Readmission Case of Diabetes

By
Sheetal

Executive Summary:

In the past few years there has been lot of the patients have readmitted in the hospital, out of all the diseases people suffering with Diabetes have readmitted more frequent that too under 30days of their discharge. Evertime that happens, the hospital have to provide free service plus compensate the patient. In the year 2011 the hospital has paid 40Million Dollars for just the diabetes patient.

Business Problem:

The business problem is to identify the probable patient which tends to readmit in less than 30 days.

Introduction:

Before I continue about my steps I have taken to identify the steps and the work I have done. Let me explain you what happen when a person is suppose to diabetes.

There are two types diabetes type 1 and type II, both of these types are chronic diseases. Diabetes affects the regulation of glucose in the body, by affecting the release of insulin. Due to which the blood cells cannot absorb glucose.

Type 1: in the type 1 the insulin is not produced.

Type II: In the initial stages the body don't react to the release of insulin and later stages the body stops producing it.

Dataset:

The Major requirement during the data analysis is the dataset and we have a decent dates with almost 1 lakh rows and 50 columns. Let view the data.

Duplicate Row:

Many patients were readmitted in the hospital more than once, and almost all the rows have the same data for the patient. iT look like a redundant data So am removing the duplicate rows and keeping only the first row.

After performing this action only 70K rows are lefts.

NULL Data:

Now lets looks at the NULL data and how we can work on it. Below are the columns having the null data.

```
race 1948
weight 68665
payer_code 31043
```

```
medical_specialty 34477
diag_1 11
diag_2 294
diag_3 1225
```

Out of 70K rows 68K rows have no value at the weight column , so it doesnt make sense to have this data in the system. SImilarly payer code and medical speciality both have missing almost half the rows am deleting Bothe columns.

Gender: Three rows have “unknown” value in the column it doesnt makes sense to keep this rows in system. So deleting this too.

Diag_1,Diag_2 &diag_3: There is one row having no values in all three columns and it doesnt make sense and have no diagonsis data, so removing the row.

Discharge_Disposition_ID: IF discharge_disposition_id is 11 that means patient expired in the hospital so it doesnt make sense to maintain these rows.

Citoglipton & Examide: These two column have same value in all the rows doesnt makes sense to have this data in the system.

Manipulating Records:

I will be amending few rows according to my sense to better use the dataset.

Number of Medication: There are 23 different medication list maintain in the dataset. Not every patient have all the medication up and down . This column will indicate the number of medication used by each patient.

Output :

```
1     31944
0     17197
2     15605
3      5703
4      1011
5        50
6         4
```

Maximum people take only one medication.

Service Utlization: This column basically mean the time and number of visits the patient made to the hospital. It would be a combination of outpatient visit , inpatient visit and emergency visit.

Diag_1,Diag_2 & Diag_3: These three column have more than 200 distinct rows of data. On checking identified that these are ICD codes and these can be segregated into smaller buckets according to the ICD codes and below is the out put.

diag

Code	Description	ID
1-39	Infectious and parasitic disease	3
140-239	Neoplasm	4
240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	5
280-289	Diseases Of The Blood And Blood-Forming Organs	6
290-319	Mental Disorders	7
320-389	Diseases Of The Nervous System And Sense Organs	8
390-459	Diseases Of The Circulatory System	9
460-519	Disease of the Respiratory System	10
520-579	Disease of the Digestive system	11
580-629	Disease of the Genitourinary system	12
630-679	Complications Of Pregnancy, Childbirth, And The Puerperium	13
680-709	Diseases Of The Skin And Subcutaneous Tissue	14
710-739	Diseases Of The Musculoskeletal System And Connective Tissue	15
740-759	Congenital Anomalies	16
760-779	Certain Conditions Originating In The Perinatal Period	17
780-799	Symptoms, Signs, And Ill-Defined Conditions	18
800-999	Injury and Poisoning	19
V01-v91	Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services	1
E000-E999	Supplementary Classification Of External Causes Of Injury And Poisoning	2

Admission_Type_ID: In this data is coded using the number and when read the value 1 ,2 and 7 have the data emergency, Urgent and trauma. All three are same So am replacing the codes to a same code. Similarly 6,5 and 8 have NUll, and NOT mapped. Replacing this data too.

Creating Dummy Variable:

Change: Change has only two distinct values NO and Ch, replacing both with 0 and 1.

Gender: it has two distinct column values Male and Female, replacing them with 1 and 0.

Diabetes_Med: It has two distinct values YES and NO, So replacing them with 0 and 1.
A1Cresult: It has 4 distinct values >7,>8,NORM and NONE. Coding >7 and >8 as 1 and Norm as 0 and NONe as -99.

max_glu_serum: It has four distinct values >200,>300,NORM and NONE. Replacing >200 and >300 as 1 and NONE as -99 and nORM as 0.

Readmitted: it has 3 distinct values >30,<30 ad NO. we are more concern about <30(meaning people beeng admitted in less than 30 days). So we will replace <30 as 1 and other two as 0.

Age: the exact age of the patient is not mentioned the information is given in the range form. So instead of using dummy it would be easy to deal with mean of the values. Replacing the data with mean.

Last but not the least created the dummy values for Diag1,Diag2,Diag3,Dischanrge disposition id and admission id.

Correlation:

Before checking the logistic regression lets have look at the column which are correlated with the readmitted data.

Below are the few columns which are negative correlated.

	correlation
metformin	-0.011851357471157502
diag3_level_5	-0.012057528231431698
discharge_disposition_id_	-0.012499160481959027
diag2_level_12	-0.013771242989716881
diag3_level_12	-0.013930473946817134
discharge_disposition_id_	-0.014137927529364639
diag1_level_12	-0.01444157565548757
diag1_level	-0.015882016077498926
diag1_level_17	-0.027369818768929858
encounter_id	-0.047614512208617614
discharge_disposition_id_	-0.09080102188233706

Positively Correlated data:

correlation-1

num_medications	0.036068014854168245
number_diagnoses	0.044235688226722
age	0.045706056074396546
discharge_disposition_id_	0.04683358592530968
time_in_hospital	0.054759085309681425
discharge_disposition_id_	0.05893715560096389
service_utilization	0.05910074916989753
discharge_disposition_id_	0.0871362469269035
number_inpatient	0.10067520866208235

Very Low correlated data:

Table 1

Column	Corre
admission_source_id_8	6.865132878115949e-05
num_procedures	7.892353137281024e-05
diag2_level_11	-3.857916538835771e-06
diag1_level_11	-7.381036913638643e-05
admission_source_id_2	-9.353666947417277e-05
admission_source_id_9	-0.0006611676555529844
diag3_level_15	-0.0010693181843194927
metformin-pioglitazone	-0.0011801942761892781

Logistic regression:

The first step before setting starting the logistic regression is I created the training set (60000) and valid set (10437).
The first trial of logistic regression had all the columns and it gave the below deviance.

Null deviance: 37458 on 59999 degrees of freedom
Residual deviance: 35601 on 59875 degrees of freedom
AIC: 35851

Training prediction:

predicted	0	1
0	54299	5607
1	49	45

valid dattion pretcition

predicted	0	1
0	9793	641
1	3	0

There were many column coefficient which were insignificant and few columns were interactive model, after many trial and error I ran the model with below columns.

Below is the model I used.

```
Diabetes.model5 <- glm(readmitted ~ diag2_level_17+discharge_disposition_id_24
+diag2_level_9+diag2_level_14+diag2_level_6+discharge_disposition_id_14+discharge_disposition_id_25+d
iag2_level_5+discharge_disposition_id_23+discharge_disposition_id_13+encounter_id+discharge_dispositio
n_id_1+diag2_level_11+discharge_disposition_id_8+discharge_disposition_id_7+discharge_disposition_id_4
+patient_nbr+diag2_level_8+discharge_disposition_id_6+discharge_disposition_id_18+number_outpatient
+nummed+repaglinide+numchange+discharge_disposition_id_2+insulin+diabetesMed+number_emergency+
discharge_disposition_id_15+num_lab_procedures+discharge_disposition_id_28+num_medications+number
_diagnoses+age+discharge_disposition_id_5+discharge_disposition_id_3+service_utilization
                  +  num_medications* num_procedures+
                    number_diagnoses*age+number_diagnoses*time_in_hospital, data=train_data, family =
binomial)
summary(Diabetes.model5)

    Null deviance: 37458  on 59999  degrees of freedom
Residual deviance: 35749  on 59957  degrees of freedom
AIC: 35835

Number of Fisher Scoring iterations: 6
```

The coefficient and other details are attached in txt file please refer that.

Lets look at the prediction data:

Training Data:

predicted	0	1
0	54305	5624
1	43	28

```
predicted    0    1
        0 9794  641
        1    2    0
```

Now in the final model number of columns used are reduced from 124 to 42, by not reducing the quality of the model.

Lets look at below test:

Pchiq Test:

```
 pchisq(35749, df = 59957, lower.tail = F)
[1] 1
```

According to the chi test we cannot reject the null hypothesi the Null hypothesis says that the models valid.

Hoslem Lemeshow-

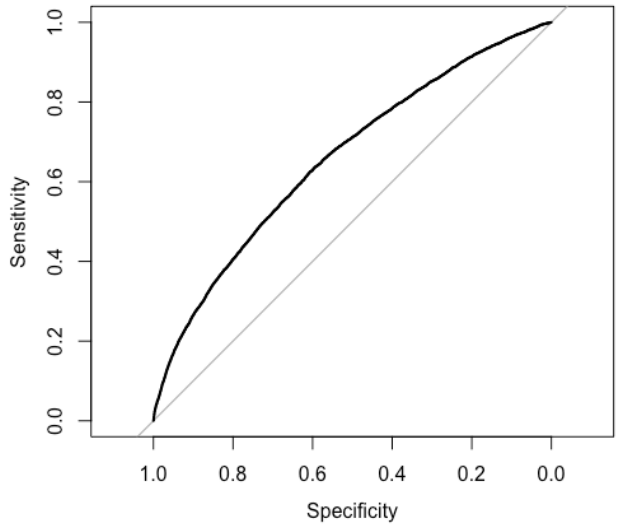
```
> hoslem.test(train_data$readmitted, fitted(Diabetes.model5), g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  train_data$readmitted, fitted(Diabetes.model5)
X-squared = 15.612, df = 8, p-value = 0.04827
```

Even the hoslem test for good ness of fit suggest that we accept the model.

ROC CURVE:



The Roc curve as about 0.65 area under the curve which is a good reason to accept the curve.

Conclusion:

The specificity for Both the training data and validation data is above 0.92, and the sensitivity is less than 0.005. So the model which I have created is good enough to predict if the patient will not be readmitted in the hospital but it is not good for reverse.