Darwin™ User Interface Guide

A SparkCognition™ Education Document

v. 2.0.4 - 12.2019

This document contains copyrighted and proprietary information of SparkCognition and is protected by United States copyright laws and international treaty provisions. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under such laws or with the prior written permission of SparkCognition Inc.

SparkCognition ™ , the sparkcognition logo, Darwin ™ , DeepArmor ® , DeepNLP ™ , MindFabric ® , SparkSecure ® and SparkPredict ™ , are trademarks of SparkCognition, Inc. and/or its affiliates and may not be used without written permission. All other trademarks are the property of their respective owners.

©SparkCognition, Inc. 2017-2019. All rights reserved.

Darwin User Interface Guide

Contents

Overview **2**
- System recommendations . . . 2
- Darwin documentation . . . 2

Logging in to Darwin **2**
- Resetting your password . . . 3
- Expanding the navigation menu . . . 3
- Updating your user profile . . . 4

Creating a new model **5**

Working with models **11**
- Searching for and sorting models . . . 11
- Viewing training sessions and results . . . 11
  - Viewing classification models . . . 12
  - Viewing regression models . . . 17
- Downloading a model . . . 19
- Deleting models . . . 20

Generating predictions **21**

Working with the Data Bank **22**
- Uploading a new dataset . . . 22
- Previewing a dataset . . . 23
- Downloading a dataset . . . 23
- Deleting datasets . . . 24
  - Deleting a single dataset . . . 24
  - Deleting multiple datasets . . . 25
- Contact Support . . . 25

Revision Table **26**

Overview

Darwin™ is a SparkCognition™ tool that automates model-building processes to solve specific problems. This tool enhances data scientist potential because it automates various tasks that are often manually performed. These tasks include data cleaning, latent relationship extraction, and optimal model determination. Darwin promotes rapid and accurate feature generation through both automated windowing and risk generation. Darwin quickly creates highly accurate, dynamic models using both supervised and unsupervised learning methods.

System recommendations

To ensure the best experience using Darwin, follow these recommendations:

- Use the most recent Google Chrome browser
- Set the screen resolution to a recommended 1440 x 900 pixels. The minimum value supported is 800 x 600.
- Ensure that your dataset .csv files are encoded to one of the following standards:
  - `utf-8`
  - `us-ascii`
  - `iso-8859-1`
  - `iso-8859-2`
  - `ascii`

  Note: Occasionally, dataset .csv files are not properly encoded and may contain both single-byte and multi-byte characters. This situation can be resolved by encoding the file as `utf-8`. See the following reference for instructions on how to save a .csv file as a `utf-8` encoded file: https://www.webtoffee.com/how-to-save-csv-excel-file-as-utf-8-encoded/

Darwin documentation

The documentation for this version of Darwin includes:

- The *Darwin Release Notes*, version 2.0.4
- The *Darwin User Interface Guide*, version 2.0.4
- The *Darwin API User Guide*, version 1.36.0
- The *Darwin Python SDK User Guide*, version 1.45.0
- The *Darwin RTE User Guide*, version 2.0.4

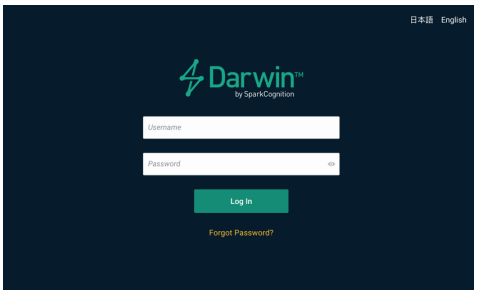
All of these documents are available for download from the Darwin support portal.

Logging in to Darwin

To log in to the Darwin user interface:

1. Obtain the login URL from your Darwin administrator or SparkCognition representative.

2. Enter the URL in your browser. The Darwin login screen appears.



3. Enter your username and password.
4. Click **Log In**. The Models page appears.

If you forgot your password and want to reset it, see the next section.

Resetting your password

Perform the following to reset your password:

1. Ensure that your username is correct.
2. Click the **Forgot Password?** link on the Login page. An email is sent to the email address associated with the username.
3. Follow the instructions in the email. If you don't get an email, check your Spam folder.

Expanding the navigation menu

The navigation menu along the left-side of the browser window shows icons by default, perform the following to expand this menu.

1. Roll your mouse over the right-hand edge of the navigation menu. A green expand icon appears.

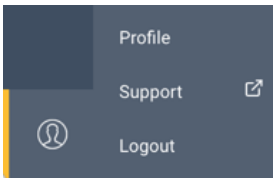
2. Click in this area to expand the navigation menu.

3. Click in the same area to collapse the menu to its icon format.

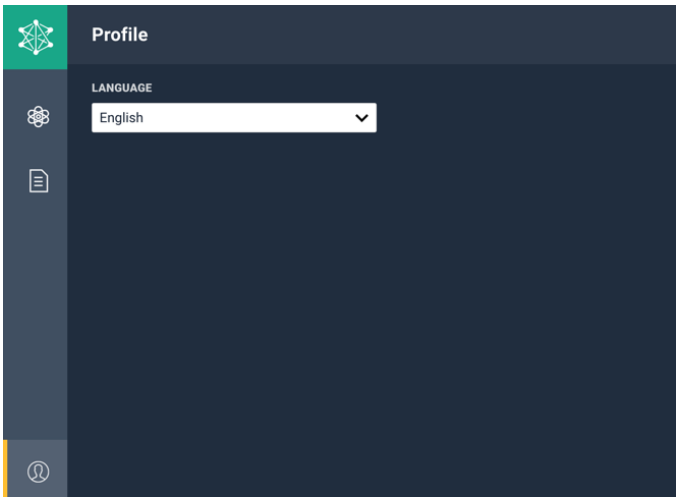
Updating your user profile

Perform the following to update your user profile:

1. After a successful login, click the user profile icon in the lower-left of the user interface. This displays links to your profile as well as a link to the Darwin Support website. You can also logout from this menu.



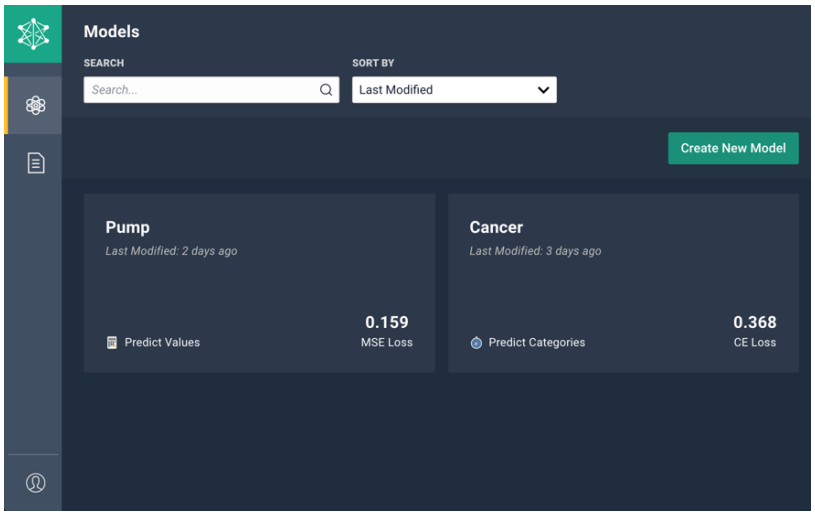
2. Select **Profile** from this menu. The Profile page appears.



3. Select the display language for the user interface.

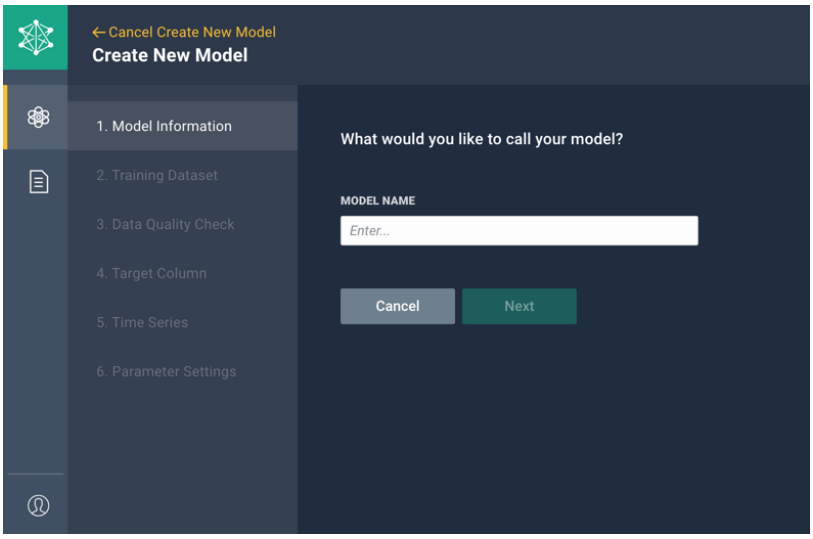
Creating a new model

One of the first tasks you need to perform is to create an initial model. The **Models** page allows you to create new models as well as search for previously created models and sort them by a variety of criteria.



To create a new model, perform the following:

1. Click **Create New Model** to begin. The Create New Model page appears.

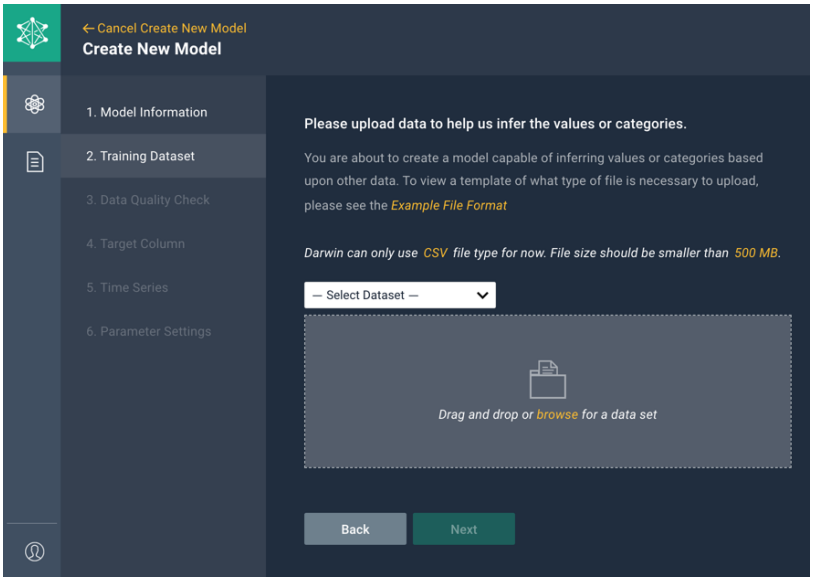


2. Enter the name of the model you want to create. We recommend a model name of less than 20 characters. Click **Next**.

3. Upload the dataset to infer the values or categories depending on your selection. There is a limit of 500MB for the dataset size. The dataset must be a .csv file type and it must have less than 1000 columns. If you need to use a larger dataset, you should use the SDK for your model building. If you have already uploaded a dataset and you want to re-use it, select it from the **Select Dataset** drop-down list. Otherwise, drag-and-drop your dataset into the appropriate area. Alternatively, you can browse for the dataset you want to upload.

   If you want to upload several datasets for use in Darwin, you can use the Data Bank tab, where you can manage all your datasets. Any dataset loaded there will appear in the **Select Dataset**

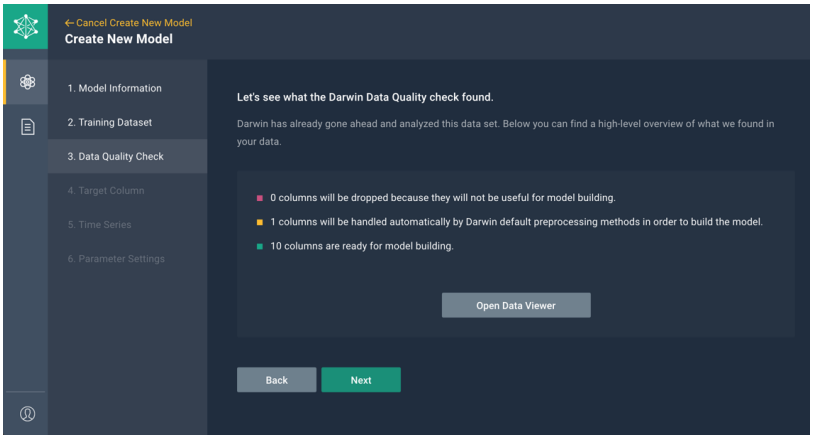
drop-down list.



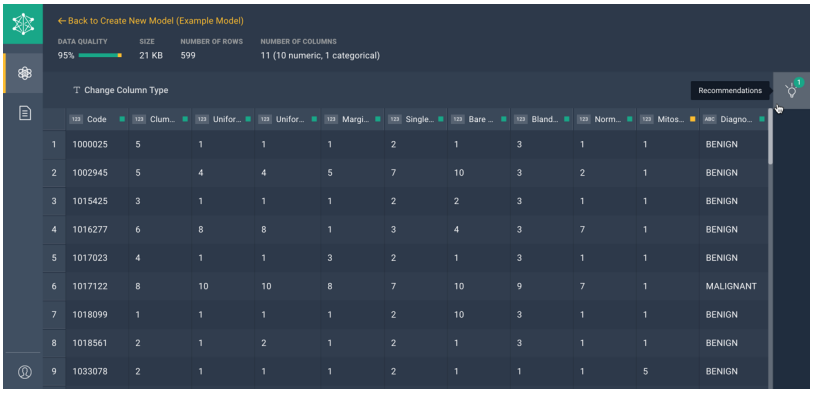
Once your dataset is loaded, its name appears. Click **Next** to continue.

Note: To view a sample of the file format your dataset should be in, click the **Example File Format** link.

4. Darwin performs a data quality check on your dataset and provides an overview of its findings.

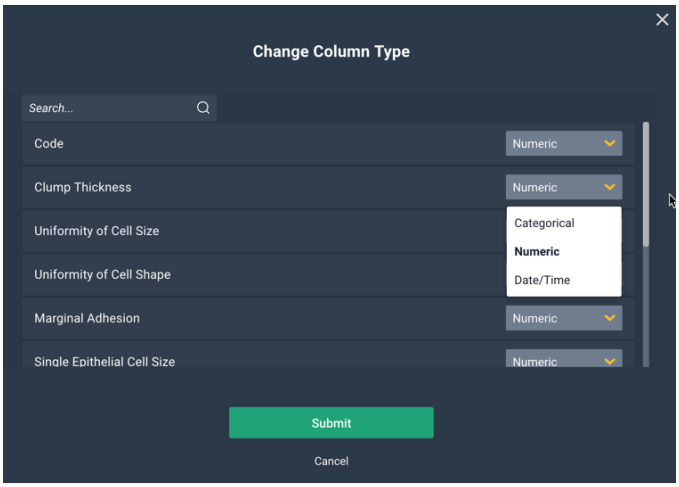


You can get further details by clicking the **Open Data Viewer** button where you can see recommendations and make adjustments to your dataset prior to model building.

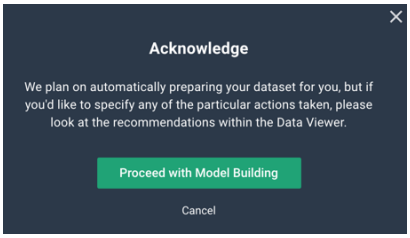


Click the **Recommendations** icon on the right side of the window to view suggested recommendations. These recommendations can be marked as yellow, indicating that they will be handled by default preprocessing methods unless you change them, or possibly marked red, indicating a column that will be dropped. These recommendations can indicate problems with low statistical variance, too many missing values, incorrect data types, or too many categories.

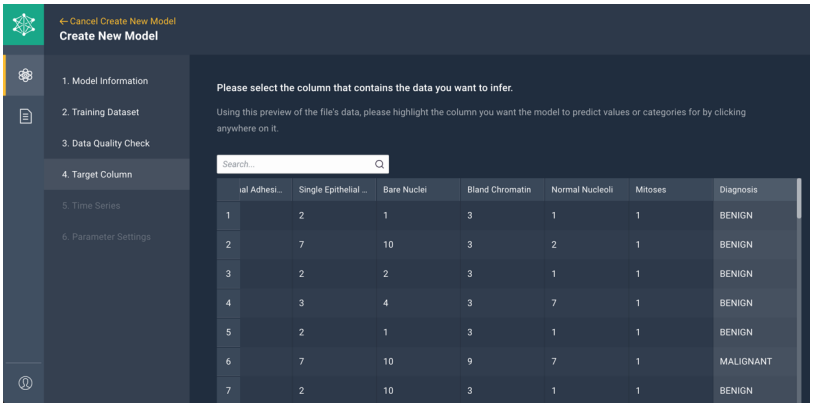
Click on a recommendation to be guided about making an adjustment. You can also make multiple changes to the column type based on this analysis. To do this, click the **Change Column Type** button to display all the columns. You can then change a column to a different column type by selecting the appropriate drop-down list and clicking **Submit**. When you are satisfied with your changes, click the **Back to Create New Model** link to continue.



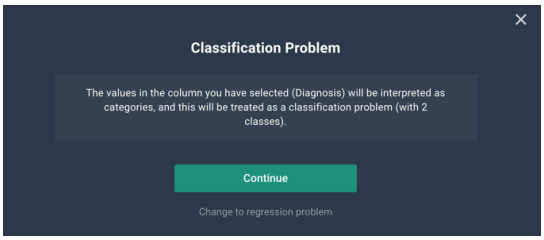
To proceed with the next step in the model building process, click **Next** and then acknowledge by clicking the **Proceed with Model Building** button.



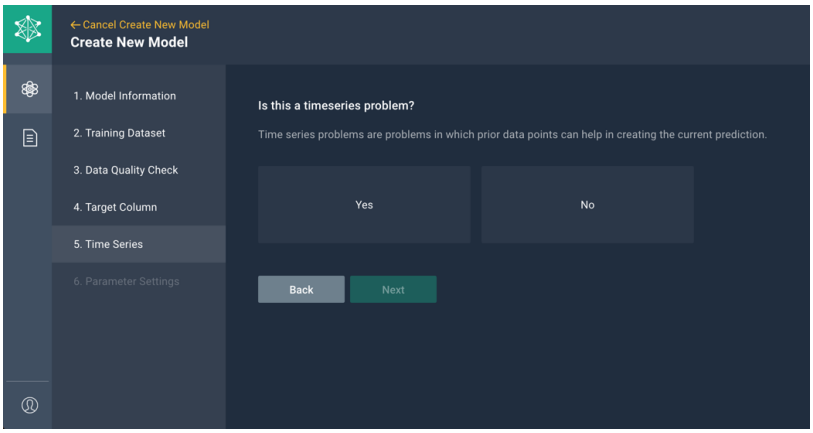
5. Select the column that contains the data you want to infer. The initial view displays the first 50 rows. You may have to scroll to the right to see the slider controls to view the latter rows. Once you have selected the appropriate column, click **Next**.



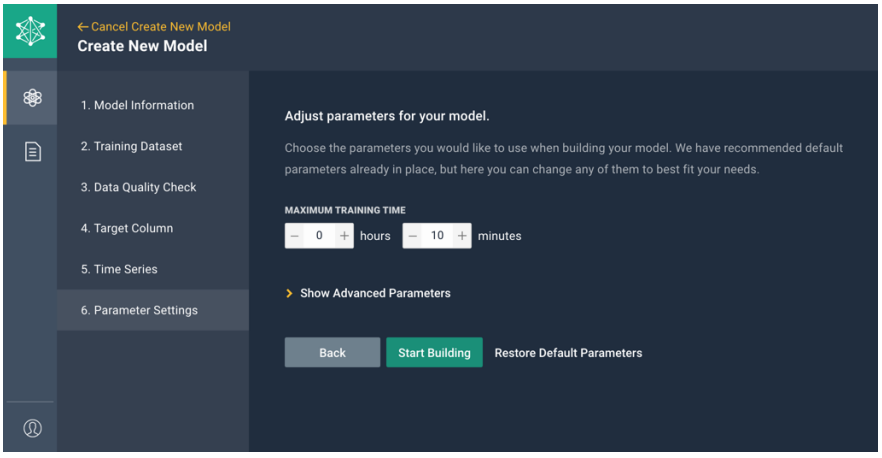
6. Based on your column selection, Darwin infers the type of problem you are trying to solve (either Classification or Regression). To accept the problem type, click **Continue**.



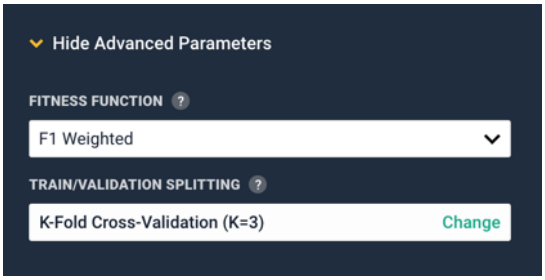
7. Answer the question about whether this is a timeseries problem. A timeseries problem is one which prior data points can help in creating the current prediction. Select **Yes** or **No**, then click **Next**.



8. Adjust the parameters for your model. You can set the maximum training time (default 10 minutes), by clicking the **+** and **-** controls as appropriate.



9. Click the **Show Advanced Parameters** link to view and edit these parameters. The following window appears with the advanced parameters.



The following parameters are available:

- **Fitness Function** - Allows you to specify the fitness function Darwin uses to control the model population's evolution during training. The final model from Darwin will be optimized for this particular objective.

  For classification problems, possible values include:

  - `Average Precision` - Measures the average precision across the spectrum of all recall values from 0 to 1. Average precision is a good metric to use for imbalanced problems, and only works on binary target columns, that is, there are two class labels being predicted.
  - `ROC Area Under Curve` - Measures the area under the Receiver Operating Characteristics curve, which plots the relationship between precision and recall for a model. ROC area under curve only works on binary target columns, that is, there are two class labels being predicted.
  - `Accuracy` - Measures the total number of correct predictions divided by the total number of predictions made.
  - `F1 Weighted` - (default) Measures the F1 score for each label and finds their average, which is weighted by the number of true instances for each label. This alters 'macro' to account for label imbalance.
  - `F1 Macro` - Measures the F1 score, but calculates metrics for each label, and finds their unweighted mean. This is recommended for imbalanced problems.
  - `F1 Micro` - Measures the F1 metrics globally by counting the total true positives, false negatives, and false positives.
  - `Balanced Accuracy` - Measures the proportion correct of each class individually and then averages those values. This is a good metric to use for imbalanced problems.
  - `Log Loss` - Measures the prediction probability of each output and how closely that maps to the actual label. In binary classification, if the actual label was 0 and the prediction probability was 0.01, the prediction would be 0.49 better than a prediction probility of 0.5. This is a very harsh penalty mechanism and will result in a model that tries to find a very defined boundary between classes.
  - `Precision Macro` - Measures precision for each label and finds their unweighted mean. This is recommended for imbalanced problems.
  - `Precision Micro` - Measures the precision metrics globally by counting the total true positives predicted.
  - `Precision Weighted` - Measures the precision score for each label, and then finds their average weighted by the number of true instances for each label. This alters 'macro' to account for label imbalance.
  - `Recall Macro` - Measures recall for each label and finds their unweighted mean. This is recommended for imbalanced problems.

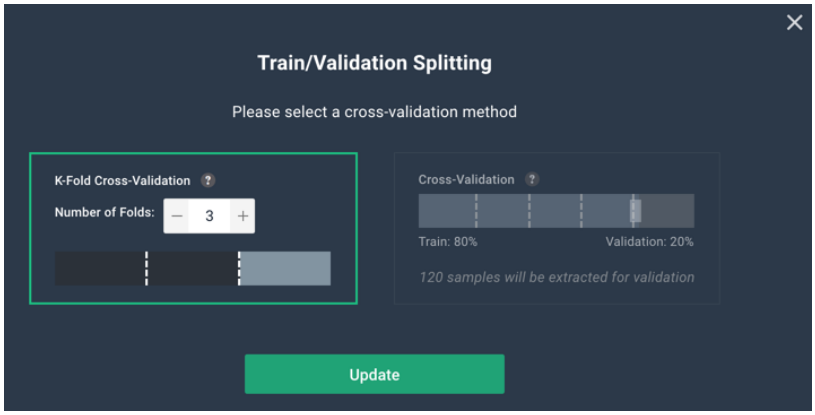
- `Recall Micro` - Measures the recall metrics globally by counting the total true positives predicted.
- `Recall Weighted` - Measures the recall score for each label and finds their average weighted by the number of true instances for each label. This alters 'macro' to account for label imbalance.

For regression problems, possible values include:

- `R2` - (default) Measures how closely the data maps to the fitted regression line. It is also known as the coefficient of determination and is useful for mapping the relationships that exist in data.
- `Mean Absolute Error` - Measures the average error for each predicted data point versus the expected value. This is useful as a good baseline metric or for capturing general trends.
- `Mean Squared Error` - Measures the square of the average error for each predicted data point versus the expected value. This is useful if you want to penalize large errors more harshly.
- `Median Absolute Error` - Measures the median error for the predicted data point versus the expected value. This is useful if your dataset has biases toward certain values.
- `Root Mean Squared Error` - Measures the square root Mean Squared Error values. This is useful if there are not a lot of outliers in your data.

- **Train/Validation Splitting** - Allows you to specify the cross-validation of the dataset during model building. There are two possible settings: a k-fold cross-validation or a percentage-based cross-validation. The default is `K-Fold Cross-Validation`.

  To change this value, click **Change** which will open the Train/Validation Splitting dialog.

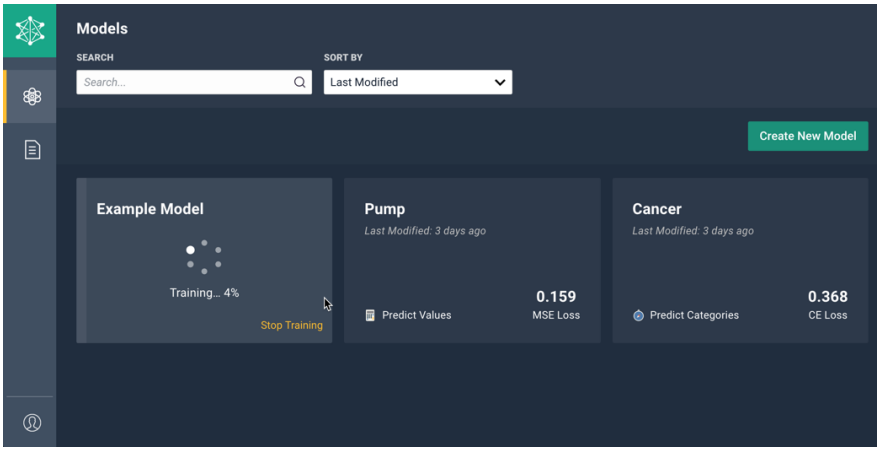


  Select your cross-validation method and adjust the input by either increasing or decreasing the number of folds or if using a percentage split, click-and-drag the slider bar to indicate the percentage split.

  Click **Update** to save your settings.

10. Once your parameters are set, click **Start Building** to begin building the model. The Models page appears with the current model showing a progress complete amount. When the model finishes training it will show a cross-entropy (CE) Loss value for classification problems or a scaled mean-squared error (MSE) Loss value for regression problems on the model's card.

    If you decide that you do not want to continue training the model, click **Stop Training**. The model will continue to finish the generation it is on and then will stop training the model.

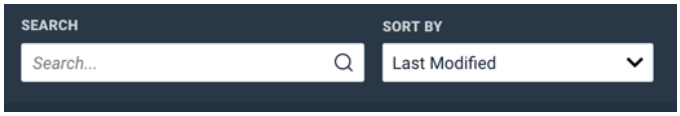


Working with models

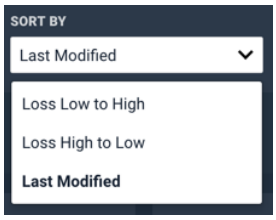
Once you have finished creating new models, a card appears on the **Models** page. Each model card displays the name, last modified time, the type of problem it is solving, and a Loss value. Clicking the card displays the training results and feature importances and allows you to run a testing dataset to generate predictions.

Searching for and sorting models

Once you have created many models, you can quickly perform a keyword search to find them or you can sort the models by a variety of criteria. To perform a keyword search, enter the keyword in the search field. The display will hide any models that do not match the entered keyword. The use of wildcard characters is not supported.



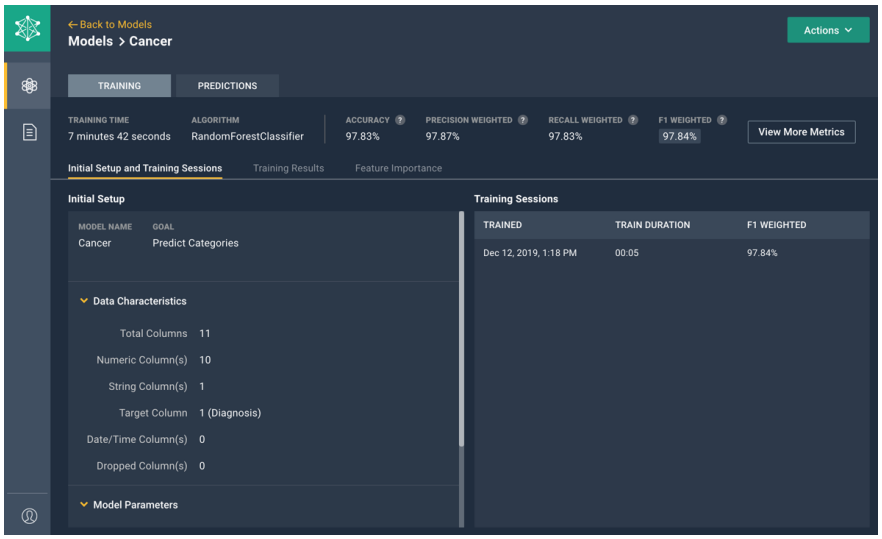
You can sort your models by a variety of criteria including loss or last modified. Click the drop-down list and make your selection. The default sort order is by Last Modified model.



Viewing training sessions and results

Click the model card to display the details screen where you can see training information and create predictions based on the model. The following sections are divided into viewing classification models and viewing regression models. The information presented focuses on viewing training sessions, results, and feature importance.

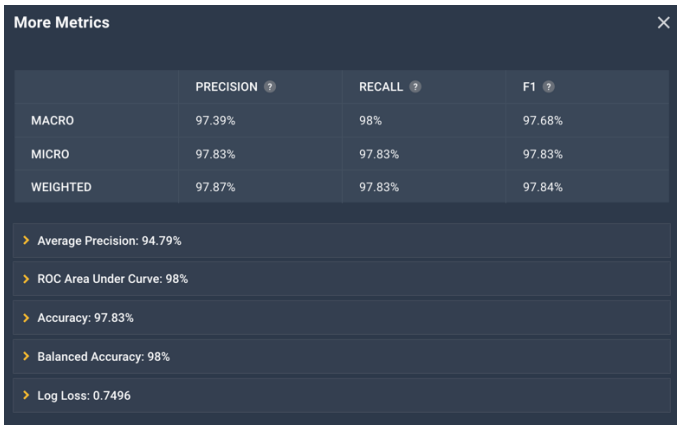
Viewing classification models



For this classification model, you can see the training time, the algorithm used to train the model and a variety of statistics including: accuracy and weighted scores for precision, recall, and F1.

- **Accuracy**: number of correct predictions divided by the total number of predictions. This metric can be misleading for imbalanced problems (many more examples of one class versus the others).
- **Precision weighted**: the percentage of how many predictions were made correctly. For example, a photograph contains 12 dogs and some cats. A program for recognizing dogs identifies 10 animals as dogs. In fact, only 8 of the selected animals are dogs, the other 2 are cats. The precision in this case is 8/10 or 0.80. So precision can be thought of as how useful the results are.
- **Recall weighted**: the percentage of total relevant results predicted correctly. For example, a photograph contains 12 dogs and some cats. A program for recognizing dogs identifies 10 animals as dogs. In fact, only 8 of the selected animals are dogs, the other 2 are cats. The recall in this case is 8/12 or 0.66, since the program identified 8 of the 12 dogs. So recall can be thought of as how complete the results are.
- **F1 weighted**: a simple score designed to gather all of the precision and recall scores and balance them into one metric. This score is a better measure to use than accuracy for imbalanced problems.

Click **View More Metrics** to see the metrics for the macro, micro, and weighted scores as well as the results of the other fitness functions available.

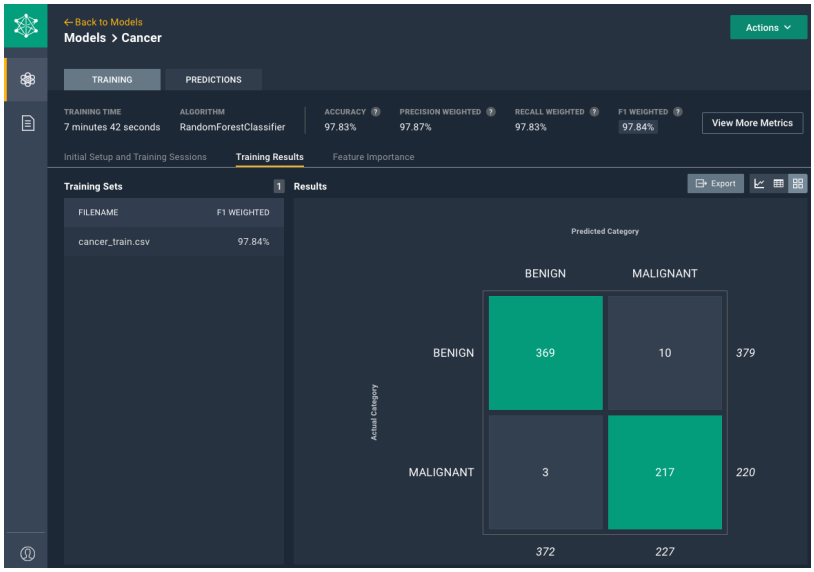


In addition, you can see the initial setup and information about the training sessions. Click the **Data**

Characteristics link to see information about the data and/or click the **Model Parameters** link to see the parameters that were used to create the model.

Viewing training results

Click **Training Results** to view the training sets used and the results.

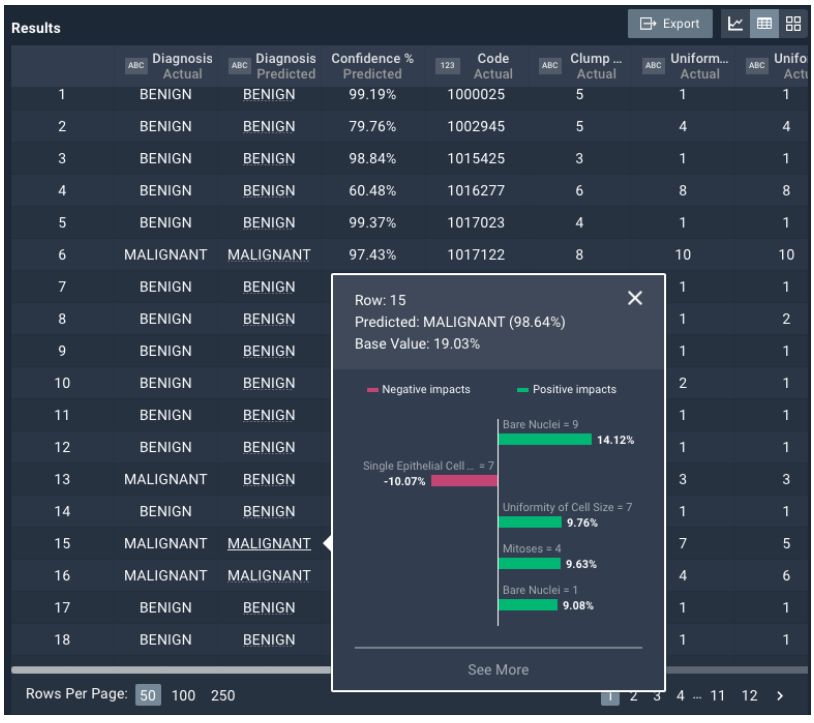


You can display the results as either a graph, a table, or a confusion matrix (default). Click the appropriate icon to switch the results display. In addition, you can export the results by clicking the **Export** button.

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. The green shaded cells show where the prediction and the actual value match. The other cells show false positives and false negatives.

Click the **Graph** icon to view the results in a graphical format. In the following image, you can see that the prediction matched the acutal value most every time. The cursor is highlighting one instance where the prediction was malignant, but the actual value was benign. Fifty visible points are displayed by default. You can adjust the display by selecting more visible points per screen and you can scroll through the pages, by clicking the page number.

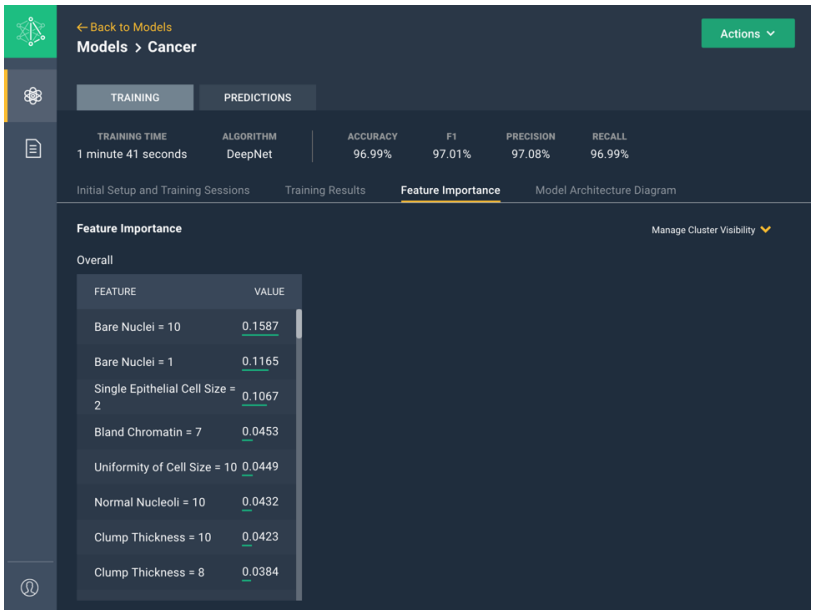
Click the **Table** icon to view the results in a tabular format. The following image shows the tabular view of the results. You can click a predicted value and see the Shapley feature importance, which tries to calculate why Darwin made that particular prediction. For row 15, Darwin predicted `MALIGNANT` with a 98.64% confidence. The base value of 19.03% represents the chance of randomly selecting an actual malignant result. The features that had either a positive or negative impact to that base value are presented to show how Darwin arrived at the final confidence value.



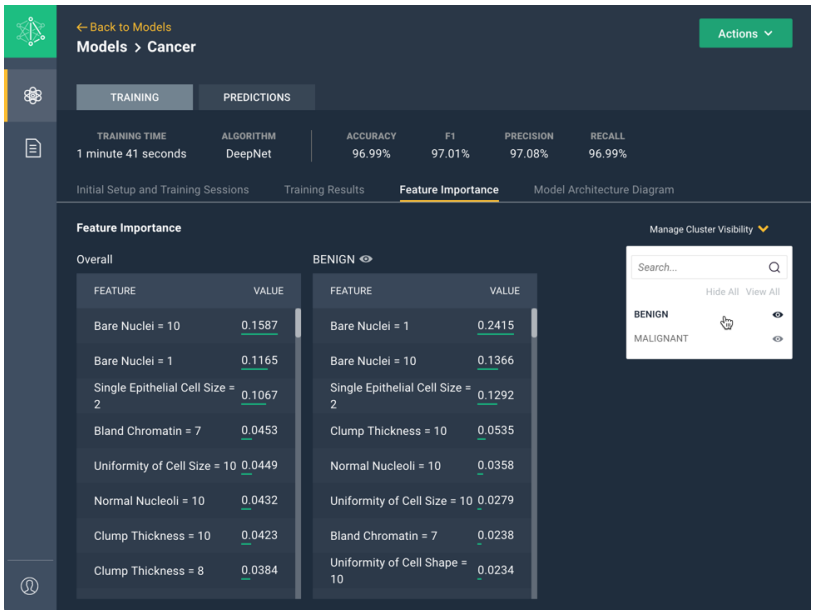
Viewing feature importance

Click **Feature Importance** to view the feature importance values. The decimal value represents how much this feature figures into the total calculation, which helps you understand the main contributing

factors that define what you are trying to predict.

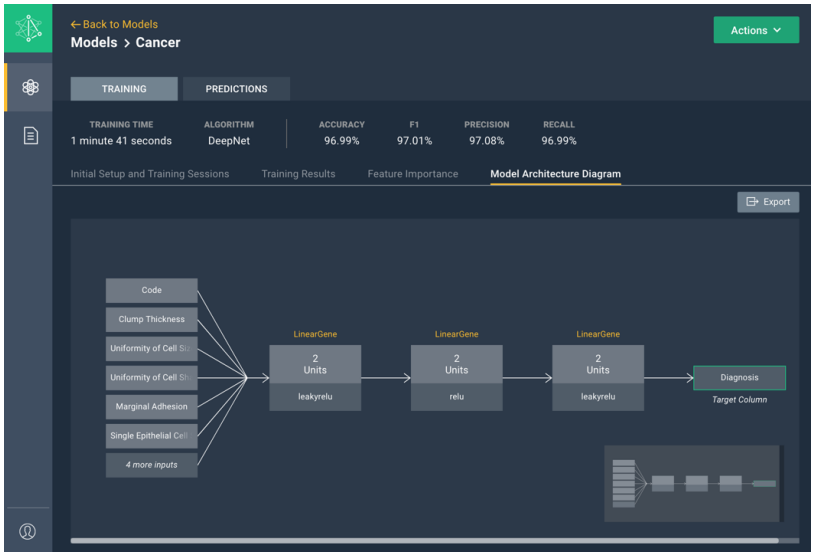


Click **Manage Cluster Visibility** to view the individual clusters from the target column. When you click on a cluster, its feature importance table is displayed next to the Overall feature importance table. The following figure shows the **BENIGN** cluster's feature importance table.



Viewing model architecture diagram

For DeepNet models, click **Model Architecture Diagram** to view the model architecture. This diagram shows the input columns that were not dropped, then the layers of the neural network showing the number of units (neurons) in that layer and the activation function used, and finally the target column.



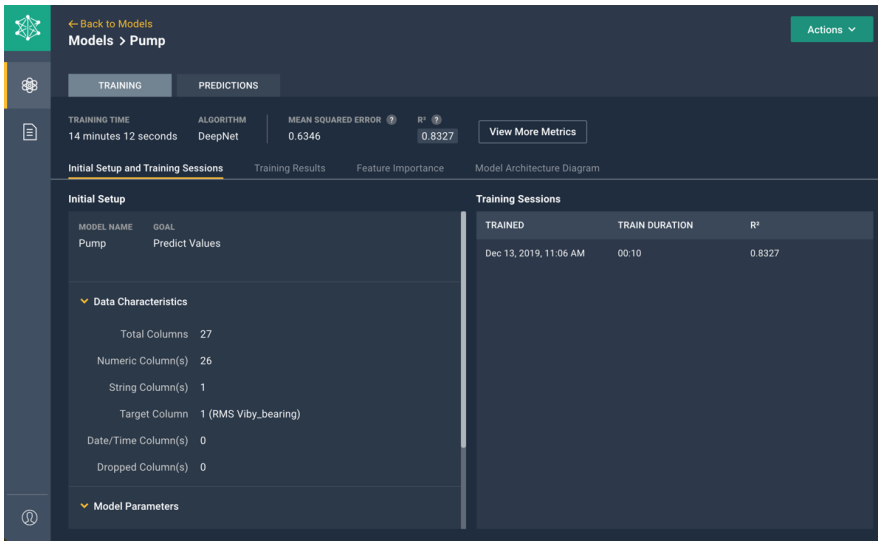
There are a variety of different types of layers and possible activation functions used in the neural network. The following list shows the different possible layer types and their associated potential activation functions:

- **LinearGene/LinearLayer**: Linear layers are single layers of linear neurons. They may be static or dynamic and can be trained on simple linear time series problems, but often are used adaptively to continue learning while deployed so they can adjust to changes in the relationship between inputs and outputs while being used. Normal Behavioral Model (NBM) activation functions include `sigmoid`, `tanh`, `gauss`, `relu`, `identity`. Non-NBM activation functions include `relu`, `identity`, `selu`, `hardtanh`, `leakyrelu`.
- **LSTMGene/LSTMLayer**: Long Short-Term Memory Units (LSTMs) are a type of recurrent neural network designed to recognize patterns in sequences of data, such as text, genomes, handwriting, spoken work, or numerical time series data. Activation functions include `relu`, `identity`, `selu`, `hardtanh`, `leakyrelu`.
- **TemporalConvolutionalNet**: Convolutions are one-dimensional filters (think of a list of numbers) that are trained to recognize short patterns over time. Temporal indicates that there is no information leakage from future to past, that is the filters at any given time cannot be applied to future data. The networks consist of several layers of these one-dimensional convolutions stacked in such a way that subsequent layers extract information at increasingly larger time scales. Activation function is always `identity`.

Click **Export** to export the diagram as a .png file.

Note: Using an XGBoost or RandomForest algorithm does not currently display a Model Architecture Diagram.

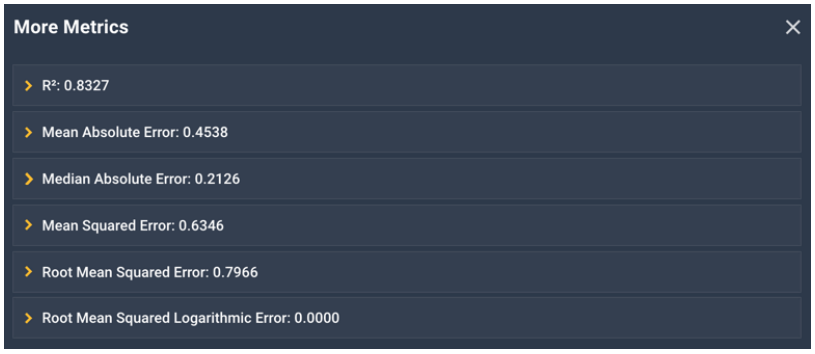
Viewing regression models



For this regression model, you can see the training time, the algorithm used to train the model and a variety of statistics including MSE and R^2.

- **Mean Squared Error (MSE)**: tells you how close a regression line is to a set of actual data points. This is done by taking the distances from the actual points to the regression line and squaring them (to get rid of negative signs). Then, the average is taken from all these errors arriving at the true mean squared error value.
- **R^2 Score**: also known as the "coefficient of determination", this score is a statistical measure of how much a target variable's variance is predicted by the created model. R^2 scores are between 0 and 1. A score of 1 means that the model perfectly predicts the target variance. A score of 0.93 indicates that 93% of variations in the target variable are explained by the independent variables present in the model.

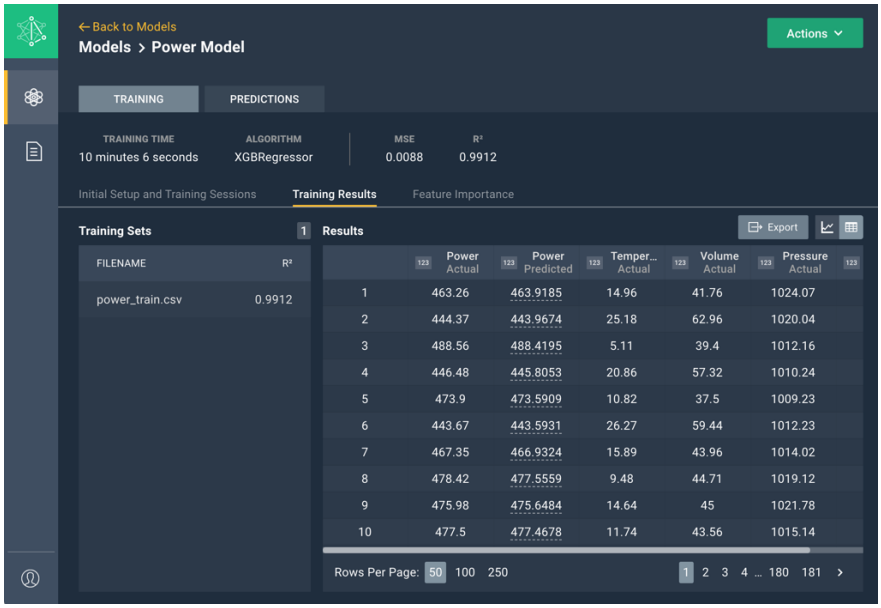
Click **View More Metrics** to see the metrics for the other fitness functions available.



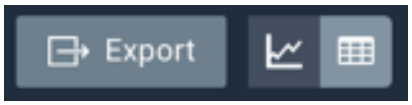
In addition, you can see the initial setup and information about the training sessions. Click the **Data Characteristics** link to see information about the data and/or click the **Model Parameters** link to see the parameters that were used to create the model.

Viewing training results

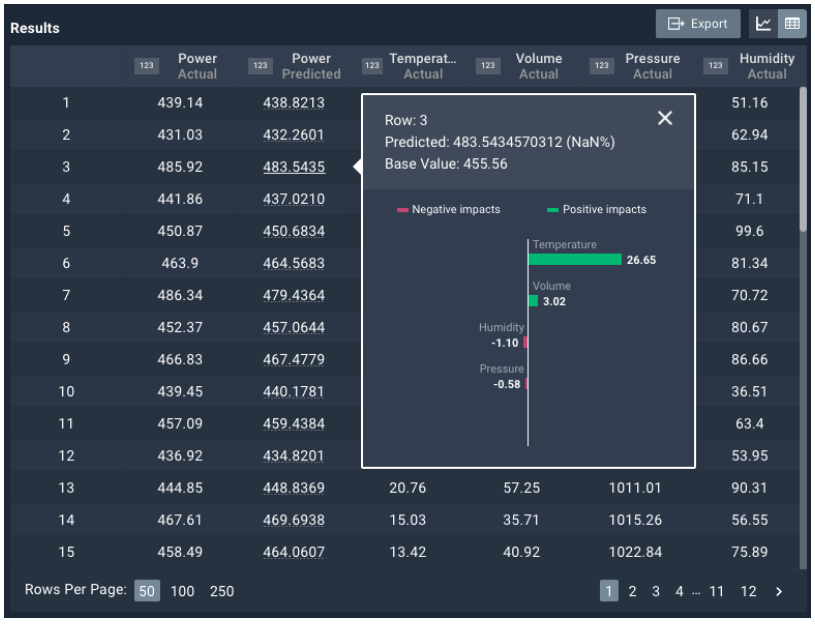
Click **Training Results** to view the training sets used and the results.



You can display the results as either a graph or a table (default). Click the appropriate icon to switch the results display. In addition, you can export the results by clicking the **Export** button.

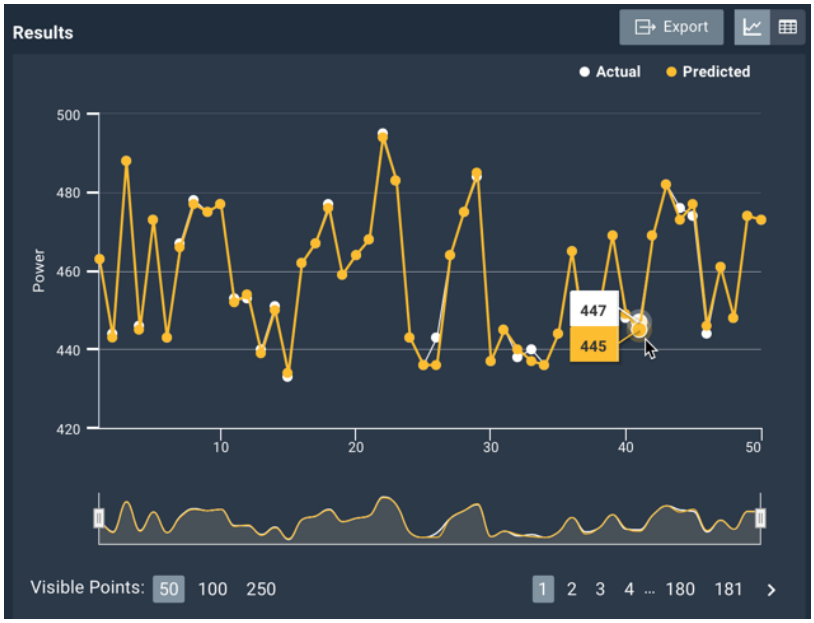


Click the **Table** icon to view the results in a tabular format. The following image shows the tabular view of the results. You can click a predicted value and see the Shapley feature importance, which tries to calculate why Darwin made that particular prediction. For row 3, Darwin predicted a Power value of 483.5435. The base value was 455.56, which is the average value of all the predictions in the dataset used. The display shows the feature importances that had a positive or negative impact to the base value and are presented to show how Darwin arrived at the final predicted value.



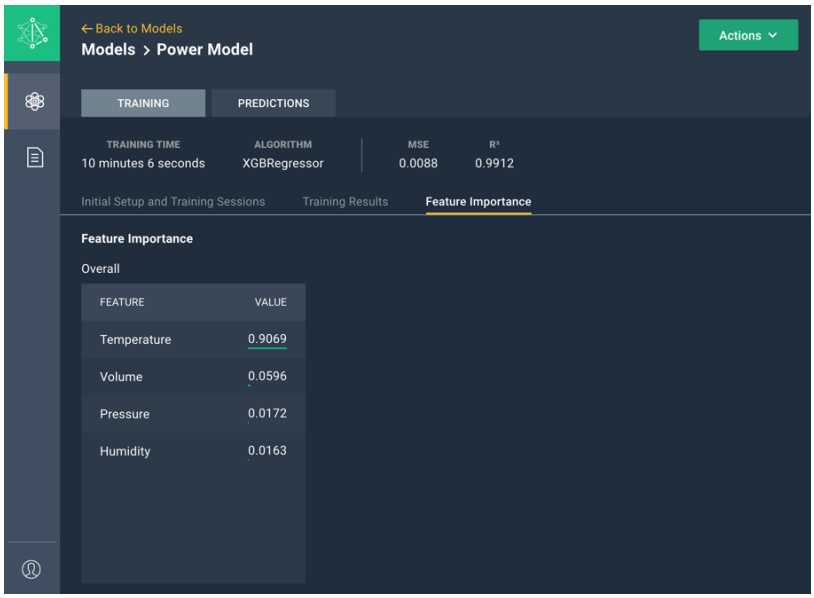
Click the **Graph** icon to view the results in a graphical format. In the following image, you can see that the predicted values are very close to the actual value. The cursor is highlighting a prediction of 445

compared to the actual value of 447. Fifty visible points are displayed by default. You can adjust the display by selecting more visible points per screen and you can scroll through the pages, by clicking the page number.



Viewing feature importance

Click **Feature Importance** to view the feature importance values. The decimal value represents how much this feature figures into the total calculation.

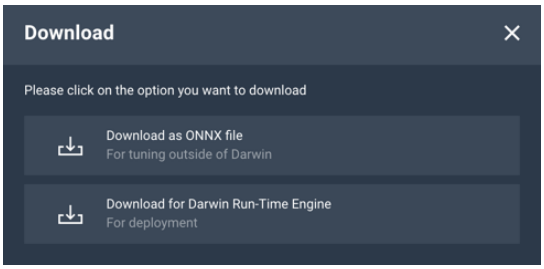


Downloading a model

After training a model, you may want to download it to use the model in the Darwin Run-Time Engine or in another program for tuning outside of Darwin.

To download a model, perform the following:

1. Click the model card of the model you want to download.
2. Click the **Actions** button and select **Download**.

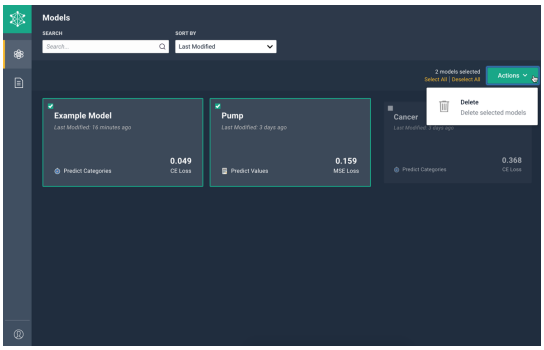


3. Click the type of download you want, either an ONNX file or as a JSON file formatted for use in the Darwin Run-Time Engine. The download consists of a `.zip` file that contains the model and the data profiler. The ONNX file format is only available for DeepNet models.

Deleting models

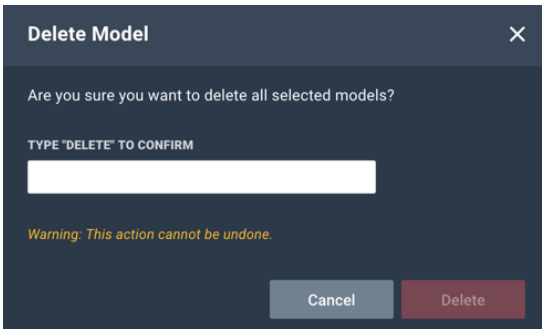
If you no longer need a model, you can delete it directly from the Models page. Perform the following:

1. Navigate to the Models page, and hover over the card containing the model you want to delete.
2. Select the check box in the upper-left corner. You can also select other models to delete by selecting the check box on the other cards.
3. Click **Actions > Delete**. The Delete Model confirmation dialog appears.



4. Type the word **DELETE** to confirm.

   Note: This action cannot be undone. Be sure you want to delete the model.

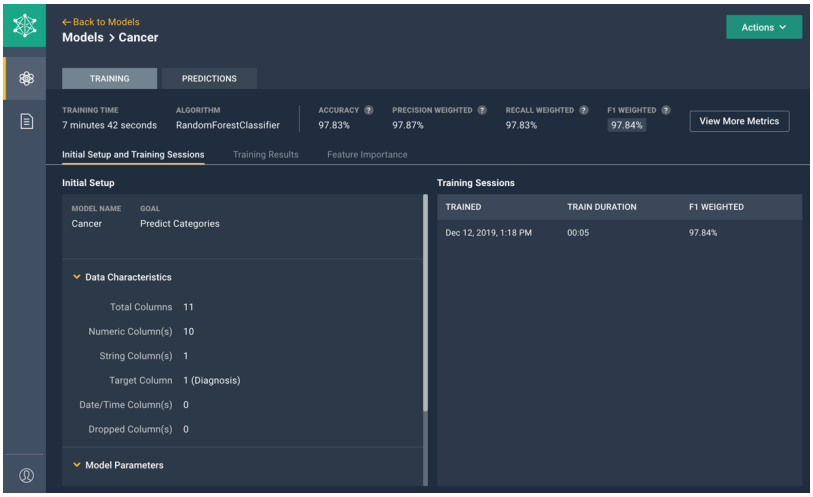


5. Click **Delete**.

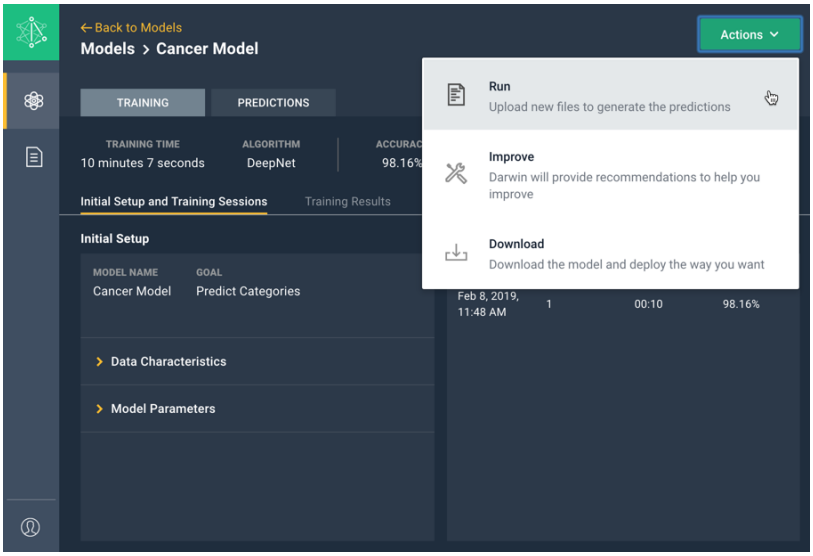
Generating predictions

Once a model has been trained, you can upload a new dataset and generate predictions using that model.

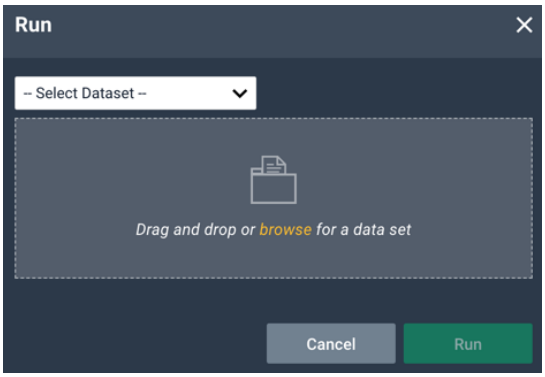
1. Navigate to the detail page of the model you want to use.



2. In the upper-right corner, click the **Actions** button and select **Run**.



3. The Run dialog appears.



4. If you have previously uploaded the testing dataset, select it from the drop-down list, otherwise drag

and drop it or browse for it. Then, click **Run**.

5. On the Predictions tab of the page, click the name of the run you want to see. The results are displayed in a table format. You can change the results display to a graph or a confusion matrix (classification only) or you can export the results by clicking the **Export** button.

Note: Odd behavior might be encounted if you try to predict categories on a numeric target or predict values on a categorical target. Ensure that the type of prediction you want is in alignment with the type of target.

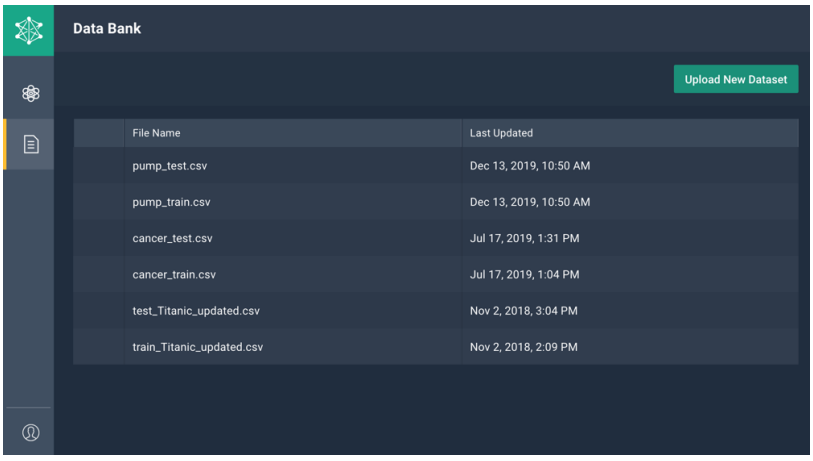
Working with the Data Bank

The data bank offers a centralized location where you can work with your datasets. You can upload, preview, download, and delete datasets from one window. You can still upload datasets as part of the creating a model workflow that is described in Creating a new model, but if you'd like to upload multiple datasets or view all the datasets in the system, the Data Bank is the place to perform those functions.

Uploading a new dataset

To upload a new dataset using the Data Bank, perform the following:

1. Click the Data Bank icon to open the Data Bank page.

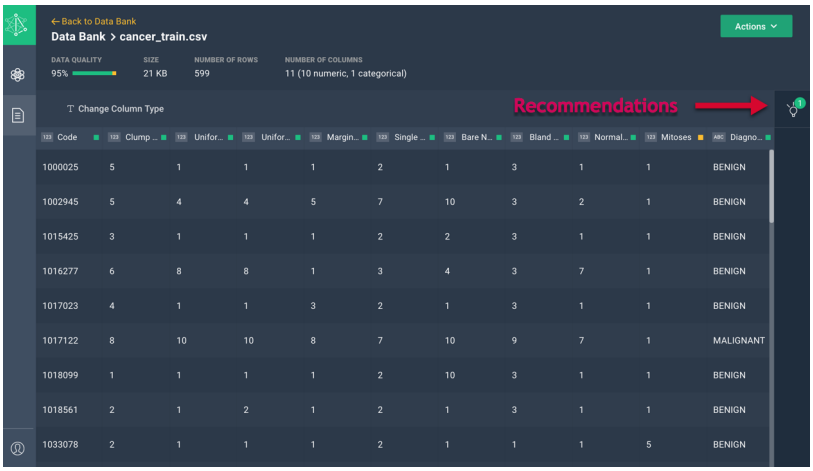


2. Click the **Upload New Dataset** button.
3. Drag and drop or browse for the dataset you want to upload. The dataset will appear in the Data Bank window and can be used in other creating or training model workflows. Datasets are listed by last updated time by default. You can change the sort order by clicking either the file name header (alphabetical sort) or the last updated header.

Previewing a dataset

To preview a dataset, perform the following:

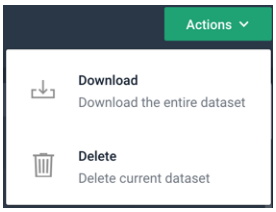
1. Click the Data Bank icon to open the Data Bank page.
2. Click on the row of the dataset you want to preview. The first 50 rows of the dataset will be displayed along with various metrics about the dataset including quality, size, number of rows, and number of columns.



3. Click the **Recommendations** icon on the right side of the window to view suggested recommendations. These recommendations can be marked as yellow, indicating processing methods that will be applied, or possibly marked red, indicating a column that will not be considered. These recommendations can indicate problems with low statistical variance, too many missing values, or incorrect data types. Click on a recommendation to be guided about making an adjustment.

   You can also make multiple changes to the column type based on this analysis. To do this, click the **Change Column Type** button to display all the columns. You can then change a column to a different column type by selecting the appropriate drop-down list and clicking **Submit**.

4. Click on a column header to sort the preview by that particular column.
5. (Optional) Click the **Actions** button to download or delete the dataset.



Downloading a dataset

To download a dataset, perform the following:

1. Click the Data Bank icon to open the Data Bank page.
2. On the row of the dataset you want to download, click the Popup menu (3 vertical dots).
3. Select **Download**. The dataset will be downloaded to the Downloads folder of your computer.

Deleting datasets

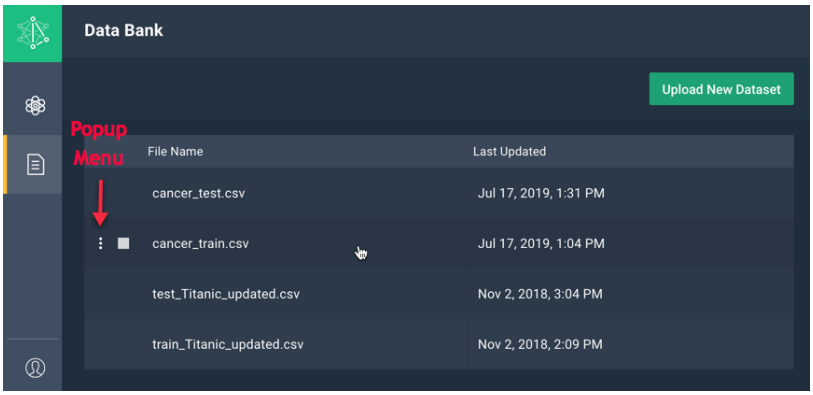
Refer to the following instructions for deleting datasets through the Data Bank:

- Deleting a single dataset
- Deleting multiple datasets

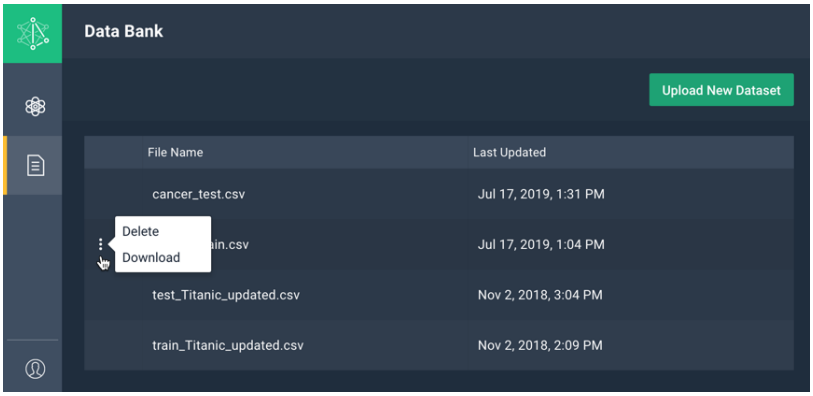
Deleting a single dataset

To delete a single dataset, perform the following:

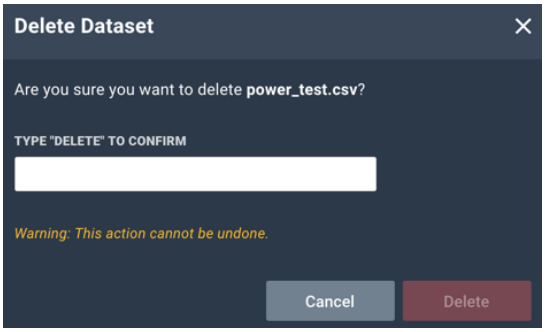
1. Click the Data Bank icon to open the Data Bank page.
2. Scroll over the row of the dataset you want to delete. A Popup menu icon and a select checkbox are displayed.



3. Click the Popup menu (3 vertical dots).



4. Select **Delete**. The Delete Dataset confirmation dialog appears.



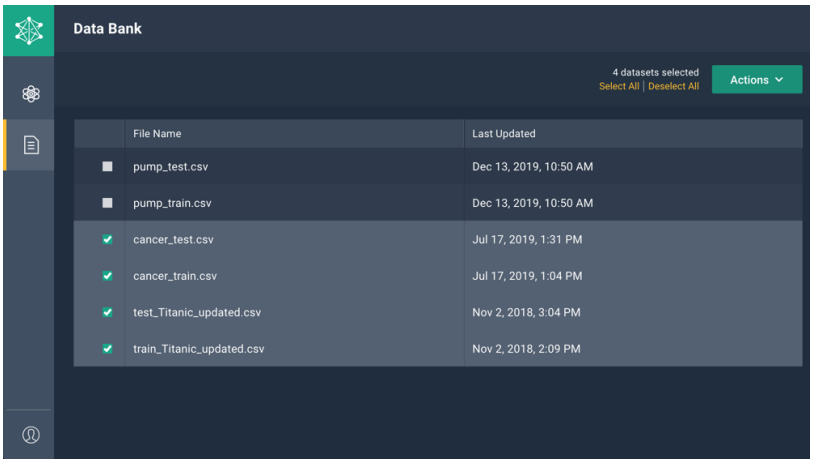
5. Type the word **DELETE** to confirm.

   Note: This action cannot be undone. Be sure you want to delete the dataset.

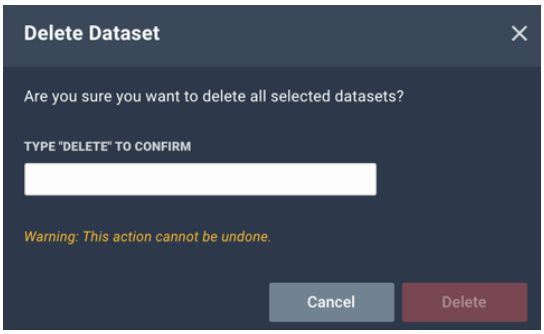
6. Click **Delete**.

Deleting multiple datasets

1. Click the Data Bank icon to open the Data Bank page.
2. Scroll over the row of one of the datasets you want to delete and click the check box. All of the check boxes appear along with an Actions drop-down list.



3. Continue selecting datasets to delete. Optionally, you can click the **Select All** link to quickly select all the datasets.
4. Click the **Actions** drop-down and select **Delete**. The Delete Dataset confirmation dialog appears.



5. Type the word **DELETE** to confirm.

   Note: This action cannot be undone. Be sure you want to delete the dataset.

6. Click **Delete**.

Contact Support

The following methods enable you to research issues, create a support ticket, or contact SparkCognition:

- Use the Darwin support portal - Read Frequently Asked Questions (FAQ), download documentation, or log your issue.

- **Email Support** - Send email to darwin_support@sparkcognition.com.
- **Phone Support** - The SparkCognition support line is +1-512-400-2001.

Revision Table

Version	Date	Notes
v 2.0	29-Jul-2019	• Added system recommendations • New model building wizard • New data quality check and recommendations • Ability to change column type • Select train/validation split for cross-validation
v 2.0.1		Internal Use Only
v 2.0.2	04-Sep-2019	• Multi-select of datasets in Data Bank • Improvements to Data Quality checks • Improvements to Data Explorer functionality • Model sorting improvements
v 2.0.3		RTE Release Only, no UI changes
v 2.0.4	19-Dec-2019	• Darwin automatically determines problem type. That step has been removed from the Create Model workflow. • New fitness functions introduced • Additional training metrics provided