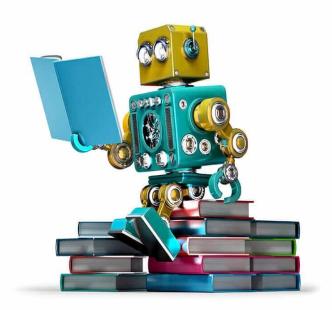
# 'Hyper'Parameters—predicted right!

# Manasa Vanga, Tejas Munot

vanga.m@husky.neu.edu, munot.t@husky.neu.edu

INFO 6210, Spring 2019, Northeastern University



Our very own Machine Learning Superhero!

#### 1. Introduction

In the good old days, computers had to be programmed and told what to do. In some cases, you even had to tell them how to do it. Life was monotonous back then. But Arthur Samuel didn't like it. He saw that there's more to it, and thus began the chapter of Machine Learning!

Machine learning is the art of teaching the machine how to learn. It can thus, identify patterns and make decisions without humans telling it to do so.

How does it do it? you ask. The answer is by building analytical models. A machine learning model is a mathematical formula having some parameters which are needed to be learned from the data through model training. Thus, we fit the model parameters. But some parameters cannot be learned directly through model training. These are called **Hyperparameters**.

Hyperparameters express the properties of the model. For example, its complexity, how fast should the model learn. These must be specified **before** the model training process starts. Thus, knowledge of the hyperparameters is important as it has a large effect on the predictive power of the statistical model and that is what we, students at Northeastern University, Boston are working on under the <u>AI SkunkWorks</u> association.

Our aim is to create a physical database which would support a website where data enthusiasts could upload their datasets and based on the millions of dataset records we have, give the user the best values of the respective hyperparameters for whatever he/she plans to achieve through the model. With this, we can help the data scientists save a lot of their time which they can utilize to find that one missing semicolon in the hundreds of lines of code! Fun isn't it?

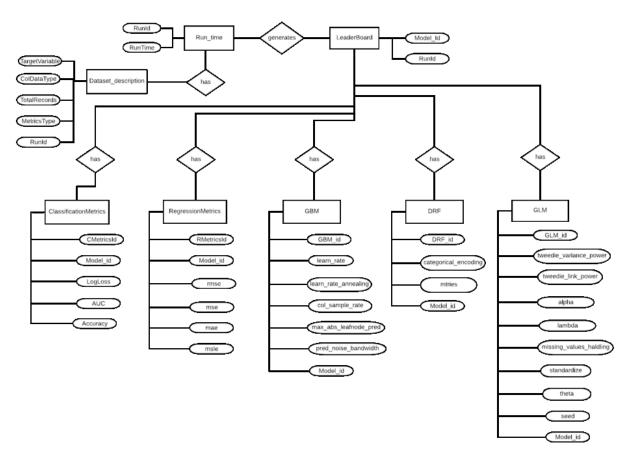
We are a team of 4 students, 2 aspiring data scientists, Prakruthi and Urja, and 2 database engineers, myself and Manasa. The data science (DS) team plans to use H2O which is a fully open source, distributed in-memory machine learning platform with linear scalability. H2O supports the most widely used statistical & machine learning algorithms including gradient boosted machines, generalized linear models, deep learning and more. H2O

also has an industry leading **AutoML** functionality that automatically runs through all the algorithms and their hyperparameters to produce a leaderboard of the best models.

#### 2. Code with Documentation

# 2.1 Conceptual Diagram

We then identified the entities, attributes, and relationships. Thus, we came up with the first version of our conceptual diagram.

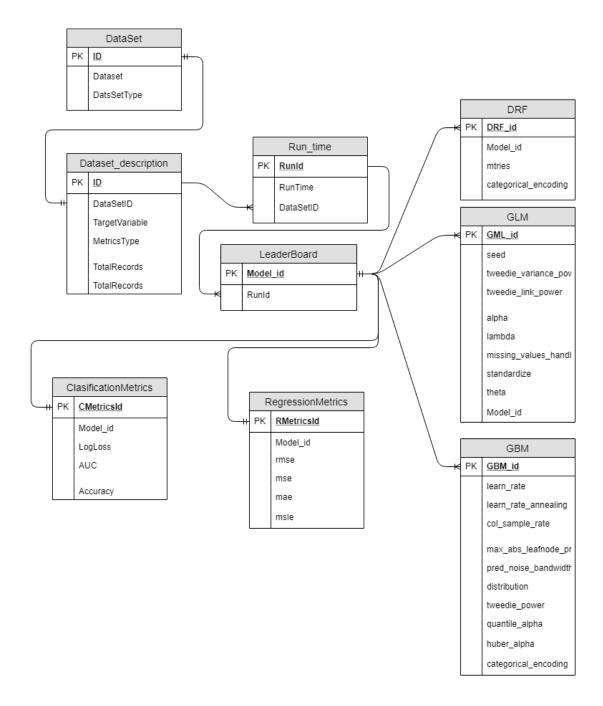


Conceptual Diagram—Iteration 1

But, as it always does, our conceptual diagram evolved throughout the course of the project. As we went ahead and addressed the uses cases for the database, we had to change the conceptual diagram

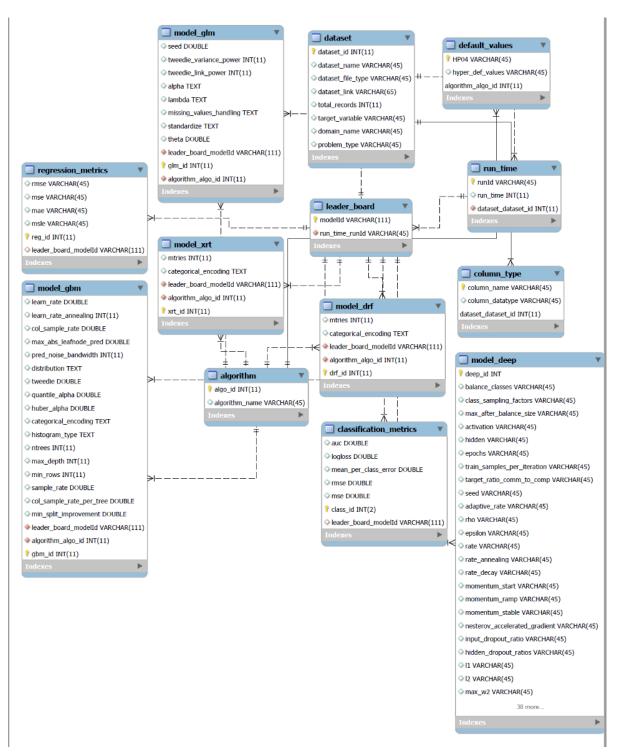
Based on the conceptual diagram, we made the Entity-Relationship Diagram (ERD) addressing the cardinalities. We were now ready to create the physical database. We used MySQL Workbench extensively. All our queries are written and executed in MySQL Workbench.

# 2.2 ER-Diagram



# Revised ER - Diagram:

#### The Final ERD is as follows:



Entity Relationship Diagram

# 2.3.Tables

# • Entity Sets:

The Final tables and their description

- 1. **Dataset**—It has all the information about the dataset. (PK—dataset id)
- 2. **Run\_time**—It has the runtimes and the run-Id's for each runtime (PK—runId; FK—dataset\_dataset\_id)
- 3. **Leader\_board**—It has the model-Id's each run has generated (PK—modelId; FK—run\_time\_runId)
- 4. **Default\_values**—It has all the hyperparameters for each algorithm and its default values (PK—HPo4)
- 5. **Column\_type**—It has the dataset's information specific to the columns and its data types given to H20. (PK—column\_name, dataset\_dataset\_id)
- 6. **Classification\_metrics**—It has the classification metrics for each classification dataset (PK—classId; FK—leader\_board\_modelId)
- 7. **Regression\_metrics**—It has the regression metrics for each regression dataset (PK—reg\_id; FK—leader\_board\_modelId)
- 8. **Algorithm**—All the algorithms stored in the table (PK—algo\_id)
- 9. **Model\_glm**—All the models from the GLM algorithm which got generated during running H20 (PK—glm\_id; FK—algorithm\_algo\_id, leader\_board\_model)
- 10. **Model\_gbm**—All the models from the GBM algorithm which got generated during running H20(PK—gbm\_id; FK—algorithm\_algo\_id, leader\_board\_model)
- 11. **Model\_xrt**—All the models from the XRT algorithm which got generated during running H20(PK—xrt\_id; FK—algorithm\_algo\_id, leader\_board\_model)

- 12.**Model\_deep**—All the models from the Deep Learning algorithm which got generated during running H2o(PK—deep\_id; FK—algorithm\_algo\_id, leader\_board\_model)
- 13.**Model\_drf**—All the models from the DRF algorithm which got generated during running H20(PK—drf\_id; FK—algorithm\_algo\_id, leader board model)

# • RelationShips:

- 1. R1 dataset\_id, runId One dataset can run for multiple runtimes. This is a one to many relationship between dataset and run time.
- 2. R2 modelId, run\_time\_runId For one run time ,multiple models from different algorithms are generated. It is a one to many relationship between leader\_board and run\_time.
- **3.** R3 PK—glm\_id; FK—(algorithm\_algo\_id, leader\_board\_model)
  For one Algorithm and leader\_board\_model, there are multiple models has its hyperparameters. It is a one to many relationship between model gbm and leader board.

Similar relationship exists for other Algorithms and is models.

#### • Normalization:

- 1. First normal form (1NF):
- Each table has a primary key
- The values in each column of a table are atomic. No multi-value attributes found.
- No repeating groups in all the tables.
- 2. Second normal form (2NF):
- All requirements for 1st NF are met.
- Partial dependencies are eliminated by creating separate entity sets.

- There is no calculated data
- 3. Third normal form (3NF):
- All requirements for 2nd NF are met.
- Fields that do not directly depend on the primary key are removed and creates as separate entities thus removed transitive dependencies.

# 2.4 Creating the physical database

After we finalized the ERD, we imported data into the tables. Then, using Forward Engineering, we got all the tables along with the foreign constraints and thus our schema was ready for querying.

## 3.Use Cases:

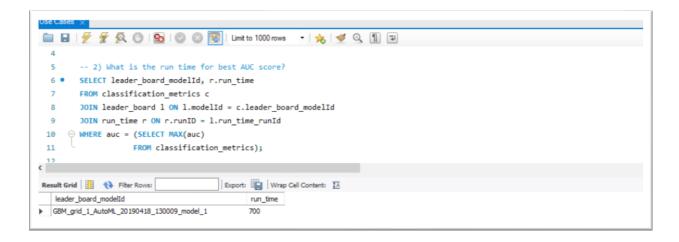
We then got together with the Data Science (DS) team to come up with some practical queries which could help the data scientists.

1. What is the maximum accuracy from all the generated models?

SELECT MAX(auc) AS MAX\_AUC, leader\_board\_modelId FROM classification\_metrics;

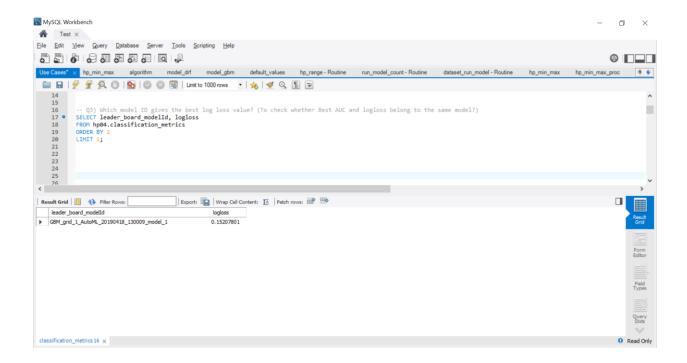
2) What is the run time for the best AUC score?

SELECT leader\_board\_modelId, r.run\_time FROM classification\_metrics c JOIN leader\_board l ON l.modelId = c.leader\_board\_modelId JOIN run\_time r ON r.runID = l.run\_time\_runId WHERE auc = (SELECT MAX(auc) FROM classification\_metrics);



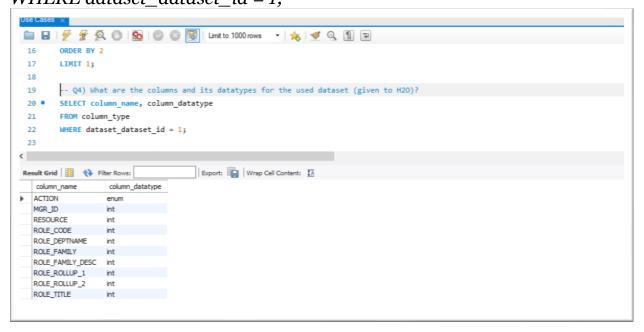
Q3) Which model ID gives the best log loss value? (To check whether Best AUC and logloss belong to the same model?)

SELECT leader\_board\_modelId, logloss, auc FROM classification\_metrics ORDER BY 2 LIMIT 1;



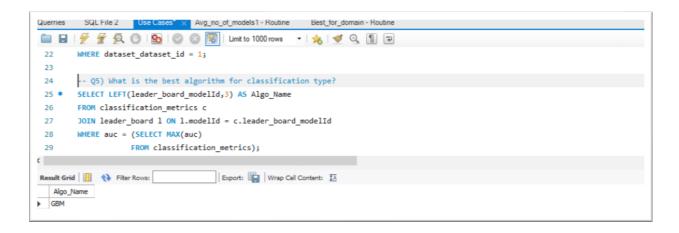
Q4) What are the columns and its datatypes for the used dataset (given to H2O)?

SELECT column\_name, column\_datatype FROM column\_type WHERE dataset\_dataset\_id = 1;



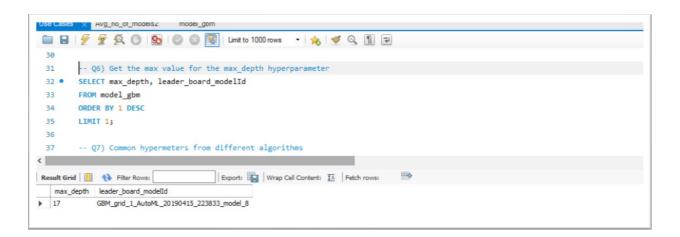
# Q5) What is the best algorithm for classification type?

SELECT LEFT(leader\_board\_modelId,3) AS Algo\_Name FROM classification\_metrics c JOIN leader\_board l ON l.modelId = c.leader\_board\_modelId WHERE auc = (SELECT MAX(auc) FROM classification\_metrics);



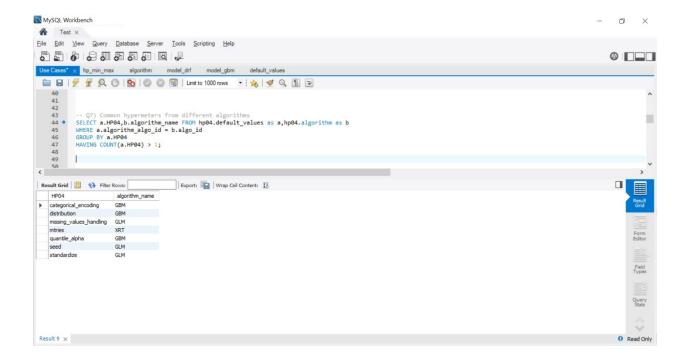
### Q6) Get the Max value of the max\_depth hyperparameter

SELECT max\_depth, leader\_board\_modelId FROM model\_gbm ORDER BY 1 DESC LIMIT 1;



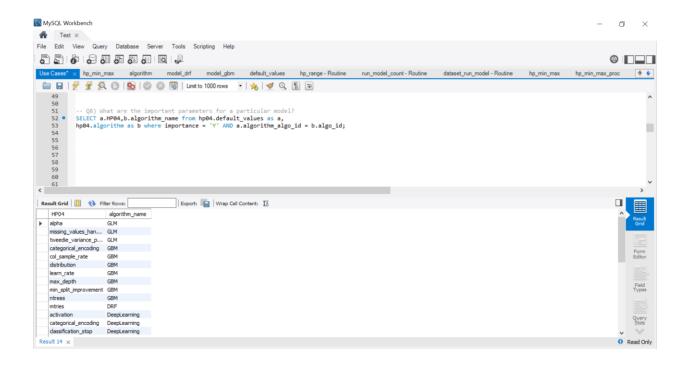
# Q7) Common hypermeters from different algorithms

SELECT a.HPo4,b.algorithm\_name FROM hpo4.default\_values as a,hpo4.algorithm as b
WHERE a.algorithm\_algo\_id = b.algo\_id
GROUP BY a.HPo4
HAVING COUNT(a.HPo4) > 1;

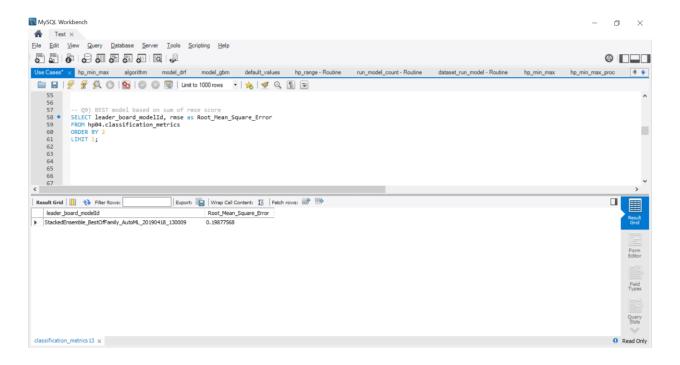


Q8) What are the important parameters for a particular model?

SELECT a.HPo4,b.algorithm\_name from hpo4.default\_values as a, hpo4.algorithm as b where importance = 'Y' AND a.algorithm\_algo\_id = b.algo\_id;

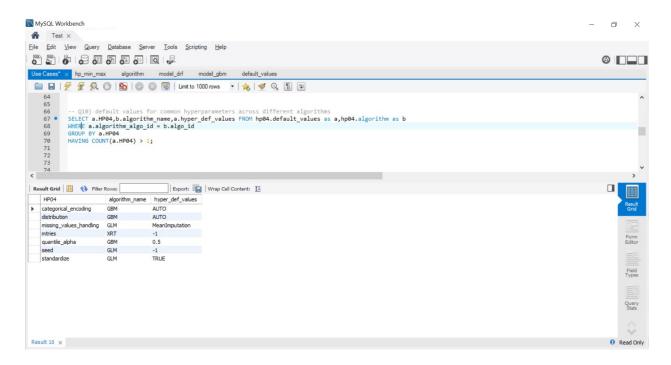


# Q9) BEST model based on rmse score: SELECT leader\_board\_modelId, rmse as Root\_Mean\_Square\_Error FROM hpo4.classification\_metrics ORDER BY 2 LIMIT 1;



# Q10) Default values for common hyperparameters across different algorithms

SELECT a.HPo4,b.algorithm\_name,a.hyper\_def\_values FROM hpo4.default\_values as a,hpo4.algorithm as b WHERE a.algorithm\_algo\_id = b.algo\_id GROUP BY a.HPo4 HAVING COUNT(a.HPo4) > 1;



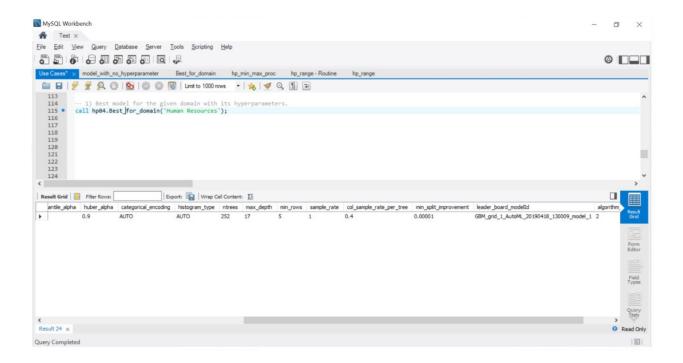
# 4. Functions and stored procedures:

1. The best model for the given domain with its hyperparameters.

Input: Domain Name

Output: The best model with its actual hyperparameter values

```
CODE:
DELIMITER $$
CREATE DEFINER= `root`@ `localhost` PROCEDURE
`Best for domain`(IN domain name1 VARCHAR(45))
BEGIN
DECLARE model VARCHAR(111);
SELECT\ c.leader\_board\_modelId\ INTO\ model
FROM classification metrics c
JOIN leader board l ON l.modelId = c.leader board modelId
JOIN run time r ON r.runID = l.run_time_runId
JOIN dataset d ON r.dataset dataset id = d.dataset id
WHERE domain name = domain name1
AND auc = (SELECT MAX(auc))
FROM classification metrics);
IF(LEFT(model,3) = GBM')
THEN
SELECT * from model gbm WHERE leader board modelId = model;
ELSEIF(LEFT(model,3) = GLM') THEN
SELECT * from model_glm WHERE leader_board_modelId = model;
ELSEIF(LEFT(model,3) = 'DRF') THEN
SELECT * from model drf WHERE leader board modelId = model;
ELSEIF(LEFT(model,3) = 'XRT') THEN
SELECT * from model xrt WHERE leader board modelId = model;
END IF;
END$$
DELIMITER;
```



2. What are the minimum and maximum values of learn\_rate, ntrees hyperparameter?

Input: Hyperparameter—learn\_rate or ntrees

Output: Minimum and Maximum values of learn\_rate or ntrees

#### CODE:

**DELIMITER \$\$** 

CREATE DEFINER= `root`@ `localhost` PROCEDURE

`hp\_min\_max\_proc`(IN parameter varchar(100))

**BEGIN** 

DECLARE default\_value varchar(45);

select distinct HPo4 from hpo4.default\_values;

IF (parameter = 'learn rate') THEN

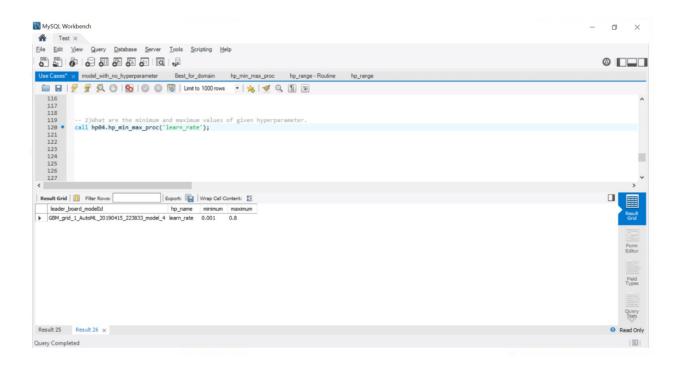
SELECT leader\_board\_modelId,'learn\_rate' as hp\_name,

min(learn\_rate) as minimum,max(learn\_rate) as maximum from hpo4.model\_gbm;

ELSEIF(parameter = 'ntrees') THEN

SELECT leader\_board\_modelId,'ntrees' as hp\_name, min(ntrees) as minimum,max(ntrees) as maximum from hpo4.model\_gbm; END IF;

# END\$\$ DELIMITER;



3. Count of all models under given algorithm and runtime.

Input: Algorithm name, runtime

CODE:

Output: Number of models created for that model and runtime

```
DELIMITER $$

CREATE DEFINER=`root`@`localhost`FUNCTION

`Avg_no_of_models1`(algorithm_name VARCHAR(45), run_times INT(11)) RETURNS decimal(10,0)

DETERMINISTIC

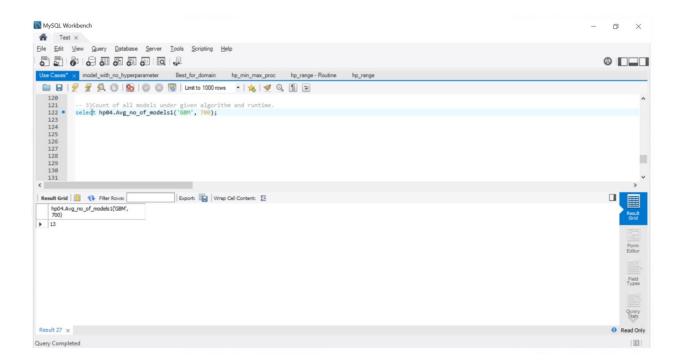
BEGIN

DECLARE lvl decimal(10);

SELECT COUNT(*) AS Total_models INTO lvl

FROM leader_board l
```

JOIN run\_time r ON l.run\_time\_runId = r.runId WHERE LEFT(modelId,3) LIKE concat(algorithm\_name,'%') AND r.run\_time = run\_times; RETURN (lvl); END\$\$ DELIMITER;



4. What are the actual, default values and the range of learn\_rate hyperparameter

Output: All hyperparameter default values

#### CODE:

**DELIMITER \$\$** 

 $\begin{tabular}{ll} CREATE DEFINER = `root` @ `localhost` PROCEDURE `hp\_range` () \\ BEGIN \end{tabular}$ 

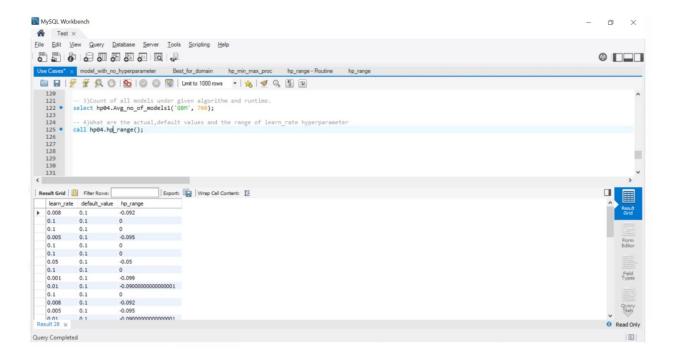
DECLARE default\_value varchar(45);

SELECT hyper\_def\_values INTO default\_value from hpo4.default\_values where HPo4 = 'learn rate';

select learn\_rate, default\_value, (learn\_rate-default\_value) as hp\_range from hpo4.model\_gbm;

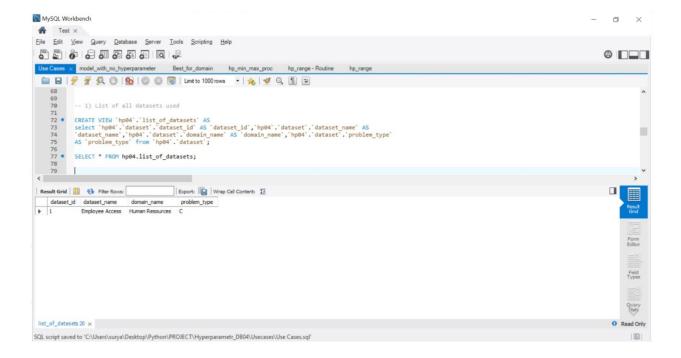
END\$\$

DELIMITER;

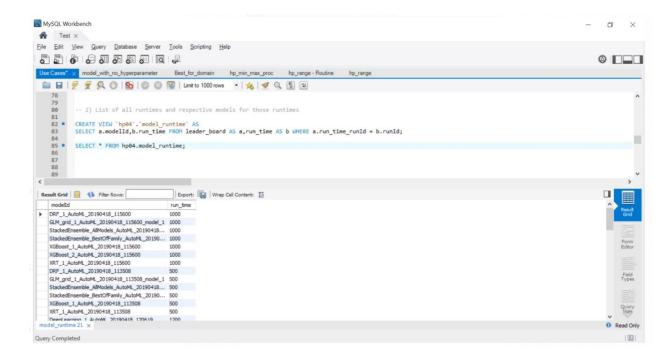


# 5. Views:

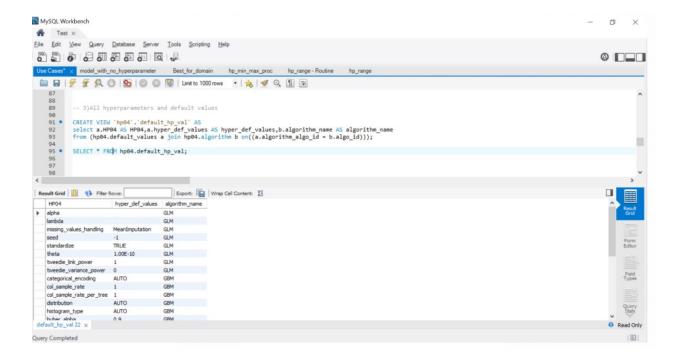
1. View of a list of all the datasets available in the database



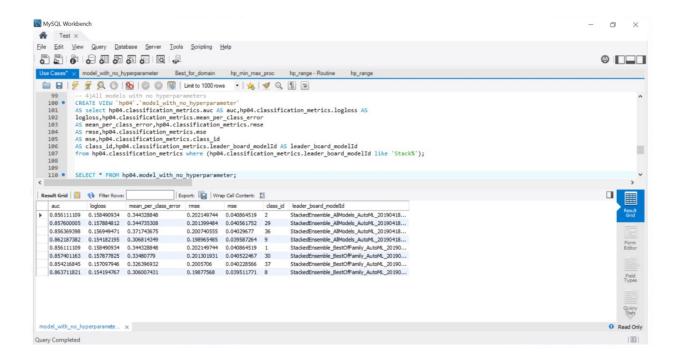
# 2. List of all runtimes and respective models for those runtimes



# 3) All hyperparameters and default values



4) All models with no hyperparameters



## 6. Conclusion:

Thus, after the project, we were able to create an actual physical database storing the hyperparameters' actual and default values. Through the demonstration of the use cases, we will be able to support a website for the same. The following points were covered:

- 1. Conceptual Diagram
- 2. ER-Diagram
- 3. Normalization
- 4. Creating a physical database
- 5. Converting JSON files into CSV files based on the model

- 6. Use Case preparation
- 7. Functions
- 8. Views
- 9. Stored Procedures
- 10. Documentation and Professionalism

# 7. Citations:

- 1. <a href="http://docs.h2o.ai/h2o/latest-stable/h2o-docs/grid-search.html">http://docs.h2o.ai/h2o/latest-stable/h2o-docs/grid-search.html</a>—H20 Hyperparameters
- 2. <a href="https://www.hindawi.com/journals/complexity/2019/6278908/">https://www.hindawi.com/journals/complexity/2019/6278908/</a>- Sampling of a Dataset in the Hyperparameter.
- 3. <a href="https://github.com/nikbearbrown/INFO">https://github.com/nikbearbrown/INFO</a> 6210—Prof. Nik's Git Hub
- 4. <a href="http://www.mysqltutorial.org/mysql-stored-procedure-tutorial.aspx">http://www.mysqltutorial.org/mysql-stored-procedure-tutorial.aspx</a>—
  Stored Procedures.
- 5. <a href="https://www.northeastern.edu/rise/presentations/hyperparameter-database/">https://www.northeastern.edu/rise/presentations/hyperparameter-database/</a> Hyperparameter Database(RISE-2019)
- 6. <a href="http://www.mysqltutorial.org/mysql-functions.aspx">http://www.mysqltutorial.org/mysql-functions.aspx</a> Functions

#### 8. License:

Copyright 2019 Tejas Munot, Manasa Vanga

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.