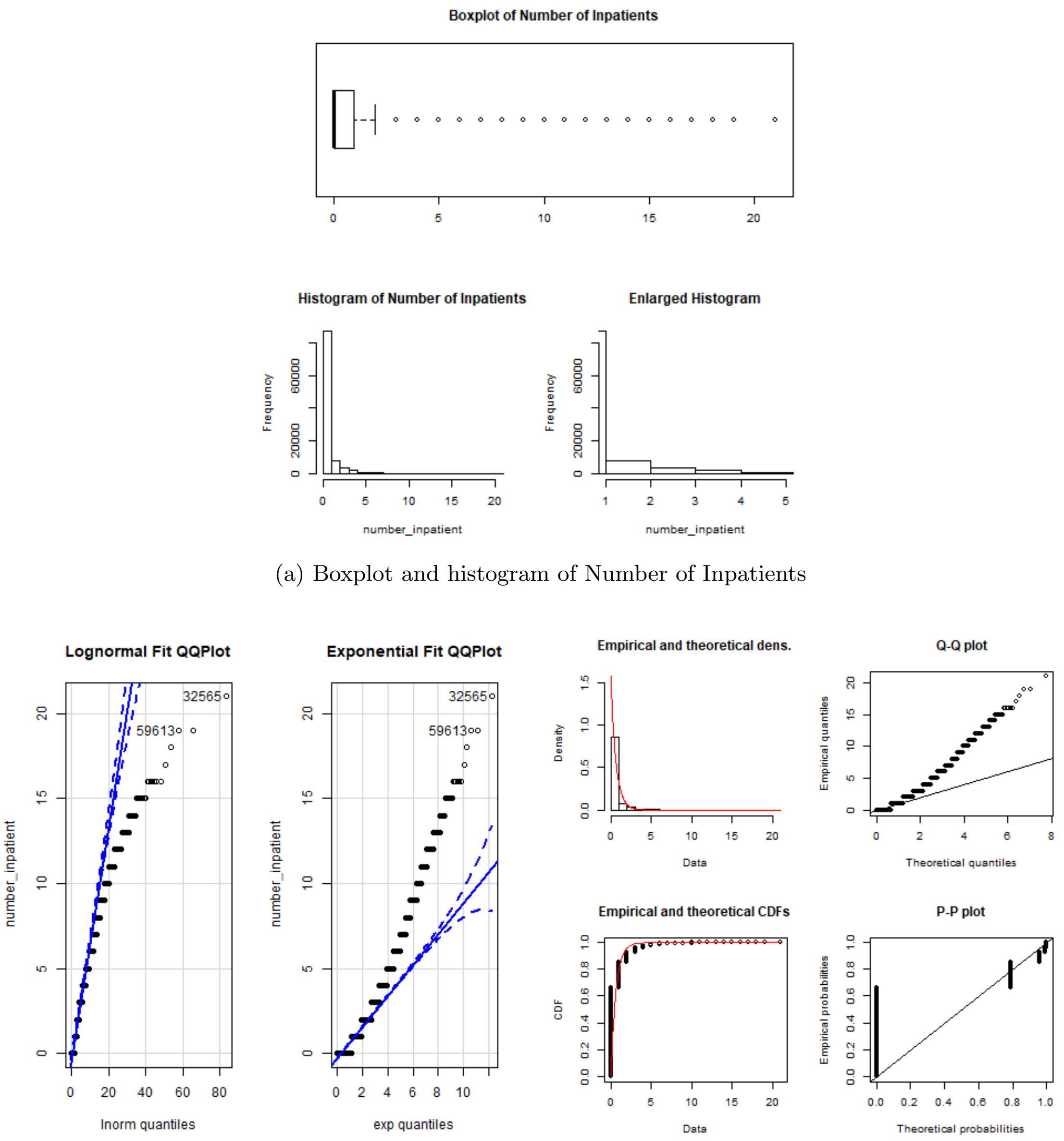
(b) Number of Medication

(c) Number of Outpatients

(d) Number of Emergency Visits

Figure 1: Boxplots and histograms of selected variables that needed cleaning.

6.2 Cleaning and Fitting Distribution to a Variable



(a) Boxplot and histogram of Number of Inpatients

(b) QQPlot of 2 Fitted Distributions

(c) Summary of Fitting Exponential Fit

Figure 2: Number of inpatients does not follow any clear distribution.

6.3 LASSO Output

```
Do LASSO on training set

Call:
rlasso.formula(formula = formula, data = data, post = post, intercept = intercept,
    model = model, control = control)

Post-Lasso Estimation:  FALSE

Total number of variables: 168
Number of selected variables: 50

Residuals:
    Min      1Q  Median      3Q     Max
-0.9572 -0.5257  0.2722  0.4078  0.9841

                                                 Estimate
(Intercept)                                         0.787
raceAsian                                           0.007
raceOther                                           0.013
age30_40                                            0.004
age50_60                                            0.002
age70_80                                           -0.017
age80_90                                           -0.005
age90_100                                           0.004
admission_type_id2                                 -0.003
discharge_disposition_id6                          -0.006
discharge_disposition_id11                          0.469
discharge_disposition_id13                          0.356
discharge_disposition_id14                          0.225
discharge_disposition_id19                          0.273
discharge_disposition_id22                         -0.008
discharge_disposition_id23                          0.066
admission_source_id4                                0.076
admission_source_id5                                0.020
admission_source_id6                                0.101
admission_source_id7                               -0.003
time_in_hospital                                   -0.003
medical_specialtyEmergency_Trauma                  -0.028
medical_specialtyFamily_GeneralPractice            -0.015
medical_specialtyGastroenterology                  -0.018
medical_specialtyGynecology                         0.083
medical_specialtyInternalMedicine                   0.008
medical_specialtyNephrology                        -0.052
medical_specialtyNeurology                          0.036
medical_specialtyObstetricsandGynecology            0.130
medical_specialtyOncology                          -0.005
```

```
medical_specialtyOrthopedics                        0.035
medical_specialtyOrthopedics_Reconstructive         0.045
medical_specialtyPediatrics_Endocrinology           0.074
medical_specialtyPediatrics_Pulmonology            -0.037
medical_specialtyPulmonology                       -0.007
medical_specialtySurgeon                            0.021
medical_specialtySurgery_Cardiovascular_Thoracic    0.088
medical_specialtySurgery_Neuro                      0.076
number_outpatient                                  -0.043
number_emergency                                   -0.040
number_inpatient                                   -0.090
number_diagnoses                                   -0.019
A1CresultNone                                      -0.013
metforminNo                                        -0.015
repaglinideNo                                       0.012
glipizideNo                                         0.002
acarboseNo                                          0.026
tolazamideSteady                                    0.247
insulinSteady                                       0.010
changeNo                                            0.015
diabetesMedYes                                     -0.058

Residual standard error: 0.4714
Multiple R-squared:  0.09004
Adjusted R-squared:  0.08864
Joint significance test:
 the sup score statistic for joint significance test is 39.51 with a p-value of 0.092

Count and Kept Significant Variables by LASSO
Count: [1] 50
                                       raceAsian                                        raceOther
                                               1                                                4
                                        age30_40                                         age50_60
                                               8                                               10
                                        age70_80                                         age80_90
                                              12                                               13
                                       age90_100                               admission_type_id2
                                              14                                               15
                       discharge_disposition_id6                       discharge_disposition_id11
                                              22                                               25
                      discharge_disposition_id13                       discharge_disposition_id14
                                              27                                               28
                      discharge_disposition_id19                       discharge_disposition_id22
                                              30                                               31
                      discharge_disposition_id23                             admission_source_id4
                                              32                                               38
                            admission_source_id5                             admission_source_id6
```

```
                                              39                                               40
                            admission_source_id7                                 time_in_hospital
                                              41                                               47
               medical_specialtyEmergency_Trauma          medical_specialtyFamily_GeneralPractice
                                              54                                               57
               medical_specialtyGastroenterology                      medical_specialtyGynecology
                                              58                                               59
               medical_specialtyInternalMedicine                      medical_specialtyNephrology
                                              64                                               65
                      medical_specialtyNeurology         medical_specialtyObstetricsandGynecology
                                              66                                               70
                       medical_specialtyOncology                     medical_specialtyOrthopedics
                                              71                                               73
     medical_specialtyOrthopedics_Reconstructive        medical_specialtyPediatrics_Endocrinology
                                              74                                               83
         medical_specialtyPediatrics_Pulmonology                     medical_specialtyPulmonology
                                              87                                               96
                        medical_specialtySurgeon medical_specialtySurgery_Cardiovascular_Thoracic
                                             102                                              104
                  medical_specialtySurgery_Neuro                                number_outpatient
                                             108                                              118
                                number_emergency                                 number_inpatient
                                             119                                              120
                                number_diagnoses                                    A1CresultNone
                                             121                                              124
                                     metforminNo                                    repaglinideNo
                                             126                                              129
                                     glipizideNo                                       acarboseNo
                                             141                                              153
                                tolazamideSteady                                    insulinSteady
                                             158                                              160
                                        changeNo                                   diabetesMedYes
                                             167                                              168
```

Do OLS on training set using selected variables from LASSO

```
Call:
lm(formula = formula, data = diabetic[train, ])

Residuals:
    Min      1Q  Median      3Q     Max
-1.0435 -0.5086  0.2420  0.4063  1.0648

Coefficients:
                                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           0.6972400  0.0551740  12.637  < 2e-16 ***
raceAsianTRUE                                         0.0661407  0.0273830   2.415 0.015724 *
```

```
raceOtherTRUE                                         0.0584128 0.0203091   2.876 0.004028 **
age30_40TRUE                                          0.0242379 0.0140346   1.727 0.084176 .
age50_60TRUE                                          0.0119649 0.0076953   1.555 0.119995
age70_80TRUE                                         -0.0247413 0.0069626  -3.553 0.000381 ***
age80_90TRUE                                         -0.0132913 0.0081256  -1.636 0.101903
age90_100TRUE                                         0.0366084 0.0169092   2.165 0.030395 *
admission_type_id2TRUE                               -0.0092038 0.0064618  -1.424 0.154357
discharge_disposition_id6TRUE                        -0.0197639 0.0091031  -2.171 0.029930 *
discharge_disposition_id11TRUE                        0.4873970 0.0204468  23.837  < 2e-16 ***
discharge_disposition_id13TRUE                        0.4220878 0.0441462   9.561  < 2e-16 ***
discharge_disposition_id14TRUE                        0.3470026 0.0719633   4.822 1.43e-06 ***
discharge_disposition_id19TRUE                        0.4798802 0.1780774   2.695 0.007047 **
discharge_disposition_id22TRUE                       -0.0468241 0.0173070  -2.706 0.006824 **
discharge_disposition_id23TRUE                        0.1528650 0.0397149   3.849 0.000119 ***
admission_source_id4TRUE                              0.1110760 0.0162737   6.826 8.92e-12 ***
admission_source_id5TRUE                              0.0763580 0.0256965   2.972 0.002965 **
admission_source_id6TRUE                              0.1226766 0.0151555   8.095 5.95e-16 ***
admission_source_id7TRUE                             -0.0031561 0.0068758  -0.459 0.646220
time_in_hospital                                     -0.0042121 0.0009403  -4.479 7.52e-06 ***
medical_specialtyEmergency_TraumaTRUE                -0.0357223 0.0097802  -3.653 0.000260 ***
medical_specialtyFamily_GeneralPracticeTRUE          -0.0314870 0.0093625  -3.363 0.000772 ***
medical_specialtyGastroenterologyTRUE                -0.0681600 0.0246459  -2.766 0.005685 **
medical_specialtyGynecologyTRUE                       0.1883527 0.0697875   2.699 0.006960 **
medical_specialtyInternalMedicineTRUE                 0.0117617 0.0076385   1.540 0.123622
medical_specialtyNephrologyTRUE                      -0.0823445 0.0159513  -5.162 2.45e-07 ***
medical_specialtyNeurologyTRUE                        0.1084242 0.0384204   2.822 0.004775 **
medical_specialtyObstetricsandGynecologyTRUE          0.1574962 0.0235543   6.687 2.32e-11 ***
medical_specialtyOncologyTRUE                        -0.0609773 0.0312031  -1.954 0.050685 .
medical_specialtyOrthopedicsTRUE                      0.0724221 0.0172129   4.207 2.59e-05 ***
medical_specialtyOrthopedics_ReconstructiveTRUE       0.0867123 0.0203620   4.259 2.06e-05 ***
medical_specialtyPediatrics_EndocrinologyTRUE         0.1250811 0.0453232   2.760 0.005788 **
medical_specialtyPediatrics_PulmonologyTRUE          -0.2519722 0.1059061  -2.379 0.017356 *
medical_specialtyPulmonologyTRUE                     -0.0504719 0.0217686  -2.319 0.020425 *
medical_specialtySurgeonTRUE                          0.1953752 0.0861878   2.267 0.023406 *
medical_specialtySurgery_Cardiovascular_ThoracicTRUE  0.1331930 0.0260901   5.105 3.32e-07 ***
medical_specialtySurgery_NeuroTRUE                    0.1231293 0.0276279   4.457 8.35e-06 ***
number_outpatient                                    -0.0522553 0.0071158  -7.344 2.13e-13 ***
number_emergency                                     -0.0476633 0.0064913  -7.343 2.14e-13 ***
number_inpatient                                     -0.0907880 0.0031278 -29.027  < 2e-16 ***
number_diagnoses                                     -0.0173267 0.0014587 -11.878  < 2e-16 ***
A1CresultNoneTRUE                                    -0.0281822 0.0069538  -4.053 5.07e-05 ***
metforminNoTRUE                                      -0.0335796 0.0071225  -4.715 2.43e-06 ***
repaglinideNoTRUE                                     0.0339130 0.0176711   1.919 0.054978 .
glipizideNoTRUE                                       0.0137484 0.0080169   1.715 0.086367 .
acarboseNoTRUE                                        0.1161125 0.0492862   2.356 0.018485 *
tolazamideSteadyTRUE                                  0.3972254 0.1780229   2.231 0.025667 *
insulinSteadyTRUE                                     0.0243666 0.0062552   3.895 9.82e-05 ***
```

```
changeNoTRUE                                          0.0185388  0.0064958   2.854 0.004320 **
diabetesMedYesTRUE                                   -0.0741878  0.0082026  -9.044  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4707 on 32505 degrees of freedom
  (42596 observations deleted due to missingness)
Multiple R-squared:  0.0943, Adjusted R-squared:  0.09291
F-statistic: 67.69 on 50 and 32505 DF,  p-value: < 2.2e-16

Predict on test set
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
 -0.063   0.495   0.583   0.577   0.664   1.277   10557

Count remaining observations
[1] 8232

MSE
[1] 0.2218215
```

6.4 LASSO with CV Output

```
    CV on LASSO and get min tuning parameter
$glmnet.fit

Call:  glmnet(x = xtrain, y = ytrain, alpha = 1)

        Df      %Dev    Lambda
  [1,]   0 0.000e+00 6.811e-03
  [2,]   3 5.616e-05 6.206e-03
  [3,]   4 1.435e-04 5.654e-03
  [4,]   6 2.410e-04 5.152e-03
  [5,]   8 3.726e-04 4.694e-03
  [6,]  14 5.220e-04 4.277e-03
  [7,]  23 7.647e-04 3.897e-03
  [8,]  29 1.047e-03 3.551e-03
  [9,]  42 1.366e-03 3.236e-03
 [10,]  50 1.703e-03 2.948e-03
 [11,]  55 2.030e-03 2.686e-03
 [12,]  65 2.323e-03 2.448e-03
 [13,]  74 2.613e-03 2.230e-03
 [14,]  82 2.891e-03 2.032e-03
 [15,]  87 3.139e-03 1.852e-03
 [16,]  95 3.369e-03 1.687e-03
 [17,]  98 3.569e-03 1.537e-03
 [18,] 103 3.745e-03 1.401e-03
 [19,] 104 3.894e-03 1.276e-03
 [20,] 110 4.023e-03 1.163e-03
 [21,] 113 4.135e-03 1.060e-03
 [22,] 116 4.231e-03 9.654e-04
 [23,] 121 4.314e-03 8.797e-04
 [24,] 125 4.385e-03 8.015e-04
 [25,] 129 4.448e-03 7.303e-04
 [26,] 132 4.499e-03 6.654e-04
 [27,] 135 4.551e-03 6.063e-04
 [28,] 135 4.606e-03 5.524e-04
 [29,] 137 4.659e-03 5.034e-04
 [30,] 142 4.704e-03 4.587e-04
 [31,] 144 4.741e-03 4.179e-04
 [32,] 146 4.782e-03 3.808e-04
 [33,] 148 4.820e-03 3.470e-04
 [34,] 148 4.856e-03 3.161e-04
 [35,] 149 4.893e-03 2.880e-04
 [36,] 150 4.927e-03 2.625e-04
 [37,] 151 4.955e-03 2.391e-04
 [38,] 152 4.980e-03 2.179e-04
 [39,] 153 5.000e-03 1.985e-04
 [40,] 154 5.019e-03 1.809e-04
```

```
[41,] 154 5.036e-03 1.648e-04
[42,] 154 5.048e-03 1.502e-04
[43,] 153 5.060e-03 1.368e-04
[44,] 153 5.070e-03 1.247e-04
[45,] 157 5.078e-03 1.136e-04
[46,] 157 5.086e-03 1.035e-04
[47,] 157 5.094e-03 9.432e-05
[48,] 159 5.097e-03 8.594e-05
[49,] 160 5.105e-03 7.831e-05
[50,] 161 5.108e-03 7.135e-05
[51,] 163 5.115e-03 6.501e-05
[52,] 163 5.118e-03 5.924e-05
[53,] 162 5.121e-03 5.397e-05
[54,] 162 5.123e-03 4.918e-05
[55,] 162 5.126e-03 4.481e-05
[56,] 163 5.128e-03 4.083e-05
[57,] 164 5.131e-03 3.720e-05
[58,] 165 5.133e-03 3.390e-05
[59,] 165 5.136e-03 3.089e-05
[60,] 166 5.142e-03 2.814e-05
[61,] 166 5.146e-03 2.564e-05
[62,] 165 5.150e-03 2.336e-05
[63,] 166 5.153e-03 2.129e-05
[64,] 167 5.155e-03 1.940e-05
[65,] 167 5.158e-03 1.767e-05
[66,] 168 5.159e-03 1.610e-05
[67,] 168 5.162e-03 1.467e-05
[68,] 168 5.163e-03 1.337e-05
[69,] 168 5.164e-03 1.218e-05
[70,] 168 5.166e-03 1.110e-05
[71,] 168 5.167e-03 1.011e-05
[72,] 168 5.168e-03 9.215e-06
[73,] 168 5.169e-03 8.397e-06
[74,] 168 5.170e-03 7.651e-06
[75,] 168 5.171e-03 6.971e-06
[76,] 168 5.172e-03 6.352e-06
[77,] 168 5.173e-03 5.788e-06
[78,] 168 5.174e-03 5.273e-06
[79,] 168 5.174e-03 4.805e-06
[80,] 169 5.175e-03 4.378e-06
[81,] 169 5.176e-03 3.989e-06
[82,] 169 5.177e-03 3.635e-06
[83,] 169 5.178e-03 3.312e-06
[84,] 169 5.178e-03 3.018e-06
[85,] 169 5.179e-03 2.750e-06
[86,] 169 5.180e-03 2.505e-06
[87,] 169 5.180e-03 2.283e-06
```

```
 [88,] 169 5.181e-03 2.080e-06
 [89,] 169 5.182e-03 1.895e-06
 [90,] 169 5.182e-03 1.727e-06
 [91,] 169 5.183e-03 1.573e-06
 [92,] 169 5.183e-03 1.434e-06
 [93,] 169 5.184e-03 1.306e-06
 [94,] 169 5.184e-03 1.190e-06
 [95,] 169 5.185e-03 1.084e-06
 [96,] 169 5.186e-03 9.881e-07
 [97,] 169 5.186e-03 9.004e-07
 [98,] 169 5.187e-03 8.204e-07
 [99,] 169 5.187e-03 7.475e-07
[100,] 169 5.187e-03 6.811e-07

$lambda.min
[1] 0.006810838

Do LASSO using min tuning parameter then predict
          Length Class     Mode
a0          1    -none-    numeric
beta      193    dgCMatrix S4
df          1    -none-    numeric
dim         2    -none-    numeric
lambda      1    -none-    numeric
dev.ratio   1    -none-    numeric
nulldev     1    -none-    numeric
npasses     1    -none-    numeric
jerr        1    -none-    numeric
offset      1    -none-    logical
call        5    -none-    call
nobs        1    -none-    numeric

Call:  glmnet(x = xtrain, y = ytrain, alpha = 1, lambda = cv.lambda)

     Df %Dev   Lambda
[1,]  0    0 0.006811
       1
 Min.   :0.5577
 1st Qu.:0.5577
 Median :0.5577
 Mean   :0.5577
 3rd Qu.:0.5577
 Max.   :0.5577

MSE
[1] 0.2481553
```

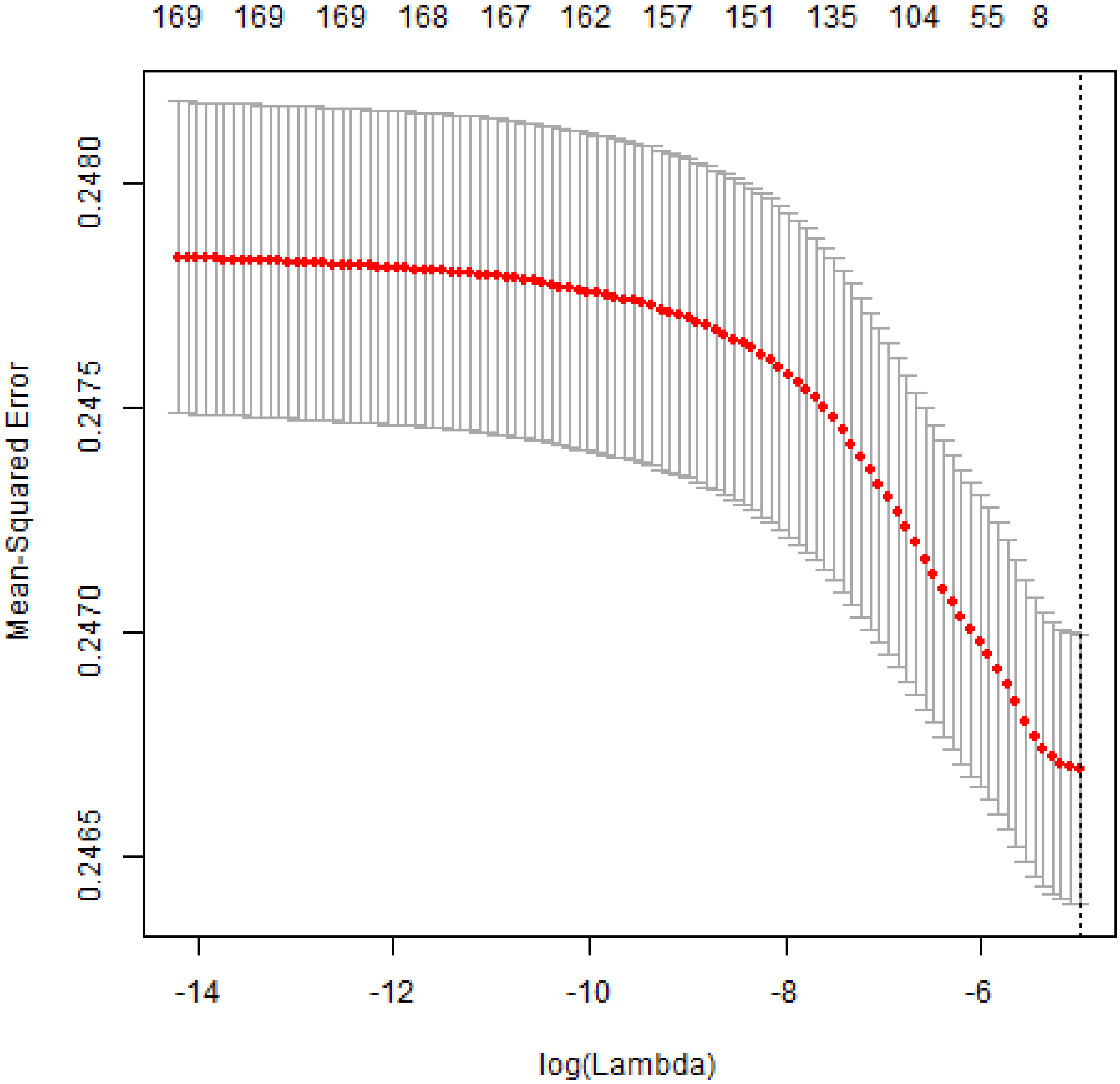


Figure 3: Plot of decreasing MSE of LASSO from using CV

6.5 Logistic Regression Output

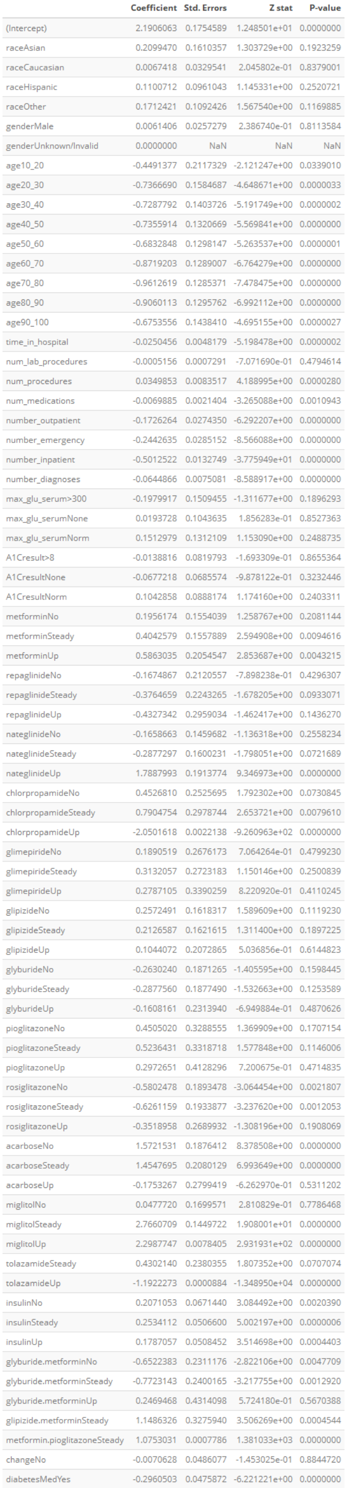
```
    Do logit on training set

Call:
glm(formula = formula, family = "binomial", data = diabetic,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0240  -1.2210   0.8441   1.0317   2.4128

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   2.871e+00  3.300e+02   0.009 0.993060
raceAsian                     3.743e-01  9.840e-02   3.804 0.000142 ***
raceCaucasian                -1.660e-02  2.001e-02  -0.830 0.406753
raceHispanic                  1.444e-01  5.715e-02   2.526 0.011536 *
raceOther                     2.261e-01  6.506e-02   3.475 0.000510 ***
genderMale                    4.417e-02  1.561e-02   2.829 0.004671 **
age10_20                     -8.267e-01  2.579e-01  -3.206 0.001348 **
age20_30                     -6.874e-01  2.495e-01  -2.755 0.005865 **
age30_40                     -7.421e-01  2.447e-01  -3.032 0.002427 **
age40_50                     -7.961e-01  2.429e-01  -3.277 0.001047 **
age50_60                     -8.114e-01  2.425e-01  -3.346 0.000819 ***
age60_70                     -8.987e-01  2.424e-01  -3.708 0.000209 ***
age70_80                     -9.362e-01  2.424e-01  -3.862 0.000112 ***
age80_90                     -8.820e-01  2.427e-01  -3.635 0.000278 ***
age90_100                    -5.559e-01  2.464e-01  -2.257 0.024029 *
time_in_hospital             -1.222e-02  3.037e-03  -4.024 5.71e-05 ***
num_lab_procedures           -1.376e-03  4.414e-04  -3.118 0.001820 **
num_procedures                3.998e-02  5.020e-03   7.965 1.66e-15 ***
num_medications              -3.172e-04  1.317e-03  -0.241 0.809687
number_outpatient            -2.392e-01  1.742e-02 -13.731  < 2e-16 ***
number_emergency             -2.685e-01  1.934e-02 -13.884  < 2e-16 ***
number_inpatient             -4.011e-01  9.499e-03 -42.231  < 2e-16 ***
number_diagnoses             -7.072e-02  4.411e-03 -16.033  < 2e-16 ***
max_glu_serum>300            -2.177e-01  9.659e-02  -2.254 0.024193 *
max_glu_serumNone             1.955e-02  6.560e-02   0.298 0.765740
max_glu_serumNorm             1.282e-01  8.107e-02   1.581 0.113882
A1Cresult>8                  -9.200e-02  4.826e-02  -1.906 0.056602 .
A1CresultNone                -5.329e-02  4.070e-02  -1.309 0.190385
A1CresultNorm                 7.489e-02  5.217e-02   1.435 0.151150
metforminNo                  -6.302e-02  1.028e-01  -0.613 0.539903
metforminSteady               7.152e-02  1.028e-01   0.696 0.486480
metforminUp                   1.306e-01  1.255e-01   1.041 0.297969
repaglinideNo                 1.153e-01  3.456e-01   0.334 0.738736
repaglinideSteady            -1.255e-01  3.513e-01  -0.357 0.720977
```

```
repaglinideUp                 2.991e-01  4.163e-01   0.718 0.472485
nateglinideNo                -1.872e-01  6.570e-01  -0.285 0.775677
nateglinideSteady            -2.521e-01  6.633e-01  -0.380 0.703906
nateglinideUp                -1.297e-01  8.229e-01  -0.158 0.874745
chlorpropamideNo             -1.093e+01  1.970e+02  -0.055 0.955760
chlorpropamideSteady         -1.102e+01  1.970e+02  -0.056 0.955383
chlorpropamideUp             -2.273e+01  2.199e+02  -0.103 0.917686
glimepirideNo                 6.359e-02  1.769e-01   0.359 0.719291
glimepirideSteady             5.575e-02  1.791e-01   0.311 0.755565
glimepirideUp                 2.423e-01  2.199e-01   1.102 0.270551
glipizideNo                   1.494e-01  1.080e-01   1.382 0.166829
glipizideSteady               4.782e-02  1.081e-01   0.443 0.658101
glipizideUp                   3.082e-02  1.363e-01   0.226 0.821118
glyburideNo                   9.738e-02  1.041e-01   0.935 0.349738
glyburideSteady               6.429e-02  1.044e-01   0.616 0.538001
glyburideUp                   1.205e-01  1.314e-01   0.917 0.359055
pioglitazoneNo                3.473e-01  2.294e-01   1.514 0.129976
pioglitazoneSteady            2.739e-01  2.306e-01   1.188 0.235001
pioglitazoneUp                4.524e-02  2.761e-01   0.164 0.869868
rosiglitazoneNo              -4.764e-01  2.850e-01  -1.671 0.094648 .
rosiglitazoneSteady          -5.898e-01  2.862e-01  -2.061 0.039306 *
rosiglitazoneUp              -2.177e-01  3.388e-01  -0.643 0.520471
acarboseNo                    1.067e+01  1.970e+02   0.054 0.956794
acarboseSteady                1.037e+01  1.970e+02   0.053 0.958028
acarboseUp                    9.287e+00  1.970e+02   0.047 0.962393
miglitolNo                    1.093e+01  1.103e+02   0.099 0.921021
miglitolSteady                1.079e+01  1.103e+02   0.098 0.922084
miglitolUp                    2.313e+01  2.257e+02   0.102 0.918381
tolazamideSteady              5.755e-01  4.221e-01   1.363 0.172735
tolazamideUp                 -1.153e+01  1.970e+02  -0.059 0.953311
insulinNo                     1.309e-01  4.116e-02   3.179 0.001477 **
insulinSteady                 2.232e-01  3.162e-02   7.058 1.69e-12 ***
insulinUp                     7.435e-02  3.240e-02   2.295 0.021755 *
glyburide.metforminNo        -1.156e+01  1.384e+02  -0.084 0.933443
glyburide.metforminSteady    -1.167e+01  1.384e+02  -0.084 0.932823
glyburide.metforminUp        -1.071e+01  1.384e+02  -0.077 0.938332
glipizide.metforminSteady    -5.477e-01  6.600e-01  -0.830 0.406605
metformin.pioglitazoneSteady  1.090e+01  1.970e+02   0.055 0.955870
changeNo                     -2.583e-02  2.959e-02  -0.873 0.382670
diabetesMedYes               -2.813e-01  2.831e-02  -9.937  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 100993  on 73409  degrees of freedom
Residual deviance:  96365  on 73336  degrees of freedom
```

```
  (1742 observations deleted due to missingness)
AIC: 96513

Number of Fisher Scoring iterations: 10



Predict using test set
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
 0.0000  0.4870  0.5755  0.5516  0.6392  1.0000     452

MSE
[1] 0.3941668

Confusion matrix

logit.pred.1 FALSE TRUE
       FALSE  3286 2218
       TRUE   5188 8097

Accuracy
[1] 0.6058332

NOTICE, sum of MSE and accuracy =
[1] 1
```

6.6 Multinomial Logistic Neural Net Output

```
    Fit multinomial logistic neural net and get number of features
    in model, probabilities, and effective DF

# weights:  228 (150 variable)
initial  value 80649.128111
iter  10 value 68791.634046
iter  20 value 68361.773829
iter  30 value 68139.598662
iter  40 value 67256.600972
iter  50 value 66691.445833
iter  60 value 66432.318248
iter  70 value 66202.732847
iter  80 value 66043.533915
iter  90 value 65995.649884
iter 100 value 65964.331781
final  value 65964.331781
stopped after 100 iterations

[1] 75
      <30                  >30                  NO
 Min.   :0.004494   Min.   :0.05597   Min.   :0.008356
 1st Qu.:0.076149   1st Qu.:0.28270   1st Qu.:0.488833
 Median :0.094028   Median :0.32938   Median :0.575306
 Mean   :0.105448   Mean   :0.34346   Mean   :0.551095
 3rd Qu.:0.121691   3rd Qu.:0.39125   3rd Qu.:0.637871
 Max.   :0.415550   Max.   :0.96748   Max.   :0.899200
[1] 148

Confusion matrix
Confusion Matrix and Statistics

          Reference
Prediction  <30  >30   NO
       <30    0    0    0
       >30  488 1453  950
       NO  1483 4907 9056

Overall Statistics

               Accuracy : 0.5731
                 95% CI : (0.5659, 0.5803)
    No Information Rate : 0.5457
    P-Value [Acc > NIR] : 4.1e-14

                  Kappa : 0.121
 Mcnemar’s Test P-Value : < 2e-16

Statistics by Class:

                     Class: <30 Class: >30 Class: NO
Sensitivity              0.0000    0.22846    0.9051
Specificity              1.0000    0.87994    0.2330
Pos Pred Value              NaN    0.50259    0.5863
Neg Pred Value           0.8925    0.68231    0.6714
Prevalence               0.1075    0.34684    0.5457
Detection Rate           0.0000    0.07924    0.4939
Detection Prevalence     0.0000    0.15766    0.8423
Balanced Accuracy        0.5000    0.55420    0.5690

MSE
[1] 0.4268964
```



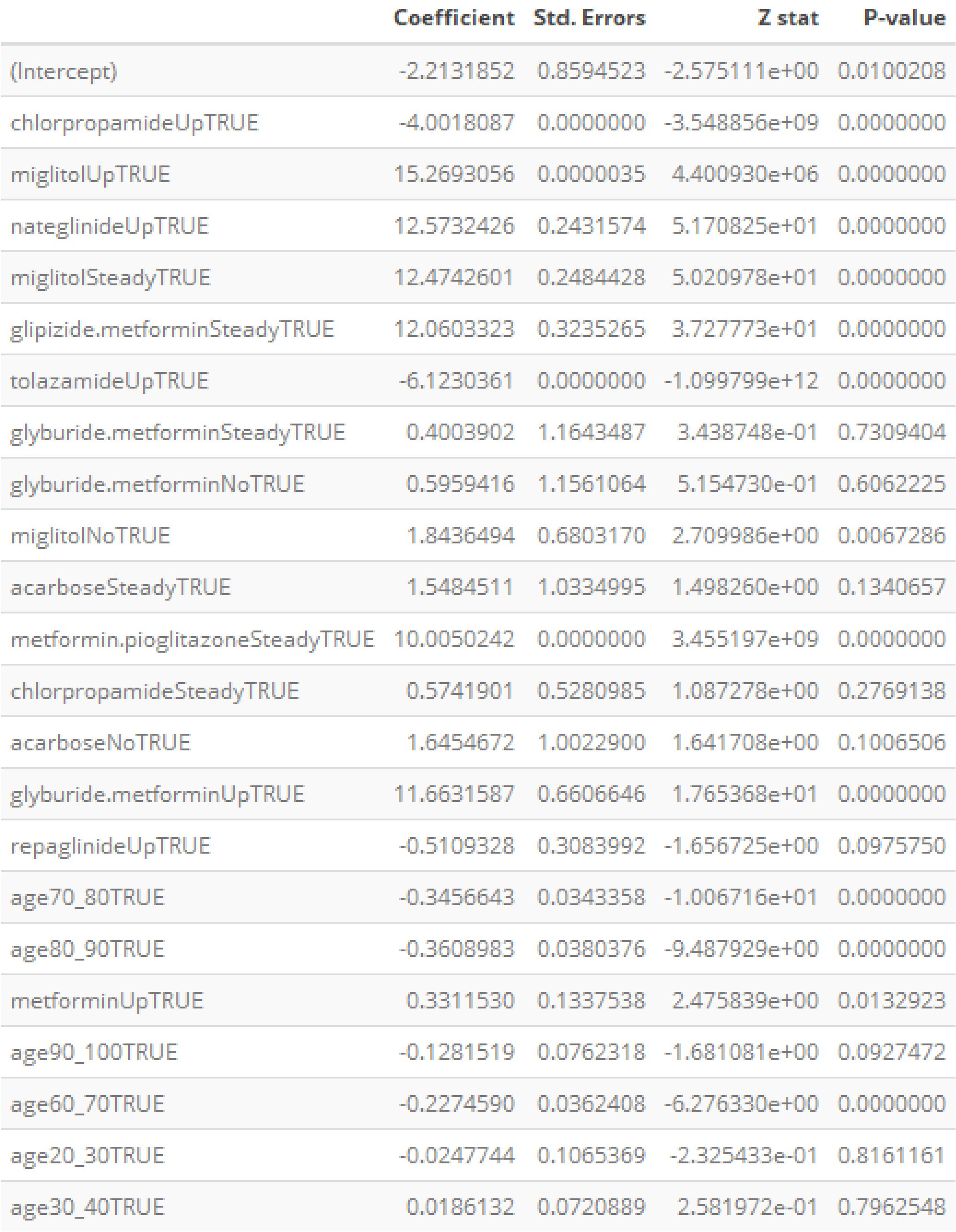
	Coefficient	Std. Errors	Z stat	P-value
(Intercept)	2.1906063	0.1754589	1.248501e+01	0.0000000
raceAsian	0.2099470	0.1610357	1.303729e+00	0.1923259
raceCaucasian	0.0067418	0.0329541	2.045802e-01	0.8379001
raceHispanic	0.1100712	0.0961043	1.145331e+00	0.2520721
raceOther	0.1712421	0.1092426	1.567540e+00	0.1169885
genderMale	0.0061406	0.0257279	2.386740e-01	0.8113584
genderUnknown/Invalid	0.0000000	NaN	NaN	NaN
age10_20	-0.4491377	0.2117329	-2.121247e+00	0.0339010
age20_30	-0.7366690	0.1584687	-4.648671e+00	0.0000033
age30_40	-0.7287792	0.1403726	-5.191749e+00	0.0000002
age40_50	-0.7355914	0.1320669	-5.569841e+00	0.0000000
age50_60	-0.6832848	0.1298147	-5.263537e+00	0.0000001
age60_70	-0.8719203	0.1289007	-6.764279e+00	0.0000000
age70_80	-0.9612619	0.1285371	-7.478475e+00	0.0000000
age80_90	-0.9060113	0.1295762	-6.992112e+00	0.0000000
age90_100	-0.6753556	0.1438410	-4.695155e+00	0.0000027
time_in_hospital	-0.0250456	0.0048179	-5.198478e+00	0.0000002
num_lab_procedures	-0.0005156	0.0007291	-7.071690e-01	0.4794614
num_procedures	0.0349853	0.0083517	4.188995e+00	0.0000280
num_medications	-0.0069885	0.0021404	-3.265088e+00	0.0010943
number_outpatient	-0.1726264	0.0274350	-6.292207e+00	0.0000000
number_emergency	-0.2442635	0.0285152	-8.566088e+00	0.0000000
number_inpatient	-0.5012522	0.0132749	-3.775949e+01	0.0000000
number_diagnoses	-0.0644866	0.0075081	-8.588917e+00	0.0000000
max_glu_serum>300	-0.1979917	0.1509455	-1.311677e+00	0.1896293
max_glu_serumNone	0.0193728	0.1043635	1.856283e-01	0.8527363
max_glu_serumNorm	0.1512979	0.1312109	1.153090e+00	0.2488735
A1Cresult>8	-0.0138816	0.0819793	-1.693309e-01	0.8655364
A1CresultNone	-0.0677218	0.0685574	-9.878122e-01	0.3232446
A1CresultNorm	0.1042858	0.0888174	1.174160e+00	0.2403311
metforminNo	0.1956174	0.1554039	1.258767e+00	0.2081144
metforminSteady	0.4042579	0.1557889	2.594908e+00	0.0094616
metforminUp	0.5863035	0.2054547	2.853687e+00	0.0043215
repaglinideNo	-0.1674867	0.2120557	-7.898238e-01	0.4296307
repaglinideSteady	-0.3764659	0.2243265	-1.678205e+00	0.0933071
repaglinideUp	-0.4327342	0.2959034	-1.462417e+00	0.1436270
nateglinideNo	-0.1658663	0.1459682	-1.136318e+00	0.2558234
nateglinideSteady	-0.2877297	0.1600231	-1.798051e+00	0.0721689
nateglinideUp	1.7887993	0.1913774	9.346973e+00	0.0000000
chlorpropamideNo	0.4526810	0.2525695	1.792302e+00	0.0730845
chlorpropamideSteady	0.7904754	0.2978744	2.653721e+00	0.0079610
chlorpropamideUp	-2.0501618	0.0022138	-9.260963e+02	0.0000000
glimepirideNo	0.1890519	0.2676173	7.064264e-01	0.4799230
glimepirideSteady	0.3132057	0.2723183	1.150146e+00	0.2500839
glimepirideUp	0.2787105	0.3390259	8.220920e-01	0.4110245
glipizideNo	0.2572491	0.1618317	1.589609e+00	0.1119230
glipizideSteady	0.2126587	0.1621615	1.311400e+00	0.1897225
glipizideUp	0.1044072	0.2072865	5.036856e-01	0.6144823
glyburideNo	-0.2630240	0.1871265	-1.405595e+00	0.1598445
glyburideSteady	-0.2877560	0.1877490	-1.532663e+00	0.1253589
glyburideUp	-0.1608161	0.2313940	-6.949884e-01	0.4870626
pioglitazoneNo	0.4505020	0.3288555	1.369909e+00	0.1707154
pioglitazoneSteady	0.5236431	0.3318718	1.577848e+00	0.1146006
pioglitazoneUp	0.2972651	0.4128296	7.200675e-01	0.4714835
rosiglitazoneNo	-0.5802478	0.1893478	-3.064454e+00	0.0021807
rosiglitazoneSteady	-0.6261159	0.1933877	-3.237620e+00	0.0012053
rosiglitazoneUp	-0.3518958	0.2689932	-1.308196e+00	0.1908069
acarboseNo	1.5721531	0.1876412	8.378508e+00	0.0000000
acarboseSteady	1.4547695	0.2080129	6.993649e+00	0.0000000
acarboseUp	-0.1753267	0.2799419	-6.262970e-01	0.5311202
miglitolNo	0.0477720	0.1699571	2.810829e-01	0.7786468
miglitolSteady	2.7660709	0.1449722	1.908001e+01	0.0000000
miglitolUp	2.2987747	0.0078405	2.931931e+02	0.0000000
tolazamideSteady	0.4302140	0.2380355	1.807352e+00	0.0707074
tolazamideUp	-1.1922273	0.0000884	-1.348950e+04	0.0000000
insulinNo	0.2071053	0.0671440	3.084492e+00	0.0020390
insulinSteady	0.2534112	0.0506600	5.002197e+00	0.0000006
insulinUp	0.1787057	0.0508452	3.514698e+00	0.0004403
glyburide.metforminNo	-0.6522383	0.2311176	-2.822106e+00	0.0047709
glyburide.metforminSteady	-0.7723143	0.2400165	-3.217755e+00	0.0012920
glyburide.metforminUp	0.2469468	0.4314098	5.724180e-01	0.5670388
glipizide.metforminSteady	1.1486326	0.3275940	3.506269e+00	0.0004544
metformin.pioglitazoneSteady	1.0753031	0.0007786	1.381033e+03	0.0000000
changeNo	-0.0070628	0.0486077	-1.453025e-01	0.8844720
diabetesMedYes	-0.2960503	0.0475872	-6.221221e+00	0.0000000

Figure 4: Summary of model feature coefficients (Coefficient, Std. Error, Z-stat, and p-value) WITHOUT variable selection (75 features are currently being used in this model)

6.7 Multinomial Logistic Neural Net with Variable Selection Output

```
    Look at first few most important variables
                           Overall                 Variables
chlorpropamideUp          4.965912          chlorpropamideUp
miglitolUp                4.008221                miglitolUp
nateglinideUp             3.957236             nateglinideUp
miglitolSteady            3.852238            miglitolSteady
glipizide.metforminSteady 2.905403 glipizide.metforminSteady
tolazamideUp              2.859611              tolazamideUp

Fit multinomial logistic neural net and get number of features
    in model, probabilities, and effective DF
# weights:  72 (46 variable)
initial  value 82562.910718
iter  10 value 71352.520457
iter  20 value 70110.276094
iter  30 value 69758.372756
iter  40 value 69752.485797
final  value 69752.431522
converged
[1] 23
      <30                 >30                  NO
 Min.   :2.100e-07   Min.   :0.0000001   Min.   :0.0000001
 1st Qu.:8.956e-02   1st Qu.:0.3282274   1st Qu.:0.5310456
 Median :1.061e-01   Median :0.3444211   Median :0.5494634
 Mean   :1.049e-01   Mean   :0.3406163   Mean   :0.5545020
 3rd Qu.:1.154e-01   3rd Qu.:0.3535272   3rd Qu.:0.5822084
 Max.   :2.860e-01   Max.   :0.9999996   Max.   :0.9999980
[1] 46
Confusion Matrix and Statistics

          Reference
Prediction   <30   >30    NO
       <30     0     0     0
       >30     4    26    22
       NO   1999  6445 10293

Overall Statistics

               Accuracy : 0.5492
                 95% CI : (0.5421, 0.5563)
    No Information Rate : 0.549
    P-Value [Acc > NIR] : 0.4796

                  Kappa : 0.0017

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

```
Statistics by Class:

                     Class: <30 Class: >30 Class: NO
Sensitivity              0.0000   0.004018   0.99787
Specificity              1.0000   0.997889   0.00354
Pos Pred Value              NaN   0.500000   0.54934
Neg Pred Value           0.8934   0.656028   0.57692
Prevalence               0.1066   0.344404   0.54899
Detection Rate           0.0000   0.001384   0.54782
Detection Prevalence     0.0000   0.002768   0.99723
Balanced Accuracy        0.5000   0.500954   0.50070

MSE
[1] 0.4507957
```



	Coefficient	Std. Errors	Z stat	P-value
(Intercept)	-2.2131852	0.8594523	-2.575111e+00	0.0100208
chlorpropamideUpTRUE	-4.0018087	0.0000000	-3.548856e+09	0.0000000
miglitolUpTRUE	15.2693056	0.0000035	4.400930e+06	0.0000000
nateglinideUpTRUE	12.5732426	0.2431574	5.170825e+01	0.0000000
miglitolSteadyTRUE	12.4742601	0.2484428	5.020978e+01	0.0000000
glipizide.metforminSteadyTRUE	12.0603323	0.3235265	3.727773e+01	0.0000000
tolazamideUpTRUE	-6.1230361	0.0000000	-1.099799e+12	0.0000000
glyburide.metforminSteadyTRUE	0.4003902	1.1643487	3.438748e-01	0.7309404
glyburide.metforminNoTRUE	0.5959416	1.1561064	5.154730e-01	0.6062225
miglitolNoTRUE	1.8436494	0.6803170	2.709986e+00	0.0067286
acarboseSteadyTRUE	1.5484511	1.0334995	1.498260e+00	0.1340657
metformin.pioglitazoneSteadyTRUE	10.0050242	0.0000000	3.455197e+09	0.0000000
chlorpropamideSteadyTRUE	0.5741901	0.5280985	1.087278e+00	0.2769138
acarboseNoTRUE	1.6454672	1.0022900	1.641708e+00	0.1006506
glyburide.metforminUpTRUE	11.6631587	0.6606646	1.765368e+01	0.0000000
repaglinideUpTRUE	-0.5109328	0.3083992	-1.656725e+00	0.0975750
age70_80TRUE	-0.3456643	0.0343358	-1.006716e+01	0.0000000
age80_90TRUE	-0.3608983	0.0380376	-9.487929e+00	0.0000000
metforminUpTRUE	0.3311530	0.1337538	2.475839e+00	0.0132923
age90_100TRUE	-0.1281519	0.0762318	-1.681081e+00	0.0927472
age60_70TRUE	-0.2274590	0.0362408	-6.276330e+00	0.0000000
age20_30TRUE	-0.0247744	0.1065369	-2.325433e-01	0.8161161
age30_40TRUE	0.0186132	0.0720889	2.581972e-01	0.7962548

Figure 5: Summary of model prediction coefficients after variable selection (only 23 variables are in the model now)

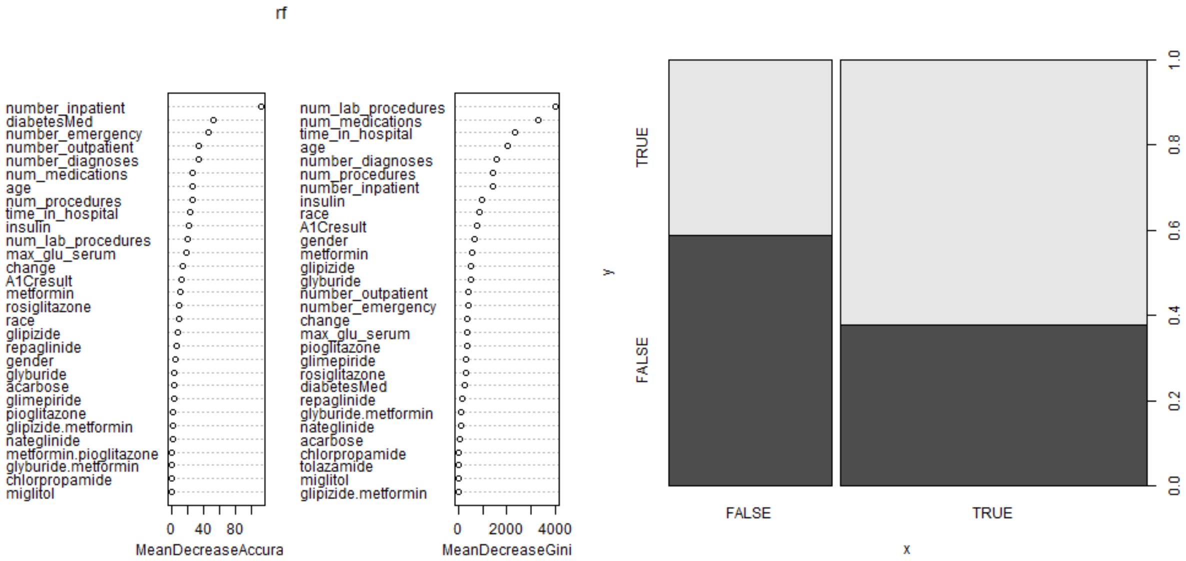
6.8 Boosting Output

```
    Do boosting
                                          var     rel.inf
number_inpatient             number_inpatient 21.75878071
age                                       age 12.36010246
num_lab_procedures         num_lab_procedures 10.14851319
num_medications               num_medications  8.29298172
number_diagnoses             number_diagnoses  6.76682515
time_in_hospital             time_in_hospital  5.30771716
race                                     race  4.83601697
num_procedures                 num_procedures  3.95956107
insulin                               insulin  3.54441477
number_emergency             number_emergency  3.52579776
number_outpatient           number_outpatient  2.87797778
max_glu_serum                   max_glu_serum  2.32475087
A1Cresult                           A1Cresult  2.32310869
metformin                           metformin  1.72937369
diabetesMed                       diabetesMed  1.72220080
glyburide                           glyburide  1.29364723
glipizide                           glipizide  1.24217629
rosiglitazone                   rosiglitazone  1.07453978
glimepiride                       glimepiride  0.97885574
gender                                 gender  0.78470464
pioglitazone                     pioglitazone  0.77553551
repaglinide                       repaglinide  0.72779240
nateglinide                       nateglinide  0.32406589
change                                 change  0.30248183
glyburide.metformin       glyburide.metformin  0.29873755
acarbose                             acarbose  0.25288221
chlorpropamide                 chlorpropamide  0.24566728
miglitol                             miglitol  0.14551590
tolazamide                         tolazamide  0.07527494
glipizide.metformin       glipizide.metformin  0.00000000
metformin.pioglitazone metformin.pioglitazone  0.00000000

Accuracy
[1] 0.6229997
```

6.9 Random Forest Output

```
    Do random forest on training set

Call:
 randomForest(formula = formula, data = diabetic.sub, mtry = 6,      importance = TRUE, subset = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 38.89%
Confusion matrix:
      FALSE  TRUE class.error
FALSE 13348 15531   0.5377956
TRUE   9443 25900   0.2671816

Confusion matrix using the test set

rf.predict FALSE  TRUE
     FALSE  5616  3940
     TRUE   6774 11195

Accuracy
[1] 0.6107539

Calculate importance of the variables
                           FALSE        TRUE MeanDecreaseAccuracy MeanDecreaseGini
race                   4.6284285   7.4435099           8.74739671       835.918054
gender                 3.1502932   1.7869977           3.58911239       659.026793
age                    4.9244659  29.0625270          25.52790441      2014.824776
time_in_hospital      -4.0617969  31.3605333          23.09593546      2314.993611
num_lab_procedures   -11.2414670  34.2483927          19.59796832      4019.685069
num_procedures        -2.6343842  34.2995860          25.10503325      1441.782356
num_medications       -0.3902545  26.2530286          25.99414835      3283.299655
number_outpatient      2.3758663  40.5651178          33.21641365       415.413304
number_emergency       0.7802175  53.0933453          45.22694477       412.850035
number_inpatient      70.9510170 103.3032669         112.52779742      1410.396966
number_diagnoses      11.2153799  29.6881999          32.69286023      1550.073592
max_glu_serum         -7.2050757  27.4410638          18.32469048       361.504021
A1Cresult             -9.0345667  21.5179086          11.45422412       761.823169
metformin             -8.6196178  19.3310203          10.77093243       551.282538
repaglinide           -1.2862058   7.9108846           4.94820978       133.404901
nateglinide            0.8536031  -0.4070501           0.30084491        73.151080
chlorpropamide         2.3643441  -2.7800565          -0.34598861        12.662631
glimepiride           -0.1840706   2.6918065           2.12508358       317.245198
glipizide              1.1774832   5.7152612           6.97194293       499.344082
glyburide              2.4058391   1.1793523           2.69463318       485.456669
pioglitazone           0.8966737   0.8970810           1.39005299       355.601467
```

```
rosiglitazone            2.7309728   6.7136552          9.09893790      316.217569
acarbose                 0.1899245   2.8706574          2.19997618       31.243677
miglitol                 0.1268356  -1.0914318         -0.57210057        3.732658
tolazamide               1.1980305  -4.1086007         -2.22201171        5.852963
insulin                 -8.8461944  26.3286609         21.35251220      967.721379
glyburide.metformin     -0.6319962   0.4986783         -0.05217561       74.084428
glipizide.metformin      0.4453654   0.5748033          0.70089943        1.863285
metformin.pioglitazone   0.0000000   0.0000000          0.00000000        0.000000
change                  -5.8175768  13.0138040         13.62706569      367.774503
diabetesMed            -12.9584598  46.7028824         51.67921992      230.409146
```



(a) Individual Variable Importance

(b) Final Prediction of Response

Figure 6: Plots obtained by fitting a random forest.

6.10 R Script

```
## Brian Kang
## 05/17/2019
## ECON484
## Final Project

# import data————————————————————————————————————————
rm(list = ls())  # reset working vars
setwd(”C:/Users/slexi/Documents/ECON484”)  # set working directory
temp <- read.csv(”diabetic_data.csv”, na.strings = ”?”)  # save temp data
temp2 <- temp  # backup
#temp <- temp2  # recover backup

#names(temp)
#head(temp)
#str(temp)
#sapply(temp, class)
#summary(temp)

# install packages when needed
#install.packages(”naniar”)
#install.packages(”car”)
#install.packages(”fitdistrplus”)
#install.packages(”hdm”)
#install.packages(”stringr”)
#install.packages(”caret”)
#install.packages(”kableExtra”)
#webshot::install_phantomjs()
#install.packages(”magick”)

# clean data—————————————————————————————————————————
# unique type int identifiers should be type factors
temp$encounter_id <- as.factor(temp$encounter_id)
temp$patient_nbr <- as.factor(temp$patient_nbr)
temp$admission_type_id <- as.factor(temp$admission_type_id)
temp$discharge_disposition_id <- as.factor(temp$discharge_disposition_id)
temp$admission_source_id <- as.factor(temp$admission_source_id)
```

```r
# replace meaningless identifiers to NA
#library(naniar)
#replace_with_na(temp, replace = list(admission_type_id=c(5,6,8)))

# c(5,6,8) replaces only some to NA
temp$admission_type_id[temp$admission_type_id==5] <- NA
temp$admission_type_id[temp$admission_type_id==6] <- NA
temp$admission_type_id[temp$admission_type_id==8] <- NA

# c(18,25,26) replaces only some to NA
temp$discharge_disposition_id[temp$discharge_disposition_id==18] <- NA
temp$discharge_disposition_id[temp$discharge_disposition_id==25] <- NA
temp$discharge_disposition_id[temp$discharge_disposition_id==26] <- NA

# c(9,15,17,20,21) replaces only some to NA
temp$admission_source_id[temp$admission_source_id==9] <- NA
temp$admission_source_id[temp$admission_source_id==15] <- NA
temp$admission_source_id[temp$admission_source_id==17] <- NA
temp$admission_source_id[temp$admission_source_id==20] <- NA
temp$admission_source_id[temp$admission_source_id==21] <- NA

# rename levels of vars "age"
for (ii in 0:(length(levels(temp$age))-1)) {
  nm <- paste("[",ii*10,"-",(ii+1)*10,")", sep = "")
  chng <- paste(ii*10,"_",(ii+1)*10, sep = "")
  levels(temp$age)[levels(temp$age)==nm] <- chng
}
# rename levels of vars "weight"
for (ii in 0:(length(levels(temp$weight))-2)) {
  nm <- paste("[",ii*25,"-",(ii+1)*25,")", sep = "")
  chng <- paste(ii*25,"_",(ii+1)*25, sep = "")
  levels(temp$weight)[levels(temp$weight)==nm] <- chng
}
# rename levels of vars "medical_specialty"
library(stringr)
levels(temp$medical_specialty) <- str_replace_all(
  levels(temp$medical_specialty), "[&/\\-]", "_")
```

```
# delete rows with three unknown genders
delete <- which(temp$gender==”Unknown/Invalid”)
temp <- temp[-delete,]
# delete rows with num_lab_procedures >97
outlier <- boxplot(temp$num_lab_procedures, plot = F)$out
delete <- outlier[outlier>97]
temp <- temp[-which(temp$num_lab_procedures %in% delete),]
# delete rows with num_medication >45
outlier <- boxplot(temp$num_medications, plot = F)$out
delete <- outlier[outlier>45]
temp <- temp[-which(temp$num_medications %in% delete),]
# delete rows with number_outpatient >2
outlier <- boxplot(temp$number_outpatient, plot = F)$out
delete <- outlier[outlier>2]
temp <- temp[-which(temp$number_outpatient %in% delete),]
# delete rows with number_emergency >3
outlier <- boxplot(temp$number_emergency, plot = F)$out
delete <- outlier[outlier>3]
temp <- temp[-which(temp$number_emergency %in% delete),]
# delete rows with number_inpatient >=5
outlier <- boxplot(temp$number_inpatient, plot = F)$out
delete <- outlier[outlier>=5]
temp <- temp[-which(temp$number_inpatient %in% delete),]

# modeling--------------------------------------------------------------
diabetic <- temp # rename data after cleaning
temp3 <- diabetic # backup

# Question:
# Did the treatment and medication actually work?
# Predict readmitted or not, using variables related to treatment
# and/or examination in hospital and variables related to medication

# Function to reset dataset for each model
reset <- function() {
  diabetic <- temp3 # recover backup
  isReadmitted <- ifelse(diabetic$readmitted %in% c(”<30”,”>30”),F,T)
  diabetic <- cbind(diabetic,isReadmitted)
  # split data into train & test
  set.seed(987)
  train <- sample(1:nrow(diabetic), nrow(diabetic)*0.8)  # 80% for training

  # get which variables have <2 factor levels
  get <- which(sapply(diabetic[train,], function(x) length(unique(x))<2))
  # exclude encounter_id, patient_nbr, weight, payer_code, diag_1, daig_2,
  # diag_3, readmitted, isReadmitted
  # Reason: unrelated to question OR too many factors
  # also exclude acetohexamide, tolbutamide, troglitazone,
  # glimepiride.pioglitazone, metformin.rosiglitazone
  # Reason: causes "<2 level" error from sampling
  varnames <- paste(c(names(diabetic
    [,-c(get,1,2,6,11,19,20,21,30,33,38,45,46,50,51)])), collapse = "+")
  formula <- paste(c("isReadmitted",varnames), collapse = "~")
  return(list(diabetic, train, get, varnames, formula))
}
resetData <- reset()
diabetic <- resetData[[1]]
train <- resetData[[2]]
get <- resetData[[3]]
varnames <- resetData[[4]]
formula <- resetData[[5]]

# Model 1: LASSO
#sink("lasso_output.txt")  # start outputing to text file
library(hdm)
lasso.1 <- rlasso(formula , data = diabetic[train,], post = F)
#cat("Do LASSO on training set\n")
summary(lasso.1, all = F)

# get ceoffs that matter and make OLS formula
x <- which(coef(lasso.1)[-1]!=0)
#cat("\nCount and Kept Significant Variables by LASSO\nCount: ")
length(x)
#x
x <- paste(names(x), collapse = "+")
formula <- paste(c("isReadmitted", x), collapse = "␣~␣")
```

```
# name all extra variables created from doing OLS
diabetic$raceAsian <- diabetic$race == "Asian"
diabetic$raceHispanic <- diabetic$race == "Hispanic"
diabetic$raceOther <- diabetic$race == "Other"
diabetic$genderMale <- diabetic$gender == "Male"
diabetic$age30_40 <- diabetic$age == "30_40"
diabetic$age50_60 <- diabetic$age == "50_60"
diabetic$age70_80 <- diabetic$age == "70_80"
diabetic$age80_90 <- diabetic$age == "80_90"
diabetic$age90_100 <- diabetic$age == "90_100"
diabetic$admission_type_id2 <- diabetic$admission_type_id == "2"
diabetic$discharge_disposition_id5 <-
  diabetic$discharge_disposition_id == "5"
diabetic$discharge_disposition_id6 <-
  diabetic$discharge_disposition_id == "6"
diabetic$discharge_disposition_id11 <-
  diabetic$discharge_disposition_id == "11"
diabetic$discharge_disposition_id13 <-
  diabetic$discharge_disposition_id == "13"
diabetic$discharge_disposition_id14 <-
  diabetic$discharge_disposition_id == "14"
diabetic$discharge_disposition_id19 <-
  diabetic$discharge_disposition_id == "19"
diabetic$discharge_disposition_id22 <-
  diabetic$discharge_disposition_id == "22"
diabetic$discharge_disposition_id23 <-
  diabetic$discharge_disposition_id == "23"
diabetic$admission_source_id4 <- diabetic$admission_source_id == "4"
diabetic$admission_source_id5 <- diabetic$admission_source_id == "5"
diabetic$admission_source_id6 <- diabetic$admission_source_id == "6"
diabetic$admission_source_id7 <- diabetic$admission_source_id == "7"
diabetic$medical_specialtyEmergency_Trauma <-
  diabetic$medical_specialty == "Emergency_Trauma"
diabetic$medical_specialtyFamily_GeneralPractice <-
  diabetic$medical_specialty == "Family_GeneralPractice"
diabetic$medical_specialtyGastroenterology <-
  diabetic$medical_specialty == "Gastroenterology"
```

```
diabetic$medical_specialtyGynecology <-
  diabetic$medical_specialty == "Gynecology"
diabetic$medical_specialtyHematology <-
  diabetic$medical_specialty == "Hematology"
diabetic$medical_specialtyInternalMedicine <-
  diabetic$medical_specialty == "InternalMedicine"
diabetic$medical_specialtyNephrology <-
  diabetic$medical_specialty == "Nephrology"
diabetic$medical_specialtyNeurology <-
  diabetic$medical_specialty == "Neurology"
diabetic$medical_specialtyObstetricsandGynecology <-
  diabetic$medical_specialty == "ObstetricsandGynecology"
diabetic$medical_specialtyObstetrics <-
  diabetic$medical_specialty == "Obstetrics"
diabetic$medical_specialtyOncology <-
  diabetic$medical_specialty == "Oncology"
diabetic$medical_specialtyOrthopedics <-
  diabetic$medical_specialty == "Orthopedics"
diabetic$medical_specialtyOrthopedics_Reconstructive <-
  diabetic$medical_specialty == "Orthopedics_Reconstructive"
diabetic$medical_specialtyOtolaryngology <-
  diabetic$medical_specialty == "Otolaryngology"
diabetic$medical_specialtyPediatrics_Endocrinology <-
  diabetic$medical_specialty == "Pediatrics_Endocrinology"
diabetic$medical_specialtyPediatrics_Pulmonology <-
  diabetic$medical_specialty == "Pediatrics_Pulmonology"
diabetic$medical_specialtyPulmonology <-
  diabetic$medical_specialty == "Pulmonology"
diabetic$medical_specialtySurgeon <-
  diabetic$medical_specialty == "Surgeon"
diabetic$medical_specialtySurgery_Cardiovascular <-
  diabetic$medical_specialty == "Surgery_Cardiovascular"
diabetic$medical_specialtySurgery_Cardiovascular_Thoracic <-
  diabetic$medical_specialty == "Surgery_Cardiovascular_Thoracic"
diabetic$medical_specialtySurgery_Neuro <-
  diabetic$medical_specialty == "Surgery_Neuro"
diabetic$medical_specialtySurgery_Vascular <-
  diabetic$medical_specialty == "Surgery_Vascular"
```

```
diabetic$A1CresultNone <- diabetic$A1Cresult == "None"
diabetic$A1CresultNorm <- diabetic$A1Cresult == "Norm"
diabetic$metforminNo <- diabetic$metformin == "No"
diabetic$repaglinideNo <- diabetic$repaglinide == "No"
diabetic$glipizideNo <- diabetic$glipizide == "No"
diabetic$pioglitazoneUp <- diabetic$pioglitazone == "Up"
diabetic$acarboseNo <- diabetic$acarbose == "No"
diabetic$tolazamideSteady <- diabetic$tolazamide == "Steady"
diabetic$insulinSteady <- diabetic$insulin == "Steady"
diabetic$metforminSteady <- diabetic$metformin == "Steady"
diabetic$changeNo <- diabetic$change == "No"
diabetic$diabetesMedYes <- diabetic$diabetesMed == "Yes"

# OLS regression on training set
olsLasso.1 <- lm(formula, data = diabetic[train,])
#cat("\nDo OLS on training set using selected variables from LASSO\n")
summary(olsLasso.1)$coefficients[,1]
# prediction on test data to predict patient readmission or not
prob.lasso.1 <- predict(olsLasso.1, newdata = diabetic[-train,])
#cat("Predict on test set\n")
summary(prob.lasso.1)
#cat("\nCount remaining observations\n")
length(na.omit(prob.lasso.1)) # count remaining observations
# test error
mse.1 <- mean((prob.lasso.1-diabetic[-train,]$isReadmitted)^2, na.rm=T)
#cat("\nMSE\n")
mse.1
#sink() # stop writing to text file

# ---------------------------------------------------------------
# Model 2: LASSO with CV choosing Tuning Parameter
#sink("lassoCV_output.txt") # start outputing to text file
resetData <- reset() # reset data
diabetic <- resetData[[1]]
train <- resetData[[2]]
get <- resetData[[3]]
varnames <- resetData[[4]]
formula <- resetData[[5]]
```

```r
# split into train & test
# takeout intercept
xtrain <- model.matrix(as.formula(formula), data = diabetic[train,])[,-1]
xtest <- model.matrix(as.formula(formula), data = diabetic[-train,])[,-1]
set.seed(987)
# nrow unequal so adjust
ytrain <- diabetic[sample(train, nrow(xtrain)),]$isReadmitted

# cross validation then fit LASSO
library(glmnet)
set.seed(987)
cv.lasso.1 <- cv.glmnet(xtrain, ytrain, alpha = 1)  # 1 for lasso
#cat("CV on LASSO and get min tuning parameter\n")
cv.lasso.1[c(8,9)]
cv.lambda <- cv.lasso.1$lambda.min  # get smallest tuning parameter
#png(filename="lassoCV.png")  # save plot
plot(cv.lasso.1)
#dev.off()
lasso.2 <- glmnet(xtrain, ytrain, alpha = 1, lambda = cv.lambda)
#cat("\nDo LASSO using min tuning parameter then predict\n")
summary(lasso.2)
lasso.2
# prediction on test data to predict patient readmission or not
pred.lasso.2 <- predict(lasso.2, s = cv.lambda, newx = xtest)
summary(pred.lasso.2)
test <- (1:nrow(diabetic))[-train]  # test data
# test error
mse.2 <- mean((pred.lasso.2-diabetic
            [sample(test,length(pred.lasso.2)),]$isReadmitted)^2, na.rm=T)
#cat("\nMSE\n")
mse.2
#sink()  # stop writing to text file

# ---------------------------------------------------------------------
# Model 3: Logistic Regression
#sink("logit_output.txt")  # start outputing to text file
resetData <- reset()  # reset data
```

```
diabetic <- resetData[[1]]
train <- resetData[[2]]
get <- resetData[[3]]
# update formula
# also exclude admission_type_id, discharge_disposition_id, admission_source_id
# and medical_specialty
# Reason: error in dataset jams logit
varnames <- paste(c(names(diabetic
  [,-c(get,1,2,6,7,8,9,11,12,19,20,21,33,38,45,46,50,51)])), collapse = "+")
formula <- paste(c("isReadmitted",varnames), collapse = "~")

# fit logit using training data
logit.1 <- glm(formula, data = diabetic, family = "binomial", subset=train)
#cat("Do logit on training set\n")
summary(logit.1)  # very sparse results, many individually insignificant
#plot(logit.1)

# predict probability of readmission using test data
logit.prob.1 <- predict(logit.1, newdata = diabetic[-train,], type = "response")
#cat("\nPredict using test set\n")
summary(logit.prob.1)
hist(logit.prob.1)

# predicting whether patient will be readmitted or not
# if prob > 1/2 then patient will not readmit
logit.pred.1 <- rep(F, nrow(diabetic[-train,]))
logit.pred.1[logit.prob.1 > 0.5] <- T

# test error
mse.3 <- mean((logit.pred.1-diabetic$isReadmitted[-train])^2)
#cat("\nMSE\n")
mse.3

#cat("\nConfusion matrix\n")
# confusion matrix
table(logit.pred.1, diabetic$isReadmitted[-train])
#cat("\nAccuracy\n")
mean(logit.pred.1 == diabetic$isReadmitted[-train])
```

```
#cat("\nNOTICE, sum of MSE and accuracy = \n")
mse.3 + mean(logit.pred.1 == diabetic$isReadmitted[-train]) # =1!
#sink() # stop writing to text file

# --------------------------------------------------------------------
# Model 3.5: Logit with K-fold CV
## Author: Matt Kelly
## Date: 06/05/2019
set.seed(987)
k = 10
folds = sample(1:k, nrow(diabetic), replace = TRUE)
cv.error.10 = matrix(NA, nrow = k, ncol = 19)
diab.pred.function = function(object, newdata, id,...){
  form = as.formula(object$call[[2]])
  mat = model.matrix(form, newdata)
  coefi = coef(object,id=id)
  xvars = names(coefi)
  mat[,xvars]%*%coefi
}
## end --------------------------------------------------------------


# --------------------------------------------------------------------
# Instead of predicting whether patient was readmitted or not,
# predict the range of days between previous and next readmission
# factors: None, <30 days, >30 days
# Model 4: Multinomial Logistic using Neural Network
#sink("nn_output.txt") # start outputing to text file
library(nnet)
library(pscl)
resetData <- reset() # reset data
diabetic <- resetData[[1]]
train <- resetData[[2]]
get <- resetData[[3]]
# use formula from above logit but for readmitted, not isReadmitted
formula <- paste(c("readmitted",varnames), collapse = "~")

# fit multinomial
```

```
nn.1 <- multinom(formula, data = diabetic[train,])
#cat("Fit multinomial logistic neural net and get number of features
#    in model, probabilities, and effective DF\n")
length(nn.1$coefnames)  # number of features
summary(nn.1$fitted.values)
nn.1$edf  # effective DF exhausted up by model
#nn.1$deviance  # residual deviance, minus twice log-likelihook
#nn.1$AIC  # AIC for fit

# predict probability of days between readmission using test data
nn.pred.1 <- predict(nn.1, newdata = diabetic[-train,], type = "probs")
# predict intervals between readmission using test data
nn.class.1 <- predict(nn.1, newdata = diabetic[-train,])
#cat("\nConfusion matrix\n")
# confusion matrix
library(caret)
caret::confusionMatrix(as.factor(nn.class.1),
                       as.factor(diabetic[-train,]$readmitted))
# test error
mse.4 <- mean(as.character(nn.class.1) !=
                as.character(diabetic[-train,]$readmitted), na.rm = T)
#cat("\nMSE\n")
mse.4
#sink()  # stop writing to text file

# calculate z score and p values
c <- summary(nn.1)$coefficients
se <- summary(nn.1)$standard.errors
z <- c/se
p <- (1-pnorm(abs(z),0,1))*2  # I am using two-tailed z test
z
p
summ <- as.data.frame(rbind(c[2,],se[2,],z[2,],p[2,]))
rownames(summm) <- c("Coefficient","Std.␣Errors","Z␣stat","P-value")
summ <- t(summm)

# make neat table
library(knitr)
library(kableExtra)
library(dplyr)
library(magick)
#summ %>%
#  mutate_if(is.numeric, function(x) {
#    cell_spec(x, bold = T,
#              color = spec_color(x, end=0),
#              font_size = spec_font_size(x, end = 12))
#  }) %>%

#save_kable(
  kable(summ, escape = F) %>%
  kable_styling(fixed_thead = T, bootstrap_options =
                  c("striped", "condensed", "responsive"),
                full_width = F, font_size = 12)
#  , "nn1.png")

#------------------------------------------------------------------
# Model 4.5: Same model but with variable selection
#sink("nn_output2.txt")  # start outputing to text file
# calculate important variables
impvars <- varImp(nn.1)
impvars$Variables <- row.names(impvars)
impvars <- impvars[order(-impvars$Overall),]
#cat("Look at first few most important variables\n")
head(impvars)

# choose variables that matter
imp1 <- names(summ)[which(p[2,-1]<0.001)]  # individual significance
imp2 <- impvars$Variables[which(impvars$Overall>1)]  # overall importance
critvars <- union(imp1, imp2)

# make formula
varnames <- paste(critvars, collapse = "+")
formula <- paste(c("readmitted",varnames), collapse = "~")

# name all extra variables created
diabetic$age10_20 <- diabetic$age == "10_20"
```

```
diabetic$age20_30 <- diabetic$age == ”20_30”
diabetic$age30_40 <- diabetic$age == ”30_40”
diabetic$age40_50 <- diabetic$age == ”40_50”
diabetic$age50_60 <- diabetic$age == ”50_60”
diabetic$age60_70 <- diabetic$age == ”60_70”
diabetic$age70_80 <- diabetic$age == ”70_80”
diabetic$age80_90 <- diabetic$age == ”80_90”
diabetic$age90_100 <- diabetic$age == ”90_100”
diabetic$metforminUp <- diabetic$metformin == ”Up”
diabetic$repaglinideNo <- diabetic$repaglinide == ”No”
diabetic$repaglinideSteady <- diabetic$repaglinide == ”Steady”
diabetic$repaglinideUp <- diabetic$repaglinide == ”Up”
diabetic$nateglinideNo <- diabetic$nateglinide == ”No”
diabetic$chlorpropamideSteady <- diabetic$chlorpropamide == ”Steady”
diabetic$pioglitazoneUp <- diabetic$pioglitazone == ”Up”
diabetic$rosiglitazoneNo <- diabetic$rosiglitazone == ”No”
diabetic$rosiglitazoneSteady <- diabetic$rosiglitazone == ”Steady”
diabetic$rosiglitazoneUp <- diabetic$rosiglitazone == ”Up”
diabetic$acarboseNo <- diabetic$acarbose == ”No”
diabetic$acarboseSteady <- diabetic$acarbose == ”Steady”
diabetic$acarboseUp <- diabetic$acarbose == ”Up”
diabetic$miglitolSteady <- diabetic$miglitol == ”Steady”
diabetic$insulinUp <- diabetic$insulin == ”Up”
diabetic$glyburide.metforminNo <- diabetic$glyburide.metformin == ”No”
diabetic$glyburide.metforminSteady <-
  diabetic$glyburide.metformin == ”Steady”
diabetic$changeNo <- diabetic$change == ”No”
diabetic$chlorpropamideUp <- diabetic$chlorpropamide == ”Up”
diabetic$miglitolUp <- diabetic$miglitol == ”Up”
diabetic$nateglinideUp <- diabetic$nateglinide == ”Up”
diabetic$glipizide.metforminSteady <-
  diabetic$glipizide.metformin == ”Steady”
diabetic$tolazamideUp <- diabetic$tolazamide == ”Up”
diabetic$miglitolNo <- diabetic$miglitol == ”No”
diabetic$metformin.pioglitazoneSteady <-
  diabetic$metformin.pioglitazone == ”Steady”
diabetic$glyburide.metforminUp <- diabetic$glyburide.metformin == ”Up”
```

```
#cat("\nFit multinomial logistic neural net and get number of features
#     in model, probabilities, and effective DF\n")
# do multinomial logistic neural nets
nn.2 <- multinom(formula, data = diabetic[train,])
length(nn.2$coefnames)  # number of features
summary(nn.2$fitted.values)
nn.2$edf  # effective DF exhausted up by model
#nn.2$deviance  # residual deviance, minus twice log-likelihook
#nn.2$AIC  # AIC for fit

# predict probability of days between readmission using test data
nn.pred.2 <- predict(nn.2, newdata = diabetic[-train,], type = "probs")
# predict intervals between readmission using test data
nn.class.2 <- predict(nn.2, newdata = diabetic[-train,])
# confusion matrix
caret::confusionMatrix(as.factor(nn.class.2),
                       as.factor(diabetic[-train,]$readmitted))
# test error
mse.4.5 <- mean(na.omit(as.character(nn.class.2) !=
                           as.character(diabetic[-train,]$readmitted)))
#cat("\nMSE\n")
mse.4.5
#sink()  # stop writing to text file

# calculate z score and p values
c2 <- summary(nn.2)$coefficients
se2 <- summary(nn.2)$standard.errors
z2 <- c2/se2
p2 <- (1-pnorm(abs(z2),0,1))*2  # I am using two-tailed z test
z2
p2
summ2 <- as.data.frame(rbind(c2[2,],se2[2,],z2[2,],p2[2,]))
rownames(summ2) <- c("Coefficient","Std.␣Errors","Z␣stat","P-value")
summ2 <- t(summ2)

# make neat table
#save_kable(
  kable(summ2, escape = F) %>%
  kable_styling(fixed_thead = T, bootstrap_options =
                c("striped", "condensed", "responsive"),
              full_width = F, font_size = 12)
#   , "nn2.png")

# ----------------------------------------------------------------------
# Model 5: Boosting
#sink("boosting.txt")  # start outputing to text file
resetData <- reset()  # reset data
diabetic <- resetData[[1]]
train <- resetData[[2]]
get <- resetData[[3]]
varnames <- resetData[[4]]
formula <- resetData[[5]]

## Author: Tatsuya Okuda
## Date: 06/08/2019 ----------------------------------------------------
set.seed(987)
diabetic.sub = diabetic[,-c(get,1,2,6,7,8,9,11,12,19,20,21,33,38,45,46,50)]
varnames <- paste(c(names(diabetic
  [,-c(get,1,2,6,7,8,9,11,12,19,20,21,33,38,45,46,50,51)])), collapse = "+")
formula <- paste(c("isReadmitted",varnames), collapse = "~")
formula = as.formula(formula)

diabetic = na.omit(diabetic.sub) #omit NA
train = sample(1:nrow(diabetic.sub), floor(nrow(diabetic.sub)*0.7))
test = setdiff(1:nrow(diabetic.sub), train)

library(gbm)

boosting = gbm(formula,data=diabetic.sub[train,], distribution = "bernoulli",
              n.trees = 1000, interaction.depth = 4)
#cat("Do boosting\n")
summary(boosting)

boosting.pred = predict.gbm(boosting, newdata = diabetic.sub[test,],
                            n.trees = 1000, type = "response")
prediction = rep(0,length(test))
```

```
prediction[boosting.pred>0.5] = ”TRUE”
prediction[boosting.pred<=0.5] = ”FALSE”

diabetic.sub$isReadmitted[diabetic.sub$isReadmitted==0] = ”FALSE”
diabetic.sub$isReadmitted[diabetic.sub$isReadmitted==1] = ”TRUE”

correct = sum(prediction == diabetic.sub$isReadmitted[test])
total = length(test)
accuracy = correct/total
#cat(”\nAccuracy\n”)
accuracy
## end --------------------------------------------------------------------
#sink()  # stop writing to text file

# ------------------------------------------------------------------------
# Model 6: Random Forest
#sink(”randomforest.txt”)  # start outputing to text file
resetData <- reset()  # reset data
diabetic <- resetData[[1]]
train <- resetData[[2]]
get <- resetData[[3]]
varnames <- resetData[[4]]
formula <- resetData[[5]]

## Author: Tatsuya Okuda
## Date: 06/08/2019 -------------------------------------------------------
diabetic$isReadmitted = as.factor(diabetic$isReadmitted) #factor

#remove medical specialty because RF does not work for too many levels
set.seed(987)
diabetic.sub = diabetic[,-c(get,1,2,6,7,8,9,11,12,19,20,21,33,38,45,46,50)]
diabetic.sub = na.omit(diabetic.sub) #omit NA
train = sample(1:nrow(diabetic.sub), floor(nrow(diabetic.sub)*0.7))
test = setdiff(1:nrow(diabetic.sub), train)

varnames <- paste(c(names(diabetic
  [,-c(get,1,2,6,7,8,9,11,12,19,20,21,33,38,45,46,50,51)])), collapse = ”+”)
formula <- paste(c(”isReadmitted”,varnames), collapse = ”~”)
formula = as.formula(formula)

library(randomForest)
#cat("Do random forest on training set\n")
rf = randomForest(formula, data = diabetic.sub, subset = train,
                  mtry = 6, importance = TRUE)
rf

rf.predict = predict(rf, newdata = diabetic.sub[test,])
#cat("\nConfusion matrix using the test set\n")
table(rf.predict, diabetic.sub$isReadmitted[test])

num.correct = sum(rf.predict==diabetic.sub$isReadmitted[test])
num.total = length(test)
accuracy = num.correct/num.total
#cat("\nAccuracy\n")
accuracy

#cat("\nCalculate importance of the variables\n")
importance(rf)
#sink() # stop writing to text file
#png(filename="rfImportance.png") # save plot
varImpPlot(rf)
#dev.off()
## end ----------------------------------------------------------
#png(filename="randomforest.png") # save plot
# plot the random forest
plot(rf.predict, diabetic.sub$isReadmitted[test])
# abline(0,1) # not used because we are predicting binary response
#dev.off()

#####################DON'T RUN###########################
# make all variables into factors
#factorVars <- numeric(ncol(trainedDiabetic))
#for (ii in 1:ncol(trainedDiabetic)) {
#   if (is.factor(trainedDiabetic[,ii])) {
#     factorVars[ii] <- ii
#   }
#}
#for (ii in 1:length(factorVars)) {
#  if (factorVars[ii]!=0) {
#     trainedDiabetic[,ii] <- as.factor(trainedDiabetic[,ii])
#  }
#}
####################################################

# observing data--------------------------------------------------------------
# races with weight data not available
summary(temp$race[which(is.na(temp$weight))])
# races with weight data not available
summary(temp$race[which(!is.na(temp$weight))])

# highest number of lab procedures in one encounter
temp[temp$num_lab_procedures==132,]
sort(temp$num_lab_procedures,decreasing = T)
#png(filename="cleandat1.png")  # save plot
#layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
min(boxplot(temp$num_lab_procedures,horizontal = T,
            main = "Boxplot of Number of Lab Procedures")$out)
hist(temp$num_lab_procedures, main = "Histogram of Number of Lab Procedures",
     xlab = "num_lab_procedures")
hist(temp$num_lab_procedures, xlim = c(60,140), main = "Enlarged Histogram",
     xlab = "num_lab_procedures")  # DELETE >97
#dev.off()

# highest number of medications in one encounter
temp[temp$num_medications==81,]
sort(temp$num_medications,decreasing = T)
#png(filename="cleandat2.png")  # save plot
#layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
min(boxplot(temp$num_medications,horizontal = T,
            main = "Boxplot of Number of Medication")$out)
hist(temp$num_medications, main = "Histogram of Number of Medication",
     xlab = "num_medication")
hist(temp$num_medications, xlim = c(30,60), main = "Enlarged Histogram",
     xlab = "num_medication")  # DELETE >45
```

```
#dev.off()

# highest number of outpatient visits
temp[temp$number_outpatient==42,]
sort(temp$number_outpatient,decreasing = T)
#png(filename="cleandat3.png")  # save plot
#layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
boxplot(temp$number_outpatient, horizontal = T,
          main = "Boxplot␣of␣Number␣of␣Outpatients")
hist(temp$number_outpatient, main = "Histogram␣of␣Number␣of␣Outpatients",
      xlab = "number_outpatient")
hist(temp$number_outpatient, xlim = c(1,4), main = "Enlarged␣Histogram",
      xlab = "number_outpatient")  # DELETE >2
#dev.off()

# highest number of emergency visits in one year
temp[temp$number_emergency==76,]
sort(temp$number_emergency,decreasing = T)
#png(filename="cleandat4.png")  # save plot
#layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
min(boxplot(temp$number_emergency, horizontal = T,
               main = "Boxplot␣of␣Number␣of␣Emergency␣Visits")$out)
hist(temp$number_emergency, main = "Histogram␣of␣Number␣of␣Emergency␣Visits",
      xlab = "number_emergency")
hist(temp$number_emergency, xlim = c(2,8), main = "Enlarged␣Histogram",
      xlab = "number_emergency")  # DELETE >3
#dev.off()

# highest number of inpatient visits
temp[temp$number_inpatient==21,]
sort(temp$number_inpatient,decreasing = T)
#png(filename="cleandat5.png")  # save plot
#layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
min(boxplot(temp$number_inpatient, horizontal = T,
               main = "Boxplot␣of␣Number␣of␣Inpatients")$out)
hist(temp$number_inpatient, main = "Histogram␣of␣Number␣of␣Inpatients",
      xlab = "number_inpatient")
hist(temp$number_inpatient, xlim = c(1,5), main = "Enlarged␣Histogram",
    xlab = "number_inpatient")  # DELETE >=5
#dev.off()

# try fit different distributions
library(car)
#png(filename="fitdat1.png")  # save plot
par(mfrow=c(1,2))
qqPlot(temp$number_inpatient, dist = "lnorm", main = "Lognormal␣Fit␣QQPlot",
       ylab = "number_inpatient")  # log is not normally distributed
qqPlot(temp$number_inpatient, dist = "exp", main = "Exponential␣Fit␣QQPlot",
       ylab = "number_inpatient")  # lognormal still looks like better fit
#dev.off()

# fit exponential distribution
library(fitdistrplus)
#png(filename="fitdat2.png")  # save plot
par(mfrow=c(1,1))
plot(fitdist(temp$number_inpatient, "exp"))
#dev.off()

# explore
hist(temp$number_diagnoses)
sort(temp$number_diagnoses,decreasing = T)
plot(temp$age)
plot(temp$race)
plot(temp$race:temp$age, xlab = "Grouping␣of␣Race␣w.r.t.␣Age")
```