**Use-Cases Document**

**Case 01:** ***Suggest latest 10 datasets used by other user's based on an industry domain (Tag given to a dataset)***

**INP**: *Name of the Tag: Finance/Healthcare/Retail/Tourism etc.*

**OUT**: *Name of the Dataset, Data-size, no. of Rows, columns etc.*

**Operation**: Get the tag-Id corresponding to the tag name from the Tag\_Map table. Find the matching details pertaining to Name, Data-size, no. of rows and columns etc. from the Dataset\_Metadata table corresponding to the tag-Id and display the top 10 results filtered on the basis of descending Starttime (EPOCH Time-Format starting 1970)

**Case 02:** ***Suggest latest 10 datasets based on Model\_Species(Regression/Classification) and the model-type used ( Linear/Logistic/GLM etc.)***

**INP:** *Model\_Type\_Name, Model\_Species*

**OUT:** *Name of the Dataset, Data-size, no. of Rows, columns etc.*

**Operation:** Find all the run-Ids matching the input Model\_species from the leaderboard\_Metadata table. Find the corresponding Model-Ids for each run-id from the Model\_map table. Filter these model-ids based on matching model\_ Ids for which the model\_type Id matches the input (in the ID\_Map table). Display the Dataset names by joining the Model\_Run, Data\_Map and Dataset\_Metadata table

**Case 03: *List the max repeated model for a particular run***.  
**INP:** *run\_id*  
**OUT:** count *of the max repeated Model*  
**Operation:** Find the count of model types ids corresponding to the input Run-Id and select the topmost value

**Case 04:** ***Details of the conditions under which model was tested.***

**INP:** *Model\_ID*

**OUT:** *Run\_ID, Duration of run, Model\_Name,Presence in StackedEnsemble BestOfFamily*

**Operation:** Get the matching Run\_Id from the ModelRun table for the input Model\_ID. Get the other details related to this model\_ID from the Leaderboard\_Metadata such as Start and End time of run. Append a calculated field to find duration (Start-End time). Get the model\_name and presence in StackedEnsemble from the ID\_Map table for this Model\_Id

**Case 05:** ***Get average no. of models generated for a particular run\_time and datasets with the same tag having almost similar data size in terms of no. of rows and columns.***

**INP:** *Tag\_Name, Run\_time, No. of Rows, Columns in user's dataset*

**OUT:** *Average no. of models generated*

**Operation:** From the tags table get the Ids corresponding to the input tag name. Get the Dataset Ids for these tag\_Ids from the Tag\_Map table. Select only those dataset Ids for which the  the no. of rows and columns for all these datasets and the input run-time from the leaderboard\_Metadata table ( input values of rows and column lying within a particular range, lets say +/- 10). Group the data as per Run\_Id and calculate the total count of model of a particular type/total count of models in the run.

**Case 06:** ***For the datasets from a particular industry (tag), size (observations) and type of analysis to be done (Regression/Classification/Time-series forecasting), which is the best model?***

**INP:** *Tag\_Name, No. of rows, columns, model\_species*

**OUT:** *Best Model's name and performance metric values (RMSE for Regression/ AUC for Classification)*

**Operation:** Get the dataset Ids for the Tag\_Name from the Tag\_Map table. Select only those dataset Ids for which the no. of rows and columns for all these datasets and the input run-time from the leaderboard\_Metadata table (input values of rows and column lying within a particular range, lets say +/- 10) and model\_species matches the input. From the Data\_Map table find the Run\_Id for each of these Dataset\_Ids. Get all the Model\_IDs for these Run\_Ids from the Model\_Run table and filter only the model\_Ids associated with the input model type from the ID\_Map table. Compare the performance values for all these model Ids

**Case 07: *For each domain (Tag), list the best datasets (having least no. of null values) at the top.***  
**INP:** *Tag*  
**OUT:** *Dataset\_Id, sum of null values column*  
**Operation:** Determine all the Dataset\_Ids for the given input Tag\_Id and find the details of the corresponding dataset\_ids from the Dataset\_Variabe\_Details table. Group by the Dataset Id and calculate the sum of null values column and Order by Desc.

**Case 08:** ***For regression model, enlist the no. of linear, logistic, GBM and GLM models generated for a particular runtime.***  
**INP:** *Runtime*  
**OUT:** *No. of models in each Model\_Type*  
**Operation:** Get run\_Ids of all models for the input runtime. Filter the run\_Ids matching with {the Model Ids for the particular model\_type Id and model\_species(Regression in this case)} and calculate the count of model\_IDs grouped by each model\_type

**Case 09:** ***Which is the most commonly applied best model-type across run-times for a given dataset?***

**INP:** *Model-type\_ID, Dataset\_ID*

**OUT:** *Dataset Name, count(Model\_Type\_Id)*

**Operation:** Filter Run\_Ids from Data\_Map table for the given dataset\_Id. Now find all the Model\_Ids for each of the Run\_Ids from Model\_Run table. Find the best\_model Id based on the logic applied in Case 05. Find the corresponding Model\_Type\_Ids and calculate the count of distinct model\_type\_Ids.

**Case 10: *Display the hyperparameter values of all the models (of same type) for a particular dataset & run-time. (Applicable for analysing cases where 2-3 GBM models are created during a 300/500 sec runtime).***

**INP:** *Dataset\_Id, Run-time, Model\_Type\_ID*

**OUT:** *Hyperparameter ID, values*

**Operation:** Filter the Run\_Ids in the Data\_Map table based on the (Run-Ids matching input run-time and model\_species from leaderboard\_Metadat table and Dataset Id of input dataset). For each of those run\_Ids find the model-ids from the Model\_Run table and Get the hyperparam\_Id and value from the hyperparameter\_values table.

**Case 11:** ***Determine no. of times the dataset has been used for the study.***  
**INP:** *None*  
**OUT:** *Count of times when it was used*  
**Operation:** For all the dataset\_Ids find all the Run\_Ids from the Data\_Map table. Count the no. of Run\_Ids grouped by each dataset\_Id

**Case 12:** ***Display the hyperparameter values of all the models (of same type) for a particular dataset & across all run-times.***

**INP:** *Dataset\_Id, Model\_Type\_ID*

**OUT:** *Hyperparameter ID, values*

**Operation:** Filter the Run\_Ids in the Data\_Map table based on the (Run-Ids matching model\_species from leaderboard\_Metadata table and Dataset Id of input dataset). For each of those run\_Ids find the model-ids from the Model\_Run table and Get the hyperparam\_Id and value from the hyperparameter\_values table.

For each model\_Id also display the RMSE, MAE,MAPE etc.. by joining with leaderboard table

**Case 13:** ***Display the default values of the hyperparameters for a particular model\_Type.***

**INP:** *Model\_Type\_Name*

**OUT:** *List of Hyperparameter names and their values*

**Operation:** Get the model\_Type\_Id for the model\_type name from the Model\_Map table and join with hyperparameter\_Default table to display the required values.

**Case 14:** ***Calculate change in hyperparameter values (denote increase/decrease using if/case statement) for a particular Model\_Type and run-time.***

**INP:** *Model\_Type\_Id, Run-time*

**OUT:**  *Model\_Name, change in value*

**Operation:** For a particular model-type and run-time get all the Run-Ids from Leaderboard\_Metadata table. For each run id get the model-Ids and join with Hyerparameter\_value and hyperparameter\_default value to calculate increase/decrease in the value compared to default.

**Case 15: *What is the best model type for a tag?***  
**INP:** *Model\_Type\_Id, Tag*  
**OUT:** *Tag, Count(Model\_Type\_Id)*  
**Operation:** From the Tag\_Map table find all the dataset\_Ids corresponding to the input tag name.Filter Run\_Ids from Data\_Map table for these dataset\_Ids. Now find all the Model\_Ids for each of the Run\_Ids from Model\_Run table. Find the best\_model Id based on the logic applied in Case 06. Find the corresponding Model\_Type\_Ids and calculate the count of distinct model\_type\_Ids