Summer Internship Presentation

Nikhil Shekhar
Navdeep Gill
08/12/2016

H_2O.ai

Assignment Details:

Develop a framework that can help benchmark multiple machine learning libraries

Background

Education

- Masters in Computer Science at State university of New York at Buffalo - **Graduating December 2016**.
- Bachelor of Engineering in Computer Science.

Work Experience

- 4 years in Big Data Eco-space in companies - Impetus technologies, Center of Excellence at Cognizant.
- Worked on multiple Big data projects during the 4 years for clients like Capital One, American express, AT&T, Cisco, Lexus.
- 1 year in Java - J2EE technologies.

Project details

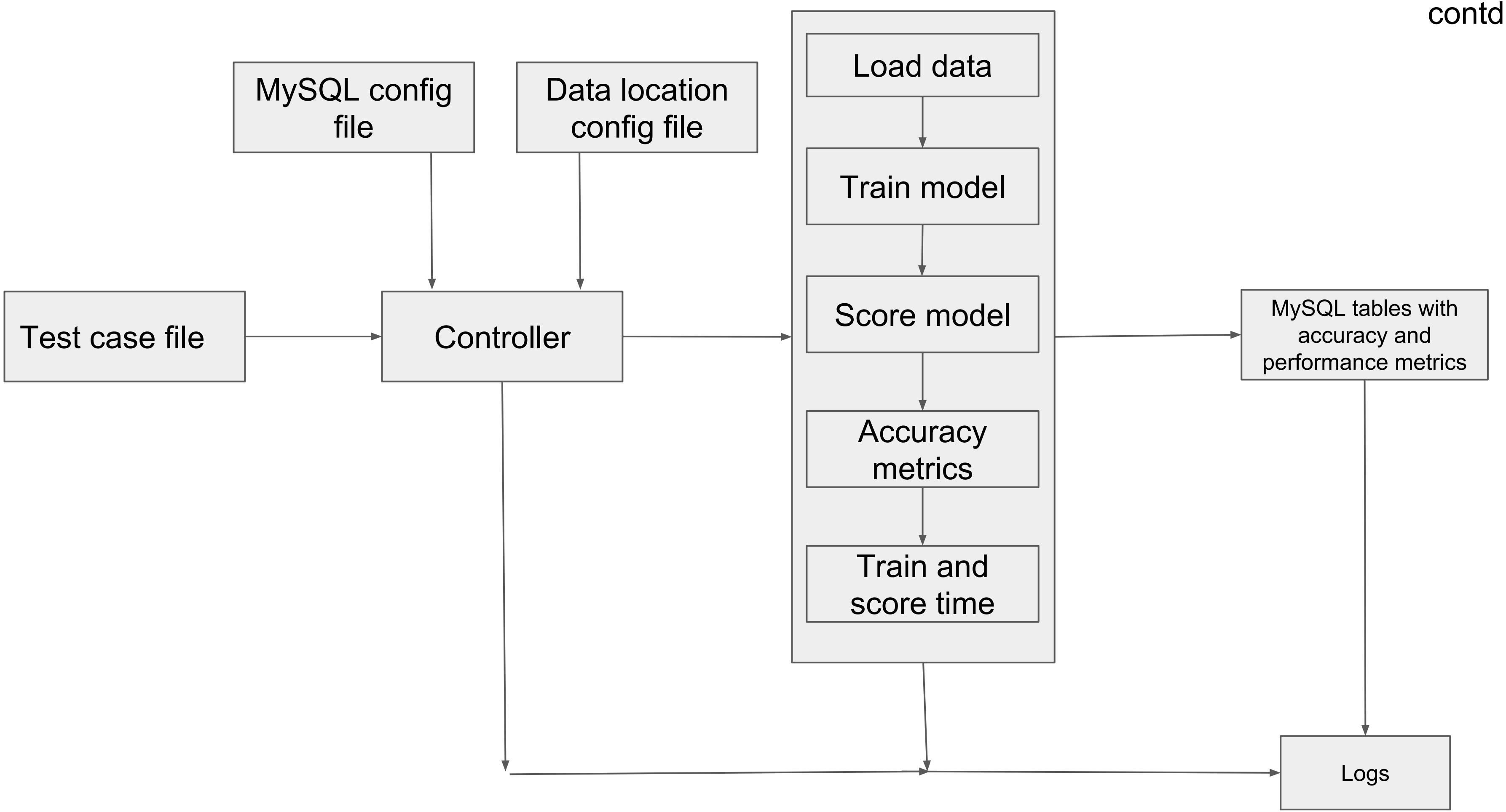
ML Frameworks used for benchmarking efforts:

- Spark-ml
- Xgboost
- scikit

Benchmark workflow

- Read the test case file and get the Algorithm details and the hyper-parameters for the same.
- Connect to EC2 S3 storage and read the data files needed for the test case.
- Generate the feature matrix
- Run the algorithm
- Collect the accuracy metrics
- Collect the performance metrics - training and scoring time
- Push the metrics to MySQL for each test case and each run
- Logs are written to a configurable directory

contd...



MySQL config file

Data location config file

Test case file

Controller

Load data

Train model

Score model

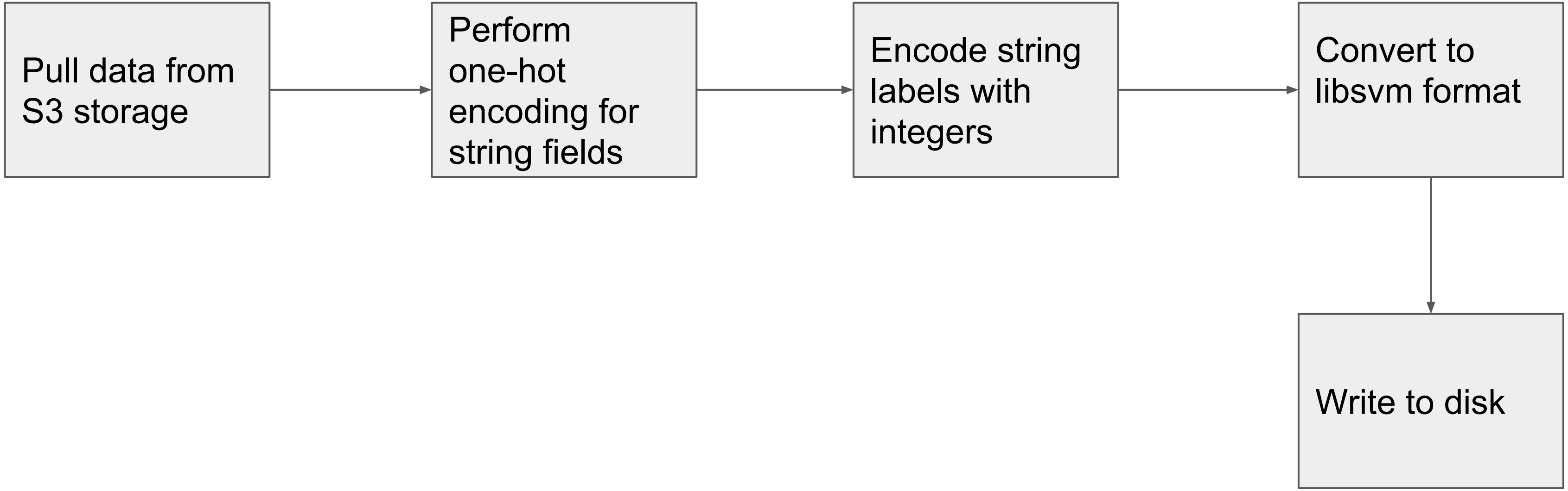
Accuracy metrics

Train and score time

MySQL tables with accuracy and performance metrics

Logs

Preprocessing data for xgboost and scikit



Pull data from S3 storage

Perform one-hot encoding for string fields

Encode string labels with integers

Convert to libsvm format

Write to disk

Benchmark - MLlib

Algorithms supported:

- Linear regression
- Logistic regression - Supports only binary classification.
- Random forest classifier
- Random forest regressor
- Gradient boosted machine classifier - Supports only binary classification.
- Gradient boosted machine regressor
- Perceptron classifier
- KMeans
- Principal component analysis

Achieved through Java API’s on Spark - 1.6.2

Benchmark - xgboost

Algorithms supported:

- Gradient Boosting algorithm that supports regression and classification.

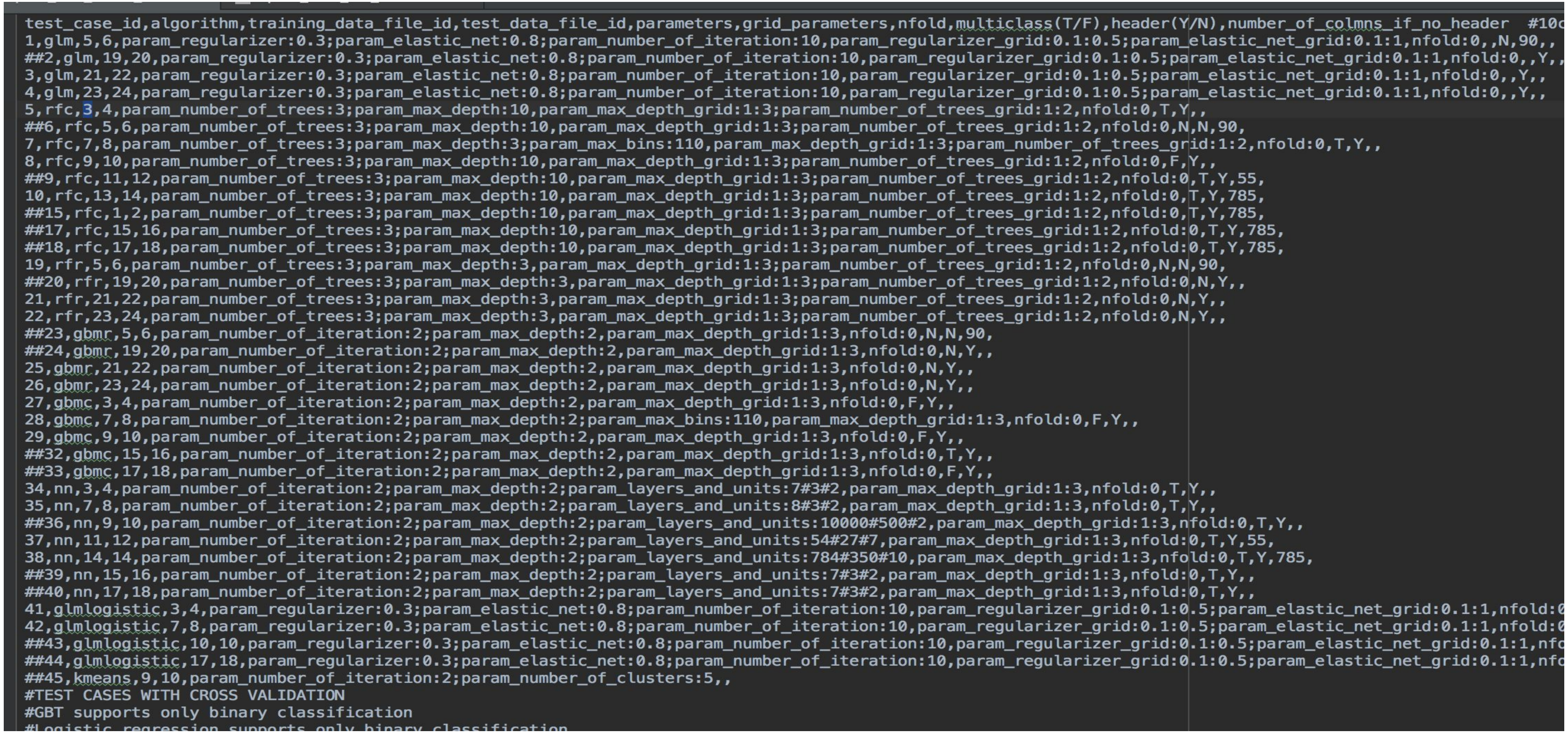
- A pre-processing step is needed which pulls data from S3, performs one-hot encoding and creates a libsvm file which is fed to xgboost
- Achieved through Java API’s on xgboost - 0.6

Benchmark - scikit learn

Algorithms supported:

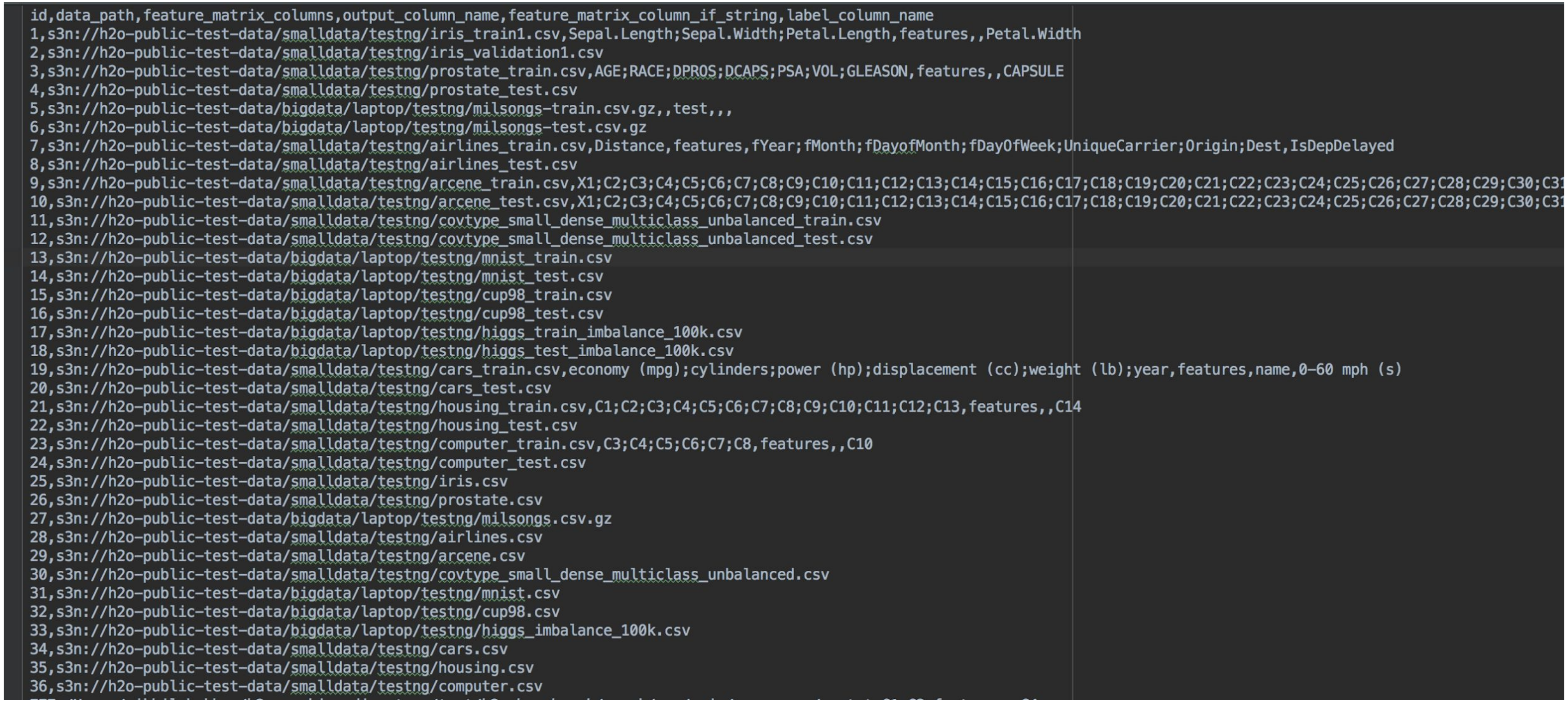
  - Linear regression
  - Logistic regression
  - Random forest classifier
  - Random forest regressor
  - Gradient boosted machine classifier
  - Gradient boosted machine regressor
  - Naive Bayes
  - KMeans
- A pre-processing step is needed which pulls data from S3 and performs one-hot encoding
- Achieved through Python API’s on scikit - 0.17

Test cases file



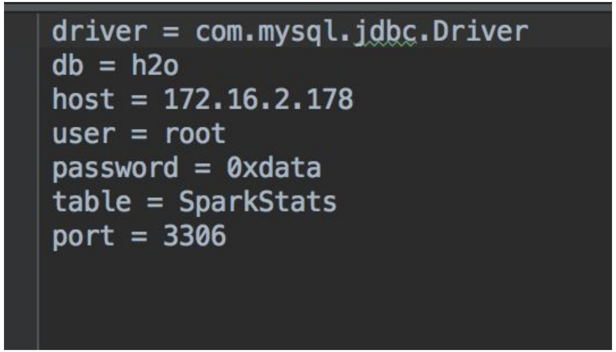
```
test_case_id,algorithm,training_data_file_id,test_data_file_id,parameters,grid_parameters,nfold,multiclass(T/F),header(Y/N),number_of_colmns_if_no_header  #10c
1,glm,5,6,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfold:0,,N,90,,
##2,glm,19,20,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfold:0,,Y,,
3,glm,21,22,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfold:0,,Y,,
4,glm,23,24,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfold:0,,Y,,
5,rfc,3,4,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,,
##6,rfc,5,6,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,N,N,90,
7,rfc,7,8,param_number_of_trees:3;param_max_depth:3;param_max_bins:110,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,,
8,rfc,9,10,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,F,Y,,
##9,rfc,11,12,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,55,
10,rfc,13,14,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,785,
##15,rfc,1,2,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,785,
##17,rfc,15,16,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,785,
##18,rfc,17,18,param_number_of_trees:3;param_max_depth:10,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,T,Y,785,
19,rfr,5,6,param_number_of_trees:3;param_max_depth:3,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,N,N,90,
##20,rfr,19,20,param_number_of_trees:3;param_max_depth:3,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,N,Y,,
21,rfr,21,22,param_number_of_trees:3;param_max_depth:3,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,N,Y,,
22,rfr,23,24,param_number_of_trees:3;param_max_depth:3,param_max_depth_grid:1:3;param_number_of_trees_grid:1:2,nfold:0,N,Y,,
##23,gbmr,5,6,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,N,N,90,
##24,gbmr,19,20,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,N,Y,,
25,gbmr,21,22,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,N,Y,,
26,gbmr,23,24,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,N,Y,,
27,gbmc,3,4,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,F,Y,,
28,gbmc,7,8,param_number_of_iteration:2;param_max_depth:2;param_max_bins:110,param_max_depth_grid:1:3,nfold:0,F,Y,,
29,gbmc,9,10,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,F,Y,,
##32,gbmc,15,16,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,T,Y,,
##33,gbmc,17,18,param_number_of_iteration:2;param_max_depth:2,param_max_depth_grid:1:3,nfold:0,F,Y,,
34,nn,3,4,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:7#3#2,param_max_depth_grid:1:3,nfold:0,T,Y,,
35,nn,7,8,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:8#3#2,param_max_depth_grid:1:3,nfold:0,T,Y,,
##36,nn,9,10,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:10000#500#2,param_max_depth_grid:1:3,nfold:0,T,Y,,
37,nn,11,12,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:54#27#7,param_max_depth_grid:1:3,nfold:0,T,Y,55,
38,nn,14,14,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:784#350#10,param_max_depth_grid:1:3,nfold:0,T,Y,785,
##39,nn,15,16,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:7#3#2,param_max_depth_grid:1:3,nfold:0,T,Y,,
##40,nn,17,18,param_number_of_iteration:2;param_max_depth:2;param_layers_and_units:7#3#2,param_max_depth_grid:1:3,nfold:0,T,Y,,
41,glmlogistic,3,4,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfold:0
42,glmlogistic,7,8,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfold:0
##43,glmlogistic,10,10,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfo
##44,glmlogistic,17,18,param_regularizer:0.3;param_elastic_net:0.8;param_number_of_iteration:10,param_regularizer_grid:0.1:0.5;param_elastic_net_grid:0.1:1,nfo
##45,kmeans,9,10,param_number_of_iteration:2;param_number_of_clusters:5,,
#TEST CASES WITH CROSS VALIDATION
#GBT supports only binary classification
#Logistic regression supports only binary classification
```

Data file details



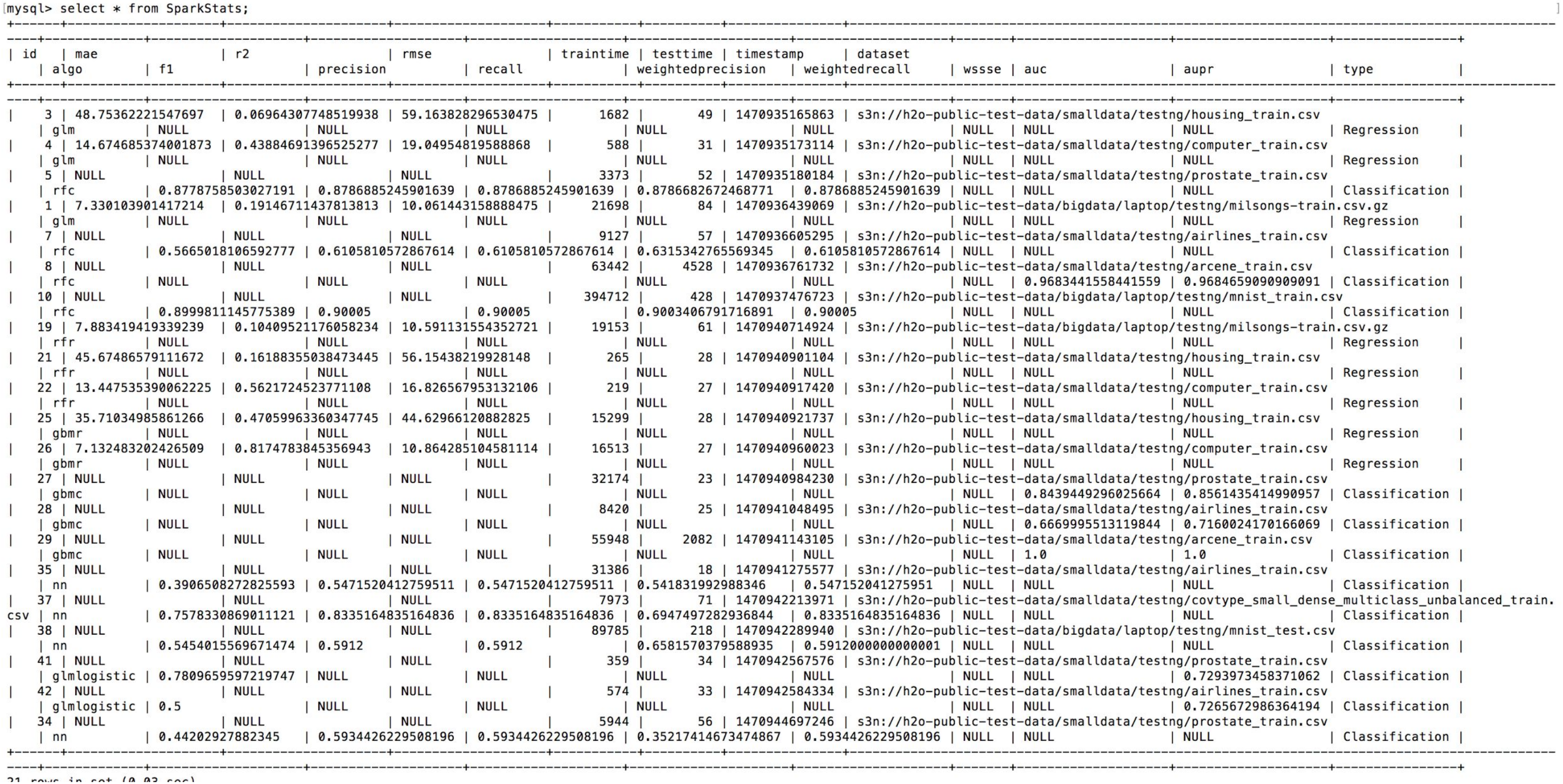
```
id,data_path,feature_matrix_columns,output_column_name,feature_matrix_column_if_string,label_column_name
1,s3n://h2o-public-test-data/smalldata/testng/iris_train1.csv,Sepal.Length;Sepal.Width;Petal.Length,features,,Petal.Width
2,s3n://h2o-public-test-data/smalldata/testng/iris_validation1.csv
3,s3n://h2o-public-test-data/smalldata/testng/prostate_train.csv,AGE;RACE;DPROS;DCAPS;PSA;VOL;GLEASON,features,,CAPSULE
4,s3n://h2o-public-test-data/smalldata/testng/prostate_test.csv
5,s3n://h2o-public-test-data/bigdata/laptop/testng/milsongs-train.csv.gz,,test,,,
6,s3n://h2o-public-test-data/bigdata/laptop/testng/milsongs-test.csv.gz
7,s3n://h2o-public-test-data/smalldata/testng/airlines_train.csv,Distance,features,fYear;fMonth;fDayofMonth;fDayOfWeek;UniqueCarrier;Origin;Dest,IsDepDelayed
8,s3n://h2o-public-test-data/smalldata/testng/airlines_test.csv
9,s3n://h2o-public-test-data/smalldata/testng/arcene_train.csv,X1;C2;C3;C4;C5;C6;C7;C8;C9;C10;C11;C12;C13;C14;C15;C16;C17;C18;C19;C20;C21;C22;C23;C24;C25;C26;C27;C28;C29;C30;C31
10,s3n://h2o-public-test-data/smalldata/testng/arcene_test.csv,X1;C2;C3;C4;C5;C6;C7;C8;C9;C10;C11;C12;C13;C14;C15;C16;C17;C18;C19;C20;C21;C22;C23;C24;C25;C26;C27;C28;C29;C30;C31
11,s3n://h2o-public-test-data/smalldata/testng/covtype_small_dense_multiclass_unbalanced_train.csv
12,s3n://h2o-public-test-data/smalldata/testng/covtype_small_dense_multiclass_unbalanced_test.csv
13,s3n://h2o-public-test-data/bigdata/laptop/testng/mnist_train.csv
14,s3n://h2o-public-test-data/bigdata/laptop/testng/mnist_test.csv
15,s3n://h2o-public-test-data/bigdata/laptop/testng/cup98_train.csv
16,s3n://h2o-public-test-data/bigdata/laptop/testng/cup98_test.csv
17,s3n://h2o-public-test-data/bigdata/laptop/testng/higgs_train_imbalance_100k.csv
18,s3n://h2o-public-test-data/bigdata/laptop/testng/higgs_test_imbalance_100k.csv
19,s3n://h2o-public-test-data/smalldata/testng/cars_train.csv,economy (mpg);cylinders;power (hp);displacement (cc);weight (lb);year,features,name,0-60 mph (s)
20,s3n://h2o-public-test-data/smalldata/testng/cars_test.csv
21,s3n://h2o-public-test-data/smalldata/testng/housing_train.csv,C1;C2;C3;C4;C5;C6;C7;C8;C9;C10;C11;C12;C13,features,,C14
22,s3n://h2o-public-test-data/smalldata/testng/housing_test.csv
23,s3n://h2o-public-test-data/smalldata/testng/computer_train.csv,C3;C4;C5;C6;C7;C8,features,,C10
24,s3n://h2o-public-test-data/smalldata/testng/computer_test.csv
25,s3n://h2o-public-test-data/smalldata/testng/iris.csv
26,s3n://h2o-public-test-data/smalldata/testng/prostate.csv
27,s3n://h2o-public-test-data/bigdata/laptop/testng/milsongs.csv.gz
28,s3n://h2o-public-test-data/smalldata/testng/airlines.csv
29,s3n://h2o-public-test-data/smalldata/testng/arcene.csv
30,s3n://h2o-public-test-data/smalldata/testng/covtype_small_dense_multiclass_unbalanced.csv
31,s3n://h2o-public-test-data/bigdata/laptop/testng/mnist.csv
32,s3n://h2o-public-test-data/bigdata/laptop/testng/cup98.csv
33,s3n://h2o-public-test-data/bigdata/laptop/testng/higgs_imbalance_100k.csv
34,s3n://h2o-public-test-data/smalldata/testng/cars.csv
35,s3n://h2o-public-test-data/smalldata/testng/housing.csv
36,s3n://h2o-public-test-data/smalldata/testng/computer.csv
```
]

MySQL config



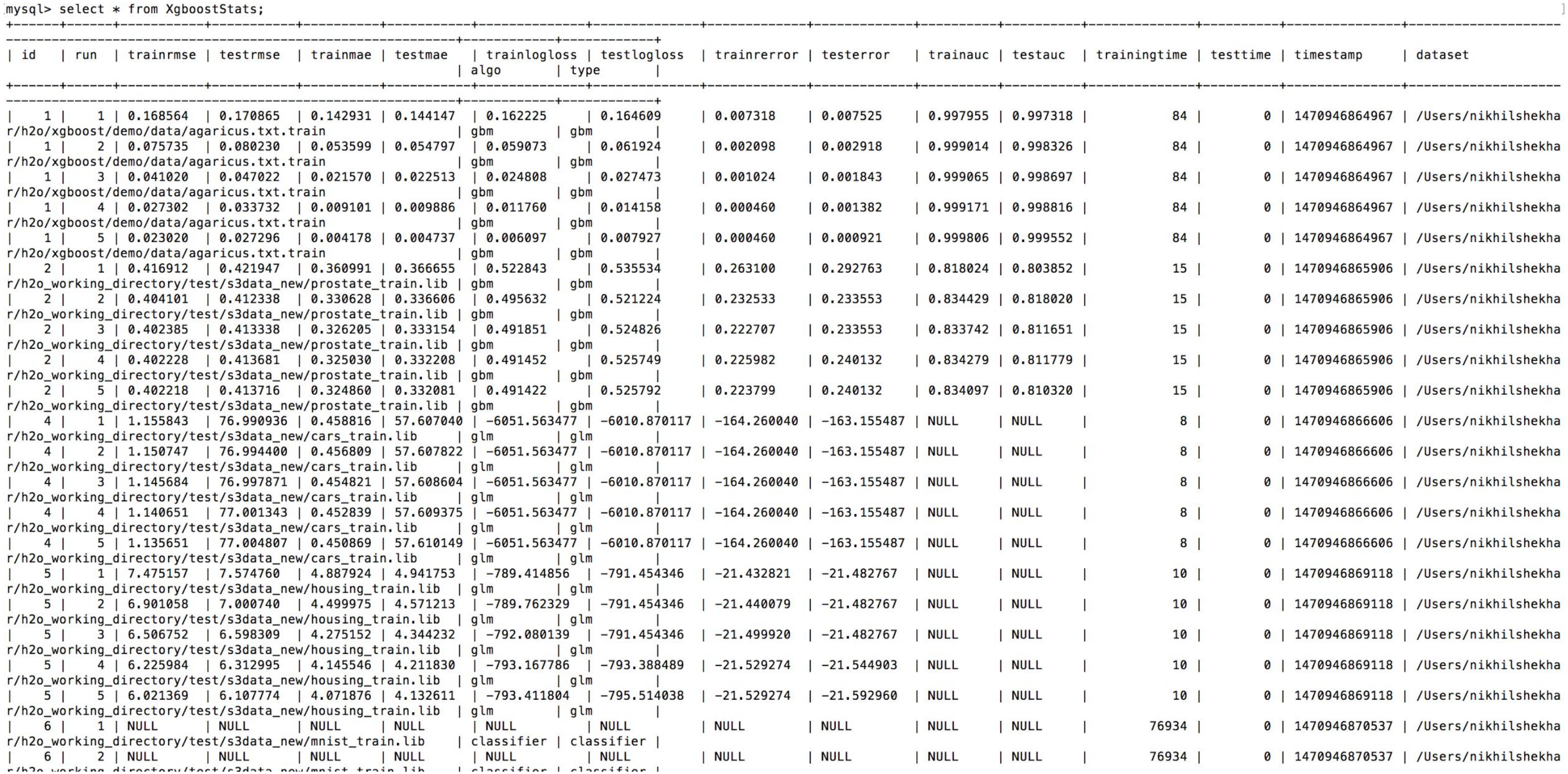
```
driver = com.mysql.jdbc.Driver
db = h2o
host = 172.16.2.178
user = root
password = 0xdata
table = SparkStats
port = 3306
```

MySQL metrics table



```
[mysql> select * from SparkStats;
+------+--------------------+---------------------+--------------------+-----------+----------+---------------+---------------------------------------------------------------------------------------------
----+-------------+--------------------+--------------------+--------------------+---------------------+--------------------+-------+--------------------+--------------------+----------------+
| id   | mae                | r2                  | rmse               | traintime | testtime | timestamp     | dataset
    | algo        | f1                 | precision          | recall             | weightedprecision   | weightedrecall     | wssse | auc                | aupr               | type           |
+------+--------------------+---------------------+--------------------+-----------+----------+---------------+---------------------------------------------------------------------------------------------
----+-------------+--------------------+--------------------+--------------------+---------------------+--------------------+-------+--------------------+--------------------+----------------+
|    3 | 48.75362221547697  | 0.06964307748519938 | 59.163828296530475 |      1682 |       49 | 1470935165863 | s3n://h2o-public-test-data/smalldata/testng/housing_train.csv
    | glm         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|    4 | 14.674685374001873 | 0.43884691396525277 | 19.04954819588868  |       588 |       31 | 1470935173114 | s3n://h2o-public-test-data/smalldata/testng/computer_train.csv
    | glm         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|    5 | NULL               | NULL                | NULL               |      3373 |       52 | 1470935180184 | s3n://h2o-public-test-data/smalldata/testng/prostate_train.csv
    | rfc         | 0.8778758503027191 | 0.8786885245901639 | 0.8786885245901639 | 0.8786682672468771  | 0.8786885245901639 | NULL  | NULL               | NULL               | Classification |
|    1 | 7.330103901417214  | 0.19146711437813813 | 10.061443158888475 |     21698 |       84 | 1470936439069 | s3n://h2o-public-test-data/bigdata/laptop/testng/milsongs-train.csv.gz
    | glm         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|    7 | NULL               | NULL                | NULL               |      9127 |       57 | 1470936605295 | s3n://h2o-public-test-data/smalldata/testng/airlines_train.csv
    | rfc         | 0.5665018106592777 | 0.6105810572867614 | 0.6105810572867614 | 0.6315342765569345  | 0.6105810572867614 | NULL  | NULL               | NULL               | Classification |
|    8 | NULL               | NULL                | NULL               |     63442 |     4528 | 1470936761732 | s3n://h2o-public-test-data/smalldata/testng/arcene_train.csv
    | rfc         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | 0.9683441558441559 | 0.9684659090909091 | Classification |
|   10 | NULL               | NULL                | NULL               |    394712 |      428 | 1470937476723 | s3n://h2o-public-test-data/bigdata/laptop/testng/mnist_train.csv
    | rfc         | 0.8999811145775389 | 0.90005            | 0.90005            | 0.9003406791716891  | 0.90005            | NULL  | NULL               | NULL               | Classification |
|   19 | 7.883419419339239  | 0.10409521176058234 | 10.591131554352721 |     19153 |       61 | 1470940714924 | s3n://h2o-public-test-data/bigdata/laptop/testng/milsongs-train.csv.gz
    | rfr         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|   21 | 45.67486579111672  | 0.16188355038473445 | 56.15438219928148  |       265 |       28 | 1470940901104 | s3n://h2o-public-test-data/smalldata/testng/housing_train.csv
    | rfr         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|   22 | 13.447535390062225 | 0.5621724523771108  | 16.826567953132106 |       219 |       27 | 1470940917420 | s3n://h2o-public-test-data/smalldata/testng/computer_train.csv
    | rfr         | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|   25 | 35.71034985861266  | 0.47059963360347745 | 44.62966120882825  |     15299 |       28 | 1470940921737 | s3n://h2o-public-test-data/smalldata/testng/housing_train.csv
    | gbmr        | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|   26 | 7.132483202426509  | 0.8174783845356943  | 10.864285104581114 |     16513 |       27 | 1470940960023 | s3n://h2o-public-test-data/smalldata/testng/computer_train.csv
    | gbmr        | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | NULL               | Regression     |
|   27 | NULL               | NULL                | NULL               |     32174 |       23 | 1470940984230 | s3n://h2o-public-test-data/smalldata/testng/prostate_train.csv
    | gbmc        | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | 0.8439449296025664 | 0.8561435414990957 | Classification |
|   28 | NULL               | NULL                | NULL               |      8420 |       25 | 1470941048495 | s3n://h2o-public-test-data/smalldata/testng/airlines_train.csv
    | gbmc        | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | 0.6669995513119844 | 0.7160024170166069 | Classification |
|   29 | NULL               | NULL                | NULL               |     55948 |     2082 | 1470941143105 | s3n://h2o-public-test-data/smalldata/testng/arcene_train.csv
    | gbmc        | NULL               | NULL               | NULL               | NULL                | NULL               | NULL  | 1.0                | 1.0                | Classification |
|   35 | NULL               | NULL                | NULL               |     31386 |       18 | 1470941275577 | s3n://h2o-public-test-data/smalldata/testng/airlines_train.csv
    | nn          | 0.3906508272825593 | 0.5471520412759511 | 0.5471520412759511 | 0.541831992988346   | 0.547152041275951  | NULL  | NULL               | NULL               | Classification |
|   37 | NULL               | NULL                | NULL               |      7973 |       71 | 1470942213971 | s3n://h2o-public-test-data/smalldata/testng/covtype_small_dense_multiclass_unbalanced_train.
csv | nn          | 0.7578330869011121 | 0.8335164835164836 | 0.8335164835164836 | 0.6947497282936844  | 0.8335164835164836 | NULL  | NULL               | NULL               | Classification |
|   38 | NULL               | NULL                | NULL               |     89785 |      218 | 1470942289940 | s3n://h2o-public-test-data/bigdata/laptop/testng/mnist_test.csv
    | nn          | 0.5454015569671474 | 0.5912             | 0.5912             | 0.6581570379588935  | 0.5912000000000001 | NULL  | NULL               | NULL               | Classification |
|   41 | NULL               | NULL                | NULL               |       359 |       34 | 1470942567576 | s3n://h2o-public-test-data/smalldata/testng/prostate_train.csv
    | glmlogistic | 0.7809659597219747 | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | 0.7293973458371062 | Classification |
|   42 | NULL               | NULL                | NULL               |       574 |       33 | 1470942584334 | s3n://h2o-public-test-data/smalldata/testng/airlines_train.csv
    | glmlogistic | 0.5                | NULL               | NULL               | NULL                | NULL               | NULL  | NULL               | 0.7265672986364194 | Classification |
|   34 | NULL               | NULL                | NULL               |      5944 |       56 | 1470944697246 | s3n://h2o-public-test-data/smalldata/testng/prostate_train.csv
    | nn          | 0.44202927882345   | 0.5934426229508196 | 0.5934426229508196 | 0.35217414673474867 | 0.5934426229508196 | NULL  | NULL               | NULL               | Classification |
+------+--------------------+---------------------+--------------------+-----------+----------+---------------+---------------------------------------------------------------------------------------------
----+-------------+--------------------+--------------------+--------------------+---------------------+--------------------+-------+--------------------+--------------------+----------------+
21 rows in set (0.03 sec)
```

contd..



```
mysql> select * from XgboostStats;
+------+------+-----------+-----------+----------+-----------+--------------+--------------+-------------+-------------+----------+----------+--------------+----------+---------------+--------------------
-----------------------------------------------------------+------------+------------+
| id   | run  | trainrmse | testrmse  | trainmae | testmae   | trainlogloss | testlogloss  | trainrerror | testerror   | trainauc | testauc  | trainingtime | testtime | timestamp     | dataset
                                                           | algo       | type       |
+------+------+-----------+-----------+----------+-----------+--------------+--------------+-------------+-------------+----------+----------+--------------+----------+---------------+--------------------
-----------------------------------------------------------+------------+------------+
|    1 |    1 | 0.168564  | 0.170865  | 0.142931 | 0.144147  | 0.162225     | 0.164609     | 0.007318    | 0.007525    | 0.997955 | 0.997318 |           84 |        0 | 1470946864967 | /Users/nikhilshekha
r/h2o/xgboost/demo/data/agaricus.txt.train                 | gbm        | gbm        |
|    1 |    2 | 0.075735  | 0.080230  | 0.053599 | 0.054797  | 0.059073     | 0.061924     | 0.002098    | 0.002918    | 0.999014 | 0.998326 |           84 |        0 | 1470946864967 | /Users/nikhilshekha
r/h2o/xgboost/demo/data/agaricus.txt.train                 | gbm        | gbm        |
|    1 |    3 | 0.041020  | 0.047022  | 0.021570 | 0.022513  | 0.024808     | 0.027473     | 0.001024    | 0.001843    | 0.999065 | 0.998697 |           84 |        0 | 1470946864967 | /Users/nikhilshekha
r/h2o/xgboost/demo/data/agaricus.txt.train                 | gbm        | gbm        |
|    1 |    4 | 0.027302  | 0.033732  | 0.009101 | 0.009886  | 0.011760     | 0.014158     | 0.000460    | 0.001382    | 0.999171 | 0.998816 |           84 |        0 | 1470946864967 | /Users/nikhilshekha
r/h2o/xgboost/demo/data/agaricus.txt.train                 | gbm        | gbm        |
|    1 |    5 | 0.023020  | 0.027296  | 0.004178 | 0.004737  | 0.006097     | 0.007927     | 0.000460    | 0.000921    | 0.999806 | 0.999552 |           84 |        0 | 1470946864967 | /Users/nikhilshekha
r/h2o/xgboost/demo/data/agaricus.txt.train                 | gbm        | gbm        |
|    2 |    1 | 0.416912  | 0.421947  | 0.360991 | 0.366655  | 0.522843     | 0.535534     | 0.263100    | 0.292763    | 0.818024 | 0.803852 |           15 |        0 | 1470946865906 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/prostate_train.lib | gbm        | gbm        |
|    2 |    2 | 0.404101  | 0.412338  | 0.330628 | 0.336606  | 0.495632     | 0.521224     | 0.232533    | 0.233553    | 0.834429 | 0.818020 |           15 |        0 | 1470946865906 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/prostate_train.lib | gbm        | gbm        |
|    2 |    3 | 0.402385  | 0.413338  | 0.326205 | 0.333154  | 0.491851     | 0.524826     | 0.222707    | 0.233553    | 0.833742 | 0.811651 |           15 |        0 | 1470946865906 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/prostate_train.lib | gbm        | gbm        |
|    2 |    4 | 0.402228  | 0.413681  | 0.325030 | 0.332208  | 0.491452     | 0.525749     | 0.225982    | 0.240132    | 0.834279 | 0.811779 |           15 |        0 | 1470946865906 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/prostate_train.lib | gbm        | gbm        |
|    2 |    5 | 0.402218  | 0.413716  | 0.324860 | 0.332081  | 0.491422     | 0.525792     | 0.223799    | 0.240132    | 0.834097 | 0.810320 |           15 |        0 | 1470946865906 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/prostate_train.lib | gbm        | gbm        |
|    4 |    1 | 1.155843  | 76.990936 | 0.458816 | 57.607040 | -6051.563477 | -6010.870117 | -164.260040 | -163.155487 | NULL     | NULL     |            8 |        0 | 1470946866606 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/cars_train.lib     | glm        | glm        |
|    4 |    2 | 1.150747  | 76.994400 | 0.456809 | 57.607822 | -6051.563477 | -6010.870117 | -164.260040 | -163.155487 | NULL     | NULL     |            8 |        0 | 1470946866606 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/cars_train.lib     | glm        | glm        |
|    4 |    3 | 1.145684  | 76.997871 | 0.454821 | 57.608604 | -6051.563477 | -6010.870117 | -164.260040 | -163.155487 | NULL     | NULL     |            8 |        0 | 1470946866606 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/cars_train.lib     | glm        | glm        |
|    4 |    4 | 1.140651  | 77.001343 | 0.452839 | 57.609375 | -6051.563477 | -6010.870117 | -164.260040 | -163.155487 | NULL     | NULL     |            8 |        0 | 1470946866606 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/cars_train.lib     | glm        | glm        |
|    4 |    5 | 1.135651  | 77.004807 | 0.450869 | 57.610149 | -6051.563477 | -6010.870117 | -164.260040 | -163.155487 | NULL     | NULL     |            8 |        0 | 1470946866606 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/cars_train.lib     | glm        | glm        |
|    5 |    1 | 7.475157  | 7.574760  | 4.887924 | 4.941753  | -789.414856  | -791.454346  | -21.432821  | -21.482767  | NULL     | NULL     |           10 |        0 | 1470946869118 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/housing_train.lib  | glm        | glm        |
|    5 |    2 | 6.901058  | 7.000740  | 4.499975 | 4.571213  | -789.762329  | -791.454346  | -21.440079  | -21.482767  | NULL     | NULL     |           10 |        0 | 1470946869118 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/housing_train.lib  | glm        | glm        |
|    5 |    3 | 6.506752  | 6.598309  | 4.275152 | 4.344232  | -792.080139  | -791.454346  | -21.499920  | -21.482767  | NULL     | NULL     |           10 |        0 | 1470946869118 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/housing_train.lib  | glm        | glm        |
|    5 |    4 | 6.225984  | 6.312995  | 4.145546 | 4.211830  | -793.167786  | -793.388489  | -21.529274  | -21.544903  | NULL     | NULL     |           10 |        0 | 1470946869118 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/housing_train.lib  | glm        | glm        |
|    5 |    5 | 6.021369  | 6.107774  | 4.071876 | 4.132611  | -793.411804  | -795.514038  | -21.529274  | -21.592960  | NULL     | NULL     |           10 |        0 | 1470946869118 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/housing_train.lib  | glm        | glm        |
|    6 |    1 | NULL      | NULL      | NULL     | NULL      | NULL         | NULL         | NULL        | NULL        | NULL     | NULL     |        76934 |        0 | 1470946870537 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/mnist_train.lib    | classifier | classifier |
|    6 |    2 | NULL      | NULL      | NULL     | NULL      | NULL         | NULL         | NULL        | NULL        | NULL     | NULL     |        76934 |        0 | 1470946870537 | /Users/nikhilshekha
r/h2o_working_directory/test/s3data_new/mnist_train.lib    | classifier | classifier |
```

Learning outcomes

- Deep exposure to h2o, Spark ML and xgboost machine learning libraries.
- Stumbled upon multiple unresolved JIRA’s for spark and had to look for alternative solution to solve the problem at hand.
- Exposure to multiple datasets stored in different formats.
- Dealing with missing labels,fields in Spark-ml,xgboost,scikit.
- Reading data from s3 buckets programmatically using Java and Python.

Work in progress

- Run Spark in multi-node environment (Our own machines to start with, but EC2 if we ever decide to publish this framework)
- Run xgboost in a distributed environment with spark (Code is available in “ml-benchmark” repo, but needs testing)
- Tune the hyper-parameters for optimal performance of algorithms
- Add more test cases to each of the libraries (Utilize more “Kaggle like” datasets. Right now we have whatever is available in S3 for H2O)
- Run algorithms on large datasets (Besides the airlines dataset…)
- Integrating h2o accuracy suite - First, need to publish the test classes to maven central repository.
    - Work around is to paste in necessary classes into the “ml-benchmark” repo for H2O
- Benchmark.ai ? -> A combination of “ml-benchmark” and “db-benchmark”
- Other frameworks to benchmark?

Acknowledgements

- Navdeep Gill
- Bill Gallmeister
- Anmol Bal
- Nancy Jordan

Thank You