Paper 082-2007
Find the best prospects for a new product by using a data mining model

Ting Millette, Carmel, CA

ABSTRACT

The market department of a financial firm keeps records on customers, including demographic information and their type of accounts. The firm is launching a new product and wishes to determine who are the best prospects among the existing customers for this new product. You are asked to provide a list of 1000 best prospects.

Regression, decision tree and neural network models are built to use for scoring the prospective customers, A confusion matrix is then used to determine what the cut-off point for the scores for customers you should use to determine who qualifies as a good target. The models are compared by assessing model performance and validating the model to new data.

INTRODUCTION

Financial markets generate large volumes of data. Analyzing these data to reveal valuable information and making use of the information in decision making presents great opportunities and grand challenges for data mining. Data Mining is used to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results.

Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis. Data mining is being used both to increase revenues (through improved marketing) and to reduce costs (through detecting and preventing waste and fraud). It provides powerful tools that can reveal complex and hidden relationships in large amounts of data. Data mining already has had a major impact on business and finance. Worldwide, organizations of all types are achieving measurable payoffs from this technology.

OBJECTIVES

This paper will introduce how to build up a data mining model using SAS Enterprise Miner, how to assess model performance, and how to validate a model by targeting the 1000 best customers for a new product.

DEFINITIONS

This section clarifies some of the important concepts used in this paper.

REGRESSION

A decision tree predicts values of continuous variables. In the logistic regression model, the dependent variable is assumed to be a binary variable that whose probability can, via the logistic function, be modeled as a linear function of one or more independent variables plus an error term introduced to account for all other factors. In this study the dependent variable is whether or not the customer is a good prospect or not and the independent variables are the customer's age, gender, region, income, marital status, children, car, saving account, current account and mortgage, which are variables that are assumed to affect the customer future behavior patterns

DECISION TREE

Decision tree is a predictive model; that is, a mapping of observations about an item to conclusions about the item's target value. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents the predicted value of target variable given the values of the variables represented by the path from the root. A decision tree describes a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications. A decision tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the derived subset. A random forest classifier uses a number of decision trees, in order to improve the classification rate.

NEURAL NETWORKS

A Neural network is a complex nonlinear modeling technique based on a model of a human neuron. A neural net is used to predict outputs (dependent variables) from a set of inputs (independent variables) by taking linear combinations of the inputs and then making nonlinear transformations of the linear combinations using an activation function. It can be shown theoretically that such combinations and transformations can approximate virtually any type of response function. Thus, neural nets use large numbers of parameters to approximate any model. Neural nets are often applied to predict future outcome based on prior experience. For example, a neural net application could be used to predict who will respond to a direct mailing.

CHI SQUARE

The chi square is a statistic that assesses how well a model fits the data. In data mining, it is most commonly used to find homogeneous subsets for fitting categorical trees.

CONFUSION MATRIX

A confusion matrix shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where misfit is occurring.

DEPENDENT VARIABLES

The dependent variables (outputs or responses) of a model are the variables predicted by the equation or rules of the model using the independent variables (inputs or predictors).

INDEPENDENT VARIABLES

The independent variables (inputs or predictors) of a model are the variables used in the equation or rules of the model to predict the output (dependent) variable.

LEAST SQUARES

The most common method of training (estimating) the weights (parameters) of a model by choosing the weights that minimize the sum of the squared deviations of the predicted values of the model from the observed values of the data.

LOGISTIC REGRESSION

Logistic regression is used for predicting a binary variable (with values such as yes/no or 0/1). An example of its use is modeling the odds that a borrower will default on a loan based on the borrower's income, debt and age.

R SQUARED

A number between 0 and 1 that measures how well a model fits its training data. One is a perfect fit; however, zero implies the model has no predictive ability. R squared is computed as the square of the covariance between the predicted and observed values divided by the variances of the predicted and observed values.

TEST DATA

A data set independent of the training data set, used to fine-tune the estimates of the model parameters (i.e., weights).

TRAINING DATA

A data set used to estimate or train a model.

VALIDATION

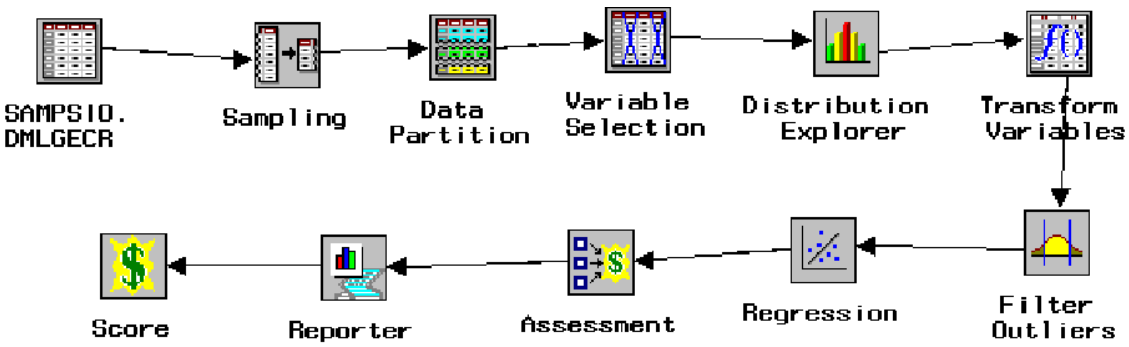
The process of testing the models with a data set different from the training data set.

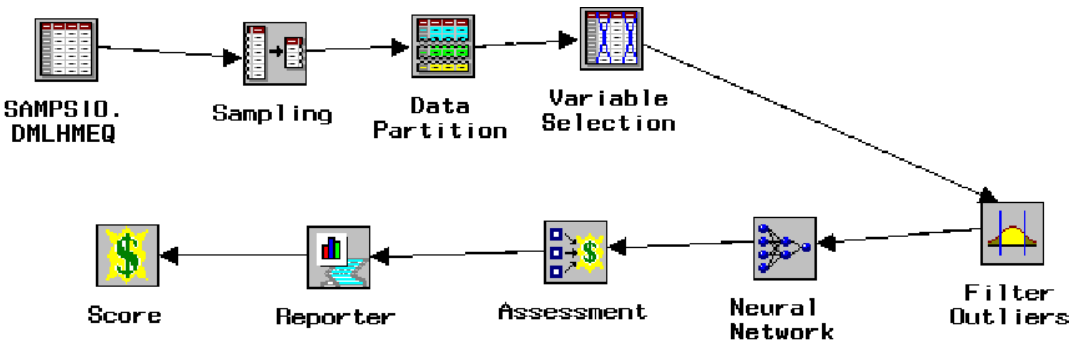
DATA PARTITION

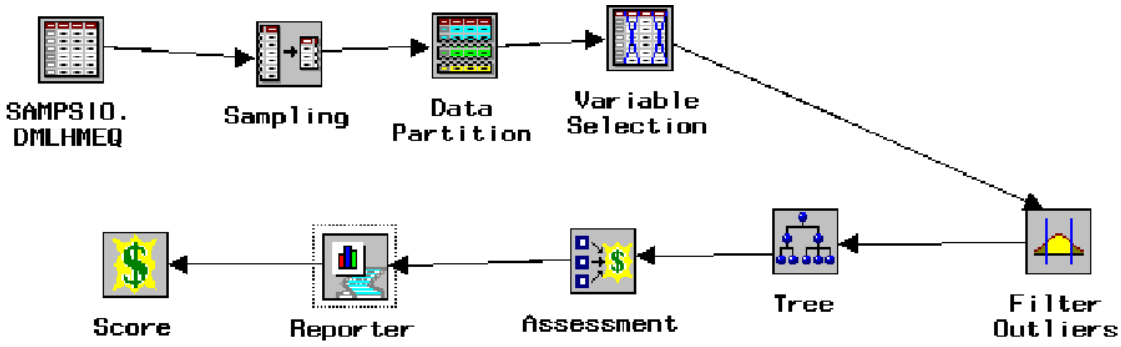
When a regression model is used as a classification tool, in order to determine how small the training data can be before the performance of the regression model degrades you experiment with the partition of the data by reducing the training data size from 50% to 40%, 30% and 20%. Then the respective cumulative response charts comparing the decision tree and neural network model are compared. As a result of the data partition process, the extent to which the training data size can be reduced to 30% before the degraded performance of the regression model occurs.

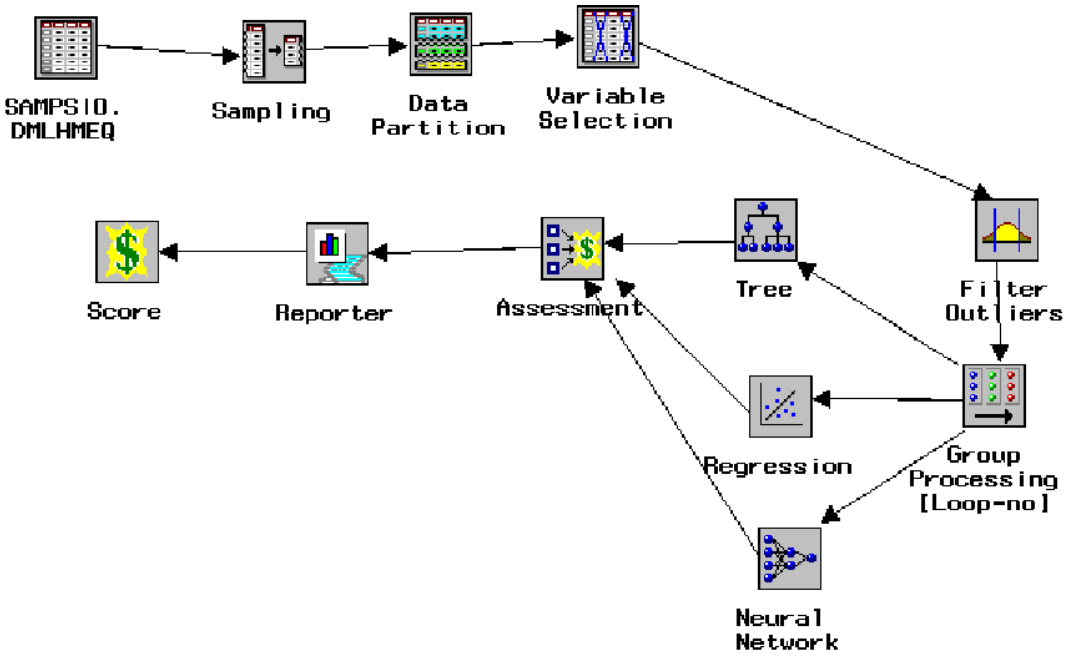
MODEL SETUP

The following charts show how to build up the models by logistic regression, decision tree and neural network respectively.






SAMPSIO.
DMLHMEQ
Sampling
Data
Partition
Variable
Selection
Score
Reporter
Assessment
Tree
Filter
Outliers


SAMPSIO.
DMLHMEQ
Sampling
Data
Partition
Variable
Selection
Score
Reporter
Assessment
Tree
Filter
Outliers
Regression
Group
Processing
[Loop-no]
Neural
Network

MODEL ASSESSMENT

1. Confusion matrix

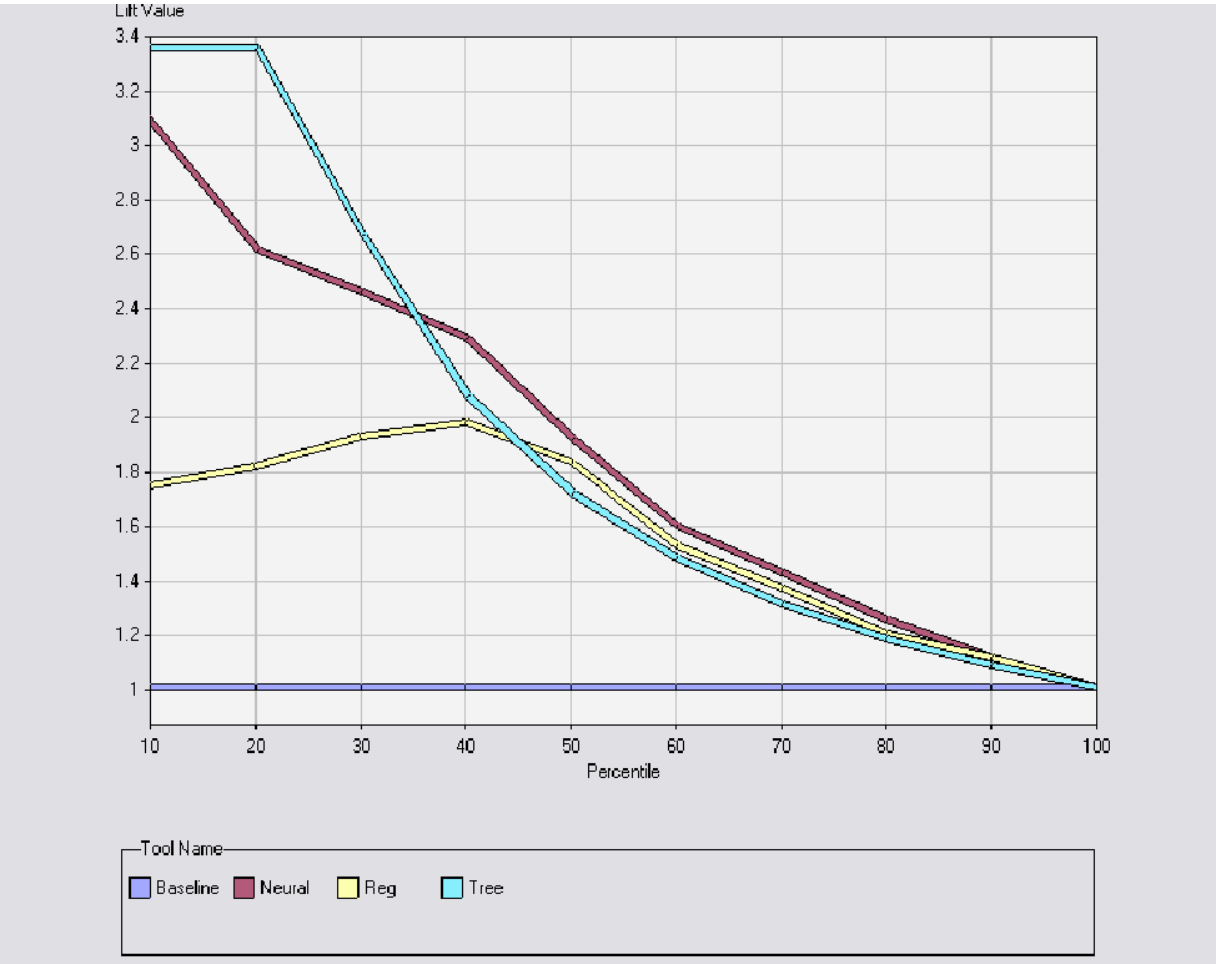
	Actual			
		buyer	Non-buyer	
	buyer	500	500	1000
	Non-buyer	500	8500	9000
Predicted		1000	9000	10000

How good is the model? The confusion matrix is optimized by maximizing the prediction rate (positive and negative) divided by the misclassification rate. Through this process, it is determined that the optimal cutoff score was 0.55 and it produced the following classification rates:

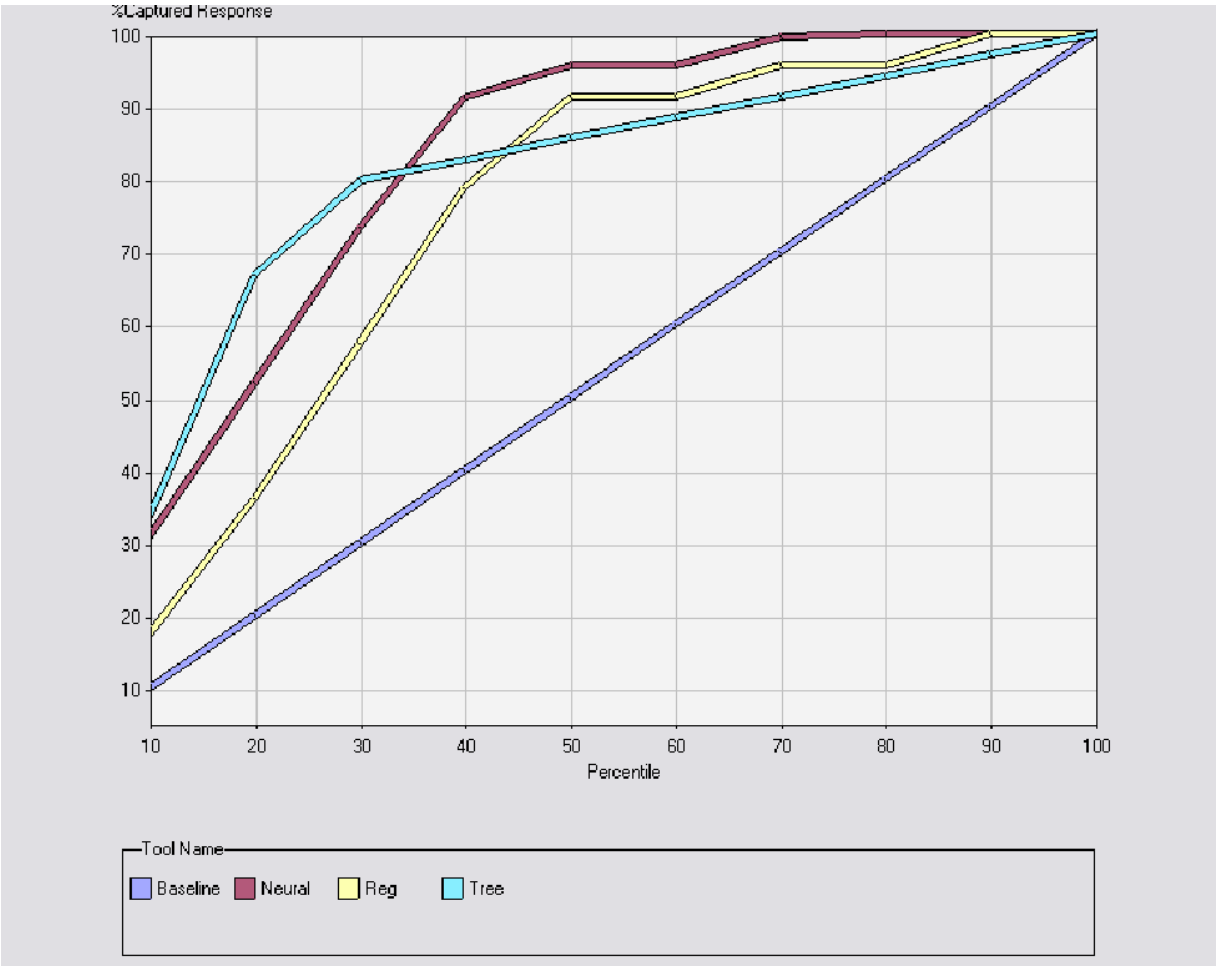
Correct prediction rate=90%
Misclassification rate= 10%
False positive rate=50%
False negative rate=5.5%

2. Cumulative response or lift charts or profit charts

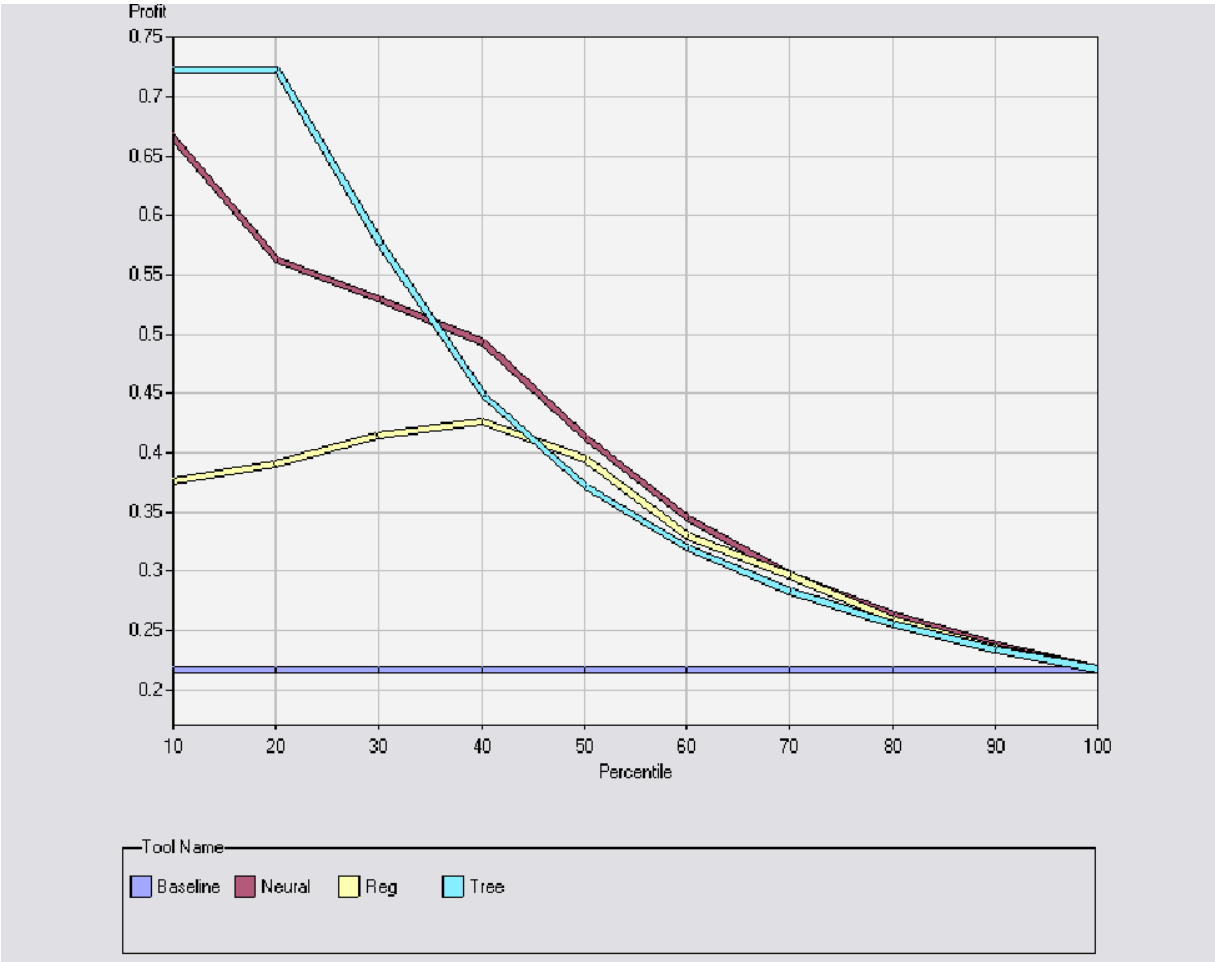
Regression has a lift of 175%, and neural network has a lift of 310% and decision tree has a lift of 335% compared to no modeling. Therefore the decision tree outperforms the regression and neural network model.



Regression can capture 90% of the buyers by targeting 50% of the potential customers, neural network can capture 90% of the buyers by targeting 40% of the potential customers, decision tree can capture 80% of the buyers by targeting 30% of the potential customers and can capture 90% of the buyers by targeting 65% of the potential customers.



Better modeling yields higher profits. The 0.75k profit earned by using the decision tree outperforms a non-targeted model for the top 20% percentile. The 0.56k profit earned by using the neural network outperforms a non-targeted model for the top 20% percentile. The 0.39k profit earned by using regression outperforms a non-targeted model for the top 20%. Again through this analysis we can tell that the decision tree outperforms the regression and neural network model.



In the regression model we can either use t tests to examine the importance of each predictor and R square or the F test to measure the model fitness. R-square and average square error are better measurement of model fit than F test. R-square can be calculated as the squared correlation between predicted values and actual values; the higher the R-square, the better the model. The lower the Average squared error, the better the model.

In the following tables it can be seen that the decision tree outperforms the regression and neural network model with smaller average squared error and smaller misclassification.

Tool	Target	Target Event	Root ASE	Valid:Root ASE	Test:Root ASE
Neural Network	buyer	1	0.4198573	0.375608677	0.377804163
Regression	buyer	1	0.42919849	0.390205283	0.359309433
Tree	buyer	1	0.29295261	0.305571021	0.341392826

Tool	Target	Target Event	Misclassification Rate	Valid:Misclassification Rate	Test:Misclassification Rate
Neural Network	buyer	1	0.230769231	0.177570094	0.175925926
Regression	buyer	1	0.237762238	0.177570094	0.166666667
Tree	buyer	1	0.097902098	0.112149533	0.148148148

MODEL VALIDATION

In this study, after model assessment, it is concluded that decision tree is the optimal model. However, You want your model to perform well on new data, not just current data. You implement the model by applying to the validation data set to verify the superiority of the selected model. After running the decision tree model, every customer gets a predicted score, and by sorting the predicted scores the best 1000 prospects are identified. You can see the improvement comparing to no-modeling: The predicted probability is 90% versus a random probability of 44% with cutoff point 0.55; the decision tree has a lift of 335%, which means it is more than 3 times more accurate than with no modeling. By evaluating the cumulative captured response and profit or return on investment chart you can maximize the profit by the optimal model.

CONCLUSION

This paper introduced how to build up the data mining models by regression, decision tree and neural network. Compare the models by assessing model performance and generalize the model by validating the model. Targeting the potential customers by optimal data mining model can help financial institution reduce costs and increase revenue.

REFERENCES

Herb Edelstein, Building Profitable Customer Relationship with Data Mining, Potomac, Maryland, 2003

Wang, H., and A. S. Weigend, Data Mining for Financial Decision Making, Decision Support Systems, Seattle, WA, 2003

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ting Millette
Carmel, CA 93923
831-277-1276
Email: wuting99@gmail.com