In proceedings of the Decision Science Institute, 2004. Distinguished Paper Award

# MODELING CONSUMER SITUATIONAL CHOICE OF LONG DISTANCE COMMUNICATION WITH NEURAL NETWORKS

G. Peter Zhang, Department of Managerial Sciences, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30303, gpzhang@gsu.edu, 404-651-4065
Michael Y. Hu, Department of Marketing, Graduate School of Management, Kent State University, Kent, OH 44242, mhu@bsa3.kent.edu, 330-672-1261
Murali Shanker, Department of M&IS, Graduate School of Management, Kent State University, Kent, OH 44242, mshanker@kent.edu, 330-672-1165
Ming S. Hung, Optimal Solutions Technologies, Inc., Solon, OH, mhung@optimize.com

## ABSTRACT

This study shows how artificial neural networks can be used to model consumer choice. Our study focuses on two key issues in neural network modeling – model and feature selection. Using the cross-validation approach, we address these two issues together and specifically examine the effectiveness of a backward feature selection algorithm for consumer situational choices of communication modes. Results indicate that the proposed heuristic for feature selection is robust with respect to validation sample variation. In fact, the feature selection approach produces the same best subset of features as the all-possible-subset approach.

***Keywords*: Consumer choices, feature selection, model building, cross validation, prediction risk.**

## INTRODUCTION

Understanding consumer choice is crucial to effective marketing management. Previous studies have shown that consumer choice is a function of consumer demographics, psychographics, the consumption motives and goals [3] [19], and the specific consumption situational context [28]. Accurate information on the relative importance of these variables makes it possible for firms to more effectively price and promote their products and services.

Logit models are traditionally used for predicting consumer choices [6] [27]. These models are useful for understanding and predicting brand choice behavior and examining the effects of marketing mix and demographic variables on consumers' choice of products. The limitation of this model, however, is the essentially linear form of the utility function that is used to calculate the probability or odds-ratio of making a specific choice. Although nonlinear terms such as interactions could be added into the model, the inclusion of such terms requires knowledge of the underlying structure of the utility function.

Artificial neural networks are a promising modeling tool to overcome the above-mentioned limitation of logit model as well as other linear parametric models used in modeling consumer choices. Neural networks have enjoyed increasing popularity and have been applied to a large number of business problems. For example, [17] uses neural networks and traditional time series methods to forecast state tax revenues. [12] compares a number of nonlinear methods for

predicting earnings surprise and returns. In marketing, we have found applications of neural networks in forecasting market share [2], predicting market response [10] [35], modeling repeat purchase purchasing in direct marketing [4], market segmentation [24] [36], and consumer brand choice modeling [5] [37].

As West et al. [37] point out in modeling consumer choice with neural networks, market researchers typically treat neural network models as a black box. It is evident that market researchers are not able to fully appreciate the power of neural network models. As a result these models have made very limited inroads into the standard toolbox of these researchers. The hesitation on the part of market researchers is also due to their limited understanding of how neural networks can be used for feature selection. Most neural network applications in consumer choice models rely on logit or logistic regression for variable/feature selection before subjecting the reduced set of features to neural network models for prediction purposes [37].

In this paper, we present a case study on neural network model building for consumer situational choice for long distance communication. Building a successful model that relates important consumer characteristics such as demographic and situational factors to their choice among various modes of communication is a valuable exercise to telecommunication companies. The model can help focus their effort in planning, advertising, and target marketing and thus enhance a company's competitive position. In the process, we highlight the critical issues we have identified previously in building neural networks:

- Model selection. Selection of an appropriate model is a non-trivial task. One must balance *model bias* (accuracy) and *model variance* (consistency). A more complex model tends to offer smaller bias but greater variance. Among neural networks, a larger network tends to fit a training data set better but may perform poorly when it is applied to new data.

- Feature selection. All modeling efforts should strive to achieve parsimony. So the goal here is to build a model with the fewest number of independent variables yet producing equal or comparable predictive power as larger models. For neural networks, as statistical parameter or model testing is difficult to apply, more computational intensive methods must be employed to determine the variables that should be included in a model.

## MODEL AND FEAURE SELCTION

Model selection addresses the issue of what is the appropriate neural network model for a given sample. Theoretically, model selection is based on the trade-off between *model bias* and *model variance* [15]. The bias of a model relates to the predictive accuracy of the model, whereas variance refers to the variability in the predictions. A model with low bias — by having many hidden nodes, for example — tends to have high variance. On the other hand, a model with low variance tends to have high bias.

The model bias measures the extent to which the average of the estimation function over all possible data sets with the same size differs from the desired function. The model variance, on the other hand, measures the sensitivity of the estimation function to the training data set. The

well-known *bias-plus-variance* decomposition of the prediction error is the guiding principle for model building and selection.

Feature selection is an important component of model building. The issue of feature selection is also closely related to the bias-variance or learning-generalization tradeoff discussed above. The objective of feature selection is to identify a small number of features that contribute most to model learning. Since exhaustive search through all possible subsets of feature variables is often computationally prohibitive, most of the feature selection methods use stepwise search algorithms such as forward addition and backward elimination approaches similar to those commonly used in linear statistical modeling. The forward addition approach successively adds one variable at a time, starting with one variable, until no attractive candidate remains. The backward elimination approach starts with all variables in the model and successively eliminates one at a time until only the ``good'' ones are left. Most of the feature selection algorithms in neural network research are based on the backward sequential method.

## RESEARCH DESIGN

### Data

The American Telephone and Telegraph Company maintained a residential consumer diary panel to study the consumer choice behavior in selecting long distance communication modes over time [26]. The company embarked on a major research effort to understand the effect of situational influences on consumer choices of communication modes. It is envisioned that the usage of long distance phone calling is largely situational since the service is readily available within a household and is relatively inexpensive. A demographically proportional national sample of 3,990 heads of households participated in the study over a twelve-month period. The sample was balanced with respect to income, marital status, age, gender, population density and geographic region. Each participant has to record the specifics on a weekly basis of one long distance (50 miles or more) communication situation.

The communication modes being reported are of three types, long distance telephone calling (LD), letter or card writing. Since long distance telephone calling is verbal and the other two are non-verbal, letter and card in this study are combined into one category. The dependent variable, COMMTYPE, is coded as `1' for LD and `0' for `letter and card'.

In a pre-diary survey, each respondent was asked to provide information on the usage rate of LD (MEANCALL) and written communications (MEANLET) in a typical month. Each diarist also provided information on five communication situation related variables for a specific communication that has taken place in a diary week. The selection of these factors is based on research findings in [21] [26]. These input variables will be treated as initial feature variables in our modeling effort. The seven factors are presented as follows:

1. MEANLET: Average number of cards and letters combined in a typical month
2. MEANCALL: Average number of calls in a typical month
3. TYCALL: the nature of the communication decision, whether it is `impulse' (coded as `0') or `planned' (coded as `1');

4. REASON: Reason for communication, `ordinary' (coded as `1') or `emergency' (coded as `0');
5. RECEIVER: Receivers of the communication, `relatives' (coded as `1') or `friends' (coded as `0');
6. NUMCALLS: Total number of LD calls made and received in a particular week, and
7. NUMLET: Total number of letters/cards sent and received in a particular week.

**Methodology**

As detailed in the following sections, we propose a backward-elimination procedure for feature selection. An experiment was conducted to evaluate our procedure, and it consisted of training neural networks with all possible combinations of the feature variables and computing the prediction risks of each trained network. Results from the backward elimination procedure were then compared with those from all possible combinations.

The methods to determine the appropriate network architecture can be summarized as follows.
1. Eliminating arcs whose weights are small or nonsignificant.
2. Eliminating arcs whose saliency measure is small. Saliency is typically based on the partial derivative of the SSE with respect to the arc. Methods differ in the approximation of this derivative. The *optimal brain damage* of [25] defines saliency of arc *i* as $H_{ii}w_i^2/2$ where $H_{ii}$ is the *i*-th diagonal element of the *Hessian* matrix, the matrix of second derivatives (of SSE with respect to arc weights), and $w_i$ is the weight of arc *i*.
3. Building networks with different numbers of hidden nodes and then selecting one using some performance measure on validation sample. For example, the measure used by Moody and Utans [29] is the prediction risk discussed below and it is the mean squared error on the validation set, adjusted by the number of weights.

We use a method for feature selection based on our measure of prediction risk, which is quite similar to that of Moody and Utans [29]. Moody and Utans' variable elimination approach is based on sensitivity analysis under the framework of prediction risk. Given a trained network of *n* features and *h* hidden nodes, denoted as $M_n^h$, the prediction risk can be estimated as the mean sum of squared errors (*SSE*) of a validation set *V*. That is,

$$MSE\left(M_n^h\right) = \frac{1}{|V|} SSE\left(M_n^h\right) = \frac{1}{|V|} \sum_{p=1}^{|V|} \sum_{j=1}^{l} \left(Y_j^p - T_j^p\right)^2 \quad (1)$$

where $|V|$ is the number of patterns in the validation set: $V=(Y,T)$, where *T* is the matrix of target values, *Y* the output of the network, and *l* the number of output nodes of the neural network $M_n^h$. As the validation sets in our study are all of the same size, we use the sums of square error $SSE\left(M_n^h\right)$ as a measure of prediction risk in our research. The procedure is detailed below:

1. Start with all *n* features and train a network over a range of hidden nodes; i.e., $h=0, 1, \ldots k$.
2. Select the optimal hidden nodes $h^*$ which yields the smallest sums of squared errors $SSE\left(M_n^h\right)$.

3. Reduce the number of features by 1, and train every possible *(n-1)* feature network with $h^*$ hidden nodes. Let $SSE^*\left(M_{(n-1)}^{h^*}\right)$ indicate the network with the smallest *SSE* of the *(n-1)* networks.

4. If $SSE'\left(M_{(n-1)}^{h'}\right) \leq SSE\left(M_n^{h'}\right)$, then $n=(n-1)$, and go to Step 3; otherwise, go to Step

5. Use the features selected in Step 3, train networks over the range of hidden nodes used in Step 1 and select the optimal hidden nodes $h^*$ again.

## RESULTS

To evaluate the effectiveness of the feature selection procedure, we consider all possible subsets of the seven potential feature variables identified. With all-possible-subset results, we are able to compare results from our feature selection method to those obtained from the best combination of features. In this AT&T situational choice study, we are able to consider all possible subsets due to the small number of feature variables. It will be very difficult if not impossible to experiment all possible subsets when the number of features is large.

A neural network was set up for each of the 127 possible subsets of the seven input variables. Each network was then trained using 8 different architectures (0 to 7 hidden nodes). These correspond to a total of 1,016 networks. Table 1 shows the minimum sum of squared errors (SSE) across all hidden nodes and subsets of feature variables for each validation sample. In validation sample 1, among the seven 1-variable networks, variable 6 (not shown) with 4 hidden nodes is tied with variable 6 with 3 hidden nodes with *SSE* equal to 103.87. Among the 6-variable networks, the network with 2 hidden nodes has the minimum *SSE* of 68.62. The network with the smallest *SSE* among all combination of variables and hidden nodes is shown in bold.

Results from validation sample 2 are similar to those from sample 1. Both indicate that the 6-variable network with variables 2, 3, 4, 5, 6 and 7, and 2 hidden nodes has the smallest *SSE*. Validation set 3 shows a slight difference from the other two samples. The 4-variable (variables 4, 5, 6, and 7) with two hidden nodes has the smallest *SSE*.

Next, we experiment with the backward elimination procedure. The seven input variables were trained in eight network architectures, hidden nodes from 0 to 7. With validation sample 1, Table 1 shows that the network with two hidden nodes has the smallest *SSE* of 73.73 for seven variables. With the number of hidden nodes fixed at 2, we then proceeded to examine the *SSE*s from the seven 6-variable networks. As shown in Table 3 (not shown), the network with variables 2, 3, 4, 5, 6, and 7 has the smallest *SSE*, 68.62. Further elimination of variables resulted in an increase in *SSE*. The set of variables 2, 3, 4, 5, 6, and 7 is then used to train networks with 0 to 7 hidden nodes, and the minimum *SSE* corresponds to the network with two hidden nodes. So the recommended feature set, based on validation sample 1, is variable combination of 2, 3, 4, 5, 6, and 7. The best network architecture is the one with two hidden nodes. This is the same "best'" selection suggested by the all-subset experiment.

Overall results indicate that the feature selection procedure identifies the same "best'' models as the all-possible-combination approach in all three validation samples. This suggests that the

Table 1: Minimum SSE across Hidden Nodes and Number of Variables

| # of Variables | Number of Hidden Nodes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Validation Sample 1 | | | | | | | | |
| 1 | 114.68 | 106.13 | 106.13 | 103.87 | 103.87 | 115.04 | 114.74 | 115.24 |
| 2 | 101.40 | 84.45 | 77.78 | 78.81 | 79.54 | 80.27 | 81.80 | 80.83 |
| 3 | 98.74 | 79.82 | 73.72 | 74.70 | 76.30 | 77.31 | 77.48 | 76.72 |
| 4 | 95.45 | 76.91 | 70.82 | 71.54 | 73.03 | 73.18 | 73.74 | 73.97 |
| 5 | 92.88 | 74.38 | 68.68 | 70.23 | 69.95 | 73.18 | 74.66 | 75.45 |
| 6 | 92.24 | 75.37 | **68.62** | 70.73 | 72.37 | 72.88 | 73.32 | 75.29 |
| 7 | 92.29 | 75.51 | 73.73 | 74.38 | 77.65 | 78.31 | 80.84 | 82.72 |
| Validation Sample 2 | | | | | | | | |
| 1 | 115.19 | 103.11 | 103.11 | 98.27 | 98.27 | 110.73 | 109.94 | 110.01 |
| 2 | 87.17 | 80.58 | 69.54 | 70.37 | 70.17 | 70.86 | 71.76 | 72.37 |
| 3 | 86.21 | 79.44 | 67.70 | 68.09 | 68.66 | 70.25 | 70.47 | 70.85 |
| 4 | 83.27 | 75.63 | 64.50 | 65.06 | 66.24 | 67.17 | 67.31 | 68.06 |
| 5 | 82.74 | 74.29 | 63.19 | 64.78 | 64.98 | 66.51 | 69.43 | 70.18 |
| 6 | 82.88 | 73.63 | **61.80** | 63.87 | 64.25 | 64.63 | 65.93 | 66.79 |
| 7 | 83.14 | 73.67 | 66.46 | 67.73 | 71.31 | 74.24 | 74.65 | 75.46 |
| Validation Sample 3 | | | | | | | | |
| 1 | 118.07 | 108.24 | 108.24 | 108.17 | 108.17 | 111.93 | 111.89 | 112.19 |
| 2 | 96.29 | 84.18 | 75.00 | 75.19 | 75.74 | 76.64 | 76.51 | 76.97 |
| 3 | 94.76 | 83.90 | 75.08 | 74.04 | 75.62 | 74.89 | 75.04 | 77.15 |
| 4 | 91.91 | 79.41 | **72.06** | 72.48 | 72.74 | 73.20 | 74.67 | 75.80 |
| 5 | 91.26 | 78.85 | 73.11 | 73.23 | 72.66 | 75.55 | 76.11 | 78.29 |
| 6 | 91.52 | 79.74 | 74.03 | 75.55 | 76.09 | 75.21 | 77.68 | 77.04 |
| 7 | 91.73 | 80.57 | 76.80 | 76.13 | 78.08 | 78.10 | 78.66 | 80.14 |

feature selection algorithm based on the prediction risk is quite robust judged from generalization ability for new observations. In addition, we find that networks with 2 or 3 hidden nodes are appropriate for our application.

## CONCLUSIONS

Our cross-validation experimental results suggest that the feature selection approach based on the prediction risk idea is very robust. The variables selected by the selection procedure correspond precisely to those identified by the all-possible-subset approach. Presumably the all-possible-subset procedure should be the most comprehensive and reliable approach for feature selection. Therefore, we have provided credence to the effectiveness of our backward selection algorithm.

## REFERENCES

References available upon request from the first author