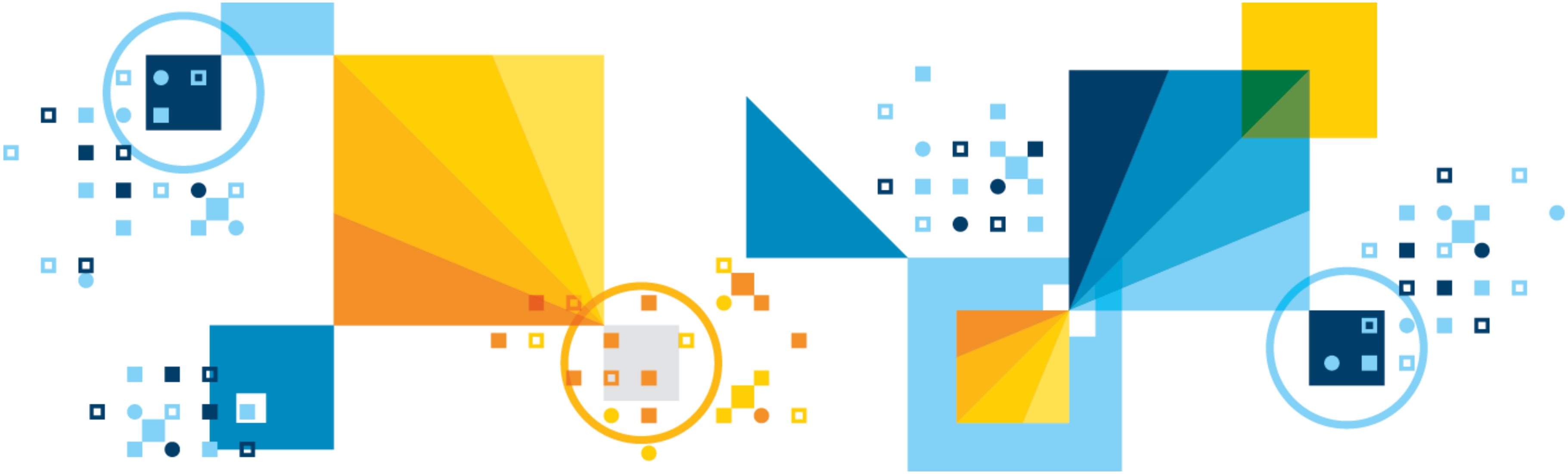
Predictive Modeling Fundamentals I
Lesson 3



Setting the Stage….

Why this is important to know…

1. Fundamental introduction to Data Mining and its application to business problems
2. Ability to utilize software tools for advanced analytics

After this session, you will be able to…

1. Introduction to the Common Modeling Techniques
2. Differentiate between unsupervised and supervised learning
3. Understand the SPSS Modeler algorithms available

Speaking to you today…

Armand Ruiz

Product Manager

Mikhail Lakirovich

Product Marketing Manager

Agenda

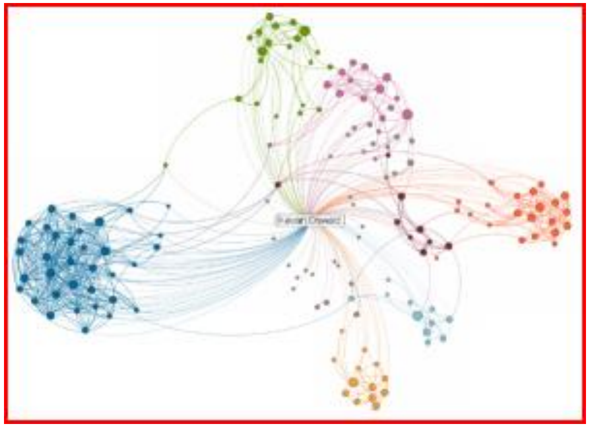
- Introduction to Common Modeling Techniques
- Unsupervised Learning - Cluster Analysis
- Supervised Learning - Classification & Prediction
- Classification - Training & Testing
- Sampling Data in Classification
- Predictive modeling Algorithms in SPSS Modeler
- Automated Selection of Algorithms

Introduction to Common Modeling Techniques

- Supervised Learning
- Describes and distinguishes classes for future prediction (on new data) based on training data
- Classification & Prediction
- Common Methods: Decision Trees, Regression, Nearest Neighbors, Neural Networks
- Unsupervised Learning
- Analyzes data where labels are unknown to create groups/classes for objects that are similar to each other (within the group) but dissimilar to objects in other clusters
- Cluster analysis
- Common Methods: K-means, Hierarchical, Two-Step
- Association
- Analyzing data for events or instances that occur together (i.e. diapers and beer commonly purchased together)
- Association Rules
- Common Methods: Apriori, CARMA

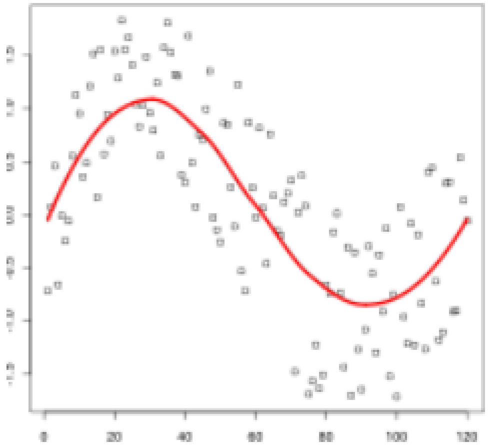
Unsupervised Learning – Cluster Analysis

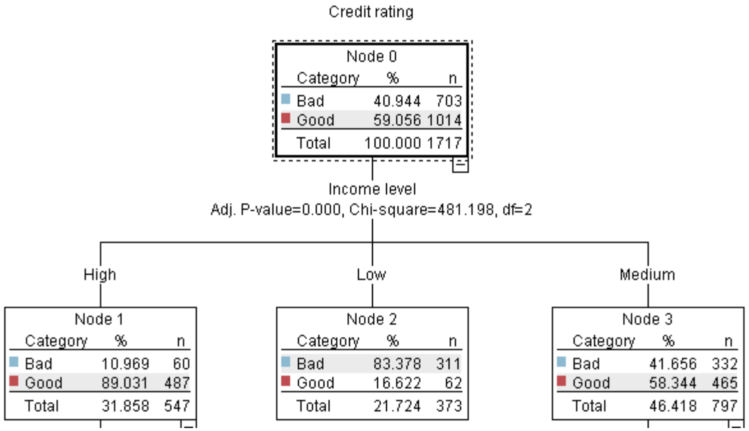
- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
  - Classes are not predefined – the model does not learn to classify new classes from existing classified data



Supervised Learning – Classification & Prediction

- Classification:
  - Predicting class label (e.g. churn, fraud, purchase – yes/no)
  - Constructs a classification model based on the training set and uses it in classifying new data
- Prediction:
  - models continuous variables and predicts unknown or missing values




Credit rating

Node 0

Category	%	n
🟦 Bad	40.944	703
🟥 Good	59.056	1014
Total	100.000	1717

Income level
Adj. P-value=0.000, Chi-square=481.198, df=2

High

Node 1

Category	%	n
🟦 Bad	10.969	60
🟥 Good	89.031	487
Total	31.858	547

Low

Node 2

Category	%	n
🟦 Bad	83.378	311
🟥 Good	16.622	62
Total	21.724	373

Medium

Node 3

Category	%	n
🟦 Bad	41.656	332
🟥 Good	58.344	465
Total	46.418	797

Classification – Training & Testing

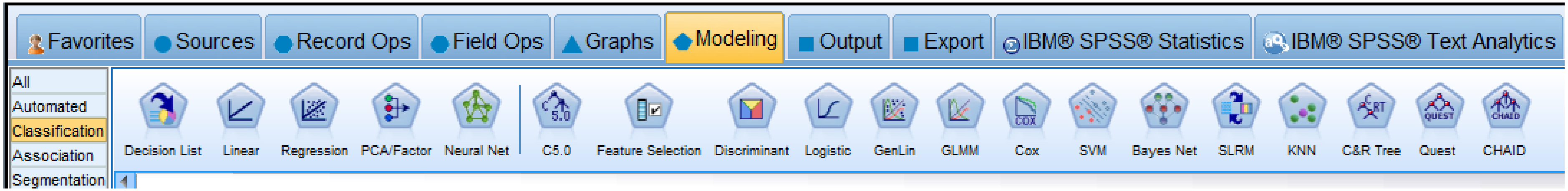
- Splitting the data set into Training and Testing
  - Approximately 66%-75% for training and 34%-25% for testing
- Training the model
  - On the data with existing classes – supervised learning
- Testing the model
  - On the portion of the data that was not included in the training phase
- Evaluating the model
  - Comparing the accuracy of the model on the training and testing sets
  - Accuracy rate is the percentage of sample that is correctly classified by the model
  - High accuracy for both training and testing data sets
  - High accuracy on training and low on testing -> overfitting problem
- Using the model
  - or classifying future or unknown objects

Sampling Data in Classification

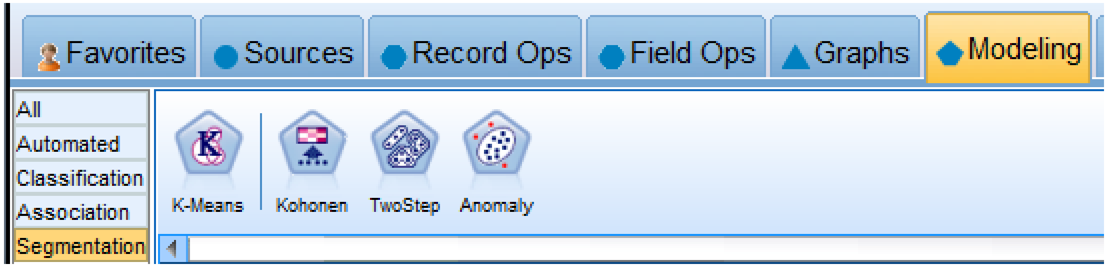
- **Why Sample?**
  - Numerosity Reduction: dealing with a smaller subset of massive dataset that is representative of the population
- ***Simple samples***
  - I take 30% of my original sample
  - May not be appropriate for unbalanced data (1000 positive and 100 negative cases)
- ***Complex samples***
  - ***-Clustered samples***: used to sample groups or clusters rather than individual units.
- ***Stratified samples***
  - ***-Stratified samples***: Used to select samples independently within non-overlapping subgroups of the population, or strata.
  - -For example, you can ensure that men and women are sampled in equal proportions, or that every region or socioeconomic group within an urban population is represented.
  - -You can also specify a sample size for each strata

Predictive Modeling Algorithms in SPSS Modeler

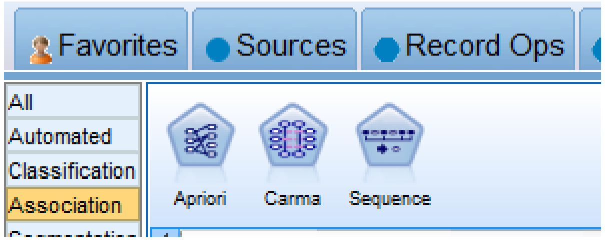
- **Classification/Prediction Algorithms**



- **Clustering Algorithms**



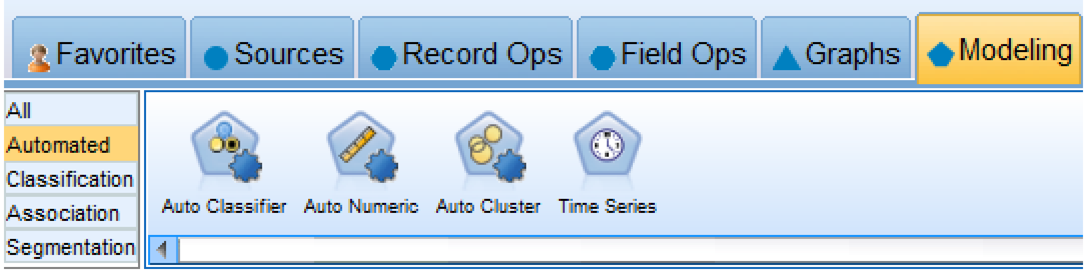
- **Association Rules Algorithms**

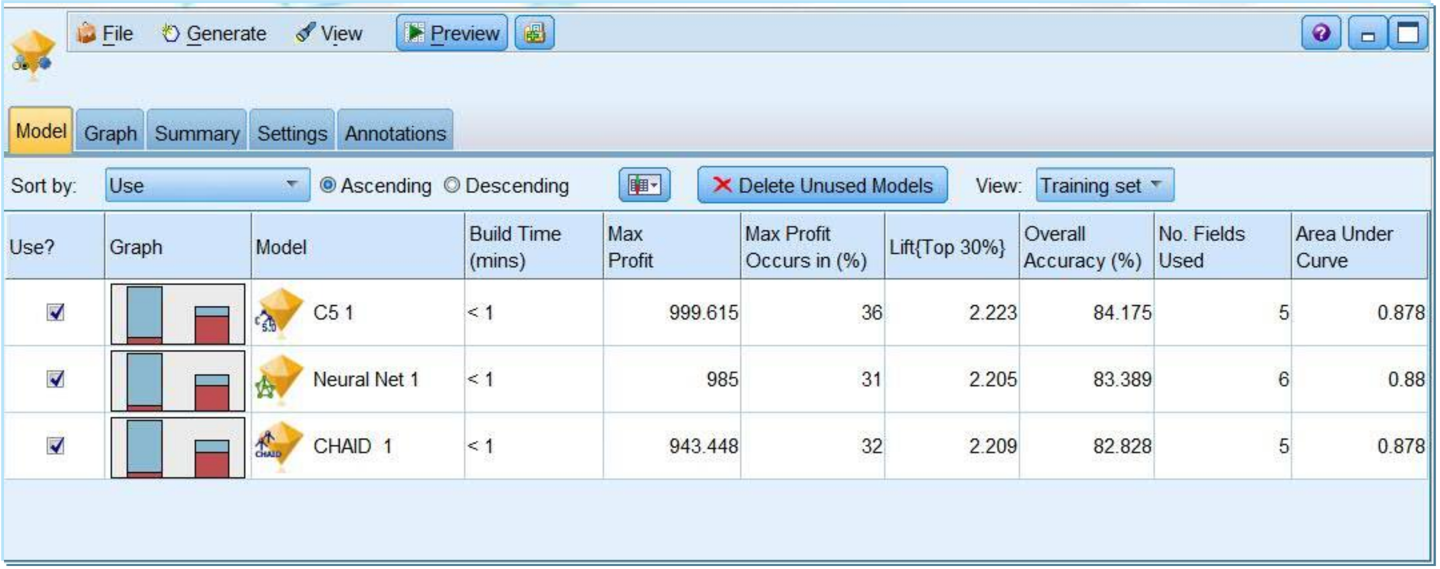


Too many choices…which one to pick?

- **How do you select the right algorithm for your project?**

- **Automated Algorithms in Modeler**
  - Modeler selects the best algorithms for the project given the data and the task





Sort by: Use ● Ascending ○ Descending View: Training set

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
☑		C5 1	< 1	999.615	36	2.223	84.175	5	0.878
☑		Neural Net 1	< 1	985	31	2.205	83.389	6	0.88
☑		CHAID 1	< 1	943.448	32	2.209	82.828	5	0.878

Lab 3:

- Build a Logistic Regression Model
- Use the Auto-Modelling