SparkCognition™ Darwin™ Release Notes

v 1.7 - 05.16.2019

Darwin Release Notes v 1.7

Darwin release 1.7 has incorporated customer feedback to provide improvements in the speed and accuracy of model building as well as providing a streamlined deployment technology that enables more efficient scaling of Darwin algorithms. The following changes are completed and rolled into the Darwin release 1.7 for immediate use:

New Features in 1.7

- Optimized Darwin GPU usage to parallelize model training, resulting in 3-5x speed improvements
- Modified genetic algorithm seeding strategy to favor smaller parameter sizes and prevent overfitting
- Added a new core component to the user interface called the Data Bank, which allows for an exploration of datasets in the UI
- Implemented streamlined deployment technology that enables Darwin algorithms to more efficiently scale
- The SDK and API now have independent version numbers to allow for release outside of the normal Darwin product release window. As of these release notes, the Darwin API is at version 1.33.0 and the Darwin SDK is at version 1.43.0

Fixed Issues in 1.7

- Fixed an issue in which the SHAP predicted_proba was often incorrect for regression
- Resolved an intermittent issue in XGBoost parameters to achieve slightly higher model fitnesses
- Canceling a model through the UI no longer gets stuck in a Stopping state
- Fixed an issue where the UI showed a model being finished with training when it was not actually done
- The start_index and end_index of analyze_predictions was not working in SDK
- The "Download Dataset" function was not handling multi-part cleaned datasets correctly for big data
- Fixed an issue where the RTE installation was failing on Windows machines
- Fixed an issue where some parameters were not saved after resuming model training
- The fitness function exposed in the API was not getting used
- In on-prem instances, models built in the UI never showed progress of more than 0% even though they were building properly

Known Issues in 1.7

- If the best genome was an sklearn genome after *create_model* or *POST /train/model*, calling *resume_training* or *PATCH /train/model/{model_name}* will result in an error. To work around this issue, you need to re-run *create_model* or *POST /train/model* for the original training time plus the amount of time you wanted to resume training for.
- 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.
- There is a discrepancy between the MSE values returned in the Darwin UI compared to sklearn methods. In the Darwin UI, MSE is computed using the transformed (scaled) values rather than those in the original domain.
- For on-prem instances, the maximum size of datasets for use in the SDK/API is 500 MB and the maximum size of datasets for use in the UI is 50 MB. These limitations will be improved in a future release.
- 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 models nor models with TCN architectures. It only supports supervised and NBM models.
- Analyze predictions is not supported for big datasets (> 500 MB).
- Non-target columns that begin with the same prefix as the target are considered to be the target. To avoid this issue, ensure that the columns do not begin with the same prefix as the target. For example, if your target column is "temperature", ensure that no other columns begin with the string "temperature". A column named "temperature_change" would be interpreted as the target and cause this issue.
- When large datasets are cleaned, they are divided into 1 GB parts. There is an issue where the download dataset function will only download part 1 of a multi-part cleansed dataset. This does not affect the training or running of models using large datasets.
- Dataset names cannot contain Japanese characters. There is a known issue with the way Darwin works with filenames that contain characters outside of the Latin-1/ISO-8859-1 character set. This

does not affect model building or training, but when viewing the results of a model, the user interface will report an error and the Data Characteristics and Model Architecture Diagram will be unavailable. This issue will be fixed in a future release.

- If a dataset column header contains a period/full-stop/decimal character (.), an open bracket ([), or a closed bracket (]), it will not display correctly in the user interface. This does not affect the model building or training process, just the display in the user interface. This issue will be fixed in a future release.
- Data submitted to *run_model* must have the same number of columns and column headers as data submitted to *create_model*, otherwise an error message is returned.

General Notes

- Darwin will split the training set into a train and validation set using an 80/20 split by default:
  - For classification problems, the split will be created using stratified shuffling.
  - For regression problems, the split will be created using random shuffling.
  - For problems with a timestamp (regression or classification problems), no reordering will be done and the last 20% of the input data will be used as validation data. So if sparse time-series data is used for modeling and the important points for predictions are clustered densely together, there is the potential that the resulting model may only train on non-useful data. If this issue is occurring, try removing the time stamp from the data set.

  You can change the size of the validation set by adjusting the `val_size` parameter in *create_model* in the SDK or in *POST /train/model* through the API.
- 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, these columns are dropped in test and training sets.
- If the target has more numeric values than the *max_int_unique* set point, the problem is treated as a regression and will use MSE.
- Darwin only drops duplicated columns in data sets with less than 5000 rows.
- Any data set can only have a single (one) date time column or be indexed by date/time, otherwise an error message is returned.
  Note: Affects *create_model, analyze_data*.

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 1.0	02.05.2018
v 1.1	02.22.2018
v 1.2	03.29.2018
v 1.3	05.23.2018
v 1.3.1	06.14.2018
v 1.4	07.31.2018
v 1.5	10.15.2018
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