Predicting Hospital Readmission

BST260 Project

Jacob Rosenthal, Selena Huang, Jonathan Waring, Erica Moreira

Overview

The dataset we will be working with 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:

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

The dataset is avaliable from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008).

Question of Interest

The project is a 2-way classification task in which we try to predict if a patient will either:

1. Not be readmitted
2. Readmitted in $<$30 days

Data Preparation and Preprocessing

After cleaning our data, the dataset used for analysis contained 42 columns.

New Features: - Healthcare Utilization is calculated as the sum of number of outpatient, emergency, and inpatient encouters the patient has had in the past year - Sum on non-insulin diabetes medications - ICD codes have over 700 different "levels" in our dataframe. We will group these ICD codes into 20 different categories based on the Strack et al. 2014 paper (https://www.hindawi.com/journals/bmri/2014/781670/). We will then one hot encode these categories for each patient.

One Hot Encodings - Admission types, discharge dispositions, admission source will be grouped and one hot encoded —

Train Test Split

The dataset is now ready, and we will split it into training(80%) and test(20%) datasets using caret's `createDataPartition` function.

SMOTE Algorithm For Unbalanced Classification Problems

"SMOTE is an oversampling method which creates new instances of the minority class by forming convex combinations of neighboring instances. It effectively draws lines between minority points in the feature space, and samples along these lines. This allows us to balance our data-set without as much overfitting, as we create new synthetic examples rather than using duplicates. This however does not prevent all overfitting, as these are still created from existing data points." (https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2)

Hyperparameter optimization

In machine learning, hyperparameter optimization is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is set before the learning process begins. By contrast, the values of other parameters are derived via training.

Training the Model

Variable Importance According to XGBoost

Simple ROC Curve

ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. You can read more about it here:(https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5).

KNN