Predicting readmission probability for diabetes inpatients

STAT 471/571/701 Modern Data Mining

Due midnight, April 7th, 2019

Table of Contents

[Executive Summary 2](#_Toc5366916)

[Background 2](#_Toc5366917)

[The data 2](#_Toc5366918)

[Methods used 3](#_Toc5366919)

[Limitation 3](#_Toc5366920)

[Data Analysis 3](#_Toc5366921)

[(i) Data Summary 3](#_Toc5366922)

[Description of variables 3](#_Toc5366923)

[EDA 5](#_Toc5366924)

[ii) Analyses 6](#_Toc5366925)

[Lasso Model 6](#_Toc5366926)

[Backward Selection 7](#_Toc5366927)

[Model Comparision 7](#_Toc5366928)

[iii) Conclusion 8](#_Toc5366929)

[iii) Appendix 9](#_Toc5366930)

[Readmission vs Gender 9](#_Toc5366931)

[Readmission vs Race 10](#_Toc5366932)

[Readmission vs Age 11](#_Toc5366933)

[Collaboration 12](#_Toc5366934)

# Executive Summary

## Background

Hospital readmission is very expensive and it is an emerging indicator of quality of health care. On average, Medicare spends double the amount for an episode with one readmission. Moreover, diabetes is a chronic medical condition affecting millions of Americans, but if managed well, with good diet, exercise and medication, patients can lead relatively normal lives. However, if improperly managed, diabetes can lead to patients being continuously admitted and readmitted to hospitals. Readmissions are especially serious - they represent a failure of the health system to provide adequate support to the patient and are extremely costly to the system. As a result, the Centers for Medicare and Medicaid Services announced in 2012 that they would no longer reimburse hospitals for services rendered if a patient was readmitted with complications within 30 days of discharge.

Given these policy changes, being able to identify and predict those patients most at risk for costly readmissions has become a pressing priority for hospital administrators. In this project, we shall explore how to use the techniques we have learned in order to help better manage diabetes patients who have been admitted to a hospital. Our goal is to avoid patients being readmitted within 30 days of discharge, which reduces costs for the hospital and improves outcomes for patients. If we could identify important factors relating to the chance of a patient being readmitted within 30 days of discharge, effective intervention could be done to reduce the chance of being readmitted. Also if we could predict one’s chance being readmitted well, actions can be taken.

## The data

The original data is from the [Center for Clinical and Translational Research](https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008) at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are over 100,000 unique hospital admissions in this dataset, from ~70,000 unique patients. The data includes demographic elements, such as age, gender, and race, as well as clinical attributes such as tests conducted, emergency/inpatient visits, etc. Refer to the original documentation for more details on the dataset. Three former students Spencer Luster, Matthew Lesser and Mridul Ganesh, brought this data set into the class and did a wonderful final project. We will use a subset processed by the group but with a somewhat different objective.

All observations have five things in common:

1. They are all hospital admissions
2. Each patient had some form of diabetes
3. The patient stayed for between 1 and 14 days.
4. The patient had laboratory tests performed on him/her.
5. The patient was given some form of medication during the visit.

The data was collected during a ten-year period from 1999 to 2008. There are over 100,000 unique hospital admissions in the data set, with ~70,000 unique patients.

## Methods used

In order to analyse data, I first cleaned the data and deleted unecessary variables.Then I label readmission into “1” as “Yes, readmitted within 30 days” and “0” as “No”. Then, we did exploratory data analysis to have an overview of the data.

The data is seperated into two sets. One set is used and trained to generate models, while the other set of data is used to test the model and see if the model fit the data well. In order to find posible factors that are related to readmission, I used two method to build 2 models. The first model is Lasso model, and the second one is backward selection model.We weighted the cost of mislabel a readmission to be twice as much as mislabel a non-readmission We compared the two models based on their prediction error and ROC curve.

The final model contains 7 variables: 1. number\_inpatient (Number of inpatient visits by the patient in the year prior to the current encounter) 2. number\_diagnoses (Total no. of diagnosis entered for the patient) 3. insulin 4. diabetesMed (Indicates if any diabetes medication was prescribed) 5. disch\_disp\_modified 6. diag1\_mod (ICD9 codes for the primary diagnoses of the patient.) 7. diag3\_mod (ICD9 codes for the tertiary diagnoses of the patient.)

## Limitation

Readmission is not only related to factors inside of the hospital. In our model, we only related the readmission rate to patient himself/herself or factors in hospital. However, whether or not a patient is readmitted may be very related to factors outside of the hospital. Do patients receive good care after leaving the hospital; did they taking medicine on time; and their excercise frequency and emotional mood can also be a potential factors to patients readmission rate.

# Data Analysis

## (i) Data Summary

### 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.

**1) Patient identifiers:**

1. encounter\_id: unique identifier for each admission
2. patient\_nbr: unique identifier for each patient

**2) 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.

**3) Admission and discharge details:**

1. 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).
2. discharge\_disposition\_id indicates where the patient was discharged to after treatment.

**4) Patient Medical History:**

1. num\_outpatient: number of outpatient visits by the patient in the year prior to the current encounter
2. num\_inpatient: number of inpatient visits by the patient in the year prior to the current encounter
3. num\_emergency: number of emergency visits by the patient in the year prior to the current encounter

**5) Patient admission details:**

1. medical\_specialty: the specialty of the physician admitting the patient
2. 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)
3. time\_in\_hospital: the patientâs length of stay in the hospital (in days)
4. number\_diagnoses: Total no. of diagnosis entered for the patient
5. num\_lab\_procedures: No. of lab procedures performed in the current encounter
6. num\_procedures: No. of non-lab procedures performed in the current encounter
7. num\_medications: No. of distinct medications prescribed in the current encounter

**6) Clinical Results:**

1. max\_glu\_serum: indicates results of the glucose serum test
2. A1Cresult: indicates results of the A1c test

**7) Medication Details:**

1. diabetesMed: indicates if any diabetes medication was prescribed
2. change: indicates if there was a change in diabetes medication
3. 24 medication variables: indicate whether the dosage of the medicines was changed in any manner during the encounter

**8) 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.

### EDA

In our dataset, we have 101766 observations. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008.We have 31 variables but encouter\_id and patient\_nbr are unrelevent to our model, so we exclude those two. Readmitted as our output variable, we modified it into “1”: patient readmit within 30 days, “0”, patient does not readmit within 30 days.

From Bar plot in Appendix I, we can see that, more females are readmitted than male, but the probability of readmission is about the same (0.1266991 for female and 0.1243728 for male).

Among all race, Caucasians and African American are the most likely to readmit, and more Caucasians are readmitted.

There are more 60-79 yrs old people being readmitted. As age goes up, the ratio of readmission goes up.

In our dataset, we find 2273 “?” in race, here, we consider “?” as mixed race, and we will take “?” into consideration. We also find 3 Unknown/Invalid sex in gender column. We will also take this 3 observations into consideration as representation of transgenders.

Name shown below will be our input, and readmitted will be our response variable.

## [1] "race" "gender" "time\_in\_hospital"   
## [4] "num\_lab\_procedures" "num\_procedures" "num\_medications"   
## [7] "number\_outpatient" "number\_emergency" "number\_inpatient"   
## [10] "number\_diagnoses" "max\_glu\_serum" "A1Cresult"   
## [13] "metformin" "glimepiride" "glipizide"   
## [16] "glyburide" "pioglitazone" "rosiglitazone"   
## [19] "insulin" "change" "diabetesMed"   
## [22] "disch\_disp\_modified" "adm\_src\_mod" "adm\_typ\_mod"   
## [25] "age\_mod" "diag1\_mod" "diag2\_mod"   
## [28] "diag3\_mod" "readmitted"

## ii) Analyses

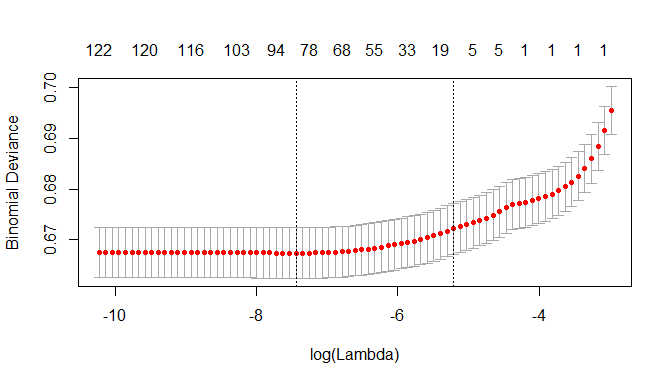
We seperate the data set into 2 subsets. 70% of the data is assigned into training set and the remaining 30% is assigned to testing data set. There are 71236 variables in training set, and 30530 variables in testing set.

### Lasso Model

The first model we built is a logistic regression with Lasso method. We used cv.glmnet function, and binomial deviance from 10-fold cross validation. We get the following graph. Then we choosed variables of coef.1se, and fit those variables into a logistic regression.

glm(readmitted~ time\_in\_hospital + number\_inpatient + number\_diagnoses + insulin + diabetesMed + diabetesMed + disch\_disp\_modified + diag1\_mod + diag3\_mod ,family=binomial, data=data.train)

After eliminate variables, and make sure every variables are statistically significant at 0.05 level, we get our lasso model - fit.lasso with 7 variables



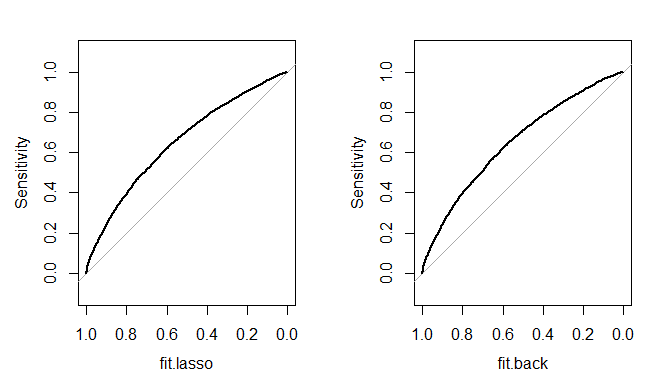
### Backward Selection

The second method we used to build the model is backward selection. Started with the full model, kept one variable with that largest p-value until each variable is significant at 0.05 level. We get our backward selection model - fit.back with 15 variables.

### Model Comparision

We use Bayes Rule to choose a threshold for out model. 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 calculation, we get out threshold set to be 0.33. After calculation, fit.lasso has a mce of 0.1165084, and fit.back has a mce of 0.1169014. We choose the model with lower mce. Thus, we choose lasso model.

The following graph shows the ROC curve and AUC value for both fit.lasso and fit.back. Two ROC curve looks very similar and we have AUC = 0.65 for both model. So, we still use the lasso model because it has a lower prediction error. 

## AUC.fit.lasso. AUC.fit.back.  
## 1 0.65 0.65

## iii) Conclusion

In conclusion, our final model contains 7 variables.

number\_inpatient (Number of inpatient visits by the patient in the year prior to the current encounter): As other variables remain the same, increase in one unit of number\_inpatient will result in increase of readmission by 0.258085. The hospital should limit the inpatient visits to reduce the readmission probability.

number\_diagnoses (Total no. of diagnosis entered for the patient): As other variables remain the same, increase in one unit of number\_diagnoses will result in increase of readmission by 0.027515. The more disgnosis entered for the patient, the patient is more likely to readmit.

insulin: All insulin level has a negative effect on readmission. Insulin is good for diabetes, so the negative relationship makes sense.

diabetesMed (Indicates if any diabetes medication was prescribed): diabetesMed Yes has a higher positive effect on readmission

disch\_disp\_modified, All level has positive relationship with readmission.

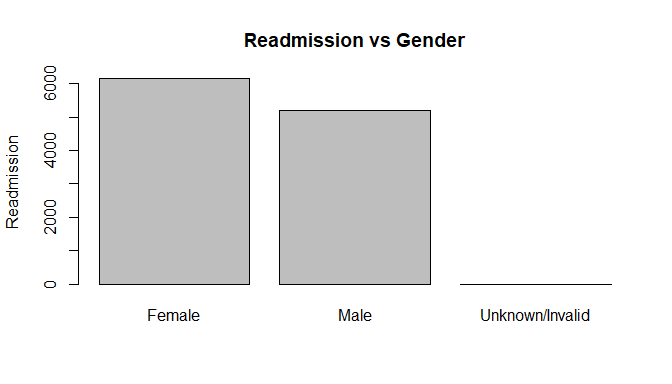
diag1\_mod (ICD9 codes for the primary diagnoses of the patient.): All levels of diag1\_mod have positive relationship with readmission

diag3\_mod (ICD9 codes for the tertiary diagnoses of the patient.): Most diad3\_mod level have positive relationship with readmission

Beside these variables, I suggest hospitals to pay more attention to patient that are 60 yrs old and above. Based on our bar plot from EDA, we find the probability increases as people getting older, which also makes sense that older people are more vulnerable. We can also have some further analysis on factors outside of the hospital to see if we can find some valuable information.

## iii) Appendix

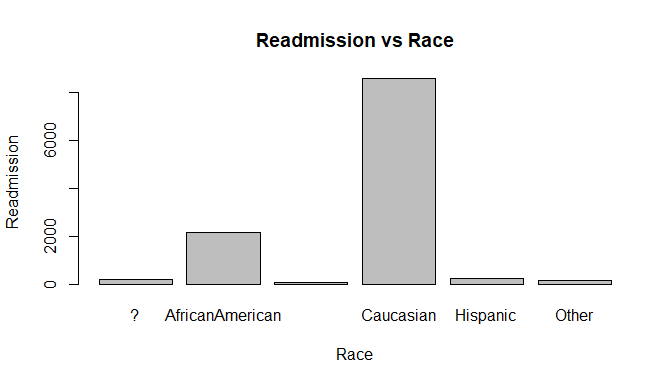
### Readmission vs Gender

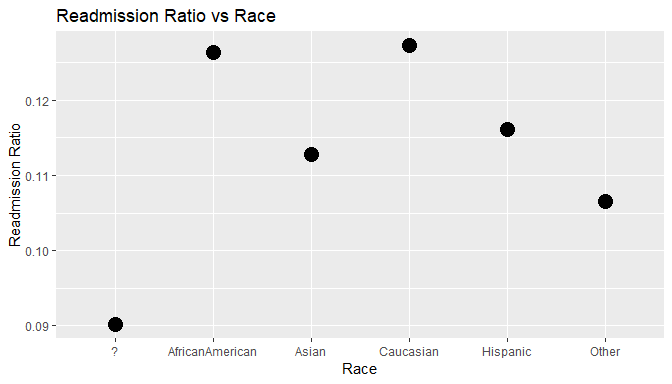


Readmission rate vs gender

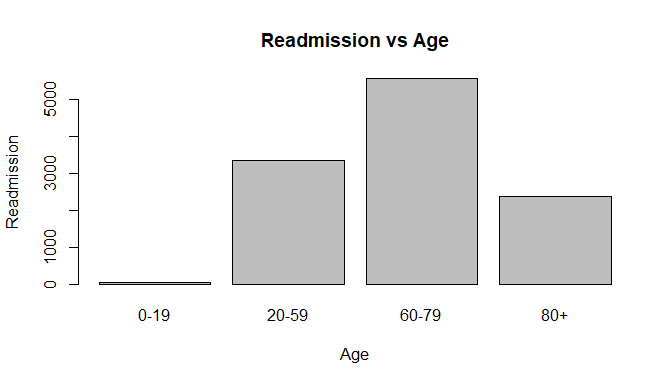
## # A tibble: 3 x 2  
## gender ratio  
## <fct> <dbl>  
## 1 Female 0.127  
## 2 Male 0.124  
## 3 Unknown/Invalid 0

### Readmission vs Race

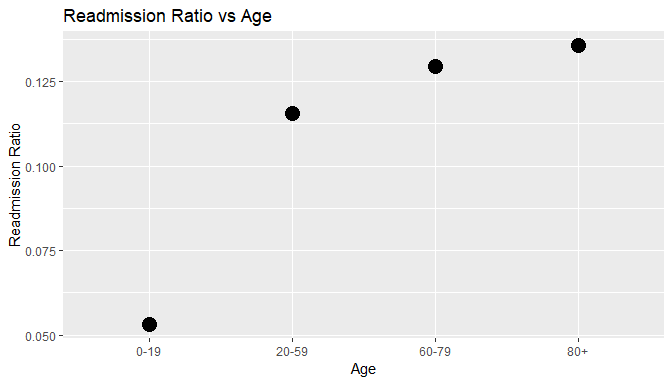




### Readmission vs Age



ratio\_age <- readmission %>% group\_by(age\_mod) %>% summarise(ratio=sum(readmitted=="1")/sum(readmitted=="0"))   
ratio\_age %>% ggplot(aes(x = ratio\_age$age\_mod, y = ratio\_age$ratio)) +   
 geom\_point(size=5) +   
 labs(x="Age",y="Readmission Ratio", title="Readmission Ratio vs Age")



# Collaboration

This is an **individual** assignment. We will only allow private Piazza posts for questions. If there are questions that are generally useful, we will release that information.