

WHITE PAPER

CallTech™: Modeling and Optimization Methodology

December 2009

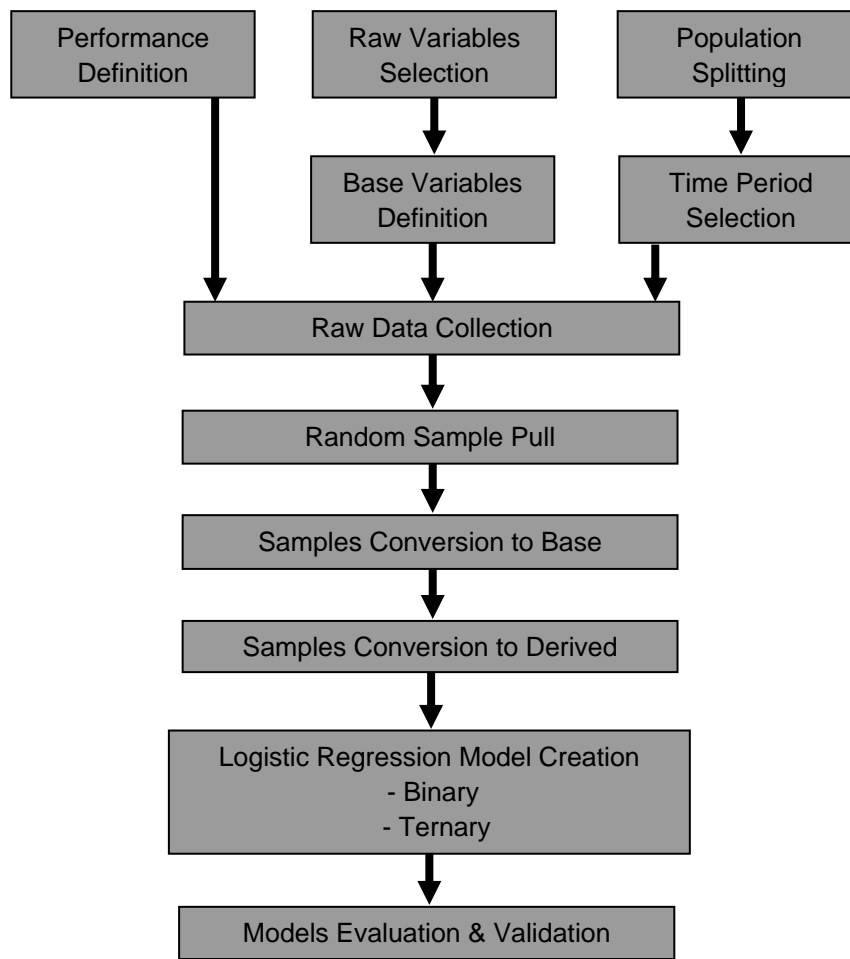
WHITE PAPER

Overview	3
Model Development	4
Daily Call Assignment Optimization	6
Summary	7
References	7

Overview

A unique approach to modeling and model-based optimization has been created in CallTech™ to leverage the customer's data to the greatest extent in solving the “Best Time to Call” problem. This process has two major phases: model development and daily optimal call assignment. This paper provides a description of both phases and the unique requirements placed on CallTech models by the manner in which they are employed. CallTech may be used for either collections or telemarketing outbound calling. In what follows, the term “record” is used to denote either an account (collections) or a prospect (telemarketing) record. In either usage, the objective is to maximize the number of right-party contact calls placed given the available operator resources. Two types of continuous logistic regression models are employed in CallTech: binary (promise-to-pay versus non-promise for collections, offer acceptance versus rejection for telemarketing) and ternary (call classification as right party contact, wrong party contact, or non-contact).

Model Development



Model Development

The Model Development diagram illustrates the development process. With a new installation of CallTech, the precise definition of call performance must be defined. Each call must be classified as either a right party contact, a wrong party contact, or a non-contact, and each right party contact call must be further classified as resulting in “promise made” (“offer accepted” for telemarketing) or not. There is often some degree of latitude in selection of precisely how to classify calls, and it is quite important that the selected classification scheme be uniformly adhered to, which may require Telephone Service Representative retraining in some instances.

For new CallTech installations, the potential modeling data available from a variety of customer sources is reviewed and a set of raw variables to be used in modeling is defined. The methods which will be employed to create the set of base variables from the raw variables are then selected and the base variable set is defined. An example of a raw variable is the occurrence/non-occurrence of a call for a particular record on a single day. A corresponding base variable would be the number of calls in the last seven days for that record. Some raw variables may be included in the base variable set without any conversion, such as current account balance. Both numerical and character variables may be included in the base variables set, and the numerical variables are allowed to have missing/invalid values. The raw and base variables sets chosen will be used uniformly for the entire population.

It is also necessary with a new installation of CallTech to define how the total calling population will be split into subpopulations. Each subpopulation is referred to as a Strategy, Strategy Class, or SC. Examples of population splits are by delinquency level (for collections) or by previous customer versus new prospect (for telemarketing). Once the family of Strategy Classes has been defined, then the starting and ending times of the calling time periods, sometimes referred to as TP, for each Strategy Class must be chosen, for the day(s) of the week for which they apply. These time periods will each have separate calling lists, and the optimization process (described later) will choose which records to assign to each time period. Ideally each hour of the calling day would be chosen as a separate time period, in order to allow the optimization process the greatest possible latitude to maximize performance. However, the definition of both Strategy Classes and time periods is driven by the availability of data on which to build models. Often time periods and even Strategy Classes must be combined to allow for sufficient sample sizes to support creation of models which are useful for prediction.

After the performance, the raw and base variables, the Strategy Classes, and the time periods are defined, raw data collection can proceed. This process continues until sufficient data is collected to support model building, and then random sampling begins. The binary models are based on the right-party contacts call results, and the samples for these are taken at the Strategy Class level. The ternary models are based on all available call results, and the samples for these are taken at the time period level. In other words, a single binary model will be developed for each Strategy Class, and a separate ternary model will be developed for each time period associated with a Strategy Class. For both types of models, a development sample is randomly selected; a validation sample is also selected if there is sufficient data. The raw data samples are then converted to base variables samples.

As has been mentioned, the same raw variables and base variables are used for the entire population, all Strategy Classes and time periods. At this point the base variables samples are converted to the derived variables samples that are actually presented to the model development programs. These programs accept only numerical variables, and the derived variables are not allowed to have missing or invalid variables. The model creation programs employ continuous logistic regression, which means that in the scorecards a single parameter

is associated with each variable. The numerical base variables are converted to derived variables as follows: the original variable is truncated at its 1% and 99% values in its tail regions, the mean value is used to replace missing/invalid values, the variable is transformed by squaring, a missing value flag is created if there are any missing values, the variable is discretized into value ranges, and a flag is created for each range. A single base variable may therefore give rise to as many as 13 derived (model candidate) variables: the truncated and imputed variable, its square, a missing value flag, and up to 10 value range flags. Note that the value range flags are treated as independent model candidate variables and are not mutually constrained as in traditional risk models. The base character variables are converted into a set of flags by a weight-of-evidence grouping process and also by associating a flag with each individual character value. It is important to note that the creation of the derived variables is done separately for each development sample, which results in a different set of derived variables (candidates for model inclusion) for each model developed. This tailoring of the modeling process maximizes the extraction of useful information from the available data. The derived variables creation process often results in a large set of candidates for a given model, and this set is reduced to a manageable number by using only those candidate variables with the highest information values.

Once the derived variables samples are ready, the model development is performed. The intended purpose of the CallTech models is markedly different from the purpose of traditional risk or behavior models. Risk models are intended to provide rank ordering of candidates, and are often discretized and constrained in order to support legal restrictions on model usage. Such legal restrictions do not exist for CallTech models, whose purpose is to create accurate estimates of probabilities. The CallTech modeling process is almost totally automated, which would be impossible with a discretized, constrained modeling approach. In addition, accurate estimation of probabilities is actually more analytically challenging than rank ordering, and the useful lifetime of probability estimation models is typically much shorter than that of rank ordering models. This shorter lifetime further reinforces the value of having an automated modeling process. The logistic regression modeling process employed in CallTech utilizes a single forward loop, followed by a single pruning pass at the end of that loop. Special methods are employed to insure that numerical stability problems do not interfere with the modeling process, to obtain the best local fit of probability estimates, and to avoid probability distribution tail region errors. The details of the methods employed are proprietary to ALI Solutions, but a general introduction to logistic regression is found in Reference 1 covering both binary and ternary models. Unlike binary models, ternary logistic regression models have two mutually dependent scorecards which are linked through the probability equation.

Upon completion of models creation, the models are reviewed to determine how well they can be expected to perform. For each model, summary statistics (mean, minimum and maximum, etc.) are automatically computed for all modeling candidate (derived) variables, and frequency and coarse classification analyses of all base and derived variables are performed. Each model's performance, as measured by the Kolmogorov-Smirnov (KS) statistic, is checked on the development sample, and compared to KS on the validation sample if one is available. KS is actually a rank ordering measure, and does not provide a direct assessment of the goodness of local fit of the probability estimates produced by a model. Goodness-of-fit tables are created for each model's probability estimators in case a detailed review is needed, but long practice has proven that use of KS is adequate for assessment of CallTech model performance. If the KS values for a given model are low relative to the overall richness and strength of the modeling variable set available, or if model percentage validation falls below expectations, then all the variables that entered the model are carefully examined. This examination typically reveals one or more raw or base variables that are invalidly defined or unreliably populated. Such variables are then eliminated and the modeling process is repeated.

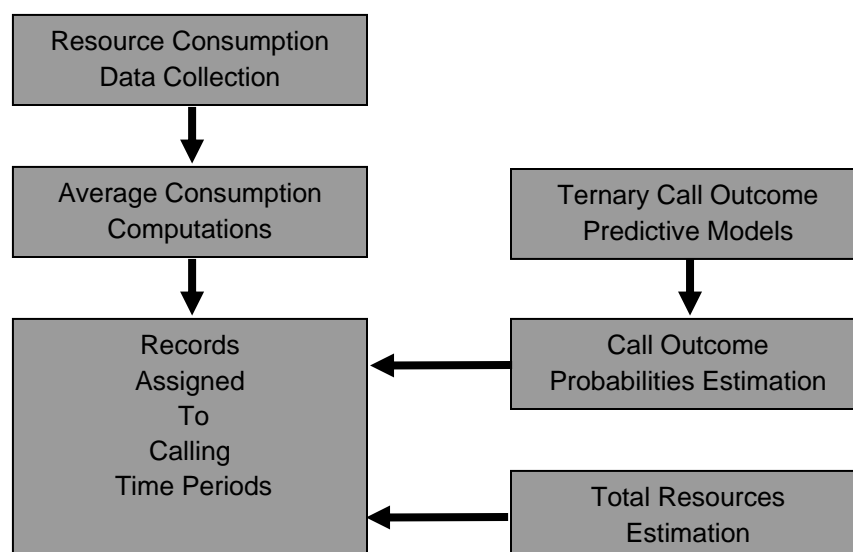
For established CallTech installations, the Strategy Class/time period structure can be refined and new models can be created as needed to keep the model set synchronized with that structure. The modeling process is also repeated at regular intervals to make sure the probability estimates continue to represent the current population.

Daily Call Assignment Optimization

An optimization process is applied on a daily basis for each Strategy Class, as shown in the Call Assignment Methodology diagram. In parallel to collecting performance and raw modeling variables data, the actual durations of right-party, wrong-party, and non-contact calls are also collected; they provide the basis for computing the average values of these three consumers of operator time, per Strategy Class. The appropriate time period models are applied to the total list of records available for calling during the coming day for a given Strategy Class, and the right-party contact and wrong-party contact probabilities are estimated for all time periods for each record. The last input needed for the optimization is the estimate of how many operator hours will be available, per time period for a given Strategy Class, on the coming day. All of this information is provided to the optimal call period assignment process separately for each Strategy Class.

The optimization process assigns the records for a given Strategy Class to its time periods in a way that maximizes total right-party contacts with the expected resources. It is possible to apply a weight to the optimizer's objective function to modify the process. An example of such a weighting factor is probability of promise-to-pay (or probability of offer acceptance, for telemarketing). The default mode of the optimization allows records to be assigned to, at most, one time period of a Strategy Class; some records will not be assigned if the total list exceeds resource capacity. The option is also available to break up the set of time periods for a given Strategy Class into a sequence of blocks (also called chunks), restricting individual records to being assigned to at most one time period within a single block, but allowing single records to be assigned to more than one block. That option is needed when deeper list penetration is required. The optimization problem solved is a particular integer programming problem that falls in the category of generalized linear network problems. A specially adapted version of the simplex method is employed to solve the optimization problem very quickly.

Call Assignment Methodology



The optimization process has a natural preference for records which have higher right-party contact probabilities. This preferential behavior creates a critical need for the upper tail region of the distribution of right-party contact probability to be accurately modeled, for every time period. That unique need is addressed within the model development process.

Summary

An overview of the key elements of the CallTech modeling and optimization processes has been provided. The specific requirements for CallTech models distinguish them from more traditional risk or behavior models, and require that different analytic approaches and techniques be used.

References

1. **Applied Logistic Regression**, David W. Hosmer and Stanley Lemeshow, John Wiley & Sons, New York (2000).

© Copyright 2009 ALI Solutions™. All Rights Reserved Worldwide. The information described in this document is furnished as proprietary information and may not be copied or sold without the written permission of ALI Solutions.