MSc Project Report

Improving Boosting Classification Performance using Cluster Centroid Under Sampling & Bayesian Hyperparameter Optimization for Predict Patients' Readmission

Dileepa Nuwan Samayawardena
2019

A Report Submitted as Part of the Requirements for the Degree of MSc Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

Abstract

Healthcare has been considered as one of the main stream priorities in any countries. This field has been raising gradually towards the scientific decision making to improve the quality indicators of the health care services connecting to the healthcare 4.0. An immense amount of data transactions is processed in every second through out the several data platforms. That data from healthcare will be significance sources to derive insights and that insights will direct to the proactive set of actions for better improvements. Meanwhile, hospital readmission is an important measurement in terms of improving quality in healthcare industry and bending the healthcare cost curve to minimize the unnecessary expenditures. The hospital will be penalized in terms of financial cost if they have higher number of readmission rate than expected. This concept is not a common approach everywhere in global and this concept is practiced in limited locations in global based on conceptual manner. Therefore, developing a data driven prototype to properly demonstrate this concept will helpful to implement this concept in other locations globally.

When developing data driven system to prototype this, Understanding the knowledge pool of the data related to the hospital readmission is a significance objective for any researchers in the field of machine learning and big data. Based on the literatures, limited studies have been focused in this area of study and there is lack of concepts have been developed to model the data related to the hospital readmission. Therefore, this research has been focused a secondary data source related to the hospital readmission and done a powerful study to improve the binary classification performances to predict the hospital readmission behaviors of the patients. Imbalance nature of the class variable has been a significance noise for any development to come up with an accurate outcome in the domain of machine learning. Once, this has been properly done with a correct framework, fast and novel machine learning algorithms will provide an improved performance to predict readmission correctly. The concepts called hyperparameter optimization has been a significance approach to optimize the model performance further and fast approaches haven't been touched properly in this domain. Not only that, explaining predictions based on the Local Interpretable Model-agnostic Explanations (LIME) will be a key advantage for the users to do a proactive intervention successfully.

As the results summary, this research has proven the way of applying a concepts and frameworks on cluster centroid under sampling to class balancing inheritance issue related to the hospital readmissions data. And, vertically (leaf-wise) developed boosting algorithms with fast and accurate hyperparameters optimization method like Bayesian hyperparameter optimization have provided better outcomes than traditional way of doing this. The usage of Local Interpretable Model-agnostic Explanations has been tested and proven on top of the hospital readmission data while setting up the platform to further research in this domain. Finally, data driven prototype has been developed in a cloud platform with a well-defined data architecture.

Key Words: Readmission Prediction, Class Balancing, Gradient Boosting, Bayesian Hyperparameter optimizations, LIME