StatSoft®

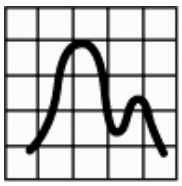
data analysis • data mining • quality control • web-based analytics

Financial Institutions

and

STATISTICA

Case Study: Credit Scoring



STATISTICA
Solutions for Business Intelligence, Data Mining, Quality Control, and Web-based Analytics

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o
China: StatSoft China
France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z.o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda
S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Table of Contents

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o
China: StatSoft China
France: StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z.o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda
S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Introduction: What is Credit Scoring?

In the financial industry, consumers regularly request credit to make purchases. The risk for financial institutions to extend the requested credit depends on how well they distinguish the good credit applicants from the bad credit applicants. One widely adopted technique for solving this problem is "Credit Scoring."

Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders. Further, it helps to assess the risk in lending. Credit scoring is a dependable assessment of a person's credit worthiness since it is based on actual data.

A lender commonly makes two types of decisions: first, whether to grant credit to a new applicant or not, and second, how to deal with existing applicants, including whether to increase their credit limits or not. In both cases, whatever the techniques used, it is critical that there is a large sample of previous customers with their application details, behavioral patterns, and subsequent credit history available. Most of the techniques use this sample to identify the connection between the characteristics of the consumers (annual income, age, number of years in employment with their current employer, etc.) and how "good" or "bad" their subsequent history is.

Typical application areas in the consumer market include: credit cards, auto loans, home mortgages, home equity loans, mail catalog orders, and a wide variety of personal loan products.

Credit Scoring: Business Objectives

The application of scoring models in today's business environment covers a wide range of objectives. The original task of estimating the risk of default has been augmented by credit scoring models to include other aspects of credit risk management: at the pre-application stage (identification of potential applicants), at the application stage (identification of acceptable applicants), and at the performance stage (identification of possible behavior of current customers). Scoring models with different objectives have been developed. They can be generalized into four categories as listed below.

1. Marketing Aspect

Purposes

1.1. Identify credit-worthy customers most likely to respond to promotional activity in order to reduce the cost of customer acquisition and minimize customer dissatisfaction.

1.2. Predict the likelihood of losing valuable customers and enable organizations to formulate effective customer retention strategy.

Examples

Response scoring. The scoring models that estimate how likely a consumer would respond to a direct mailing of a new product.

Retention/attrition scoring. The scoring models that predict how likely a consumer would keep using the product or change to another lender after the introductory offer period is over.

2. Application Aspect

Purposes

2.1. Decide whether to extend credit, and how much credit to extend.

2.2. Forecast the future behavior of a new credit applicant by predicting loan-default chances or poor-repayment behaviors at the time the credit is granted.

Example

Applicant scoring. The scoring models that estimate how likely a new applicant of credit will become default.

3. Performance Aspect

Purpose

3.1. Predict the future payment behavior of existing debtors in order to identify/isolate bad customers to direct more attention and assistance to them, thereby reducing the likelihood that these debtors will later become a problem.

Example

Behavioral scoring. Scoring models that evaluate the risk levels of existing debtors.

4. Bad Debt Management

Purpose:

4.1. Select optimal collections policies in order to minimize the cost of administering collections or maximizing the amount recovered from a delinquent's account.

Example

Scoring models for collection decisions: Scoring models that decide when actions should be taken on the accounts of delinquents and which of several alternative collection techniques might be more appropriate and successful.

Thus, the overall objective of credit scoring is not only to determine whether the applicant is credit worthy, but also to attract quality credit applicants who can subsequently be retained and controlled while maintaining an overall profitable portfolio.

Case Study: Consumer Credit Scoring

Case Description

In credit business, banks are interested in learning whether a prospective consumer will pay back their credit. The goal of this study is to model or predict the probability with which a credit applicant can be categorized as a good or bad customer.

The techniques explained in this case will illustrate how to build a credit-scoring model using *STATISTICA Data Miner* to identify the inputs or predictors that differentiate "risky" customers from others (based on patterns pertaining to previous customers), identify predictive techniques that perform well on test data, and later deploy those models to predict new risky customers.

Data File

The example data set used in this case, *CreditScoring.sta*, contains 1,000 cases and 20 variables (or predictors) with information pertaining to past and current customers who borrowed from a German bank (source:http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html) for various reasons. The data set contains information related to the customers' financial standing, reason to loan, employment, demographic information, etc. The example data file is found in the *STATISTICA* example data folder

For each customer, the binary outcome "creditability" is also available. This variable contains information about whether each customer's credit is deemed *Good* or *Bad.* The data set has a distribution of 70% credit worthy (good) customers and 30% not credit worthy (bad) customers. Customers who have missed 90 days of payment can be thought of as bad risks, and customers who have missed no payment can be thought of as good risks. Other typical measures for determining good and bad customers are the amount over the overdraft limit, current account turnover, number of months of missed payments, or a function of these and other variables.

Following is the complete list of variables used in this data set:

Category	Variables
1. Basic Personal Information	Age, Sex, Telephone, Foreign worker
2. Family Information	Marital Status, Number of dependents
3. Residential Information	Years at current address, Type of apartment
4. Employment Status	Years in current occupation, Occupation
5. Financial Status	Most valuable available assets, Further running credits, Balance of current account, Number of previous credits at this bank
6. Security information	Value of savings or stocks, Guarantors
7. Others	Purpose of credit, Amount of credit in Deutsche Marks (DM)

In this example, we will look at how well the variables listed above enable us to discriminate between whether someone has *Good* or *Bad Credit Standing*. If we can discriminate between these two groups, we can then use the predictive model to classify or predict new cases where we have the above-mentioned information but do not know the person's credit standing. This would be useful, for example, to decide whether to qualify a person for a loan.

Data Analysis with *STATISTICA*

Data Preparation

With *STATISTICA Data Miner,* it is straightforward to apply powerful modeling tools to data and judge the value of resulting models based on their predictive or descriptive value. This does not diminish the role of careful attention to data preparation efforts. Data is the main resource for data mining – therefore it should be prepared properly before applying any data-mining tool. Otherwise, it would be just a case of Garbage-In Garbage-Out (GIGO). Since major strategic decisions are impacted by these results, any error might give rise to huge losses. Thus, it is important to preprocess the data and improve the accuracy of the model so that one can make the best possible decision.

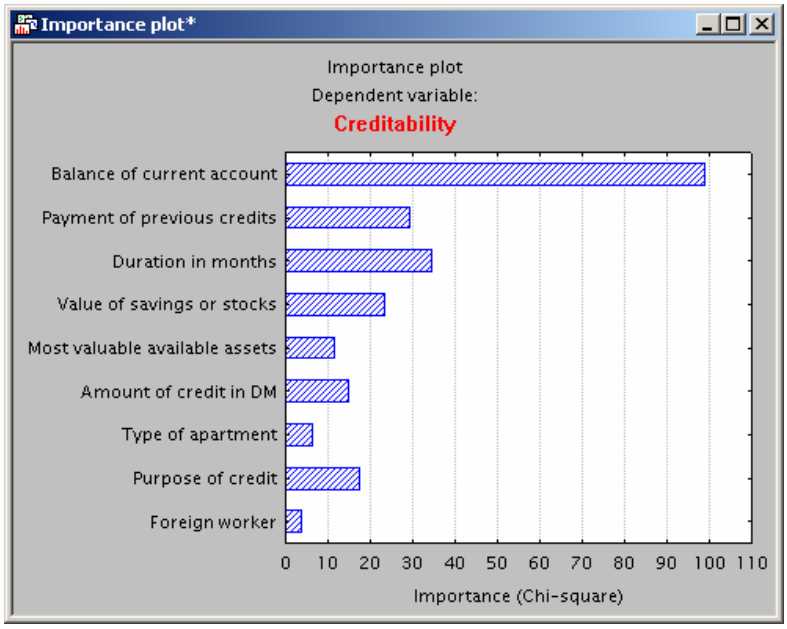
The following aspects of the data were noted during this stage

- Insight into data: Descriptive statistics (by looking at distributions, means, minimum and maximum values, quartiles, etc.)
- There are no outliers in the data
- There are no missing values in the data
- No transformations are required
- Feature selection – Variables reduced from 20 to 10

Feature Selection

In order to reduce the complexity of the problem, the data set can be transformed into a data set of lower dimension. The *Feature Selection and Variable Screening* tool available in *STATISTICA Data Miner* automatically found important predictors that clearly discriminate between good and bad customers.

The bar plot and spreadsheet of the predictor importance give insight into the variables that are related to the prediction of the dependent variable of interest. For example, shown below is the bar plot of predictor importance for the dependent variable "Creditability."



In this case, variables *Balance of current account*, *Payment of previous credits*, and *Duration in months* stand out as the most important predictors.

These predictors will be further examined using a wide array of data mining and machine learning algorithms available in *STATISTICA Data Miner* such as:

- *Standard Classification Trees with Deployment*
- *Standard Classification CHAID with Deployment*
- *Boosting Classification Trees with Deployment*
- *STATISTICA Automated Neural Networks with Deployment*
- *Support Vector Machine with Deployment (Classification)*
- *MARSplines for Classification with Deployment*

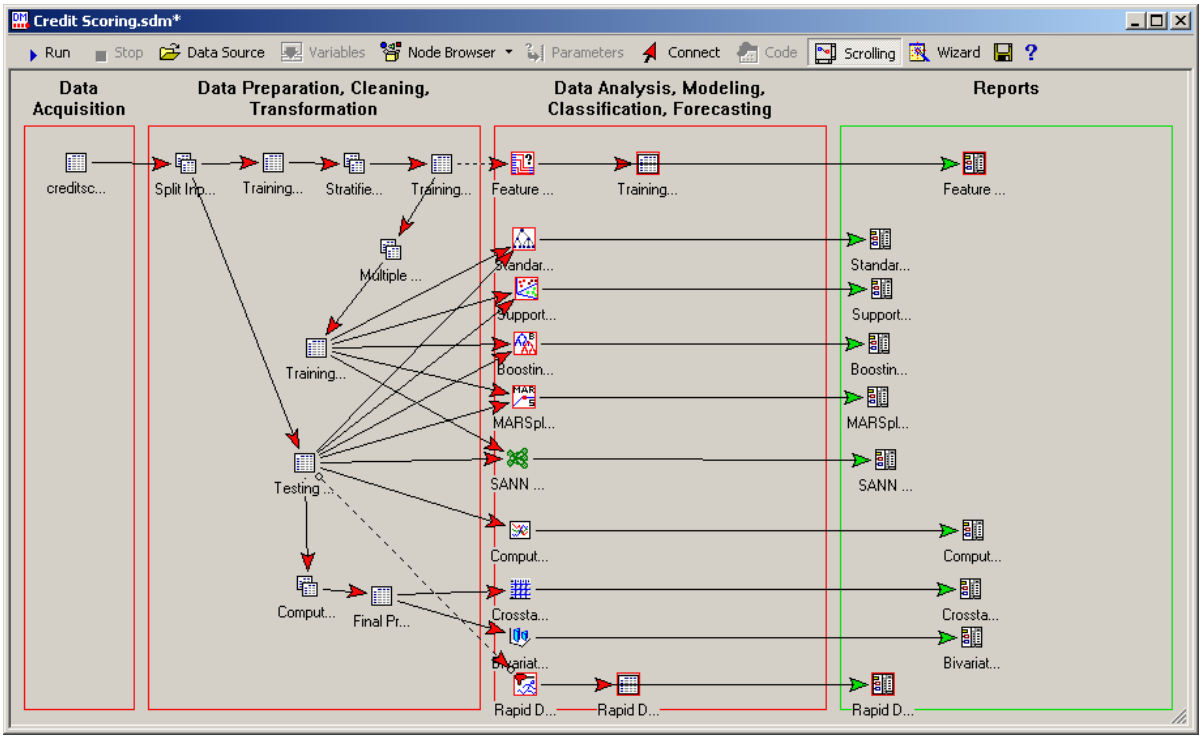
The novelty and abundance of available techniques and algorithms involved in the modeling phase make this the most interesting part of the data mining process. Classification methods are the most commonly used data mining techniques that are applied in the domain of credit scoring to predict the risk level of credit takers. Moreover, it is good practice to experiment with a number of different methods when modeling or mining data. Different techniques may shed new light on a problem or confirm previous conclusions.

STATISTICA Data Miner is a comprehensive and user-friendly set of complete data mining tools designed to enable users to more easily and quickly analyze their data to uncover hidden trends, explain known patterns, and predict the future. From querying databases and drilling down, to generating final reports and graphs, it offers ease of use without sacrificing power or comprehensiveness. Moreover, *STATISTICA Data Miner* features the largest selection of algorithms on the market for classification, prediction, clustering, and modeling as well as an intuitive icon-based interface. It offers simple techniques such as *C&RT* and *CHAID* to more advanced techniques such as *Neural Networks*, *Boosted Trees*, *Random Forests*, *Support Vector Machines*, *MARSplines*, etc.

***STATISTICA Data Miner* Workspace**

The *Data Miner* workspace depicts the flow of the analyses; all tools of *STATISTICA Data Miner* are available as icons via simple drag-and-drop.

The following diagram illustrates how the *Data Miner* workspace looks after all the analyses were performed.



The following steps summarize the data preparation and analysis flow:

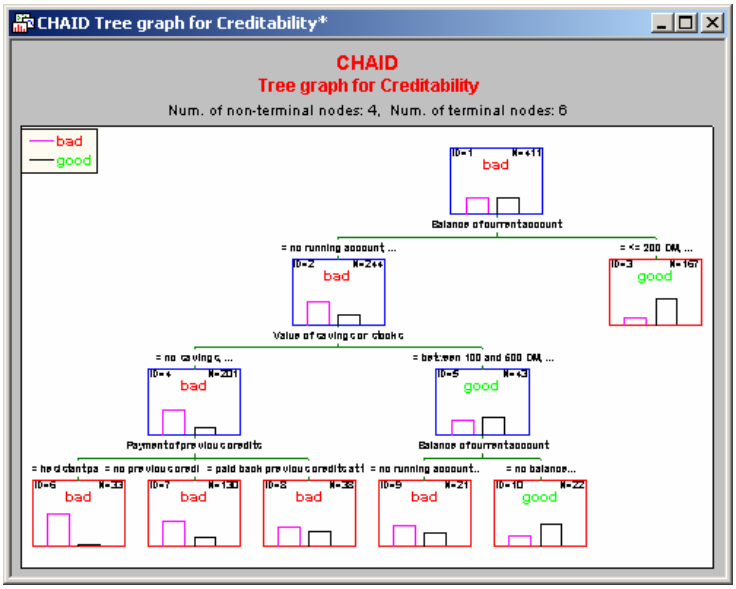
1. Split the original data set into two subsets; 34% of cases were retained for testing and 66% of cases were used for model building.
2. Used Stratified Random Sampling method to extract equal numbers of observations for both good and bad risk customers.
3. Used Feature Selection tool to rank the best predictor variables for distinguishing good and bad customers.
4. Reduced the number of possible predictors from 20 to 10 based on the results of Feature Selection.
5. Used different advanced Predictive Models (Machine Learning algorithms) to detect and understand relationships among words.
6. Used comparative tools such as Lift Charts, Gains Charts, Cross tabulation, etc., to find the best model for prediction purposes.
7. Applied the model to the Test Set (hold-out sample) to validate prediction accuracy.

Analyzing Results

Next, we will review the analysis results to better understand the characteristics of bad and good customers. Let's first start with the *CHAID* decision tree results.

Decision Tree - *CHAID*

Decision trees are powerful and popular tools for classification and prediction. The fact that decision trees can readily be summarized graphically makes them particularly easy to interpret.



***CHAID* decision tree for "Creditability"**

Note that the results you will see on your computer may vary because of different training and testing samples that will be created every time you update the project, at which point the input data are split into training and testing samples. However, in general, the results should be similar with respect to the major split variables and types of splits depicted in the tree shown above.

Looking at the tree shown here, you can see that the *CHAID* algorithm created a tree with 6 terminal nodes (highlighted in red), resulting from 4 *if-then* conditions to predict good/bad customers. Terminal nodes (or terminal leaves as they are sometimes called) are those where no further splits could be applied to further improve the predictive accuracy of the solution (given the current parameters that were selected to guide the tree-building process). The tree starts with the top decision node (also called the root node) with 411 cases in the training data set with approximately equal proportions of customers from both "good" and "bad" categories obtained by using the Stratified Random Sampling tool. The legend identifying which bars in the node histograms correspond to the two categories is located in the upper-left corner of the graph.

The interpretation of the tree is quiet straightforward. The rightmost node resulting from the first split contains 167 instances with a majority of cases associated with *good* customers. Since further splits from this point wouldn't help to improve the predictive accuracy of the model (depending on the defined settings), this node becomes the "terminal" node without any further splits. The leftmost node containing 244 instances is further split based on the predictor *Value of savings or stock*, resulting in two more nodes and so on.

Next, "decision rules" can be generated by following the path to each terminal node. For example, we can say that:

> **IF** Balance of current account = > no running account, no balance
> **AND** Value of Savings or Stocks = > no savings, less than 100 DM
>
> **THEN** Creditability = "**bad**"

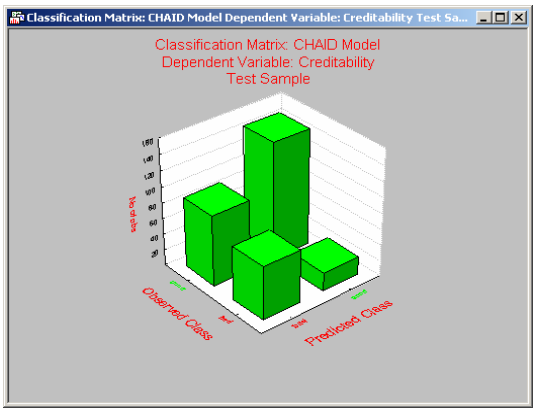
Classification Matrix - *CHAID* Model

The *Classification matrix* compares the actual classifications and predicted classifications (those that are dominant within the respective terminal node), to summarize the classification accuracy (or misclassification rate) for the different outcome categories.

The program computes the matrix of predicted and observed classification frequencies for testing the data set, which are displayed in a results spreadsheet along with the bivariate histogram as shown below.

Data: 2-Way Summary Table: Observed Frequenci...

Creditability	Classification Matrix: CHAID Model Predicted bad	Predicted good	% Correctly Predicted
Observed bad	61	31	66.30
Observed good	88	149	62.87
Totals	149	180	63.82



Classification Matrix: *CHAID* Model

The classification matrix shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified as the other category.

In this case, the overall model could correctly predict whether the customer's credit standing was good or bad with 63.82% accuracy (61+ 149)/(61+31+88+149). Note that our main goal is to reduce the proportion of bad credits predicted as good credits. The percent of correct predictions for the "bad" category is 66.30%.

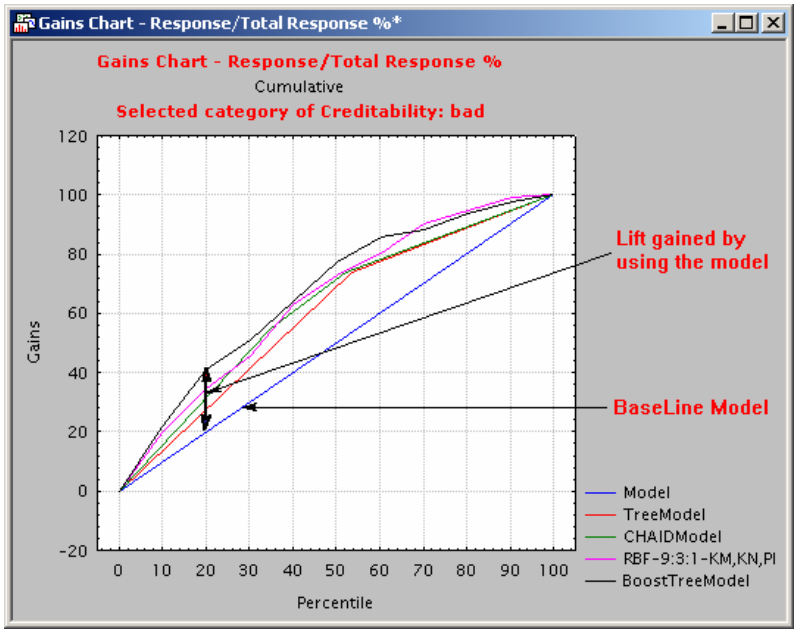
Comparative Evaluation of the Models

It is good practice to experiment with a number of different methods when modeling or mining data rather than relying on a single model for final deployment. Different techniques may shed new light on a problem or confirm previous conclusions.

Gains Chart

The gains chart provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting categorical dependent variable. Specifically, the chart summarizes the utility that one can expect by using the respective predictive models, as compared to using baseline information only.

The following overlaid gains charts were generated (for multiple predictive models) based on models trained in *STATISTICA Data Miner* using the *Compute Overlaid Lift Charts from All Models* node.



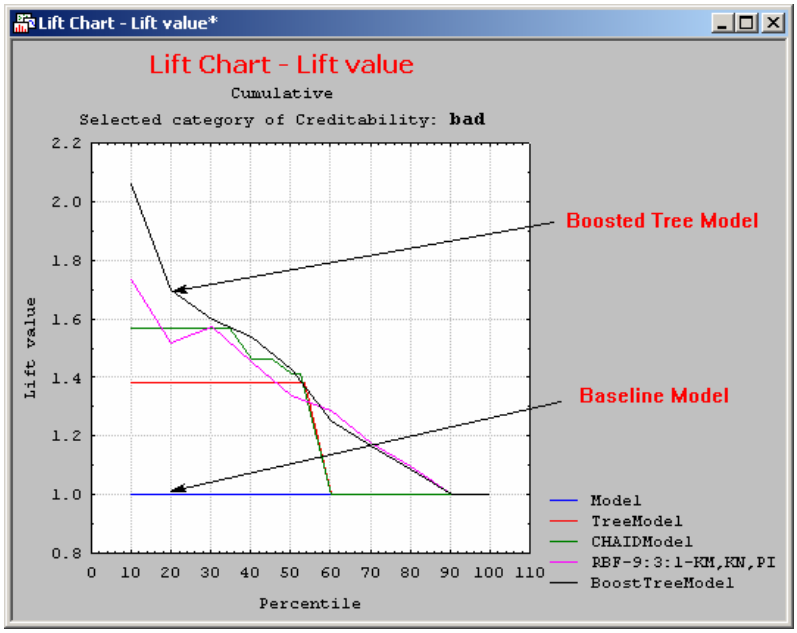
Gains Chart for "Creditability" = "Bad"

This chart depicts that the *Boosting Trees with Deployment* model is the best among the available models for prediction purposes. For this model, if you consider the top two deciles (after sorting based on the confidence of prediction), you would correctly classify approximately 40 percent of the cases in the population belonging to category "bad." The baseline model serves as a comparison to gauge the utility of the respective models for classification.

Corresponding values of Gains/Lift can be computed for each percentile of the population (in this case loan applicants sorted based on the confidence level of prediction) to determine the percentile of cases that should be targeted to achieve a certain percentage of predictive accuracy. You can see from the above graph that the gains values for different percentiles can be connected by a line and it will typically ascend slowly and merge with the baseline if all customers (100%) were selected.

Lift Chart

The following lift chart depicts that the *Boosting Trees with Deployment* model is the best among the available models for prediction purposes.



Lift Chart for "Creditability" = "Bad"

If you consider the top two deciles, you would end up with a sample that has almost 1.7 times the number of 'bad' customers when compared to the baseline model. In other words, the relative gain or lift value by using *Boosting Trees with Deployment* model is approximately 1.7.

Classification Matrix - *Boosting Trees*

Similar to what we did with the *CHAID* analysis, we can look at a classification matrix displaying the actual number of cases belonging to each class, and assigned by the model to that or other classes.

Data: 2-Way Summary Table: Observed Frequencies (Compute B...

Creditability	Classification Matrix: Boosted Trees		
	Predicted bad	Predicted good	% correctly predicted
Observed: bad	68	24	73.91
Observed: good	89	148	62.45
	157	172	65.65

Classification Matrix: Boosted Trees Model

The classification matrix for the testing data set shows the number of cases that were correctly classified (on the diagonal of the matrix) and those that were misclassified as the other category.

In this case, the overall model could correctly predict whether the customer's credit standing was good or bad with 65.65% accuracy. Our main goal is to reduce the proportion of bad credit. The percent of correct predictions for the "bad" category when using the *Boosted Trees* model is 73.91%.

Deploying the Model for Prediction

The final stage involves using the best model and applying it to new data in order to predict the good/bad customers. In this case, we will deploy the *Boosting Classification Trees* model that gave us high predictive accuracy on the test set when compared to the other models. *STATISTICA* provides a convenient way to deploy predictive models. You just need to save the PMML deployment code for the best performing model, and then use that code via the *Rapid Deployment* node in *STATISTICA Data Miner* to predict (classify) the credit risk of new loan applicants. Then the predicted/classified applicants can be sorted by the probability of the prediction to decide beforehand who would be more likely to default on a loan. This could save institutions such as banks enormous amounts of money.

Conclusion

The purpose of this example is to demonstrate how easy it is to train and use predictive models when the user has all the necessary tools available to guide him/her at each step of the model building process. *STATISTICA* also provides numerous tools for *Data Preparation/Cleaning.* The techniques provided in *STATISTICA Data Miner* represent some of the most advanced predictive techniques currently available in the market. *STATISTICA Data Miner* offers a very large selection of graphs and charts that can be combined with all other functionality of the program, allowing an analyst to use “visual data mining” techniques, or perhaps even use visual techniques (graphical methods) exclusively throughout the project. Once the model is finalized, solutions computed via *STATISTICA Data Miner* can be deployed as complete projects accessible via a single click of a button.