Introduction to IBM SPSS Modeler 18.2 On-demand Software Training

These training modules are intended to assist end-users in learning to use IBM SPSS Modeler software. Though this training contains examples of advanced and predictive analytics, it is not intended to be used in the place of formal advanced and predictive analytics training.

Contents:

Part 2 – Introduction to Exploratory Data Analysis (EDA)

Module 4 – (:39 mins) Data Understanding and Preparation

Summary: Briefly discuss data science origins and trends and terms. Discuss Modeler data types and roles. Run Data Audit and review output. Filter fields from 25 fields to 5 fields. Run Data Audit again. Review the Quality tab. Coerce outliers and extremes by using the Generate button to automatically produce Filler nodes to trim outlier and extreme values. Identify negative values in two different fields. Create Filler nodes to replace negative values with the means of each field. Create a quintile binning field for a continuous field. Examine bins using Distribution graph. Examine means of continuous variable across quintiles. Create new segment variable by reclassifying quintiles into Low Med High segments. Run Means node on new segment field. Transform skewed continuous variables then run a Histogram to review transformations. Discuss use of comments to annotate streams.

Nodes Introduced: [Histogram Transform Means]

Module 5 – (:25 mins) Basic Univariate Analysis

Summary: Reiterate why IBM SPSS Modeler is a powerful tool and how it can enable an analytics culture/discipline and the importance of identifying "Useful Variance" to enable informed business action. Discuss univariate analysis (e.g., frequencies, descriptive statistics, and visualizations). Run and review a Data Audit. Generate graph node automatically by double-clicking a graph inside Data Audit output. Review how outliers can be coerced, and missing values can be imputed from within the Data Audit output. Discuss Graphs tab. Run 3 histograms. Note skew of a field. Transform field with a Derive node by taking the square root of the field. Compare histograms of the transformed field and original field. Use a Graphboard node to create a Boxplots by a discrete field. Run descriptive statistics. Create 2 Distribution graphs and a Matrix cross-tab. Create and sort a Modeler node usage frequency table.

Nodes Introduced: [Graphboard Box Plot Distribution Statistics]

Module 6 – (:32 mins) Basic Multivariate Analysis

Summary: Module 6 steps things up a bit and gets closer to the heart of advanced and predictive analysis. Discuss various forms of multi-variate analysis in addition to key terms. Show how comments can be used to highlight and demarcate elements of a complex stream. Discuss 4 types of multi-variate analysis: Continuous x Continuous, Discrete by Discrete, Continuous by Discrete, and a 3 variable Heat Map. Reiterate the use of stream annotations. Use Statistics node to compute correlations among continuous variables. Use Graphboard to produce a scatter plot matrix (SPLOM) for 5 continuous variables. Use Plot graphs to display scatter plots. Demonstrate a Distribution graph with color to visualize relationship between 2 discrete fields and then show a Matrix crosstab and Web graph with 2 discrete variables. Run Means node to show how a continuous variable varies across a discrete variable. Plot paneled Histograms and Boxplots and a Heat Map that displays the relationship between 2 discrete variables and a continuous variable.

Nodes Introduced: [Plot SPLOM Matrix Web Heat Map]

Part 3 – Introduction to Auto Machine Learning in Modeler

Module 7 – (:33 mins) Feature Selection

Summary: Discuss importance of proper training for predictive analytics and excellent training resources available online. Discuss applied predictive analytics and some best practice approaches. Reiterate that data science and predictive analytics is a team sport. Use a business to business example to describe Feature Selection at a high level and how it can be used to prepare for predictive analytics. Merge 2 tables, run Data Audit then aggregate results. Run Feature Selection node and examine the model nugget output. Introduce various classification and regression tree algorithms (i.e., XGBoost, Random Trees, C&RT, C5.0, CHAID, Quest). Introduce and review Auto Data Prep node.

Nodes Introduced: [Feature Selection Auto Data Prep]

Module 8 – (:29 mins) Classification & Regression Trees

Summary: Reiterate importance of proper training for predictive analytics. Discuss what classification and regression trees are, how they are a data science/predictive analytics workhorse, and how they can be used. Review Feature Selection output nugget from Module 7. Review Modeler C&RT node. Launch an interactive session from the C&RT node and explore a classification tree. Grow the tree 5 levels and explain how the interactive tree can be used for Interaction Detection. Create a training/validation partition and examine it with a Distribution node. Explain the use of a partition for predictive model development. Run and examine a C5.0 classification model. Review predictor importance and the tree produced by the model. Run Evaluation and Analysis nodes to review the model's properties via a confusion matrix and a gains chart. Output the model to a table and describe the additional fields produced. Score Test partition with the C5.0 model and analyze results.

Nodes Introduced: [C&RT Analysis Evaluation]

Module 9 – (:42 mins) Auto Machine Learning

Summary: Discuss how applied data science/predictive analytics can be messy and the need to balance information risks and benefits. Discuss Modeler's capabilities in terms of automated feature selection, predictive modeling, time series forecasting, and cluster development. Also discuss general categories of automated algorithms. Review C5.0 classification model produced in Module 8. Run the same model requirements through Auto Classifier node and highlight model improvement. Discuss Auto Data prep node. Configure a regression model with Income as the target variable using the Auto Numeric node. View the top 3 regression models produced. Examine 1 of the models in detail including predictor importance. Generate multiple unsupervised cluster solutions using Auto Cluster node. Select 1 cluster solution and examine in detail. Use a Distribution node to further examine clusters. Utilize a date conversion function and the Time Interval node to prepare a time series model. Build a time series model using Time Series node. Examine results and output. Wrap up series and close.

Nodes Introduced: [Auto Classifier Auto Numeric Auto Cluster Time Series Time Intervals Time Plot]