**UT Arlington / IBM**

**Deep Learning Workshop**

****

**H2O Driverless AI**

**Arlington**

**October 22, 2019**

**James Nash**

**jjnash@us.ibm.com**



|  |  |
| --- | --- |
| **Team 01** | **http**://10.31.204.151:**12301** |
| **Team 02** | **http**://10.31.204.151:**12302** |
| **Team 03** | **http**://10.31.204.151:**12303** |
| **Team 04** | **http**://10.31.204.151:**12304** |
| **Team 05** | **http**://10.31.204.153:**12305** |
| **Team 06** | **http**://10.31.204.153:**12306** |
| **Team 07** | **http**://10.31.204.153:**12307** |
| **Team 08** | **http**://10.31.204.153:**12308** |
| **Team 09** | **http**://10.31.204.155:**12309** |
| **Team 10** | **http**://10.31.204.155:**12310** |
| **Team 11** | **http**://10.31.204.155:**12311** |
| **Team 12** | **http**://10.31.204.155:**12312** |
| **Team 13** | **http**://10.31.204.157:**12313** |
| **Team 14** | **http**://10.31.204.157:**12314** |
| **Team 15** | **http**://10.31.204.157:**12315** |
| **Team 16** | **http**://10.31.204.157:**12316** |
| **Team 17** | **http**://10.31.204.159:**12317** |
| **Team 18** | **http**://10.31.204.159:**12318** |
| **Team 19** | **http**://10.31.204.159:**12319** |
| **Team 20** | **http**://10.31.204.159:**12320** |

H2O Driverless AI is an artificial intelligence platform for automatic machine learning (ML). It automates many of the most difficult machine learning workflows such as feature engineering and the selection, validation, tuning and deployment of ML models.

By using end-to-end automation, Driverless AI achieves the high predictive accuracies comparable to expert data scientists. It also offers built-in visualization and interpretability which are critical to understanding, explaining transparent models. Accurate models are exportable to Python and Java-based modules to provide scoring services.

While Driverless AI supports a variety of hardware, it has been optimized to take advantage of graphical processing units (GPUs) supported by IBM’s POWER9 AC922 server.

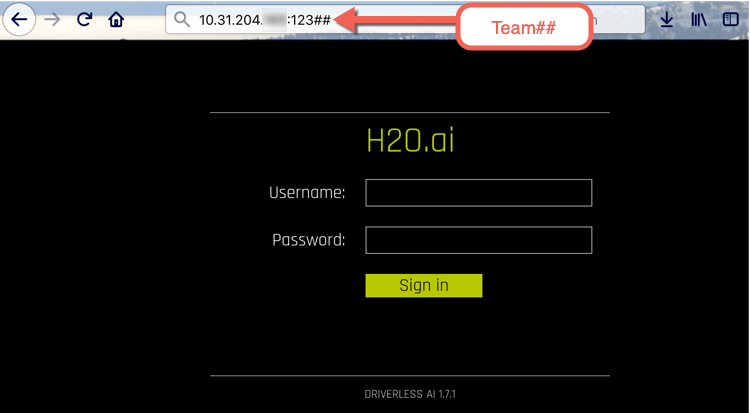
Why H2O Driverless AI?

H2O Driverless AI employs the techniques of expert data scientists in an easy to use application that helps scale your data science efforts. Driverless AI empowers data scientists to work on projects faster using automation and state-of-the-art  
computing power from GPUs to accomplish tasks in minutes that used to take months.

With Driverless AI, everyone including expert and junior data scientists, domain scientists, and data engineers can develop trusted machine learning models. This next-generation automatic machine learning platform delivers unique and advanced functionality for data visualization, feature engineering, model interpretability and low-latency deployment.

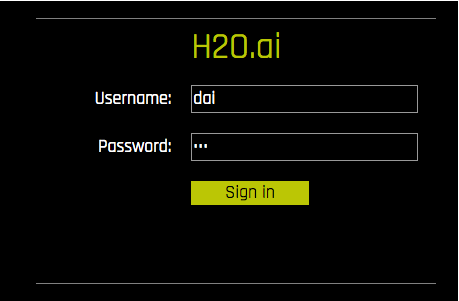


1. Browse to H2O DAI Login Page (**HTTP** not HTTPS)



1. Username: dai

Password: dai



1. We will not go through all of the top menu choices in this lab, but will list their meanings here:



PROJECTS – Managing and grouping Datasets together with Experiments in one place

DATASETS – View the source datasets available and datasets that have been imported

AUTOVIZ – Navigate to the visualizations list page

EXPERIMENTS – Datasets we have used to generate predictions. We can have multiple experiments on a single dataset

DIAGNOSTICS – Ability to diagnose a model against a different dataset

MLI – Machine Learning Interpretability – Explains the model results in human readable format

DEPLOYMENTS –

RESOURCES –

MESSAGES –

LOGOUT –

HELP – Self descriptive

PY\_CLIENT – This will download a binary file associated with running a generated Python Scoring Pipeline for an experiment. We will discuss this later.

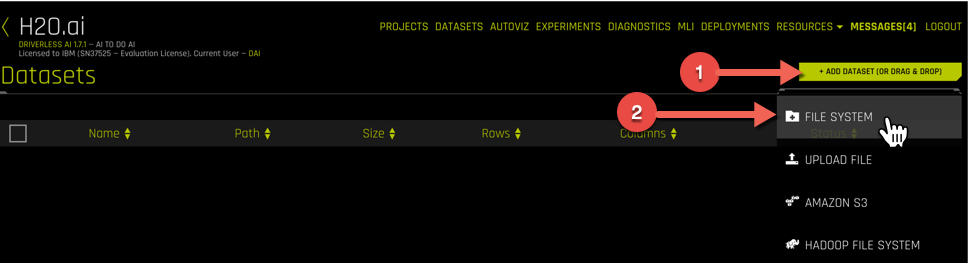
MOJO2-RUNTIME – This will download a binary file associated with running a MOJO Scoring Pipeline. MOJO is an acronym for a Java object model. We will discuss this later.

MESSAGES – Messages associated with the product. These are general and not specific to experiments (which have logs that we will see later).

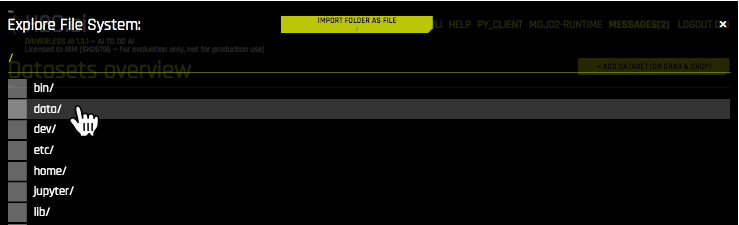
LOGOUT DAI

ADD DATASET – Allows the importation of a data set for processing

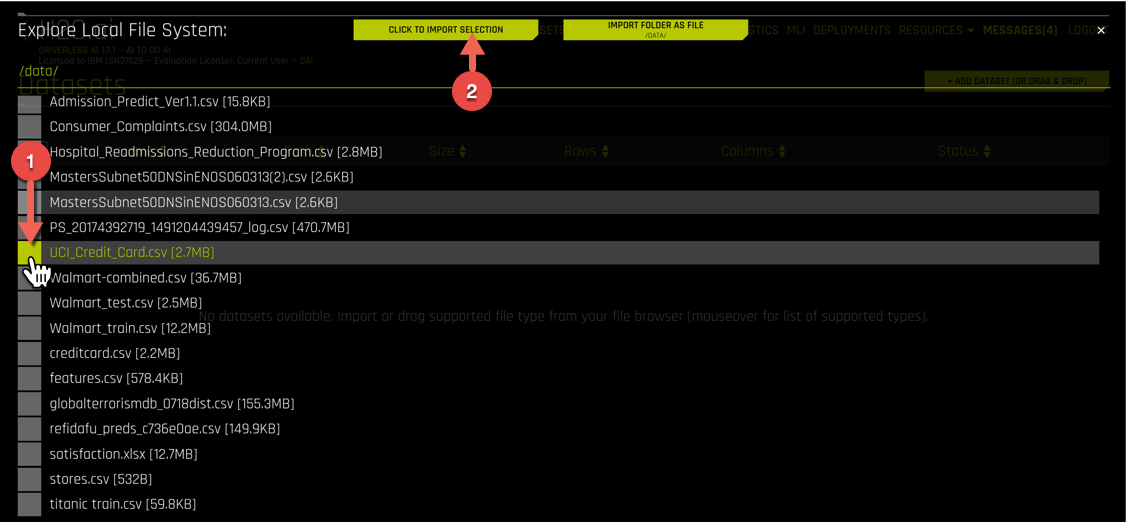
1. Click “+Add Dataset” and Click “File System”



1. Open the “data/” directory.

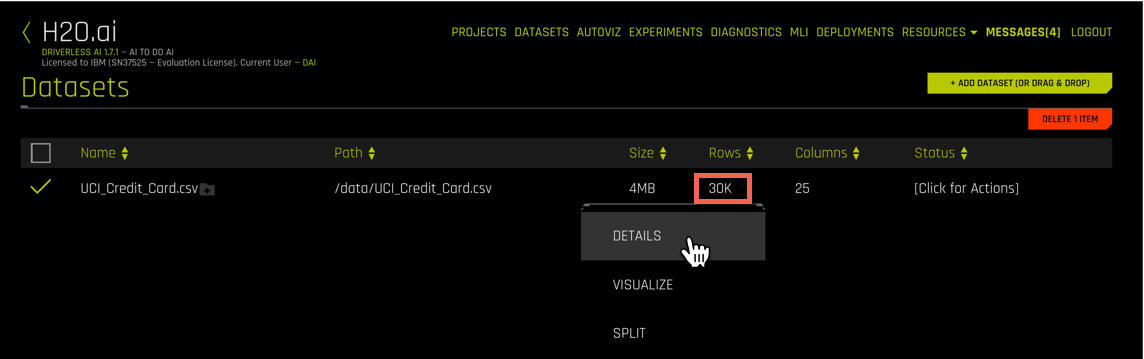


1. Select the “UCI\_Credit\_Card.csv” file and click “Import File”



Driverless AI uses tabular data, typically found in existing Comma-Separated Value (CSV) typically generated by spreadsheets.

1. Note the columns listed for this experiment show 30000 Rows and 25 Columns of data. Click on the UCI\_Credit\_Card.csv row and select “Details”



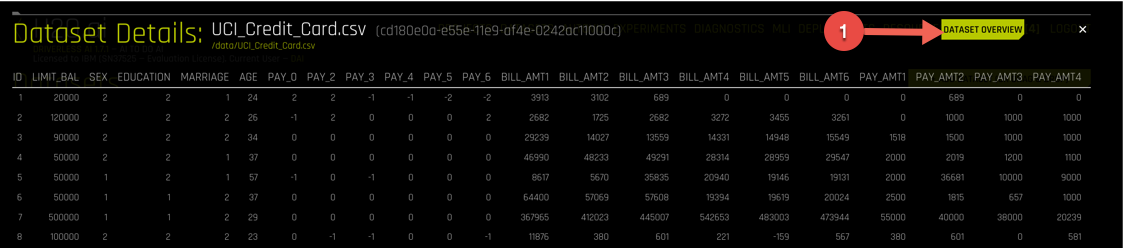
1. Explore the contents of the contents of the CSV file.

Click “Dataset Rows” to see a view of the row entries.



Use the bottom scroller to move left and right and view the data. The column titles represent the names of the columnar data within the dataset. You can also hover over the histogram graphs and a pop-up will show more details about each column.

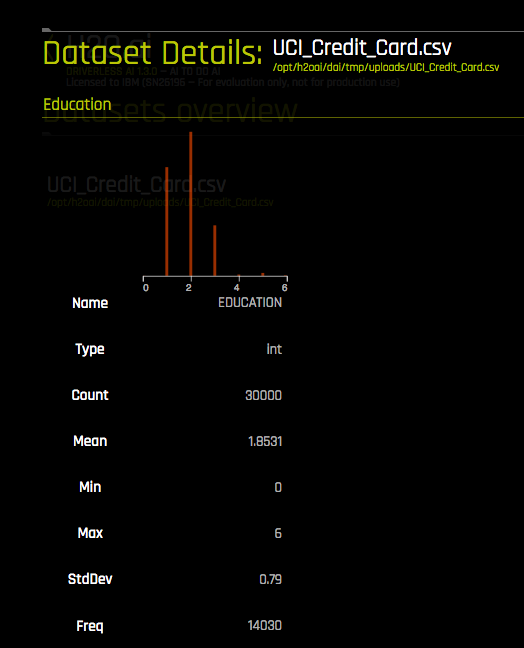
1. Click the “Dataset Overview” button to return.



The overview provides a formatted view of the contents of your CSV dataset.

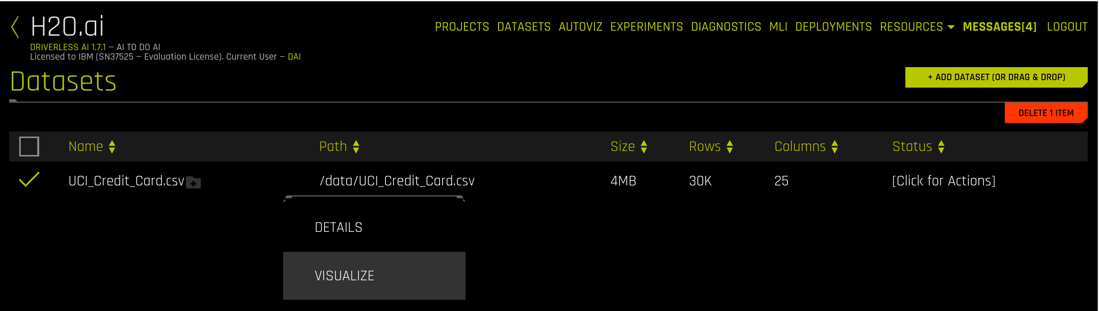
1. To filter on different columns to change the view (like Pay), simply enter the column of interest below “Dataset Details”. 

Example of filtering for “Education”. This field like many in data science products, the numeric values here map to specific meanings like: high school, bachelors, masters or graduate degree.



After you are done, close this (by hitting the ‘x’) view to return to “Datasets Overview”

1. Click the UCI\_Credit\_Card.csv row and select “Visualize”

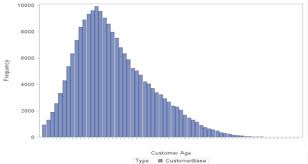
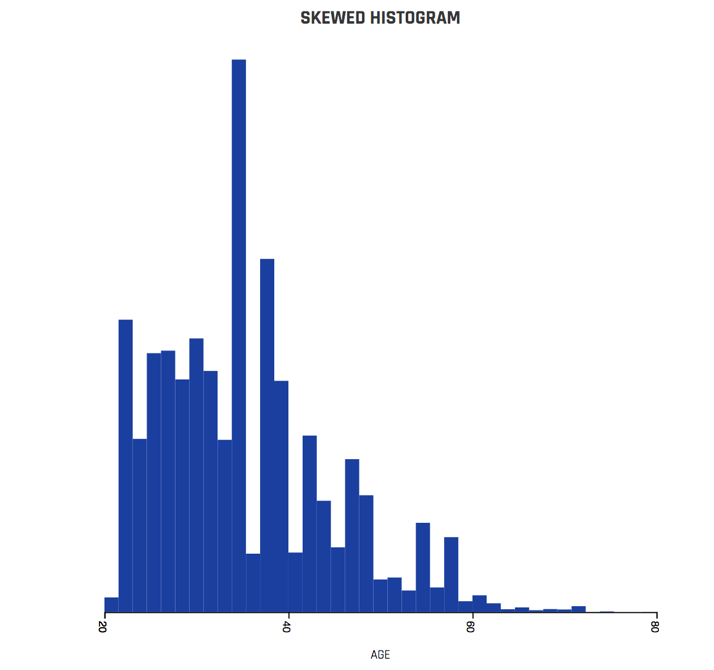


1. Explore the different visualization options available to you. Do you see any interesting relationships that jump out?



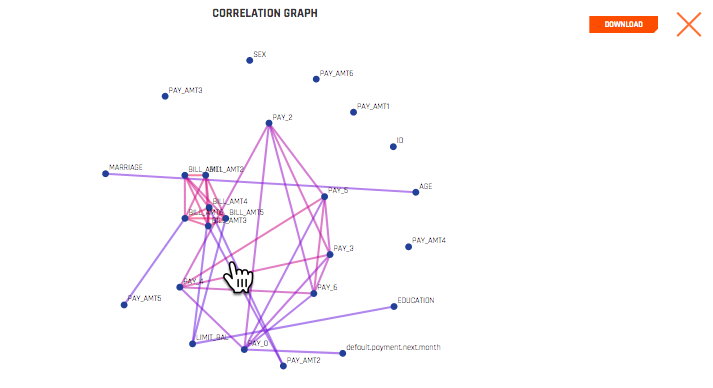
You can double-click on any graph, which will pop-up and also have choices for downloading or help. The help selection will give a general explanation of the plot type and how it is used by data engineers and scientists. Most users will be familiar with histograms and what we might be looking for there are spike or outlier values, which may be meaningful or perhaps a bad data value is present. For example, if we had a human age in years and one entry showed 255, it would point us back to fixing the dataset for an error.

Here is an example of the AGE values in the dataset as viewed by a Skewed Histogram on the left. We don’t have someone with an unusual age (no 2-year olds with credit histories), but what we do see is that the data is somewhat spikey with respect to ages. In a normal population of consumers, we might expect something more like the distribution on the right. A data analyst would have to think about whether our dataset represents the population accurately or if people just tend to default at certain ages.

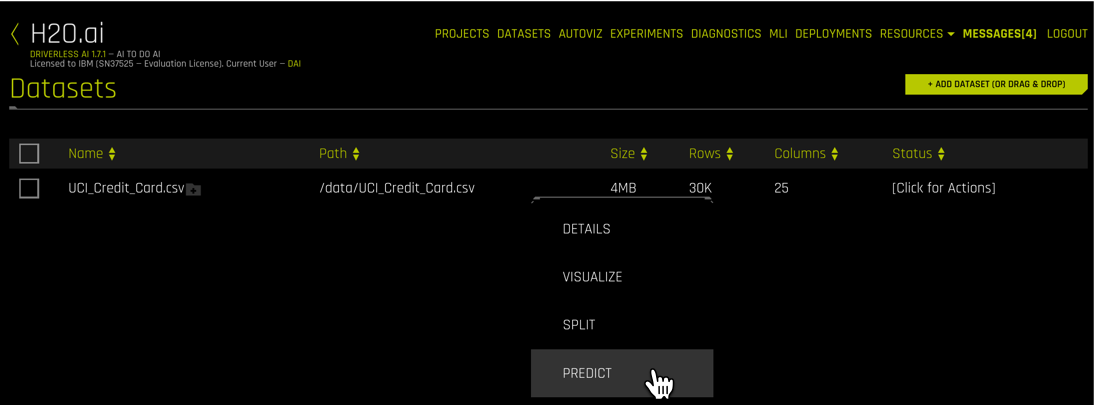


H2O Age Histogram Normal population distribution

Some generated graphs are interactive, like the Correlation Graph below, where you can grab one of fields (as dots) and move them around.



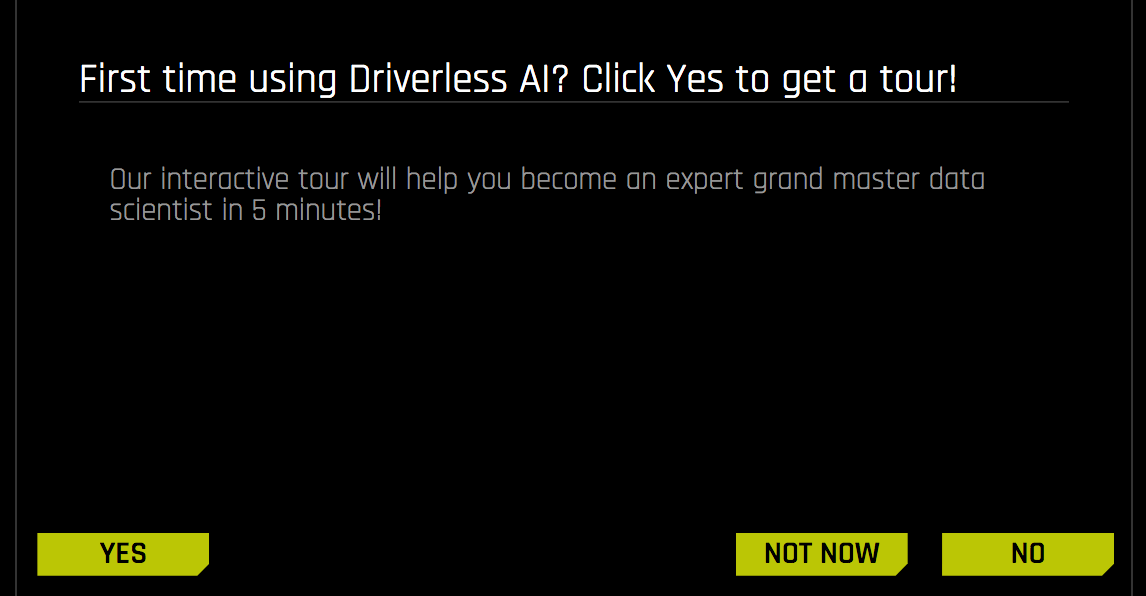
1. Click “H2O.ai” to return to the Datasets overview page. 
2. Click the UCI\_Credit\_Card.csv row and select “Predict”



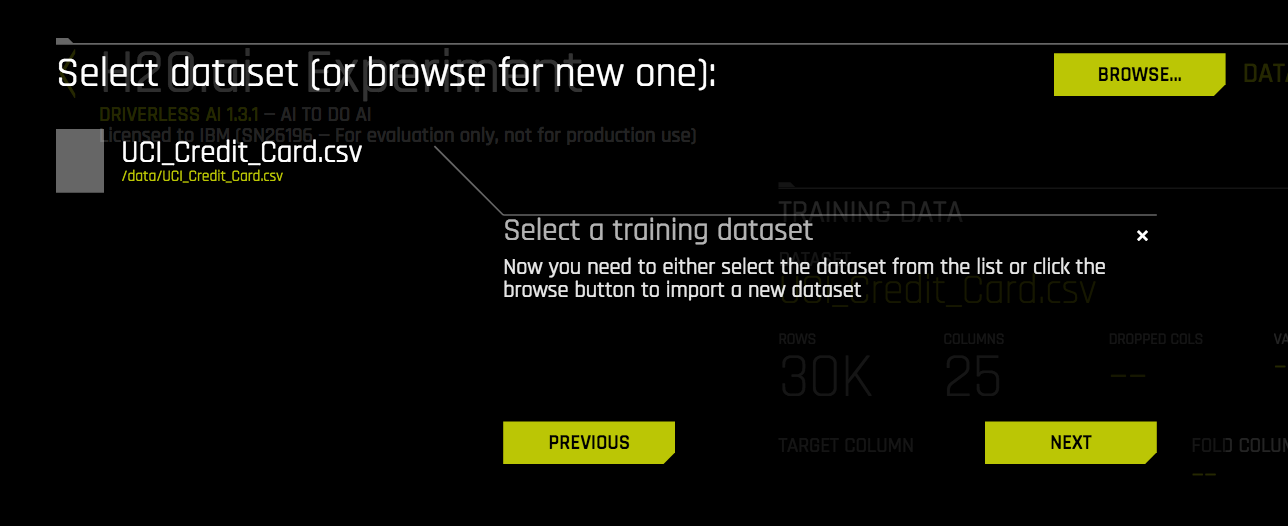
1. Walk through a brief 5-minute tour of the Prediction interface.

If your tour is not available, you can watch/download a replay here:

<http://10.31.193.184/PowerAI/h2o-tour.mp4>

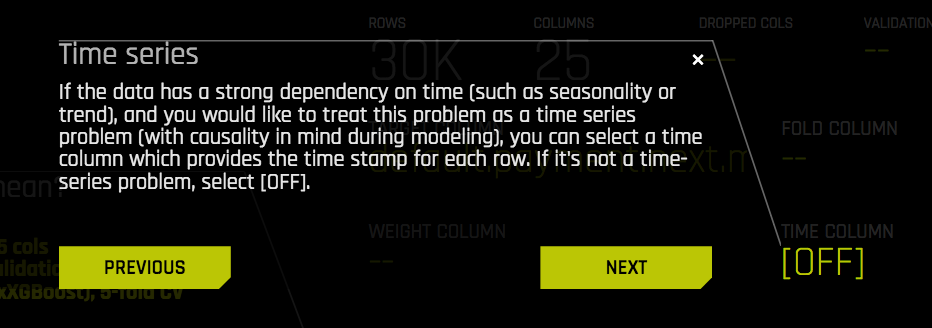


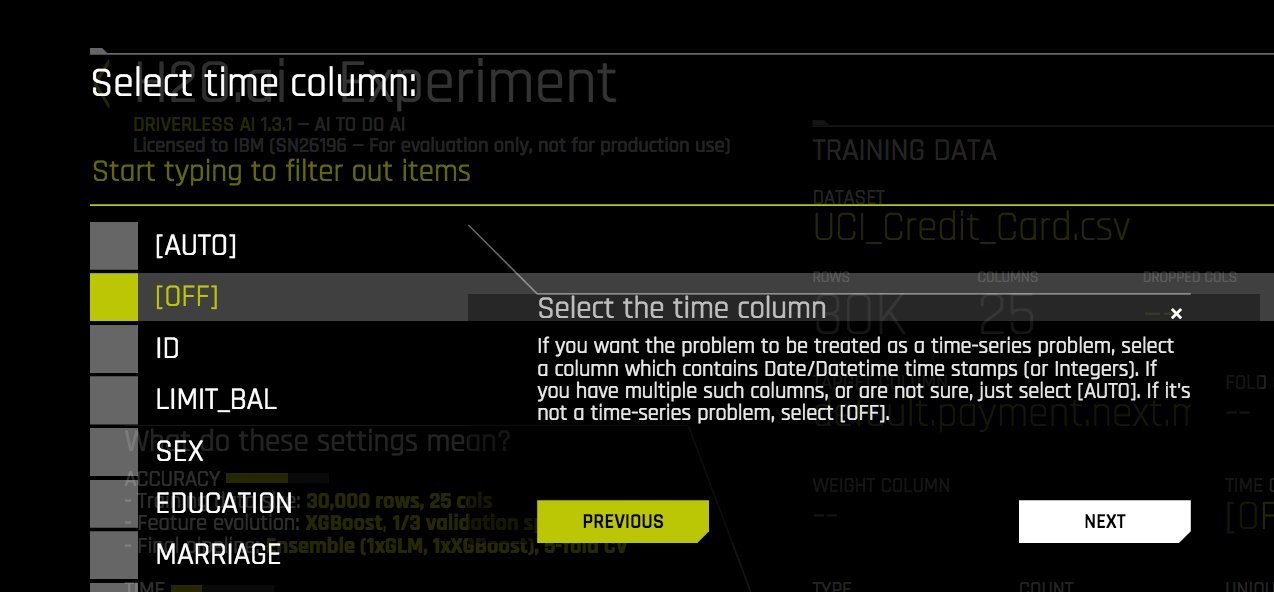
1. First choice in the tour is to show the selection of a dataset, select NEXT when you want to proceed to the next tour menu



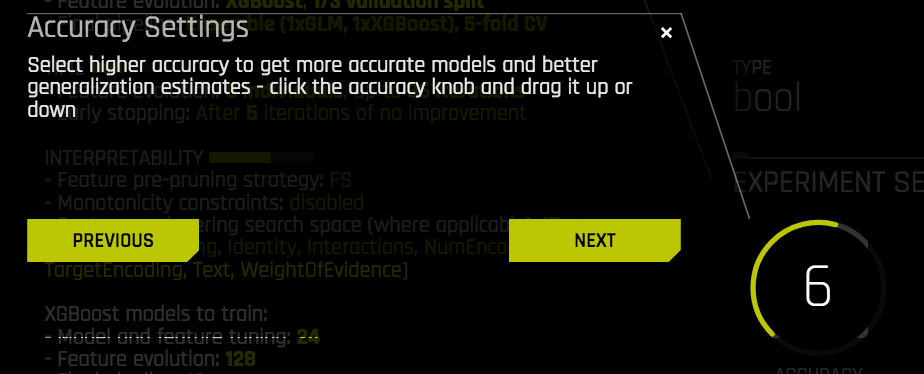
1. The next choice is to select a target column to be **predict**. We would like to predict if the cardholder is going to default on next month’s payment. You may need to scroll down to see this option.

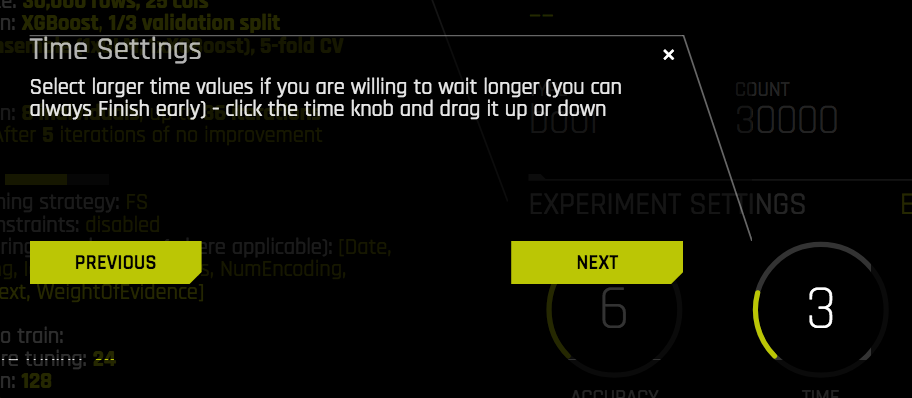


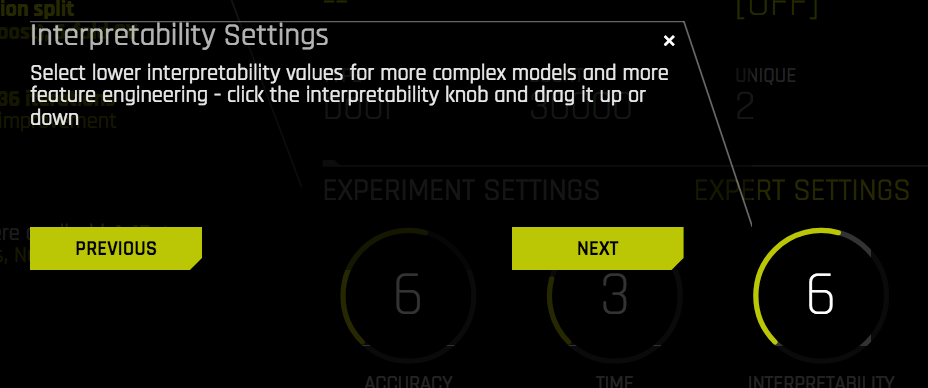
1. The next choice covers if the data has a dependency upon time, and the selection of the column to be treated as a time-series



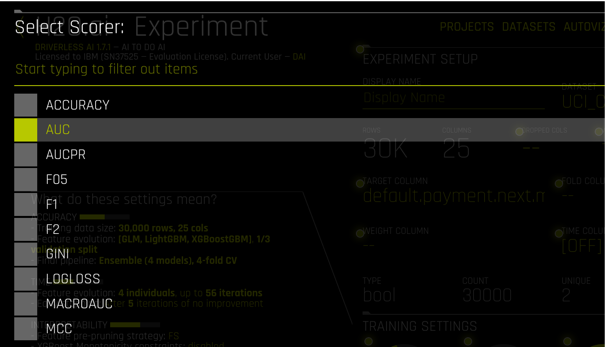
1. In the center of the interface, there are three primary “dials” for tuning. The first is an accuracy setting with a range from 1 to 10. The higher the number, the more accurate the model will be, but will also take more time to generate a model. Lower value runs faster but potentially less accuracte.

l

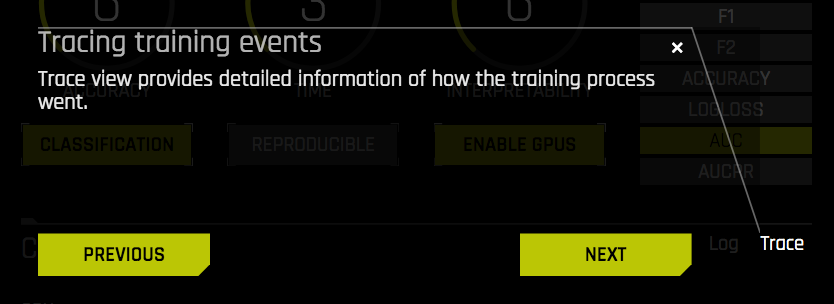
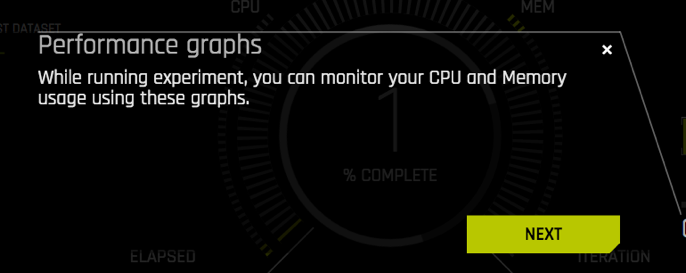
1. The second dial is the Time setting. 
2. The third dial is the Interpretability setting. The higher the number, the more features will be used to generate the model. Lower values will have more complex models and more feature engineering.

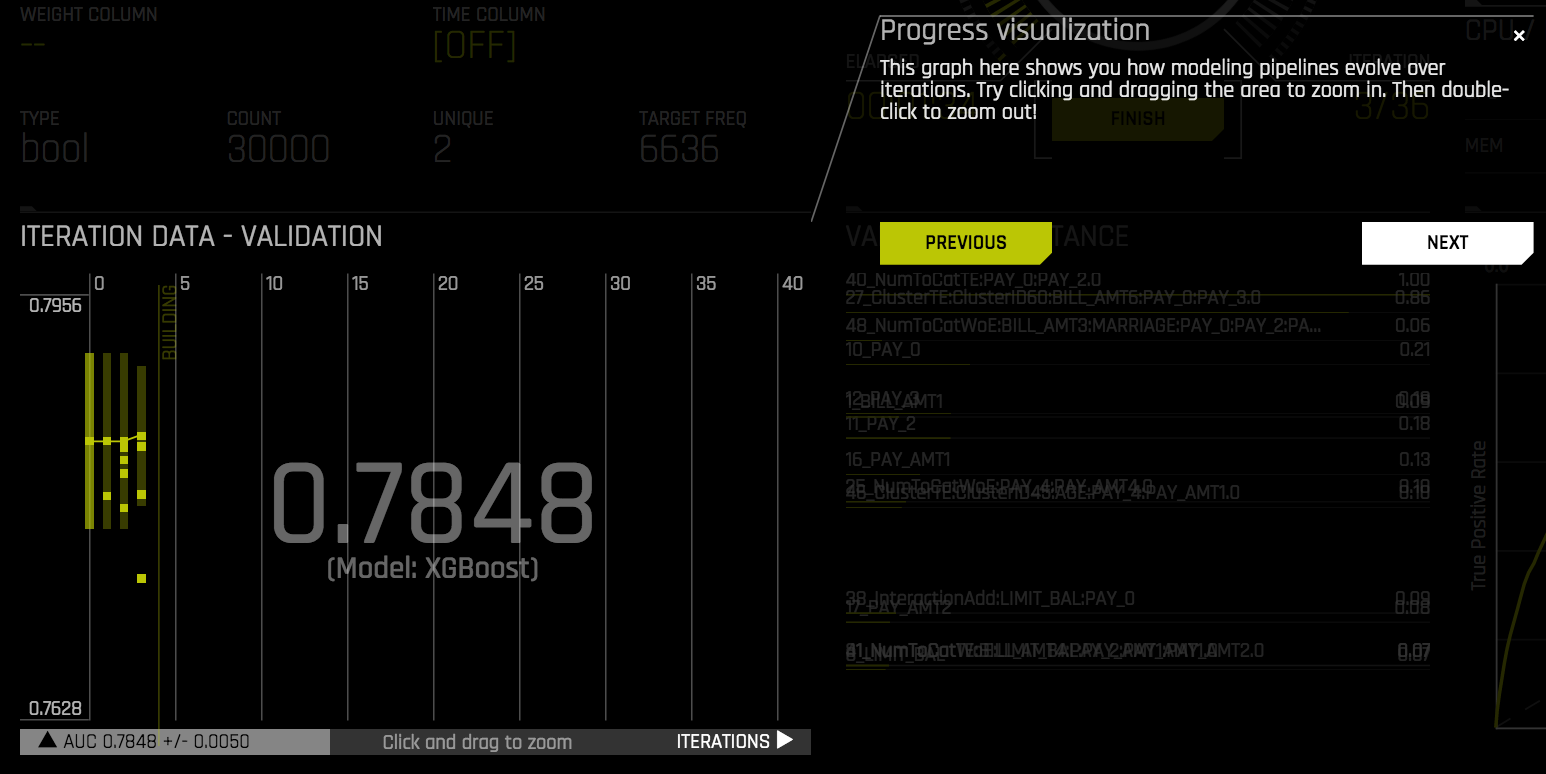


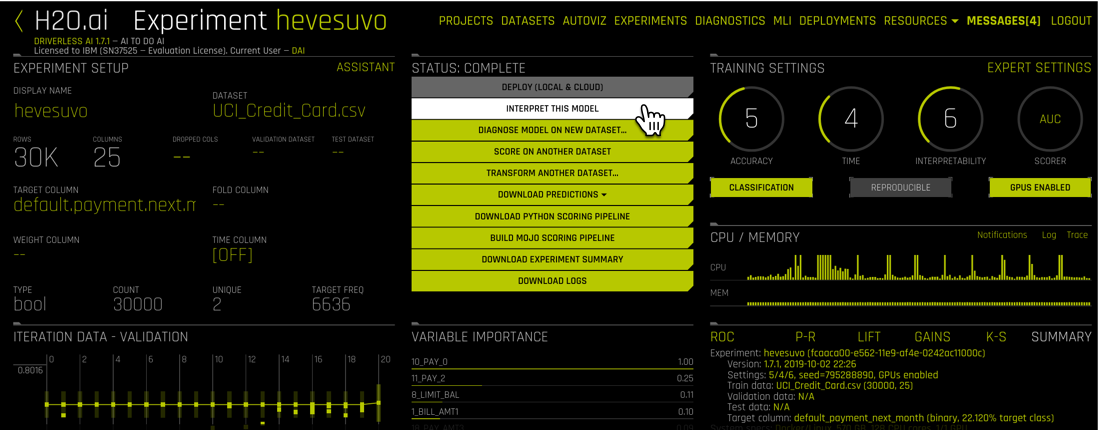
1. The fourth dial “SCORER” dial allows you to change the scoring selection. This is also automated depending upon analysis of the dataset and other selections, but advanced users can choose other options. AUC was selected by default.



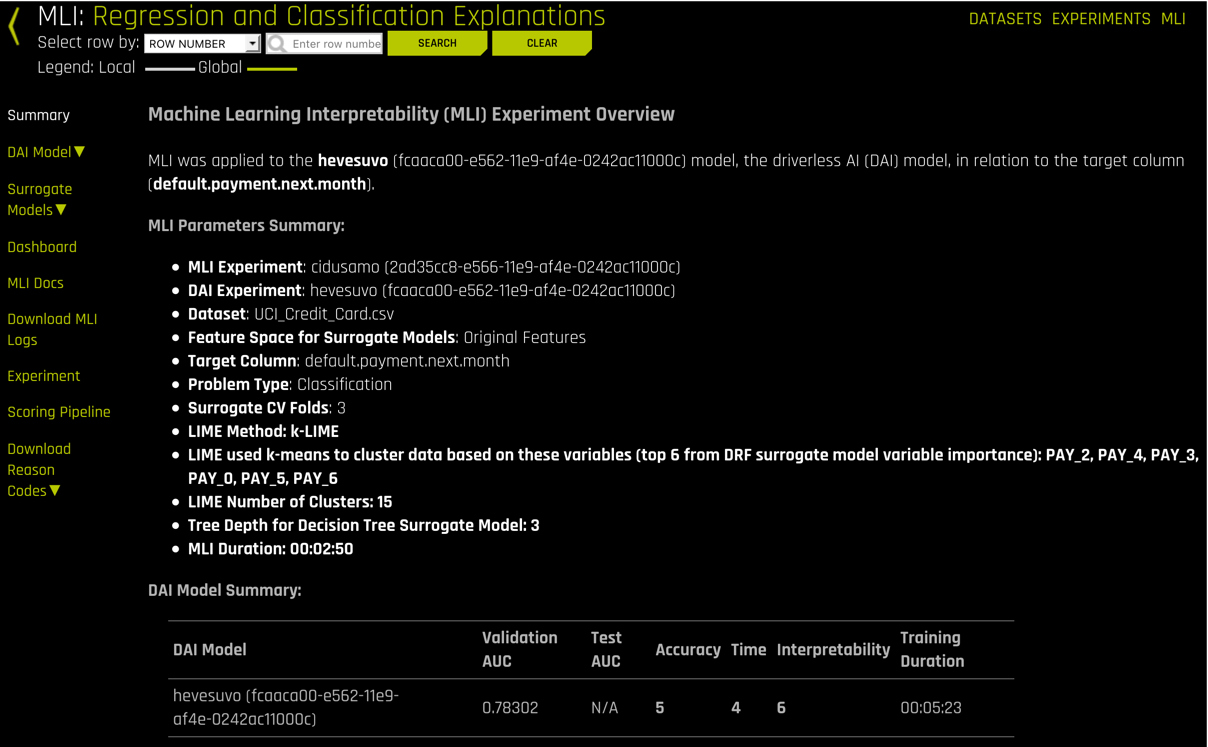
1. You can now select the Launch Experiment and it will run in the background as you continue through the tour
2. While an experiment is running, Performance Graphs and Tracing outputs are updated on the main interface



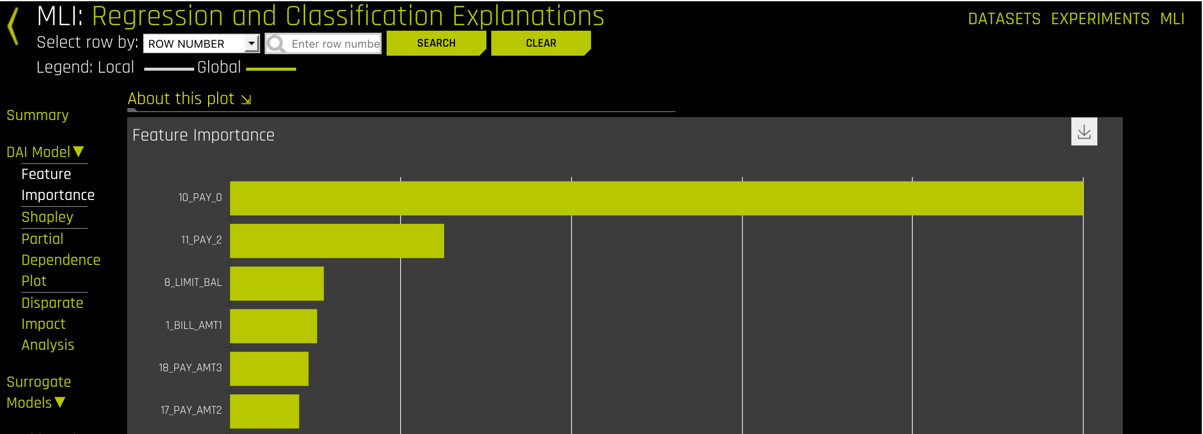
1. As the experiment runs, a progress graph shows the results as the model applies features and iterates through them
2. Follow steps outlined in the tour to run an experiment on your dataset.



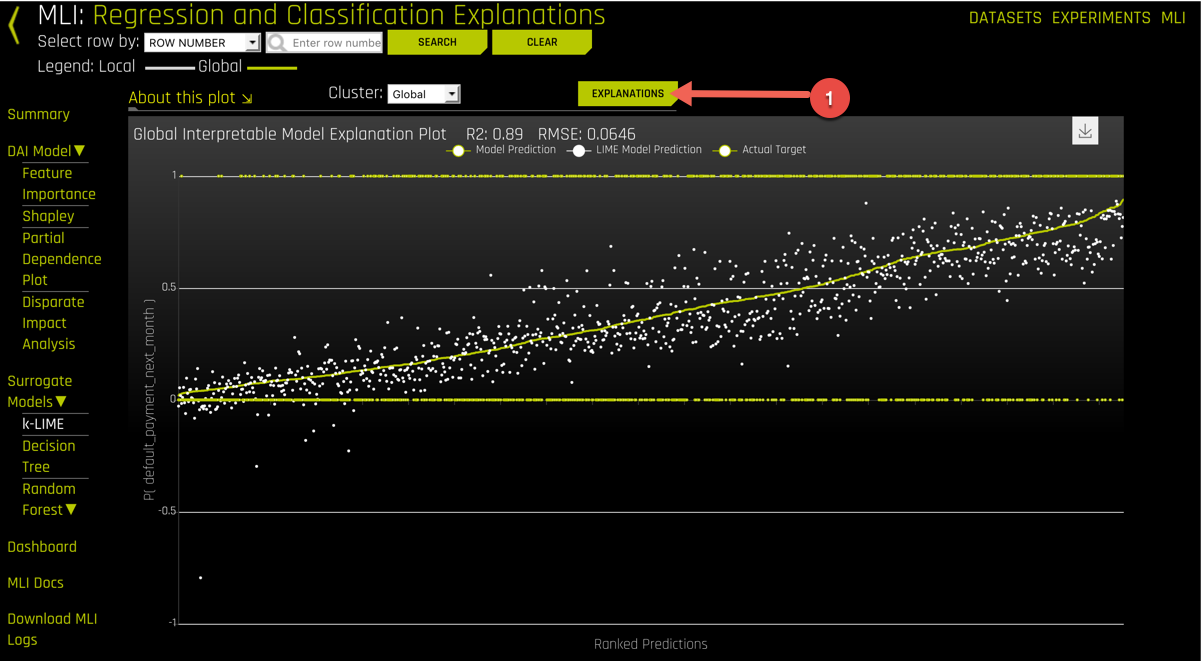
1. Interpret your results. This may take a few minutes 
2. Summary view. What might you change?



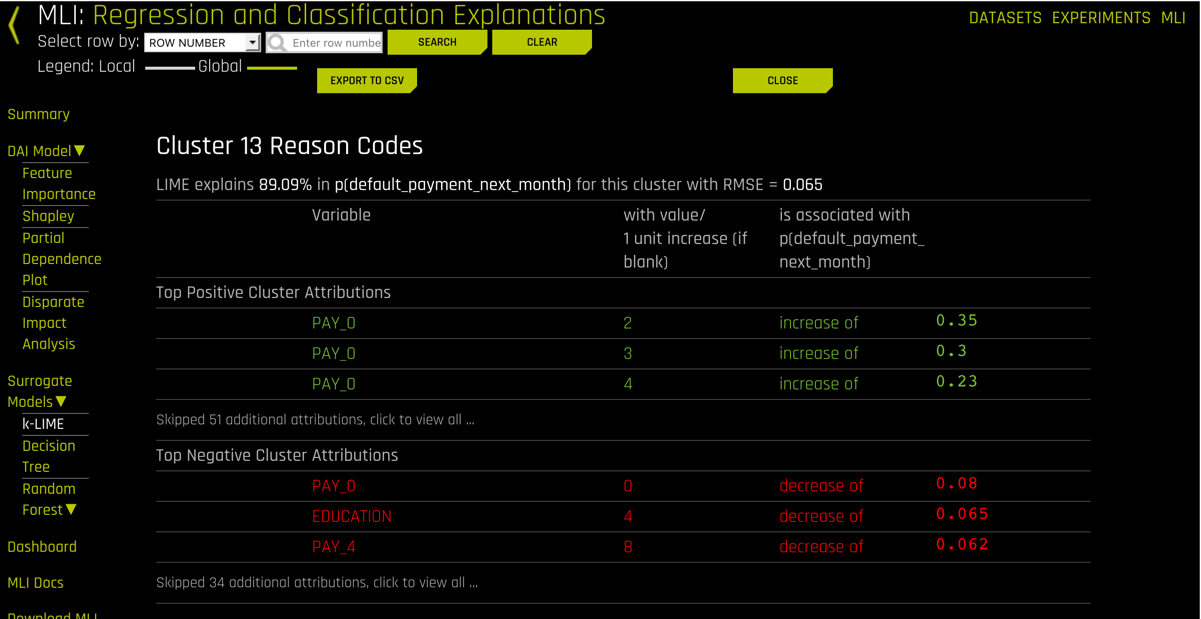
1. Let’s run through some of the menus for interpretation. Select DAI Model -> Feature Importance. This displays the weighting given by feature



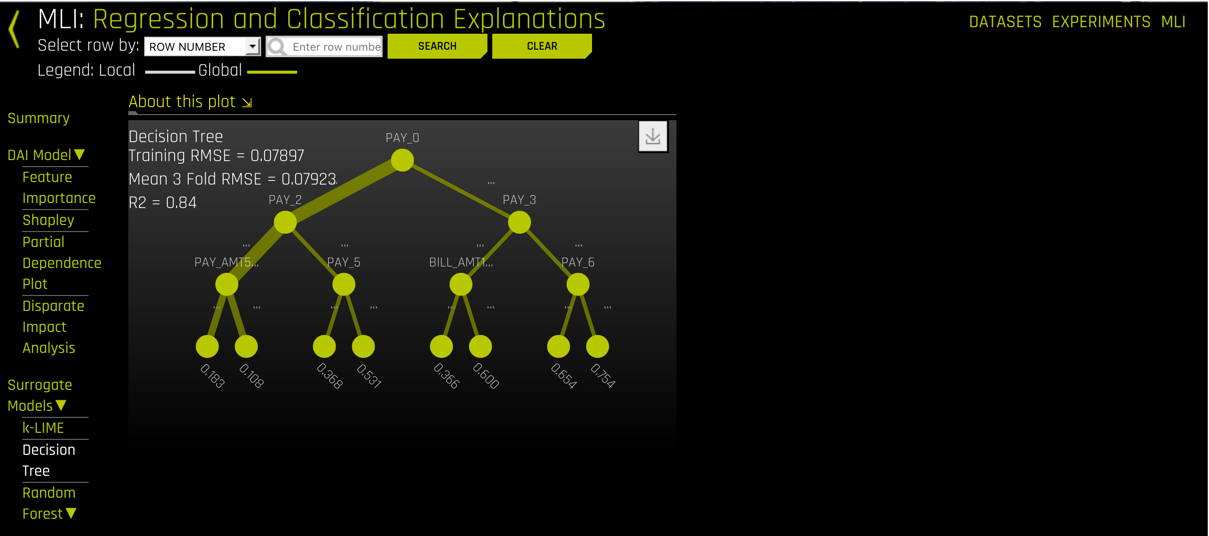
1. Select Surrogate Models -> KLIME. This plot

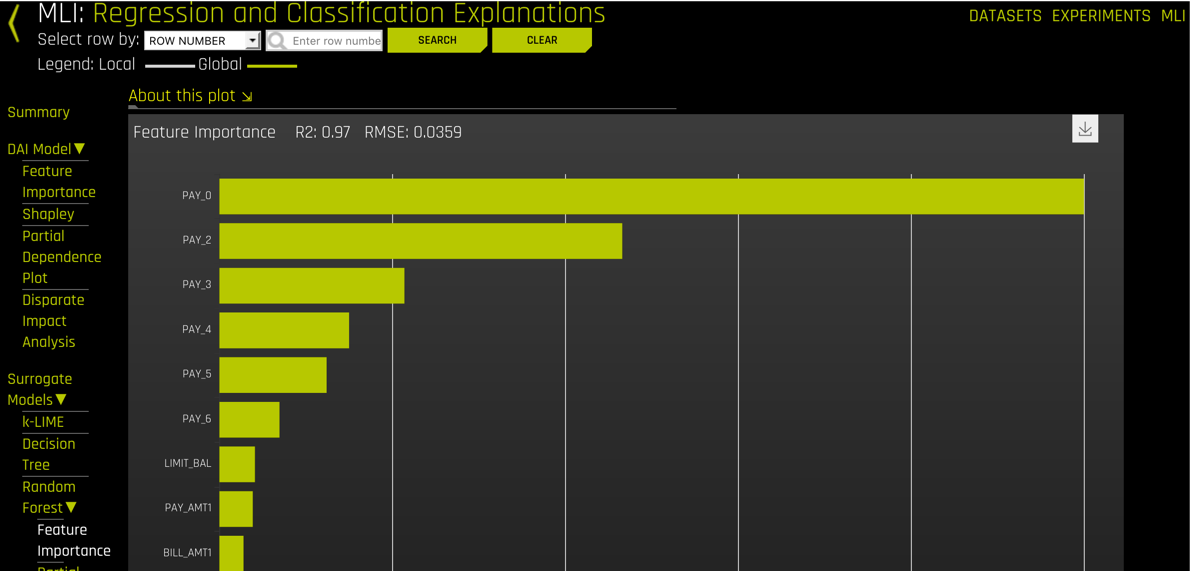
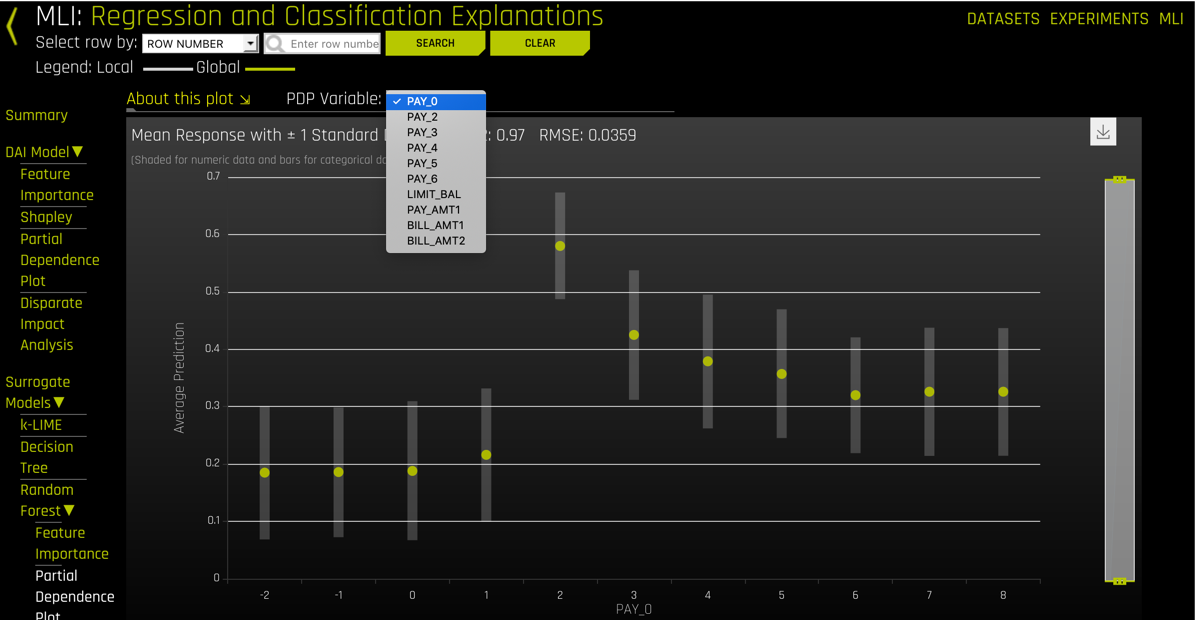
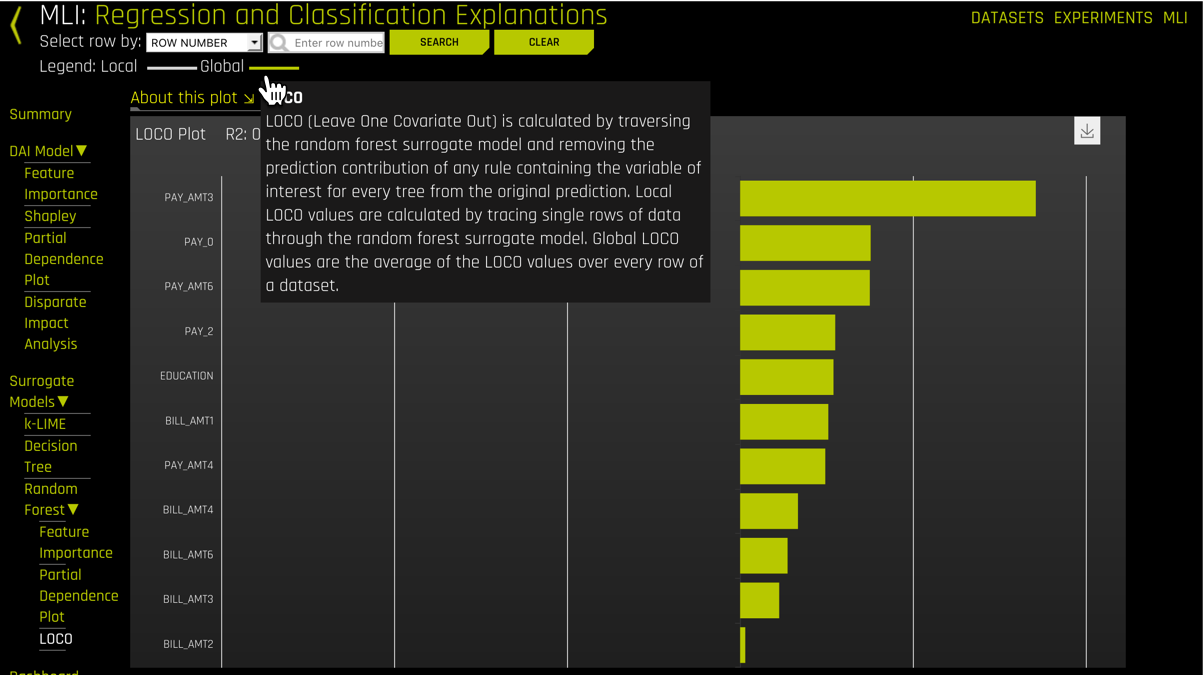


1. Selecting Explanation provides a view with the importance of various fields upon the accuracy of the model

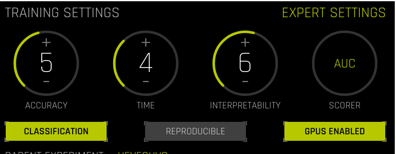


1. Under Surrogate Models, select Decision Tree

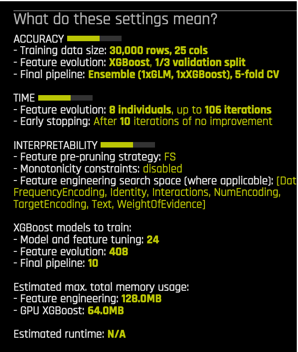


1. Under Surrogate Models, select Random Forest -> Feature Importance
2. Select Random Forest -> Partial Dependence Plot. You can filter off of column data
3. Select Random Forest -> LOCO
4. Select Dashboard to see all these views on the same page 
5. Go back to the Experiments launch page and run again with different values (“Restart from Last Checkpoint”)
   1. What if you increase Accuracy
   2. What if you increase Time
   3. What is the effect of increasing Interpretability

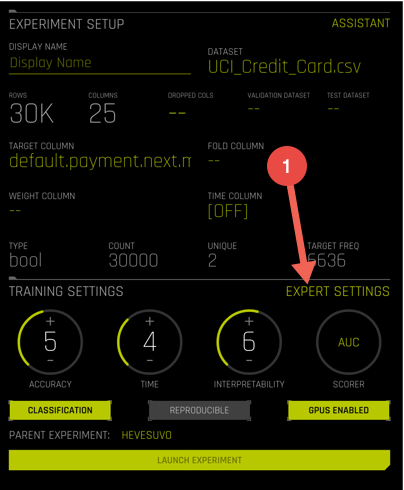




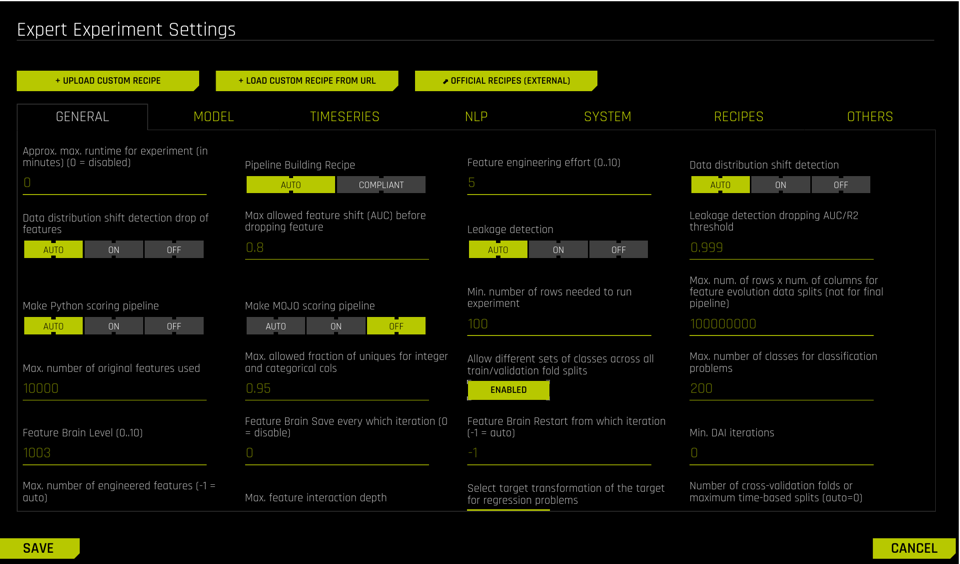
* + - Note how one change updates multiple settings



1. Launch your experiment again and view the “Expert Setting”



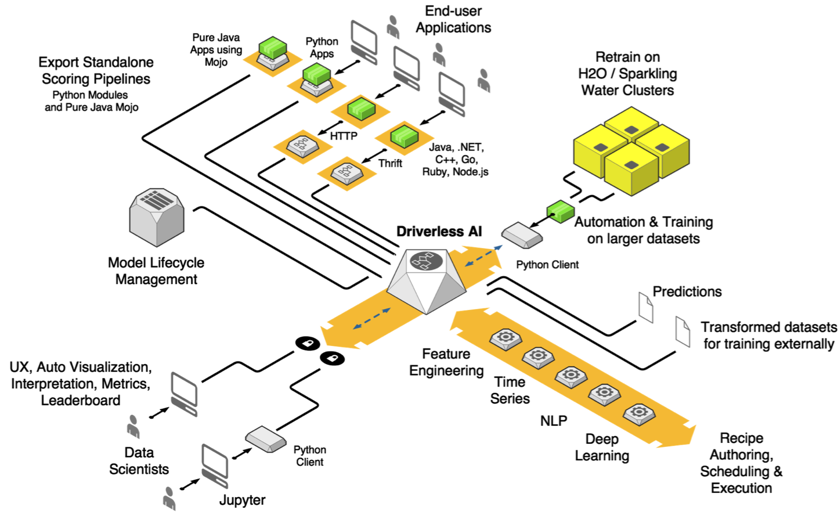
1. Note additional settings available to Data Scientists and resident Experts.



**NOT PART OF THIS LAB ---- ADDITIONAL DETAILS POTENTIAL DEMO**

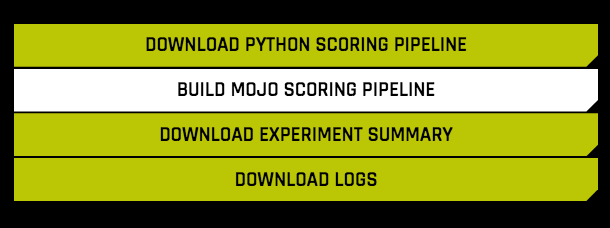
1. Deploying your model

Look at options for PY\_CLIENT and MOJO2-RUNTIME from the menus





1. Below the top center menu is a series of selections:



The most important of these is taking the built scoring model in these experiments which can then be used by other services. These would transform a testing experiment into a production model. There are two services available.

1. Phython Scoring Model. These models are currently only deployable to Lintel (Linux x86) environments. So you could prototype and experiment with Driverless AI on POWER and deploy to Lintel.
2. MOJO Scoring Model. This is a Java Object Model and is built into a Java jar file. This model can be deployed to POWER or x86, anything supporting Java.

We will build a MOJO model and test it. First, Select the Build MOJO Scoring Pipeline. This will take a minute or two.

Select “Log” on the right hand side of the H2O dashboard. You should see something like this:

2018-10-09 19:43:10 **INFO** : Creating MOJO pipeline in **'/opt/h2oai/dai/./tmp/h2oai\_experiment\_fikeciti/mojo\_pipeline**'...

2018-10-09 19:43:10 Building MOJO runtime environment...

2018-10-09 19:43:10 **INFO** : Building MOJO...

2018-10-09 19:44:07 **INFO** : Assembling MOJO pipeline...

We now know where to look for the model file.

1. Log into your docker host with a terminal (use your assigned IP, something like 10.31.204.151)
2. Determine your docker session by executing: docker ps | grep 88

To identify the CONTAINER ID you are using.

ac922a ~]# docker ps | more

CONTAINER ID IMAGE PORTS NAMES

f3fb86822c2b pai 0.0.0.0:8801->8890/tcp team01

**5adf5040f100** h2oai/dai-.. 8888/tcp…:12301->12345/tcp practical\_bell

Look for the Port # for your H2O session. Then look left for the Container ID. For the example above, if your Port is 12301, then your Container ID would be **5adf5040f100,**

1. Log into the docker container

docker login [Container ID]

1. Change to the MOJO directory from above

**cd /opt/h2oai/dai/./tmp/h2oai\_experiment\_fikeciti/mojo\_pipeline**

1. Untar the file with the model

tar xvf …