

# SAP FORUM 2010

## CLAREZA PARA UM NOVO BRASIL



Apice Consultoria  
08/03/2010  
José Domingos



**SAP** WORLD TOUR 10

# SAP FORUM 2010

## CLAREZA PARA UM NOVO BRASIL

### Agenda



1. **Apice Consultoria – Serviços SAP em *Banking, Modelling e Finance***
2. Predictive Workbench (BO-PW)
  - 2.1. Análise preditiva: definição e aplicações
  - 2.2. Arquitetura tecnológica
  - 2.3. Funcionalidades: base de dados, *inputs*, modelos e *outputs*
3. Sumário do *case*
4. Demo



A Apice conta com vários profissionais com experiência em soluções SAP:



Criar Valor para  
os Clientes

- 1 Contabilidade e Controladoria
- 2 Tesouraria e Gerenciamento de riscos
- 3 Planejamento Financeiro e Orçamentário
- 4 *Banking*
- 5 Tecnologia e BI
- 6 Cadeia de Suprimentos Financeira

# SAP FORUM 2010

CLAREZA PARA UM NOVO BRASIL

## Agenda



1. Apice Consultoria – Serviços SAP em *Banking, Modelling e Finance*
2. **Predictive Workbench (BO-PW)**
  - 2.1. **Análise preditiva: definição e aplicações**
  - 2.2. **Arquitetura tecnológica**
  - 2.3. **Funcionalidades: base de dados, *inputs*, modelos e *outputs***
3. Sumário do *case*
4. Demo

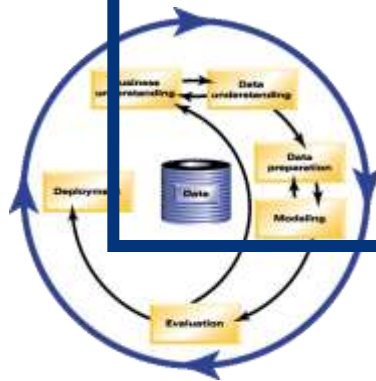


## O que é análise Preditiva?

Análise Preditiva ajuda a conectar dados históricos com a tomada de decisão, através de interpretações e conclusões de condições atuais (“as-is”) e que ajudam a prever os eventos futuros mais prováveis.

**Gareth Herschel,**

Research Director, Gartner Group



## O que é análise preditiva? “Insights” baseados em dados



## Análises preditivas aplicadas:

### Estimação de modelo de *default*

- Quais variáveis explicativas estão fortemente correlacionados com o *default* de clientes *corporate* e varejo?

### Identificando grupos de clientes

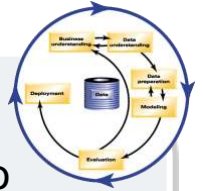
- Como consigo agrupar os meus clientes da carteira de crédito de maneira a gerar subgrupos o mais homogêneo possível?

### Descobrendo tendências de variáveis macro

- Quais são as tendências, ciclos e/ou sazonalidades das variáveis macroeconômicas que impactam a concessão de crédito das diferentes carteiras?

### Entendendo oportunidades de negócio

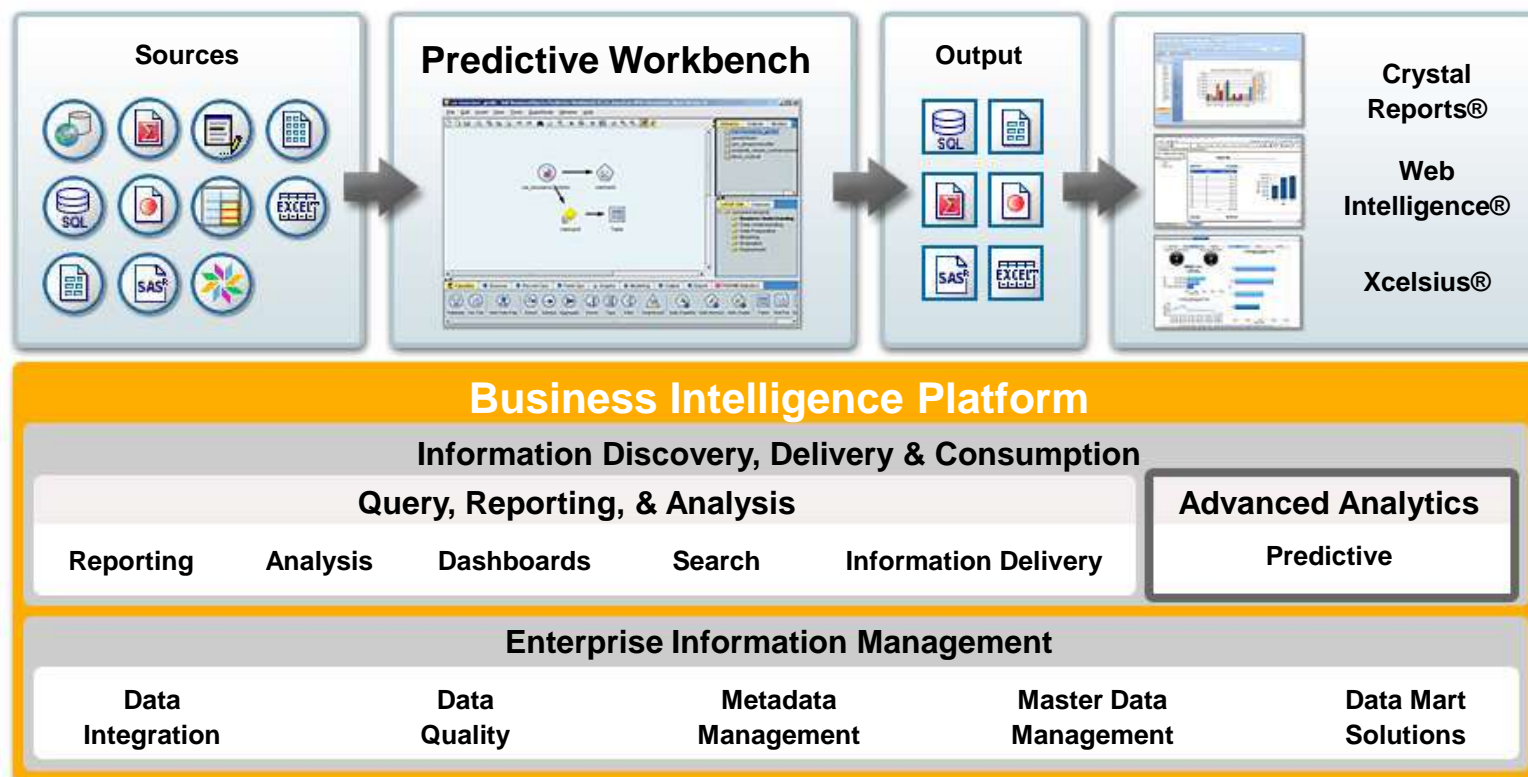
- Quais produtos estão inter-relacionados? Quais são as oportunidades “cross-sell” e “up-sell” ?



## Integrado com plataforma SAP Business Objects

### SAP BusinessObjects Predictive Workbench

- Baseado no SPSS Statistics
- Adaptado pela SAP BusinessObjects, e vendido com pacotes de integração para suportar o uso de universos e queries com fonte de dados



## Máximo de benefícios em menor tempo

Produtividade máxima para analistas

- Utilize os dados para descobrir novos fatos
- Resolva problemas de processo de negócio
- Use os resultados para alinhar os objetivos estratégicos

Flexibilidade de uso por qualquer usuário

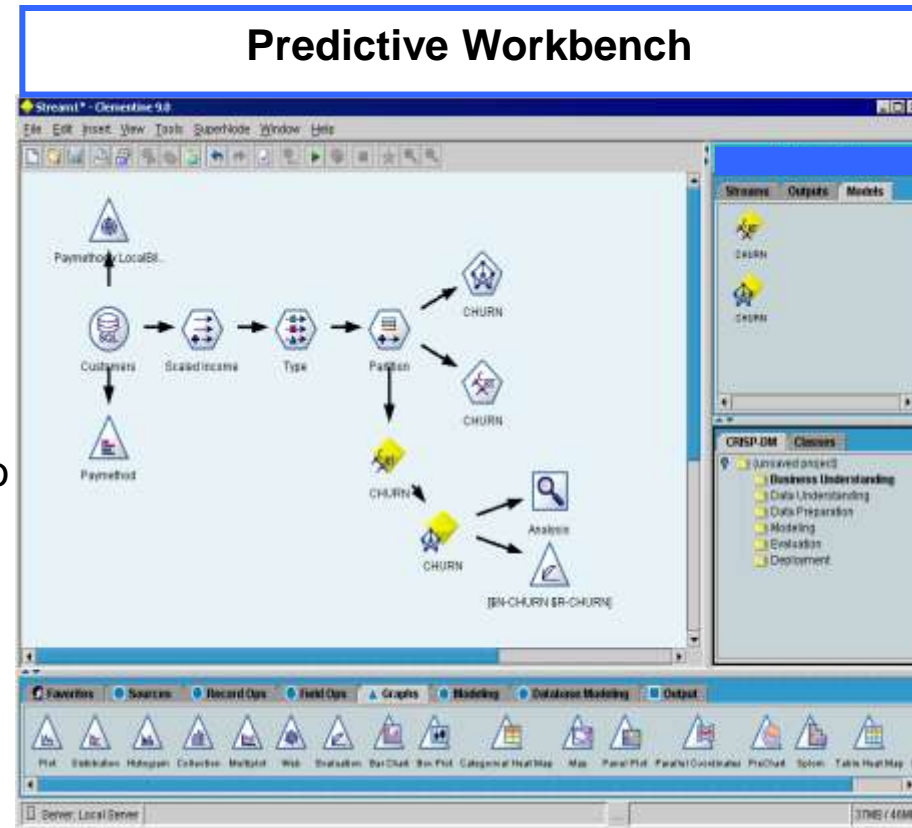
- Eficiência na modelagem
- Ferramenta fácil de usar, “user-friendly”
- Fique à frente da concorrência
- Faça um planejamento mais racional e assertivo

Inteligência preditiva integrada ao BI

- Apresentação de “insight” preditivos para os usuários
- Manual completo para utilização dos algoritmos



### Predictive Workbench



## Predictive Workbench

### Data sources – Universes e ainda...

The Sources palette contains the following nodes:



The Database node can be used to import data from a variety of other packages using ODBC (Open Database Connectivity), including Microsoft SQL Server, DB2, Oracle, and others.



The Variable File node reads data from free-field text files—that is, files whose records contain a constant number of fields but a varied number of characters. This node is also useful for files with fixed-length header text and certain types of annotations.



The Fixed File node imports data from fixed-field text files—that is, files whose fields are not delimited but start at the same position and are of a fixed length. Machine-generated or legacy data are frequently stored in fixed-field format.



The SPSS Import node reads data from the .sav file format used by SPSS, as well as cache files saved in Clementine, which also use the same format.



The Dimensions Data Import node imports survey data based on the Dimensions Data Model used by SPSS market research products. The Dimensions Data Library must be installed to use this node.



The SAS Import node imports SAS data into Clementine.



The Excel Import node imports data from any version of Microsoft Excel. An ODBC data source is not required.



The User Input node provides an easy way to create synthetic data—either from scratch or by altering existing data. This is useful, for example, when you want to create a test dataset for modeling.

## Predictive Workbench

### Modelos

#### Classification models



The Classification and Regression (C&R) Tree node generates a decision tree that allows you to predict or classify future observations. The method uses recursive partitioning to split the training records into segments by minimizing the impurity at each step, where a node is considered "pure" if 100% of cases in the node fall into a specific category of the target field. Target and predictor fields can be range or categorical; all splits are binary (only two subgroups).



The QUEST node provides a binary classification method for building decision trees, designed to reduce the processing time required for large C&R Tree analyses while also reducing the tendency found in classification tree methods to favor predictors that allow more splits. Predictor fields can be numeric ranges, but the target field must be categorical. All splits are binary.



The CHAID node generates decision trees using chi-square statistics to identify optimal splits. Unlike the C&R Tree and QUEST nodes, CHAID can generate nonbinary trees, meaning that some splits have more than two branches. Target and predictor fields can be range or categorical. Exhaustive CHAID is a modification of CHAID that does a more thorough job of examining all possible splits but takes longer to compute.



The Decision List node identifies subgroups, or **segments**, that show a higher or lower likelihood of a given binary outcome relative to the overall population. For example, you might look for customers who are unlikely to churn or are most likely to respond favorably to a campaign. You can incorporate your business knowledge into the model by adding your own custom segments and previewing alternative models side by side in order to compare the results. Decision List models consist of a list of rules in which each rule has a condition and an outcome. Rules are applied in order, and the first rule that matches determines the outcome.



Linear regression is a common statistical technique for summarizing data and making predictions by fitting a straight line or surface that minimizes the discrepancies between predicted and actual output values.



The Factor/PCA node provides powerful data-reduction techniques to reduce the complexity of your data. Principal components analysis (PCA) finds linear combinations of the input fields that do the best job of capturing the variance in the entire set of fields, where the components are orthogonal (perpendicular) to each other. Factor analysis attempts to identify underlying factors that explain the pattern of correlations within a set of observed fields. For both approaches, the goal is to find a small number of derived fields that effectively summarizes the information in the original set of fields.

#### Segmentation models



The K-Means node clusters the dataset into distinct groups (or clusters). The method defines a fixed number of clusters, iteratively assigns records to clusters, and adjusts the cluster **centers** until further refinement can no longer improve the model. Instead of trying to predict an outcome, k-means uses a process known as unsupervised learning to uncover patterns in the set of input fields.

#### Association models



The Generalized Rule Induction (GRI) node discovers association rules in the data. For example, customers who purchase razors and aftershave lotion are also likely to purchase shaving cream. GRI extracts rules with the highest information content based on an index that takes both the generality (support) and accuracy (confidence) of rules into account. GRI can handle numeric and categorical inputs, but the target must be categorical.

## Predictive Workbench

### Modelos

#### Classification Module

The Classification module helps organizations to predict a known result, such as whether a customer will buy or leave or whether a transaction fits a known pattern of fraud. Modeling techniques include machine learning (neural networks), decision trees (rule induction), subgroup identification, statistical methods, and multiple model generation. The following nodes are included:



The Binary Classifier node creates and compares a number of different models for binary outcomes (yes or no, churn or don't, and so on), allowing you to choose the best approach for a given analysis. A number of modeling algorithms are supported, making it possible to select the methods you want to use, the specific options for each, and the criteria for comparing the results. The node generates a set of models based on the specified options and ranks the best candidates according to the criteria you specify.



The Numeric Predictor node estimates and compares models for continuous numeric range outcomes using a number of different methods. The node works in the same manner as the Binary Classifier node, allowing you to choose the algorithms to use and to experiment with multiple combinations of options in a single modeling pass. Supported algorithms include neural networks, C&R Tree, CHAID, linear regression, generalized linear regression, and support vector machines (SVM). Models can be compared based on correlation, relative error, or number of variables used.



The Neural Net node uses a simplified model of the way the human brain processes information. It works by simulating a large number of interconnected simple processing units that resemble abstract versions of neurons. Neural networks are powerful general function estimators and require minimal statistical or mathematical knowledge to train or apply.



The C5.0 node builds either a decision tree or a rule set. The model works by splitting the sample based on the field that provides the maximum information gain at each level. The target field must be categorical. Multiple splits into more than two subgroups are allowed.



The Feature Selection node screens predictor fields for removal based on a set of criteria (such as the percentage of missing values); it then ranks the importance of remaining predictors relative to a specified target. For example, given a dataset with hundreds of potential predictors, which are most likely to be useful in modeling patient outcomes?



Discriminant analysis makes more stringent assumptions than logistic regression but can be a valuable alternative or supplement to a logistic regression analysis when those assumptions are met.



Logistic regression is a statistical technique for classifying records based on values of input fields. It is analogous to linear regression but takes a categorical target field instead of a numeric range.



The generalized linear model expands the general linear model so that the dependent variable is linearly related to the factors and covariates via a specified link function. Moreover, the model allows for the dependent variable to have a non-normal distribution. It covers the functionality of a wide number of statistical models, including linear regression, logistic regression, loglinear models for count data, and interval-censored survival models.



The Bayesian Network node enables you to build a probability model by combining observed and recorded evidence with real-world knowledge to establish the likelihood of occurrences. In the current Clementine 12.0 release, the node focuses on Tree Augmented Naive Bayes (TAN) and Markov Blanket networks that are primarily used for classification.



The Cox regression node enables you to build a survival model for time-to-event data in the presence of censored records. The model produces a survival function that predicts the probability that the event of interest has occurred at a given time ( $t$ ) for given values of the predictor variables.



The Support Vector Machine (SVM) node enables you to classify data into one of two groups without overfitting. SVM works well with wide datasets, such as those with a very large number of predictor fields.



The Self-Learning Response Model (SLRM) node enables you to build a model in which a single new case, or small number of new cases, can be used to reestimate the model without having to retrain the model using all data.



The Time Series node estimates exponential smoothing, univariate Autoregressive Integrated Moving Average (ARIMA), and multivariate ARIMA (or transfer function) models for time series data and produces forecasts of future performance. A Time Series node must always be preceded by a Time Intervals node.

## Predictive Workbench

### Modelos

#### Segmentation Module

The Segmentation module is recommended in cases where the specific result is unknown (for example, when identifying new patterns of fraud, or when identifying groups of interest in your customer base). Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics, and it distinguishes clustering models from the other modeling techniques in that there is no predefined output or target field for the model to predict. There are no right or wrong answers for these models. Their value is determined by their ability to capture interesting groupings in the data and provide useful descriptions of those groupings. Clustering models are often used to create clusters or segments that are then used as inputs in subsequent analyses (for example, by segmenting potential customers into homogeneous subgroups).

This following nodes are included:



The Kohonen node generates a type of neural network that can be used to cluster the dataset into distinct groups. When the network is fully trained, records that are similar should appear close together on the output map, while records that are different will appear far apart. You can look at the number of observations captured by each unit in the model nugget to identify the strong units. This may give you a sense of the appropriate number of clusters.



The TwoStep node uses a two-step clustering method. The first step makes a single pass through the data to compress the raw input data into a manageable set of subclusters. The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters. TwoStep has the advantage of automatically estimating the optimal number of clusters for the training data. It can handle mixed field types and large datasets efficiently.



The Anomaly Detection node identifies unusual cases, or outliers, that do not conform to patterns of "normal" data. With this node, it is possible to identify outliers even if they do not fit any previously known patterns and even if you are not exactly sure what you are looking for.

#### Association Module

The Association module is most useful when predicting multiple outcomes—for example, customers who bought product X also bought Y and Z. Association models associate a particular conclusion (such as the decision to buy something) with a set of conditions. Association rule algorithms automatically find the associations that you could find manually using visualization techniques, such as the Web node. The advantage of association rule algorithms over the more standard decision tree algorithms (C5.0 and C&RT) is that associations can exist between any of the attributes. A decision tree algorithm will build rules with only a single conclusion, whereas association algorithms attempt to find many rules, each of which may have a different conclusion. The following nodes are included:



The Apriori node extracts a set of rules from the data, pulling out the rules with the highest information content. Apriori offers five different methods of selecting rules and uses a sophisticated indexing scheme to process large datasets efficiently. For large problems, Apriori is generally faster to train than GRI; it has no arbitrary limit on the number of rules that can be retained, and it can handle rules with up to 32 preconditions. Apriori requires that input and output fields all be categorical but delivers better performance because it is optimized for this type of data.



The CARMA model extracts a set of rules from the data without requiring you to specify In (predictor) or Out (target) fields. In contrast to Apriori and GRI, the CARMA node offers build settings for rule support (support for both antecedent and consequent) rather than just antecedent support. This means that the rules generated can be used for a wider variety of applications—for example, to find a list of products or services (antecedents) whose consequent is the item that you want to promote this holiday season.



The Sequence node discovers association rules in sequential or time-oriented data. A **sequence** is a list of item sets that tends to occur in a predictable order. For example, a customer who purchases a razor and aftershave lotion may purchase shaving cream the next time he shops. The Sequence node is based on the CARMA association rules algorithm, which uses an efficient two-pass method for finding sequences.

## Predictive Workbench

### Gráficos

The Graphs palette contains the following nodes:



The Graphboard node offers many different types of graphs in one single node. Using this node, you can choose the data fields you want to explore and then select a graph from those available for the selected data. The node automatically filters out any graph types that would not work with the field choices.



The Plot node shows the relationship between numeric fields. You can create a plot by using points (a scatterplot) or lines.



The Multiplot node creates a plot that displays multiple Y fields over a single X field. The Y fields are plotted as colored lines; each is equivalent to a Plot node with Style set to Line and X Mode set to Sort. Multiplots are useful when you want to explore the fluctuation of several variables over time.



The Distribution node shows the occurrence of symbolic (categorical) values, such as mortgage type or gender. Typically, you might use the Distribution node to show imbalances in the data, which you could then rectify using a Balance node before creating a model.



The Histogram node shows the occurrence of values for numeric fields. It is often used to explore the data before manipulations and model building. Similar to the Distribution node, the Histogram node frequently reveals imbalances in the data.



The Collection node shows the distribution of values for one numeric field relative to the values of another. (It creates graphs that are similar to histograms.) It is useful for illustrating a variable or field whose values change over time. Using 3-D graphing, you can also include a symbolic axis displaying distributions by category.



The Web node illustrates the strength of the relationship between values of two or more symbolic (categorical) fields. The graph uses lines of various widths to indicate connection strength. You might use a Web node, for example, to explore the relationship between the purchase of a set of items at an e-commerce site.



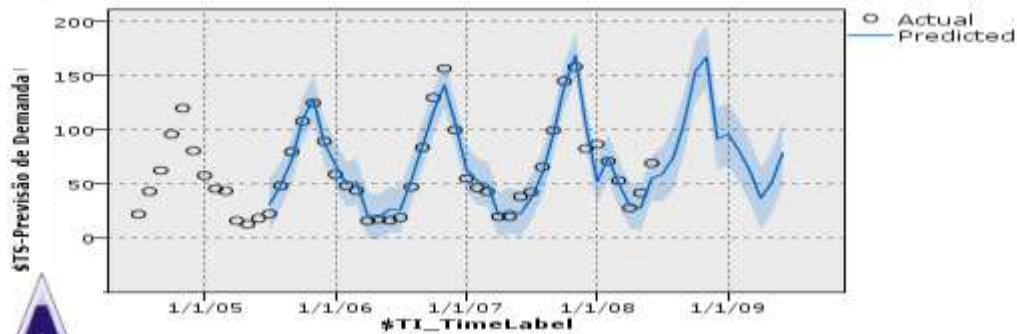
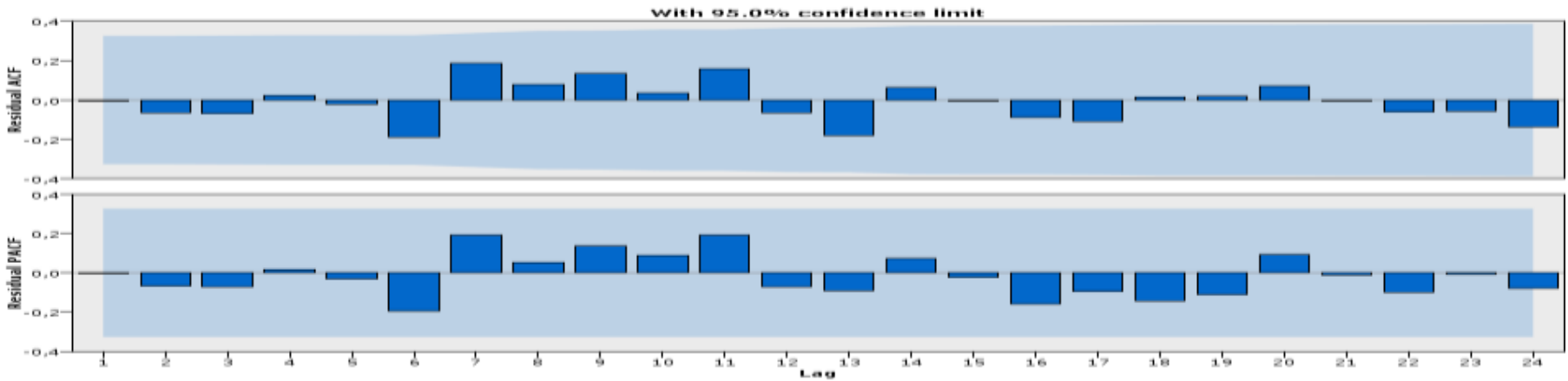
The Evaluation node helps to evaluate and compare predictive models. The evaluation chart shows how well models predict particular outcomes. It sorts records based on the predicted value and confidence of the prediction. It splits the records into groups of equal size (**quantiles**) and then plots the value of the business criterion for each quantile from highest to lowest. Multiple models are shown as separate lines in the plot.



The Time Plot node displays one or more sets of time series data. Typically, you would first use a Time Intervals node to create a *TimeLabel* field, which would be used to label the x axis.

## Séries temporais - ARIMA

lay plot for model: Previsão de PIB mensal



\$TI_Year	\$TI_Month	Previsão de PIB mensal
2008	7	58.934
2008	8	74.465
2008	9	108.279
2008	10	153.904
2008	11	167.014
2008	12	91.402
2009	1	95.490
2009	2	79.776
2009	3	61.973
2009	4	36.763
2009	5	50.566
2009	6	77.994

# SAP FORUM 2010

## CLAREZA PARA UM NOVO BRASIL

### Agenda



1. Apice Consultoria – Serviços SAP em *Banking, Modelling e Finance*
2. Predictive Workbench (BO-PW)
  - 2.1. Análise preditiva: definição e aplicações
  - 2.2. Arquitetura tecnológica
  - 2.3. Funcionalidades: base de dados, *inputs*, modelos e *outputs*
3. **Sumário do case**
4. Demo





## Objetivos da demo

- Através de regressão logística, selecionar variáveis que explicam *default*
- Criar classes de rating, utilizando *clusters* e árvores de decisão

# SAP FORUM 2010

## CLAREZA PARA UM NOVO BRASIL

### Agenda



1. Apice Consultoria – Serviços SAP em *Banking, Modelling e Finance*
2. Predictive Workbench (BO-PW)
  - 2.1. Análise preditiva: definição e aplicações
  - 2.2. Arquitetura tecnológica
  - 2.3. Funcionalidades: base de dados, *inputs*, modelos e *outputs*
3. Sumário do *case*
4. **Demo**





# SAP FORUM 2010

CLAREZA PARA UM NOVO BRASIL

**Thank you!**

THE BEST-RUN BUSINESSES RUN SAP™

