Analytics Explorer Production Prediction Workflow

IHS Markit has developed a revolutionary methodology to predict production performance and optimum engineering variables, presenting the results in maps, charts, and different error metrics to account for risk analysis from multiple model simulations. The methodology uses a mix of proprietary machine learning algorithms and modifications of existing algorithms to better adjust to nature of oil and gas data.

The following example is from the Appalachian Basin, Marcellus Shale play. We run close to 1000 model simulations. Our partner Dell has provided the laptops and workstations that IHS Markit is using for the development of such models and tools. The software applications involved in this project are The Kingdom Suite and Analytics Explorer powered by TIBCO Spotfire and IHS Markit DNA analytics-ready data.

The starting point is the Kingdom project zone data that was prepared using a variety of Kingdom tools: zone attribute calculations, data extracted from grids, and zone computations. In Analytics Explorer, we can merge this zone data with data from other data sources such as DNA to create a master table.

The Zones table in Kingdom for this project includes the following columns (attributes) which were calculated using these various Kingdom tools and DNA data:

Acoustic Impedance
Analysis Name
Analysis Type
Average Stage Length
Breakdown Pressure (psia) Median
Completed Horizontal Length
Density
First_12months_Gas
Fluid Loading
Fluid System
G - EURg
ISIP (psia) Median
K Average Azimuth
K Average Inclination
K CUM_GAS 2-6mo
K Heel Formation
K Heel Location X
K Heel Location Y
K Heel Location Z
K Inclination at Heel
K Inclination at Toe
K Lateral Length K Maximum
Azimuth
K Maximum Inclination
K Maximum Porpoise Deviation
K MD at Inclination
K Mean Porpoise Deviation
K Midpoint Location X
K Midpoint Location Y
K Midpoint Location Z
K Minimum Azimuth
K Minimum Inclination
K Porpoise Count
K Primary Formation
K Segment Count
K Toe Difference
K Toe Formation
K Toe Location X
K Toe Location Y
K Toe Location Z
K Toe Up or Toe Down
Lateral Spacing
Marcellus Depth
Marcellus_150ftDown_footage
Marcellus_150ftDown_percentage
Marcellus_90ftDown_footage
Marcellus_90ftDown_percentage
MARCELLUS_footage
MARCELLUS_percentage
P Wave Velocity
Peak Rate Gas (MMSCF/D)
Permeability
Poisson Ratio
Porosity
Proppant Mesh Size
Proppant Type
Pump Rate (BPM) Median
S Wave Velocity
Sand Loading
SW
TOC
Total Number of Stages
Vsh
Youngs's Modulus

Note: This example uses Kingdom data. The process is the same for Harmony data.

The steps in the production prediction workflow include the following:

1. Prepare your data and open Analytics Explorer
2. Run Machine Learning Predictions
3. Run the saved model on another table

Prepare your data and open Analytics Explorer

Again, in this example, the source of data is the Kingdom project Zones table. You can open Kingdom Analytics directly from the Zones table in Well Explorer. Previously, we created a production well subset with the wells producing from the Marcellus formation:

1. In Well Explorer, select the well subset. Then select Zones and the specific zone.
2. With the zone table displayed for the active well, select **Project > Kingdom Analytics > Open Active Table**. Spotfire will open to the selected Zone table. In our example, this zone is `Zone:` `Marcellus_Production (Author).`

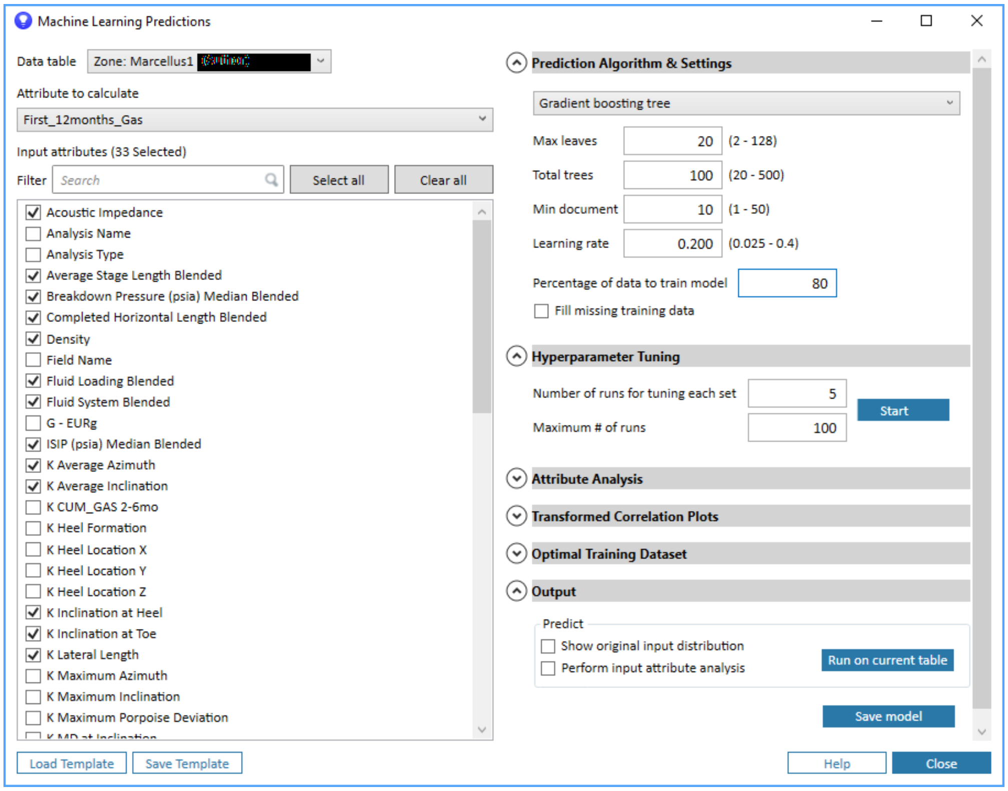
Run Machine Learning Predictions

The first step is to run the machine learning predictions: **Tools > IHS Markit Analytics Explorer > Machine Learning Predictions**. This algorithm calculates a selected metric based on specified inputs. In this example, the selected metric is the first 12 months of gas production.

For **Attribute to calculate**, select the production metric you want to calculate. For example, the first year of production, the first six months of production, the EUR for a specified time period, etc.

Input Attributes are the attributes that you want included in creating the model. This process involves an iterative process to determine the attributes to be input into the model. This selection will vary from interpreter to interpreter. The geologist, reservoir engineer, production engineer will likely consider different attributes when calculating their chosen metric. In our example, the initial run included 33 selected attributes and the default prediction algorithm and settings (hyperparameters).

Analytics Explorer has embedded Principal Component Analysis (PCA). PCA can help to determine unique and non-correlated attributes to use in the model to avoid overfitting. We highly recommend running PCA before predicting production. The results of the PCA together with the subject matter experts' knowledge are the best tools for attribute selection. For more information about PCA, please refer to the help files.



Machine Learning Predictions

Data table: Zone: Marcellus1 [redacted]

Attribute to calculate: First_12months_Gas

Input attributes (33 Selected)

Filter: Search

☑ Acoustic Impedance
☐ Analysis Name
☐ Analysis Type
☑ Average Stage Length Blended
☑ Breakdown Pressure (psia) Median Blended
☑ Completed Horizontal Length Blended
☑ Density
☐ Field Name
☑ Fluid Loading Blended
☑ Fluid System Blended
☐ G - EURg
☑ ISIP (psia) Median Blended
☑ K Average Azimuth
☑ K Average Inclination
☐ K CUM_GAS 2-6mo
☐ K Heel Formation
☐ K Heel Location X
☐ K Heel Location Y
☐ K Heel Location Z
☑ K Inclination at Heel
☑ K Inclination at Toe
☑ K Lateral Length
☐ K Maximum Azimuth
☐ K Maximum Inclination
☐ K Maximum Porpoise Deviation
☐ K MD at Inclination

Prediction Algorithm & Settings

Gradient boosting tree

Max leaves: 20 (2 - 128)
Total trees: 100 (20 - 500)
Min document: 10 (1 - 50)
Learning rate: 0.200 (0.025 - 0.4)

Percentage of data to train model: 80

☐ Fill missing training data

Hyperparameter Tuning

Number of runs for tuning each set: 5
Maximum # of runs: 100

Predict
☐ Show original input distribution
☐ Perform input attribute analysis

Note that the interpreter's attribute selections can be saved to a template to allow the interpreter to modify, run, save, and load quickly and efficiently. After making your selection, click Save Template below the attribute list. You can also save the model which includes the data table, all attributes, the algorithm, and the hyperparameter settings.

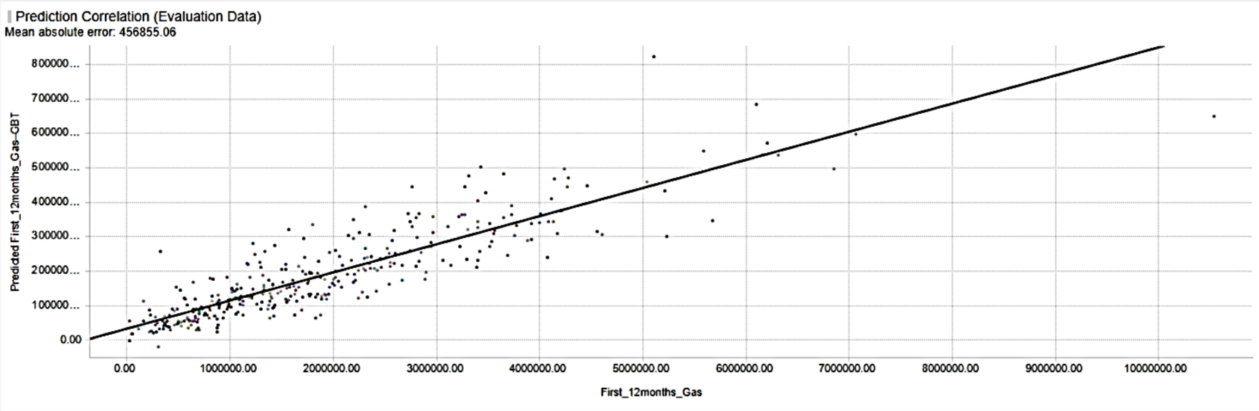
In general, the default algorithm and settings will provide optimal results. The defaults provide a solid starting point for analysis. After you select your input attributes, simply click Run on current table. The resulting visualizations are the Prediction Correlation Plot and the Importance per Variable graph.

The Machine Learning Prediction produces 2 initial key charts: **Prediction Correlation Plot**, and **Importance per Variable**. Each are discussed below.

Prediction Correlation Plot

This plot reveals how well the predicted values of the selected attribute correlate with the actual data from the testing wells (By default we use 80% of the samples to train the model and 20% of the samples to test the models). However, in a non-linear progression problem, the prediction correlation may not be the best indicator. Therefore, the Mean Absolute Error (MAE) is displayed for the production metric values displayed on the Y axis.

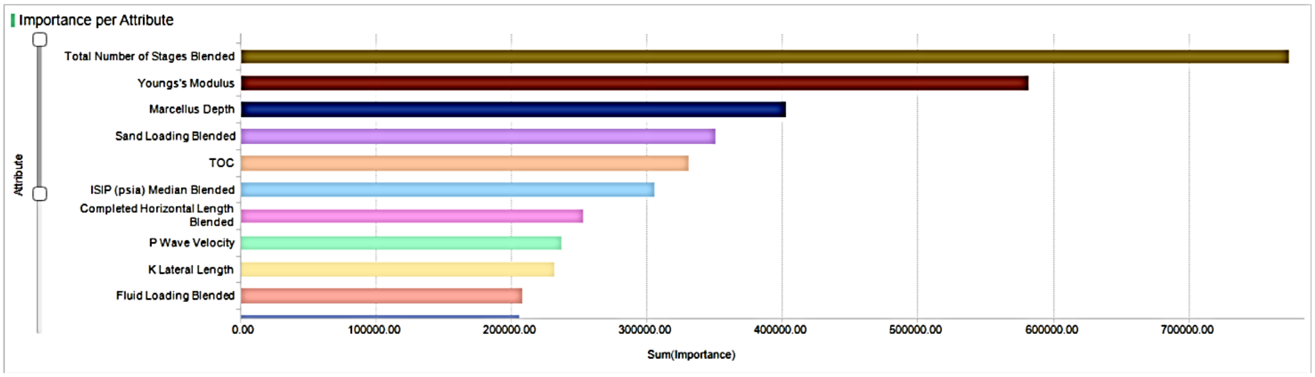
In this example the average gas production for the first year of production is approximately 7 million cubic feet gas. The calculated MAE given the attributes selected will be approx. 456,855 cubic feet gas, or 4%.



The next graph shows the rated impact of each selected variable on the prediction model.

Importance per Variable

The order of importance of the variables denotes the variables that have the greatest impact on the selected attribute, in this example, 12-month production.



The next question is "Can we fine-tune the hyperparameters to produce an even better model?"

You can also adjust/run the following to fine-tune your model:

- Prediction algorithm & Settings
- Hyperparameter Tuning
- Attribute Analysis
- Transformed Correlation Plots

After you have created and saved your model, you can **Run** the saved model on another table.

Prediction algorithm & Settings

Our data science research teams and analysts determined optimal defaults for the prediction algorithm (Gradient boosting tree) and hyperparameter settings:

- Max leaves (20)
- Total trees (100)
- Min document (10)
- Learning rate 0.200
- Percentage of data to train model (80)

The range of each parameter is displayed beside the value. In most cases, these default parameters will give the best results. A good starting point is to run the model with the default settings to provide a baseline for further tuning. However, you can adjust the hyperparameters with Hyperparameter Tuning to fine-tune the values to those which produce the minimal MAE with your dataset. For a description of each parameter see the online help.

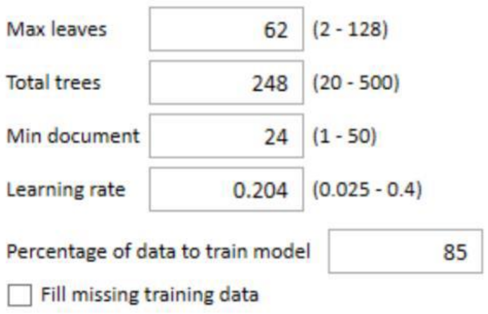
Hyperparameter Tuning

The hyperparameters tuning options are advanced options to tune the parameter values to your data. If you want a greater degree of precision, you can increase the maximum # of runs and the number of runs for tuning each set. Again, the default parameters have proven to yield optimal and consistent results.

• **Maximum # of runs**—the default is 100, which means that the application will change the hyperparameters 100 times in pursuit of the model that produces the smallest MAE.

• **Number of runs** for tuning each set for each run. The default is 100. The application will create 5 models. For each model, the Prediction Settings will be different because the data used to run the model is randomly selected.

To begin, Start the hyperparameter tuning using the default tuning parameters of 5 and 100. After you run the Hyperparameter Tuning, the hyperparameter fields will populate with the values that produced the lowest MAE. The "best" model is the one that produces the lowest MAE. In our example, after running the hyperparameter tuning using the default values, the resulting settings were:



Max leaves	62	(2 - 128)
Total trees	248	(20 - 500)
Min document	24	(1 - 50)
Learning rate	0.204	(0.025 - 0.4)

Percentage of data to train model 85

☐ Fill missing training data

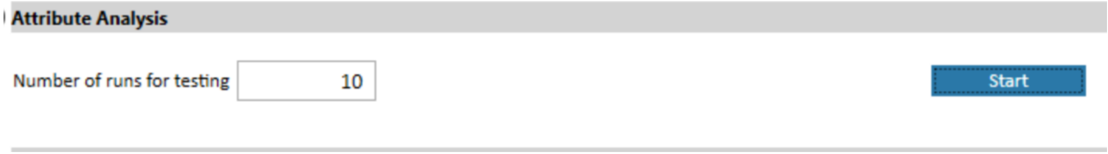
You can adjust the hyperparameter tuning parameters but note that a large increase in the values may take a long time to compute and yield minimum reduction in the MAE. When you have determined the optimal input attributes and hyperparameter settings, you can save the model to use as a starting point for later models.

Now, can the MAE inform the user how many or which attributes to use in the analysis? Out of the 33 attributes we started with, can we identify which ones had the greatest impact?

Attribute Analysis

Attribute Analysis computes attribute importance (Importance per Attribute chart) and the MAE based on the number of attributes used in the calculation selected in priority order.

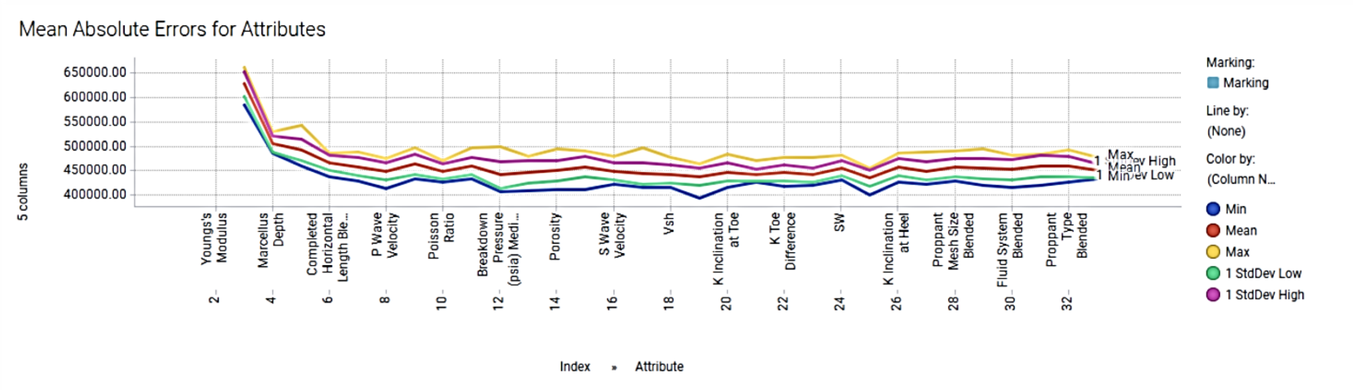
First a model is created using the top 3 attributes and the MAE is calculated. This process is repeated a default of 10 times (Number of runs for testing).

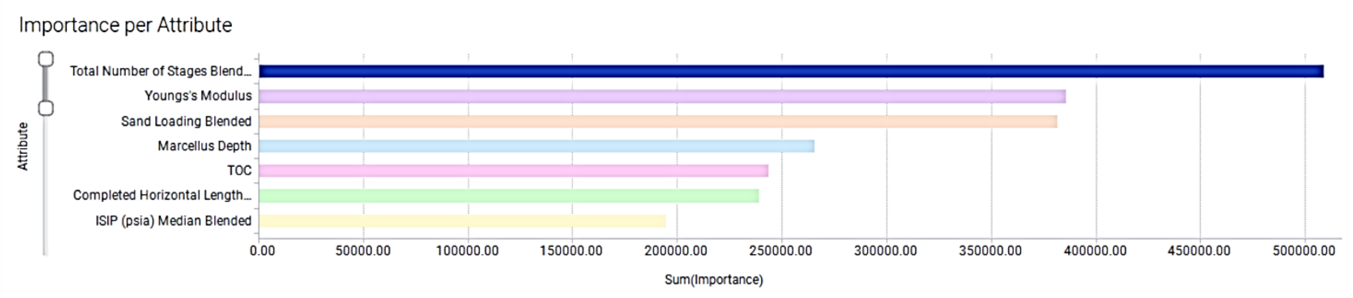


Attribute Analysis

Number of runs for testing: 10

Then, the next attribute in order of importance is added and another 10 models are created. This process is repeated with each attribute. The output is the Mean Absolute Errors for Attributes graph, and the Importance per Attribute graph.





The Mean Absolute Errors graph shows the MAE as each attribute is added. Note that the graph starts at 3 attributes, the minimum number. At 3 attributes, the MAE is high. In this example, close to 750000 out of the total estimated average production of approximately 7 million cubic feet, so around 10%. Adding the fourth attribute drops the MAE substantially, to around 500000. At around the 6th attribute, the line starts to flatten, signifying that adding additional attributes does not significantly affect the MAE.

The lowest MAE is at 19 attributes. Now based on the results we can build a new model using the identified top 19 attributes and the tuned hyperparameter values. In our example the 19 top attributes were the following:

Acoustic Impedance
Breakdown Pressure Median
Completed Horizontal Length
Fluid Loading
ISIP Median
K Lateral Length
Marcellus Depth
P Wave Velocity
Permeability
Poisson Ratio
Porosity
S Wave Velocity
Sand Loading
TOC
Total Number of Stages
Vsh
Youngs's Modulus

Again, we can save the new attribute selections as a template (**Save Template**) and save the model itself (**Save model**) which includes all selections and settings.

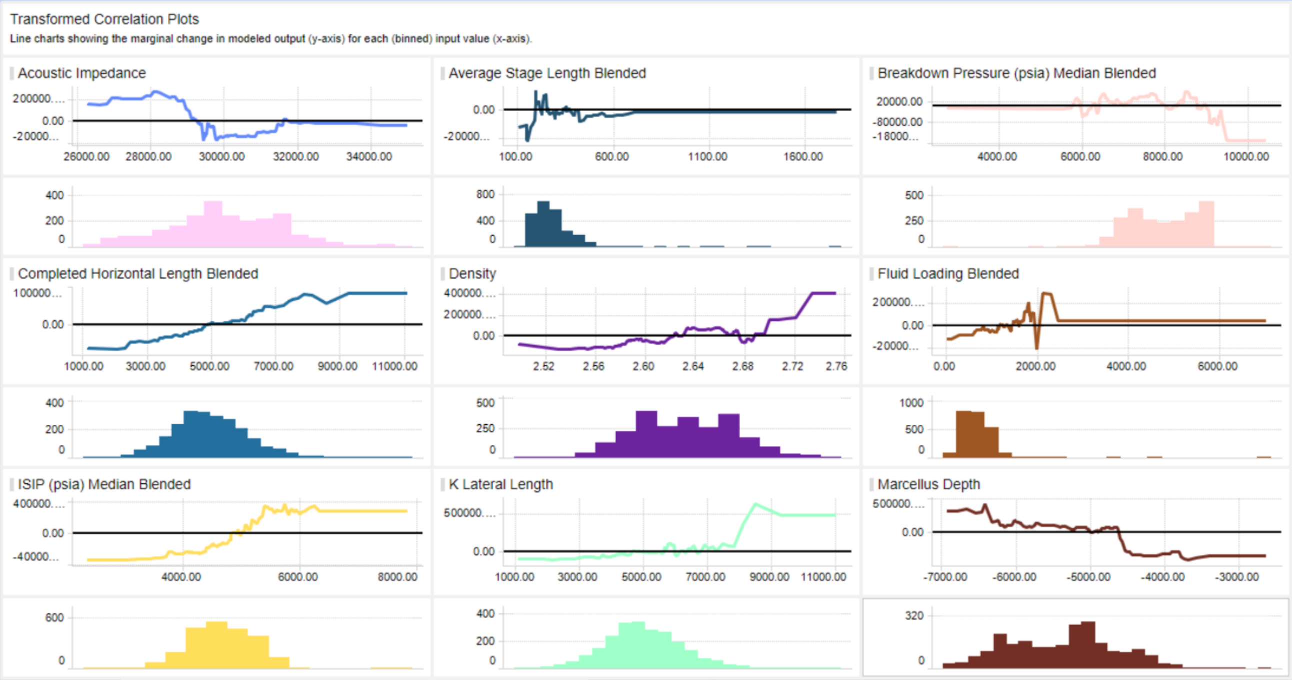
Transformed Correlation Plots

Transformed correlation plots relate how each of the attributes contributes to the model (affect the production, positive or negative).

When we build the transformed correlation plots, we divide each of the attributes into 100 bins. For example, in the entire data set, the range of lateral length goes from 3000 to 10000. Within this range we create 100 bins. Every bin is analyzed independently computing how each bin of wells contributes to production. This process is applied successively to each attribute. If you increase the number of bins, the computations will take longer, and individual bin analysis may be irrelevant.

The image below shows the transformed correlation plots in our Marcellus shale example. The model was rerun using the 19 top attributes as identified by the Attribute Analysis.

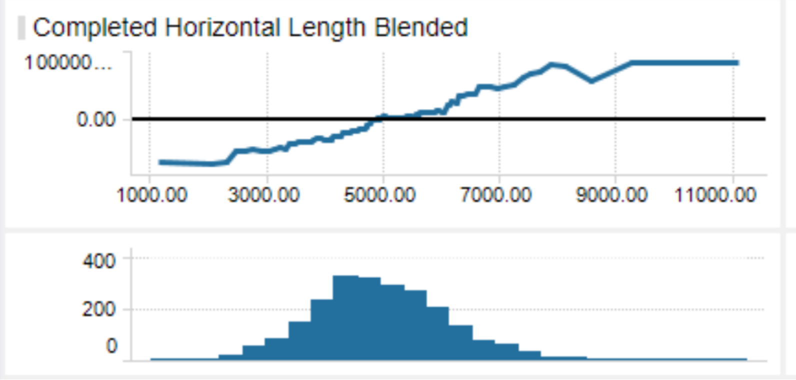
The horizontal line at 0.00 is the average first year of production. The X axis is the value of each attribute. The Y axis is the first year of production. The plots show how each attribute affects production. The histogram below the graph shows the distribution of the data.



Transformed Correlation Plots

Line charts showing the marginal change in modeled output (y-axis) for each (binned) input value (x-axis).

Let's take completed lateral length as an example. The completed lateral length range (X axis) is 0 to 11000 feet. The first year of production is the Y axis.



Let’s take completed lateral length as an example. The completed lateral length range (X axis) is 0 to 11000 feet. The first year of production is the Y axis.

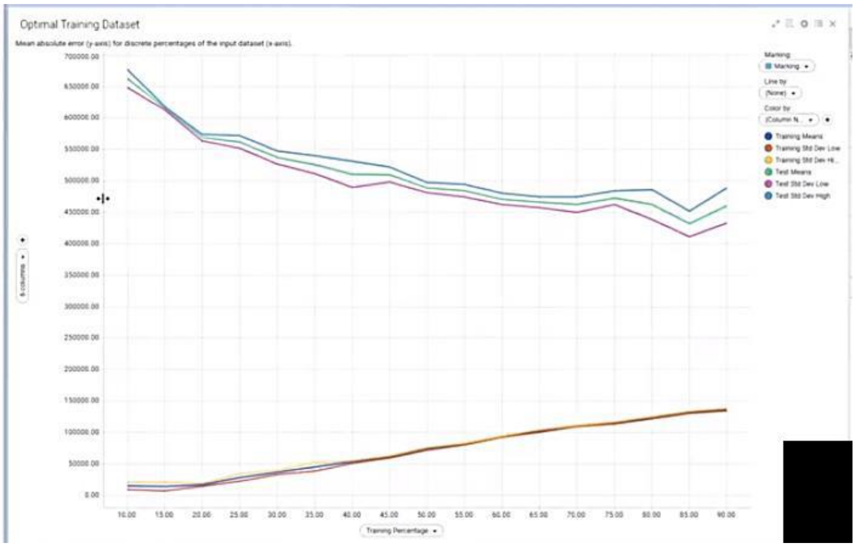
The average first year of production is reached at a lateral length of approximately 5000 feet. The peak production rate occurs at approximately 8000 feet. Longer lengths do not increase the production rate. We can therefore conclude that a lateral length of 7800 ft is optimal based on the input data we have. We also plot the histogram of the input data right below the transformation plot of each variable. In this way the interpreter can analyze if a potential peak in production is due to a solid set of points or to an outlier.

Optimal Training Dataset

In most cases, this option will just confirm that the current parameters minimize MAE. Running this function generates a chart that correlates the mean absolute error with discrete percentages of the input dataset.

The graph below is the output after running the **Optimal Training Dataset** option. The bottom lines reflect the MAE when using successive increments of the training/testing samples to predict the same data (notice that when predicting the same data set the MAE values are low, but this approach does not necessarily reflect how well the model will behave with new data coming in). The top lines show the error when using different training/testing samples. For example, the first point trains the model with 10% of the samples to predict the remaining 90% of those samples; the resulting error is high. However, the error decreases as we use more testing data and less training data. Even though the MAE is higher, the top lines better quantify the error and the risk to predict production in a new well location in the Marcellus basin.

The default training dataset is 80%. In our example, after we ran hyperparameter tuning, the application determined that 85% minimized MAE. In the upper lines on the graph below, you can see the dip at a training dataset of 85%.


Optimal Training Dataset

Mean absolute error (y-axis) for discrete percentages of the input dataset (x-axis).

Marking: Marking

Line by: (None)

Color by: (Column N...

[redacted]

Run the saved model on another table

The production predictive model has been created using a single value of every attribute per well. However, the final trusted model can now be applied to a different table such as the grid table to create a more detailed production prediction map honoring the full geological lateral facies variations or to include the details of seismic attributes. In this example, we are going to run our final 19-attribute model on the Grids table. The saved model includes the attributes, selected algorithm, and all hyperparameters. Another example where you can use the final trusted prediction model in another table is in cases where the model belongs to a specific basin where you have good data control and you want to apply the same model to another basin that can be considered of similar geological characteristics of the one used to create the model.

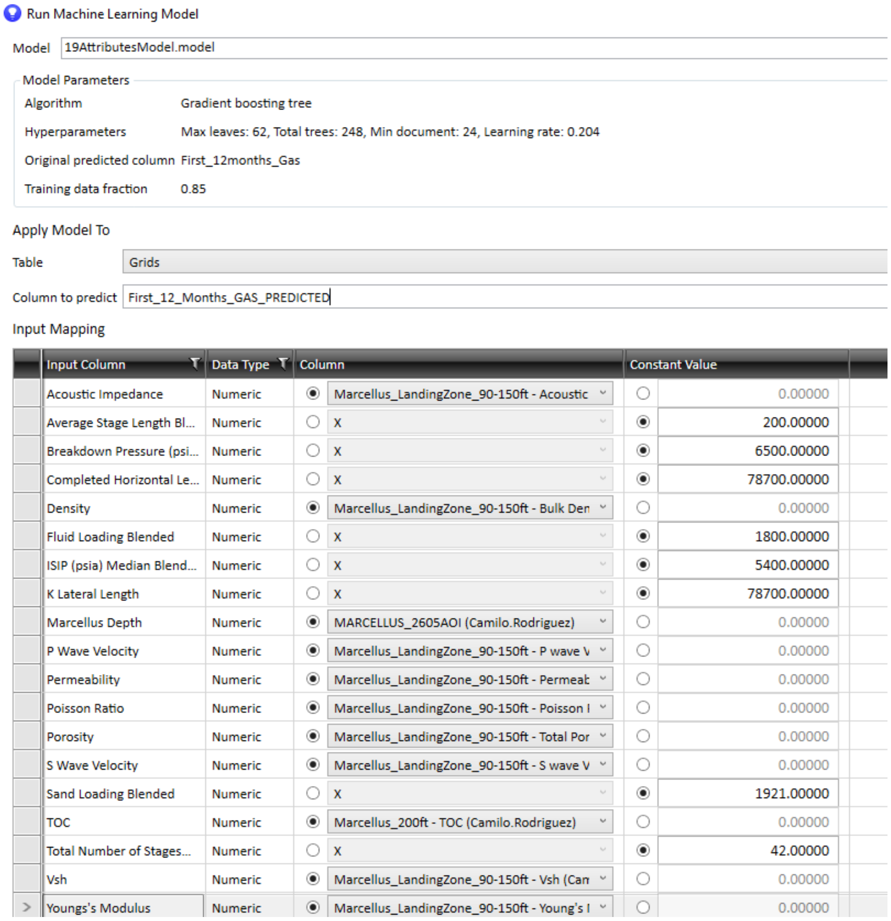
When running the model on another table, the interface exposes all the attributes used to create the final model. The user then matches every attribute to the corresponding grid. The engineering parameters, which don't have any corresponding grids can be defined as constant values. Such values can be taken from the optimum values interpreted from the Transformed Correlation Plots.

To run the saved model on another dataset (in this case the Grids table)

1. Select **Tools > IHS Markit Analytics Explorer > Run Machine Learning Model** and browse to the saved model file (*.model). The dialog box displays a table with all the model inputs.
2. Now select the table you want to apply the model to. In this example, we are applying the model to the Grids table.
3. Select or enter the attribute to predict. If you enter a new attribute name, that attribute will be added as a new column in the table. In our example we are adding a new attribute, `First_12_Months_GAS_PREDICTED`.
4. For each input, you can select a column in the table, or specify a constant value. The application tries to map the input columns with the columns in the table based on name. For example, for Acoustic Impedance, if the grid table has a column with Acoustic Impedance in the column header, that column in the grid table will be selected.

   For some input columns (attributes) you can specify a specific constant value that was determined from the transformed correlation plots. For example, we know that the optimal lateral length is 78700 feet. We also have values for other engineering parameters.

   The figure below displays the parameters for our example:



Run Machine Learning Model

Model: 19AttributesModel.model

Model Parameters

Algorithm	Gradient boosting tree
Hyperparameters	Max leaves: 62, Total trees: 248, Min document: 24, Learning rate: 0.204
Original predicted column	First_12months_Gas
Training data fraction	0.85

Apply Model To

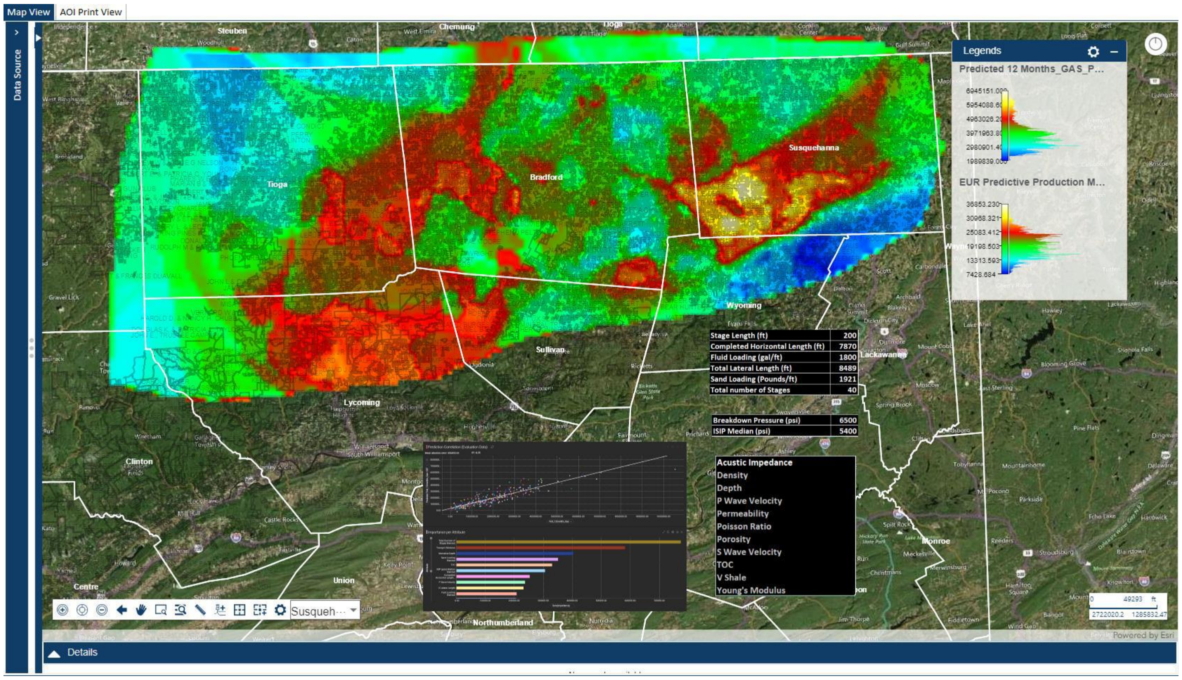
Table: Grids

Column to predict: First_12_Months_GAS_PREDICTED

Input Mapping

Input Column	Data Type		Column		Constant Value
Acoustic Impedance	Numeric	●	Marcellus_LandingZone_90-150ft - Acoustic	○	0.00000
Average Stage Length Bl...	Numeric	○	X	●	200.00000
Breakdown Pressure (psi...	Numeric	○	X	●	6500.00000
Completed Horizontal Le...	Numeric	○	X	●	78700.00000
Density	Numeric	●	Marcellus_LandingZone_90-150ft - Bulk Den	○	0.00000
Fluid Loading Blended	Numeric	○	X	●	1800.00000
ISIP (psia) Median Blend...	Numeric	○	X	●	5400.00000
K Lateral Length	Numeric	○	X	●	78700.00000
Marcellus Depth	Numeric	●	MARCELLUS_2605AOI (Camilo.Rodriguez)	○	0.00000
P Wave Velocity	Numeric	●	Marcellus_LandingZone_90-150ft - P wave V	○	0.00000
Permeability	Numeric	●	Marcellus_LandingZone_90-150ft - Permeab	○	0.00000
Poisson Ratio	Numeric	●	Marcellus_LandingZone_90-150ft - Poisson I	○	0.00000
Porosity	Numeric	●	Marcellus_LandingZone_90-150ft - Total Por	○	0.00000
S Wave Velocity	Numeric	●	Marcellus_LandingZone_90-150ft - S wave V	○	0.00000
Sand Loading Blended	Numeric	○	X	●	1921.00000
TOC	Numeric	●	Marcellus_200ft - TOC (Camilo.Rodriguez)	○	0.00000
Total Number of Stages...	Numeric	○	X	●	42.00000
Vsh	Numeric	●	Marcellus_LandingZone_90-150ft - Vsh (Can	○	0.00000
Youngs's Modulus	Numeric	●	Marcellus_LandingZone_90-150ft - Young's I	○	0.00000

Now we run the model on the Grids table and create the predicted first 12 months of gas production. After the run, you can open the Grids table, right click on the new column, and save the column as a grid back to Kingdom which you can then display in the base map and Spatial Explorer.



Stage Length (ft)	200
Completed Horizontal Length (ft)	7870
Fluid Loading (gal/ft)	1800
Total Lateral Length (ft)	8489
Sand Loading (Pounds/ft)	1921
Total number of Stages	40

Breakdown Pressure (psi)	6500
ISIP Median (psi)	5400

Acustic Impedance
Density
Depth
P Wave Velocity
Permeability
Poisson Ratio
Porosity
S Wave Velocity
TOC
V Shale
Young's Modulus