Kathy Byun

Raye Liu

Population segmentation for health system management

Mimic-III clinical database consists of health data of forty thousand patients in Beth Israel Deaconess Medical Center in Boston from 2001 to 2012. The dataset includes information such as diagnosis, demographics, prescriptions, ICU stays, procedures, vital signs and mortality.

International Statistical Classification of Diseases and Related Health Problems, ICD-9 code, is a list of codes used to classify disease, symptoms and external causes of disease. ICD-9 code was used to categorize diseases.

Diabetes Matrix Description

The matrix contains diabetes patients administered for ICU. The matrix has 10,310 rows × 931 columns. Each row represents a patient. Patients.csv, D\_ICD\_DIAGNOSES.csv, DIAGNOSES\_ICD.csv, ADMISSIONS.csv and ICUSTAYS.csv are joined to create the matrix with diagnoses matched to patients, ICD9 code, disease categories and demographic information.

Numerical columns:

LOS is the length of the ICU stay of a patient. The length of stay is measured in days.

Hospitalization is the number of hospital visits of each patient.

Categorical/ Binary columns:

Admission type columns: elective, emergency, newborn, urgent

Marital status columns: divorced, life partner, married, NaN, separated, single, unknown (default), widowed

Insurance columns: government, Medicaid, Medicare, private, self-pay

Ethnicity columns: there are 40 ethnicity columns. ASIAN, WHITE, BLACK/AFRICAN AMERICAN, HISPANIC OR LATINO, etc.

Gender: 1 for males. 0 for females

Dead: 1 for death. 0 otherwise

ICD9 code columns: Column names are 3-4 digit alphanumerics. 1 if a patient has a disease with the ICD9 code, 0 otherwise

Ordinal column:

ORDINAL\_AGE column: the following 10 age groups are assigned to 1-10 in ascending order. AGE: 0-10, AGE: 11-20, AGE: 21-30, AGE: 31-40, AGE: 41-50, AGE: 51-60, AGE: 61-70, AGE: 71-80, AGE: 81-89, AGE: 90+

Low Rank Model

Generalized low rank model represents an array with a low rank matrix and includes principal components analysis, matrix completion, nonnegative matrix factorization and k-means. The diabetes patient array was fitted to generalized low rank model k-means clustering to find clusters of similar patients.

3 approaches of generalized low rank model k-means clustering are used.

1.K-means with different loss functions

2.K-means with different loss functions and proximal gradient descent algorithm

3.K-means with different weighted loss functions and proximal gradient descent algorithm (currently working on 3)

Objective function and cross validation error of K-means with different loss functions

k=6: objective value = 457844.11504416, Train error = 0.048182, test error = 0.0478441

k=5 : objective value = 458976.40695412154, Train error = 0.0482869, test error = 0.0482868

k=4 : objective value = 458055.4999498729, Train error = 0.0482435 test error = 0.0481639

k=3 : objective value = 457943.0327361399, Train error = 0.048235, test error = 0.0482053

k= 6 has the smallest objective value. k=3 has the second smallest objective value. Train test errors are similar across the different k values.

Objective function and cross validation error of K-means with different loss functions and proximal gradient descent algorithm

k= 6, inner iteration = 40: objective value = 458147.6334159647, Train error = 0.0482166, test error = 0.0485213, Greater objective value

k=3, inner iteration = 40: objective value = 460455.0771161948, Train error = 0.0485515, test error = 0.0484703, Greater objective value

Predictive Performance of Low Rank Model

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Since the objective value of our prediction is quantitative variable LOS, therefore we use the regression tree. We do separate training for different clusters produced by the general low rank model training, as well as for the whole data set in Diabetes\_Numerical.csv. The train-test set split is 75%-25% of the target dataframe, and the mean square errors are:

K-means with different loss functions

k = 6

Cluster 1: 91.79407462701668

Cluster 2 : 108.23295410287105

Cluster 3: 211.3686790483595

Cluster 4: 135.3733607617661

Clutter 5: 205.66740605137792

Cluster 6: 82.73244305555494  
Max: 211.3686790483595

k=5

Cluster 1: 115.16327501139055

Cluster 2:131.60936177178976

Cluster 3: 92.46493992424074

Cluster 4: 110.83058711641635

Clutter 5: 110.06729997609855  
Max: 131.60936177178976

k=4

Cluster 1: 71.07275675466538

Cluster 2: 126.62211856517415

Cluster 3: Score: 114.20062898677885

Cluster 4: 136.53558997669276

Max: 136.53558997669276

k=3

Cluster 1: 87.1130393644

Cluster 2: 165.41786262812693

Cluster 3: 117.60504860109836

Max: 165.41786262812693

K-means with different loss functions and proximal gradient descent algorithm

k= 6 and inner iteration = 40

Cluster 1: 88.02653900067114

Cluster 2: 121.7822222875

Cluster 3: 105.74482755074999

Cluster 4: 145.01850502939322

Cluster 5: 88.82467452813678

Cluster 6: 76.90571254204545

Max: 145.01850502939322 less than than k=6

k=3 and inner iteration = 40

Cluster 1: 131.58260460431006

Cluster 2: 86.46730035293763

Cluster 3: 174.09246571365017

Max score: 174.09246571365017 greater than k=3

Linear Regression

As a linear approach to modeling the relationship between a numerical response the rest of the explanatory variables, multiple linear regression, linear regression models are used as a baseline evaluation for the performances of other forthcoming models we are trying. The performance of the linear regression model is considerably poor on the diabetes patients’ subgroup dataset, even after tuning the model with non-linear terms. We are trying to examine the outliers in the coefficients in the training results.

Mean square errors for K-means with different loss functions:

k = 6

Cluster 1: 6.4553564721143735e+22

Cluster 2: 2.6163306751921842e+22

Cluster 3: 4.5454524460558475e+19

Cluster 4: 4.554160043204208e+20

Clutter 5: 6.192207418609172e+22

Cluster 6: 1.9742893196383015e+22  
Max: 6.4553564721143735e+22

k=5

Cluster 1: 3.133211643586004e+20

Cluster 2: 1.5483866834806618e+21

Cluster 3: 2.920527363935486e+21

Cluster 4:5.451606501473484e+22

Clutter 5: 3.448475511420542e+23

Max: 3.448475511420542e+23

k=4

Cluster 1: 1.5978054899701247e+21

Cluster 2: 8.714307290317095e+21

Cluster 3: 1.6016118399737103e+23

Cluster 4: 6.465389092682245e+19

Max: 1.6016118399737103e+23

k=3

Cluster 1: 2.8919513891721223e+22

Cluster 2: 1.6551507799146446e+23

Cluster 3: 2.0447264347063947e+21

Max: 1.6551507799146446e+23

K-means with different loss functions and proximal gradient descent algorithm

k=6 and inner iteration = 40

Cluster 1: 1.6125562789649733e+24

Cluster 2: 2.0137381692478077e+24

Cluster 3: 9.50596938056577e+21

Cluster 4: 2.0060717411479823e+21

Cluster 5: 7.245536375425792e+22

Cluster 6: 2.413190339823639e+23

Max: 2.0137381692478077e+24 greater than k=6

k=3 and inner iteration = 40

Cluster 1: 7.365972421278354e+22

Cluster 2: 3.3915861941652265e+21

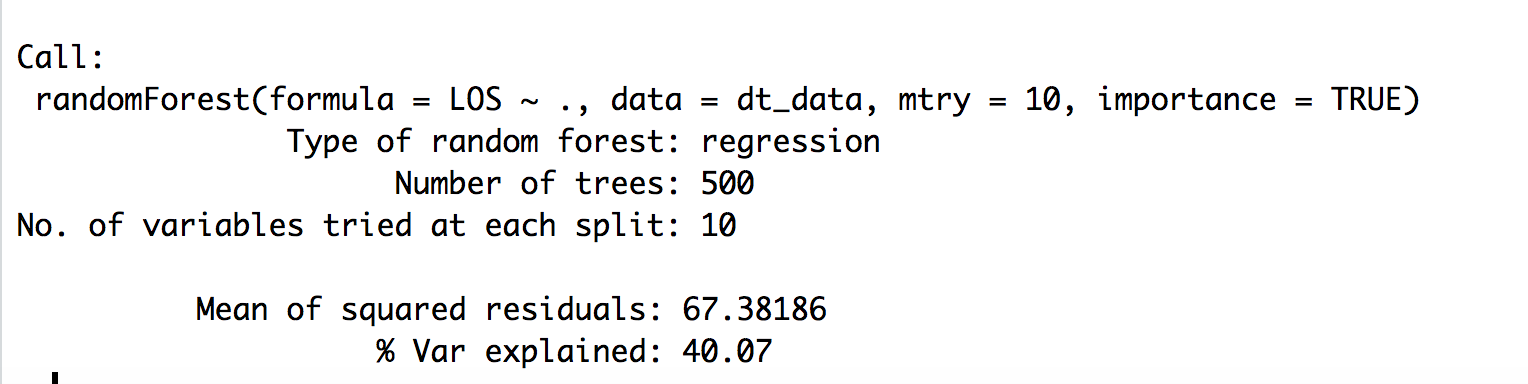
Cluster 3: 2.9180117119324733e+22

Max: 7.365972421278354e+22 less than k=3

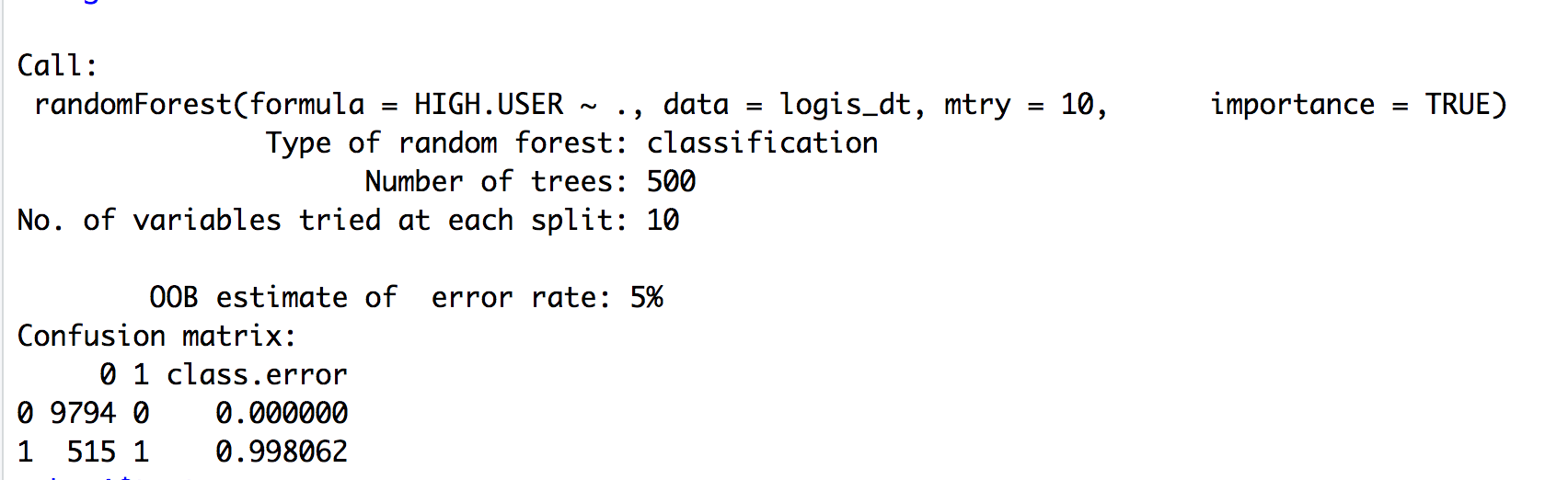
Random Forests

We use 500 trees and 10 variables at each split in the random forest training and obtain the mean square errors and the R^2 values of approximately 0.4. For the regression, we try to predict the value of LOS which is the length of stays in ICU measured in days. For the classification, we divide the patients into two groups: the group of High-Resource-Users and the group of Non-High-Resource-Users. We define the patients with the top 5% largest LOS values as the “1” group, standing for high emergency health care resources consumers, the other 95% percent of patients are “0” group. This approach resembles the HRUPoRT of Canadian population survey data from the paper *“Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System”* by Dr. L. C. Rosella. The threshold value of the top 5% LOS is 25.0924, whereas 516 out of 10310 patients have a LOS value greater than or equal to this value.

Regression:

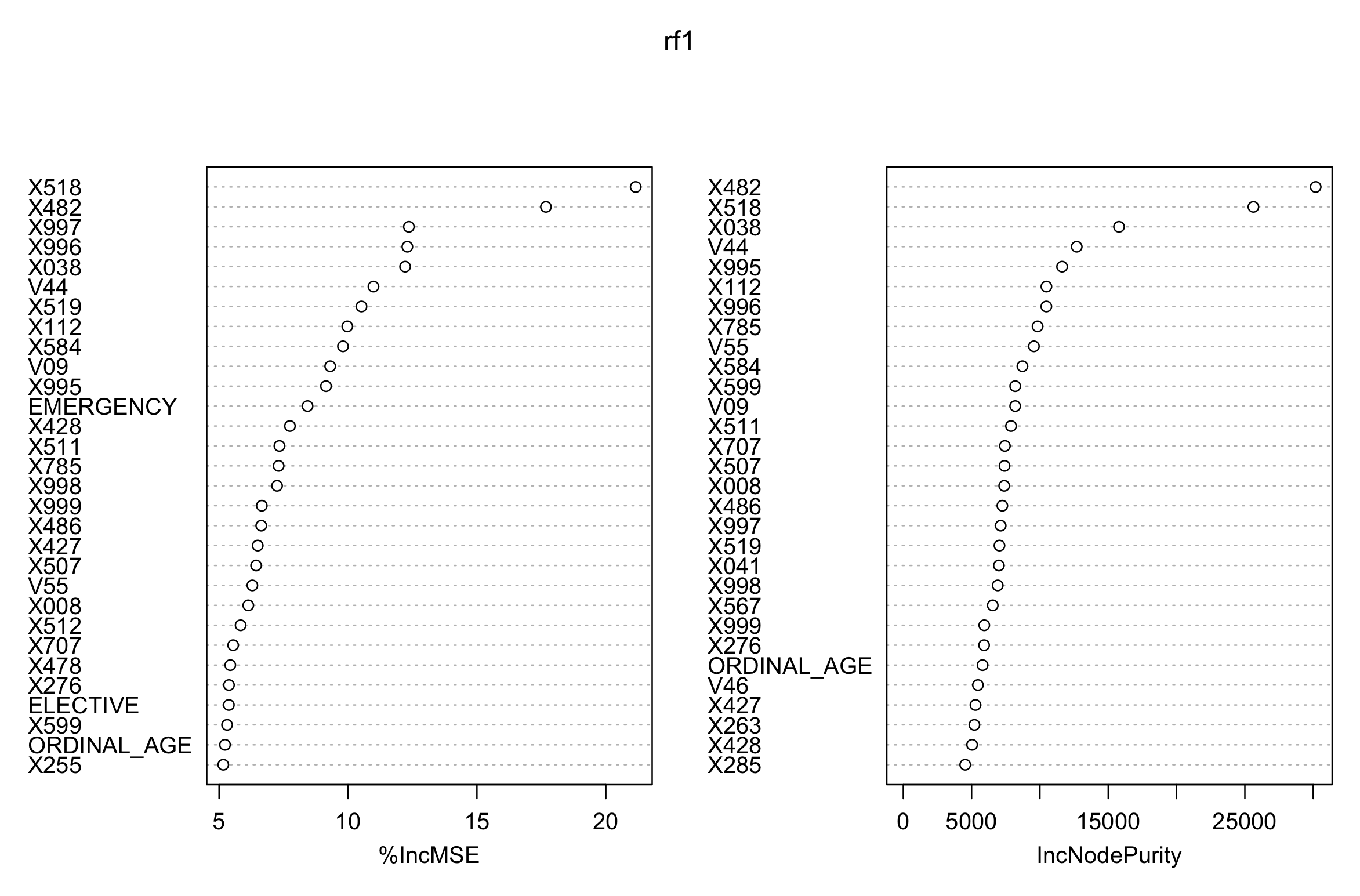


Classification:



From the confusion matrix of the classification trees we can see the true negative rate is 1 and the true positive rate is almost 0, saying that it fails to identify almost all the High-Resource Users. The training is to be improved.

Random forest evaluates each factors’ contribution towards the mean square errors and puts weights on each of them. From the diagram below, we select the top 20 (or another number) of the factors and put them in training again for the general low rank model.



Important features are features with highest %incMSE

ICD-9 codes

518: Other diseases of the lung. A disorder characterized by the collapse of part or the entire lung.

482: Other bacterial pneumonia.

997 : Complications affecting specified body system not elsewhere classified.

Ex.997.0 Nervous system complications

996: Complications peculiar to certain specified procedures.

Ex. 996.0 Mechanical complication of unspecified cardiac device, implant, and graft

038: Septicemia: blood poisoning, especially that caused by bacteria or their toxins.

112: Candidiasis: A condition in which candida albicans, a type of yeast, grows out of control in moist skin areas of the body.

Logistic Regression

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

In our training approach, we define the top 5% percent of the patients with large LOS values as the “1” group, standing for high emergency health care resources consumers, the other 95% percent of patients are “0” group. This approach resembles the HRUPoRT of Canadian population survey data from the paper *“Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System”* by Dr. L. C. Rosella.

Training result: to be out.

Center of clusters

The following features have different values among the clusters. Features not included in the tables have the same values among the clusters.

ICD-9 codes

14: Personal history of allergy to unspecified medicinal agent

184: Malignant neoplasm of other and unspecified female genital organs.

301: Personality disorders

310: Specific nonpsychotic mental disorders due to brain damage.

878: Open wound of genital organs (external) including traumatic amputation

Centers of clusters for k=6

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | LOS | HOSPITALIZATION | 14 | 184 | 301 | 310 |
| 0 | 3.0341833 | 1.246904 | 1 | 0 | 1 | 0 |
| 1 | 2.9640 | 1.292386 | 1 | 1 | 1 | 1 |
| 2 | 2.9531 | 1.231171 | 1 | 0 | 1 | 0 |
| 3 | 2.9515 | 1.3219 | 1 | 1 | 1 | 0 |
| 4 | 3.0576 | 1.2329 | 1 | 0 | 0 | 0 |
| 5 | 3.2224 | 1.3235 | 0 | 1 | 1 | 0 |

Centers of clusters for k=3

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | LOS | HOSPITALIZATION | 878 |
| 0 | 2.991951 | 1.246904 | 0 |
| 1 | 3.074003 | 1.292386 | 1 |
| 2 | 2.922126 | 1.231171 | 0 |

Reference (formal quotations to be continued):

Johns Hopkins ACG System

Predicting High Health Care Resource Utilization