KDD’99 Competition: Knowledge Discovery Contest

Jim Georges, Ph.D.
SAS Institute Inc
Jamboree Ctr, Ste 900, 5 Park Plz
Irvine, California 92614
Jim.Georges@sas.com

Anne H. Milley, M.A.
SAS Institute Inc
95 Glastonbury Blvd.
Glastonbury, CT 06033
Anne.Milley@sas.com

1. Introduction

The 1998 KDD-cup competition focused on building a supervised classification model for selecting likely donors to a charitable organization when solicited via mail. The model’s target, expected gift amount to a June 1997 donation campaign, was to be predicted from inputs reflecting the status of potential donors prior to solicitation. The models were built using a 95,412 case training data set with known responses. To judge model efficacy, expected gift amounts were calculated for a ’validation’ data set with concealed responses and submitted to competition organizers. Cases with an expected gift in excess of $0.68 were selected and the actual response amounts for these cases (known only to the organizers) were summed. The model corresponding to the largest response sum was declared the winner.

In the 1999 KDD-cup competition, emphasis shifts from producing a single prediction model to extracting unspecified interesting findings of commercial value from the 1998 competition data. This paper discusses SAS Institute’s findings.

In this paper, “interesting findings” are conceptual features of the knowledge acquired in the process of revenue maximization. The derived concepts enhance model understanding and allow for better approaches in future undertakings. To this end, Section 2 reports on a close examination of the training data and presents patterns later shown to enhance donation prediction. Section 3 uses these findings to build a model with larger validation revenue than the 1998 competition winners.

Some analysts and managers are reluctant to deploy a black-box model--like a neural network--in a direct marketing campaign because they usually want to understand of the segments targeted for solicitation. Section 4 addresses this common concern by explaining, at least approximately, the characteristics that define individuals selected for solicitation by the model in Section 3. The technique is general in nature and therefore can, in fact, be used with *any* classification model.

Before a model is deployed, an estimate of expected revenue is usually obtained, often by scoring an independent sample with known responses. Typically, little attention is paid to the variability in this estimate from one independent sample (eg the validation data) to another (eg the real world). By making certain reasonable assumptions about the probability distribution of an individual’s response, estimates of the variability in total revenue for a particular model can be calculated. These calculations provide a deeper understanding of the results possible in a given campaign. Section 5 develops this idea, provides a lower-bound estimate for variability total revenue in direct marketing models, and applies this estimate to the model of Section 3.

2. Exploratory Data Analysis

Successful statistical prediction models are built from a combination of three elements: problem-specific knowledge, historical data, and analytical savvy. When an analyst finds an irregularity in the data, additional problem-specific knowledge (such as how the data were prepared) or additional data (such as repairing defective data) may help to rectify the anomaly. Unfortunately, the rules of competition differ from those of real-world data analysis. Discoveries made in an exploratory analysis may only be reported and not clarified with authorities on the problem.

An innocuous example of a data anomaly can be seen in Figure 1, which plots the distribution of birth date as specified by the variable DOB. Four anomalies are immediately apparent. First, the number of individuals born on even years is twice that of odd years. Second, there are no individuals born before 1910. Third, there are spikes in the distributions on years ending in a zero (for example 1970). Fourth, there are a surprising number of extremely recent birth dates, some as late as 1997.

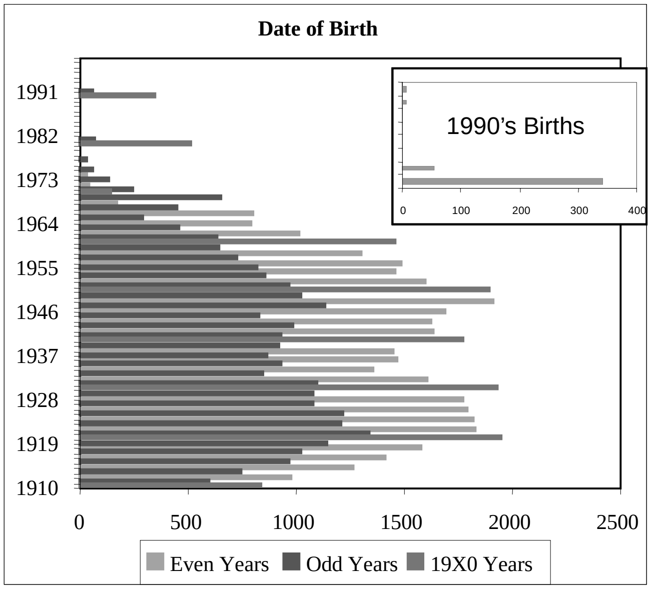


Figure 1. Date-of-birth (DOB) distribution of even/odd years.

While DOB was not used in subsequent modeling, problems in this variable could be indicative of larger problems relating to dates in general. Data integrity is vital for worthwhile analysis and for obtaining meaningful results. Third-party providers of the data may be at fault or the processes in place to handle data extraction, transformation and loading (ETL) may be questionable.

A more serious data integrity concern relates to the definition of the subpopulation composing the training data. Recency coding for the campaign of interest indicates the subpopulation consists exclusively of lapsing donors. According to the documentation accompanying the data, a lapsing donor is a previous donor who made his/her last donation between 13 and 24 months ago inclusive. For the June 1997 campaign, no donations should have been received after June 1, 1996. The data, however, tell a different story.

Figure 2 plots last gift amount (LASTGIFT) versus last gift date (LASTDATE) from July 1996 to February 1997. Almost 3,700 (4 percent) of the last gifts were received after June 1996. Clearly, not all of the subpopulation are lapsing donors.

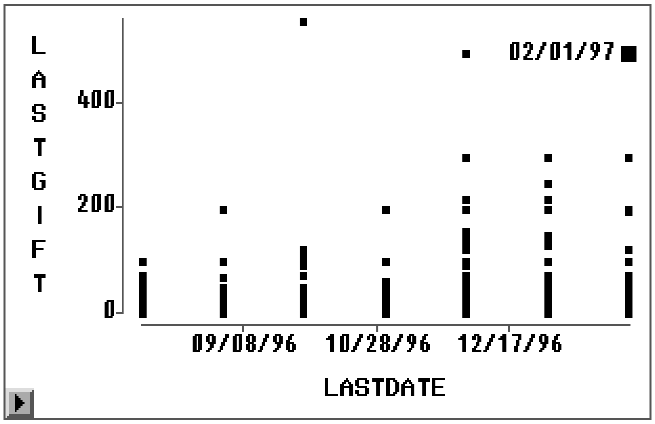


Figure 2. A plot of last gift amount versus last gift date.

Plotting last gift date versus control number, as in Figure 3, yields an even more interesting discovery: most of the gifts received after March 1996 occur with cases assigned either a low control number or a high control number.

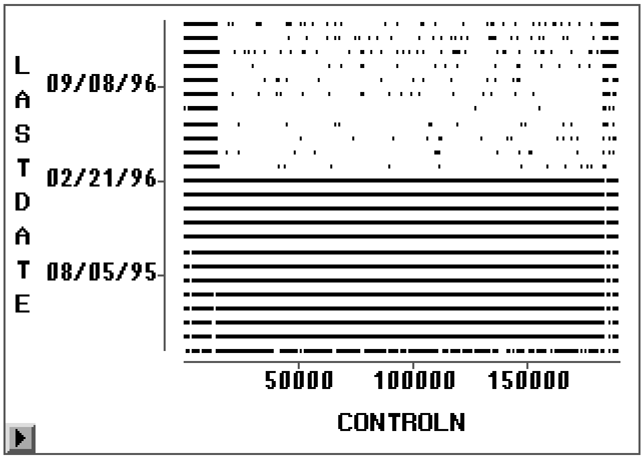


Figure 3. A surprising association between last gift date and control number.

Typically, control number is assigned sequentially as donors enter the donor database. Alternatively, the numbers may be assigned completely at random. An association between control number and one input suggests looking for others. The data provide several other examples. As Figure 4 illustrates, plots of gift amounts for the 96NK and 96TK campaigns versus control number give a similar pattern.

Finally and most importantly, as Figure 5 shows, a plot of the number of gifts in the 1997 NK campaign (that is, the target variable) versus control number yields the same pattern. This implies that the *control number is a potentially significant predictor of the target variable.*

Each horizontal bar in Figure 5 represents about 5,000 cases. The vertical reference lines indicate the expected number of responders out of 5000 (based on the marginal response rate), plus and minus two standard deviations.

In a paid analysis, such a finding would typically result in an interruption of analytic endeavors until an explanation could be provided. Here one may only speculate as to the reason for the pattern. It could simply be an artifact of sort order of the data when extracted. Or the finding could indicate a concatenation of data from at least two distinct subpopulations in the database: one or more with strong response properties, another with weak response properties. Combined they would form a richer (and more challenging) analysis data set. Another possibility is that the pattern could reflect an effort to artificially diminish the signal present in the data in order to make the competition more interesting. It is the authors' hope that such issues will be addressed as part of the results of this contest. Independent of cause, the exploratory analysis has revealed a surprising predictor of donation propensity.

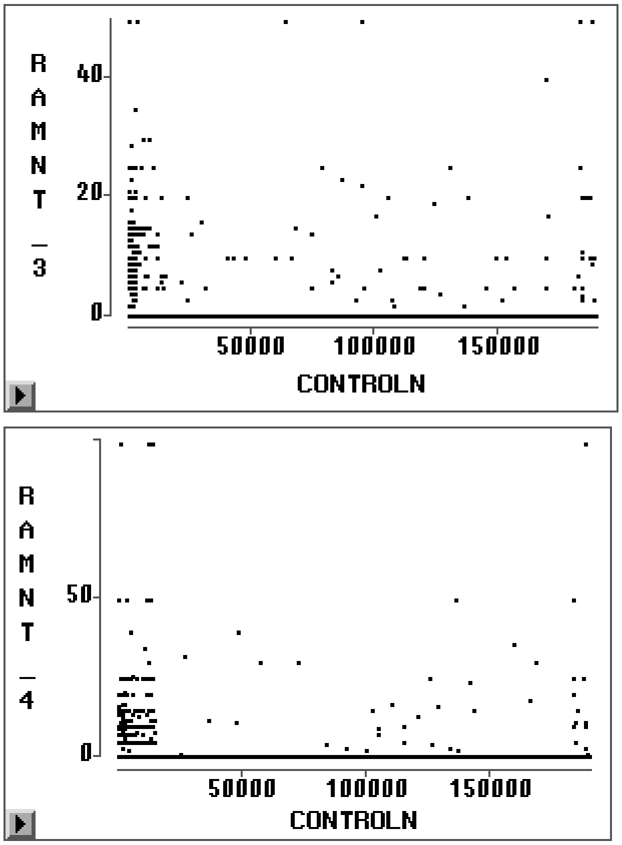
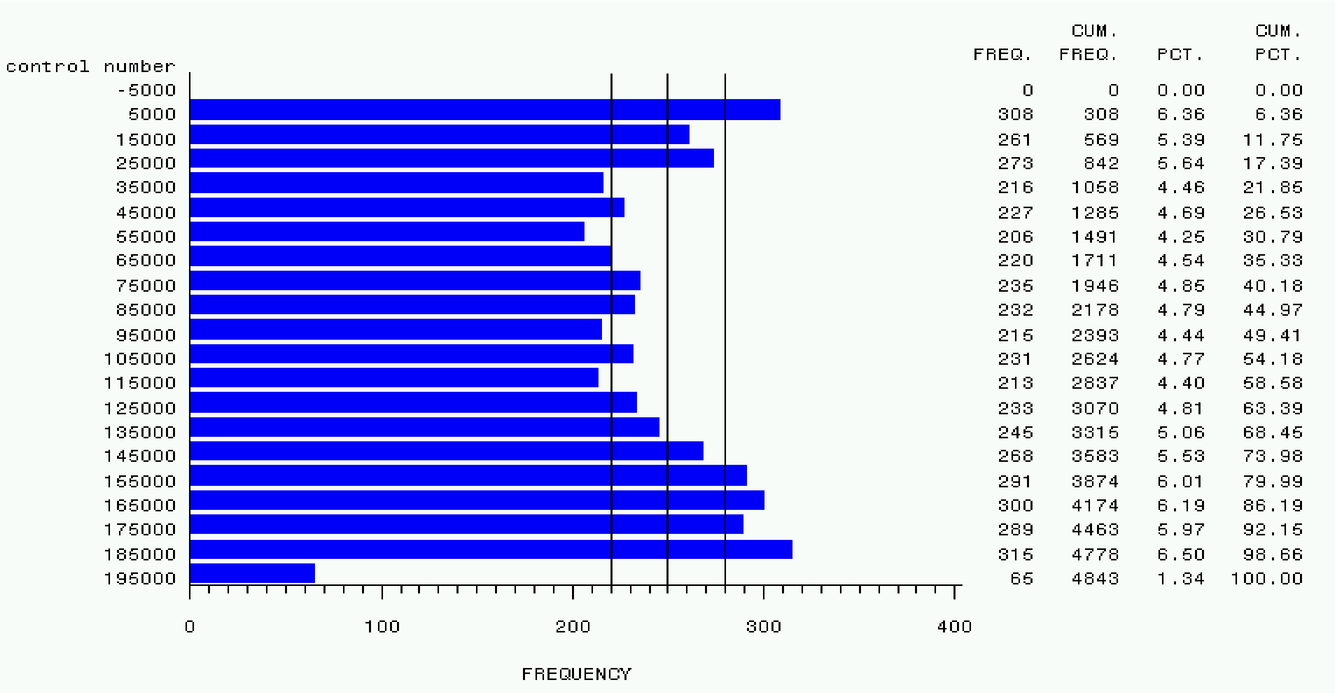


Figure 4. Gift amounts for 96NK and 96TK campaigns versus control number



```
                         CUM.                CUM.
control number   FREQ.   FREQ.    PCT.       PCT.
   -5000            0       0     0.00       0.00
    5000          308     308     6.36       6.36
   15000          261     569     5.39      11.75
   25000          273     842     5.64      17.39
   35000          216    1058     4.46      21.85
   45000          227    1285     4.69      26.53
   55000          206    1491     4.25      30.79
   65000          220    1711     4.54      35.33
   75000          235    1946     4.85      40.18
   85000          232    2178     4.79      44.97
   95000          215    2393     4.44      49.41
  105000          231    2624     4.77      54.18
  115000          213    2837     4.40      58.58
  125000          233    3070     4.81      63.39
  135000          245    3315     5.06      68.45
  145000          268    3583     5.53      73.98
  155000          291    3874     6.01      79.99
  165000          300    4174     6.19      86.19
  175000          289    4463     5.97      92.15
  185000          315    4778     6.50      98.66
  195000           65    4843     1.34     100.00
```

Figure 5. 1997 NK response counts versus control number.

3. A Two-Stage Prediction Model

The main analytic objective for a charitable organization is to accurately estimate the probability distribution of gift amount for every solicitation to a potential donor. This probability distribution can be envisioned as a mixture of a point mass at zero (indicating no gift) and a continuous distribution with positive support (indicating gift amount). Analyst's efforts are usually directed at predicting the expected gift, the product of the gift probability and the gift amount. Typically, the prediction is done in one of two ways: directly, by estimating a single quantity, the expected gift itself, or indirectly, by first separately estimating the expected donation probability and the expected gift amount, and then multiplying the separate estimates. SAS Institute's 1998 KDD-cup entry was an ensemble of both direct and indirect models.

In this analysis, an indirect or two-stage model is considered. Separate estimates of gift probability and gift amount are obtained using two multi-layer perceptron (MLP) neural networks. The gift amount model is built first using the cases where a gift actually occurred (about 5% of the data) and inputs reflecting historical patterns in the gift amount. These inputs are selected from a list of potential inputs by fitting a class-probability decision tree to gift data and eliminating all inputs not used in the tree. Then the gift probability model is built using all the cases and inputs reflecting recency, frequency, amount (RFA), and demographic characteristics as well as inputs reflecting the patterns noted in the exploratory data analysis. The inputs for the gift probability model were again chosen using a class-probability tree. One of the selected inputs, the output from the gift amount model (itself an RFA measure), serves to couple the two models.

Similar two-stage approaches occur in the econometrics literature, often referred to as limited-dependence models. For example, Heckman's two-step estimation procedure (Greene [1993]) uses the transformed output from a probit regression model (modeling the probability of an event) as an input to an ordinary least squares regression model (modeling the amount of response given the event). This is done to produce consistent parameter estimates in the regression model. While consistency in parameter estimates is important for statistical inference, its benefit in statistical prediction is less clear. In fact, for flexible, over-parameterized, highly sensitive models like neural networks, the parameter estimates are inscrutable. The predicted values themselves are the focus.

The first stage MLP has an input layer with five inputs fully connected to a hidden layer with 20 hidden units (with linear combination and hyperbolic tangent activation). The hidden layer is fully connected to a single target unit (with linear combination and identity activation). Two inputs, AVGGIFT and LASTGIFT, are as described in the data documentation. The other inputs, AMPERGFT, PGIFTH, and MYAMNT are functions of several variables:

AMPERGFT is the average gift from 94NK to 96NK.
PGIFTH is the ratio of NGIFTALL to NUMPROM.
MYAMNT is the sum of RAMNT_8, 9, 12, and 14.

The training data for the first-stage model, 4843 cases with TARGET_B=1, were split in half for training and tuning/selecting. Both early stopping and weight decay were used to protect against overfitting. Model weights were estimated via the double-dogleg optimization algorithm. Figure 6 shows predicted versus actual target. The plot shows some overestimation for smaller gift amounts and considerable under estimation for large gift amounts. Nevertheless, because these disagreements affect few cases, the agreement is deemed adequate for this analysis.

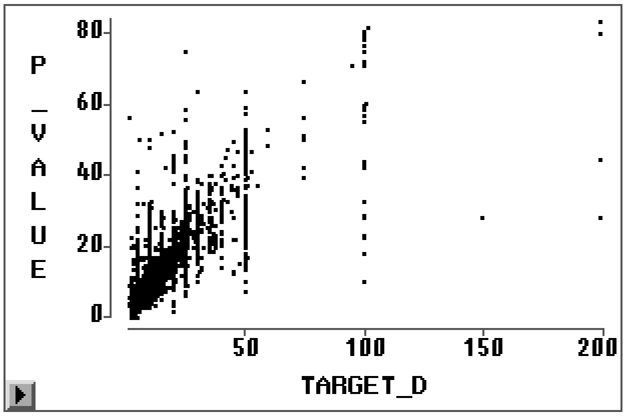


Figure 6. Predicted target versus actual target for first-stage neural network.

The second stage MLP has an input layer with eight inputs fully connected to a hidden layer with 20 hidden units (with linear combination and hyperbolic tangent activation). The hidden layer is fully connected to a single target unit (with linear combination and logistic activation). Four of the eight inputs, LASTDATE, FISTDATE, INCOME and CONTROLN, are as described in the data documentation. The remaining four inputs, PGIFTH, MYAMNT, P_TARGET, and SUSPECT, are functions of several variables:

- PGIFTH and MYAMNT are described above.
- P_TARGET is the predicted gift amount (from stage 1).
- SUSPECT estimates the ratio of gifts to solicitations from June 1996 to June 1997 and is given by the following equation:

  SUSPECT = max (min ((NGIFTALL * ((NUMPRM12 + RDATE) / NUMPROM) - RDATE) / NUMPRM12, 1), 0) where RDATE is the number of gifts specified by RAMNT_3 through RAMNT_24.

The training data for the second-stage model were split for training and tuning/selecting. Both early stopping and weight decay were used to protect against overfitting. Model weights were estimated using the double-dogleg optimization algorithm.

The expected gift amount for each case was calculated as the product of predictions from the first- and second-stage models. Expected gift amounts in excess of $0.68 were selected for mailing. The net revenue of the model using the now public validation data is $14,877.77. This is $165.53 more than the Gold-medal winner at KDD-98.

4. Understanding the Model

The two-stage model described in Section 3 has 341 parameters. Trying to understand the model by examining parameter values alone is probably impossible. One might argue, however, that trying to understand a model strictly from its parameters is analogous to trying to understand a symphony by examining the individual notes in its score. Just as a symphony is meant to be played, this model is meant to make predictions. These predictions make the model interesting, not the parameters. Understanding the predictions leads to understanding the model.

For database marketing models, two possible actions are typical: mail or not mail. Each point in the input space can be coded **zero** or **one** based on the model's predicted gift amount. Under reasonable regularity conditions, regions coded **one** can be separated from regions coded **zero** (to a specified level of accuracy) by intersecting hyperplanes orthogonal to carefully chosen axes in the input space. It may take only a few hyperplanes to approximately describe the location of most of the ones and zeroes for a given model. While not a substitute for the original model, such a description may give a reasonable idea of who is receiving a solicitation.

The intersecting hyperplanes may be thought of as branches and leaves of a decision tree. Many complicated models can be approximately understood by using decision trees to describe the regions in space where mail changes to not mail (the decision boundary). Table 1 presents such a description for the model discussed in Section 3.

The first column lists the fraction of individuals mailed in the given segment. The second describes the size of the segment relative to the rest of the data. Subsequent columns describe the membership rules for the given segment. For example, 94% of the population with amount per gift in excess of $21.10, income category 4 and above and control number less than 87,782 will receive a solicitation. Similarly 93% of the population with the amount per gift in excess of $10.23, income category greater than or equal to 4, and control number greater than or equal to 87,882 will receive a solicitation. These segments account for almost 36% of the population and more than half of the total mailing. Such descriptions form the basis of business rules describing the targeted subpopulation. (A description of the mechanism used to assign control numbers would also prove useful!).

Further scrutiny of the segment definitions reveals that the targeted population forms a single continuous region with

Table 1. Solicitation Segments.

% Mailed	% Pop	AMPERGFT	FISTDATE	LASTDATE	PGIFTH	MYAMNT	INCOME	CONTROLN
94%	5.7%	≥ 20.10					≥ 4	< 87782
93%	30.1%	≥ 10.23					> 4	≥ 87782
82%	1.9%	[7.56, 10.23)		≥ 12/16/96				≥ 127828
77%	5.2%	≥ 10.23				≤ 20.50	< 3	
75%	3.0%	≥ 10.23		≥ 10/16/95		< 20.50	< 3	≥ 127291
74%	1.0%	[10.23, 20.50)	≥ 3/16/94				≥ 6	< 87782
72%	1.4%	[7.56, 10.23)			≥ 0.245			< 127828
72%	11.7%	[10.23, 20.50)	< 3/16/94				≥ 4	< 87782
71%	0.9%	≥ 10.23		≥ 2/15/96	≥ 0.144	< 20.50	< 3	< 127291
35%	3.0%	[7.56, 10.23)			< 0.245			< 127828
33%	1.7%	≥ 10.23		< 10/16/95		< 20.50	< 3	≥ 127291
28%	7.4%	[7.56, 10.23)		< 12/16/96				
27%	0.7%	≥ 10.23		≥ 2/15/96	< 0.144	< 20.50	< 3	< 127291
23%	3.7%	[10.23, 20.50)	≥ 3/16/94				[4 , 6)	< 87782
16%	10.9%	≥ 10.23		< 2/15/96		< 20.50	< 3	< 127291
12%	11.6%	< 7.56		< 12/16/96				

respect to the input space. Solicitations are generally sent to higher-income donors or lower-income donors who have donated recently or donated larger amounts to card promotions 96GK, 96XK, 95NK or calendar promotion 96CC.

The 16-leaf tree has overall accuracy of about 83 percent. Accuracy may improve by increasing the number of branches; improvements may increase slowly increasing the number of leaves. For the example above, a 230-leaf tree achieves accuracy of only about 89 percent. Accuracy generally improves slowly when the decision boundary is a function of many variables. In this case, many steps may be needed to fit even a simple linear association.

5. Beyond Expectations

A more complete understanding of model performance may be obtained by examining not only expected gift amount but also the variability of gift amount between independent samples. We will do this by showing that the sum of individual gifts satisfies the Liapounov condition for the central limit theorem (CLT), and therefore has an asymptotically normal distribution. Under the CLT, the expected gift total is simply the sum of the expected gift amounts for each solicitation and the variance of the expected gift total is simply the sum gift variances for each solicitation. Using these facts, (and assuming correctly specified models) rough prediction limits on total revenue may be calculated for a given number of solicitations.

As mentioned earlier, the amount of gift X_i from an individual for a given solicitation may be thought of as a random variable whose distribution may be written as

$$\nu_i = (1 - p_i)\delta_{\{0\}} + p_i\phi_i$$

where $\delta_{\{0\}}$ is a point mass measure at zero, ϕ_i is normal measure with parameters μ_i and σ^2_i, and p_i is a constant representing the probability receiving a gift.

Assume that there is a negligible number of people who always give when solicited and always give the same amount of money (expressed as $p_i = 1$ and $\sigma^2_i = 0$) or people who never give any money (expressed as $p_i = 0$). Both cases imply var(X_i) = 0 and the CLT does not hold for sums of such independent variables. (This part of the sum, all virtually constants, would contribute to the degenerate part of the limit). Hence, the following (physically practical) restrictions are put on the parameters:
$0 << p_i \leq 1$, $\sigma^2_i << \infty$, (donors have positive probability of making a gift and their gift amounts have finite variance) and either $\sigma^2_i = 0$ and $p_i << 1$ (donors who give the same amount do not always give) or $\sigma^2_i >> 0$ (donors do not always give the same amount). (Note: the symbol >> is read bounded away from, uniformly in i). Moreover assume that $0 << \mu_i << \infty$.

These assumptions imply that the third central moment of X_i is bounded uniformly in i and var(X_i) is bounded away from zero uniformly in i. Thus the CLT holds by the Liapounov condition (Chung [1974]) and from the assumed form of the distribution of X_i , the random variable ΣX_i has expected value $\Sigma p_i\mu_i$ and variance $\Sigma\, p_i(\sigma^2_i + \mu_i^2(1 - p_i))$.

As expected, variance associated with the sum increases with the number of terms in the sum. In dollar terms, assume N=58,000 (the approximate mailing size for the model of Section 3),σ^2_i = 0, p_i=0.05 (the marginal gift probability) and μ_i=$13.50. Then the standard deviation of this extremely simplified situation is more than $700. Estimates from the two-stage model of Section 3 for μ_i and σ^2_i (respectively, the expected gift amount and estimated mean squared error from the first stage model), and. p_i (the gift probability from the second-stage model), give more realistic estimates of profit variance. Plugging these estimates into the variance equation and taking the square root yields a standard deviation of about $1,200. (Note that both cases fail to include the variability of fitting the model itself). Coupling this result with the expected total gift $\Sigma p_i\mu_I \approx$ $16,000 leads to a 95% prediction interval spanning $13,600 to $18,400. The wide interval suggests that there is little statistical difference between the model of Section 3 and last year's winning models (and even less between last year's top finishers).

A final comment: assume two models are of approximately equal expected revenue but different mailing depths. Smaller mailing size will, in general, lead to smaller variability in expected total gift because the variance sum, will have fewer terms. A smaller variance in expected return is usually identified as a better risk. Therefore, for models of equal expected revenue, the model with the smaller mailing depth should be preferred. This suggests as an area for further development, a selection rule that incorporates both profit maximization and risk minimization as determinants of optimum mailing depth.

Acknowledgments. We would like to acknowledge Renata Cioczek-Georges, Ph.D. uniformly in *i* and thank our colleagues Warren Sarle and Padraic Neville.

References

Chung, K. L., 1974, *A Course in Probability Theory*, Academic Press, Orlando, 2nd edition.
Greene, W. H., 1993, *Econometric Analysis*, Macmillan Publishing Company, New York, 2nd edition.
Potts, W. J. E., *Neural Network Modeling Course Notes*, Cary, NC: SAS Institute Inc., 1999.
Sarle, W. S., Neural Network FAQ, Part 2: Learning, [ftp://ftp.sas.com/pub/neural/FAQ.html].
Sarle, W. S., Neural Network FAQ, Part 3: Generalization, [ftp://ftp.sas.com/pub/neural/FAQ.html].