SparkCognition™ Darwin™ Release Notes

v 2.0.4 - 12.19.2019

Darwin Release Notes v 2.0.4

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 Run Time Engine User Guide*, version 2.0.4

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

New Features introduced in version 2.0.4

- Streamlining of model creation workflow. Users no longer have to select the type of problem they are solving. Darwin determines this based on the dataset used and the selected target column.
- All the additional metrics that Darwin calculated are displayed on the Model Results page.
- Additional fitness metrics available for model creation.
- More informational dialogs to indicate how Darwin will build a model and indicating any issues found with the target column.

- Support for more than 30 categories in a column now. However, if 50% or more of the values are unique, a column will be dropped.
- Improvements made to anomaly detection pipeline.
- Model type can be specified in the SDK/API if a user wants to run or analyze a particular model.
- Cleaning process can be specified in the SDK/API.
- Passwords are now PCI compliant and for on-prem installations, the Darwin admin can change passwords.
- Resume training on new data will recompute the loss on the new dataset so that it has a new baseline to learn from.
- Users are notified if columns have a pending datatype conversion.

New Features introduced in version 2.0.3

- For on-prem installations, the API timeout has been increased to improve the performance in high latency situations.
- Changes were made to ensure that the whole dataset is used for running a model and not just a particular chunk of the dataset.

New Features introduced in version 2.0.2

- Multi-select of datasets in the Data Bank
- Improvements to the Data Explorer functionality
- Improvements to data ingestion to detect malformed .csv files

New Features introduced in version 2.0

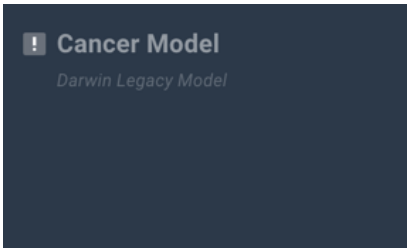
- New data analysis and recommendation service providing cleaning and analysis of datasets as they are uploaded as well as a guided UI workflow for creating models
- New Forecasting Models accessible through the SDK
- Manual and Automatic Train/Validation splitting
- Overfitting improvements
- Support links and a feedback form available from the UI
- Improved UI Error messaging
- UTC time is used for SDK/API; local time is used for UI

Fixed Issues in 2.0.4

- Fixed issues that resulted in algorithm type and fitness results not appearing.
- Fixed numerous UI issues that resulted in "Mismatch Error".
- Darwin no longer fails silently when a test set's columns do not match the training set.
- Feature importances now accurately reflect what is happening in a model.
- Fixed a date reporting issue on model cards.

Known Issues in 2.0.x

- Models created in version 1.x are not compatible with version 2.0.x. Those models will display an error similar to the following and will need to be re-created.



- Date/Time columns are currently dropped. If you want to create a time series problem, you must perform one of the following:
  - In the SDK, set `recurrent=True`, or
  - In the UI, select `Yes` on the Time Series problem page.

  In both of these cases, Darwin will automatically perform nested cross-validation, use forward fill imputation, and turn on recurrent architectures to treat the problem as a time series problem.
- The maximum number of features (columns) that can be displayed in the UI is 1000. For best usability, ensure that your datasets have less than 1000 features. However, the maximum number of features that can be used in the SDK is 4000.
- Darwin will not convert a numerical column into a categorical column when integers are expressed as floating point values (e.g., 3.0). The workaround is to remove the ".0" at the end of the numerical value so that Darwin interprets them as integers, which will then be able to be converted into a categorial column.
- In the SDK, the auth_login_user() method does not handle upper/lower case for the username.
- The Clean Data functionality is not able to clean datasets with unseen categories for categorical columns.
- The RTE cannot run a dataset with a space in its path.
- A single row dataset cannot be processed through the user interface.
- Date/Time columns will be displayed incorrectly in the Model Results page. In these cases, the date columns will be shown as categorical.
- The Mean Squared Error (MSE) loss value presented in the UI is the scaled MSE loss, whereas the MSE presented in the Model Results view is the unscaled MSE loss.
- When uploading a dataset, ensure that you wait until the upload is complete. Other UI functionality may not be available while the dataset upload is in progress.
- When downloading large datasets, you are not notified until the download is complete. Do not log out of Darwin until the download is complete. (Download times may vary depending on bandwidth.)
- When downloading an artifact using the Runtime Engine (RTE), it is not being downloaded to the user-defined path. The RTE is saving the artifact in a temporary folder on the local machine. The download confirmation will output the temp folder path.

- When exporting a model, the ONNX format is only available for neural network models. The JSON format is available for all model downloads, including neural network models.
- The Darwin RTE does not support unsupervised anomaly detection nor models with TCN architectures. It only supports supervised and NBM models.
- Analyze predictions is not supported for large datasets (> 500 MB).

General Notes

- Re-training or resuming training on a model should be done with the original dataset, since a different dataset may not have the same categories for each feature as the original dataset.
- Any created models can only specify either *zero* or a single *Target* column.
- Because Darwin cannot one hot encode categorical columns with more than *max_unique_values* in training and test sets (set to 30), these columns are dropped in test and training sets.
- Darwin only drops duplicated columns in data sets with less than 5000 rows.
- Users must now set `recurrent = True` with the SDK and API in order to see the LSTM and TCN models used.
- The Darwin UI may not show data when header names contain more than 63 characters. Model building will operate as expected, but components of the visualization may not appear as designed.
- 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.
- Any data set can only have a single (one) date time column or be indexed by date/time, otherwise an error message is returned.

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
v 1.6	01.16.2019
v 1.6.1	02.06.2019
v 1.6.2	03.25.2019
v 1.7	05.16.2019
v 2.0	07.29.2019
v 2.0.1	Internal use only
v 2.0.2	09.04.2019

Version	Date
v 2.0.2.1	10.14.2019
v 2.0.3	RTE release only
v 2.0.4	12.19.2019