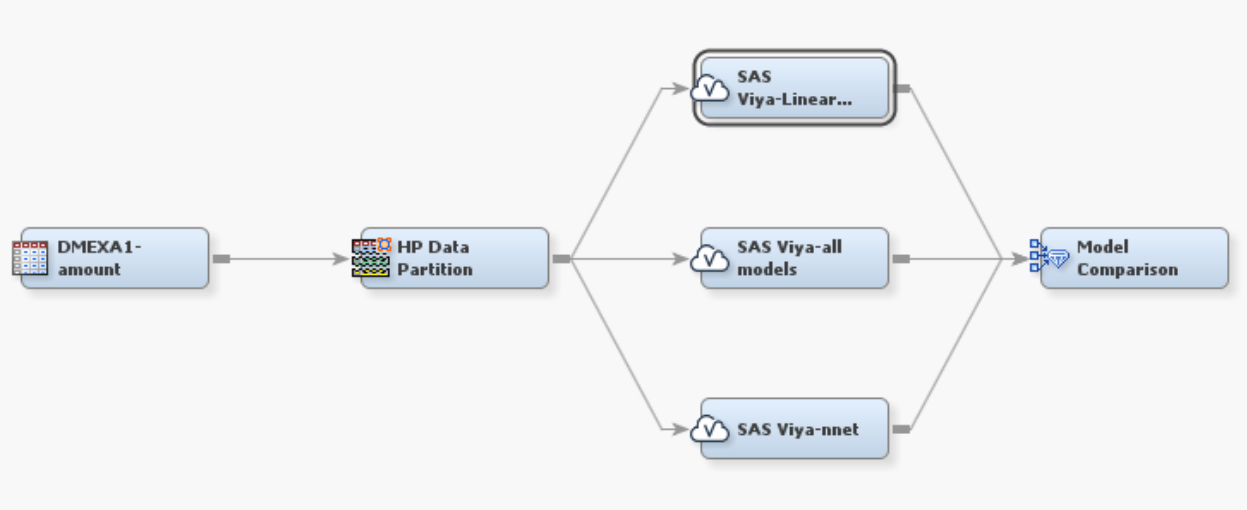
Parallel Training in SAS® Viya™ Using SAS® Enterprise Miner™

Process Flow Diagram:



Data:
The DMEXA1 data set in the SAMPSIO SAS library is used to create the data source. You can generate the data in SAS Enterprise Miner 14.3 by selecting **Help > Generate Sample Data Sources** to open the Generate Sample Data Sources window and then selecting **DM Example-1**. The target is set to an interval variable, **Amount**, which represents dollars spent by a customer.

Goal:
The goal is to run multiple models in parallel in SAS Viya and select the best ("champion") model. SAS Viya Code nodes (introduced in SAS Enterprise Miner 14.2) are used to run different models that use different SAS® Visual Data Mining and Machine Learning procedures. Finally, a Model Comparison node is used to select a champion model based on the specified criterion for model selection.

Flow:
The Input Data node (titled "DMEXA1-amount" in the flow diagram) is connected to the HP Data Partition node, which partitions the data into two parts: 70% training and 30% validation. The training data are used to build a model, and the validation data are used to avoid overfitting the final model.

Three SAS Viya Code nodes are set up to train in parallel:

- The "SAS Viya-Linear Regression" node trains a single linear regression model via the REGSELECT procedure, a statistical procedure in SAS Visual Data Mining and Machine Learning.

- The code in the "SAS Viya-all models" node runs multiple SAS Visual Data Mining and Machine Learning procedures: the NNET procedure to fit a neural network model, the GENSELECT procedure to fit a generalized linear model, the FOREST procedure to fit a forest model, the

GRADBOOST procedure to fit a gradient boosting model, the TREESPLIT procedure to fit a regression tree, and the REGSELECT procedure to fit a linear regression model. Since there are multiple models, the NAME= argument is specified for each %EM_VIYA_ASSESS call. Then the %EM_VIYA_MODELSELECTION macro is invoked to select the best model to export to subsequent nodes in the flow. This selection is based on the following properties specified in the SAS Viya Code node: **Interval Selection Statistic** (because the target is interval) and **Selection Table**. The best model is selected based on the **Mean Absolute Error** in the **Train** partition. The SAS Viya Model Selection table in the **Results** for this node shows that the forest model was selected, because it has the lowest mean absolute error in the **Train** partition.

- The "SAS Viya-nnet" node trains a neural network by using the NNET procedure. In PROC NNET, the AUTOTUNE statement automatically tunes the hyperparameters, searching via an optimization algorithm for the best number of hidden layers, hidden units, and values of the regularization parameters.

The Model Comparison node compares the three models (one each from the three SAS Viya Code nodes) and selects the champion model. (Keep in mind that the "SAS Viya-all models" node exports only the forest model, which it chose as the best model.) Because all three models come from SAS Viya Code nodes, the **Selection Data** property should be set to **SAS Viya** to assess on the entire data partitions. You can then choose the **SAS Viya Selection Statistic** for selecting the champion model. Additional assessment statistics are available for SAS Viya models because SAS Visual Data Mining and Machine Learning supports an extensive set via the ASSESS procedure. By default, the champion model is selected based on the average square error (because the target is interval) of the validation partition.