Diabetes Occurrence Prediction

Continuous Assessment 1 - Problem Solving Using Pattern Recognition

Srikar Namburu (A0198490N)

Vishaag Suriyanarayanan (A0198504N)

Institute of System Science, National University of Singapore

Problem Statement and Objective

A hospital readmission is when a patient who is discharged from the hospital gets re-admitted again within a certain period of time. Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely. For this reason, the Centers for Medicare & Medicaid Services established the Hospital Readmissions Reduction Program which aims to improve the quality of care for patients and reduce health care spending by applying payment penalties to hospitals that have more than expected readmission rates for certain conditions. Although diabetes is not yet included in the penalty measures, the program is regularly adding new disease conditions to the list, now totaling 6 for FY2018.

In 2011, American hospitals spent over $41 billion on diabetic patients who got readmitted within 30 days of discharge. Being able to determine the factors that lead to higher readmission in such patients, and correspondingly being able to predict which patients will get readmitted can help hospitals save millions of dollars while improving quality of care.

Goals

To find out the factors that are the significant predictors of hospital readmission in patients who have diabetes.

Dataset

We have used the Diabetes 130-US hospitals for years 1999-2008 Data Set from the UCI Machine Learning Repository.

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

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

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

Tools / Techniques

Tools used:

1. Google Collab - for writing code, training models
2. Pandas, Numpy, Scipy, Matplotlib python packages - for data wrangling and visualization
3. Sci-kit learn for machine learning models

Techniques:

1. SMOTE (Synthetic Minority Over-sampling TEchnique)
2. Decision Tree Classifier
3. Random Forest Classifier
4. Xtreme Gradient Boosting
5. K-Nearest Neighbours

Design of Models

For all the models, after parameter tuning, the following gave the best performance -

1. Decision Tree Classifier -

   ```
   max_depth=25, criterion = "gini", min_samples_split=10
   ```

2. Random Forest -

   ```
   n_estimators = 10, max_depth=25, criterion = "gini",
   min_samples_split=10
   ```

3. XGBoost - Default Model Design
4. K-Nearest Neighbours

   ```
   n_neighbors=9
   ```

Performance of models

S.No	Model	Train Data Accuracy	Validation Data Accuracy	Test Data Accuracy
1	Decision Tree	97.47%	91.60%	91.36%
2	Random Forest	95.67%	94.81%	94.47%
3	XGBoost	94.72%	94.76%	94.47%
4	K-NN	78.84%	72.53%	73.14%

Findings

From our analysis, we found out that the following 5 features played a key role in determining the recurrence.

1. Insulin Levels
2. Diabetes Medication
3. Gender
4. Number of Medications
5. Discharge Disposition