**Hyperparameters Database**

Bhavya Manvitha J, Manogjna Sindhu Potluri

**Abstract:** The aim of the project is to build a website where the user can get the hyperparameters for a particular dataset.This can be achieved by building a database for hyperparameters with the models of their particular algorithm. This databse is created by running many datasets with huge number of hyperparameter values.

**Keywords:**Hyperparameters, H2O, AutoML, Gradient Boost Machine(GBM), Generalised Linear Model(GLM), Extermely Randomised Trees(XRT), Distributed Random Forests(DRF), Grid search.

**1 Introduction**

The steps involved in finding the hyperparameters are:

* Select a dataset which is either classification or regression
* Load the dataset through H2O and run AutoML to find the leaderboard
* The leaderboard contains the best models ranked by their efficiency.
* The metrics based on which the hyperparameters are generated includes RMSE, MSE, AUC, logloss
* Metadata is generated for runtime, execution time and start time
* The parameters generated are stored into a json file.

**2 Exploring each step**

**2.1 Selecting a dataset**

A dataset of 11 columns containing liver patients information was selected. It was identified as a classification problem.The dataset contains various fields like age,gender,protein,total bilirubin,etc. which are used to determine whether the person is a liver patient or not. The target variable is the dataset column and it is a categorical column.

**2.2 Loading the dataset and analysis**

The dataset is loaded into the python environment and H2O algorithm is performed on it. The AutoML generates a leaderboard of all the models from the algorithms that H2O supports.These models are ranked according to their efficiency. This efficiency is determined based on various metrics like RMSE,MSE,AUC,logloss

**RMSE**

The Root Mean Square Error helps us to find out how distributed our data is around the line of best fit.The less the RMSE value, the better the model fit.

**AUC**

The Area Under the Curve of an ROC helps us to determine the capability of the model to distinguish between classes.Higher AUC results in a better model predictability.

**Logloss**

The logarithmic loss is used to analyse the performance of the model.Ideally the loss must be 0 for the model to perform better.

**2.3 Leaderboard analysis**

As H2O supports only selected algorithms, the leaderboard contains the models derived through these algorithms only.

**Gradient Boost Machine(GBM)**

They improve the learning capability of a model.It is a ensemble method that performs forward learning.The first boosting algorithm,Adaboost is generalised to obtain a gradient boost machine.The efficiency of the model is determined using a loss function. GBM can quickly overfit the data which is it’s main drawback.

**Generalised Linear Model(GLM)**

It is a normal linear regression model but with more flexibility.The target variable need not be normally distributed.Eventhough they are error distributed, GLM can be performed.

**Extreme Randomised Trees(XRT)**

They are a class of random trees. An extra step of randomization is performed to obtain the extreme random trees or extra trees.

**Distributed Random Forests (DRF)**

DRF is an amazing arrangement and relapse instrument. At the point when given a lot of information, DRF produces a group of trees as opposed to a single tree. Every one of these trees is a week one based on a subset of lines and sections. More trees will decrease the change. Both classification and regression take the normal expectation over the majority of their trees to make a last forecast for both class and numeric esteem.

**Grid Search**

The way toward finding most ideal hyperparameters in ML is called hyperparameter optimisation.Grid search is one basic algorithm.In this strategy, we construct a model for each blend of different hyperparameters and assess each model. The model which gives the most astounding precision is choosen.