Predicting Diabetes Hospital Readmission

Udacity Machine Learning Engineer Nanodegree Proposal | 1.8.22

Domain Background

Readmission risk prediction is a promising machine learning use case with potential to improve both patient outcomes and reduce health care costs. Typically, it is the responsibility of a medical provider to determine when a patient is ready to be discharged from an inpatient facility. This is a crucial decision because discharging a patient too early is likely to lead to poorer health outcomes. Conversely, failing to discharge a patient in a timely manner increases cost unnecessarily and puts pressure on an already overburdened healthcare system. A high performance readmission model could serve as a practical consultative tool to augment a provider's medical expertise when evaluating patients for discharge.

Problem Statement

Given a dataset of historical diabetic inpatient encounters, predict whether each will result in a readmission. This is a standard classification task where the model output can be used to rank records based on readmission probability.

Datasets and Inputs

The dataset I will use can be access from the UCI Machine Learning Repository. It contains 10 years of diabetic inpatient encounters with 50 corresponding features of patient and clinical attributes. Also included is a “readmitted” target variable indicating whether the encounter resulted in a readmission in <30 days, >30 days, or no readmission. See link to raw data below:
https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

Solution Statement

The implemented solution will attempt to fit multiple classification models (1 linear, 1 non-linear) using Sagemaker built-in algorithms. Additionally, the solution will make use of Sagemaker’s hyperparameter tuning and batch transform/inference functionality.

Benchmark Model

As a general comparison, I reviewed the meta-analysis Risk Prediction Models for Hospital Readmission: A Systematic Review:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603349/

While none of the models reviewed exactly match our specific problem statement pertaining to diabetes patients, the analysis does offer good directional information. Depending on the type of readmission and population, readmissions models typically perform somewhere in the AUC (C-statistic) range of 0.60 to 0.70.

Evaluation Metrics

Models will be evaluated on the following criteria:

- Accuracy (ACC)
- Area under ROC (AUC)

Area under ROC (AUC) curve is a useful evaluation metric because it describes a classifier's ability to rank positive class instances ahead of negative class instances based on their predicted probabilities. In this way, AUC evaluates a classifier without the (arbitrary) need to define a classification threshold.

We will also want to look at Accuracy (ACC). Accuracy can be a useful evaluation metric in that it offers an intuitive sense of model performance. This is particularly true if we sample our training data to have a balance number of target instances (equal number of readmit and non-readmit encounters).

Project Design

1. Write script to ingest raw data from URL
2. Explore and visualize raw data
3. Preprocess data into numeric features
4. Perform feature selection on numeric features
5. Segment selected features into train/test/validation partitions
6. Fit and tune a LinearLearner model on the train/validation data
7. Tune a batch transformer and evaluate test data
8. Calculate AUC/ACC evaluation metrics
9. Repeat steps 6-8 using Sagemaker XGBoost classifier
10. Compare modul performance (LinearLearner v. XGBoost)