Predicting readmission probability for diabetes inpatients

Modern Data Mining

Contents


```
knitr::opts_chunk$set(
    echo = TRUE,
    fig.height = 4,
    fig.width = 7
)
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(caret, pROC, leaps, dplyr, ggplot2, glmnet, car, data.table)   #add your packages here
```

Executive Summary

Diabetes is well known in the United States, being a medical conditions that affect millions of people for a majority of their lives. In some cases, there are times in which readmission to hospitals may occur, which are extremely costly. Under such a pretense, 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.

Therefore, it would be beneficial to understand what factors heavily influence such an event, and help provide a prognosis on who may be at risk of being readmitted.

Using the readmission data set produced by a group in STAT 571/701, we will attempt to produce a model that could provide such a diagnosis. With over 30 numeric and categorical variables being considered, our goal is to create an understandable and accessible model that can help hospitals and patients limit such a costly experience.

This data set contains over 100000 observations of around 70000 unique patients who could have been readmitted to the hospital multiple times, and contains detailed information on their medical history, admission

and discharge details, patient demographics, and identifiers. Patients spend 1-14 days in the hospital, have a large range of up to 81 medications used, and may have had multiple procedures performed on them.

For our methodology, we first used various models, including Linear Modeling, and Linear Regression, in order to determine which one was the best one. By splitting it into training and test data with K-fold cross validation, we determined the validity and flexibility of our model as well.

Some issues that have come up during the project are similar to many others you may see - missing data which was ignored, but largely has been cleaned.

Analysis Process

Data Summary / EDA

Beginning with the data summary, as previously described, we continue with the readmission.csv data set. First, it would be beneficial to understand general information about each group.

One thing to note beforehand is the number of categorical variables that are found here. There are numerous anti-diabetic medications, race and gender variables, and diagnosis. Let's be sure to change them into factors.

Now, it becomes easier for us to determine overarching characteristics of each variable.

First, we take notice that there over 101,766 observations with 31 variables. The number of unique observations of the variable $patient_nbr$ also indicates that there at 71,518 unique patients in this data set. It becomes clear here that there are numerous readmission of the same patients in this dataset, which affects the way in which we approach our modeling.

Furthermore, Based on our own intuition, there are some particular variables that would be obvious candidates as having a strong influence for readmission. Some interesting characteristics to first note down from the summary of the data is that (i) the range of a hospital stay was between 1-14 days (ii) the number of medications ranged from 1 to 81 (iii) the number of and diagnosis ranged from 1 to 16

Note that each one of these variables showed variability in comparison to others.

We would also like to note some minor information is missing from some variables like race and gender, upon which we ignored and kept in the dataset.

Based on the goal of our study, we also modified our *readmitted* variable, so that the only two unique possible values are either "NO" or "<30", as ">30" would not result in a large cost for the hospital.

Simplicity is key within our models, so it would beneficial to continue to look at the other variables and determine which ones may be better to eliminate.

We will first remove $encounter_id$ and $patient_nbr$ for the sake of our analysis, as these are just identifiers that should not be included in the analysis.

```
readmis_cln = subset(readmis, select=-c(encounter_id, patient_nbr))
readmis_cln$readmitted <- as.numeric(readmis_cln$readmitted)
```

Analysis

To begin with capturing important factors, let us start off with a linear model to get quick results.

```
fit.first.lm <- lm(readmitted ~., readmis_cln)
summary(fit.first.lm)
```

```
##
## Call:
## lm(formula = readmitted ~ ., data = readmis_cln)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01238  0.05835  0.08955  0.12673  0.85873
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                     1.955e+00
## raceAfricanAmerican                                            -1.065e-02
## raceAsian                                                      -1.290e-02
## raceCaucasian                                                  -1.319e-02
## raceHispanic                                                   -1.187e-02
## raceOther                                                      -5.317e-03
## genderMale                                                     -2.035e-03
## genderUnknown/Invalid                                           7.545e-02
## time_in_hospital                                               -4.327e-04
## num_lab_procedures                                             -4.478e-05
## num_procedures                                                  1.374e-03
## num_medications                                                -4.303e-04
## number_outpatient                                               1.076e-03
## number_emergency                                               -6.667e-03
## number_inpatient                                               -3.633e-02
## number_diagnoses                                               -9.010e-04
## max_glu_serum>300                                              -8.666e-04
## max_glu_serumNone                                               1.272e-02
## max_glu_serumNorm                                               1.495e-03
## A1Cresult>8                                                     2.355e-03
## A1CresultNone                                                  -7.178e-03
## A1CresultNorm                                                   1.482e-03
## metforminNo                                                     1.565e-02
## metforminSteady                                                 2.396e-02
## metforminUp                                                     3.465e-02
## glimepirideNo                                                   1.273e-02
## glimepirideSteady                                               2.314e-02
## glimepirideUp                                                   1.073e-02
## glipizideNo                                                     3.189e-02
## glipizideSteady                                                 3.074e-02
## glipizideUp                                                     1.596e-02
## glyburideNo                                                    -2.175e-02
## glyburideSteady                                                -2.053e-02
## glyburideUp                                                    -1.750e-02
## pioglitazoneNo                                                  2.893e-02
## pioglitazoneSteady                                              3.392e-02
## pioglitazoneUp                                                  2.044e-02
## rosiglitazoneNo                                                -4.987e-02
## rosiglitazoneSteady                                            -4.570e-02
## rosiglitazoneUp                                                -4.082e-02
## insulinNo                                                       1.099e-02
## insulinSteady                                                   1.397e-02
## insulinUp                                                       8.547e-03
## changeNo                                                       -1.407e-03
```

```
## diabetesMedYes                                                  -1.987e-02
## disch_disp_modifiedDischarged to home with Home Health Service -1.611e-02
## disch_disp_modifiedDischarged/Transferred to SNF               -3.687e-02
## disch_disp_modifiedOther                                        -3.918e-02
## adm_src_modOther                                                 6.390e-03
## adm_src_modPhysician Referral                                   -4.806e-03
## adm_src_modTransfer from Home Health                             1.290e-02
## adm_typ_modEmergency                                            -6.379e-03
## adm_typ_modOther                                                -7.146e-04
## adm_typ_modUrgent                                               -5.108e-03
## age_mod20-59                                                    -3.564e-02
## age_mod60-79                                                    -4.539e-02
## age_mod80+                                                      -4.255e-02
## diag1_mod250.8                                                   6.082e-02
## diag1_mod276                                                     3.112e-02
## diag1_mod38                                                      6.349e-02
## diag1_mod410                                                     3.985e-02
## diag1_mod414                                                     4.302e-02
## diag1_mod427                                                     5.238e-02
## diag1_mod428                                                     3.097e-02
## diag1_mod434                                                    -6.860e-03
## diag1_mod435                                                     6.251e-02
## diag1_mod486                                                     7.406e-02
## diag1_mod491                                                     4.390e-02
## diag1_mod493                                                     5.690e-02
## diag1_mod518                                                     8.305e-02
## diag1_mod577                                                     2.744e-02
## diag1_mod584                                                     4.255e-02
## diag1_mod599                                                     6.620e-02
## diag1_mod682                                                     6.581e-02
## diag1_mod715                                                     3.761e-02
## diag1_mod780                                                     5.685e-02
## diag1_mod786                                                     5.950e-02
## diag1_mod820                                                     1.012e-02
## diag1_mod996                                                     4.385e-02
## diag1_modOther                                                   4.258e-02
## diag2_mod250.01                                                 -3.203e-02
## diag2_mod250.02                                                 -1.253e-02
## diag2_mod276                                                    -1.426e-02
## diag2_mod285                                                     1.486e-02
## diag2_mod401                                                     3.615e-03
## diag2_mod403                                                    -1.928e-02
## diag2_mod411                                                    -8.968e-03
## diag2_mod413                                                     3.163e-03
## diag2_mod414                                                    -5.029e-04
## diag2_mod424                                                    -8.793e-03
## diag2_mod425                                                    -1.023e-02
## diag2_mod427                                                    -6.077e-03
## diag2_mod428                                                    -1.325e-02
## diag2_mod486                                                     1.182e-02
## diag2_mod491                                                    -2.208e-02
## diag2_mod496                                                    -3.570e-03
## diag2_mod518                                                     1.528e-02
## diag2_mod584                                                     2.985e-03
```

```
## diag2_mod585                                -1.377e-02
## diag2_mod599                                 6.288e-04
## diag2_mod682                                -1.852e-02
## diag2_mod707                                -1.829e-02
## diag2_mod780                                 2.826e-03
## diag2_modOther                              -1.142e-02
## diag3_mod250                                -2.068e-03
## diag3_mod250.02                             -2.694e-02
## diag3_mod250.6                              -6.343e-02
## diag3_mod272                                 6.081e-04
## diag3_mod276                                -7.872e-03
## diag3_mod285                                -2.705e-04
## diag3_mod401                                -1.706e-03
## diag3_mod403                                -3.749e-02
## diag3_mod414                                 4.913e-04
## diag3_mod424                                -1.189e-03
## diag3_mod425                                -4.748e-03
## diag3_mod427                                -7.635e-03
## diag3_mod428                                -8.888e-03
## diag3_mod496                                -2.090e-02
## diag3_mod585                                -3.540e-02
## diag3_mod599                                -7.677e-03
## diag3_mod707                                -2.151e-02
## diag3_mod780                                -8.255e-03
## diag3_modOther                              -1.451e-02
## diag3_modV45                                 6.112e-03
##                                            Std. Error
## (Intercept)                                 5.898e-02
## raceAfricanAmerican                         6.948e-03
## raceAsian                                   1.389e-02
## raceCaucasian                               6.660e-03
## raceHispanic                                9.525e-03
## raceOther                                   1.034e-02
## genderMale                                  1.993e-03
## genderUnknown/Invalid                       1.785e-01
## time_in_hospital                            4.114e-04
## num_lab_procedures                          5.894e-05
## num_procedures                              7.429e-04
## num_medications                             1.643e-04
## number_outpatient                           7.841e-04
## number_emergency                            1.094e-03
## number_inpatient                            8.267e-04
## number_diagnoses                            6.514e-04
## max_glu_serum>300                           1.190e-02
## max_glu_serumNone                           9.117e-03
## max_glu_serumNorm                           1.009e-02
## A1Cresult>8                                 6.133e-03
## A1CresultNone                               5.183e-03
## A1CresultNorm                               6.663e-03
## metforminNo                                 1.313e-02
## metforminSteady                             1.314e-02
## metforminUp                                 1.602e-02
## glimepirideNo                               2.239e-02
## glimepirideSteady                           2.269e-02
```

```
## glimepirideUp                                                    2.803e-02
## glipizideNo                                                      1.339e-02
## glipizideSteady                                                  1.344e-02
## glipizideUp                                                      1.719e-02
## glyburideNo                                                      1.342e-02
## glyburideSteady                                                  1.349e-02
## glyburideUp                                                      1.697e-02
## pioglitazoneNo                                                   2.855e-02
## pioglitazoneSteady                                               2.872e-02
## pioglitazoneUp                                                   3.492e-02
## rosiglitazoneNo                                                  3.324e-02
## rosiglitazoneSteady                                              3.341e-02
## rosiglitazoneUp                                                  4.047e-02
## insulinNo                                                        5.165e-03
## insulinSteady                                                    4.008e-03
## insulinUp                                                        4.048e-03
## changeNo                                                         3.667e-03
## diabetesMedYes                                                   3.534e-03
## disch_disp_modifiedDischarged to home with Home Health Service   3.168e-03
## disch_disp_modifiedDischarged/Transferred to SNF                 3.288e-03
## disch_disp_modifiedOther                                         2.984e-03
## adm_src_modOther                                                 4.093e-03
## adm_src_modPhysician Referral                                    3.482e-03
## adm_src_modTransfer from Home Health                             6.122e-03
## adm_typ_modEmergency                                             4.019e-03
## adm_typ_modOther                                                 5.220e-03
## adm_typ_modUrgent                                                3.503e-03
## age_mod20-59                                                     1.141e-02
## age_mod60-79                                                     1.149e-02
## age_mod80+                                                       1.167e-02
## diag1_mod250.8                                                   1.180e-02
## diag1_mod276                                                     1.163e-02
## diag1_mod38                                                      1.193e-02
## diag1_mod410                                                     1.086e-02
## diag1_mod414                                                     1.067e-02
## diag1_mod427                                                     1.099e-02
## diag1_mod428                                                     1.001e-02
## diag1_mod434                                                     1.155e-02
## diag1_mod435                                                     1.343e-02
## diag1_mod486                                                     1.066e-02
## diag1_mod491                                                     1.135e-02
## diag1_mod493                                                     1.329e-02
## diag1_mod518                                                     1.328e-02
## diag1_mod577                                                     1.323e-02
## diag1_mod584                                                     1.221e-02
## diag1_mod599                                                     1.205e-02
## diag1_mod682                                                     1.141e-02
## diag1_mod715                                                     1.178e-02
## diag1_mod780                                                     1.151e-02
## diag1_mod786                                                     1.051e-02
## diag1_mod820                                                     1.336e-02
## diag1_mod996                                                     1.158e-02
## diag1_modOther                                                   9.264e-03
## diag2_mod250.01                                                  9.006e-03
```

```
## diag2_mod250.02                                                 8.084e-03
## diag2_mod276                                                    5.986e-03
## diag2_mod285                                                    9.267e-03
## diag2_mod401                                                    6.898e-03
## diag2_mod403                                                    7.683e-03
## diag2_mod411                                                    8.624e-03
## diag2_mod413                                                    1.101e-02
## diag2_mod414                                                    7.708e-03
## diag2_mod424                                                    1.053e-02
## diag2_mod425                                                    9.519e-03
## diag2_mod427                                                    6.360e-03
## diag2_mod428                                                    6.061e-03
## diag2_mod486                                                    9.678e-03
## diag2_mod491                                                    9.460e-03
## diag2_mod496                                                    7.056e-03
## diag2_mod518                                                    9.668e-03
## diag2_mod584                                                    9.044e-03
## diag2_mod585                                                    8.660e-03
## diag2_mod599                                                    7.191e-03
## diag2_mod682                                                    9.545e-03
## diag2_mod707                                                    8.498e-03
## diag2_mod780                                                    9.184e-03
## diag2_modOther                                                  4.749e-03
## diag3_mod250                                                    9.420e-03
## diag3_mod250.02                                                 1.241e-02
## diag3_mod250.6                                                  1.314e-02
## diag3_mod272                                                    1.138e-02
## diag3_mod276                                                    1.014e-02
## diag3_mod285                                                    1.277e-02
## diag3_mod401                                                    9.487e-03
## diag3_mod403                                                    1.131e-02
## diag3_mod414                                                    1.049e-02
## diag3_mod424                                                    1.331e-02
## diag3_mod425                                                    1.309e-02
## diag3_mod427                                                    1.050e-02
## diag3_mod428                                                    1.036e-02
## diag3_mod496                                                    1.107e-02
## diag3_mod585                                                    1.174e-02
## diag3_mod599                                                    1.165e-02
## diag3_mod707                                                    1.259e-02
## diag3_mod780                                                    1.243e-02
## diag3_modOther                                                  9.194e-03
## diag3_modV45                                                    1.243e-02
##                                                               t value Pr(>|t|)
## (Intercept)                                                    33.142  < 2e-16
## raceAfricanAmerican                                            -1.532 0.125406
## raceAsian                                                      -0.928 0.353190
## raceCaucasian                                                  -1.981 0.047648
## raceHispanic                                                   -1.247 0.212552
## raceOther                                                      -0.514 0.606985
## genderMale                                                     -1.021 0.307309
## genderUnknown/Invalid                                           0.423 0.672604
## time_in_hospital                                               -1.052 0.292915
## num_lab_procedures                                             -0.760 0.447466
```

```
## num_procedures                                                    1.850 0.064307
## num_medications                                                  -2.619 0.008815
## number_outpatient                                                 1.372 0.169941
## number_emergency                                                 -6.093 1.11e-09
## number_inpatient                                                -43.950  < 2e-16
## number_diagnoses                                                 -1.383 0.166594
## max_glu_serum>300                                                -0.073 0.941963
## max_glu_serumNone                                                 1.395 0.163025
## max_glu_serumNorm                                                 0.148 0.882163
## A1Cresult>8                                                       0.384 0.700998
## A1CresultNone                                                    -1.385 0.166090
## A1CresultNorm                                                     0.222 0.824018
## metforminNo                                                       1.192 0.233452
## metforminSteady                                                   1.824 0.068192
## metforminUp                                                       2.163 0.030570
## glimepirideNo                                                     0.569 0.569469
## glimepirideSteady                                                 1.020 0.307903
## glimepirideUp                                                     0.383 0.701975
## glipizideNo                                                       2.381 0.017247
## glipizideSteady                                                   2.287 0.022173
## glipizideUp                                                       0.928 0.353189
## glyburideNo                                                      -1.621 0.105054
## glyburideSteady                                                  -1.522 0.128051
## glyburideUp                                                      -1.032 0.302307
## pioglitazoneNo                                                    1.013 0.310905
## pioglitazoneSteady                                                1.181 0.237592
## pioglitazoneUp                                                    0.585 0.558305
## rosiglitazoneNo                                                  -1.500 0.133506
## rosiglitazoneSteady                                              -1.368 0.171336
## rosiglitazoneUp                                                  -1.009 0.313180
## insulinNo                                                         2.128 0.033328
## insulinSteady                                                     3.487 0.000489
## insulinUp                                                         2.111 0.034747
## changeNo                                                         -0.384 0.701190
## diabetesMedYes                                                   -5.623 1.88e-08
## disch_disp_modifiedDischarged to home with Home Health Service   -5.086 3.67e-07
## disch_disp_modifiedDischarged/Transferred to SNF                -11.212  < 2e-16
## disch_disp_modifiedOther                                        -13.130  < 2e-16
## adm_src_modOther                                                  1.561 0.118504
## adm_src_modPhysician Referral                                    -1.380 0.167527
## adm_src_modTransfer from Home Health                              2.108 0.035060
## adm_typ_modEmergency                                             -1.587 0.112482
## adm_typ_modOther                                                 -0.137 0.891122
## adm_typ_modUrgent                                                -1.458 0.144809
## age_mod20-59                                                     -3.124 0.001784
## age_mod60-79                                                     -3.951 7.80e-05
## age_mod80+                                                       -3.646 0.000266
## diag1_mod250.8                                                    5.154 2.56e-07
## diag1_mod276                                                      2.675 0.007472
## diag1_mod38                                                       5.321 1.03e-07
## diag1_mod410                                                      3.668 0.000244
## diag1_mod414                                                      4.031 5.55e-05
## diag1_mod427                                                      4.767 1.88e-06
## diag1_mod428                                                      3.093 0.001982
```

```
## diag1_mod434                                -0.594 0.552684
## diag1_mod435                                 4.654 3.26e-06
## diag1_mod486                                 6.950 3.68e-12
## diag1_mod491                                 3.868 0.000110
## diag1_mod493                                 4.282 1.85e-05
## diag1_mod518                                 6.252 4.07e-10
## diag1_mod577                                 2.074 0.038054
## diag1_mod584                                 3.485 0.000492
## diag1_mod599                                 5.492 3.99e-08
## diag1_mod682                                 5.766 8.12e-09
## diag1_mod715                                 3.193 0.001410
## diag1_mod780                                 4.941 7.80e-07
## diag1_mod786                                 5.664 1.49e-08
## diag1_mod820                                 0.757 0.448842
## diag1_mod996                                 3.786 0.000153
## diag1_modOther                               4.596 4.31e-06
## diag2_mod250.01                             -3.556 0.000377
## diag2_mod250.02                             -1.550 0.121235
## diag2_mod276                                -2.382 0.017236
## diag2_mod285                                 1.604 0.108822
## diag2_mod401                                 0.524 0.600264
## diag2_mod403                                -2.509 0.012093
## diag2_mod411                                -1.040 0.298359
## diag2_mod413                                 0.287 0.773814
## diag2_mod414                                -0.065 0.947974
## diag2_mod424                                -0.835 0.403582
## diag2_mod425                                -1.075 0.282335
## diag2_mod427                                -0.955 0.339329
## diag2_mod428                                -2.186 0.028828
## diag2_mod486                                 1.221 0.221926
## diag2_mod491                                -2.334 0.019593
## diag2_mod496                                -0.506 0.612883
## diag2_mod518                                 1.581 0.113937
## diag2_mod584                                 0.330 0.741341
## diag2_mod585                                -1.590 0.111763
## diag2_mod599                                 0.087 0.930320
## diag2_mod682                                -1.940 0.052341
## diag2_mod707                                -2.152 0.031362
## diag2_mod780                                 0.308 0.758352
## diag2_modOther                              -2.404 0.016198
## diag3_mod250                                -0.220 0.826201
## diag3_mod250.02                             -2.170 0.029976
## diag3_mod250.6                              -4.829 1.38e-06
## diag3_mod272                                 0.053 0.957384
## diag3_mod276                                -0.776 0.437565
## diag3_mod285                                -0.021 0.983099
## diag3_mod401                                -0.180 0.857291
## diag3_mod403                                -3.315 0.000917
## diag3_mod414                                 0.047 0.962654
## diag3_mod424                                -0.089 0.928836
## diag3_mod425                                -0.363 0.716835
## diag3_mod427                                -0.727 0.467139
## diag3_mod428                                -0.858 0.390753
## diag3_mod496                                -1.888 0.058985
```

```
## diag3_mod585                                                    -3.016 0.002560
## diag3_mod599                                                    -0.659 0.509864
## diag3_mod707                                                    -1.708 0.087602
## diag3_mod780                                                    -0.664 0.506760
## diag3_modOther                                                  -1.578 0.114494
## diag3_modV45                                                     0.492 0.622987
##
## (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
## metforminNo
## metforminSteady                                                .
## metforminUp                                                    *
## glimepirideNo
## glimepirideSteady
## glimepirideUp
## glipizideNo                                                    *
## glipizideSteady                                                *
## glipizideUp
## glyburideNo
## glyburideSteady
## glyburideUp
## pioglitazoneNo
## pioglitazoneSteady
## pioglitazoneUp
## rosiglitazoneNo
## rosiglitazoneSteady
## rosiglitazoneUp
## 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                                              ***
## diag2_mod250.01                                             ***
## diag2_mod250.02
## diag2_mod276                                                *
## diag2_mod285
## diag2_mod401
## diag2_mod403                                                *
## diag2_mod411
## diag2_mod413
## diag2_mod414
## diag2_mod424
## diag2_mod425
## diag2_mod427
## diag2_mod428                                                *
## diag2_mod486
## diag2_mod491                                                *
## diag2_mod496
## diag2_mod518
## diag2_mod584
## diag2_mod585
## diag2_mod599
## diag2_mod682                                                .
```

```
## diag2_mod707                                                  *
## diag2_mod780
## diag2_modOther                                                *
## diag3_mod250
## diag3_mod250.02                                               *
## diag3_mod250.6                                                ***
## diag3_mod272
## diag3_mod276
## diag3_mod285
## diag3_mod401
## diag3_mod403                                                  ***
## diag3_mod414
## diag3_mod424
## diag3_mod425
## diag3_mod427
## diag3_mod428
## diag3_mod496                                                  .
## diag3_mod585                                                  **
## diag3_mod599
## diag3_mod707                                                  .
## diag3_mod780
## diag3_modOther
## diag3_modV45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.309 on 101642 degrees of freedom
## Multiple R-squared:  0.03786,    Adjusted R-squared:  0.0367
## F-statistic: 32.52 on 123 and 101642 DF,  p-value: < 2.2e-16
```

```
Anova(fit.first.lm)
```

```
## Anova Table (Type II tests)
##
## Response: readmitted
##                       Sum Sq     Df   F value    Pr(>F)
## race                     0.5      5    1.0680  0.375794
## gender                   0.1      2    0.6130  0.541728
## time_in_hospital         0.1      1    1.1062  0.292915
## num_lab_procedures       0.1      1    0.5771  0.447466
## num_procedures           0.3      1    3.4227  0.064307 .
## num_medications          0.7      1    6.8603  0.008815 **
## number_outpatient        0.2      1    1.8835  0.169941
## number_emergency         3.5      1   37.1208 1.114e-09 ***
## number_inpatient       184.5      1 1931.6464 < 2.2e-16 ***
## number_diagnoses         0.2      1    1.9134  0.166594
## max_glu_serum            0.4      3    1.4839  0.216645
## A1Cresult                1.0      3    3.3635  0.017816 *
## metformin                1.2      3    4.0185  0.007197 **
## glimepiride              0.5      3    1.6351  0.178867
## glipizide                0.7      3    2.4342  0.062858 .
## glyburide                0.3      3    0.9022  0.439086
## pioglitazone             0.3      3    0.9217  0.429269
## rosiglitazone            0.3      3    1.0691  0.360742
```

```
## insulin                1.3      3    4.4429  0.003979 **
## change                 0.0      1    0.1472  0.701190
## diabetesMed            3.0      1   31.6207 1.879e-08 ***
## disch_disp_modified   22.4      3   78.2889 < 2.2e-16 ***
## adm_src_mod            1.3      3    4.4581  0.003895 **
## adm_typ_mod            0.3      3    1.1618  0.322669
## age_mod                2.7      3    9.5095 2.818e-06 ***
## diag1_mod             19.8     23    9.0094 < 2.2e-16 ***
## diag2_mod              6.3     24    2.7531 8.497e-06 ***
## diag3_mod              9.2     20    4.8157 5.805e-12 ***
## Residuals           9707.6 101642
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It becomes clear here that there are numerous variables in which we must reduce from - the simpler the model, the better. We continue to try to narrow down which important factors will capture the chance of readmission.

Conclusion

From the *Goals* section above, your study should respond to the following:

Analyses suggested

1) Identify important factors that capture the chance of a readmission within 30 days.

Before anything, lets separate our training, validation and testing datasets.

```
# Split the data:
N <- length(readmis_cln$readmitted)
n1 <- floor(.6*N)
n2 <- floor(.2*N)
set.seed(10)

# Split data to three portions of .6, .2 and .2 of data size N
idx_train <- sample(N, n1)
idx_no_train <- (which(! seq(1:N) %in% idx_train))
idx_test <- sample( idx_no_train, n2)
idx_val <- which(! idx_no_train %in% idx_test)
data.train <- readmis_cln[idx_train,]
data.test <- readmis_cln[idx_test,]
data.val <- readmis_cln[idx_val,]
```

```
readmis_cln <- data.train
```

First we will attempt removing some of the variables that seemed non-important according the the lm model, as shows by the Anova above. However, in order to understand whether we can indeed remove them we will make a simple anova() to compare the this reduced model with the complete one. We are using a high threshold of .2 to remove the variables.

```
readmis_cln.1 <- readmis_cln
readmis_cln.1$readmitted <- readmis_cln$readmitted-1
readmis_cln.2 = subset(readmis_cln.1, select=-c(race, time_in_hospital, gender, num_lab_procedures, max_
fit.logit.1 <- glm(readmitted ~., readmis_cln.1, family=binomial(logit))
fit.logit.2 <- glm(readmitted ~., readmis_cln.2, family=binomial(logit))
anova(fit.logit.1, fit.logit.2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: readmitted ~ race + gender + time_in_hospital + num_lab_procedures +
##     num_procedures + num_medications + number_outpatient + number_emergency +
##     number_inpatient + number_diagnoses + max_glu_serum + A1Cresult +
##     metformin + glimepiride + glipizide + glyburide + pioglitazone +
##     rosiglitazone + insulin + change + diabetesMed + disch_disp_modified +
##     adm_src_mod + adm_typ_mod + age_mod + diag1_mod + diag2_mod +
##     diag3_mod
## Model 2: readmitted ~ num_procedures + num_medications + number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     A1Cresult + metformin + glimepiride + glipizide + insulin +
##     diabetesMed + disch_disp_modified + adm_src_mod + age_mod +
##     diag1_mod + diag2_mod + diag3_mod
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1     60935      40680
## 2     60960      40703 -25  -22.465   0.6087
```

As we can see, there's no evidence that the removed variables were important as Pr(>Chi)=.23. So lets further simplify our model, maintaining these excluded variables out. Our current model has 19 variables. Now we can run the full Chisquare Anova() and further explore other non-important variables.

```
Anova(fit.logit.2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: readmitted
##                       LR Chisq Df Pr(>Chisq)
## num_procedures            3.66  1  0.0558437 .
## num_medications           3.21  1  0.0733733 .
## number_outpatient         0.26  1  0.6081440
## number_emergency         10.72  1  0.0010606 **
## number_inpatient        794.79  1  < 2.2e-16 ***
## number_diagnoses          2.71  1  0.0996328 .
## A1Cresult                 8.21  3  0.0418079 *
## metformin                14.00  3  0.0029091 **
## glimepiride               8.94  3  0.0300578 *
## glipizide                 5.08  3  0.1659159
## insulin                  12.28  3  0.0064861 **
## diabetesMed              28.27  1  1.053e-07 ***
## disch_disp_modified     149.86  3  < 2.2e-16 ***
## adm_src_mod               7.85  3  0.0491076 *
## age_mod                  14.42  3  0.0023883 **
## diag1_mod               145.77 23  < 2.2e-16 ***
## diag2_mod                58.57 24  0.0001013 ***
```

```
## diag3_mod              67.10 20  5.384e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here *number_outpatient* and *glimepiride* seems to be rather unimportant, but lets make sure we can remove both of them when comparing with the full model:

```
readmis_cln.3 <- subset(readmis_cln.2, select=-c(glimepiride, number_outpatient))
fit.logit.3 <- glm(readmitted ~., readmis_cln.3, family=binomial(logit))
anova(fit.logit.2, fit.logit.3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: readmitted ~ num_procedures + num_medications + number_outpatient +
##     number_emergency + number_inpatient + number_diagnoses +
##     A1Cresult + metformin + glimepiride + glipizide + insulin +
##     diabetesMed + disch_disp_modified + adm_src_mod + age_mod +
##     diag1_mod + diag2_mod + diag3_mod
## Model 2: readmitted ~ num_procedures + num_medications + number_emergency +
##     number_inpatient + number_diagnoses + A1Cresult + metformin +
##     glipizide + insulin + diabetesMed + disch_disp_modified +
##     adm_src_mod + age_mod + diag1_mod + diag2_mod + diag3_mod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     60960      40703
## 2     60964      40712 -4  -9.1724  0.05693 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit.logit.1, fit.logit.3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: readmitted ~ race + gender + time_in_hospital + num_lab_procedures +
##     num_procedures + num_medications + number_outpatient + number_emergency +
##     number_inpatient + number_diagnoses + max_glu_serum + A1Cresult +
##     metformin + glimepiride + glipizide + glyburide + pioglitazone +
##     rosiglitazone + insulin + change + diabetesMed + disch_disp_modified +
##     adm_src_mod + adm_typ_mod + age_mod + diag1_mod + diag2_mod +
##     diag3_mod
## Model 2: readmitted ~ num_procedures + num_medications + number_emergency +
##     number_inpatient + number_diagnoses + A1Cresult + metformin +
##     glipizide + insulin + diabetesMed + disch_disp_modified +
##     adm_src_mod + age_mod + diag1_mod + diag2_mod + diag3_mod
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1     60935      40680
## 2     60964      40712 -29  -31.638    0.336
```

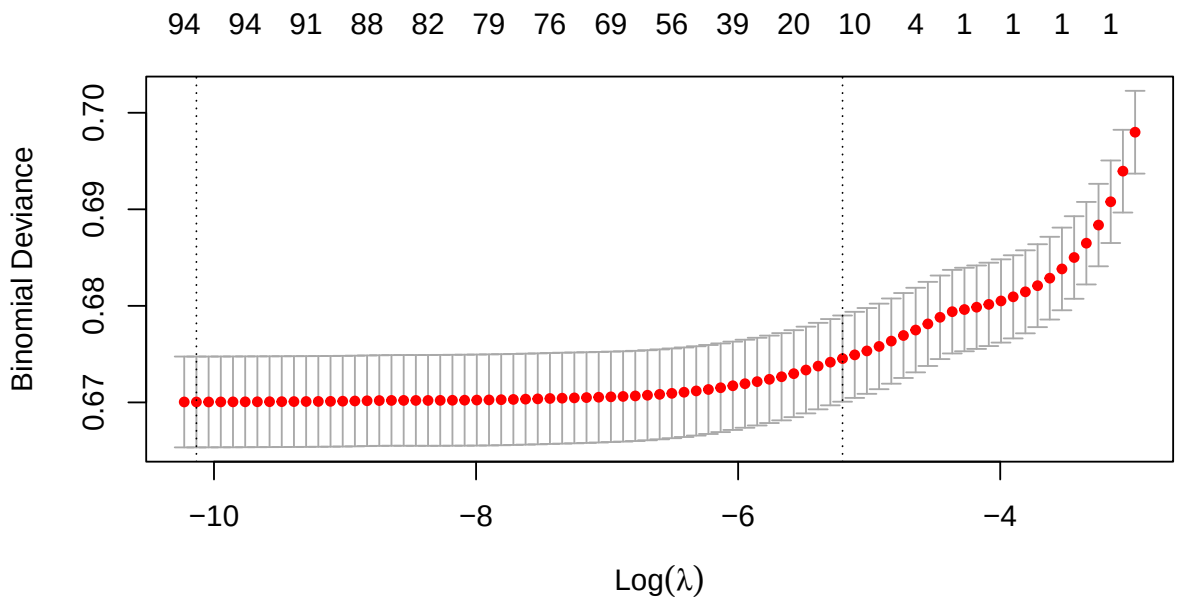
Observe that even when comparing with the full model (made with *readmis_cln* with 29 variables) there is no evidence support keeping all the variables we removed. So now our subset of variables is *readmis_cln.3* with 17 variables.

Now, lets run a Lasso model to understand which variables indeed contribute the most with our model

```
X <- model.matrix(readmitted~., data=readmis_cln.3)[,-1] # for each factor: num of levels -1
dim(X)
```

```
## [1] 61059    94
```

```
Y <- readmis_cln.3$readmitted
set.seed(10) # to have same sets of K folds
fit1.cv <- cv.glmnet(X, Y, alpha=1, family="binomial", nfolds = 10, type.measure = "deviance")
plot(fit1.cv)
```




```
fit1.cv$lambda.1se
```

```
## [1] 0.005497939
```

I am interested in the model as much reduced as possible, therefore the one with highest *Lambda*, so I'll take the one for 1se.

```
coef.min <- coef(fit1.cv, s="lambda.1se")  #s=c("lambda.1se","lambda.min") or lambda value
coef.min <- coef.min[which(coef.min !=0),]   # get the non=zero coefficients
var.min <- rownames(as.matrix(coef.min))[-1]
```

Since our original data is composed for factor variables, the Lasso response gives the dummy for each factor level. So simplifying we have that the 1se variables in the lasso are:

```
varslasso <- c("num_medications", "number_emergency", "number_inpatient", "number_diagnoses", "A1Cresul
          "diag1_mod", "diag2_mod", "diag3_mod")
```

In other words, the 1se Lasso lambda excluded the following variables:

```
excluded <- readmis_cln.3 %>% select(-varslasso, -"readmitted")
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(varslasso)` instead of `varslasso` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
excluded <- names(excluded)
excluded
```

```
## [1] "num_procedures" "glipizide"      "adm_src_mod"
```

The set of available predictors is not limited to the raw variables in the data set. You may engineer any factors using the data, that you think will improve your model's quality.

2) For the purpose of classification, propose a model that can be used to predict whether a patient will be a readmit within 30 days. Justify your choice. Hint: use a decision criterion, such as AUC, to choose among a few candidate models.

So based on the lasso results I will make 5 models: - one of 13 variables excluding these 3 - 3 models of 14 variables, adding each one individually - one model of 16 variables

With these 5 models, I will compute the AUC for each and pick the best one (highest AIC)

```
readmis_fin.1 <- readmis_cln.3 %>% select(-excluded)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(excluded)` instead of `excluded` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
readmis_fin.2 <- readmis_cln.3 %>% select(c(varslasso, "num_procedures", "readmitted"))
readmis_fin.3 <- readmis_cln.3 %>% select(c(varslasso, "glipizide", "readmitted"))
readmis_fin.4 <- readmis_cln.3 %>% select(c(varslasso, "adm_src_mod", "readmitted"))
readmis_fin.5 <- readmis_cln.3

fit.fin.1 <- glm(readmitted ~., readmis_fin.1, family=binomial(logit))
fit.fin.2 <- glm(readmitted ~., readmis_fin.2, family=binomial(logit))
fit.fin.3 <- glm(readmitted ~., readmis_fin.3, family=binomial(logit))
fit.fin.4 <- glm(readmitted ~., readmis_fin.4, family=binomial(logit))
fit.fin.5 <- glm(readmitted ~., readmis_fin.5, family=binomial(logit))

fit1.fitted.test <- predict(fit.fin.1, data.test, type="response")
fit2.fitted.test <- predict(fit.fin.2, data.test, type="response")
fit3.fitted.test <- predict(fit.fin.3, data.test, type="response")
fit4.fitted.test <- predict(fit.fin.4, data.test, type="response")
fit5.fitted.test <- predict(fit.fin.5, data.test, type="response")

fit1.test.auc <- auc(data.test$readmitted, fit1.fitted.test)
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
fit2.test.auc <- auc(data.test$readmitted, fit2.fitted.test)
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
fit3.test.auc <- auc(data.test$readmitted, fit3.fitted.test)
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
fit4.test.auc <- auc(data.test$readmitted, fit4.fitted.test)
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
fit5.test.auc <- auc(data.test$readmitted, fit5.fitted.test)
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
fit1.test.auc
```

```
## Area under the curve: 0.6484
```

```
fit2.test.auc
```

```
## Area under the curve: 0.6486
```

```
fit3.test.auc
```

```
## Area under the curve: 0.6488
```

```
fit4.test.auc
```

```
## Area under the curve: 0.6482
```

```
fit5.test.auc
```

```
## Area under the curve: 0.6489
```

We can see the the AUC changes very little between models. Thus, I will pick the one with the least variables - the simpler model

```
fit1.test.roc <- roc(data.test$readmitted, fit1.fitted.test)
```

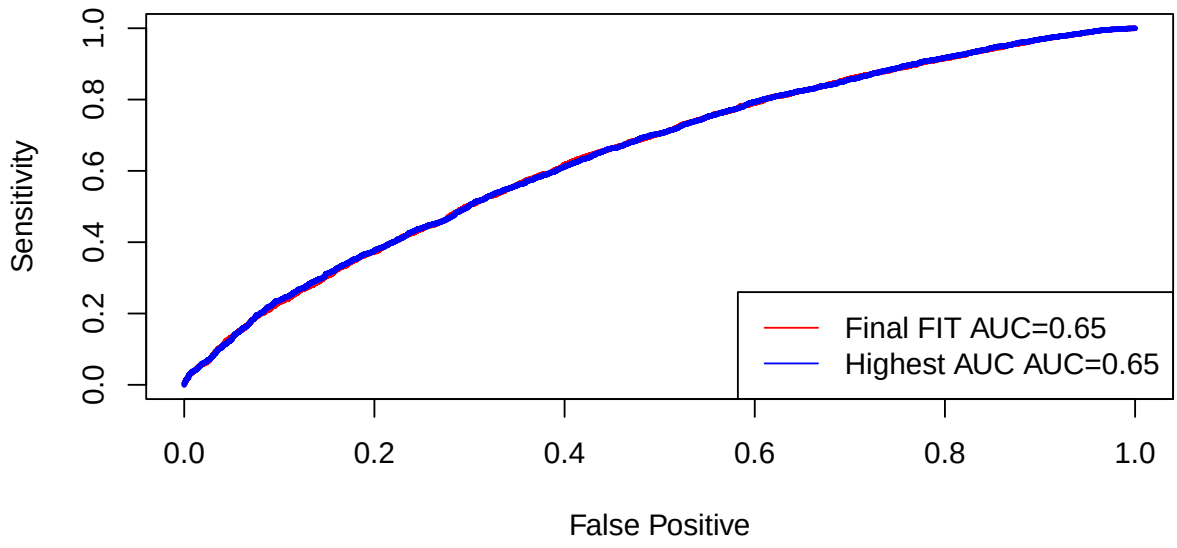
```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```

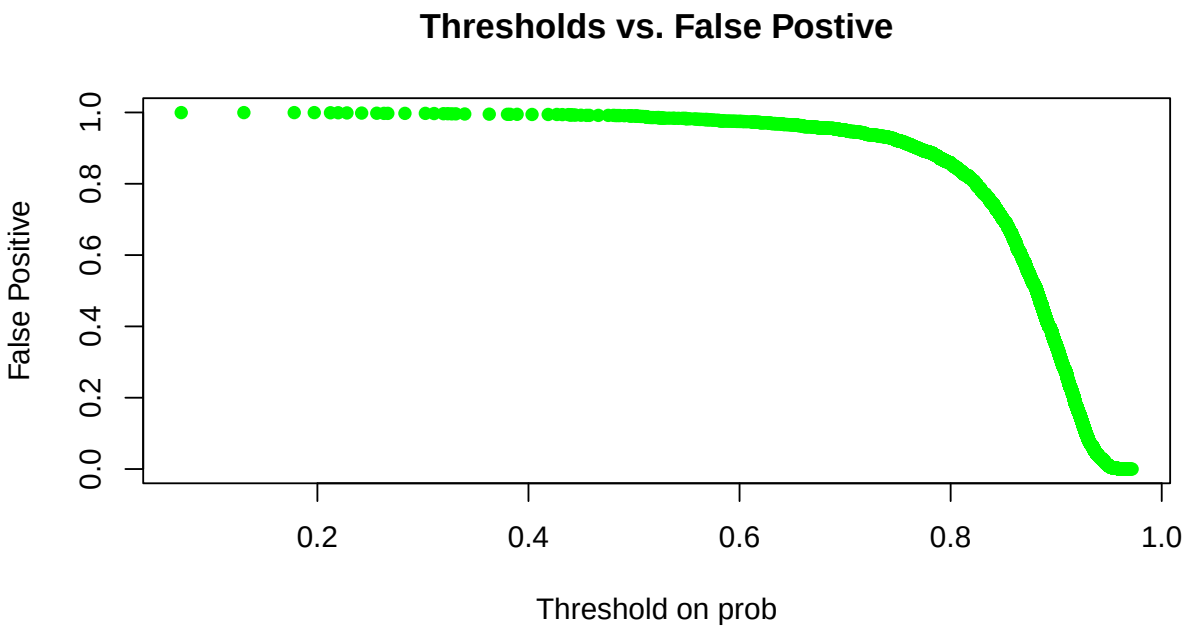
```
fit5.test.roc <- roc(data.test$readmitted, fit5.fitted.test)
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
plot(1-fit1.test.roc$specificities,
     fit1.test.roc$sensitivities, col="red", lwd=3, type="l",
     xlab="False Positive", ylab="Sensitivity")
lines(1-fit5.test.roc$specificities, fit5.test.roc$sensitivities, col="blue", lwd=3)
legend("bottomright",
       c(paste0("Final FIT AUC=", round(fit1.test.roc$auc,2)),
         paste0("Highest AUC AUC=", round(fit5.test.roc$auc, 2))),
         col=c("red", "blue"), lty=1)
```




```
plot(fit1.test.roc$thresholds, 1-fit1.test.roc$specificities, col="green", pch=16,
xlab="Threshold on prob",
ylab="False Positive",
main = "Thresholds vs. False Postive")
```



We can observe by the ROC that both models should perform virtually the same in all levels of threshholds. So picking the simplest one is the better call here, i.e. *fit.fin.1*

3) Based on a quick and somewhat arbitrary guess, we estimate **it costs twice as much** to mislabel a readmission than it does to mislabel a non-readmission. Based on this risk ratio, propose a specific classification rule to minimize the cost. If you find any information that could provide a better cost estimate, please justify it in your write-up and use the better estimate in your answer.

Based on Bayes' rule, we can stablish a threshhold of *33.33%* probability to determine a readmission. with the model being *fit.fin.1* as mentioned in the prior item.

```
ratio_of_costs <- 1/2
P <- ratio_of_costs/(1+ratio_of_costs)
P
```

```
## [1] 0.3333333
```

Now, let's test how good our model is based on that threshold, and testing it against the validation data set.

```
fit1.fitted.validation <- predict(fit.fin.1, data.test, type="response")

fit1.pred.33 <- ifelse(fit1.fitted.validation > 1/3, "2", "1")

confusionMatrix(data = as.factor(fit1.pred.33), # predicted value
reference = as.factor(data.test$readmitted),
positive = levels(as.factor(data.test$readmitted))[2])
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction     1     2
##          1    10     9
##          2  2299 18035
##
##                Accuracy : 0.8866
##                  95% CI : (0.8822, 0.8909)
##     No Information Rate : 0.8866
##     P-Value [Acc > NIR] : 0.4967
##
##                   Kappa : 0.0068
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.999501
##             Specificity : 0.004331
##          Pos Pred Value : 0.886938
##          Neg Pred Value : 0.526316
##              Prevalence : 0.886552
##          Detection Rate : 0.886110
##    Detection Prevalence : 0.999066
##       Balanced Accuracy : 0.501916
##
##        'Positive' Class : 2
##
```

Thus,

Suggestion: You may use any of the methods covered so far in parts 1) and 2), and they need not be the same. Also keep in mind that a training/testing data split may be necessary.

4) We suggest you to split the data first to Training/Testing/Validation data:

- Use training/testing data to land a final model (If you only use LASSO to land a final model, we will not need testing data since all the decisions are made with cross-validations.)
- Evaluate the final model with the validation data to give an honest assessment of your final model.