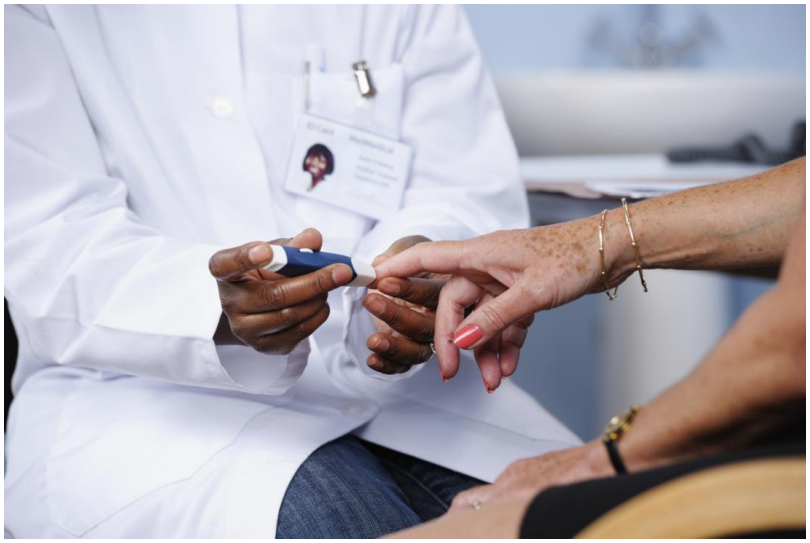
Predicting diabetes related hospital readmission

STAT 471/571/701, Fall 2017

Joseph Haymaker

Due: November 19, 2017 at 11:59PM

Contents



Executive Summary

For most of 2017 health care has been in the forefront of public discourse due to president Trump's campaign promise to "repeal and replace Obamacare" (the Affordable Care Act). The majority of the deleterious macro effects these changes would have on the American public have highlighted losses in coverage and the disproportionate effect to lower income and elderly individuals (source 1, source 2). To a lesser extent, these appraisals have addressed changes affecting Americans with pre-existing conditions and funding cuts hospitals would experience (source). Even in spite of this ongoing national debate around the topic, healthcare costs and savings in the US are a perennially obtuse topic, even to most Americans.

This investigation seeks to focus on one particular aspect of healthcare costs and savings in the US: hospitalization costs, specifically the cost of readmitting a diabetes patient in less than 30 days. This inquiry intends to address this issue by providing healthcare providers information about a patient's estimated propensity for readmission in this time period given certain individual factors. It is our hope that the findings here shed light on patient characterisitics influencing readmission, which the care community may then turn into actionable policy and practice changes that both better serve patients as well as reduce readmission costs.

The dataset analyzed is courtesy of Virginia Commonwealth University and consists of patient observations from 130 hospitals in the US from 1999 - 2008. The original dataset contained 50 patient features across 101,766 observations. In the simplest terms the variables belong to the following categories: patient identifiers, patient demographics, admission and discharge details, patient medical history, patient admission details, clinical results, medication details, & readmission indicator (dependent variable). After cleaning many of the medication details were eliminated due to little variability, while binning was applyed to some categorical variables. Similarly, througout the process variables with large amounts of missing entries, those with an excessive amount of levels, and ones unrelated to readmission status were eliminated.

Three classification models were obtained from this data: two using logistic regression and one using the random forests technique. The implications of these models are that to prevent readmission and better serve patients the hospital should pay particular attention to the following patient characteristics and their purported effects on readmission under 30 days:

- admission source/patient referral
- patient age
- diabetes medication prescribed during visit (y/n)
- discharge disposition (discharged to home, discharged to home w/ home health services, discharged/transferred to SNF (skilled nursing facilities), & other)
- insulin dose change during visit (down, no, steady, or up)
- Metformin medication change (down, no, steady, or up),total number of diagnosis for patient
- Number of emergency visits by the patient in the year prior to the current encounter
- Number of inpatient visits by the patient in the year prior to the current encounter
- Patient's length of stay in the hospital (in days)

With respect to factors that increase likelihood of readmission <30 days, unsurprisingly, probability of short term readmission increase as number of emergency visits, number of inpatient visits, time in hospital, and number of diagnoses per patient increase. Other more abstract factors increasing the probability are physician referral for admission, diabetes medication prescription during visit, and virtually any discharge situation, though discharge to 'other' situation and skilled nursing facilities cause a greater increase in this probability. There may be something of a selection effect in that finding, however. Lastly, all ages over 19 imply an increase in the probability, though the 60-79 age range corresponds with the greatest increase in this probability.

Factors that decrease likelihood of readmission <30 days are the following: no insulin dosage prescribed, steady dosage, or increased dosage *all* have a decreased effect on likelihood of short term readmission. Compared to an emergency room patient admission, both 'other' admits and referrals from home health imply a drop in likelihood of a <30 day readmission.

By paying particular attention to these indicators hospitals may be able to identify patients of greater risk, and change their treatment plans accordingly, as well as connect them with more extensive social services when possible to avoid the current costs being incurred by these types of patients and their short term hospital readmits.

Lastly, their are several problems, solutions, and ideas for further investigation that emerge from this dataset and these findings. The most salient question from the dataset is where the 130 hospitals are located. This is important due to the differing health trends of US regions. Thus, a large representation from unhealthier regions or states would imply the sample wasn't random. To speak to the different environmental factors of the hospitals future investigation might take environmental factors into account, such as weather trends (number of sunny days per year, for example), walkability scores of the areas, percent car owners in the areas, average commute time, proximity to grocery stores/ fresh food availabilty (or food deserts), etc. These would all serve as indicators of environmental factors that either cause or reflect a more sedentary lifestyle or poor eating habits, which in turn affect health outcomes. Another aspect of future investigation might be the role of early intervention such as health eating programs and exercise regimes à la Michelle Obama's "Let's Move!" campaign. Moreover, with the data as is there remain many questions about correlation of variables. For example, a better way to ascertain if it was diabetes that caused hospitalization, or a concurrent/opportunistic complication. Lastly, one important thing to remember about this data is that it only covers outcomes up through 2008. It was only in 2014 that the major provisions of the Affordable Care Act came into effect.

Up until that point many of these people would likely have been shut out of certain types of coverage due to pre-existing condition status. Since then they may have experienced greater access to preventative care, which in turn would reflect in the readmission rate. A more current dataset would be needed to assess this.

Detailed Process of the Analysis

Introduction

In the *National Diabetes Statistics Report, 2017* the CDC estimates that 30.3 million people of all ages—or 9.4% of the U.S. population—had diabetes as of 2015 (source). Moreover, the CDC estimates both direct and indirect costs due to diabetes in the United States in 2012 was $245 billion. Average medical expenditures for people with diagnosed diabetes were about $13,700 per year, 60% of which was attributed to the illness directly. Medical expenditures among people with diagnosed diabetes were about 2.3 times higher than expenditures for people without diabetes (source). Consequently, 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.

On the healthcare provider side the estimated cost burden resulting from avoidable hospitalizations due to short-term uncontrolled diabetes is equally staggering: $2.8 billion (source). Thus it is clear that hyperglycemia management and treatment is not only a hospital expenditure question, but a public health concern. This study seeks to use large amounts of patient data to develop models that can predict a patient's likelihood of readmission in less that 30 days, and thus provide healthcare providers with more tools to anticipate and avoid these outcomes.

Data Summary

The original data is from the Center for Clinical and Translational Research at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are 101,766 hospital admissions observations in this dataset, with 50 variables describing patients. The cleaned dataset contains only 31 variables. Pairing down was done due to large amounts of missing values (`Payer code`, `weight` and `Medical Specialty`, & entries with a '?' `race` value). Many of the medication variables had little variation and thus were eliminated. Binning was performed on the diagnoses variables. Lastly, `readmitted` was changed to be a categorical variable. The remaining variables still account for many patient characteristics such as demographics, medications, test results, and partial medical history (number of hospital visits, etc.)

Data source: Clore, John, et al. " Diabetes 130-US Hospitals for Years 1999-2008 Data Set ." UCI Machine Learning Repository: Diabetes 130-US Hospitals for Years 1999-2008 Data Set, Center for Clinical and Translational Research, Virginia Commonwealth University, 2014, archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.

Modeling & Analyses

Variables of interest

For the sake of getting a better idea about the data and trends, I isolated several variables that seem likely to influence readmission. These variables are `readmitted`, `race`, `gender`, and `number_diagnoses` (note: I originally also had `age`, `change`, but they have been moved to the appendix for the sake of brevity). I have summarized and visualized my findings below.

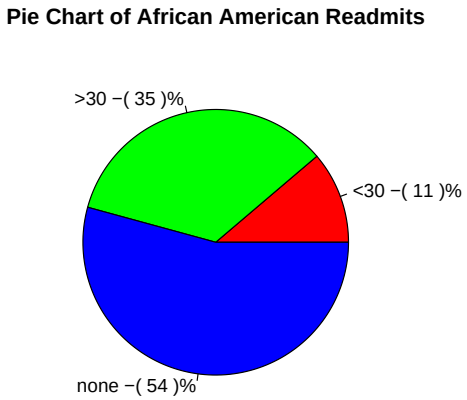
See appendix for more detailed information on variables

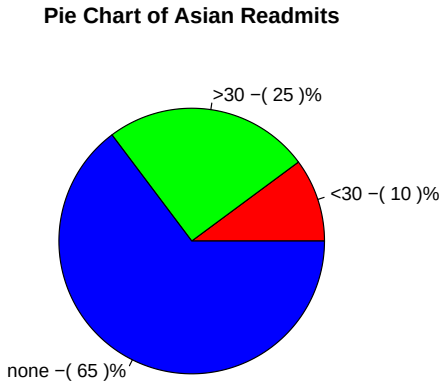
Readmitted

Before changing this dependent variable to be categorical (readmitted <30 days = 1, not readmitted <30 days = 0) it's worth noting that the patients readmitted <30 days comprise ~11% of all patients, while those readmitted over 30 days is 35%, and those not readmitted is the remaining 54%.

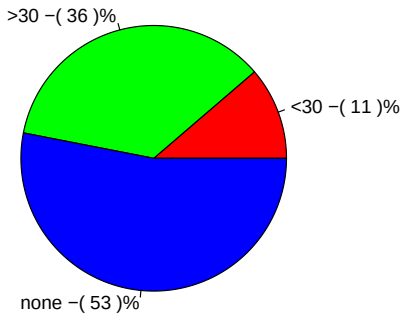
Race

Since many health issues affect certain racial groups disproportionately, it seems reasonable to explore the relationship between race and readmission. Causasians and african americans observations dwarf the remaining 3 categories; there are (roughly) 76,000 Caucasian patients and 20,000 African American patients as opposed to only 650 Asian, 2050 Hispanic, and 1500 'other' patients. Visualizing the breakdown of readmit status for these groups provides us with some interesting information:

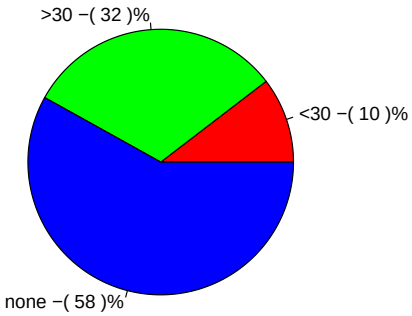


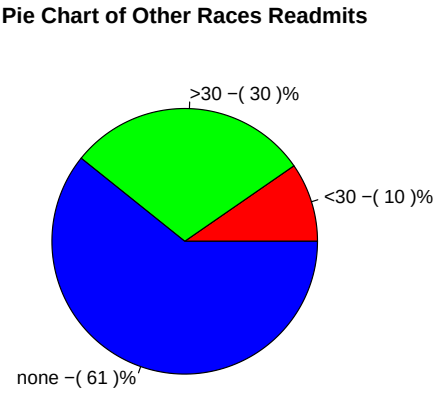


Pie Chart of Caucasian Readmits



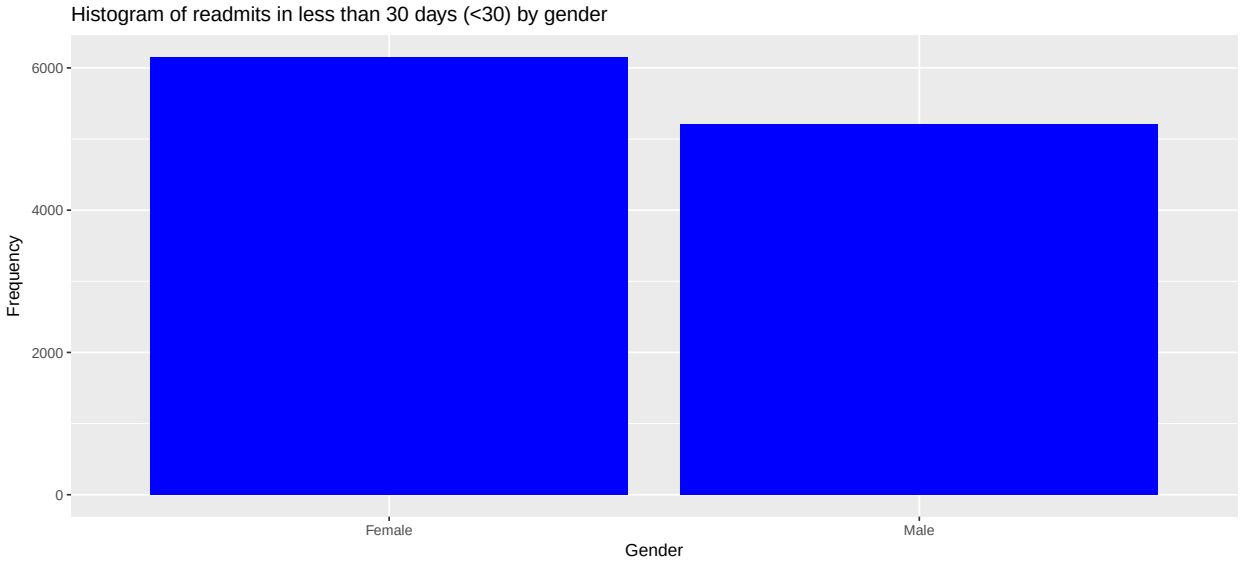
Pie Chart of Hispanic Readmits

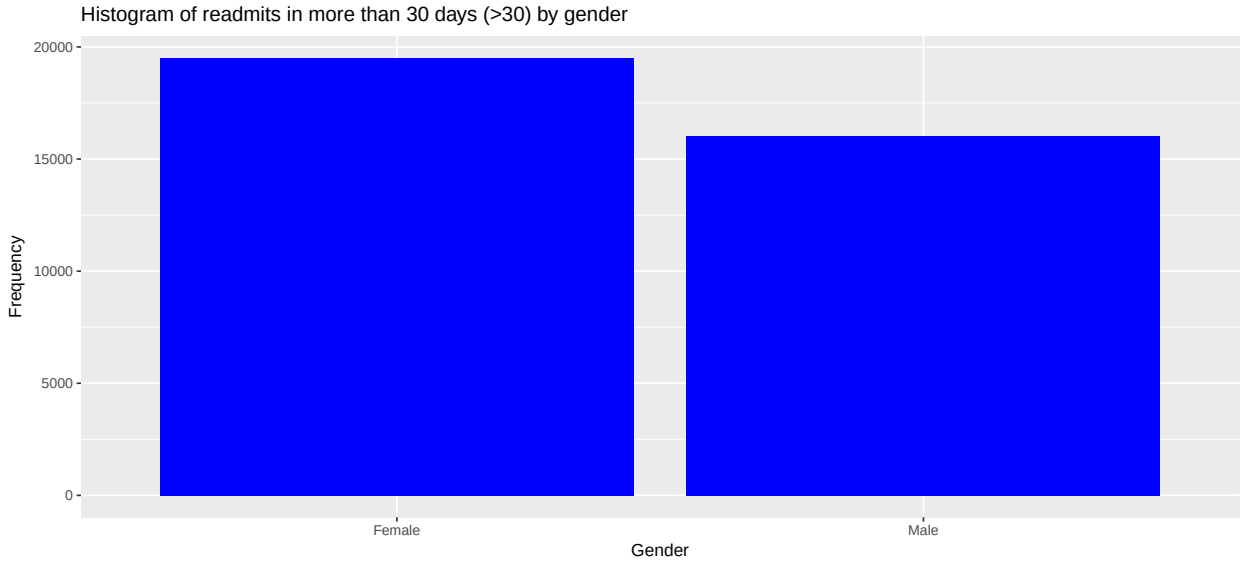




Regardless of race, roughly 10-11% of the patients are readmitted in less than 30 days. This implies race may not play such a big role in this outcome after all. All proportions are very close to one another across races, save for Asians which have the highest percentage of no readmits at 65%.

Gender



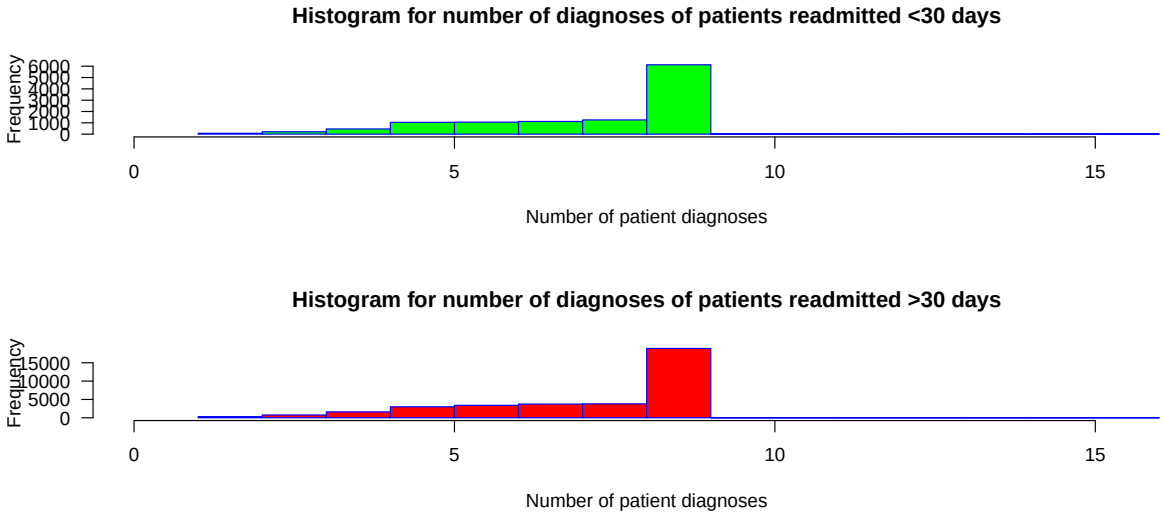


In the cleaned dataset we have 54708 female observations and 47055 male observations, which means roughly 54% of the patients under consideration were female (for all readmission categories), while ~46% were male. When comparing hospital readmits striated by gender, of the patients that were readmitted in *under* 30 days approximately 54% (6152/11357) were female, matching the overall female representation. Similarly, of patients that were readmitted *over* 30 days again 54% (19518/35545) were female. Note that the total number of patients (male & female) readmitted over 30 days is about 3 times that of those readmitted in *less* than 30 days.

There seems to be a gap between genders here implying that women are more prone to readmission, but this is quickly rebuked when we compare the genders in terms of their total observations. For patients who were readmitted in *less* than 30 days, female patients represent 11.2% (6152/54708) of the total female population, while those who are male represent a similar 11.1% (5205/47055) of the overall male population. The same is true for patients readmitted *over* 30 days: female patients account for 35.7% (19518/54708) of the total female population, while male patients comprise 34.1% (16027/47055) of the total male population.

This lends credence to the notion that gender does not contribute to likelihood of readmission.

Number of diagnosis



There consistently seems to be a large spike in frequency around 9 diagnoses, which greatly outnumbers the count of other diagnosis.

Model building

Further prepping data

After investigating collinearity of variables (see appendix) I further modified the dataset by making `readmitted` categorical (readmitted <30 days = 1, otherwise = 0), removed all `race` entries coded as "?", and removed the `encounter_id` and `patient_nbr` columns as they add nothing to the model.

Initial modeling

After some initial modeling, I opted for a more robust and methodically sound model used LASSO regularization. This method has the advantage of model building that adds constraints (a penalty) to the variable coefficients, giving us sparse model selection. This is good given the amount of features (many with multiple levels) that we have. The tuning parameter for this penalty function - λ - was chosen via cross validation, the value of which was chosen to minimize mean cross validation errors (`lambda.min`). This yielded a model with 22 variables (not double counting variable levels). Creating a model with these 22 variables and running the `Anova()` test revealed that not all variables were statistically significant. Thus, I began the process of (manual) backwards elimination by kicking out the variable with the largest P value (there is a large chance that the true value is zero (null hypothesis)).

After doing manual backwards elimination the model has 12 variables, all of which are statistically significant at the .05 level. Some of these remaining variables seem like they might exhibit colinearity. Although plotting them shows that the actuality is really not all that bad (see appendix).

In looking at the summary of the model multiple levels of the `A1Cresult` are not significant. Also under further investigation this number seems likely to be highly correlated with insulin. This is a similar case with `metformin`. For those reasons I've decided to remove them.

Lastly, for the sake of model robustness I am also choosing to remove `glipizide`, which is only significant at the .05 level according to the Anova test.

Classifier 1 (from logistic regression)

The final model specified by preliminary model building using LASSO (least absolute shrinkage and selection operator): the resulting model obtained by LASSO regularization has 9 variables, 5 of which have multiple levels:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.151458   0.584642  -7.101 1.24e-12 ***
age[10-20)        0.889354   0.605864   1.468 0.142128
age[20-30)        1.424425   0.588665   2.420 0.015531 *
age[30-40)        1.409278   0.585919   2.405 0.016162 *
age[40-50)        1.332041   0.584615   2.278 0.022697 *
age[50-60)        1.276257   0.584303   2.184 0.028945 *
age[60-70)        1.437881   0.584173   2.461 0.013840 *
age[70-80)        1.498978   0.584121   2.566 0.010282 *
age[80-90)        1.512456   0.584315   2.588 0.009642 **
age[90-100)       1.436819   0.587038   2.448 0.014382 *
time_in_hospital  0.022443   0.003758   5.971 2.35e-09 ***
num_procedures   -0.025263   0.006791  -3.720 0.000199 ***
num_medications   0.005140   0.001572   3.269 0.001079 **
number_emergency  0.032809   0.008436   3.889 0.000101 ***
number_inpatient  0.265180   0.006509  40.741  < 2e-16 ***
number_diagnoses  0.041705   0.006065   6.876 6.14e-12 ***
insulinNo        -0.168857   0.035189  -4.799 1.60e-06 ***
insulinSteady    -0.151091   0.033016  -4.576 4.73e-06 ***
insulinUp        -0.092093   0.039477  -2.333 0.019657 *
diabetesMedYes    0.116898   0.031464   3.715 0.000203 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Doing a simple prediction with this model (which would be the goal) using the last observation in the dataset sets the prediction of this patient being readmitted in 30 days is 7.7%.

Classifier and threshold

Now that we have a logistic model to serve as a classifier we have to choose a probability threshold for what will be classified as a likely readmit <30 days, and what will not be classified as such. The goal will obviously be to minimize misclassification error (false positives & false negatives out of total observations). Due to the fact that we have estimated that mislabeling a readmission will incur a cost twice that of mislabeling a non-readmission we can use Bayes rule to take into account the weight of these potential losses. According to Bayes' rule we have:

$$P(Y = 1|x) > \frac{\frac{a_{0,1}}{a_{1,0}}}{1 + \frac{a_{0,1}}{a_{1,0}}}$$

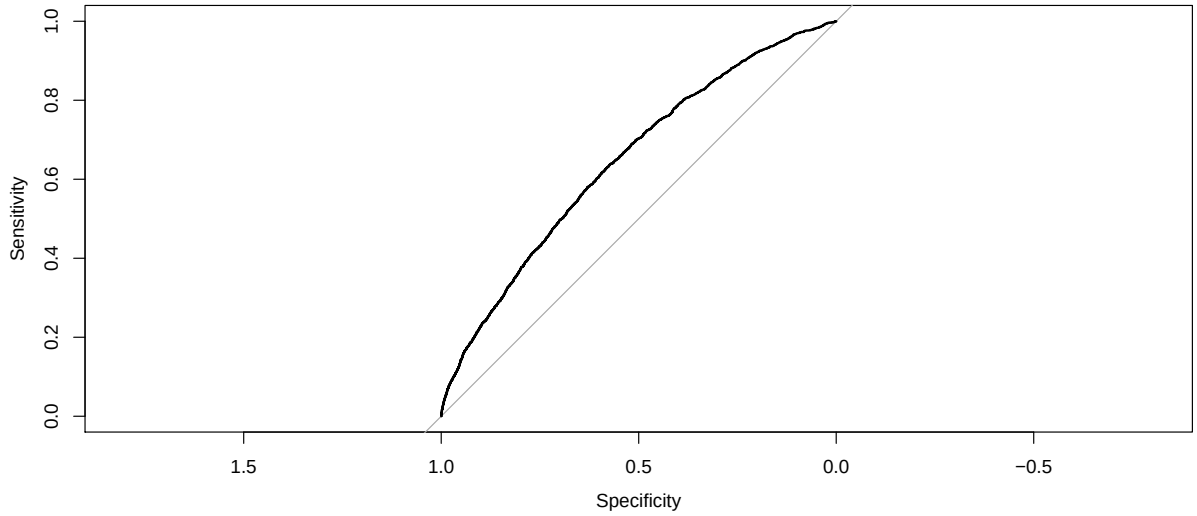
Which given our cost estimation becomes:

$$P(Y = 1|x) > \frac{\frac{1}{2}}{1 + \frac{1}{2}} = .33$$

$$logit > \log(\frac{0.33}{0.66}) = -0.69$$

After using the information about false positives costing twice as much as false negatives to establish a threshold we see that the resulting mis-classification error is 22%, i.e. roughly 1 out of 5 patients would be mis-classified with this model and threshold.

Using testing data to assess classifier



While we now have a working model to predict patient readmission <30 days, for the sake of due diligence I will obtain another classifier using logistic regression, before obtaining a third using random forests.

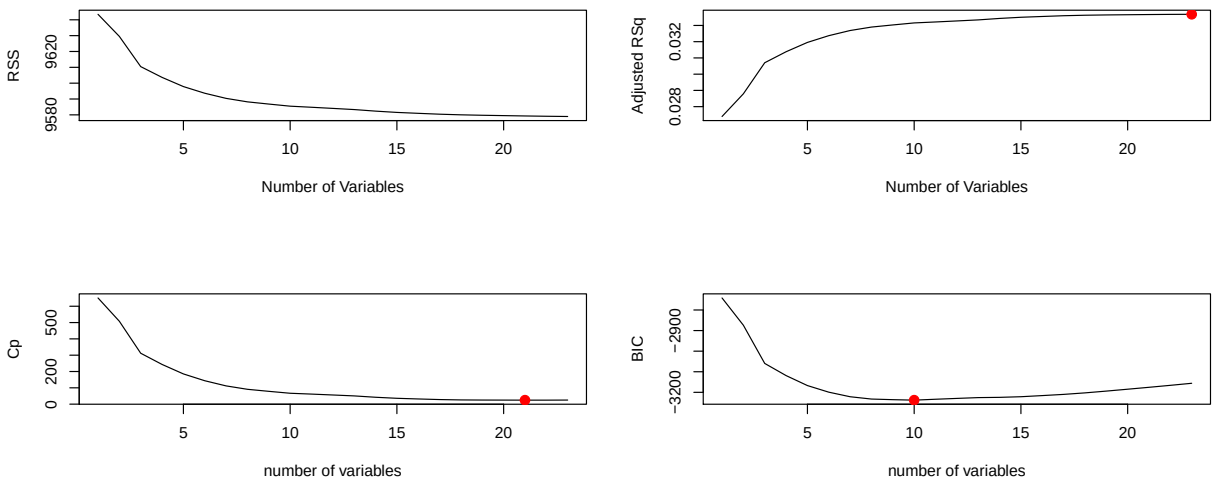
Model building (classifier 2) using non-zero coefficients from LASSO with forward & backward stepwise selection

As a reminder the non-zero coefficients given by LASSO were:

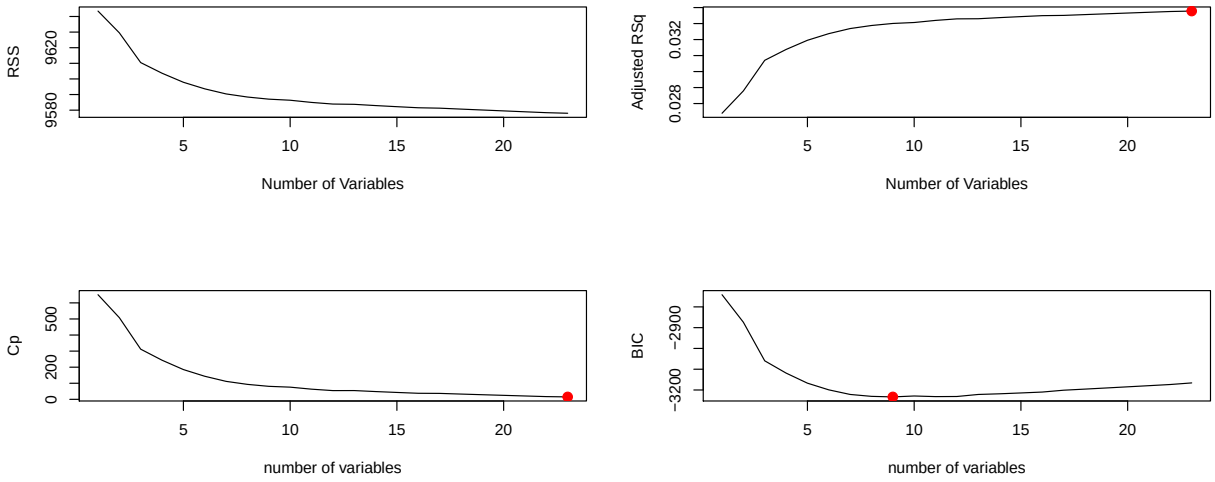
```
readmitted ~ race + gender + time_in_hospital + num_lab_procedures + num_medications + number_emergency
```

I will perform forward and backward stepwise selection on these variables to get the smallest possible model with statistically significant coefficients. This is to say using minimum BIC as criterion I will choose the appropriate number of variables from each method, then compare outputs.

Forward stepwise selection



Backward stepwise selection



Putting the variables chosen from forward & backward selection we see that the results are almost identical. The only difference is that forward selection pointed to the importance of one more variable than forward selection, `num_lab_procedures`. For this reason I've chosen to take the variables from forward selection as the inputs for a logistic regession model. Note that the BIC value is lower with this model as well.

All variables are significant at the .05 level according to the Anova test in this model, but for robustness the two with the highest P values are being removed. After removing these two variables (`num_lab_procedures` and `max_glu_serum`) all variables are significant at the .01 level according to the Anova test. The misclassification error is 22%, similar to that of the first classifier.

Classifier 2 (using logistic regression)

This formula for this classifier is:

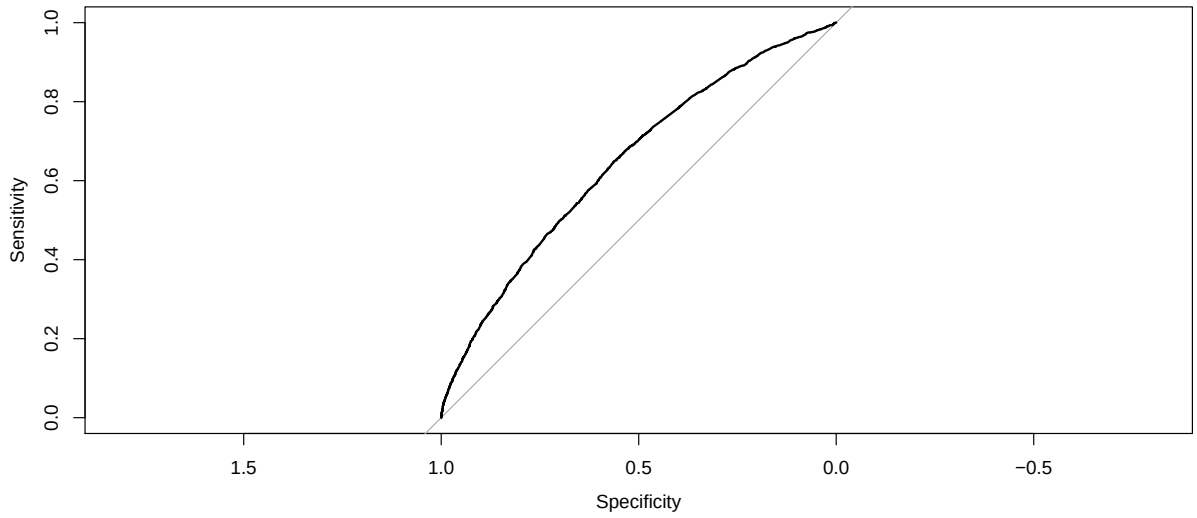
```
readmitted ~ number_inpatient + number_diagnoses +
    metformin + disch_disp_modified + adm_src_mod + age_mod
```

and the values of the coefficients are:

```
Coefficients:
                                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                   -3.128193   0.206342 -15.160  < 2e-16 ***
number_inpatient                                               0.275203   0.006243  44.079  < 2e-16 ***
number_diagnoses                                               0.042862   0.005929   7.229 4.86e-13 ***
metforminNo                                                   -0.174865   0.130759  -1.337 0.181123
metforminSteady                                               -0.257777   0.132659  -1.943 0.051998 .
metforminUp                                                   -0.400679   0.172565  -2.322 0.020239 *
disch_disp_modifiedDischarged to home with Home Health Service 0.231017   0.031307   7.379 1.59e-13 ***
disch_disp_modifiedDischarged/Transferred to SNF               0.423541   0.030362  13.950  < 2e-16 ***
disch_disp_modifiedOther                                       0.434021   0.028540  15.207  < 2e-16 ***
adm_src_modOther                                              -0.098864   0.041641  -2.374 0.017588 *
adm_src_modPhysician Referral                                  0.008675   0.023791   0.365 0.715373
adm_src_modTransfer from Home Health                          -0.116594   0.042870  -2.720 0.006534 **
age_mod20-59                                                   0.499108   0.160295   3.114 0.001848 **
age_mod60-79                                                   0.583015   0.160324   3.636 0.000276 ***
age_mod80+                                                     0.524009   0.161610   3.242 0.001185 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Using testing data to assess classifier

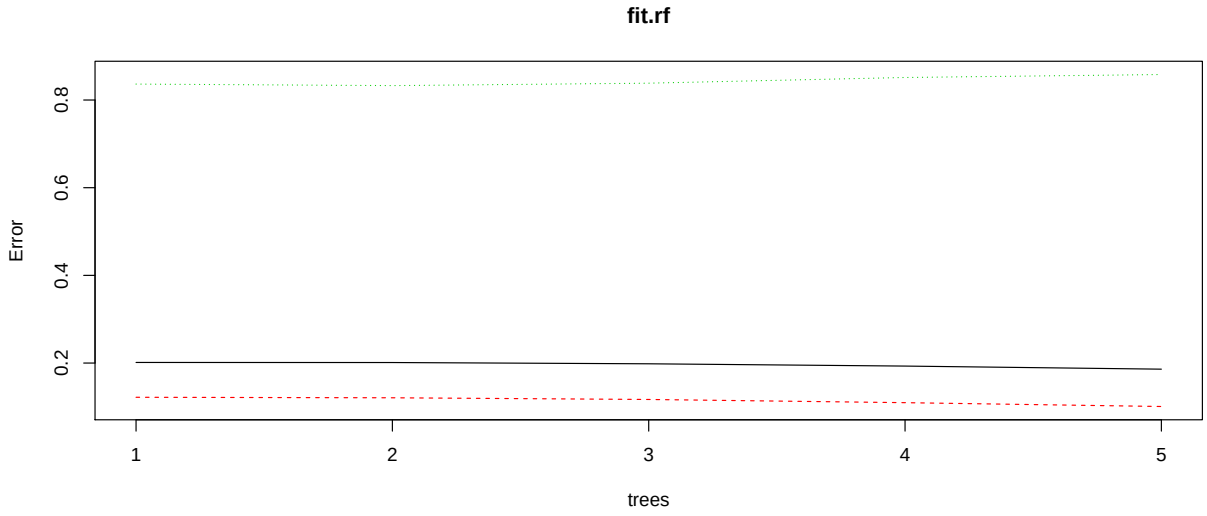
Splitting the observations into testing and training data we can then plot the ROC curve:



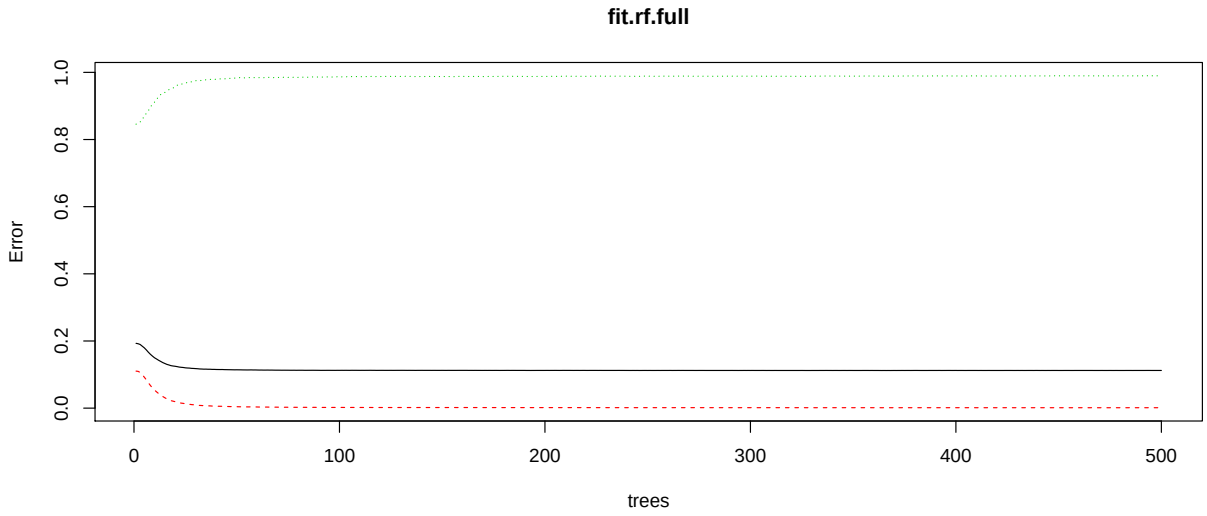
Lastly, this classifier takes the last row observation and predicts it's `readmitted` value to be 8.84%, similar to the 7.7% predicted by the first classifier.

Second model approach: random forests for classification

The second approach to model building will use random forests to develop the third and final classifier.

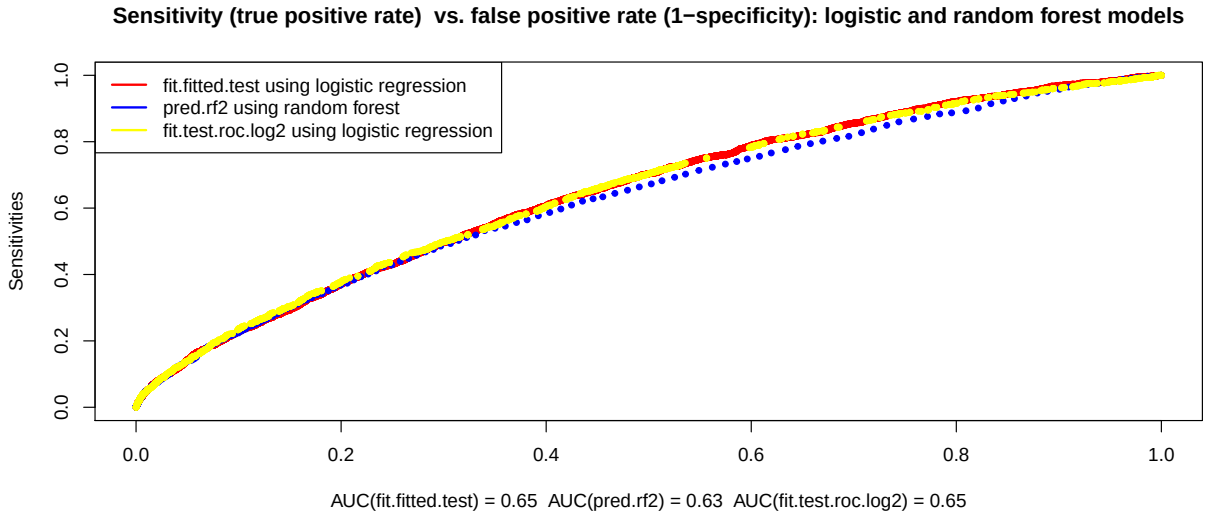


Note that the variables `diag1_mod`, `diag2_mod`, and `diag3_mod` were all removed from modeling to begin with due to the fact that they all have a large number of levels. We can see (above) using a forest with just a few number of trees results in mean prediction error designated by the black line is around 20% (less than desirable). Expanding the number of trees we get:



As the graph shows when we expand the number of trees in the forest to 500 we get a misclassification error now around 11%. To assure overfitting in this model we now split the dataset into training (80%) and testing data (20% of original data). Plotting the model using the training data is almost identical to the full dataset findings (which is good).

We now have 3 classifiers (two from logistic regression and now one using random forests), and thus have 3 ROC curves to compare:



We can see that when comparing the three classifiers that in terms of area under the curve alone the random forest model is worse than the logistic regression models by only .02. More importantly, the logistic regression models outperform the random forest model at practically every level. Since the two logistic models are almost identical, I'd give preference to the original classifier due to the fact that it maximizes the true positive rate slightly more than the second classifier at several points.

Conclusion

This study had as its goal to identify patient characteristics that affect readmission < 30 days in order to better serve patients and minimize healthcare costs. To this end, I first examined relevant variables before performing LASSO regularization with cross validation to decrease the number of variables under consideration. I then obtained 2 classifiers from logistic regression – one via manual backwards elimination, and another as a result of the comparision of backwards and forward selection. The threshold used for classification was chosen using Baye's Rule, knowing that a false positive costs twice as much as a false negative. I then created one last classifier using random forests. All of these models were checked using testing and training data. I then plotted all three classifiers' ROC curves and compared them before concluding that the first was the best in terms of maximizing the true positive rate. The implications of this model are extensive. For the sake of clarity I have summarized them in the following table:

Description	Variable name	Categories/levels (if applicable)	Effect on readmission <30 days
Admission source/patient referral	`adm_src_mod`	emergency room, other, physician, or transfer from home health	compared to emergency room referrals both 'other' and home health transfers decrease readmission, while physician referrals only slightly increase it
Patient age	`age_mod`	0-19 yrs, 20-59 yrs, 60-79 yrs, & 80+ yrs	compared to 0-19 year old patients all other age brackets increase probability, with the 60-79 bracket assiated with the greatest increase
Diabetes medication prescribed (y/n)	`diabetesMed`	no or yes	a medication prescription is associated with an increase in readmission

Description	Variable name	Categories/levels (if applicable)	Effect on readmission <30 days
Discharge disposition	`disch_disp_modified`	discharged to home, discharged to home w/ home health services, discharged/transferred to SNF (skilled nursing facilities), & other	compared to home discharge all other categories increase readmission likelihood
Insulin dose change	`insulin`	down, no, steady, or up	compared to a drop in dosage all other categories correspond to a decrease in readmission
Total diagnosis for patient	`number_diagnoses`	1-16	increases in diagnoses correspond with increase in readmission
Number of emergency visits by the patient in the year prior to the current encounter	`number_emergency`	0-76	increases in emergency visits correspond with increase in readmission
Number of inpatient visits by the patient in the year prior to the current encounter	`number_inpatient`	0-21	increases in inpatient visits correspond with increase in readmission
Patient's length of stay in the hospital (in days)	`time_in_hospital`	1-14	increases in length of stay correspond with increase in readmission

Knowing this information hospitals can now predict the likelihood that a given patient will be readmitted in less than thirty days, and adjust their intervention accordingly.

Appendix

Variable names and detail explanations

id: variables

Description of variables

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

a) Patient identifiers:

a. `encounter_id`: unique identifier for each admission
b. `patient_nbr`: unique identifier for each patient

b) Patient Demographics:

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

c) Admission and discharge details:

a. `admission_source_id` and `admission_type_id` identify who referred the patient to the hospital (e.g. physician vs. emergency dept.) and what type of admission this was (Emergency vs. Elective vs. Urgent).
b. `discharge_disposition_id` indicates where the patient was discharged to after treatment.

d) Patient Medical History:

a. `num_outpatient`: number of outpatient visits by the patient in the year prior to the current encounter
b. `num_inpatient`: number of inpatient visits by the patient in the year prior to the current encounter
c. `num_emergency`: number of emergency visits by the patient in the year prior to the current encounter

e) Patient admission details:

a. `medical_specialty`: the specialty of the physician admitting the patient
b. `diag_1`, `diag_2`, `diag_3`: ICD9 codes for the primary, secondary and tertiary diagnoses of the patient. ICD9 are the universal codes that all physicians use to record diagnoses. There are various easy to use tools to lookup what individual codes mean (Wikipedia is pretty decent on its own)
c. `time_in_hospital`: the patient's length of stay in the hospital (in days)
d. `number_diagnoses`: Total no. of diagnosis entered for the patient
e. `num_lab_procedures`: No. of lab procedures performed in the current encounter
f. `num_procedures`: No. of non-lab procedures performed in the current encounter
g. `num_medications`: No. of distinct medications prescribed in the current encounter

f) Clinical Results:

a. `max_glu_serum`: indicates results of the glucose serum test
b. `A1Cresult`: indicates results of the A1c test

g) Medication Details:

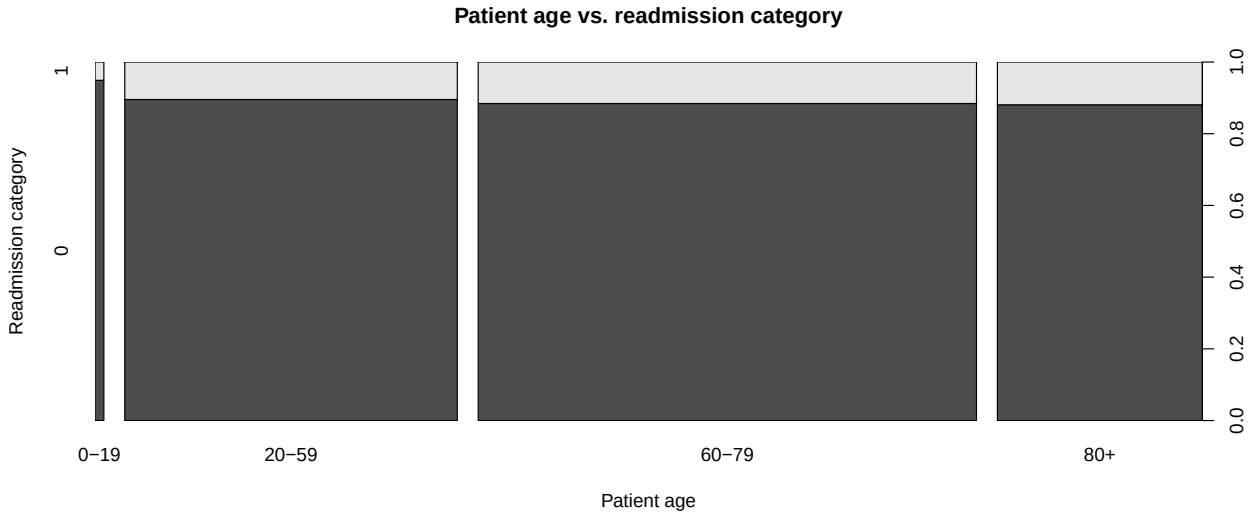
a. `diabetesMed`: indicates if any diabetes medication was prescribed
b. `change`: indicates if there was a change in diabetes medication
c. `24 medication variables`: indicate whether the dosage of the medicines was changed in any manner during the encounter

h) Readmission indicator:

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

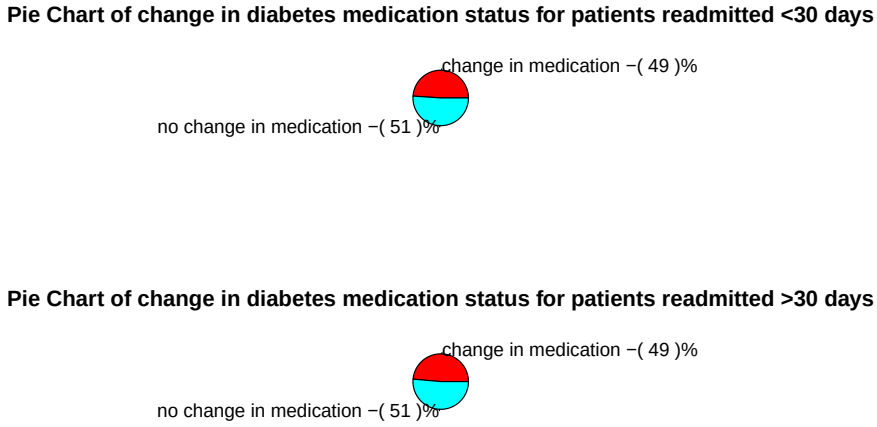
Further EDA

Age



It appears that the categories with the largest number of readmits are 60-79 and 80+, which are almost identical. As expected these percentages fall as age category does, though not dramatically.

Change (in diabetes medication)

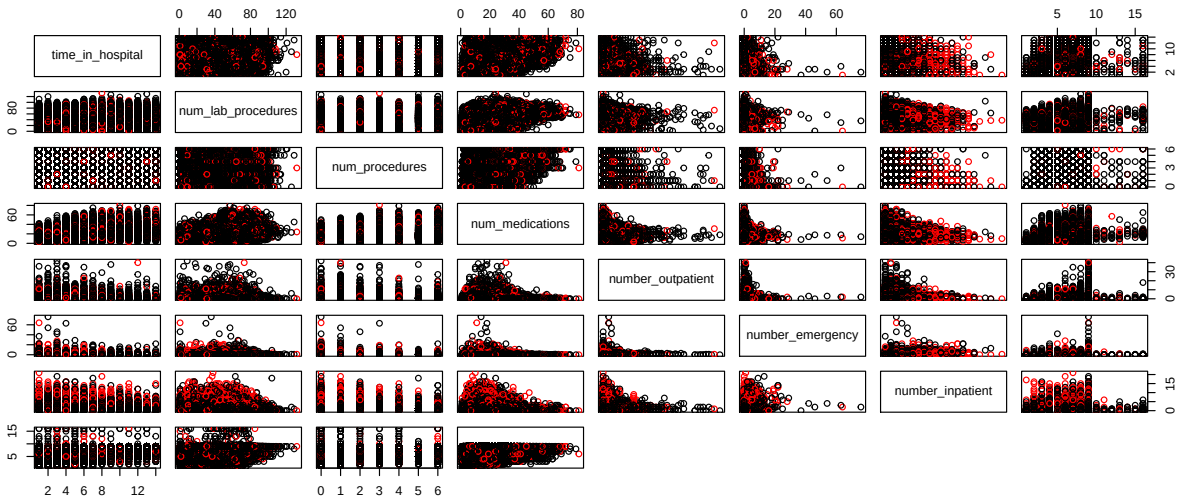


These figures show that roughly half of both types of readmits experienced a change in medication, while the other half did not. This appears to have little bearing on the outcome.

Investigating collinearity

Some of the features provided seem like they might be highly correlated with one another. In particular, diagnoses & health indicator information, as well as hospital visit information.

Before initial modeling:



After LASSO:

```
##                    time_in_hospital num_procedures num_medications
## time_in_hospital        1.000000000     0.19323405      0.46638083
## num_procedures          0.193234051     1.00000000      0.38553831
## num_medications         0.466380832     0.38553831      1.00000000
## number_emergency       -0.009798542    -0.03836918      0.01296355
## number_inpatient        0.073408368    -0.06584306      0.06499285
## number_diagnoses        0.220686659     0.07233875      0.25860468
##                    number_emergency number_inpatient number_diagnoses
## time_in_hospital       -0.009798542       0.07340837       0.22068666
## num_procedures         -0.038369184      -0.06584306       0.07233875
## num_medications         0.012963548       0.06499285       0.25860468
## number_emergency        1.000000000       0.26638244       0.05408781
## number_inpatient        0.266382440       1.00000000       0.10325182
## number_diagnoses        0.054087810       0.10325182       1.00000000
```