EECS E6690: Statistical Learning for Biological & Information Systems

Diabetes Readmission Prediction

Dwiref Oza
dso2119

Overview

- Problem Area
- Original Paper
- Dataset overview
- Data Preprocessing
- Data Visualization
- Categorization of diagnoses to ICD-9 ranges
- Logistic Regression
- Naive Bayes approach
- Observations

Paper: A study by Strack et. al

B. Strack et. al, *“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.

- Investigated the impact of HbA1c measurement on readmission rates by analyzing a database of 70,000 patient records.

Problem Area: Diabetes Readmission within 30 days

- 9.3% of the US population is affected by diabetes mellitus
- 28% of the above are undiagnosed cases
- Average readmission rate for a hospital patient 8.5% - 13.5%
- For diabetes, readmission rates vary from 14.4% - 22.7%

Dataset Overview

Data from study by Strack et. al

Database of 70,000 patient records

Patient details: Age, Gender, Race, Weight

Medical details: Test/medical procedures, admission source/type, hormone and enzyme levels, HbA1c blood test results

Columns

encounter_id <int>	patient_nbr <int>	race <fctr>	gender <fctr>	age <fctr>	weight <fctr>	admission_type_id <int>
2278392	8222157	Caucasian	Female	[0-10)	NA	6
149190	55629189	Caucasian	Female	[10-20)	NA	1
64410	86047875	AfricanAmerican	Female	[20-30)	NA	1
500364	82442376	Caucasian	Male	[30-40)	NA	1
16680	42519267	Caucasian	Male	[40-50)	NA	1
35754	82637451	Caucasian	Male	[50-60)	NA	2

| 1-8 of 50 columns

discharge_disposition_id <int>	admission_source_id <int>	time_in_hospital <int>	payer_code <fctr>
25	1	1	NA
1	7	3	NA
1	7	2	NA
1	7	2	NA
1	7	1	NA
1	2	3	NA

| 9-12 of 50 columns

medical_specialty <fctr>	num_lab_procedures <int>	num_procedures <int>	num_medications <int>	number_outpatient <int>
Pediatrics-Endocrinology	41	0	1	0
NA	59	0	18	0
NA	11	5	13	2
NA	44	1	16	0
NA	51	0	8	0
NA	31	6	16	0

ows | 13-17 of 50 columns

number_emergency <int>	number_inpatient <int>	diag_1 <fctr>	diag_2 <fctr>	diag_3 <fctr>	number_diagnoses <int>	max_glu_serum <fctr>
0	0	250.83	NA	NA	1	None
0	0	276	250.01	255	9	None
0	1	648	250	V27	6	None
0	0	8	250.43	403	7	None
0	0	197	157	250	5	None
0	0	414	411	250	9	None

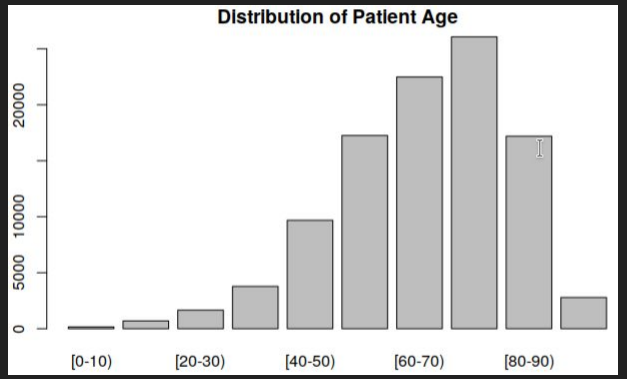
ws | 18-24 of 50 columns

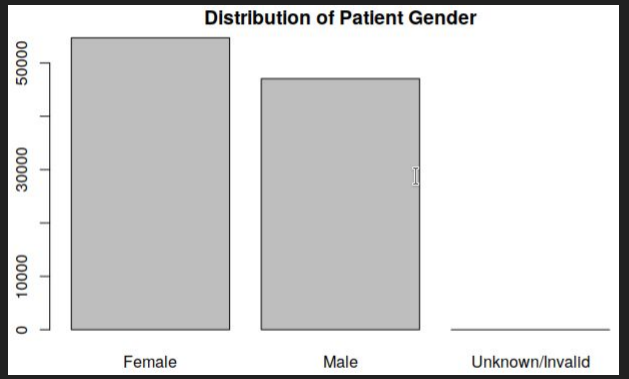
A1Cresult <fctr>	metformin <fctr>	repaglinide <fctr>	nateglinide <fctr>	chlorpropamide <fctr>	glimepiride <fctr>	acetohexamide <fctr>	glipizide <fctr>
None	No	No	No	No	No	No	No
None	No	No	No	No	No	No	No
None	No	No	No	No	No	No	Steady
None	No	No	No	No	No	No	No
None	No	No	No	No	No	No	Steady
None	No	No	No	No	No	No	No

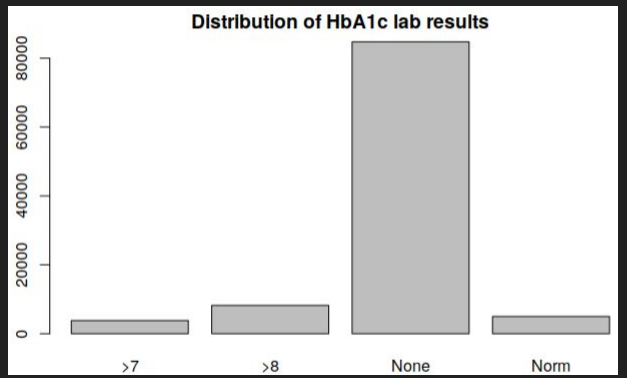
rows | 25-32 of 50 columns

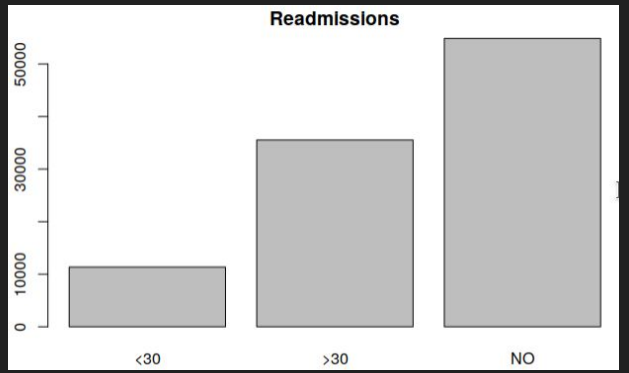
glyburide <fctr>	tolbutamide <fctr>	pioglitazone <fctr>	rosiglitazone <fctr>	acarbose <fctr>	miglitol <fctr>	troglitazone <fctr>	tolazamide <fctr>
No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No

Patient Distributions









Age: mode @ [70-80)

Gender:
Female 53%, Male 47%

HbA1c Results:
84% - no lab results

Readmissions
>50% were not readmitted

Categorical Variables

Columns 20, 21 and 22 signify diagnoses for patient visits.

The values of these columns are the International Classification of Diseases (ICD-9) medical codes. The range of these values is from 001 to 999

These can be reduced to categorical values

ICD - 9 Codes

001–139: infectious and parasitic diseases
140–239: neoplasms
390–459: diseases of the circulatory system
460–519: diseases of the respiratory system
520–579: diseases of the digestive system
580–629: diseases of the genitourinary system
710–739: diseases of the musculoskeletal system and connective tissue
800–999: injury and poisoning

(certain ranges omitted)

Categorizing Diagnosis Codes to ICD-9 Groupings

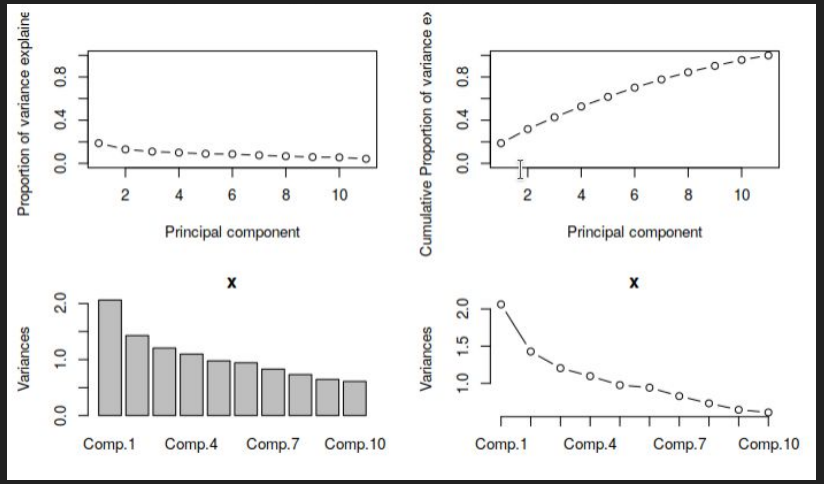
for diag1, diag2, diag3 in data:

140–239	:	“Neoplasms”
390–459	:	“Circulatory”
460–519	:	“Respiratory”
520–579	:	“Digestive”
580–629	:	“Genitourinary”
710–739	:	“Musculoskeletal”
800–999	:	“Injury”
rest	:	“Other”

This is done using

PCA on numerical columns

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
Comp.8     Comp.9    Comp.10
Standard deviation     1.4359984 1.1952241 1.0978608 1.04782538 0.98879294 0.97134088 0.9110183
0.85581637 0.80392090 0.7810781
Proportion of Variance 0.1874629 0.1298692 0.1095726 0.09981255 0.08888286 0.08577301 0.0754504
0.06658379 0.05875353 0.0554621
Cumulative Proportion  0.1874629 0.3173320 0.4269046 0.52671712 0.61559999 0.70137300 0.7768234
0.84340718 0.90216071 0.9576228
                          Comp.11
Standard deviation     0.68275115
Proportion of Variance 0.04237719
Cumulative Proportion  1.00000000
```



Using a simple function to visualize variance, cumulative variance and proportion of variance:

```r
pcaCharts <- function(x) {
    x.var <- x$sdev ^ 2
    x.pvar <- x.var/sum(x.var)
    print("proportions of variance:")
    print(x.pvar)

    par(mfrow=c(2,2))
    plot(x.pvar,xlab="Principal component", ylab="Proportion of variance explained", ylim=c(0,1), type='b')
    plot(cumsum(x.pvar),xlab="Principal component", ylab="Cumulative Proportion of variance explained", ylim=c(0,1), type='b')
    screeplot(x)
    screeplot(x,type="l")
    par(mfrow=c(1,1))
}
```

Logistic Regression

```
normal_fit <- glm(readmitted ~., data=train, family=binomial(link = 'logit'))
```

Accuracy: 43.06%

Sensitivity: 44.25%

Specificity: 22.54%

```
diag3Diabetes               -0.0039784  0.0257032  -0.155 0.876994
diag3Injury                  0.2570866  0.0644283   3.990 6.60e-05 ***
diag3Musculoskeletal        -0.0300714  0.0606998  -0.495 0.620310
diag3Genitourinary           0.0017629  0.0345151   0.051 0.959265
diag3Neoplasms               0.1100490  0.0247751   4.442 8.92e-06 ***
diag3other                   0.0614061  0.0270473   2.270 0.023188 *
admission_sourceemergency   -0.1489424  0.0297984  -4.998 5.78e-07 ***
admission_sourceother        0.0427706  0.0312404   1.369 0.170975
discharged_totransferred     0.0980901  0.0197884   4.957 7.16e-07 ***
discharged_toleft_AMA        0.0749680  0.1046037   0.717 0.473568
payer_codeself_pay          -0.1935496  0.0371738  -5.207 1.92e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 92181  on 71236  degrees of freedom
Residual deviance: 89598  on 71173  degrees of freedom
AIC: 89726

Number of Fisher Scoring iterations: 7
```

Low specificity in medical prediction tasks is extremely dangerous. Model is evidently really bad at predicting probability of readmission.

Logistic Regression w/ most significant predictors

```
normal_fit <- glm(readmitted ~ race+age+discharged_to+time_in_hospital+
            Num_lab_procedures+num_procedures+num_medications+number_outpatient+
            Number_emergency+number_inpatient+number_diagnoses+
            insulin+change+diabetesMed+diag1+diag2+diag3+A1Cresult,
          data=train, family=binomial(link = 'logit'))
```

Accuracy: 43.18%

Sensitivity: 44.33%

Specificity: 22.39%

Low specificity in medical prediction tasks is extremely dangerous. Model is evidently really bad at predicting probability of readmission.

Naive Bayes Classifier

```
     pred_nbayes
               <30        >30         NO
  <30 0.09568535 0.16378045 0.74053419
  >30 0.04492169 0.16477539 0.79030292
  NO  0.01251595 0.07321222 0.91427183
Confusion Matrix and Statistics

          Reference
Prediction   <30   >30    NO
       <30   326   479   206
       >30   558  1757  1205
       NO   2523  8427 15048

Overall Statistics

               Accuracy : 0.5611
                 95% CI : (0.5556, 0.5667)
    No Information Rate : 0.5391
    P-Value [Acc > NIR] : 5.816e-15

                  Kappa : 0.1168

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: <30 Class: >30 Class: NO
Sensitivity             0.09569    0.16478    0.9143
Specificity             0.97474    0.91126    0.2217
Pos Pred Value          0.32245    0.49915    0.5788
Neg Pred Value          0.89562    0.67026    0.6886
Prevalence              0.11160    0.34927    0.5391
Detection Rate          0.01068    0.05755    0.4929
Detection Prevalence    0.03312    0.11530    0.8516
Balanced Accuracy       0.53521    0.53802    0.5680
```

Some Observations and Key Points

- The dataset is extremely unbalanced
- Raw data is virtually untenable for predictive tasks
- Diagnosis information is too expansive to make meaningful predictions
- Converting to ICD-9 categorical data helps narrow down to clearly defined groups
- Due to the readmission class imbalance, prediction accuracy in all models tested for chances of readmission within or after 30 days is very poor.

Thank you