SparkCognition™ Darwin™ Release Notes

v 1.6 - 01.16.2019

Darwin Release Notes v 1.6

Darwin release 1.6 has incorporated customer feedback to provide improvements in the speed and accuracy of model building as well as support for larger datasets when model building. The following changes are completed and rolled into the Darwin release 1.6 for immediate use:

New Features in 1.6

- Model building using larger datasets. The maximum file size is now 500 MB for unsupervised and Normal Behavioral Modeling (NBM) and 10 GB for supervised. As a result of this support, the cleaning of data step has been separated from the model creation step, which means that you must clean your data prior to model building.
- Export models to either JSON or ONNX format. **Note:** ONNX is only supported for supervised Deepnet models. Exporting to ONNX is not currently available through the UI.
- Users can view and download the Elites of the model population (top Deepnet, top Random Forest, or top XGBoost model)
- Time-series models completed in version 1.6 are more accurate than in previous versions due to the addition of temporal convolutional networks to the architecture search
- Data can be run successfully through downloaded models and inserted into the Darwin Runtime Engine (DRE)

- NBM models can be downloaded and run in the DRE
- Users can specify the Error function they care about fitting to when creating a model
- IsolationForest Outlier detection has been added for unsupervised anomaly detection
- A minimum recommended training time is now calculated based on an input dataset
- Users can force Darwin to treat certain problems as regression even if there are limited samples

Fixed Issues in 1.6

- Better error messaging for Darwin Internal Errors
- Display which columns contain DateTime values if there are multiple datasets
- Incorrect probabilities were being reported for XGBClassifier
- On the evaluation endpoint, the actual and predicted results were being doubled
- There was no messaging when a dataset contains multiple date/timestamp columns
- Spaces and some special characters can now be used in dataset and model names
- An exception was not being generated when there was only one member of a categorical target
- The analyze and clean dataproc jobs were getting stuck in the *Running* state

Known Issues in 1.6

- Updates to the Darwin Runtime Engine (DRE) will be delayed until a few weeks after the 1.6 release date.
- Analyze predictions is not supported for big datasets.
- If you create a model using the big data pipeline, you cannot clean test data whose size is less than the big data threshold of 500 MB. To workaround this issue, ensure that your test data size is greater than 500 MB.
- 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.
  Note: Affects *create_model, run_model*.
- Setting *recurrent=true* does not work for unsupervised.
  Note: Affects *create_model*.

General Notes

- Darwin will split the training set into a train and validation set using an 80/20 split:
  - 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.
- 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 support@darwinamb.zendesk.com.
- **Phone Support** - The SparkCognition support line is +1-512-956-5576.

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