Diabetes

kmathuru96

A **(“Hospital Readmission”, n.d.)** is an episode when a patient who had been discharged from a hospital is admitted again within a specified time interval. Readmission rates have increasingly been used as an outcome measure in [health services research](https://en.wikipedia.org/wiki/Health_services_research) and as a quality benchmark for health systems. Insurance companies and other payers sometimes view unplanned hospital readmission’s as wasteful spending .To penalize these readmission rates, measures are to be taken by hospitals. Machine learning models are used to overcome and analyse this problem. “Predicting whether a patient is readmitted to hospital or not” address the problem. A data set that represents 10 years (1999-2008) of clinical care at 130 US hospitals, which has 50 features and 101766 records of patients and hospital outcomes.

# Introduction

Tracking  patients who are readmitting to a hospital after a hospital stay is one category of data which is used to evaluate the quality of hospital care. Patients with (“Diabetes”, n.d.) have high rates of readmission compared with patients without diabetes, according to a pilot study published in Clinical diabetes and Endocrinology. In the first study, the **readmission rate** was 26% in patients with diabetes vs 22% in patients without diabetes “ US has the highest prevalence of diabetes among all developed countries across the world”- (“IDF”, n.d.). Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat.  Glucose comes from the foods you eat. Insulin is a hormone that helps the glucose get into your cells to give them energy.

We have  (“Data Set”, n.d.) that represents 10 years (1999-2008) of clinical care at 130 US hospitals, which has 50 features and 101766 records of patients and hospital outcomes.  Information was extracted from the database for encounters that satisfied the criteria’s such as inpatient encounter (a hospital admission), diabetic encounter( one during which any kind of diabetes was entered to the system as a diagnosis), length of stay was at least 1 day and at most 14 days, laboratory tests performed during the encounter, medications were administered during the encounter.

The data also contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

image

## Attribute Information:

## 

### Feature name                               Type                              Description and values                                                           % missing

Encounter ID                      Numeric               Unique identifier of an encounter                                             0%

Patient number                Numeric               Unique identifier of a patient                                                      0%

Race                                      Nominal               Values: Caucasian, Asian, African American,                         2%

                                                                                 Hispanic, and other

Gender                                Nominal               Values: male, female, and unknown/invalid                         0%

Age                                        Nominal               Grouped in 10-year intervals:                                                      0%                                                                                                                                                                                          (0, 10), 10, 20), …, 90, 100)

Weight                                 Numeric                Weight in pounds.                                                                           97%

Admission type                 Nominal               Integer identifier corresponding to 9 distinct                        0%

                                                                                 values, for example, emergency, urgent, elective,

                                                                                 newborn, and not available

Discharge disposition     Nominal               Integer identifier corresponding to 29 distinct values,      0%

                                                                                   for example, discharged to home, expired, and

                                                                                    not available

Admission source             Nominal                 Integer identifier corresponding to 21 distinct values,      0%

                                                                                    for example, physician referral, emergency room,

                                                                                    and transfer from a hospital

Time in hospital                Numeric               Integer number of days between admission                         0%

                                                                                 and discharge

Payer code                          Nominal               Integer identifier corresponding to 23 distinct values,      52%

                                                                                 for example, Blue Cross/Blue Shield, Medicare,

                                                                                 and self-pay

Medical specialty             Nominal               Integer identifier of a specialty of the admitting                  53%

                                                                                 physician, corresponding to 84 distinct values, for

                                                                                 example, cardiology, internal medicine, family/

                                                                                 general practice, and surgeon

Number of lab                   Numeric               Number of lab tests performed during the                            0%

Procedures                                                          encounter

Number of                          Numeric               Number of procedures (other than lab tests)                       0%

Procedures                                                         performed during the encounter

Number of                          Numeric               Number of distinct generic names administered                0%

Medications                                                       during the encounter

Number of                          Numeric               Number of outpatient visits of the patient in the                0%

outpatient visits                                                year preceding the encounter

Number of                          Numeric               Number of emergency visits of the patient in the               0%

emergency visits                                              year preceding the encounter

Number of                          Numeric               Number of inpatient visits of the patient in the                   0%

inpatient visits                                                   year preceding the encounter

Diagnosis 1                          Nominal               The primary diagnosis (coded as first three                           0%                                                                                                                                                                              digits of ICD9); 848 distinct values.                                                                                                                                                                                                                               (International Classification of diseases)

Diagnosis 2                          Nominal               Secondary diagnosis (coded as first three digits                   0%                                                                                                                                                                             of ICD9); 923 distinct values

Diagnosis 3                          Nominal               Additional secondary diagnosis (coded as first                     1%                                                                                                                                                                            three digits of ICD9); 954 distinct values

Number of                          Numeric               Number of diagnoses entered to the system                           0%

diagnoses

Glucose serum                  Nominal               Indicates the range of the result or if the test was                0%

test result                                                            not taken. Values: “>200,” “>300,” “normal,” and                                                                                                                                                                                             “none” if not measured

A1c test result                   Nominal               Indicates the range of the result or if the test was                 0%                                                                                                                                                                           not taken. Values: “>8” if the result was greater                                                                                                                                                                                                 than 8%, “>7” if the result was greater than 7% but                                                                                                                                                                                          less than 8%, “normal” if the result was less than                                                                                                                                                                                              7%, and “none” if not measured.

Change of                            Nominal               Indicates if there was a change in diabetic                               0%

medications                                                        medications (either dosage or generic name).                                                                                                                                                                                                       Values: “change” and “no change”

Diabetes                              Nominal               Indicates if there was any diabetic medication                       0%

Medications                                                        prescribed. Values: “yes” and “no”

23 features for                  Nominal               For the generic names: metformin, repaglinide,                     0%

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

                                                                                the feature  indicates  whether the drug was prescribed

                                                                                 or there was a change, and “no” if the drug was not

                                                                                 prescribed

Readmitted                        Nominal               Days to inpatient readmission. Values: “<30” if the                0%                                                                                                                                                                           patient was readmitted in less than 30 days, “>30” if                                                                                                                                                                                         the patient was readmitted in more than 30 days,                                                                                                                                                                                              and “No” for no record of readmission.

## Preprocessing :

* When the data is loaded into IDE, the attributes has to be changed to relevant data types such as numeric or nominal, based on attribute information.
* Data standardization is done.
* Class imbalance is handled using SMOTE(Synthetic minority oversampling technique). It uses k- nearest neighbors algorithm and calculate the similarity between records, high similarity records of minority class are over-sampled.
* Omitted three attributes (weight, payer\_code, medical specialty) which have missing values more than 50 % ,remaining attributes have less than 3% of missing values. Instead of imputing missing values(which may not be perfect), records (with missing values) are deleted(as data is huge).
* Unimportant features such as patient number, encounter ID, examide are dropped(features which have unique records / no variance).
* Feature engineering is done, number of lab procedures (number of lab tests performed during the encounter)and number of procedures (number of procedures performed other than lab tests) are combined which gave total number of procedures.

# Approach :

There are three approaches for this data.

**Approach – 1 :**

  Handling the class imbalance using SMOTE, which over samples   the minority class.

  **Logistic Regression:**

             Logistic regression is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome.

**Decision Tree:**

*Decision Trees*  are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

**Random Forest:**

              Random forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the target.

**XG Boost:**

             XGBoost is an optimized distributed gradient boosting library designed to be highly ***efficient***, ***flexible*** and ***portable***. It implements machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. XGBoost provides a parallel tree boosting .

**Majority voting:**

             Majority voting is an ensemble approach which combines different models( mode of the classes for classification and mean prediction for regression of the target.)

**Stacking:**

S*tacking* (also called meta ensembling) is a model ensembling technique used to combine information from multiple predictive models to generate a new model.

**Final test scores:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | F1 – score | Recall |
| Logistic Regression | 47.5 | 51 | 48 | 48 |
| Decision Tree | 52 | 50 | 50 | 52 |
| Random Forest | 58 | 56 | 52 | 58 |
| XG Boost | 58 | 54 | 54 | 58 |
| Majority Voting | 52 | 53 | 50 | 52 |
| Stacking | 58 | 70 | 62 | 58 |

Approach – 2 :

  Random forest gives us variable importance. Using these important variables  models are built in this approach.

         **Logistic Regression:**

         **XG Boost:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | F1 – score | Recall |
| Logistic Regression | 48 | 51 | 48 | 48 |
| XG - Boost | 59 | 71 | 63 | 59 |

Approach – 3 :

  In this approach, target variable is made to binary class from multi-class

* **Target as Yes or No:**

**Logistic Regression:**

**Random Forest:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | F1 – score | Recall |
| Logistic Regression | 60 | 62 | 59 | 61 |
| Random Forest | 62 | 60 | 61 | 62 |

* **Target as >30 or <30:**

**Logistic Regression:**

**Random Forest:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | F1 – score | Recall |
| Logistic Regression | 76 | 93 | 86 | 76 |
| Random Forest | 77 | 95 | 87 | 77 |

# Conclusion:

Recall is the error metric,  we cannot classify a patient wrongly as he cannot be readmitted (treatment should be done on time, this may also cost a patient’s life).

**Recall** = True Positive / Total Actual Positive

Based on the approaches made during model building, it can be concluded that these models can be used by the client to predict hospital readmission

|  |  |  |  |
| --- | --- | --- | --- |
| Approach                  Model                              Recall |  |  |  |
| 1                Random Forest                          58             2                    XG Boost                                   59             3               Random Forest               Target - Yes/No - 61                                                                           Target : >30 / <30 - 77   |  | | --- | |  | |  |  |  |

# References

n.d. <https://en.wikipedia.org/wiki/Hospital_readmission.> <https://en.wikipedia.org/wiki/Hospital_readmission.>

n.d. <http://www.endocrinologyadvisor.com/type-2-diabetes/diabetes-hospital-readmission-rates/article/675957/.> <http://www.endocrinologyadvisor.com/type-2-diabetes/diabetes-hospital-readmission-rates/article/675957/.>

n.d. <http://www.dailymail.co.uk/health/article-3341773/America-HIGHEST-rate-diabetes-developing-world-UK-Australia-Lithuania-nations-lowest-rates-condition.html.> <http://www.dailymail.co.uk/health/article-3341773/America-HIGHEST-rate-diabetes-developing-world-UK-Australia-Lithuania-nations-lowest-rates-condition.html.>

n.d. <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.> <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.>