**HS 616 QUIZ 6**

**Research Article**

Beata Strack et al. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International. Volume 2014 (2014), Article ID 781670, 11 pages.

**URL link**

<https://www.hindawi.com/journals/bmri/2014/781670/#B9>

**Github Repo**

https://github.com/tobiolatunji/Readmission\_simulation\_R

**Report Format**

Final Report will be submitted in html (R markdown)

**Full Project Description**

This data simulation project seeks to predict <30 day readmission in patients with diabetes. The data would be simulated based on the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States.

The Health Facts data used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500.

Dataset will have 30,000 rows and 24 columns. Original dataset had 100,000 rows and 55 columns. Dimensions were reduced to improve computation speed on my local machine.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Variable type** | **Description** | **Missing values** |
| Race | Factor | Caucasian (75%), Asian (1%), African-American (19%), Hispanic (3%), and other (5%) | 2% |
| Gender | Factor | Female (53%), male (46%), unknown/invalid (1%) |  |
| Age | Numeric | Right skewed Normal Distribution, Mode ~ 85yrs |  |
| Length of stay | Numeric | Min: 1 day, Max: 14 days, Mean- 4.4 |  |
| Number of lab procedures | Numeric | Min: 1, Max: 132, Mean- 43.1 |  |
| Number of non-lab procedures | Numeric | Min:0, Max: 6, Mean- 1.34 |  |
| Number of medications | Numeric | Min: 1, Max: 81, Mean- 16.02 |  |
| Number of outpatient visits | Numeric | Min: 0, Max: 42, Mean- 0.37 |  |
| Number of emergency visits | Numeric | Min: 0, Max: 76, Mean- 0.2 |  |
| Number of inpatient visits | Numeric | Min: 0, Max: 21, Mean- 0.64 |  |
| Number of diagnoses | Numeric | Min: 1, Max: 16, Mean- 7.42 | 1% |
| Glucose serum test | Factor/4 | “>200,” “>300,” “normal,” and “none” (>95%) |  |
| HbA1c result | Factor/4 | “>8” (8%), if the result was greater than 8%, “>7” (5%), if the result was greater than 7% but less than 8%, “normal” (6%), if the result was less than 7%, and “none” (85%), if not measured. |  |
| Insulin | Factor/2 | “yes” (78%), and “no” (22%) |  |
| Change in medication | Factor/2 | “change” (47%), and “no change” (53%), |  |
| Diabetes Medication | Factor/2 | “yes” and “no” |  |
| Diagnosis 1 category | Factor/9 | “Circulatory” (30%), “Respiratory” (15%), “Digestive” (10%), “Diabetes” (8%), “Trauma” (7.5%), “Musculoskeletal” (5%), “Genitourinary” (6%), “Neoplasms” (15%), “Other (7%)” |  |
| Diagnosis 2 category | Factor/9 | “Circulatory” (32%), “Respiratory” (11%), “Digestive” (4%), “Diabetes” (13%), “Trauma” (2.5%), “Musculoskeletal” (2%), “Genitourinary” (9%), “Neoplasms” (19%), “Other (10%)” |  |
| Diagnosis 3 category | Factor/9 | “Circulatory” (30%), “Respiratory” (7.5%), “Digestive” (4%), “Diabetes” (17%), “Trauma” (2%), “Musculoskeletal” (2%), “Genitourinary” (7%), “Neoplasms” (17%), “Other (15%)” |  |
| Admission source | Factor/3 | clinic referral (10%), emergency room (57%), and other (33%) |  |
| Discharged to | Factor/3 | Home (25%), transferred (73%), AMA (2%) |  |
| Payer code | Factor/2 | “Insured” (45%) and “Self\_Pay” (3%) | 52% |

**Correlations (Numeric variables)**

Time-in-hospital is positively correlated with number of lab procedures,

Number of non-lab procedures, number of medications and number of diagnoses

Number of emergency visits correlates with number of inpatient visits

**Confounding variables**

Number of lab procedures and Number of procedures

**Response variable**

distribution and relation to correlated or confounding predictor variables.

|  |  |  |  |
| --- | --- | --- | --- |
| Readmitted | Factor/3 | “<30” (10%) if the patient was readmitted in less than 30 days, “>30” (30%) if the patient was readmitted in more than 30 days, and “No” (60%) for no record of readmission. |  |

f) List steps you plan to take in simulating the data

|  |  |
| --- | --- |
| **Variable** | **Variable type** |
| Race | Sample() |
| Gender | rbinom() |
| Age | rpois() |
| Length of stay | rnorm() |
| Number of lab procedures | rnorm() |
| Number of non-lab procedures | rnorm() |
| Number of medications | rnorm() |
| Number of outpatient visits | rnorm() |
| Number of emergency visits | rnorm() |
| Number of inpatient visits | rnorm() |
| Number of diagnoses | rnorm() |
| Glucose serum test | Sample() |
| HbA1c result | Sample() |
| Insulin | rbinom() |
| Change in medication | rbinom() |
| Diabetes Medication | rbinom() |
| Diagnosis 1 category | Sample() |
| Diagnosis 2 category | Sample() |
| Diagnosis 3 category | Sample() |
| Admission source | Sample() |
| Discharged to | Sample() |
| Payer code | rbinom() |
| Readmitted (class) | Sample() |

g) List steps you plan to take in analyzing the data.

1. Identify class variable, consider also binarizing class variable into “Readmitted” and “not\_readmitted”
2. Data exploration, summaries, preprocessing, deal with missing values and outliers
3. Correlations, tests of dependence of variables, power tests
4. Feature extraction, feature importance
5. Regression model to see significant variables
6. Dimensionality reduction PCA/Step function
7. Split data into train and test sets
8. Cross validation, grid search
9. Model and optimize hyperparameters using
   1. K-Nearest Neighbors,
   2. Logistic regression,
   3. Naive Bayes (Gaussian and Multinomial),
   4. Decision Tree,
   5. Support Vector Machines (Linear and RBF Kernel),
   6. Random Forest and
   7. Gradient Boosting methods.
10. Record scores for various models- F1, accuracy, ROC, AUC, confusion matrix, sensitivity/specificity, precision, recall
11. Visualize Results