SparkCognition™ Darwin™ Release Notes

v 1.4 - 07.31.2018

Darwin Release Notes v 1.4

Darwin release 1.4 has incorporated customer feedback to provide improvements in the speed and accuracy of model building. The following changes are completed and rolled into the Darwin release 1.4 for immediate use:

New Features in 1.4

- Island models for improved speed during model building. Genetic approach breaks novel model types into islands to speed up different training processes.
- Added GPUs to the cluster, which provides a huge speed improvement.
- User selectable loss functions
- For classification problems, Darwin now also outputs model confidence value along with the resultant class
- Specify download paths for artifacts
- Parameter validation has been added to the API server to avoid non-valid inputs causing errors
- The *lookup_model* call displays model description, which allows a user to see exactly how the model is constructed should a user want to replicate the model in another platform.

Fixed Issues in 1.4

- Fixed an issue where you could not use *run_model* on files that do not contain headers if the training file had headers.
- Better messaging added when *upload_dataset* does not find a data set.
- Fixed an issue where spaces in the model name cause a *run_model* failure.
- Fixed an issue where Darwin ceases to operate when loading certain data sets from a .csv file.
- You can no longer create a new model with the same name as an existing model. Model names are also now case-sensitive.
- Fixed issue where Target Columns containing non-alphabetic characters can cause problems in API calls.
- Fixed an issue where unsupervised model was not being saved properly.
- Interrupted training jobs with no checkpoint will now restart training rather than failing.
- Fixed an issue on having too many batches in one single epoch.
- Fixed various issues with retrieving jobs from the database.
- Fixed an issue where an extra softmax layer was added on some occasions when performing classification.

Known Issues in 1.4

- Models created from earlier versions of Darwin are incompatible with version 1.4. These models need to be re-created.
- 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*.
- 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.
- Darwin only drops duplicated columns in data sets with less than 5000 rows.
- When using *feature_eng*, the percent complete reads *0%* until the automated windowing, preprocessing, and feature engineering is complete.

  Notes:

  - Depending on data set size, this process can require substantial time to complete.
  - Affects *create_model*.
- 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*.
- *max_int_uniques* keyword parameter must be a natural number, otherwise an error message is returned.

Note: Affects *create_model, analyze_data*.

- *max_unique_values* keyword parameter must be a natural number, otherwise an error message is returned.
  Note: Affects *create_model, analyze_data*.
- *window_len* parameter (optional) to *create_model* must:
  1. be a natural number,
  2. be smaller than the size of any training set, and
  3. be smaller than the size of any test set, otherwise an error message is returned.

  Note: Affects *create_model, analyze_data*.
- *feature_select* parameter must be a real number in [0,1], otherwise an error message is returned.
  Note: Affects *create_model, analyze_data*.

General Notes

- Darwin will split the training set into a train and validation set using an 70/30 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 30% 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.

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