Factors that influence Diabetes Patients Hospital Readmission Rates

Yang Wang

18523888

Abstract

There are few national assessments systems of diabetes care during patients' hospitalization. As a result, diabetes' patients' inpatient and readmission raised people's attention. [1]. In this situation, the analysis of factors that impact diabetes patients hospital readmission rates seems necessary. This project is an extraction from a large dataset that includes 130 American hospitals diabetes patients' data ranging from the year 1999 to 2008. Given the readmission status and the data structure, multiple logistic regression was used to perform the analysis. As a comparison, random forest was conducted as well to analyze the factors that pertaining to patients with diabetes. Both the multiple logistic regression and random forest results show similar covariates that influence readmission rate.

1. Introduction

It is increasingly recognized that admission hyperglycemia in the hospitalized patients has significant relevance on outcome, in terms of both morbidity and mortality in critically ill patients.[2] However, there are few assessments systems of diabetes care during hospitalization which could decrease the patients' readmission rate. Since the readmission of diabetes' patients not only cause patient's inconvenience but also could save expensive medical costs. This project is an extraction from a large clinical dataset which includes 101,766 observations and 50 variables with 10 years data in 130

American hospitals. The motivation is to analyze factors related to readmission of patients with diabetes.

2. Background and data

2.1. Data Assembly

This study was submitted by the Center for Clinical and Translational Research and Virginia Commonwealth University, used the Health Facts database (Cerner Corporation, Kansas City, MO), which is a national data warehouse that captures and stores de-identified, longitudinal electronic health record (EHR) patient data in the United States. [1]. The database contains 101,766 observations and 50 variables. The variables include demographics (race, gender and age), diabetes medications, type of admission and admission status.

The dataset I used is a random sample that contains 2000 observations. It's an extract from the big dataset called "diabetic_data", which consists 10 years of clinical observations in 130 American hospitals. Among them, 18 hospitals are from the Midwest, 58 from the northeast, 28 from the south and 16 hospitals from the west.

My dataset was created two steps. First, after analysis of the raw data, decide which variables that seem not relevant to the analysis could be ignored. Second, generate 2000 random observations from the dataset for further analysis.

2.2. Extraction of the Initial Dataset from the Database

The motivation of this project is to find out factors that result in early readmission. The variable "readmitted" indicates the days to inpatient readmission. It is the response that contains 3 levels: "<30" if the patient was readmitted in less than 30 days, ">30" if the

patient was readmitted in more than 30 days, and “No” for no record of readmission. To better analyze the data, I define the variable “readmitted” into two levels: “1” if the patient was admitted within 30 days, “0” means the patient readmitted in more than 30 days and no record of readmission.

The selection of the variables is based on lots of considerations.

First, there is no doubt that the “Encounter ID” and “Patient number” indicates unique identifier should be ignored.

Second, variables “weight”,“payer_code” and “medical_specialty” that have more than 90% missing values may not provide useful information, so they can be deleted.

Third, variables of “Diagnosis 1”, “Diagnosis 2” and “Diagnosis 3” because they have hundreds of levels which may provide little useful information , so I decided to delete them along with the variable “Number of diagnoses”.

Fourth, variables “chlorpropamide”, “acetohexamide” , “tolbutamide” and others are dropped because almost 99% of them only contain one value “NO”. This is meaningless for analyzing the dataset.

Therefore, after above considerations, only 28 variables are left. Then I generate 2000 random observations to get the dataset called “diabete.sub”. For further analysis, two parts of the dataset were split: training and test dataset. So the final dataset have 28 variables and 2000 observations, including 1500 training dataset and 500 test dataset for prediction.

Below is the summary of the final dataset:

```
> dim(diabete.sub)
[1] 2000   28
> summary (diabete.sub)
             race                 gender          age      admission_type_id
 ?              :   0   Female         :1100   [70-80):515   Min.   :1.000
 AfricanAmerican: 402   Male           : 900   [60-70):447   1st Qu.:1.000
 Asian          :  13   Unknown/Invalid:   0   [50-60):359   Median :1.000
 Caucasian      :1518                          [80-90):320   Mean   :2.022
 Hispanic       :  33                          [40-50):176   3rd Qu.:3.000
 Other          :  34                          [30-40): 68   Max.   :8.000
                                               (Other):115
 discharge_disposition_id admission_source_id time_in_hospital num_lab_procedures
 Min.   : 1.000           Min.   : 1.000      Min.   : 1.000   Min.   :  1.00
 1st Qu.: 1.000           1st Qu.: 1.000      1st Qu.: 2.000   1st Qu.: 30.00
 Median : 1.000           Median : 7.000      Median : 3.500   Median : 44.00
 Mean   : 3.672           Mean   : 5.644      Mean   : 4.309   Mean   : 42.09
 3rd Qu.: 3.000           3rd Qu.: 7.000      3rd Qu.: 6.000   3rd Qu.: 55.00
 Max.   :28.000           Max.   :20.000      Max.   :14.000   Max.   :108.00

 num_procedures  num_medications number_outpatient number_emergency  number_inpatient
 Min.   :0.000   Min.   : 1.00   Min.   : 0.000    Min.   : 0.0000   Min.   : 0.000
 1st Qu.:0.000   1st Qu.:10.00   1st Qu.: 0.000    1st Qu.: 0.0000   1st Qu.: 0.000
 Median :1.000   Median :15.00   Median : 0.000    Median : 0.0000   Median : 0.000
 Mean   :1.318   Mean   :15.75   Mean   : 0.381    Mean   : 0.1855   Mean   : 0.642
 3rd Qu.:2.000   3rd Qu.:20.00   3rd Qu.: 0.000    3rd Qu.: 0.0000   3rd Qu.: 1.000
 Max.   :6.000   Max.   :61.00   Max.   :35.000    Max.   :19.0000   Max.   :13.000

 max_glu_serum A1Cresult    metformin    repaglinide   nateglinide   glimepiride
 >200:  40     >7  :  58   Down  :  11   Down  :   0   Down  :   1   Down  :   1
 >300:  22     >8  : 155   No    :1605   No    :1977   No    :1986   No    :1919
 None:1879     None:1687   Steady: 362   Steady:  20   Steady:  12   Steady:  77
 Norm:  59     Norm: 100   Up    :  22   Up    :   3   Up    :   1   Up    :   3



  glipizide     glyburide    pioglitazone  rosiglitazone   acarbose      insulin    change
 Down  :  12   Down  :  10   Down  :   4   Down  :   2   Down  :   0   Down  :241   Ch: 943
 No    :1732   No    :1764   No    :1874   No    :1862   No    :1999   No    :924   No:1057
 Steady: 239   Steady: 206   Steady: 122   Steady: 134   Steady:   1   Steady:615
 Up    :  17   Up    :  20   Up    :   0   Up    :   2   Up    :   0   Up    :220



 diabetesMed   readmitted
 No : 448    Min.   :0.000
 Yes:1552    1st Qu.:0.000
             Median :0.000
             Mean   :0.121
             3rd Qu.:0.000
             Max.   :1.000
```

Fig. 1 Final dataset summary

3. Methods

3.1 Multiple logistic regression

As stated above, the response is a binary variable that contains only 0 and 1, with 28 predictors, therefore, multiple logistic regression is appropriate to fit the relationship between the response “readmitted” and other predictors such as demographics, diabetes medications, type of admission and admission status.

To use multiple logistic regression, the generalized linear model function was used in R，with the code: “glm.fit<-glm(readmitted ~ .,data=diabete.train,family=binomial)”.

To assess whether the covariates were significantly associated with response “readmission”, stepwise function was used to do the model selection. The step function selects variables based on the AIC value, but it does not evaluate the AIC for all possible models but uses a search method that compares models sequentially.

3.2 Random Forest

Random forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.[4]

As discussed above, the diabetes dataset contains 28 variables, some of them are numeric variables, and others are nominal. In addition, the response is a binary variable, so it’s a good idea to use the random forest to check which covariates are significantly associated with the response.

The random forest was performed in R with the code:

“diabete.rf =randomForest(readmitted~.,mtry = 9,importance = TRUE , data = diabete.train)” .

After this, check the importance of each variable, then a decision on model selection will be made.

4. Analysis and Results

4.1 Multiple logistic regression results

The result of Multiple logistic regression analysis uses step function, it’s shown as below:

```
Step:  AIC=1076.39
readmitted ~ admission_type_id + discharge_disposition_id + time_in_hospital +
    num_procedures + number_emergency + number_inpatient + max_glu_serum +
    repaglinide + insulin

                          Df Deviance    AIC
<none>                         1046.4 1076.4
- admission_type_id        1   1048.4 1076.4
- number_emergency         1   1048.5 1076.5
- time_in_hospital         1   1049.9 1077.9
- max_glu_serum            3   1054.3 1078.3
- num_procedures           1   1052.7 1080.7
- insulin                  3   1057.0 1081.0
- discharge_disposition_id 1   1053.4 1081.4
- repaglinide              2   1055.6 1081.6
- number_inpatient         1   1060.3 1088.3
```

Fig. 2 Step function result

As shown above, this is the final part of the step function result, so the final model is: “`readmitted ~ admission_type_id + discharge_disposition_id + time_in_hospital + num_procedures + number_emergency + number_inpatient + max_glu_serum + repaglinide + insulin`”

Variables were removed according to their AIC values. The coefficients and p-values are shown as below:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.07126    1.10291  -3.691 0.000223 ***
admission_type_id         0.09400    0.06459   1.455 0.145566
discharge_disposition_id  0.03648    0.01326   2.751 0.005939 **
time_in_hospital          0.05236    0.02758   1.899 0.057620 .
num_procedures           -0.13322    0.05530  -2.409 0.015993 *
number_emergency          0.12327    0.09021   1.366 0.171810
number_inpatient          0.21881    0.05620   3.893 9.89e-05 ***
max_glu_serum>300         0.43733    1.47227   0.297 0.766435
max_glu_serumNone         1.74873    1.05088   1.664 0.096101 .
max_glu_serumNorm         2.16694    1.09825   1.973 0.048487 *
repaglinideSteady         0.04592    0.77974   0.059 0.953037
repaglinideUp            15.79368  378.16371   0.042 0.966687
insulinNo                -0.47886    0.25003  -1.915 0.055468 .
insulinSteady            -0.30409    0.25857  -1.176 0.239576
insulinUp                 0.26721    0.28528   0.937 0.348935
```

Fig.3 final model summary

The result is not as ideal as I thought, p-values are not as low as usual examples in textbooks. With the training dataset, I calculated the training error rate and test error rate:

```
> glm.pred=rep ("0",1500)
> glm.pred[glm.probs >.5]="1"
> table(glm.pred,readmitted )
        readmitted
glm.pred  0    1
       0 1324  167
       1    7    2
> mean(glm.pred== readmitted )
[1] 0.884
```

From the confusion matrix above, we can get the training error rate is 100-88.4% = 11.6%, which is good. Then I calculated the test error rate:

```
> table(glm.predts,readmitted.ts)
```

readmitted.ts

glm.predts 0 1

0 444 55

1 0 1

> mean(glm.predts== readmitted.ts)

[1] 0.89

> mean(glm.predts!=readmitted.ts)

[1] 0.11

The test error rate is 11 %, which is really good.

4.2 Random Forest Result

After performing the random forest code and importance function, I got the following diagram and plot:

```
> importance(diabete.rf)
                            %IncMSE IncNodePurity
race                     -0.4130028  4.470722e+00
gender                    0.1446166  2.715457e+00
age                       3.5912002  1.431221e+01
admission_type_id         4.4414678  5.039157e+00
discharge_disposition_id  9.7573975  8.876502e+00
admission_source_id       4.2233992  4.667827e+00
time_in_hospital          3.4034716  9.823618e+00
num_lab_procedures        6.7711785  1.852555e+01
num_procedures            2.5199964  5.895621e+00
num_medications           7.3501488  1.571720e+01
number_outpatient         2.0472243  4.353695e+00
number_emergency         11.8749436  4.899494e+00
number_inpatient          8.9039096  9.538210e+00
max_glu_serum             4.2269431  1.623688e+00
A1Cresult                -1.1361454  2.975366e+00
metformin                -0.4457380  2.526201e+00
repaglinide               6.6067001  1.660268e+00
nateglinide               0.0000000  1.863721e-02
glimepiride              -0.5466023  1.544278e+00
glipizide                -2.7262353  2.733837e+00
glyburide                 0.1038074  1.919327e+00
pioglitazone              0.2061819  1.740702e+00
rosiglitazone            -0.5005096  1.068693e+00
acarbose                  0.0000000  5.551115e-19
insulin                   3.2483486  7.139744e+00
change                    0.2239504  1.780678e+00
diabetesMed              -0.3174261  1.297692e+00
```

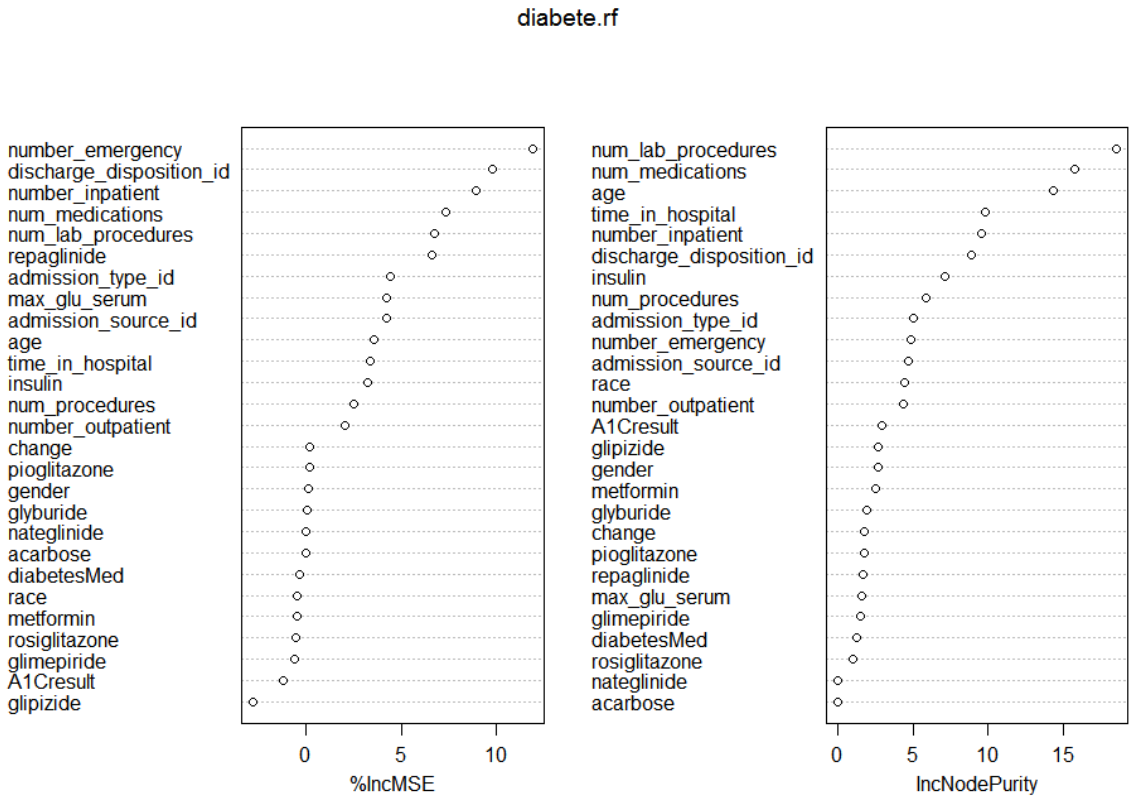

Fig.4 Random forest importance diagram and plot

The results above indicate that across all of the trees considered in the random forest, the “number_emergency”, “discharge_disposition_id” and “number_inpatient” are the three most important variables which are consistent with the multiple logistic regression result. In the random forest, the test set MSE is 0.11, which is pretty low.

5. Conclusions and Discussion

In conclusion, the comparison of multiple logistic regression and random forest results indicate that “number_emergency”, “discharge_disposition_id” and “number_inpatient” are three most significant predictors that are related to the deduction of readmission rates. For further analysis, other predictors such as “Repaglinide” and “admission_type_id” that included in both models could also be used to do the analysis.

Reference

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[2] G. E. Umpierrez, S. D. Isaacs, N. Bazargan, X. You, L. M. Thaler, and A. E. Kitabchi, “Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed diabetes,” Journal of Clinical Endocrinology and Metabolism, vol. 87, no. 3, pp. 978–982, 2002.

[3] Julian J. Faraway, Linear Models with R-Second Edition,2015

[4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008).The Elements of Statistical Learning (2nd ed.). Springer.ISBN 0-387-95284-5.