Predicting Hospital Readmittance among Diabetes Patients

Puneeth Nikin Krishnan

10/05/2020

Introduction

Diabetes is a metabolic disease that leads to high blood sugar in Patients. As of 2017 425 million people suffered from this disease. Hospital Readmission is a major concern among patients whose diabetes is poorly controlled. This leads to higher hospital bills among these patients as well as higher morbidity and mortality rates.Consequently identifying patients with high probability of readmittance will help hospitals to curate their treatment accordingly.

In this study, analysis of data from 130 Hospitals accross the US for 10 years (1999-2008) is conducted. Based on this data a model will be developed to predict whether a patient will be readmitted within 30 days, after 30 days or never. This is a classification problem. Three models will be built, namely Linear Discriminant Analysis Quadratic Discriminant Analysis and K Nearest Neighbors(KNN) and the best performing model among them will be identified.

About the Dataset

The dataset is part of the UCI MAchine Learning Repository and named " Diabetes 130-US hospitals for years 1999-2008 Data Set". Here is a ***link*** to that dataset. The dataset is first imported,

The dataset comprises of records of patients. Each record of a patient comprises of 50 data points.

Table 1: Names of Dataset Columns

encounter_id	payer_code	diag_3	glipizide	citoglipton
patient_nbr	medical_specialty	number_diagnoses	glyburide	insulin
race	num_lab_procedures	max_glu_serum	tolbutamide	glyburide-metformin
gender	num_procedures	A1Cresult	pioglitazone	glipizide-metformin
age	num_medications	metformin	rosiglitazone	glimepiride-pioglitazone
weight	number_outpatient	repaglinide	acarbose	metformin-rosiglitazone
admission_type_id	number_emergency	nateglinide	miglitol	metformin-pioglitazone
discharge_disposition_id	number_inpatient	chlorpropamide	troglitazone	change
admission_source_id	diag_1	glimepiride	tolazamide	diabetesMed
time_in_hospital	diag_2	acetohexamide	examide	readmitted

Demographics and General Patient Information

The patients Demographic information such as Age, Sex and Gender are included, along with information such as encounter id, patient number and Weight.

Table 2: Admission Type ID Description

admission_type_id	description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

Admission Type ID

admission_type_id - represents the admission type and description of the same is as follows.

Discharge Disposition ID

discharge_disposition_id - The discharge disposition refers to where the patients were discharged to,e.g home,emergency etc. Here is a description of all the id's

Table 3: Discharge Disposition ID

discharge_disposition_id	decription
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare approved swing bed
16	Discharged/transferred/referred another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	Expired at home. Medicaid only, hospice.
20	Expired in a medical facility. Medicaid only, hospice.
21	Expired, place unknown. Medicaid only, hospice.
22	Discharged/transferred to another rehab fac including rehab units of a hospital .
23	Discharged/transferred to a long term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25	Not Mapped
26	Unknown/Invalid
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere
27	Discharged/transferred to a federal health care facility.
28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).

Admission Source ID

admission_source_id - The admission source refers to the the source of the patient,e.g referals, transfer etc.Here is a description of all the id's.

Table 4: Admission Source ID

admission_source_id	description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critial access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital inpt/same fac reslt in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice

Time in Hospital

time_in_hospital - refers to the number of days from admission to discharge.

Payer Code

payer_code - corresponds to mode of payment e.g self pay, blue cross,medicare etc.

Medical Speciliality

medical_speciality - The type of physician who is admitting the patient

Numerical Features

num_lab_procedures - Number of lab tests performed.
number_procedures - Number of procedures other than lab tests.
num_medications - Number of Medications prescribed during visit.
num_outpatient - Number of outpatient visits.
number_emergency - Number of emergency visits in the preceding year.
number_inpatient - Number of inpatient visits in the preceding year.
number_diagnosis - number of diagnosis in the system.

Diagnosis

diag_1 - Primary diagnosis
diag_2 - Secondary diagnosis
diag_3 - Additional Secondary diagnosis if any.

The diagnosis are coded as ICD - 9 codes. A table of classification for the ICD-9 codes can be web-scraped from the website https://icd.codes/icd9cm . The library rvest is used for this purpose.

Table 5: ICD9 Codes

Chapter	Code Range	Description
1	001-139	Infectious And Parasitic Diseases
2	140-239	Neoplasms
3	240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
4	280-289	Diseases Of The Blood And Blood-Forming Organs
5	290-319	Mental Disorders
6	320-389	Diseases Of The Nervous System And Sense Organs
7	390-459	Diseases Of The Circulatory System
8	460-519	Diseases Of The Respiratory System
9	520-579	Diseases Of The Digestive System
10	580-629	Diseases Of The Genitourinary System
11	630-679	Complications Of Pregnancy, Childbirth, And The Puerperium
12	680-709	Diseases Of The Skin And Subcutaneous Tissue
13	710-739	Diseases Of The Musculoskeletal System And Connective Tissue
14	740-759	Congenital Anomalies
15	760-779	Certain Conditions Originating In The Perinatal Period
16	780-799	Symptoms, Signs, And Ill-Defined Conditions
17	800-999	Injury And Poisoning
18	V01-V91	Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services
19	E000-E999	Supplementary Classification Of External Causes Of Injury And Poisoning

Glucose Serum Test

max_glu_serum - refers to Glucose Serum test results
values-
“>200” - greater than 200 but less than 300
“>300” - greater than 300
“normal,” - less than 200
“none”- test not conducted

A1Cresult

A1Cresult - a1c test result reflects average blood sugar levels over 3 months, also refered to as hemoglobin A1C test.
Values - “>8” - greater than 8%
“>7” - greater than 7% less than 8%
“normal” - less than 7%
“none” - test not conducted

Medications

24 features represent type of medications. Values denote change in levels during encounter.

features- metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone

values -
“Up” - dosage increase
“Down” - dosage decreased
“Steady” - dosage maintained
“No” - Not prescribed

Change

change - change in medication
Values-
“Ch” - Change
“No” - No Change

Diabetes Medications

diabetesMed- Patient prescribed diabetes Medicine
Values-
“Yes”- medications prescribed
“No” - medications not prescribed

readmitted

The classification to be predicted
Values-
“No” - Not readmitted
“<30”- Readmitted within 30 days
“>30”- Readmitted after 30 days

Steps in Modeling

1) Cleaning and Preprocessing the dataset - This step involves identifying missing values, converting characters to factors where approprite, scaling and centering numerical features,grouping or removing levels in features that contain sparse data and identifying features with zero variablilty or near zero variability.

2) Split the Dataset- We split the dataset into two. One set is used for validation in the end and is assumed to be unseen. The remaining data is used to build the model.

3) Analysing the features- Identify relationship between features and predicted value.

4)Modelling - Build three models namely K nearest Neighbors, Linear Discriminant Analysis and Quadratic Discriminant Analysis and identify best performing model.

Cleaning and Preprocessing the Dataset

The diabetic_data dataframe cannot be directly used for modelling as most of the data needs to be reorganised. This is because some features do not have enough data while others are not converted to factor.

Missing Values

We notice that the feature weigh has 96.8% of its data missing, medical_speciality has 49% of data missing and payer_code has 39% of data missing. We drop these columns/features.

Table 6: Proportion of Missing Values

x	proportion_missing
weight	0.9685848
medical_specialty	0.4908221
payer_code	0.3955742
race	0.0223356
diag_3	0.0139831
diag_2	0.0035179
diag_1	0.0002064

repetition of patient encounters

For some patientxs there are multiple records. We only keep one record per patient.

Zero Variance/ Near zero Variance

Some features in our dataset do not have sufficient variability. For e.g examide has only one level i.e 'no' consequently it has no predictive power. Similarly there could be features whose variability is significantly low. The Caret package provides with nearZeroVar() function that can be used to identify and remove these features.

Table 7: Features with zero Variability or near zero Variability

feature	freqRatio	percentUnique	zeroVar	nzv
max_glu_serum	39.31947	0.0055930	FALSE	TRUE
repaglinide	85.66990	0.0055930	FALSE	TRUE
nateglinide	149.83122	0.0055930	FALSE	TRUE
chlorpropamide	1066.35821	0.0055930	FALSE	TRUE
glimepiride	20.09875	0.0055930	FALSE	TRUE
acetohexamide	71517.00000	0.0027965	FALSE	TRUE
tolbutamide	3763.10526	0.0027965	FALSE	TRUE
acarbose	371.43750	0.0041947	FALSE	TRUE
miglitol	3972.11111	0.0055930	FALSE	TRUE
troglitazone	23838.33333	0.0027965	FALSE	TRUE
tolazamide	2382.93333	0.0027965	FALSE	TRUE
examide	0.00000	0.0013982	TRUE	TRUE
citoglipton	0.00000	0.0013982	TRUE	TRUE
glyburide-metformin	144.63544	0.0055930	FALSE	TRUE
glipizide-metformin	10215.85714	0.0027965	FALSE	TRUE
glimepiride-pioglitazone	0.00000	0.0013982	TRUE	TRUE
metformin-rosiglitazone	35758.00000	0.0027965	FALSE	TRUE
metformin-pioglitazone	71517.00000	0.0027965	FALSE	TRUE

classifying diagnosis by ICD9 codes

One other significant problem is some features like diag_1, diag_2 and diag_3 have too many levels when converted to factors and this involves costly computation while fitting models. To avoid this we reduce number of levels by classifying to ICD9 classifications. We also seperate codes pertaining to 250.xx as they are specifically related to diabetes.

Grouping/Removing Sparse data

We identify features which have sparse data in their levels and combine them with other levels or remove them. This is done to avoid rank deficiency errors while modeling.

Race -

Table 8: level count race

race	n
Asian	497
Other	1178
Hispanic	1517
?	1948
AfricanAmerican	12887
Caucasian	53491

Race has sufficient data in each level.

Gender-

Table 9: level count Gender

gender	n
Unknown/Invalid	3
Male	33490
Female	38025

Gender has a level "Unknown/Invalid" with only three data points we remove this level.

Age-

Table 10: level count Age

age	n
[0-10)	154
[10-20)	535
[20-30)	1127
[90-100)	1900
[30-40)	2699
[40-50)	6878
[80-90)	11589
[50-60)	12466
[60-70)	15959
[70-80)	18208

The levels [0-10) and [10-20) can be merged to [0-20) as the data is sparse in level [0-10).

Admission Type ID -

The levels 4,7 and 8 are sparse and therefore merged with level 8.

Discharge Disposition ID -

The IDs 20,12,16,27,10,19,17,9,24,15,8 and 28 have sparse data and needs to be handled. We therefore merge it with level 25. Level 11 correspods to people who died. We remove this level.

Table 11: level count Admission Type

admission_type_id	n
4	9
7	21
8	291
5	3174
6	4588
2	13028
3	13916
1	36488

Table 12: level count Discharge disposition ID

discharge_disposition_id	n
20	1
12	2
16	3
27	3
10	6
19	6
17	8
9	9
24	25
15	40
8	73
28	90
14	218
13	243
23	260
7	409
4	541
25	778
5	913
11	1077
22	1409
2	1539
18	2474
6	8289
3	8784
1	44315

Admission Source ID-

Sparse data in admission source for levels11,13,14,25,10,22 and 8. Removing these levels.

Table 13: level count Admission Source

admission_source_id	n
11	1
13	1
14	2
25	2
22	4
10	7
8	11
9	95
3	136
20	153
5	514
2	909
6	1788
4	2541
17	4858
1	21849
7	37567

Diagnosis Categories-

Table 14: level count diagnosis

category_diag_1	n	category_diag_2	n	category_diag_3	n
E000-E999	1	740-759	83	740-759	75
?	11	?	293	630-679	272
740-759	41	630-679	352	E000-E999	932
630-679	585	E000-E999	570	140-239	1217
280-289	654	320-389	900	?	1224
320-389	863	V01-V91	1222	001-139	1236
V01-V91	919	001-139	1248	320-389	1240
290-319	1542	710-739	1295	710-739	1372
001-139	1704	140-239	1685	800-999	1413
680-709	1780	800-999	1829	680-709	1508
240-279	1861	290-319	1858	280-289	1732
140-239	2658	280-289	2083	290-319	2148
580-629	3450	680-709	2243	520-579	2459
710-739	4071	520-579	2722	V01-V91	2588
800-999	4702	780-799	3181	780-799	3112
780-799	5512	580-629	5089	580-629	3830
250	5764	240-279	5652	460-519	4299
520-579	6348	460-519	6510	240-279	6459
460-519	6523	250	9708	250	12574
390-459	21421	390-459	21887	390-459	20720

For category_diag_1 we notice that “E000-E999”,“?” and “740-759” are sparse. we remove them. We perform a similar exercise on category_diag_2 and category_diag_3 but instead of removing we group it with “?”.

Centering and Scaling the Numerical Predictors

The numerical predictors are standardized.This is done so that all the numerical precictors have a mean zero and can be compared easily with other predictors which have different units. To standardise the data we subtract the mean and divide by the standard deviation.

factorising the categorical data

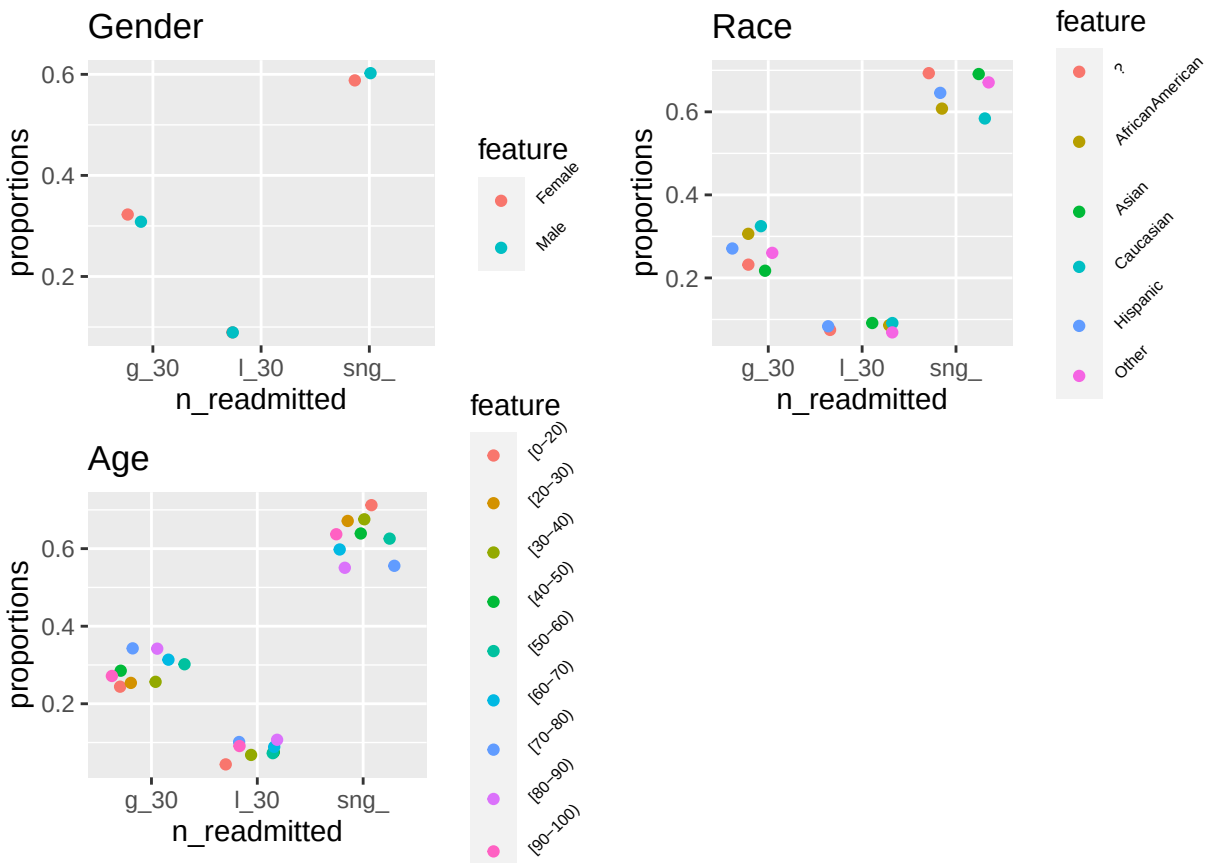
All the categorical data are mentioned as characters.However, since they are categorical data they need to be converted to factors.

Split the Dataset

The dataset is split into two. One part is the validation data which is 10% of the total dataset. This will be used in the end to validate our model.The remaining 90% data will be used to build the model.

Analysing the data

Analysis of Readmitted with Race,Gender and Age

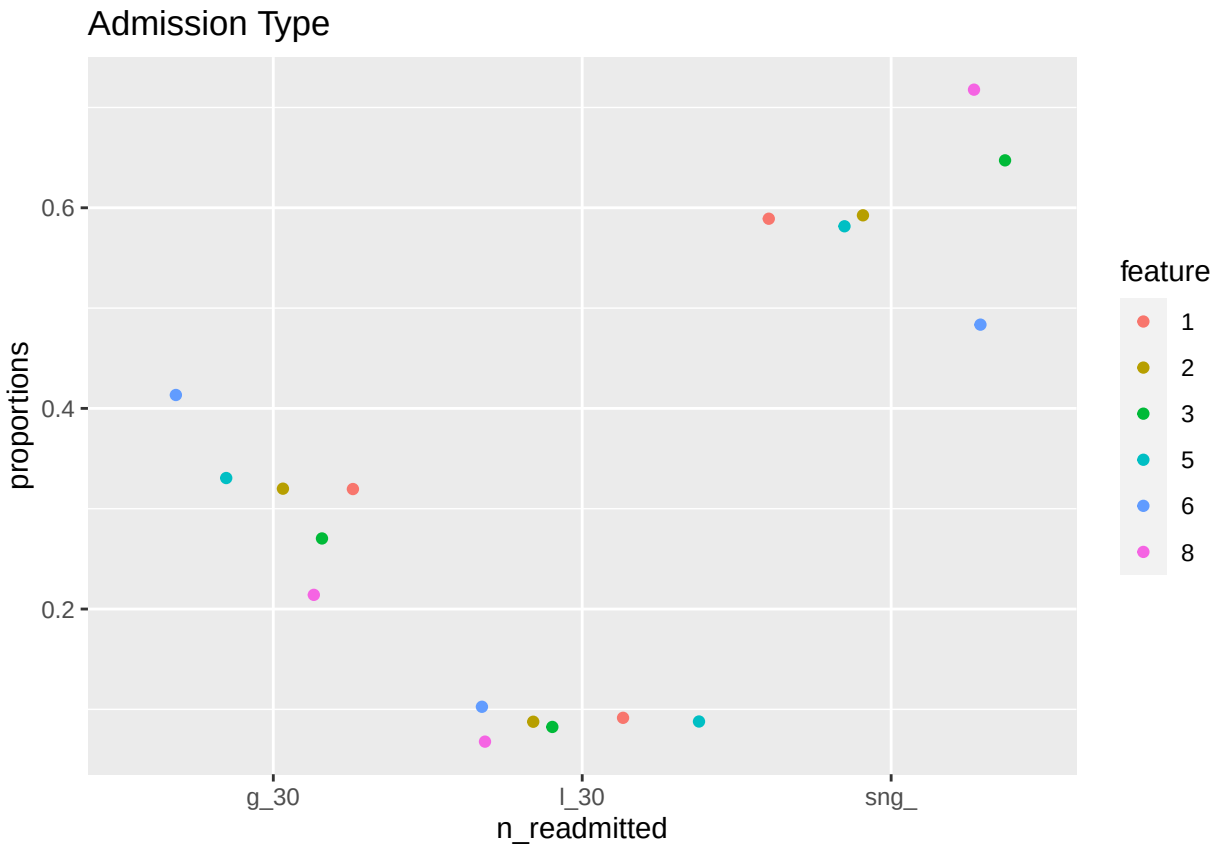


Gender- Women have a higher proportion of readmission than men.

Race- Asians have the lowest proportion of readmission among differeng races. The caucasian and African American population have a higher proportion of readmission

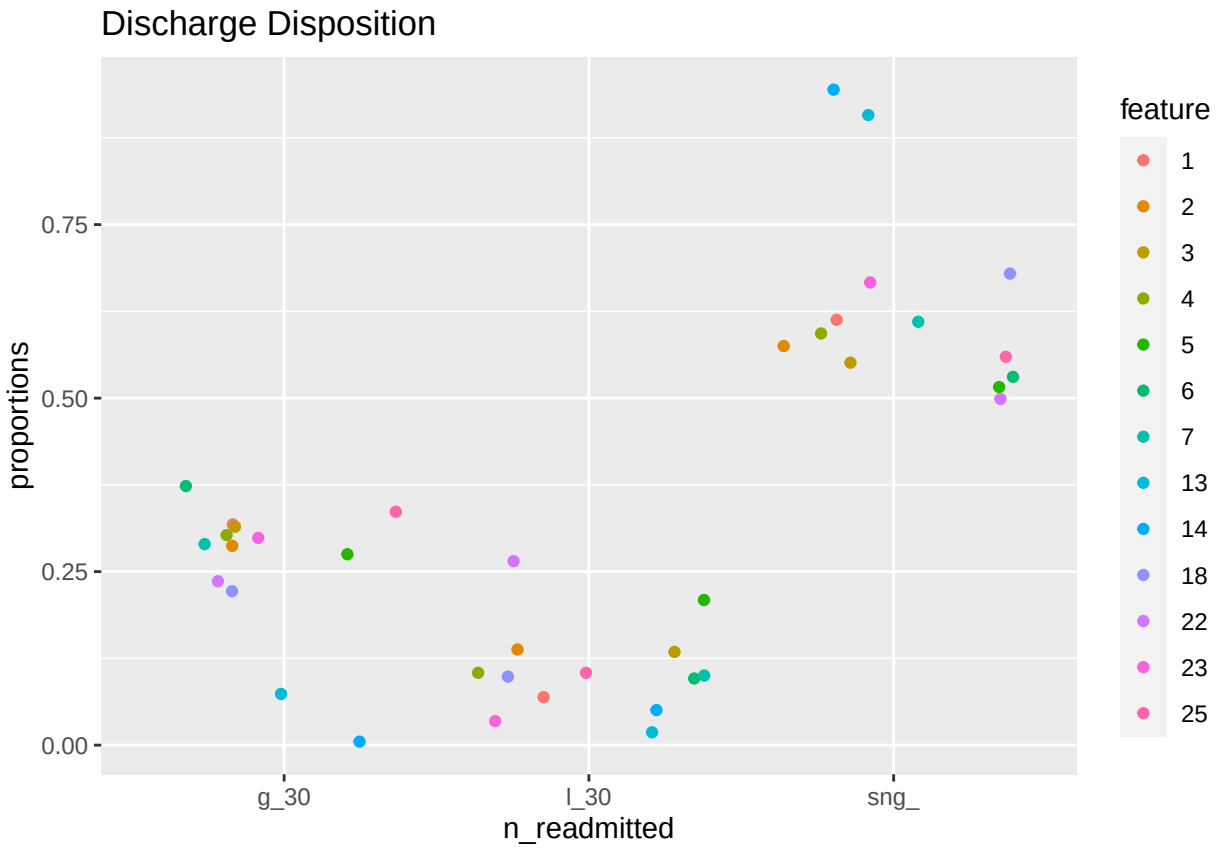
Age- People over the age of 40 have a higher proportion of readmission. It is also noticeable that older people have a higher probability of getting readmitted especially people aged between 60-90.

Analysys of readmitted with Admission type



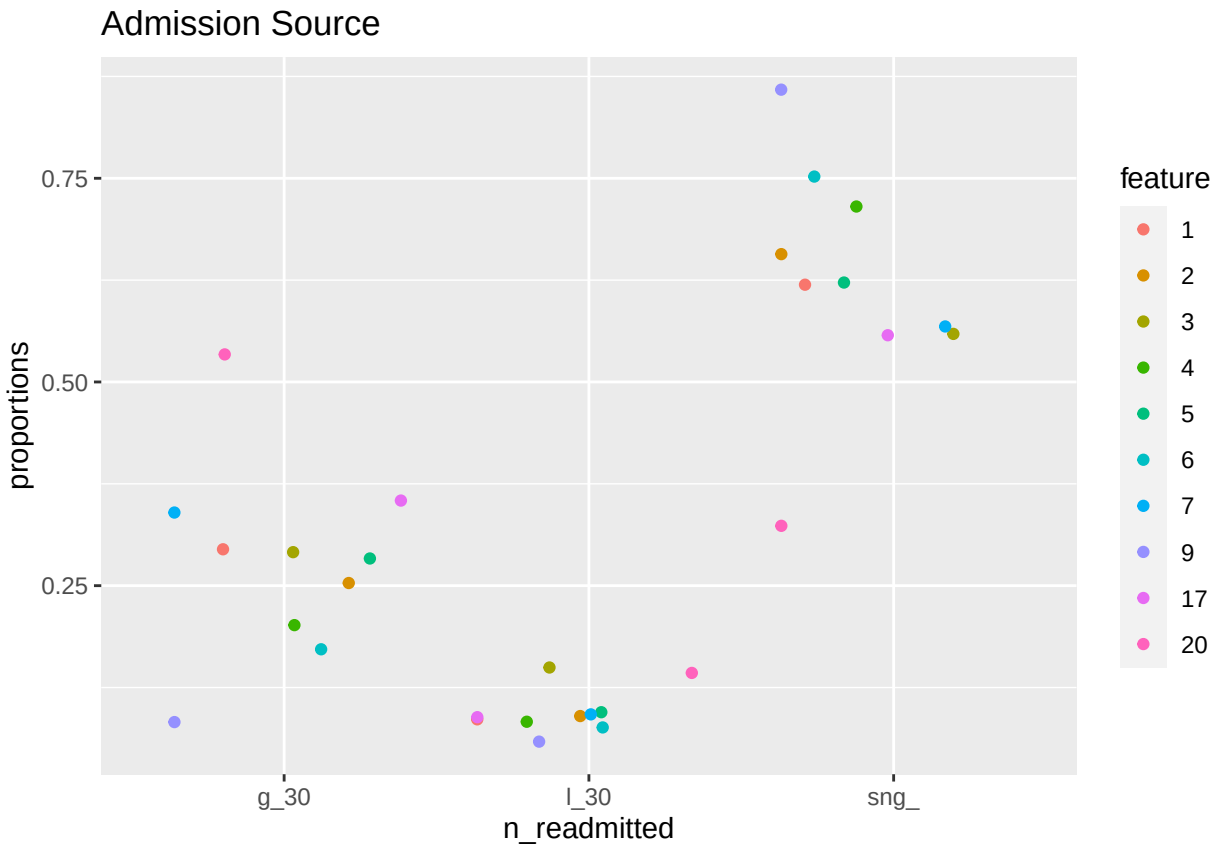
There does appear to be a relationship between admission type and readmission with category 3 showing showing lower readmission while category.(6 and 8 are null and not mapped.)

Analysis of readmitted with discharge Disposition



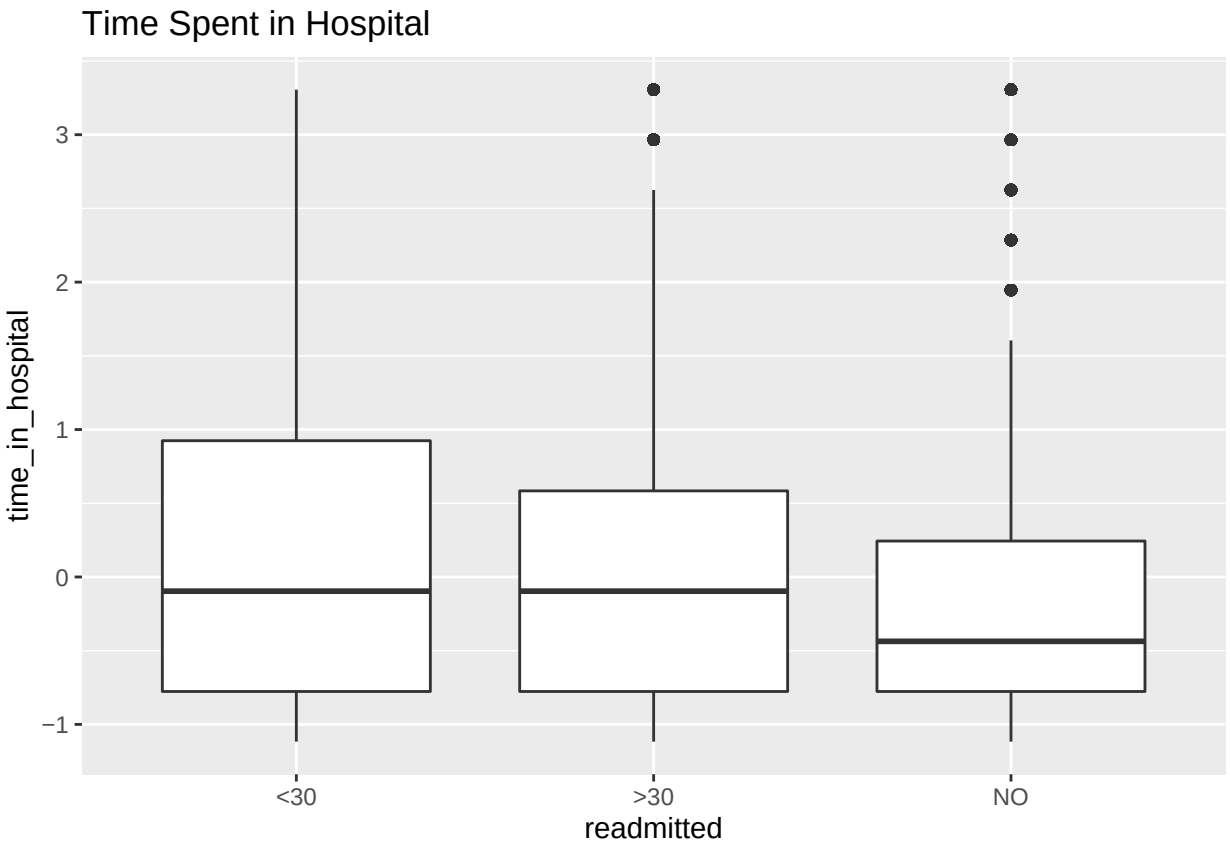
Discharge disposition definitely seems to have a relationship with readmitted. As can be noticed categories 13 and 14 have a low readmittance rate whereas category 22 has a high readmittance rate within 30 days.

Analysis of readmitted with admission source



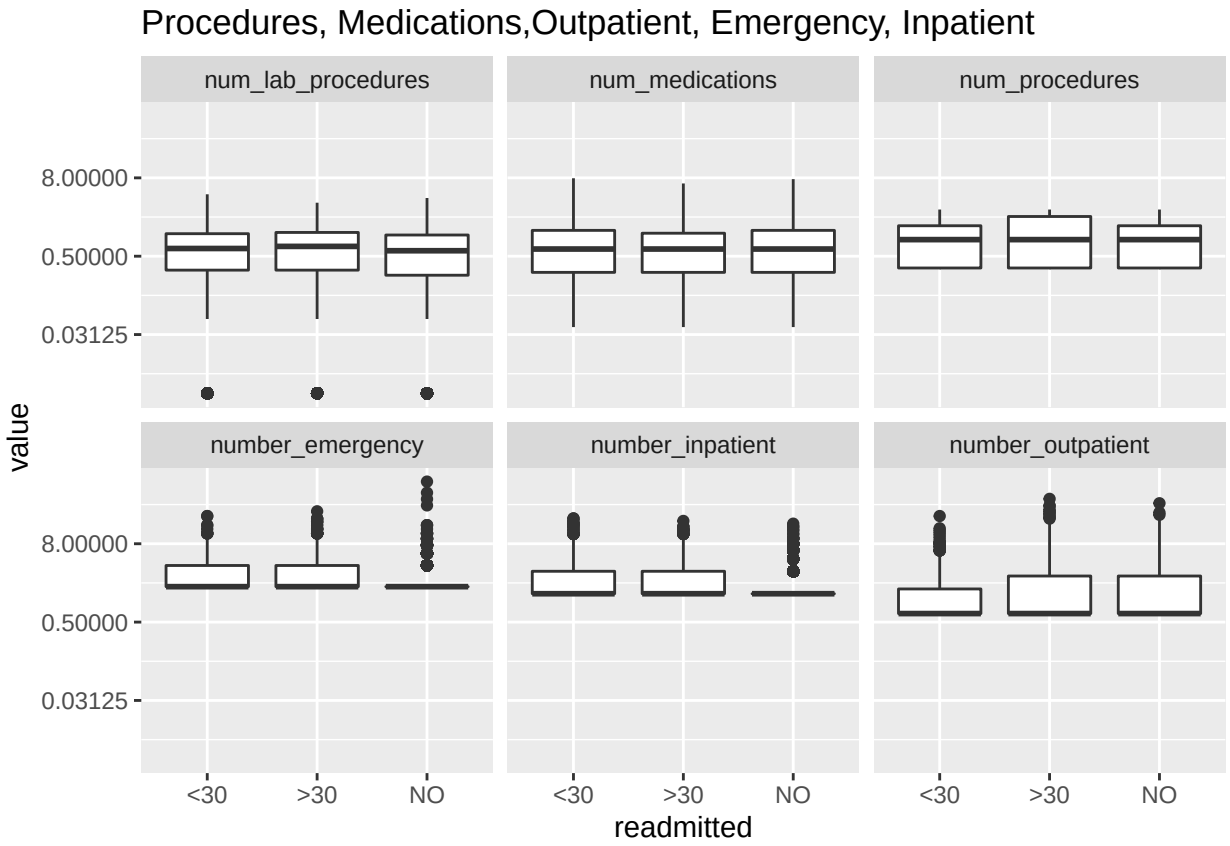
There is a relationship between admission source and readmitted. Category 3 has a high chance of getting admitted within 30 days. Category 7 has a high chance of getting readmitted after 30 days.

analysis of readmitted and time spent in hospital



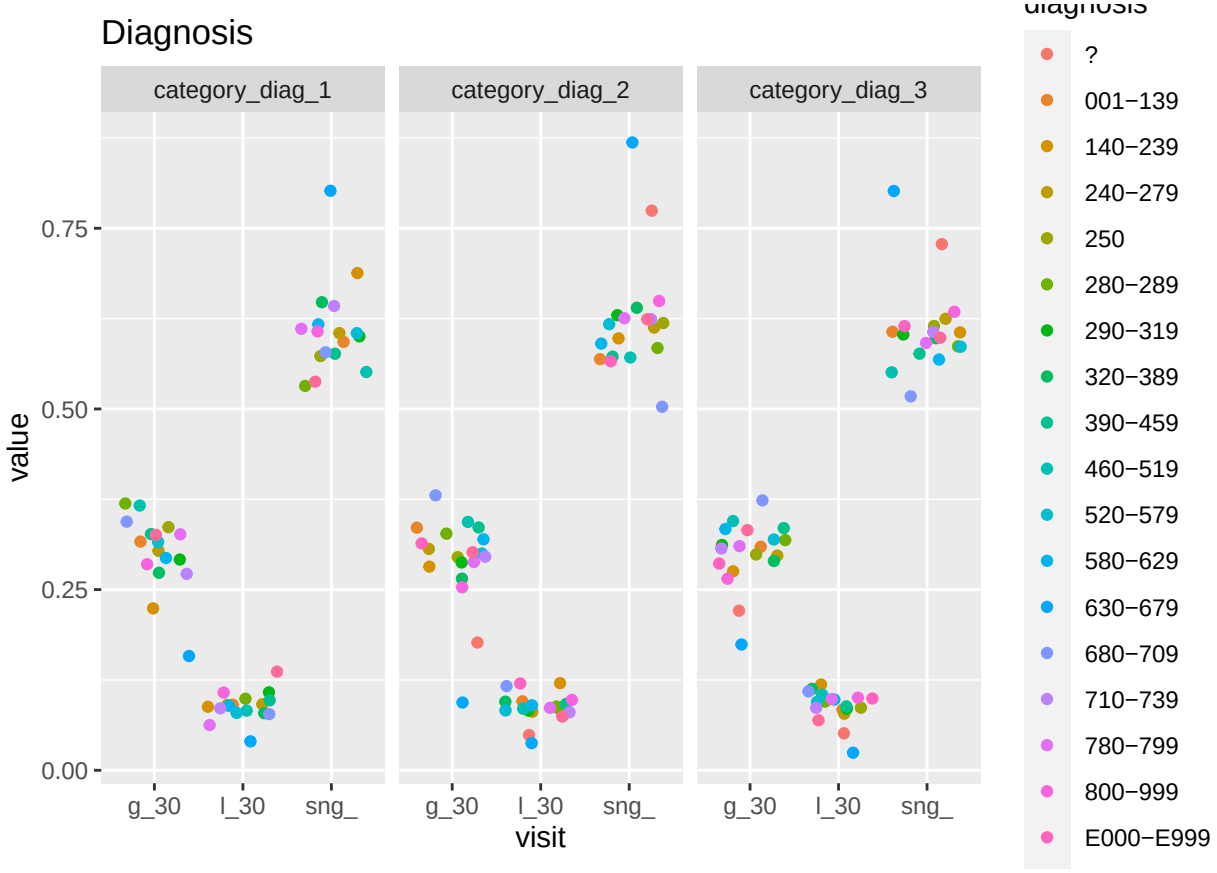
There is a relationship between time in hospital and readmitted as the median time spent in hospital for those who were readmitted is higher.

analysis of readmitted with number of lab procedures,number of procedures,number of medications, number of outpatient, number of emergency and number of inpatient



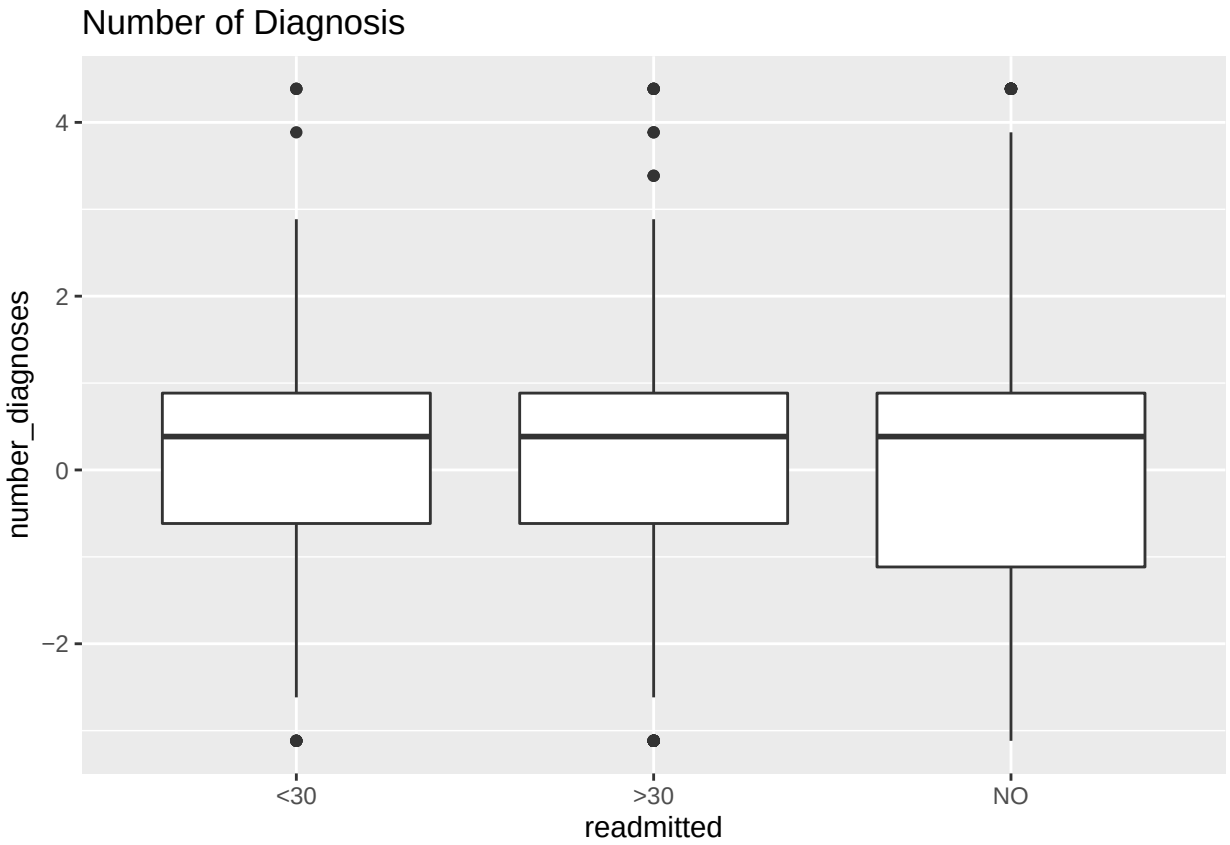
There does not seem to be a clear relationship between between tsome numerical predictors and readmission.In the case of num_lab_procedures a small reduction in median when not readmitted is noticed. number_emergency and number_inpatient have difference in interquartile distance for those readmitted and for those that are not.

analysis of readmitted with category_diag_1, category_diag_2 and category_diag_3



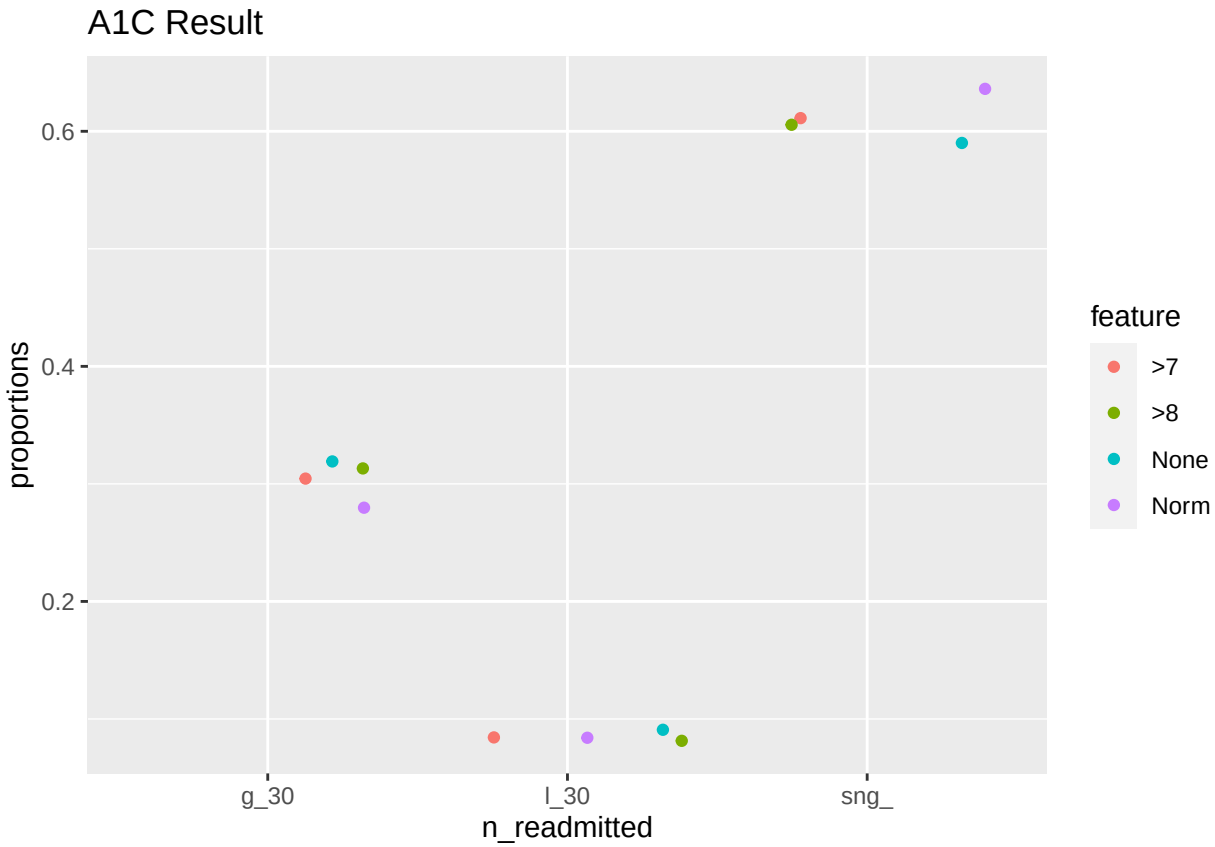
There appears to be a relation between diagnosis and readmission.Some categories like “630-679” have low readmission rate whereas “280-289”,“460-519” and “680-709” have high readmission rate.

analysis of readmitted with number of diagnosis



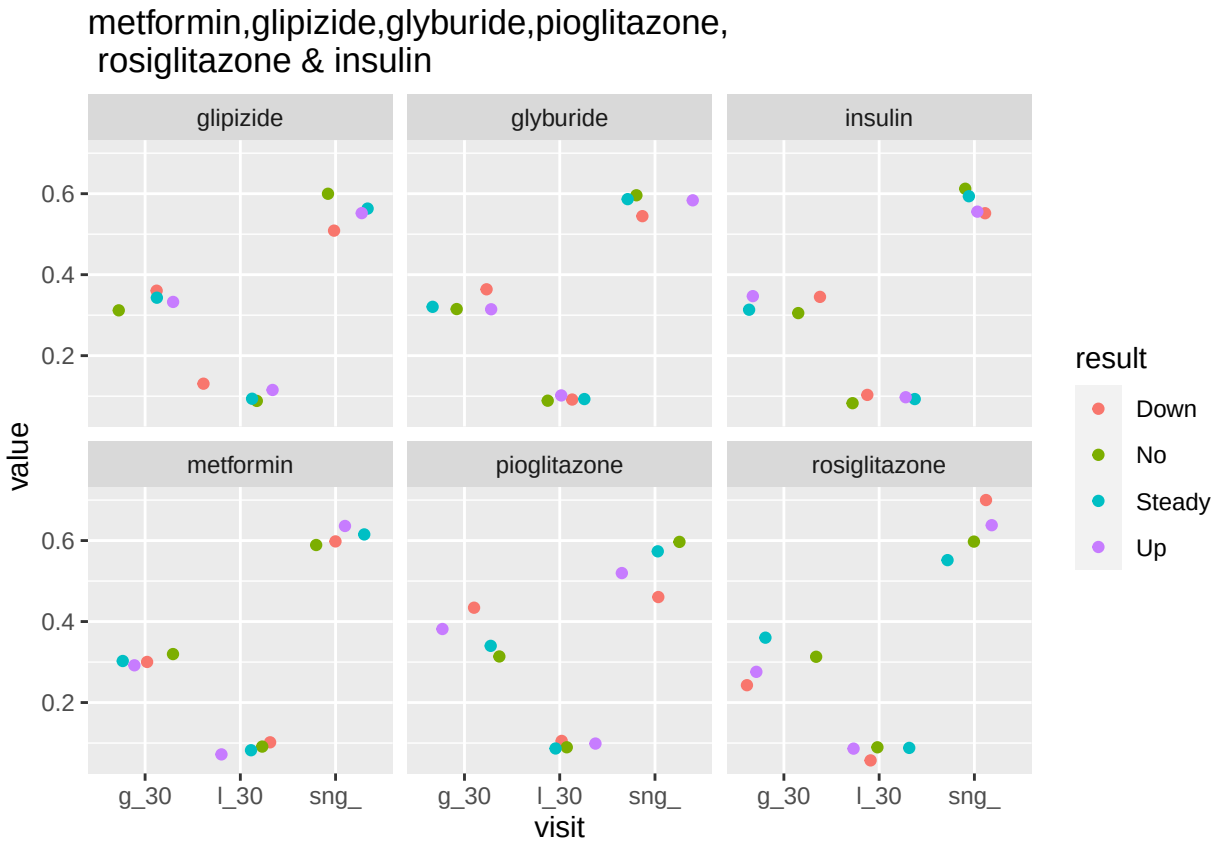
The number of diagnosis does not seem to have a significant impact on the readmission rate except a larger interquartile range for those not readmitted.

analysis of readmitted with A1Cresult



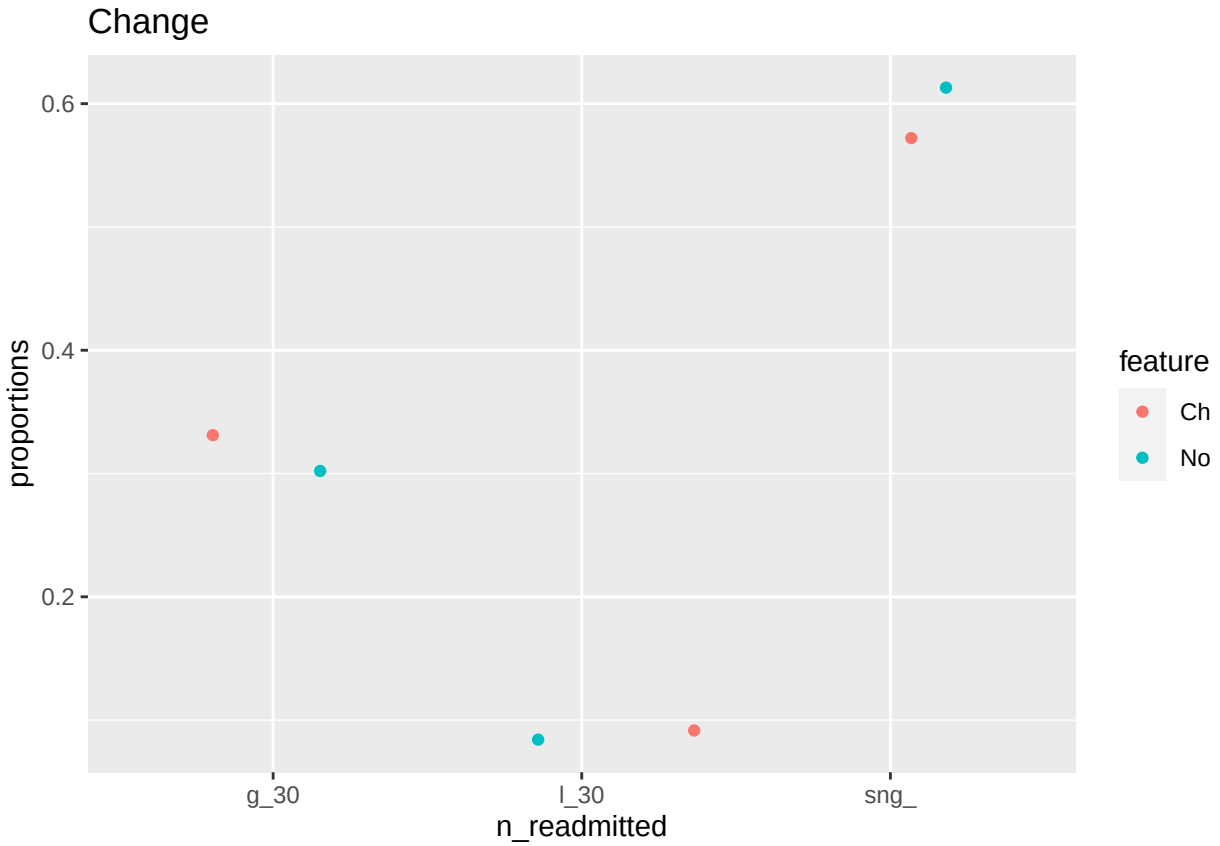
A1Cresult is related to readmission with a lower readmission rate for those with a normal report.

tests with metformin,glipizide,glyburide,pioglitazone,rosiglitazone & insulin



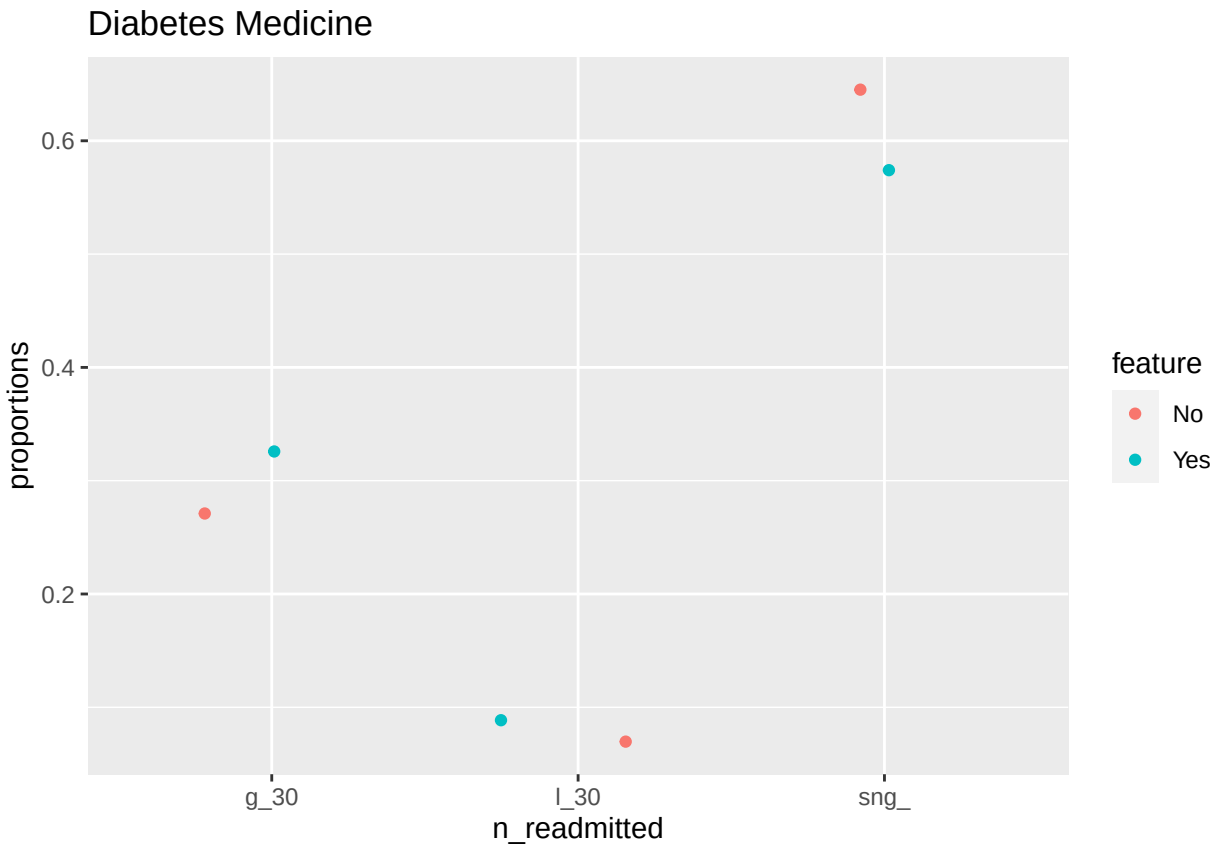
There appears to be a relation between the change in medications and readmission though not significant. There is also a different result for each medication.

analysis of readmitted with change



Patients with a change in medications have a higher chance of readmission than otherwise.

analysis of readmitted with diabetes medicine



Patients who are on diabetic medicine have a higher chance of readmission than otherwise.

Modeling

split data to train and test

The data is split into train and test set. The test set comprises of 30% of the diabetes data and will be used to test the model performance. The train set will be used to train the data. The train set has been kept large so that the model has more data to train on.

Linear Discriminant Analysis (LDA)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   <30   >30    NO
##        <30    98   104   144
##        >30   229   761   560
##        NO   1371  5137 10594
##
## Overall Statistics
##
##                Accuracy : 0.6029
##                  95% CI : (0.5959, 0.6098)
##     No Information Rate : 0.5947
##     P-Value [Acc > NIR] : 0.01115
##
##                   Kappa : 0.0917
##
##  Mcnemar's Test P-Value : < 2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity            0.057715    0.12679    0.9377
## Specificity            0.985665    0.93929    0.1548
## Pos Pred Value         0.283237    0.49097    0.6195
## Neg Pred Value         0.914218    0.69962    0.6287
## Prevalence             0.089378    0.31593    0.5947
## Detection Rate         0.005158    0.04006    0.5576
## Detection Prevalence   0.018212    0.08159    0.9002
## Balanced Accuracy      0.521690    0.53304    0.5462
```

The accuracy is 60% . The notable feature is that the sensitivity is low for both "<30" and ">30" classes. This can be attributed to the low prevalence in these classes especially in "<30" class.

Quadratic Discriminant Analysis (QDA)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  <30  >30   NO
##        <30  453 1018 1830
##        >30 1014 4162 6616
```

```
##        NO   231  822 2852
##
## Overall Statistics
##
##                Accuracy : 0.393
##                  95% CI : (0.3861, 0.4)
##     No Information Rate : 0.5947
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0888
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity             0.26678     0.6934    0.2524
## Specificity             0.83538     0.4129    0.8632
## Pos Pred Value          0.13723     0.3530    0.7303
## Neg Pred Value          0.92069     0.7447    0.4404
## Prevalence              0.08938     0.3159    0.5947
## Detection Rate          0.02384     0.2191    0.1501
## Detection Prevalence    0.17376     0.6207    0.2055
## Balanced Accuracy       0.55108     0.5532    0.5578
```

The accuracy is 39.3%.However the sensitivity for classes "<30" and ">30" have gone up. This is significant due to the fact that predicting whether a patient will get readmitted is more important.

K nearest neighbors (KNN)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   <30   >30    NO
##        <30     0     0     0
##        >30   222   638   438
##        NO   1476  5364 10860
##
## Overall Statistics
##
##                Accuracy : 0.6052
##                  95% CI : (0.5982, 0.6122)
##     No Information Rate : 0.5947
##     P-Value [Acc > NIR] : 0.001579
##
##                   Kappa : 0.0697
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity             0.00000    0.10630    0.9612
```

```
## Specificity             1.00000    0.94922    0.1117
## Pos Pred Value              NaN    0.49153    0.6136
## Neg Pred Value          0.91062    0.69695    0.6626
## Prevalence              0.08938    0.31593    0.5947
## Detection Rate          0.00000    0.03358    0.5716
## Detection Prevalence    0.00000    0.06832    0.9317
## Balanced Accuracy       0.50000    0.52776    0.5365
```

5 fold cross validation is performed to reduce run time. The tuning parameter was tested over larger number of 'k' starting from 5, however the accuracy kept rising. Here a small set is shown for representation. KNN does not have any predictive power in this case as it has very low or nearly zero sensitivity for classes"<30" and ">30". (Note- The KNN model takes a long time to run.)

Result.

We choose the Quadratic Discriminant Analysis(QDA) as the best model for predictions due to its higher sensitivity for crucial classes.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  <30  >30   NO
##        <30  142  328  583
##        >30  389 1588 2516
##        NO    98  307 1086
##
## Overall Statistics
##
##                Accuracy : 0.4002
##                  95% CI : (0.3887, 0.4117)
##     No Information Rate : 0.5947
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0897
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: <30 Class: >30 Class: NO
## Sensitivity             0.22576     0.7143    0.2595
## Specificity             0.85783     0.3966    0.8580
## Pos Pred Value          0.13485     0.3534    0.7284
## Neg Pred Value          0.91862     0.7504    0.4412
## Prevalence              0.08938     0.3159    0.5947
## Detection Rate          0.02018     0.2257    0.1543
## Detection Prevalence    0.14964     0.6385    0.2119
## Balanced Accuracy       0.54179     0.5555    0.5587
```

The overall accuracy is 40%. The model has High sensitiviy towards class ">30" which is 0.71. The sensitivity towards class "<30" is 0.22 owing to its low prevalence. However higher than the other two models.This is owed to the fact that QDA has a quadratic decision boundary.

Conclusion

In conclusion, it is apparent that with real world data it is crucial to clean and pre-process the dataset prior to fitting the model. It is also clear that it is important to identify only those features which have predictive power before fitting the model else the computation can be expensive and at the same time not doing so can affect the results as well. Among the models KNN has the highest accuracy. However this is misleading as predicting every patient as not going to be readmitted alone will give us such a high accuracy,without fitting a model.This can be attributed to the fact that class "NO", which corresponds to not readmitted, has high prevalence. Therefore we will use sensitivity as a metric to identify the best performing model. In the case of LDA and QDA, QDA has a higher sensitivity to classes "<30" and ">30", which in this case are the important classes. This could be the case as LDA assumes that the observations have a normal distribution and have a common covariance matrix for each class. Also QDA has a quadratic decision boundary. Therefore QDA was used to fit the final model. Future work involves improving sensitivity towards the crucial classes.