Hyperparameter Database for Predicting Diabetes Patient Readmission

Michelle Pradeep | pradeep.m@husky.neu.edu
Seemanthini Jois | jois.s@husky.neu.edu
Suraksha Jadhav | jadhav.su@husky.neu.edu

Abstract — Readmission of patients at a hospital is a significant worry in diabetes. In terms of money over 250 million dollars was spent on tending the readmitted diabetes patients in the year 2011. Premature recognition of patients that face a high risk of readmission can permit healthcare providers to perform a supplemental inspection and feasible prevention of further admissions. In the year 2011, it was announced that over 3.3 million patients were readmitted in the US within 30 days of being discharged and this was associated with about $41 billion in Hospital Bills. The requirement of readmission indicates that deficient care was provided to the patients at the time they were first admitted. The readmission rate has become an important measure that measures the comprehensive quality of a Hospital. The statistical metric for readmission status is AUC. Our objective here is to find the important hyperparameters and their range, along with their comparison across different algorithms. In order to show the obtained solution is optimal, we have included exploratory analyses of the dataset and conducted rigorous validation and optimization of hyperparameter using H2O.

Keywords

Diabetes, Hyperparameters, AUC, Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM).

I. Introduction

It is progressively acknowledged that managing hospitalized diabetic patients have a notable significance in terms of morbidity and mortality. This identification has led to developing formalized protocols in a Hospital's Intensive care Units. However, for a majority of non-ICU inpatient admissions, the same cannot be said. Relatively, unscientific evidence demonstrates the inpatient management is random and frequently leads to no treatment being given or wide variations in glucose when conventional methods are employed.

Although the data collected is less, recently controlled trials have shown that certain protocols driven in patients can be both safe and effective. In essence, the protocols that are being implemented in hospitals are advocated.

Nonetheless, there are a few national level evaluations of diabetes care in the patients that are hospitalized which could serve as a baseline for changes. The current analysis of a massive clinical database was taken to analyze the historical patterns of patients with diabetes admitted to hospitals in the US and to improve the safety of patients.

The databases of these Clinical data have very valuable but diverse and difficult data in terms of values that are missing, records that are incomplete, and a dimensionality that is high not only comprehended by features but also by their complexity.

Moreover, inspecting external data is more of a challenge that analyzing the results of a very well-designed trial, because no one really knows how the information was collected.

Regardless, it is important to utilize these large amounts of data to find new Information that isn't available anywhere.

II. Dataset

The data we will be working with is readmission of diabetic patient's data. The outcome is focused on the management of hyperglycemia in patients that are hospitalized. This factor has an important bearing in terms of morbidity and mortality. In the dataset, we will see the impact of various clinical practices in readmission rates. The dataset constitutes the clinical care at 130 hospitals and coherent networks in the US for the last 10 years(1999-2008). Multiple features are included such as outcomes of patient and hospital. Information was extracted from the databases for situations that satisfied the following criteria :

1) If the hospital admission is an inpatient encounter.
2) If it is a diabetic encounter. The type of diagnosis which was entered.
3) If the length of the patient's stay was at least 1 day and at most 14 days.
4) if during the encounter Laboratory tests were carried out.
5) If medications were doctored during the encounter.

Attributes in the dataset are the data contains attributes such 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 medications, diabetic medications, number of outpatients, inpatient, emergency visits in the year before the hospitalization and many others.

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

III. Method

Parameters that are stated in prior to running a machine learning algorithm and that have a colossal effect on the predictive power of statistical models are known as Hyperparameters. The aim to creating constructive models is the comprehension of the relative significance of a hyperparameter to an algorithm and also the range of the values of the hyperparameter is an integral aspect to tuning the hyperparameters in the model construction. The database for the Hyperparameters is a public resource with algorithms, tools, and data which allows the user to visualize and comprehend how to choose the hyperparameters in order to maximize the predictive power of the models. This database for the hyperparameter is created by running millions of hyperparameters values on multiple databases and calculating the individual exception of each hyperparameter on the quality of the model. Currently, the hyperparameter database analyzes the effect of hyperparameters on the following algorithms: Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM). Naïve Bayes Classifier, Stacked Ensembles, Xgboost and Deep Learning Models (Neural Networks). We can also build the models using the hyperparameter database that can predict the hyperparameters without searching or visualizing the statistical concepts such as bias/variance tradeoff.

IV. Code and Documentation

GitHub Link: The complete code with documentation can be found on the below link :
https://github.com/INFO6105-Spring19/hyperparameter-db-project-ds01

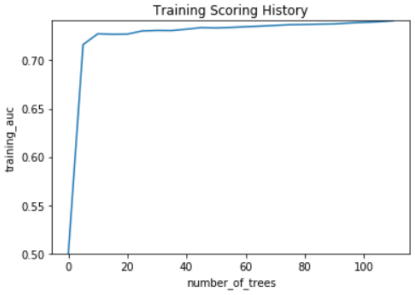
V. Results

The dataset was selected, and we ran AutoML on the dataset and decided the Target, which in our case is the readmission of the diabetic patients. We ran the AutoML for three different run times i.e.

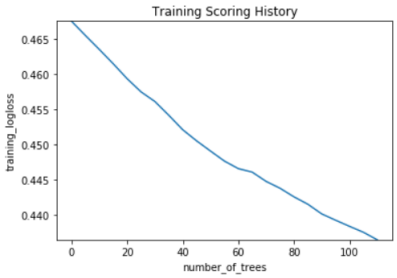
1) 500sec
2) 1000sec
3) 1500sec

After fetching the AutoML leaderboard we stored the models for future reference. The hyperparameters for all the run time is extracted and stored. These extracted Hyperparameters for each model is stored in JSON or CSV. The Meta_Data for every run is obtained. Knowledge of the relative importance of a hyperparameter to an algorithm and its range of values is crucial to hyperparameter tuning and creating effective models.

For example, we have considered a GBM model where we noticed an increase in the AUC (metric) with an increase in the number of trees (viz; Hyperparameter).



Furthermore, we have considered a GBM model where we noticed a decrease in the Log loss (metric) with an increase in the number of trees (viz; Hyperparameter).



VI. Discussion

In the current project we have found hyperparameters for the following Algorithms:

1) Gradient Boosting Machine
2) Generalized Linear Models
3) Distributed Random Forest

In the future, we are planning to include Deep learning Algorithms and Extremely Randomized Trees.

VII. Acknowledgment

We are thankful to our Prof. Nik Bear Brown, Assistant Professor, and Prabhu Subramanian, Teaching Assistant, Northeastern University, Boston and, MA for their valuable guidance, encouragement, and co-operation during the implementation of this project.

VIII. References

[1]Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian

Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[2]https://github.com/skunkworksneu/Projects/blob/master/Hyperparameter%20Database.pdf

[3]https://github.com/JeromeWynne/Predicting-Diabetic-Readmissions/blob/master/diabetes-ng.ipynb

[4]https://github.com/swengzju/Predicting-Diabetes-Patient-Readmission/blob/master/Predicting%20Diabetes%20Patient%20Readmission.ipynb

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

[6]https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568

[7]https://github.com/h2oai/h2o-3/blob/master/h2o-docs/src/product/tutorials/gbm/gbmTuning.ipynb

[8]http://h2o-release.s3.amazonaws.com/h2o/master/3233/docs-website/h2o-py/docs/h2o.model.html