Spring 5-7-2010

Predictive modeling for non-profit fundraising

Qin Chen
James Madison University

Predictive Modeling for Non-Profit Fundraising

Qin Chen

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Science

Integrated Science and Technology (ISAT)

May 2010

Acknowledgement

I would like to thank Prof. Zarrugh for his encouragement and time to read early drafts of the thesis, and his many valuable suggestions for the thesis and the presentation. I would like to thank Prof. Teate for his valuable comments for the thesis and presentations. I would like to thank Prof. Deaton for his encouragement and time. I would also like to thank every professor that I took a course with for their patience and help. I would also like to thank Dr. Wang for his tireless support during the last few years.

Table of Contents

List of Tables

List of Figures

Abstract

The objective of this thesis is to compare the predictive performance of multiple regression, logistic regression, neural networks, and support vector machines (SVM) in identifying donors and non-donors, and predicting the amount of donations for a specific solicitation campaign using a data set from the Direct Marketing Education Foundation (DMEF). Non-Profit fundraising has benefited from the use of data-mining models to identify new donors, to re-solicit existing donors, and to increase the amount of donations from a solicitation campaign. Multiple regression, logistic regression, and neural networks have been widely used for non-profit fundraising. In this thesis the support vector machines are used to identify new and repeated donors, and to predict the amount of donations. The SVM models have been used mostly in machine learning and other non-business applications. The SVM models have several advantages over multiple regression, logistic regression, and neural networks. The SVM models are free from the curse of the dimensionality as in both multiple regression and logistic regression, and are also free from local minimums during training as in neural networks. The SVM models are optimization models with easy interpretations and the potential capability of handling hundreds of thousands of variables. The thesis will show how the four methods are used and will discuss the pros and cons of using each of the methods for non-profit fund raising.

Introduction

Non-profits face increased challenges in their fundraising campaigns. These challenges include the prolonged economic recession, the rise of fund raising cost (e.g. postal, printing costs and wages), and the increased scrutiny from watchdog groups. The primary objective for non-profit fundraising is to maximize the total expected donations while keeping its expenses within a fixed or even reduced budget (Key, 2001). Non-profits have been using data-mining techniques in identifying new donors and targeting existing donors through customized messages. Haughton and Oulabi (1997) compared response modeling of the Classification and Regression Trees (CART) to that of Chi-Square Automatic Interaction Detection (CHAID) in segmenting donor groups. They pointed out both CART and CHAID could be used in model building, variable selection, and identification of interactions among variables. For a specific fundraising campaign, a development officer needs to determine, among many potential donors, who have higher probabilities to donate, when it is profitable to solicit particular groups of individuals, and what customized message to send to these groups to encourage positive responses. Goodman and Plouff (1997) outlined the use of neural networks in this process to improve non-profit fundraising.

Let y_i be the variable for the amount of donations from the person i, and let d_i be the binary indicator variable for being a donor or a non-donor, where i = 1, 2, . . . , N, in a donation dataset with a total number of N records. One record is for each potential donor in the past. The value of d_i is 1 for a donor and 0 for a non-donor. Non - profits collect many variables that describe donors' behavior. These variables include RFM (Recency, Frequency and Monetary) variables and other demographic data associated with potential donors.

The RFM measures have been widely used in direct marketing and non-profit fundraising. The recency (R) refers to how recently potential donors have responded to solicitations. The more recent the potential donors have responded to a solicitation, the higher the possibility that

they would respond to a new solicitation. The frequency (F) refers to how often or how many times the potential donors have responded to solicitations. The larger the number of times the potential donors have responded to solicitations in the past, the more likely they would respond to a new solicitation. The monetary (M) value refers to the dollar amount of contributions from the potential donors. The larger the dollar amount that the potential donors have donated in the past, the higher the possibility that they would donate again. Demographic data include gender, club membership, and household status, and so forth.

Let $\mathbf{X_i} = (x_1, x_2, ..., x_w)$ be the whole collection of the variables associated with a potential donor i, for i = 1, 2, ..., N, in the donation dataset with a total number of N records. To develop fundraising models based on a donation dataset, non-profits commonly divide the whole donation dataset into training, validation and holdout subsets through certain statistical sampling schemes. This thesis uses the training, validation, and holdout data sets interchangeable with training, validation, and holdout subset. The holdout data is sometimes called testing data where the records of potential donors are not exposed to the model development process. Let n, v, and h be the number of records in the resulting training, validation, and holdout data sets, respectively. The summation of n, v and h should be the total number of records N in the donation dataset. Non-profits use the training and validation data to develop and validate models and various offerings for different groups of potential donors, estimate the size of a promotion campaign, and estimate the average amount of contributions. The holdout data is used to test the effectiveness of the models and profitability of offerings, and let the non-profits to make the final choice of which groups of potential donors to promote, and which offering or what messages to be sent to which groups of potential donors, in a specific fundraising campaign.

In using a data-mining method to develop fundraising models, we assume a functional relationship between the dependent variable, the amount of contributions y or being a donor or non-donor d, and the independent variables **X**. This functional relationship can be given by y =

$f(\mathbf{X}, \boldsymbol{\beta})$ where $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is the vector of parameters in the model. A data-mining method explores the training data, comes up with a functional form of $f(\mathbf{X}, \boldsymbol{\beta})$, decides the number of variables, or the value of p, in the model, and estimates the values of coefficients for these variables. There is a tradeoff between using a larger number of variables versus a smaller number of variables in a model. A model with a larger number of variables tends to provide better fit for the training data; however, it may over-fit the training data and perform poorly on the holdout data. A model with a smaller number of variables that fits the training data reasonably well could provide a harmonious model to perform well on the holdout data. The SAS implementations of multiple regression and logistic regression models have various choices to select a criterion that assists users to decide the number of independent variables to be used to fit a model for the training data. A good use of a validation data could potentially improve the prediction performance of a model over the holdout data (Shtatland, et al., 2004).Neural networks use all of the variables specified by the user as inputs. The validation option of neural networks balances the performance of fitting training data and predictions over the holdout data. The SVM models with its own validation option use a certain set of observations, called support vectors, of the training data to form the estimated functional form $f(\mathbf{X}, \boldsymbol{\beta})$. The study of impact of the variable selection to the performance of data-mining models is itself a major topic of research. This thesis is to use the STEPWISE option along with the SBC(Schwarz's Bayesian Criterion for model selection) and the average squared errors in validation criteria in SAS GLMSELECT to determine the number of variables and which variables to be used in multiple regression, neural networks and the SVM models to predict the amount of contributions. In order to compare the classification performance of logistic regression, neural networks and the SVM models for donors and non-donors, each of the three models is to use the same set of variables identified by the logistic regression model.

There are two basic assumptions that guide the research and practice in non-profit fundraising: The first assumption states that the behavior of donors could be modeled by a functional relationship between the dependent variable and the set of independent variables. Non-profits often need to predict the amount of contributions in planning their fundraising activities. A regression model can be used for the purpose. The amount of contributions y is described as the function f(**X**, **β**). Non- profits also need to identify donors and non-donors. A classification model of d = f(**X**, **β**) is often used, where d = 1 for donors and d = 0 for non-donors. The second assumption states that the functional relationship between the dependent and independent variables in fundraising can be discovered through some data-mining models, such as, multiple regression, logistic regression, neural networks and the SVM models.

The objective of this thesis is to compare the predictive performance of multiple regression (Neter, et al., 1996), logistic regression (Menard, 2001), neural networks (Zahavi, and Levin, 1997), and support vector machines (SVM) (Vapnik, 1995) in identifying donors and non-donors, and predicting the amount of donations for a specific solicitation campaign using a data set from the Direct Marketing Education Foundation (DMEF). Non-profit fundraising has benefited from the use of data-mining models to identify new donors, to re-solicit existing donors, and to increase the amount of donations from a solicitation campaign. Multiple regression, logistic regression and neural networks have been widely used for non-profit fundraising. Malthouse (2002) developed a performance based prediction modeling approach and tested it on the same DMEF data set one to show possible improvement of donor contributions. In this thesis the support vector machines are used to identify new and repeated donors, and to predict the amount of donations. The SVM models have been used mostly in machine learning and other non-business applications. The SVM models have several advantages over multiple regression, logistic regression and neural networks. The SVM models are free from the curse of the dimensionality as in multiple regression and logistic regression, and are free from local

minimums during training as in neural networks. The SVM models are optimization models with easy interpretations and the potential capability of handling hundreds of thousands of variables (Vapnik, 1995). The thesis will show how the four methods are used, and discuss the pros and cons of using each of the methods for non-profit fundraising. The primary contribution of this thesis is to introduce the use of the SVM models to non-profit fundraising.

The thesis has six chapters: Chapter 1 reviews existing data-mining techniques applied to non-profit fundraising, and defines research questions for this thesis. Chapter 2 describes the DMEF data set one, the dependent, independent variables and descriptive statistics of these variables. Chapter 3 briefly describes the multiple regression, logistic regression and neural networks models for non-profit fundraising. In Chapter 3, the SVM method is introduced to identify donors and non-donors, and to predict the amount of contributions. Chapter 3 also describes the performance measures to be used to compare the predictive capabilities of the four models. Chapter 4 designs the experiments of comparisons. Chapter 5 summarizes the results of the comparisons. Finally, Chapter 6 discusses the results and presents concluding remarks.

Chapter 1: Data-Mining Techniques Applied To Non-Profit Fundraising

Haughton and Oulabi (1997) presented Classification and Regression Trees (CART) and Chi-Square Automatic Interaction Detection (CHAID) techniques in direct marketing. CHAID is used for detecting significant discriminating factors of responses versus non responses with binary indicator independent variables. Effectively recruiting pledgers as a proxy for potential legators, Cole, et al. (2005) used CHAID and logistic regressions for each of the two groups: cash donors and committed givers.

Key (2001) discussed in detail the use of SAS best subset option to reduce the number of candidate variables from over 170 to 50, then to a limited few, and use a validation data set to identify the best model among a few candidate models. Key addressed issues about variable transformations and derivations of new variables in preparing a list of candidate variables to feed the SAS best subset procedure. Key tested her approach on a data set with 10,828 records from a small, private, Catholic high school in the Northeast region of the United States, and on a large data set with 160,484 records from a large metropolitan museum in the mid-western region of the United States. Key achieved promising results from both studies.

Malthouse (2001) studied the single-split method in direct marketing, and argued that the sampling variations across splits might impair one's capability to select superior models from many competing ones. He summarized the use of Winsorization and stratified sampling to reduce the sampling variations through an empirical study.

Malthouse (2002) argued that the optimized fit of a model to a training data set is not the primary objective of a scoring model. The word "scoring" refers to the process that predictive model learned from training data assign a score for each potential donor in holdout data. The higher score for a potential donor, the higher the chance for the potential donor to donate and to donate more. The primary objective of a scoring model should be "to choose n names such that the expected total sales generated from these offers are maximized." Malthouse proposed a

performance based scoring model to estimate the parameters using weights derived from the validation data, instead of the training data for fitting a model. A moderate improvement of 3% to 4% on the average for the mailing depth of between the top 20% and 40% of the donations data sets is reported. The improvement could amount to several hundreds of thousands of dollars if the promotion is sent to several millions of potential donors. Malthouse cautioned against the inclusion of a large number of candidate variables because the computations for the weights in his method are expensive.

Goodman and Plouff (1997) indicated that 80% of the donations are normally from the top 20% of donors who contributed the largest gifts during the previous year. They pointed out that non-profit fundraising could benefit from the use of neural networks to locate the top 20% of potential donors besides its use to obtain a new donor list, cultivate repeated donors, and encourage positive responses through targeted messages. Neural networks can also help non-profits to estimate the long-term value of a donor to the institution, to select potential prospects that are likely to donate, to differentiate the potential donors who are more likely to respond to one specific appeal, and to select the potential donors who are likely to donate more than before. Deichmann, et al. (2002) compared performances of logistic regression and multiple adaptive regression splines (MARS) in direct response modeling on DMEF data set two, another educational data set from the Direct Marketing Educational Foundation. Ha, et al., (2005) described a bagging neural network model for response modeling. Based on incremental break-even decision rules, Hansotia and Rukstales (2002) developed an incremental value model to select the next customer to mail to.

In the middle of the 1960s, two groups of researchers, one led by Vapnik, Lerner, and Chervonenkis in Russia (Vapnik and Lerner, 1963; Vapnik and Chervonenkis, 1964). And the other led by Mangasarian (1964, 1969) in the United States, pioneered the development of the SVM for classification and regression. A sound statistical learning theory has guided the research

in the SVM for the last three decades. The SVM models initially are used for applications in machine learning, and recently are brought to business and direct marketing. Burges (1998) provided a comprehensive tutorial on SV classifiers. Schölkopf and Smola (2002) discussed in-depth the SV regression models. Cui and Curry (2005) proposed the use of the SV classification model in marketing, and tested their approach with simulated data for classifications. No research has been done to utilize the SVM models in non-profit fundraising as of this writing in identifying donors and non-donors or in predicting the amount of contributions.

This thesis examines multiple regression, logistic regression, neural networks and the SVM in modeling non-profit fundraising. This thesis has two major objectives: one is to predict the amount of contributions with the regression type of models; and the other is to identify donors and non-donors. To achieve the first objective to predict the amount of contributions, we compare the predictive performance of multiple regression, neural networks, and the SV regression models. To achieve the second objective to identify donors and non-donors, we compare the classification performance of logistic regression, neural networks, and the SV classification models. The brief descriptions of these models are given in Chapter 3. We use SAS PROC GLMSELECT, PROC REG; SAS Enterprise Minor; and R implementations of the SVM models in the libsvm (Chang and Lin, 2001) to carry out the comparisons.

Chapter 2 Descriptions of Direct Marketing Education Foundation (DMEF) Data Set One

2.1 DMEF Data Set One

The data set used in this thesis is the Direct Marketing Education Foundation (DMEF) Academic Data Set One, coded as DMEF01, for Non - Profit Organization Fund Raising through direct mail. The base time period of the DMEF01 is from 10/1986 to 6/1995 with the latest 1 to 10 donations by date, dollar amount and the solicitation source, the latest 1 to 11 solicitations by date and type, plus some lifetime elements of the relationship and minimal demographics, such as Sex, Zip Code, and State. The later time period of the DMEF01 is from 10/1995 to 12/1995. During the later time period, all of the potential donors in the file received at least one solicitation mailing in early 10/1995. The predictive modeling task in this thesis is to use the information about donors, non-donors, and the amount of contributions before 6/1995 to develop multiple regression, logistic regression, neural networks, and the SVM models to predict who are to be donors, therefore, to mail them the solicitations; and to predict who are the non donors, therefore, not to mail them the solicitations in early 10/1995. The amount of predicted donations is also estimated. The time gap between 6/1995 and 10/1995 is to provide time for business to develop predictive models, to select potential donors for promotion mailing, and to prepare the mailing packages. The solicitation mailings were sent out in early 10/1995. The actual responses or donations from the solicitations were recorded and used to assess the performance of the predictive models. The data set is intended for educational purposes only. Many of the data fields are with missing data. This thesis uses the data to demonstrate the use of the SVM models as a classification method to identify donors and non donors, and as a regression method to predict the amount of contributions.

2.2 Summary of the Actual Fundraising Performance

As shown in DMEF01, the latest solicitation campaign started in early 10/ 1995. All of the historical donation data were collected before 6/1995. Multiple promotion offers with distinct

promotion codes for each offer were sent to potential donors in the fall of 1995. The promotion message of each promotion code differs from each other to offer customized messages to different groups of potential donors. Therefore, a promotion offer with a higher average amount of revenue would perform better than the one with a lower average amount of revenue. Each of the 99,200 people in DMEF01 received at least one solicitation offer in the fall of 1995. The promotion codes for the campaign in the fall of 1995 can be used to identify which promotion offer was sent to whom in the data set.

The actual promotion performance for the fund-raising campaign in the fall of 1995 can be described by the number and the amount of donations in the fall of 1995. There are two variables used for the purpose, TARGRESP and TARGDOL. The first variable, TARGRESP, is the number of responses to the solicitations in the fall of 1995. Among the 99,200 people who received solicitations, over 27.4% of them donated once, 0.05% of them donated twice, one person donated three times, and 72.6% of them did not respond. The second variable, TARGDOL, is the dollar amount of donations in the fall of 1995. The dollar amount of donations in the fall of 1995 ranged from $0 to $1,500 with $2.33 as the average dollar amount of donations. The frequency distribution sorted from the largest to the smallest of TARGDOL shows that the top 11.74% of donors contributed an average of $14.10, and the second top 10.60% of donors contributed an average of $5.15. The top 22.33% of donors contributed an average of $9.86 in the fall of 1995.

2.3 Estimation of Fundraising Cost and Revenue

Among the cost items of a fundraising campaign were the costs of mailing selection, package design, printing, postage, and so forth. The total revenue was more than $228,000 as reported for the campaign in the fall of 1995. The total revenue minus the cost was the total net revenue of the campaign. For the DMEF01 data set with $2.33 average dollar amount of donations, the net revenue should be positive if the average cost per mailing of the campaign was

less than $2.33. In general, a decision should be made in terms of the estimated net revenue for the fundraising campaign. Therefore, the total revenue should be higher than the total cost for the fundraising campaign to be viable.

2.4 Variable Definitions

There are a total number of 99,200 records of potential donors and 77 fields or variables for each record in the DMEF01 data set. Almost all of the indicator variables and the variables associated with contribution and solicitation dates are not used in this study because of the large amount of missing data in these fields. The models in this thesis use twelve variables out of the 77 variables plus another fourteen variables derived from the twelve variables based on Malthouse (2002). The definitions of the variables are in Appendix H.

This thesis is to study the classification performance of the logistic regression, neural networks, and the SVM models for identifying donors and non-donors; and to study the predictive performance of the multiple regression, neural networks, and the SVM models for predicting the amount of contributions for the fundraising campaign in the fall of 1995. Let TARGDON be a binary indicator dependent variable used to classify donors and non-donors. TARGDON = 0 if TARGRESP = 0 and TARGDOL = 0 for non-donors, and TARGDON = 1 for donors responded the solicitations in the fall of 1995.

The variables in DMEF01 excluded from this study include ZIPCODE, STATCODE, CHNGDATE, MEMBCODE, PREFCODE, REINDATE, and any variable with missing data. Twelve original variables are included in this study based on Malthouse (2002). These twelve variables are used as independent variables and are used to form another fourteen transformed variables as suggested by Malthouse (2002). The recency variables measure how recently donors have responded to solicitations. The recency variables are derived from the date of 6/1995 minus the corresponding contribution date in the past in month. The monetary variables measure how much money donors have contributed. The monetary variables are derived

from the variables related to the amount of contributions prior to 6/1995. The frequency variables measure how many times donors have donated. The frequency variables are derived from the variables related to the number of times of contributions prior to 6/1995.

2.5 Data Preprocessing and Preparations of Training, Validation, and Holdout Data Sets

The SAS program in Appendix C preprocesses DMEF01 data, keeps only twelve out of over 77 variables, and then derives another fourteen variables as in Malthouse (2002). The SAS program in Appendix B uses a procedure suggested by Suhr (2010) to use SAS SURVEYSELECT to take stratified random samples from DMEF01 raw data to form the training, validation and holdout data sets used in this study. The use of stratified sampling ensures the equal proportions of donors and non-donors in training, validation and holdout data sets. The random number seeds for the stratified random samples are included in Appendix A. As suggested in Malthouse (2002), one third of the 99,200 records of the DMEF01 data are for each of the training, validation, and holdout data sets. SAS LOGISTIC, SAS REG, and SAS Enterprise Minor have no difficulty to develop models over the training data. The open source SVM program in R from libsvm (Chang and Lin, 2001), however, is extremely slow. Smaller training samples of 10% of 99,200 records are generated. All of the sampling memberships for each person in DMEF01 in each sample are stored in a SAS data set. The SAS program in Appendix D reads the stratified sampling memberships from the sampling SAS data set, and prepares training, validation and holdout data sets.

The SAS programs in Appendix E are used to estimate multiple regression and logistic regression equations, and to rank the performance for the top performance groups of donors. The R program in Appendix F trains and validates the SVM models to identify predicted donors and the predicted amount of contributions. A SAS program in Appendix G is used to summarize the performance of the SVM for the data. The data preprocessing or conversions required by neural networks are carried out with the default setting for neural networks in SAS Enterprise Minor.

The use of the SVM procedure in libsvm requires the values of independent and dependent variables to be normalized to be between 0 and 1 or -1 to +1. The following equation normalizes the input data:

$$I = Imin + (Imax - Imin) * \frac{D - Dmin}{Dmax - Dmin}$$

where I is the normalized value of the variable, Imin and Imax are the minimum (0) and the maximum (1) values of the normalized values of variables, mostly determined by the models, e.g., the SVM regression and classification may require its input data to be between 0 and 1 or -1 to +1 for efficient processing. The neural networks model in SAS Enterprise Miner automatically normalizes data for better performance.

Chapter 3 Data Mining Methods and Performance Measures for Non-profit Fundraising

3.1 Multiple Regression Models

Non-profits use multiple regression models to predict the amount of contributions in fundraising (Malthouse, 2002). By assuming a linear relationship of the amount of donations and the factors or variables involved, a general linear model for donations has the form (Neter, et al., 1996)

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

where: y_i is the dependent variable for the amount of donations from the i^{th} donor, for i = 1 to N, where N is the number of records in the data. x_{ji} is the value of the j^{th} independent variable to measure RFM (Recency, Frequency and Monetary) or demographic data for the i^{th} donor (Diamond and Noble, 2001; Malthouse, 2002). ε_i is the uncorrelated normally distributed error term with zero mean and constant variance to count for any variations that cannot be explained in the regression model. The variable x_i supposes to be independent or uncorrelated. $\beta_1, \beta_2, \ldots,$ to β_p are the parameters to represent the marginal contributions of independent variables to the amount of donations y, where *p* is the index for the number of independent variables to be included in the model. The values of the βs have to be estimated from the historical data of donors. Therefore, the multiple regression equation to be used has the form:

$$\hat{y}_i = b_0 + b_1 x_{1i} + \cdots + b_p x_{pi}$$

where, $b_0, b_1, \ldots, b_p$ are the estimates of the parameters of $\beta_0, \beta_1, \ldots,$ and β_p. The ordinary least squares procedure is commonly used to determine how many variables are to be included in the model, and how much each variable would contribute to the predicted amount of donations *y* by minimizing the following summation of squared errors (SSE).

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Several different options are available in SAS GLMSELECT to select the number of variables to be included during the model development phase. Variables can be entered into a model in the order specified by the researcher or a regression model can test the fit of the model to the training data after each variable is either added or deleted in a procedure called Stepwise regression. A specific variable x is statistically significant to be included in the model if the probability value (p –value) for the coefficient of the variable x to be as small as the one observed is smaller than the given significance level. The solution for the estimates of the coefficients (bs) involves the use of a $p \times p$ matrix; therefore, more records are required to solve the matrix if the value of p is large. This is the so-called curse of dimensionality. Malthouse (2002) discussed in more detail the multiple regression models used in direct marketing.

In order to explore the relationship between the amount of donations and the relevant variables, Malthouse (2002) derived fourteen variables from the original variables in the DMEF01 data set. Malthouse (2002) did not use any binary indicator variables in his study. Frequencies and descriptive statistics are obtained for these binary indicator variables. The results indicate that the binary indicator variables in DMEF01 data set are not useable in further analysis due to the amount of missing data associated with these variables. Malthouse(2002) also transformed a few of the original variables to derive another fourteen variables. He used 20 variables in his study in 2002. We include all of the 20 variables in Maltihouse (2002) plus additional four variables from the data. Because some of the variables are directly derived from the others, they might suffer from the multicolinearity. This is one of the areas to be evaluated in the future study for DMEF01 data set.

3.2 Logistic Regression Models

In some situations, e.g., deciding whom to send solicitations, we are only interested in the

probability of being a donor. If the probability for someone to be a donor is higher than a predefined threshold, i.e., 0.5, then the person is labeled as a predicted donor with a probability p; otherwise, the person is labeled as a predicted non-donor with a probability 1−p. In this case, a binary indicator dependent variable d_i can be used with $d_i = 1$ for a predicted donor, and $d_i = 0$ for a predicted non-donor. The independent variables x_j for j = 1 to q include real valued RFM variables, categorical and binary indicator variables, e.g., the club membership, gender, and so forth. The relationship between the dependent variable d_i and the set of independent variables is not linear. Logistic regression is often used to model the functional relationshp between the binary indicator dependent variable d_i and the set of factors or independent variables that describe the donors' behavior. There are three main reasons for using logistic regression instead of multiple regression when the dependent variable is a binary indicator variable (Boslaugh and Watters 2008, page 284): 1) The assumption of a constant variance for the error term in linear regression models is not valid when the dependent variable d_i is an indicator variable with d_i as either 1 or 0; 2) The binary indicator variable d_i has values of 0 or 1— however, a linear regression may predict values smaller than 0 and larger than 1; and 3) The logistic regression can be used to estimate the odds ratio for each of the categorical or binary indicator independent variables. Non-profits could derive valuable information from these odds ratios to assist their decision making.

The logistic regression model is a logit transformation of p as follows:

$$p(d_i = 1|(\boldsymbol{x}, \boldsymbol{b})) = \exp\left(\sum_{i=0}^{p} \beta_i x_i + \varepsilon\right) / \left(1 + \exp\left(\sum_{i=0}^{p} \beta_i x_i + \varepsilon\right)\right)$$

where β_0 is the constant when all of the values of the independent variables of x are zero, and β_j is the population parameter of the j^{th} independent variable x_j and ε is the error term. An estimated logistic regression function can be given as:

$$p(d_i = 1|(\boldsymbol{x}, \boldsymbol{b})) = \exp\left(\sum_{i=0}^{p} b_i x_i\right) / \left(1 + \exp\left(\sum_{i=0}^{p} b_i x_i\right)\right)$$

where b_0 is the constant when all of the values of the independent variables of x are zero, and b_i is the coefficient of the i^{th} independent variable. An alternative form of the logistic regression equation is:

$$logit[p(x)] = \log\left(\frac{p(x)}{1-p(x)}\right) = \sum_{i=0}^{p} b_i x_i$$

where p/(1-p) is called the odds of donors to non donors. The goal of logistic regression is to predict correctly the category of outcome for individual donors and non-donors using the most parsimonious model. To accomplish this goal, SAS Logistic uses three asymptotically equivalent Chi-Square tests to determine whether a predictor variable x is significant enough to be included in the model. This thesis uses the default setting of SAS Logistic to select the number of independent variables used in the logistic regression, neural networks and the SV classification models to classify donors and non-donors.

3.3 Neural Networks Models

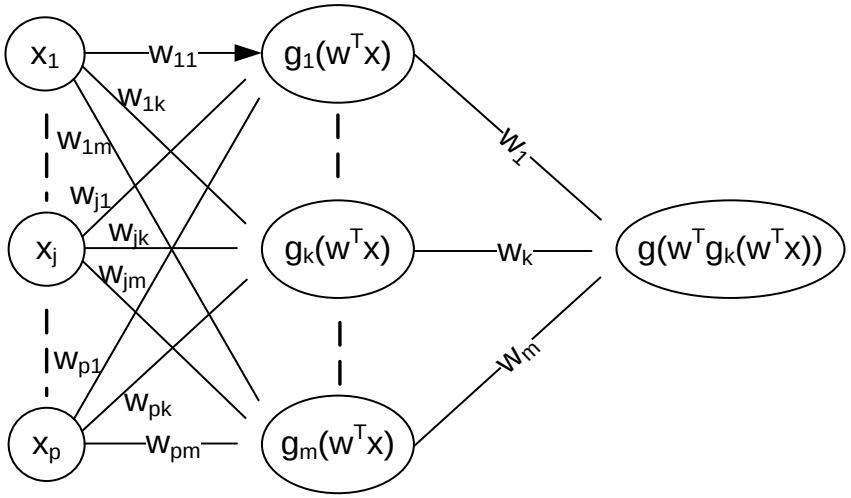
Non-profits have benefited from using neural networks (NN) models in fundraising (Goodman and Plouff, 1997). Neural networks are a mathematical model that simulates the structural aspects of the human brain to model non-linear relationships. Non-profits can use neural networks to explain the relationship between the probability of being donors and the independent variables x that describe the donors' behavior; and to explain the relationship between the amount of donations and the independent variables x. The Figure 1 below is a sketch of a simple neural networks model. It consists of an interconnected group of artificial neurons to process information using a connectionist approach to computation. The independent variables x_1 to x_p feeding into the input layer nodes to form a function of the products of the values of x_1 to x_p and the connection weights w_{i1} to w_{iq} between an input node i and a hidden layer node j. The final output node is a weighted sum of the functions from each of the hidden layer nodes. Therefore, the output of the neural networks can be given as $f(x) = g(w^T g(w^T x))$. The

objective of a neural networks model is to minimize the sum of squared errors given by the following equation. We select the sigmoid function as the activation function for the neural networks model and select a default value as the momentum factor in the study.

$$SSE = \sum_{i=1}^{n} (y_i - f(x))^2$$

In most cases, a neural networks model is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Neural networks are non-linear statistical data modeling tools. They are used to model complex relationships between the amount of donations as dependent variable and the independent variables. We use the neural network model in the SAS Enterprise Minor both as a non-linear regression method to predict the amount of donations and as a classification method to identify whether someone a donor or a non-donor. When neural network model is used to classify donors and non-donors, the output of the neural network model is either 1 for a predicted donor or 0 for a predicted non-donor. Zahavi and Levin (1997) provided more details of using neural network models in direct marketing.

Figure 1: Neural Networks Model



3.4 Support Vector Machines Models

The SVM can be used as a classification method to identify donors and non-donors. Assume the training data set is given as (x_1, y_1), (x_2, y_2),…, (x_n, y_n), where x_i is a p dimensional vector, y_i is the binary indicator dependent variable with $y_i = 1$ for donors and $y_i = -1$ for non-donors. The objective here is to separate the donors from non-donors based upon certain performance criteria. The following formulations are from Yu and Kim (2010) and Cui and Curry (2005).

When a training data set has a clear separation of donors and non-donors with a linear function, a SV classifier can be defined as the inner product between a weight vector $\boldsymbol{w}$ and an input vector $\boldsymbol{x}$ plus a bias b in the following form :

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} - b \text{ and } y_i f(\boldsymbol{x_i}) > 0, \forall(\boldsymbol{x}_i, y_i) \text{ for } i = 1, \dots, n$$

where $f(\boldsymbol{x}_i) > 0 \; if \; y_i = +1$ for donors, and $f(\boldsymbol{x}_i) < 0 \; if \; y_i = -1$ for non-donors. Therefore, $y_i f(\boldsymbol{x_i}) > 0$ if a correct classification is derived, and $y_i f(\boldsymbol{x_i}) < 0$ if a wrong classification is reached. The margin of a record $(\boldsymbol{x}_i, y_i)$ with respect to the classifier $f(x_i)$ is given by $y_i f(\boldsymbol{x_i})$. The values of $\boldsymbol{w}$ and b are learned through the training data in such a way that among all possible linear classifiers, the Optimal or Maximum Margin SV classifier provides the unique largest possible margin $1/\| \boldsymbol{w} \|$ on both sides of $f(\boldsymbol{x})$ (Vapnik, 1998). The observations (x_i, y_i) that are right on either side of the margins are called support vectors. Therefore, maximizing the margin $1/\| \boldsymbol{w} \|$ is equivalent to minimizing the norm of the weight vector $\boldsymbol{w}$ under the constraint that the rescaled margin value of $y_i f(\boldsymbol{x_i}) \geq 1$. Therefore, training in the SV classifier becomes solving a constrained optimization problem with the primal form:

$$\min \; \frac{1}{2} \| \boldsymbol{w} \|^2$$

$$\text{s.t. } y_i f(\boldsymbol{x_i}) \geq 1 \text{ for } i = 1, \dots, n.$$

The dual form of this problem can be shown as (Yu and Kim 2010):

$$\max \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j \, (\boldsymbol{x}_i . \boldsymbol{x}_j)$$

$$s.t. \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$\alpha_i \geq 0 \text{ for } i = 1, \ldots, n.$$

where α_i are the non-negative Lagrangian multipliers, and $\boldsymbol{w}^* = \sum_{i=1}^{n} \propto_i^* y_i \boldsymbol{x}_i$. The observations with nonzero α_i form the support vectors for the problem. The linear SV classifier becomes:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \boldsymbol{x} \quad - b$$

Cui and Curry (2005) pointed out the two properties of the solution above. First, relatively few observations from the training data become the support vectors. Second, the solution is free from the curse of dimensionality because it scales directly with the sample size, not with the dimensionality of the data model.

The optimal or the maximum margin SV classifier described so far can have feasible solutions only for linearly separable problems, and will be unable to be used to solve the problems which are not linearly separated, e.g., problems of non-profit fundraising. Many records could be pretty much the same; however, some of them are donors and the others are not. Among the problems with fundraising data are a lot of noise, missing data and so forth. Cortes and Vapnik (1995) introduced Soft – Margin SV classifiers. The Soft-Margin SV classifiers use positive slack variables ξ_i for i = 1 to n to indicate error terms of any misclassification, and introduce the use of a penalty term C or Cost to balance the use of the positive slack variables in the objective function. As shown in the revised formulation, the Soft-Margin SV classifiers maximize the margins of separations for records that can be separated, and minimize the numbers of misclassifications for records that cannot be separated (Cui and Curry, 2005; and Yu and Kim, 2010). The primal form of the optimization problem becomes as follows:

$$\min \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } y_i f(\boldsymbol{x_i}) \geq 1 - \xi_i \text{ for } i = 1, \ldots, n, \text{ and } \xi > 0$$

The dual form of the optimization problem has the form as follows:

$$\max \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j \, (\boldsymbol{x}_i . \boldsymbol{x_j})$$

$$s.t. \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$C \geq \alpha_i \geq 0 \text{ for } i = 1, \ldots, n.$$

For applications, e.g. non-profit fundraising, where donors and non-donors could not be separated with hyperplanes in the input data space with errors, a kernel function $\phi(x_i)$ could be used to replace $\boldsymbol{x}_i$. The kernel function $\phi(x_i)$ maps the attributes of independent variables x_i onto a higher dimensional, possibly infinite, feature space. $\phi(x_i)^T \phi(x_j) \equiv K(x_i, x_j)$ is the kernel function. Among the available kernel functions are linear, polynomial, radial basis function, sigmoid and others. The radial basis function used in this study has the form: $(x_i, x_j) = \exp(-\gamma \parallel x_i, x_j \parallel^2)$, $\gamma > 0$, where γ is a kernel parameter. The constraints become:

$$\text{s.t. } y_i \phi(\boldsymbol{x_i}) \geq 1 - \xi_i \text{ for } i = 1, \ldots, n, \text{ and } \xi > 0$$

The solution to the following SVM formulation identifies whether or not someone will be a donor.

$$f(x) = \; sign \; (w^* \cdot \phi(x) + b^*) = sign \left(\sum_{i=1}^{n} y_i \alpha_i^* K(x_i, x_j) + b^* \right)$$

where α is the vector of Lagrangian multipliers corresponding to the constraints of the SVM formulation. The computations for the solution of this SVM problem are through the inner products of $\phi(x_i)^T \phi(x_j)$ in the input data space.

3.5 Performance Measures

This thesis uses two sets of performance measures. One set of performance measures includes commonly used sensitivity, specificity, the false positive rate, the false negative rate, and the total correct classification rate to measure the classification performance of the logistic regression, neural networks and the SV classification models. For an actual donor, a model could classify him/her as either a donor or a non-donor. Likewise, for an actual non-donor, a model could classify him/her as either a non-donor or a donor. The sensitivity or the true positive rate refers to the ratio of the number of correctly predicted donors to the total number of actual donors. The specificity or the true negative rate refers to the ratio of the number of correctly predicted non-donors to the total number of actual non-donors. The false positive rate or the Type I error refers to the number of non-donors who are incorrectly predicted as donors among the total number of predicted donors. Accordingly, the false negative rate or the Type II error refers to the number of donors who are incorrectly predicted as non-donors among the total number of predicted non-donors. It is more important to reduce the number of donors who are incorrectly classified as non-donors or a smaller Type II error than otherwise to reduce the number of non-donors who are incorrectly classified as donors or a smaller Type I error because the average dollar amount of contribution from a donor is two or three times more than that of the average mailing cost. Likewise, it is more important to increase the number of correctly classified donors among the total number of actual donors or a higher sensitivity than to increase the number of correctly classified non-donors or a higher specificity among the total number of actual non-donors. The total correct classification rate refers to the ratio of the summation of the actual donors who are predicted as donors and the actual non-donors who are predicted as non-donors to the total number of potential donors in the data.

The other set of performance measures uses the average dollar amount of contributions in the top 100d% of potential donors (Maltihouse, 2002) to measure the predictive performance of

the multiple regression, neural networks and the SV regression models. To predict the dollar amount of contributions from a solicitation campaign, a model learned from training data is used to score holdout data. The scoring process generates a ranked order of potential donors based on their possibilities to donate and/or the projected amount of donations. A solicitation offer is often sent to the potential donors with the 100d% highest predicted scores (Malthouse, 2002). An average dollar amount of contributions per donor is calculated by dividing the total dollar amount of contributions by the number of potential donors in the top 100d% highest ranked group.

A Gain Chart (Figure 5 on page 31) can be drawn from the cumulative dollar amount of contributions and the sorted deciles of the donor groups to measure the effectiveness of a specific predictive model. A better predictive model should have a larger amount of contributions in the top 100d% of potential donors.

Chapter 4 Design of Experiments

In order to compare the predictive performance of multiple regression, neural networks and the SV regression models for predicting the dollar amount of contributions, the following experiments are designed with changing factors as follows:

1) The multiple regression and neural networks use both a 10% and one third of the 99,200 records as the training and validation data, while the SV regression uses only a 10% of the 99,200 records as the training and validation data due to the slower computations in the R implementation of the SV regression models. The multiple regression in SAS PROC REG has the validation option with varying percentage of the training data for validation. The neural networks use one half of its training data in validation, and the SV regression models use a ten-fold validation process in its training.
2) Ten stratified random samples are taken from DMEF01 data, each with one third of the 99,200 records as the training and validation data, and the rest as the holdout data to be used in multiple regression and neural networks.
3) Among the 24 candidate variables, SAS PROC GLMSELECT uses SBC criterion to identify the initial set of 10 to 14 variables to be significant, and the validation option in SAS PROC GLMSELECT uses the average squared validation errors to identify another smaller set of 5 to 6 variables to be significant. Each of the three models uses these two sets of variables to train their models.

In order to compare the classification performance of logistic regression, neural networks and the SV classification models for identifying donors and non-donors, the following experiments are designed with changing factors as follows:

1) The logistic regression and neural networks use both a 10% and one third of the 99,200 records as the training and validation data, while the SV regression uses only a 10% of the 99,200 records as the training and validation data due to the slower computations in

the R implementation of the SV classification models. The logistic regression in SAS does not have the option of validation in its training process. The neural networks use one half of its training data in validation, and the SV classification models use a ten-fold validation process in its training.

2) Ten stratified random samples are taken from DMEF01 data, each with one third of the 99,200 records as the training and validation data, and the rest as the holdout data to be used in logistic regression and neural networks.
3) Among the 24 candidate variables, SAS PROC LOGISTIC identifies a set of 10 to 14 variables to be significant. Each of the three models uses the same set of variables to train their models.

Malthouse (2002) split evenly the 99,200 records in DMEF01 data into the training, validation and testing data sets. We find SAS computes very efficiently all of the required training and testing results for multiple regression, logistic regression and neural networks models. The SVM regression and classification models in R, however, are extremely slow. Besides the reporting of the results for multiple regression, logistic regression and neural networks with the one third even split samples, a 10% of 99,200 records as training set is also taken in order to compare the performance of these models.

Chapter 5 Results of Comparisons

5.1 Summary of Predictive Performance for the Dollar Amount of Contributions

Multiple regression, neural networks, and the SV regression models are used to predict the dollar amount of contributions for DMEF01 data. SAS PROC REG and PROC GLMSELECT are used to carry out the computations for multiple regression models. SAS Enterprise Minor is used to carry out the computations for neural networks. The computations for the SV regression models are carried out in its R implementation in libsvm (Chang and Lin, 2001).

5.1.1 Tests with 10% of 99,200 records in training data

The first set of tests is with 10% of the 99,200 records as training and validation data, and the rest as the holdout data.

A combined training and validation sample of 9,920 records (or 10% of DMEF01 of 99,200 records) is used in this initial test. The rest of the 89,280 records are as the holdout sample. Among the 9,920 records in the training data, a 10% of 9,920 records are used as validation data through the option of PARTITION FRACTION in SAS PROC GLMSELEC. Furthermore, a ten-fold cross validation for the training is set up in the process, i.e., the training data is evenly split into ten parts; one part as validation data while the other nine parts as training data. The following process shows how the STEPWISE procedure in SAS PROC GLMSELECT identifies the number of independent variables to be included in a multiple regression model. There are 24 independent variables as potential candidates to be used besides the intercept term. The SBC criterion given by the equation n ln (SSE/n) + p ln (n) is used to determine whether an independent variable x is statistically significant enough to be included in the model to reduce the SBC value significantly. The smaller the SBC value, the better the model fit. The best model fit, however, does not ensure effective predictions of the amount of contributions.

Table 1 summarizes the process of the stepwise selection of variables in the model.

The SBC along other criteria measures the fit of the model to the training data and selects 11 variables out of 24 potential variables besides the intercept. The validation average squared errors reach a minimum of 24.30 when nine variables are in the model besides the intercept. It is worth to know that the validation average squared errors are 24.37 with 6 variables in the model besides the intercept, or only 0.288% higher than that of the model with 9 variables. Figure 3 shows the path of the reduction of the average squared errors for both training and validation data. We compare the predictive performances of the models with 11 and 6 variables plus corresponding intercepts in this initial test, and for all of the models studied.

Table 1: Variable Selections with Stepwise Regression

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Model R-Square	Adjusted R-Square	SBC	ASE	Validation ASE	F Value	Pr > F
0	Intercept		1	0	0	31222.04	32.50	28.53	0	1
1	tran13		2	0.0659	0.0658	30619.65	30.36	25.74	632.68	<.0001
2	cndol1		3	0.0829	0.0827	30464.55	29.81	25.19	165.66	<.0001
3	tran11		4	0.0871	0.0868	30432.15	29.67	24.81	41.58	<.0001
4	tran8		5	0.0909	0.0905	30403.55	29.54	24.58	37.76	<.0001
5	rcdatlrg		6	0.0938	0.0933	30384.20	29.45	24.45	28.47	<.0001
6	rccndat1		7	0.0982	0.0976	30350.18	29.31	24.37	43.2	<.0001
7	tran5		8	0.0993	0.0986	30348.34	29.27	24.40	10.94	0.0009
8		tran13	7	0.0993	0.0987	30339.31	29.27	24.40	0.07	0.7889
9	sltmlif		8	0.1017	0.101	30324.53	29.20	24.60	23.89	<.0001
10	cntrlif		9	0.105	0.1042	30300.28	29.09	24.48	33.38	<.0001
11		tran8	8	0.1043	0.1036	30298.12	29.11	24.61	6.93	0.0085
12	tran9		9	0.1065	0.1057	30285.03	29.04	24.31	22.19	<.0001
13	tran10		10	0.109	0.1081	30269.04	28.96	*24.30	25.1	<.0001
14	cntmlif		11	0.1105	0.1095	30263.53	28.91	24.49	14.61	0.0001
15	rcdatfst		12	0.1118	0.1107	30259.62	28.87	24.52	13.01	0.0003

* Optimal Value Of Criterion

Figure 2 and Table 1 shows the path of the reductions of SBC when each of the variables is either entered or removed from the model. The additional reductions of the SBC values become smaller and smaller with each additional variable added to the model. In another words, the additional contribution of a potential variable to reduce the value of SBC becomes smaller and smaller after other variables have been included in the model. Among the 24 potential candidate variables, only 11 variables plus the intercept term are significant at the end to be included in the model.

Figure 2: Sequence of Stepwise Variable Selection with SBC Criterion

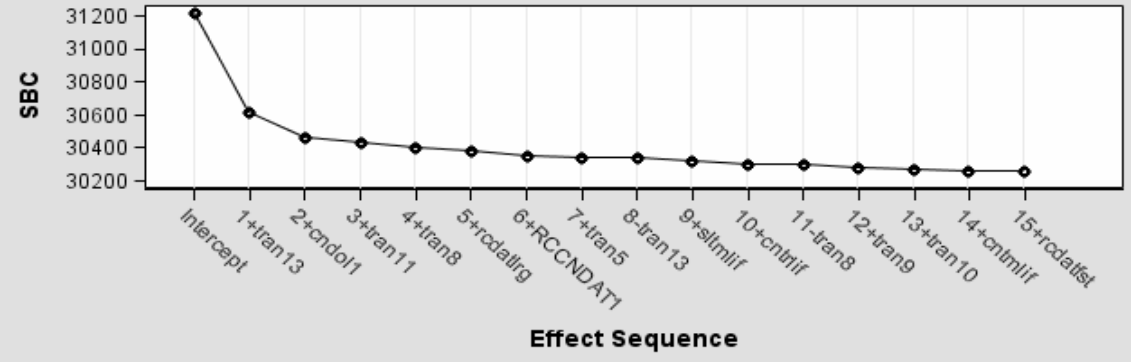


Figure 3: Average Squared Errors for Training and Validation Data

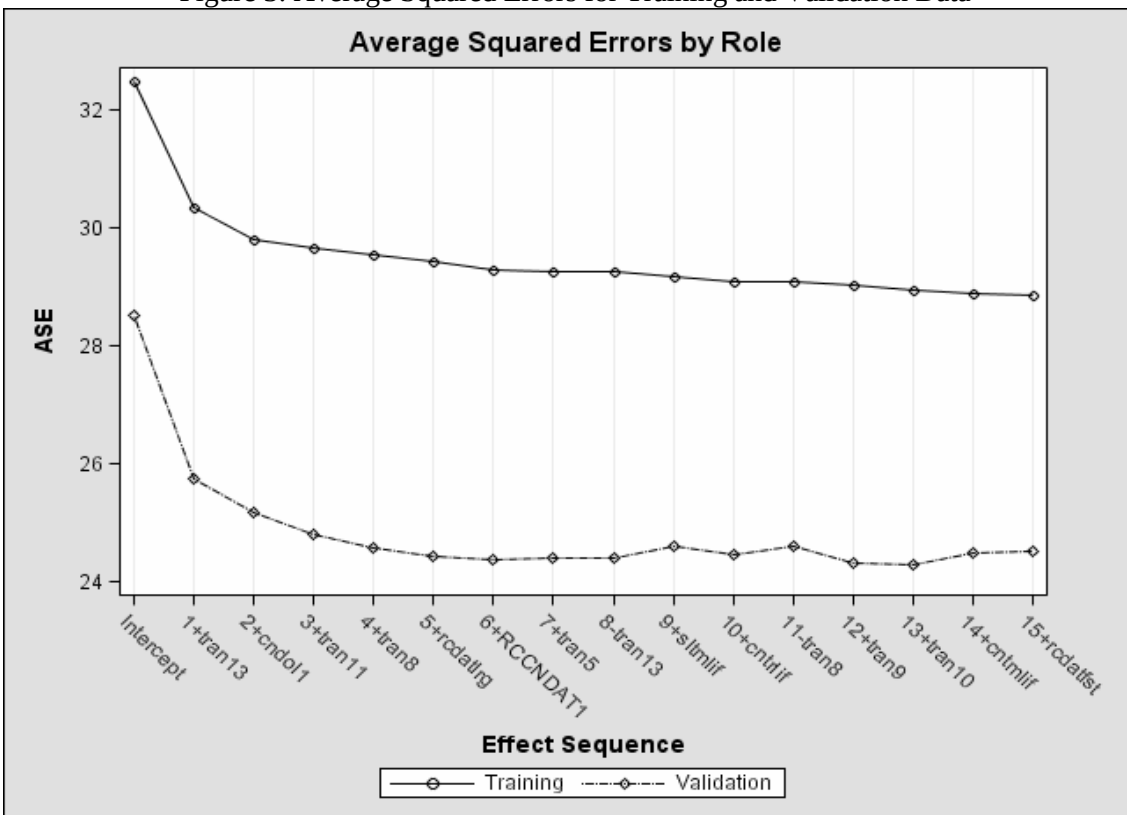


Figure 3 shows the average squared errors for the training and validation data with the addition and removal of each variable. The selection process ends when any addition and removal of a variable will not reduce the average squared errors. Multiple regression, neural networks and the SV regression models subsequently use these 11 variables to predict the dollar amount of contributions in fundraising. In certain cases, the average squared errors for the

validation data may reach a lower point, and then start to increase again while the average squared errors for the training data continues to reduce, but clearly not as significant as the amount of reductions when the initial few variables are added. We test the predictive performance of multiple regressions, neural networks and the SV regression models using the less number of variables for which the average squared errors for the validation data starts to increase.

When the predictions of the multiple regression and neural networks are applied to the holdout data, they result in $5.067 and $5.202 of contributions from the top 20% of the holdout data, respectively. The results are compatible and / or slightly better than the results reported in Malthouse (2002).

Table 2: Predictions of the Amount of Contributions with Multiple Regression

Percentile	N	Mean	Total $	Cum N	Cum Mean	Cum Total $
1	6,613	6.37	42,123.59	6,613	6.37	42,123.59
2	6,613	3.62	23,965.53	13,226	5.00	66,089.11
3	6,613	2.93	19,383.13	19,839	4.31	85,472.24
4	6,613	2.39	15,822.21	26,452	3.83	101,294.45
5	6,613	1.81	11,982.27	33,065	3.43	113,276.72
6	6,613	1.74	11,497.65	39,678	3.14	124,774.37
7	6,613	1.35	8,933.25	46,291	2.89	133,707.62
8	6,613	1.18	7,828.17	52,904	2.68	141,535.79
9	6,613	0.99	6,541.51	59,517	2.49	148,077.30
10	6,613	0.87	5,747.15	66,130	2.33	153,824.45

Tables 2 shows the results of predicting the dollar amount of donations with multiple regression models. It shows the average dollar amount of donations from each of the ten ranked deciles from $6.37 to $2.33. The top 10% of the donors contribute an average of $6.37, and the top 20% of the donors contribute an average of $5.00, and so forth, respectively.

Table 3 shows the predictive performance for the neural networks. Neural networks produce $6.72 for the top 10% of the donors, and $5.17 for the top 20% of the donors, and so forth, respectively. Neural networks outperform the multiple regression models in predicting the amount of contribution.

Table 3: Predictions of the Amount of Contributions with Neural Networks

Percentile	N	Mean	Total $	Cum N	Cum Mean	Cum Total $
1	6,613	6.72	44,453.80	6,613	6.72	44,453.80
2	6,613	3.63	23,983.15	13,226	5.17	68,436.95
3	6,613	2.93	19,343.70	19,839	4.42	87,780.64
4	6,613	2.32	15,332.28	26,452	3.90	103,112.92
5	6,613	1.87	12,392.20	33,065	3.49	115,505.12
6	6,613	1.59	10,537.88	39,678	3.18	126,043.00
7	6,613	1.35	8,947.71	46,291	2.92	134,990.71
8	6,613	1.17	7,729.25	52,904	2.70	142,719.96
9	6,613	0.99	6,569.10	59,517	2.51	149,289.06
10	6,613	0.69	4,539.90	66,130	2.33	153,828.97

Figure 4 shows the cumulative amount of contributions in each of the deciles (10%). More than 80% of the total amount of contributions is given by the top 20% of the donors. The cumulative average amount of contributions indicates the results if the solicitations are sent to everyone in the holdout data. The curve in the middle is the predicted cumulative amount of contributions with the neural networks model. The closer the curve of the predicted cumulative amount of contributions to the actual cumulative amount of contributions, the better the performance of the predictive model is.

The use of the SV ε regression to predict the dollar amount of contributions requires the choice of the values of parameters ε, the penalty term C and the γ (gamma) factor in the kernel function. The tune.svm procedure in libsvm is used for the purpose. The initial test results in Table 4 show that the SV regression models result in $4.66 as the dollar amount of donations from the top 20% of potential donors or the worst among the three tested models for predicting the amount of contributions. More experiments are needed to assess the performance of the SV regression models.

Figure 4: The Actual and Predicted Amount of Contributions

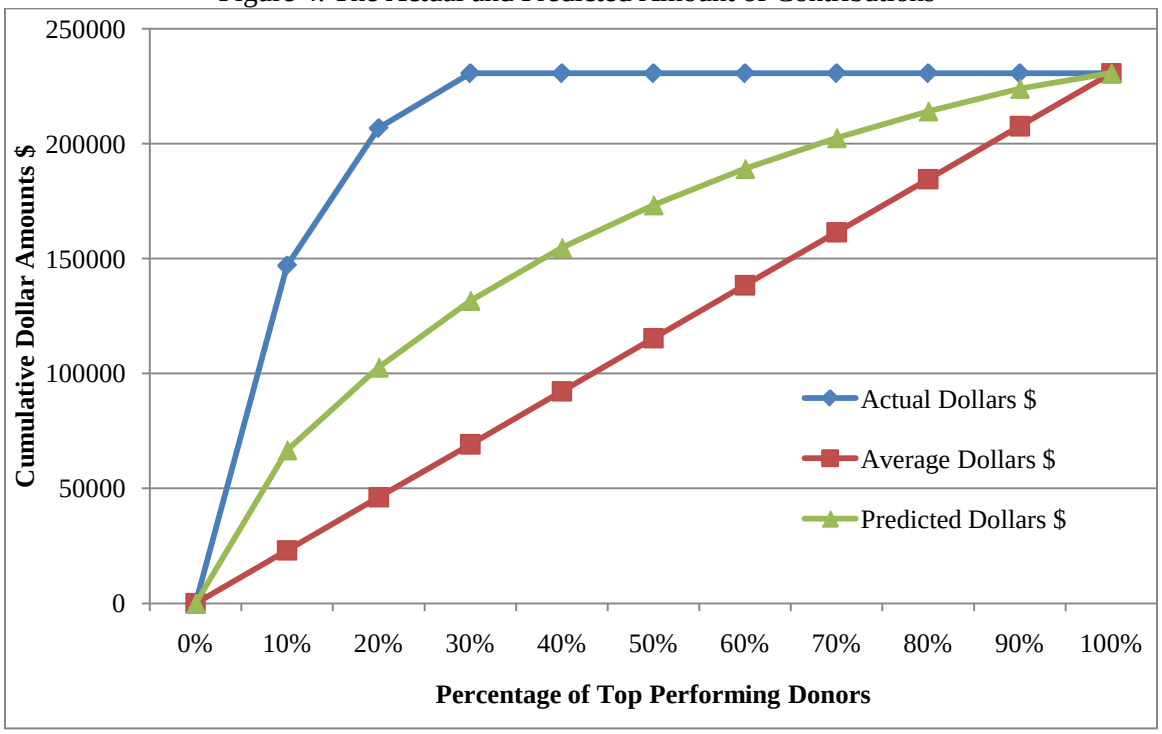


Table 4: Predictions of the Amount of Contributions with the SVMs

	Top 20% of Average $ of Donations with SVM Penalty Term C							
Gamma	0	1	2	3	4	5	6	7
-7	4.407	4.433	4.432	4.454	4.495	4.474	4.511	4.499
-6	4.548	4.548	4.430	4.443	4.499	4.642	4.616	4.588
-5	4.628	4.643	4.658	4.634	4.606	4.565	4.572	4.542
-4	4.611	4.611	4.611	4.607	4.605	4.577	4.524	4.517
-3	4.618	4.632	4.614	4.588	4.570	4.517	4.463	4.433

Figure 5 shows the grid search for the optimal values of the SV regression parameters of epsilon and gamma. The bar on the right side in Figure 5 indicates the color coding for error levels. The darker the color, the lower the error is. It seems from this specific test that the choice of gamma is not significant when the value of epsilon is smaller than 0.25. The SV regression produces its best solution when the values of the epsilon is $2^{(-1)}$ and the gamma is $2^{(-8)}$ with the value of the penalty term C as 1.

Figure 5: Grid Search of Values of C, Epsilon (ε) and Gamma (γ) in the SV Regression

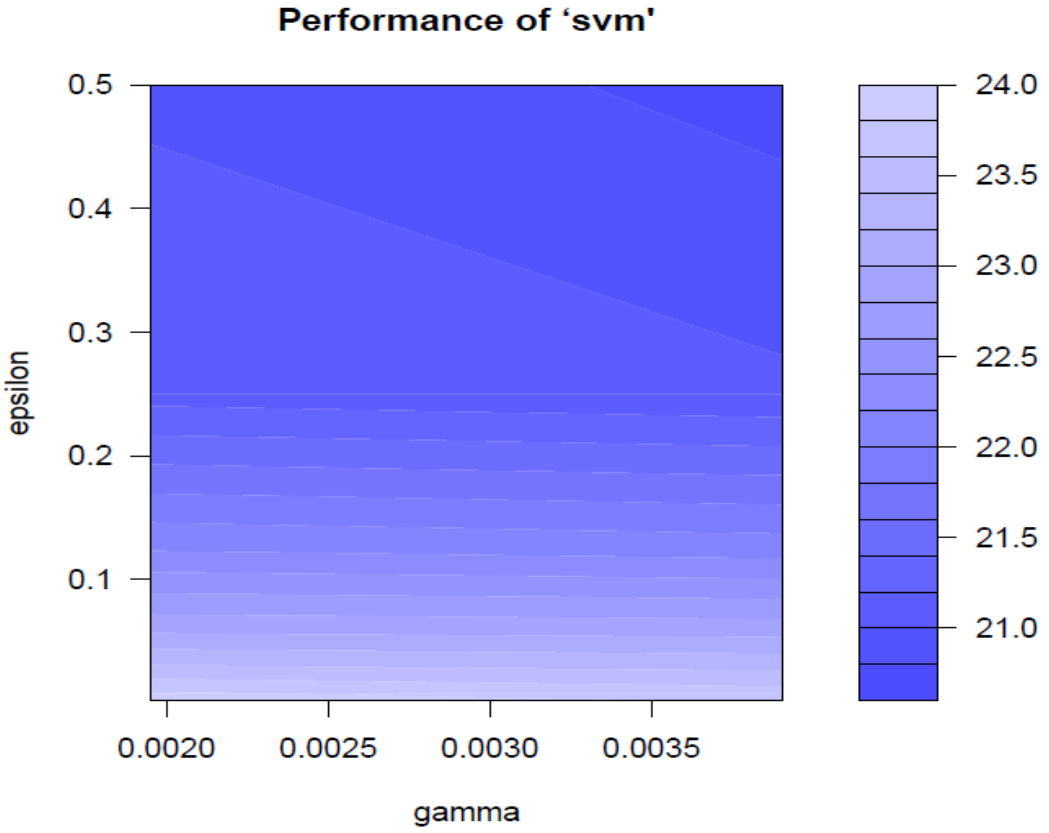


Table 5: The Dollar Amount of Contributions from Top 20% of Donors

Full	Number of Variables				Number of Variables				Average
# Variables	11	14	10		6	6	6		
Regression	5.067	5.063	5.057		5.109	5.106	5.087		5.082
Neural Nets	5.202	5.132	5.096		5.252	5.111	5.246		5.173
The SVMs	4.660								4.660

Table 5 summarizes the results of multiple tests with 10% of the 99,200 records as training data and the rest as holdout data. The multiple regression models use 10% of the training data in validation. The multiple regressions produce $5.08 on the average from the top 20% of the donors. The neural networks use 50% of the training data in validation. The neural networks produce $5.17 on the average from the top 20% of the donors. The number of variables 11, 14 and 10 are determined by the stepwise regression procedure in SAS PROC GLMSELECT. The 6 variables are determined by the smallest amount of average squared errors in the validation data

in SAS PROC GLMSELECT. Multiple regression, neural networks and the SV regression models use the same set of variables as independent variables in training. The results in Table 5 indicate that the models with a smaller number of variables can provide compatible or better results than the models specified by the commonly used stepwise regression procedure for this specific data. The SV regression models use a ten-fold validation procedure that trains the model with nine tenth of the training data each time and validates with one tenth of the training data. Due to the slower speed, the computations for the SV regression models are only carried out with 11 independent variables in the model. The SV regression models produce $4.66 on average from the top 20% of the donors, or the worst of the three models.

5.1.2 Tests with one third of 99,200 records in training data

The second set of tests is with one third of the 99,200 records as training and validation data, and the rest as the holdout data.

Tables 6 and 7 show the results from ten more tests for the multiple regressions and neural networks models. In each of the test, a different stratified sample is taken with a new random number seed. The multiple regressions produce $6.37 and $5.00 on the average from the same matching groups of donors, respectively; the neural networks produce $6.72 and $5.17 on the average from the top 10% and 20% of donors, respectively; The neural network models outperform the multiple regression models 5.55%, 3.4%, 2.5%, and 1.8% on the average for each of the top 10%, 20%, 30% and 40% of the donors, respectively. That could translate into the equal percentage of additional contributions on the average from each donor in the matching groups. It could potentially increase thousands of dollars of donations if several millions of people are solicited. The results for the SV regression models are not available due to the slower speed of computations.

Table 6: The Dollar Amount of Contributions with Multiple Regression Models

	Setid	Setid	Setid	Setid	Setid	Setid	Setid	Setid	Setid	Setid	
Group	1	02	03	04	05	06	07	08	09	10	Average
1	6.71	6.67	4.95	6.76	6.76	5.20	6.67	6.59	6.78	6.60	6.37
2	5.15	5.14	4.18	5.21	5.18	4.40	5.18	5.12	5.23	5.18	5.00
3	4.41	4.41	3.72	4.45	4.47	3.93	4.42	4.38	4.46	4.43	4.31
4	3.90	3.89	3.43	3.92	3.94	3.56	3.91	3.90	3.92	3.91	3.83
5	3.48	3.48	3.10	3.51	3.51	3.21	3.48	3.46	3.52	3.51	3.43
6	3.19	3.19	2.92	3.21	3.21	2.96	3.19	3.17	3.23	3.17	3.14
7	2.92	2.92	2.74	2.94	2.94	2.73	2.93	2.90	2.95	2.91	2.89
8	2.70	2.70	2.56	2.72	2.72	2.54	2.70	2.68	2.73	2.71	2.68
9	2.50	2.51	2.40	2.52	2.52	2.38	2.50	2.50	2.53	2.52	2.49
10	2.33	2.32	2.33	2.34	2.34	2.29	2.32	2.32	2.34	2.33	2.33

Table 7: Dollar Amount of Contributions with Neural Networks Models

	Setid	Setid	Setid	Setid	Setid	Setid	Setid	Setid	Setid	Setid	
Group	1	02	03	04	05	06	07	08	09	10	Average
1	6.86	6.72	6.60	6.69	6.84	6.33	6.64	6.77	6.97	6.81	6.72
2	5.23	5.21	5.13	5.15	5.20	4.99	5.14	5.17	5.29	5.24	5.17
3	4.46	4.47	4.39	4.43	4.45	4.27	4.38	4.44	4.51	4.44	4.42
4	3.92	3.92	3.88	3.93	3.92	3.78	3.87	3.91	3.94	3.91	3.90
5	3.50	3.51	3.49	3.51	3.51	3.39	3.48	3.51	3.52	3.50	3.49
6	3.19	3.18	3.17	3.18	3.20	3.09	3.18	3.19	3.21	3.18	3.18
7	2.93	2.92	2.92	2.92	2.94	2.85	2.90	2.92	2.94	2.92	2.92
8	2.70	2.70	2.69	2.72	2.71	2.65	2.69	2.70	2.72	2.70	2.70
9	2.51	2.51	2.51	2.53	2.52	2.47	2.50	2.50	2.52	2.51	2.51
10	2.33	2.32	2.33	2.34	2.34	2.30	2.32	2.32	2.34	2.34	2.33

5.2 Summary of Classification Performance for Donors and Non-Donors

The logistic regression, neural networks, and the SV classification models are used to identify donors and non-donors. The SAS PROC LOGISTIC is used to carry out the computations for logistic regression models. The SAS Enterprise Miner is used to carry out the computations for neural networks. The R implementation of the SV Classification models is used to compute the SV classifications.

Table 8: Identifications of Donors and Non-Donors

	Logistic		Neural Networks		The SVMs	
	Testing	Training	Testing	Training	Testing	Training
Actual & Predicted Non Donors	62038	6906	61178	6827	61585	6938
Actual Donors & Predicted Not	19425	2134	18518	2004	18370	1970
Actual Non Donors & Predicted Yes	2755	293	3615	372	3208	261
Actual & Predicted Donors	5062	587	5969	717	6117	751
Total Records	89280	9920	89280	9920	89280	9920
True Positive or Sensitivity	20.67%	21.57%	24.38%	26.35%	24.98%	27.60%
True Negative or Specificity	95.75%	95.93%	94.42%	94.83%	95.05%	96.37%
False Positive	35.24%	33.30%	37.72%	34.16%	34.40%	25.79%
False Negative	23.85%	23.61%	23.24%	22.69%	22.98%	22.11%
True Positive/Total	5.67%	5.92%	6.69%	7.23%	6.85%	7.57%
True Negative/Total	69.49%	69.62%	68.52%	68.82%	68.98%	69.94%
Total Correct	75.16%	75.53%	75.21%	76.05%	75.83%	77.51%

5.2.1 Tests with 10% of 99,200 records in training data

Table 8 summarizes the results of classifying donors and non-donors with the logistic regression, neural networks and the SV classification models in an initial test. The training data is 10% of the 99,200 records. The logistic regression model selects 11 variables with the stepwise procedure in SAS PROC LOGISTIC for the training data without validation. The neural networks and the SV classification models use the same 11 variables in training their models. The neural networks use 50% of the training data in validation, and the SV classification models use a 10-fold validation. Table 8 shows that the SV classification models outperform both logistic regression and neural networks just slightly over a half percentage point in this initial test. It is important to show that the true positive (sensitivity) rates for the holdout data are 24.98%, 24.38% and 20.67% for the SV regression, neural networks, and logistic regression models, respectively. As we discussed before, each person in the data contributed $2.33 on the average, or several times of the average cost of the solicitations. The SV classification models' capability to locate more actual donors than both neural networks and logistic regression would translate

into increased amount of donations and recruit more potential donors. The true negative (specificity) rates for the holdout data are 95.05%, 94.42% and 95.75% for the SV classification, neural networks, and logistic regression models, respectively. The logistic regression has the highest true negative rate and the lowest true positive rate. That implies the utility of this specific logistic regression model to select donors and non-donors are 3% to 4% lower than both the neural networks and the SV classification models, even though the total correct prediction rates are very close for these three models. The SV classification models have the lowest false positive (34.4%) and false negative (22.98%) rates, and the highest true positive rate (6.85%), as compared to that of the neural networks and logistic regression models.

In order to identify the best parameters for the penalty term C and the value of γ (gamma) term in training the SV classification models, a grid search as suggested in Chang and Lin (2001) is used. The grid search uses the tune.svm procedure in e1071 to search over the ranges for the penalty term C from 2^1 to 2^9 and the value of gamma from $2^{(-9)}$ to $2^{(-1)}$. A laptop computer HP compaq 2210b with Core 2 Duo T7700 Processor of 2.4GHz and 1GB RAM takes 42.85 seconds of CPU system times and another 76,211.48 seconds, or close to 22 hours, of user times to finish the search.

Figure 6 shows the SV classification models have the highest classification performance when the C (Cost) is 2^8 and the value of gamma is $2^{(-5)}$. As shown in Table 9, the best SV classification model (Best SVMs) produces the results far better than that of logistic regression and neural networks. The true positive rate is 25.75%, which is an improvement of more than 5% over that of logistic regression (20.67%), and more than 1% over that of neural networks (24.38%). The SV classification model achieves better or compatible results for true negative rate of 94.71%. This improvement is very important because the SV classification model could provide an optimal solution without the trials and errors as in neural networks. Table 9 also shows the results for training with duplicated records (The SVMs II) for each of the donors in

training set. The default proportion of donors in the DMEF01 is 27.4%. The increased proportion of donors' records in the training data should increase the probability that more donors be identified. The results for the SVMs II in Table 9 show the increase of more than 10% of the true positive rate from 25.75% to 35.78% when each donor is trained twice instead of once. The true negative rate decreases about 5% from 94.71% to 89.26%. The results of The SVMs I are from the initial test of the SV classifying models.

The bar on the right side in Figure 6 indicates the level of misclassification. The darker the color, the smaller the classification error is. It seems that the penalty term C is not sensitive when the value of gamma is smaller than 0.1 and C is larger than 200.

An economical analysis should be conducted to compare the amount of increased revenue and the increased cost due to the increased number of donors and non-donors.

Figure 6: Grid Search for Values of Cost and Gamma (γ) in the SV Classification

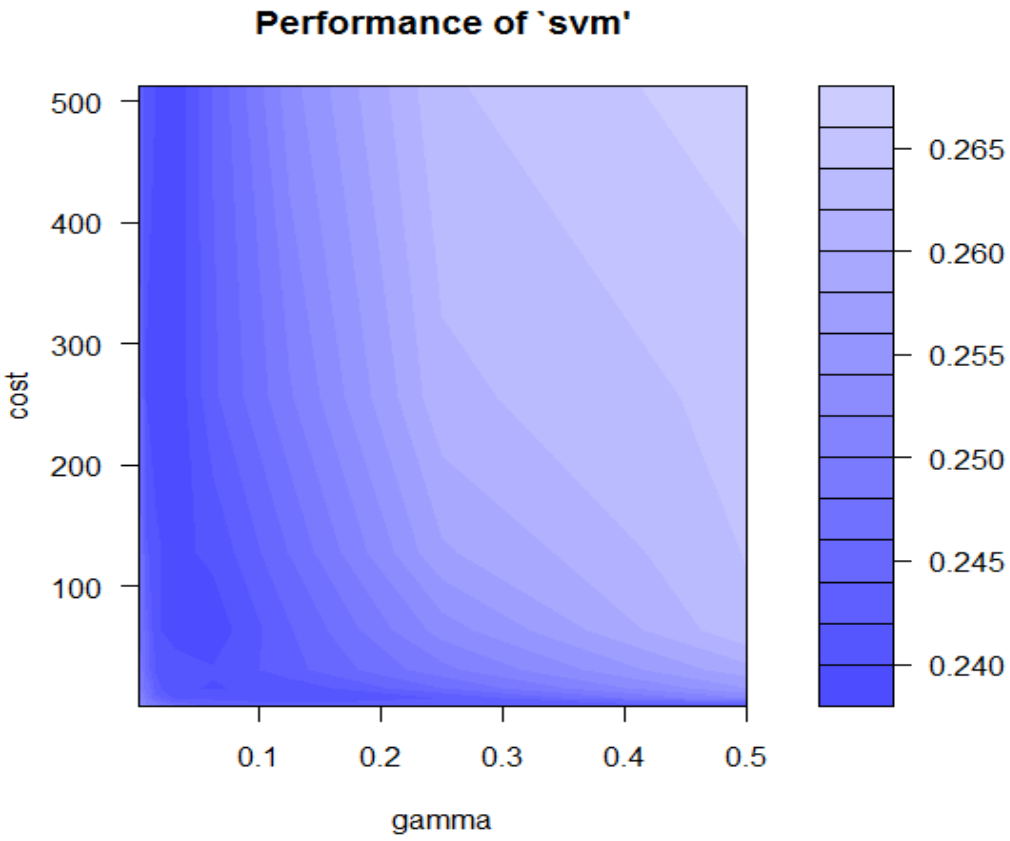


Table 9: Identifications of Donors and Non-Donors with the SVMs

	The SVMs I		Best SVMs		The SVMs II	
	Testing	Training	Testing	Training	Testing	Training
Actual & Predicted Non Donors	61585	6938	61368	6940	57835	6581
Actual Donors & Predicted Not	18370	1970	18181	1906	15729	1579
Actual Non Donors & Predicted Yes	3208	261	3425	259	22696	2197
Actual & Predicted Donors	6117	751	6306	815	8758	1142
Total Records	89280	9920	89280	9920	89280	9920
True Positive or Sensitivity	24.98%	27.60%	25.75%	29.95%	35.78%	42.01%
True Negative or Specificity	95.05%	96.37%	94.71%	96.40%	89.26%	91.43%
False Positive	34.40%	25.79%	35.20%	24.12%	44.25%	35.06%
False Negative	22.98%	22.11%	22.86%	21.55%	21.38%	19.34%
True Positive/Total	6.85%	7.57%	7.06%	8.22%	9.81%	11.52%
True Negative/Total	68.98%	69.94%	68.74%	69.96%	64.78%	66.35%
Total Correct	75.83%	77.51%	75.80%	78.18%	74.60%	77.87%

5.2.2 Tests with one third of 99,200 records in training data

Tables 10 and 11 describe results of more tests using logistic regression and neural network models to classify donors and non-donors. The results are compatible with that of the comparisons shown in Table 8. The neural networks outperform the logistic regression models in the 10 tests by more than 3% as measured by the true positive rates (24.47% for neural networks and 21.03% for logistic regression), or the possibility of selecting correct donors out of the total actual donors.

Table 10: Identifications Donors and Non-Donors with Logistic Regression

	Setid 1	Setid 02	Setid 03	Setid 04	Setid 05	Setid 06	Setid 07	Setid 08	Setid 09	Setid 10	Average
True Positive	21.20%	20.81%	21.27%	20.68%	21.19%	21.62%	21.29%	20.48%	20.91%	20.87%	21.03%
True Negative	95.76%	95.84%	95.79%	95.83%	95.86%	95.58%	95.66%	96.09%	95.64%	95.75%	95.78%
False Positive	34.61%	34.59%	34.37%	34.79%	34.06%	35.09%	35.02%	33.59%	35.56%	35.00%	34.68%
False Negative	23.72%	23.80%	23.70%	23.83%	23.71%	23.66%	23.72%	23.83%	23.81%	23.80%	23.76%
True Positive/Total	5.82%	5.71%	5.83%	5.67%	5.81%	5.93%	5.84%	5.62%	5.74%	5.72%	5.77%
True Negative/Total	69.49%	69.55%	69.52%	69.55%	69.57%	69.37%	69.43%	69.73%	69.41%	69.49%	69.51%
Total Correct	75.31%	75.26%	75.35%	75.22%	75.38%	75.30%	75.26%	75.35%	75.14%	75.21%	75.28%

Table 11: Identifications Donors and Non-Donors with Neural Networks

	Setid 1	Setid 02	Setid 03	Setid 04	Setid 05	Setid 06	Setid 07	Setid 08	Setid 09	Setid 10	Average
True Positive	23.85%	23.72%	25.50%	26.30%	23.27%	24.25%	25.19%	23.46%	25.48%	23.84%	24.49%
True Negative	95.33%	94.99%	94.27%	95.01%	95.46%	94.80%	95.05%	95.48%	94.89%	94.96%	95.02%
False Positive	34.13%	35.84%	37.31%	33.42%	34.06%	36.21%	34.20%	33.78%	34.67%	35.88%	34.97%
False Negative	23.19%	23.28%	23.00%	22.67%	23.30%	23.20%	22.92%	23.25%	22.89%	23.26%	23.10%
True Positive/Total	6.54%	6.51%	6.99%	7.21%	6.38%	6.65%	6.91%	6.43%	6.99%	6.54%	6.72%
True Negative/Total	69.18%	68.94%	68.41%	68.95%	69.28%	68.80%	68.98%	69.29%	68.86%	68.91%	68.96%
Total Correct	75.73%	75.44%	75.40%	76.17%	75.66%	75.45%	75.89%	75.72%	75.85%	75.45%	75.68%

Chapter 6 Discussions and Concluding Remarks

This thesis studies the predictive modeling for non-profit fundraising with several data mining methods. We compare the predictive performance of multiple regression, neural networks and the SV regression models in predicting the dollar amount of contributions. The neural networks outperform the multiple regression and the SV regression models in the two sets of tests conducted in this study. One set of tests is with 10% of 99,200 records as training data and the other set of tests is with one third of 99,200 records as training data. The SV regression models do not produce the results as we expected, partly because we could not have enough experiments due to the slow speed of computations in R. The SAS Enterprise Miner has the SVM on experimental level in its Sever version. We hope to have access to it in the future to fully explore the capability of the SV regression models for predicting the dollar amount of contributions in fundraising. We find that the results from training the 10% and one third of the 99,200 records are very compatible for multiple regression and neural networks models. It indicates a smaller training data could provide compatible model to leave a larger portion of the whole data for promotion, thus maximize the potential revenue. The testing results also indicate the importance of the validation during the training process. A predictive model with less number of variables determined by the average squared errors of validation produces compatible and better results than the models with more variables determined by the SBC criterion.

We also compare the classification performance of logistic regression, neural networks and the SV classification models in identifying donors and non-donors. The SV classification models outperform both neural networks and the logistic regression models in only one set of tests conducted in this study. The set of tests is with 10% of 99,200 records as training data. Among the major advantages of using the SV classification models include that the SVM models are solved through optimization process, thus avoid local minimums. The interpretations of the results from the SVM models are much straight forward than that of neural networks. As

indicated in the results, the true positive rates of correct classifications by the SV classification models are 5% higher than that of the logistic regression models, and around 1% higher than that of the neural networks. That could translate into the equal proportion of potential donors being identified correctly, and could benefit non-profits in their future fundraising activities.

Our major contribution in this thesis is to introduce the SVM models to non-profit fundraising. Our results clearly show the SV classification models can be a valuable alternative for classifying donors and non-donors. Continued effort is needed to explore the potential of the SV regression models as a regression method to predict the dollar amount of contributions in fundraising. One important technical issue in using the SVM models is the choice of the values for ε, γ and the penalty term C. Due to the hardware we have available at this point, we could not explore more of the impact of the combinations of these parameters in predicting the amount of the contributions or in classifying donors and non-donors. Recent researches in the literature show promises to use centroids of k-means to reduce the number of records of the problems for classification purposes (Wang, et al., 2005). Recent researches in the literature also show promises to use the semidefinite programming to identify the values of the parameters of ε, γ and the penalty term C (Lanckriet, et al., 2004). One possible direction of the research to improve the accuracy of classifying donors and non-donors is to use ensemble models where results from multiple models are synthesized to find donors and non-donors (Ha, et al., 2005).

Appendix A: Random Number Seeds

The following list of random number seeds is prepared in Excel@ with =ROUNDUP(RAND()*1000000,0) to be used in PROC SURVEYSELECT. One seed is for Testset, and another seed is for Trainset in order to take one third of the DMEF01 data set as the Testset and Trainset. The rest of the DMEF01 data set is used as Validset.

Random number seeds

No	seeds		No	seeds
1	9117481		11	6621954
2	619829		12	182924
3	4037405		13	9733365
4	859112		14	1406498
5	6586222		15	7501209
6	7404325		16	511477
7	8554983		17	4290198
8	8885021		18	5061750
9	2561098		19	3485255
10	3393504		20	7594695

Appendix B: SAS Program for Taking Stratified Samples from DMEF01 Data Set

A SAS program based on Suhr (2010) is written to use PROC SURVEYSELECT in order to take stratified training, validation and testing samples.

```
LIBNAME INFONE 'C:\Users'; DATA RAWSET; SET INFONE.DMEF1DAT; RUN;
PROC FREQ DATA = RAWSET; TABLES TARGDON/OUT=NEWFREQ NOPRINT; RUN;
DATA NEWFREQ2 ERROR; SET NEWFREQ;
SAMPNUM=(PERCENT * 4960)/100;
_NSIZE_= ROUND(SAMPNUM,1); SAMPNUM=ROUND(SAMPNUM,.01);
IF _NSIZE_=0 THEN OUTPUT ERROR; IF _NSIZE_=0 THEN DELETE;
OUTPUT NEWFREQ2; run;
DATA NEWFREQ3; SET NEWFREQ2;
KEEP TARGDON _NSIZE_; RUN;
PROC SORT DATA = NEWFREQ3; BY TARGDON; RUN;
PROC SORT DATA = RAWSET; BY TARGDON; RUN;

PROC SURVEYSELECT DATA=RAWSET SEED=7594695 OUT=TESTSET
SAMPSIZE=NEWFREQ3;
STRATA TARGDON; ID ACCNTNMB TARGDON; RUN;
DATA TESTSET; SET TESTSET; SETID = 3; RUN;
DATA MERGE1; SET RAWSET TESTSET;RUN;
PROC SORT DATA=MERGE1; BY ACCNTNMB; RUN;
DATA MERGE2; SET MERGE1; BY ACCNTNMB;
FIRST=FIRST.ACCNTNMB; LAST=LAST.ACCNTNMB; RUN;
DATA TRNVASET; SET MERGE2; IF FIRST=1 AND LAST=1;RUN;
```

PROC DELETE DATA = NEWFREQ NEWFREQ2 NEWFREQ3 SAMPFL SAMPFREQ
ERROR MERGE1 MERGE2 MERGE3; **RUN**;

PROC FREQ DATA = TRNVASET; TABLES TARGDON/OUT=NEWFREQ NOPRINT;
RUN;
DATA NEWFREQ2 ERROR; SET NEWFREQ; SAMPNUM=(PERCENT * **33066**)/**100**;
NSIZE= ROUND(SAMPNUM,**1**); SAMPNUM=ROUND(SAMPNUM,**.01**);
IF _NSIZE_=**0** THEN OUTPUT ERROR; IF _NSIZE_=**0** THEN DELETE;
OUTPUT NEWFREQ2; RUN;
DATA NEWFREQ3; SET NEWFREQ2; KEEP TARGDON _NSIZE_; **RUN**;
PROC SORT DATA = NEWFREQ3; BY TARGDON; **RUN**;
PROC SORT DATA = TRNVASET; BY TARGDON; **RUN**;
PROC SURVEYSELECT DATA=TRNVASET SEED=**4290198**
OUT=TRAINSET SAMPSIZE=NEWFREQ3; STRATA TARGDON;
ID ACCNTNMB TARGDON; **RUN**;
DATA TRAINSET; SET TRAINSET; SETID = **1**; **RUN**;
DATA MERGE1; SET TRNVASET TRAINSET; **RUN**;
PROC SORT DATA=MERGE1; BY ACCNTNMB; **RUN**;
DATA MERGE2; SET MERGE1; BY ACCNTNMB;
FIRST=FIRST.ACCNTNMB; LAST=LAST.ACCNTNMB; **RUN**;
DATA VALIDSET; SET MERGE2; IF FIRST=**1** AND LAST=**1**; SETID=**2**; **RUN**;
PROC DELETE DATA = NEWFREQ NEWFREQ2 NEWFREQ3 SAMPFL SAMPFREQ
ERROR MERGE1 MERGE2 MERGE3; **RUN**;

```
DATA SMP09(KEEP=ACCNTNMB TARGDON SETID); SET TRAINSET VALIDSET
TESTSET; RUN;
PROC DELETE DATA = TRAINSET VALIDSET TESTSET TRNVASET; RUN;
PROC SORT DATA=SMP09; BY ACCNTNMB; RUN;
DATA SMP09; SET SMP09; SETID09 = SETID; RUN;
DATA SMP09(KEEP=ACCNTNMB SETID09); SET SMP09; RUN;
LIBNAME INFONE "C:\USERS"; DATA SSAMPLES; SET INFONE.SSAMPLES; RUN;
PROC SORT DATA=SSAMPLES; BY ACCNTNMB; RUN;
DATA SSAMPLES; MERGE SSAMPLES SMP09;
BY ACCNTNMB; RUN;
LIBNAME OUTFILE "C:\USERS"; DATA OUTFILE.SSAMPLES; SET SSAMPLES; RUN;
LIBNAME INFONE 'C:\USERS'; DATA RAWSET; SET INFONE.RAWSET; RUN;
PROC SORT DATA=RAWSET; BY ACCNTNMB; RUN;
DATA RAWSET; MERGE RAWSET SMP09; BY ACCNTNMB; RUN;
LIBNAME OUTFILE "C:\USERS"; DATA OUTFILE.RAWSET; SET RAWSET; RUN;
/* SAS EXPORT CVS FILE FOR R SVM */
 LIBNAME INFONE 'C:\USERS';
  PROC EXPORT DATA=INFONE.RAWSET
       OUTFILE= "C:\USERS\RAWSET.CSV" DBMS=CSV REPLACE; RUN;
```

Appendix C: SAS Program to Prepare Initial Raw Data Set

```
LIBNAME INFONE 'C:\USERS';
DATA RAWSETA; SET INFONE.DMEF1DAT(KEEP =ACCNTNMB TARGRESP
TARGDOL CNDOL1 CNTMLIF CNTRLIF CONLARG CONTRFST SLTMLIF DATEFST
DATELRG CNDAT1 SLDAT1);
IF (TARGRESP=0 AND TARGDOL=0) THEN TARGDON=0; ELSE TARGDON=1;

/**** THE BASE YEAR 1900 WITH SAS# OF 1146 FOR 9506 AND
   1146-84*12 - 12 = 126 = 9506-8412 MONTHS */
RCDATFST=1146-INT(DATEFST/100)*12-MOD(DATEFST,100);
RCDATLRG=1146-INT(DATELRG/100)*12-MOD(DATELRG,100);
RCCNDAT1=1146-INT(CNDAT1/100)*12-MOD(CNDAT1,100);
RCSLDAT1=1146-INT(SLDAT1/100)*12-MOD(SLDAT1,100);
TRAN1=CNTRLIF/CNTMLIF;/*DOLLARS PER CONTRIBUTION*/
TRAN2=CNTRLIF/SLTMLIF;/*DOLLARS PER SOLICITATION*/
TRAN3=CNTMLIF/SLTMLIF;/*CONTRIBUTIONS PER SOLICITATION*/
TRAN4=LOG(RCCNDAT1+1);
TRAN5=SQRT(CNTRLIF);/*SQUARE ROOT OF MONETARY VALUE*/
TRAN6=1/(CNTRLIF);/*INVERSE OF MONETARY VALUE*/
TRAN7=1/(TRAN3);/*INVERSE OF TRAN3, SOLICITATIONS PER CONTRIBUTION*/
TRAN8=SQRT(TRAN2);/*SQUARE ROOT OF TRAN2, DOLLARS PER SOLICITATION*/
TRAN9=(SLTMLIF)*(SLTMLIF);/*SQUARE OF SOLICITATIONS*/
TRAN10=LOG(TRAN1);/*LOG OF TRAN1, DOLLARS PER CONTRIBUTION*/
TRAN11=CNTRLIF/(RCCNDAT1+1);
TRAN12=(CNTMLIF/(RCCNDAT1+1));
```

```
TRAN13=LOG(CNTRLIF/(RCCNDAT1+1));

TRAN14=LOG(CNTMLIF/(RCCNDAT1+1));

RUN;

DATA RAWSET; SET RAWSETA (DROP = DATEFST DATELRG CNDAT1 SLDAT1);

RUN;

LIBNAME INFONE 'C:\USERS'; DATA INFONE.RAWSET; SET RAWSET; RUN;
```

Appendix D: SAS Program to Prepare TRAINSET, VALIDSET, TRNVASET and TESTSET

```
libname infone 'C:\Users;

DATA RAWSET; SET INFONE.RAWSET; RUN;

DATA TRAINSET; SET RAWSET; IF (SETID=1); RUN;

DATA VALIDSET; SET RAWSET; IF (SETID=2); RUN;

DATA TRNVASET; SET RAWSET; IF (SETID=1 | SETID=2); RUN;

DATA TESTSET; SET RAWSET; IF SETID=3; RUN;
```

Appendix E: SAS Program for Multiple Regression, Logistic Regression, and Extracting Performance Measures

```
PROC GLMSELECT DATA=TESTSET PLOTS=ALL;
  MODEL TARGDOL =  CONLARG CONTRFST RCSLDAT1 CNDOL1
        CNTMLIF CNTRLIF RCCNDAT1 RCDATFST RCDATLRG SLTMLIF TRAN1
        TRAN2 TRAN3  TRAN4 TRAN5 TRAN6 TRAN7 TRAN8 TRAN9 TRAN10 TRAN11
        TRAN12 TRAN13 TRAN14
           / DETAILS=ALL STATS=ALL SELECTION=STEPWISE;
SCORE DATA = TRNVASET OUT=PREDDON; RUN;
/* ------------------------- */
ODS HTML; ODS GRAPHICS ON;
PROC GLMSELECT DATA=TRAINSET PLOTS=ALL;
  PARTITION FRACTION(VALIDATE=0.10);
  MODEL TARGDOL =  CONLARG CONTRFST RCSLDAT1 CNDOL1
  CNTMLIF CNTRLIF RCCNDAT1 RCDATFST RCDATLRG SLTMLIF TRAN1 TRAN2
TRAN3 TRAN4  TRAN5 TRAN6 TRAN7 TRAN8 TRAN9 TRAN10 TRAN11 TRAN12
TRAN13 TRAN14
           / DETAILS=ALL STATS=ALL CVMETHOD=RANDOM(10)  CVDETAILS=ALL
             SELECTION=STEPWISE;
SCORE DATA = TRNVASET OUT=PREDDON; RUN;
ODS GRAPHICS OFF; ODS HTML CLOSE;
/* ----------------------------------*/

/**** VARIALES FROM PROC GLMSELECT  **/
PROC REG DATA=TRNVASET OUTEST=TRNESTIM;
```

```
MODEL TARGDOL  =
/ CLB;  OUTPUT OUT=TRNOUSET; RUN;

ROC RANK DATA=PREDDON GROUPS=10 OUT=RANKED_A5; VAR P_TARGDOL;
   RANKS RANK_TDL; RUN;
PROC SORT DATA=RANKED_A5; BY RANK_TDL; RUN;
PROC MEANS DATA=RANKED_A5 MEAN N NMISS MAX P99 P95 P90 Q3 SUM RANGE
FW=8;
VAR TARGDOL; BY RANK_TDL; OUTPUT OUT=RANKMEAN; RUN;
```

Appendix F: R Program for the SVMs

```
library(MASS) # for write.matrix()

library("class")  # to load e1071

library("e1071")  # to load svm

rawset<- read.csv(file="C:/Users/rawset.csv",head=TRUE,sep=",")

TrainSet<-subset(rawset, rawset$setid ==1)

TrnVaSet<-subset(rawset, rawset$setid !=3 )

ValidSet <-subset(rawset, rawset$setid !=2 )

TestSet<-subset(rawset, rawset$setid !=3 )

model.svm <- svm(targdol ~ ontrfst+conlarg+cntmlif+cntrlif+sltmlif+cndol1+rcdatfst+rcdatlrg+
    RCCNDAT1+RCSLDAT1+tran1+tran2+tran3+tran4+tran5+tran6+
    tran7+tran8+tran9+tran10+tran11+tran12+tran13+tran14,
data =TrainSet, scale = T, kernel = "radial", epsilon = 0.125,
shrinking = T, cross = 10, probability = T, fitted = TRUE, degree = 1, coef0 = 0,
cost =2^(4), gamma = 2^(-4) , nu = 0.5, type="eps-regression")


###  C-classification  is used for classifying donors versus non donors  /* epsilon = 0.01  */

### eps-regression is used for regression

tunesvm <- tune.svm(targdol ~., data = TrainSet, gamma = 2^(-4:-2), cost = 2^(2:4))

predict.svm <- predict(model.svm, newdata=testset)

write.matrix(predict.svm,"predict.txt",sep=",")
```

Appendix G: SAS Program to Summarize Performance Measures for the SVMs

```
FILENAME INFPRED 'C:\USERS\PREDICT.TXT';

DATA PREDICT; INFILE INFPRED;

INPUT PRED; NEWID =_N_;RUN;

DATA TESTSET; SET TESTSET; NEWID =_N_;RUN;

DATA SVMPRED; MERGE VATEST PREDICT; BY NEWID; RUN;
```

Appendix H: Variable Definitions

The following are the definitions of the fields or variables.

ACCNTNMB is the donors' account number. The donors' account number is used to identify a specific donor during the model development and solicitation processes.

TARGRESP is the number of times of donations in the Fall 1995.

TARGDOL is the total dollar amount of donations in the Fall 1995. The TARGRESP and TARGDOL are the dependent variables in this study. They are used to derive a binary response variable, TARGDON, for developing predictive models. TARGDON is st to 0 if both TARGRESP and TARGDOL are equal to zero; otherwise TARGDON is set to 1.

CONTRFST is the dollar amount of the first contribution.

DATEFST is the data of the first contribution.

CONLARG is the dollar amount of the largest Contribution.

DATELRG is the date of the largest contribution.

CNTMLIF is the total number of times contributed lifetime.

CNTRLIF is the total dollar amount of contributions lifetime.

SLTMLIF is the number of times of solicitations lifetime.

CNDAT1 is the Latest Contribution Date. RECDAT1, the recency from the latest contribution date to 6/1995 is given by RECDAT1= Date of 6/1995 - CNDAT1 in months.

CNDOL1 is the dollar amount of the latest Contribution.

SLDAT1 is the latest Solicitation Date.

The following fourteen variables are derived from the twelve variables above based on Malthouse (2002).

TRAN1 is the dollars per contribution (CNTRLIF/CNTMLIF).

TRAN2 is the dollars per solicitation (CNTRLIF/SLTMLIF).

TRAN3 is the contributions per solicitation (CNTMLIF/SLTMLIF).

TRAN4 is LOG (RCCNDAT1+**1**).

TRAN5 is square root of monetary value (SQRT (CNTRLIF)).

TRAN6 is inverse of monetary value (**1**/ (CNTRLIF)).

TRAN7 is inverse of tran3, solicitations per contribution (**1**/ (TRAN3)).

TRAN8 is square root of TRAN2, dollars per solicitation (SQRT (TRAN2)).

TRAN9 is square of solicitations ((SLTMLIF)*(SLTMLIF)).

TRAN10 is log of TRAN1, dollars per contribution (LOG (TRAN1)).

TRAN11 is monetary value/ (RCCNDAT1+**1**) (CNTRLIF/ (RCCNDAT1+**1**)).

TRAN12 is CNTMLIF/ (RCCNDAT1+**1**).

TRAN13 is LOG (CNTRLIF/ (RCCNDAT1+**1**)).

TRAN14 is LOG (CNTMLIF/ (RCCNDAT1+**1**)).

SEX is Gender.

SEXF2M=1 means both of male and female; SEXCOM=1 means company; SEXFEM=1 means female; SEXMAL=1 means male and otherwise unknown.

RCCNDAT1 is months since latest contribution. This was computed from the CNDAT1 variable, latest contribution data.

RCDATFST is months since first contribution data. This was computed from the DATFST variable, the data of the first contribution.

RCDATLRG is months since largest contribution data. This was computed from the DATELRG variable, the date of the largest contribution.

RCSLDAT1 is the months since latest solicitation data. This was computed from the SLDAT1 variable, the latest Solicitation Date.

```
RCCNDAT1=1146-int(CNDAT1/100)*12-mod(CNDAT1,100); /* base year is 1900
and 9506 = sas number 1146*/
rcdatfst=1146-int(datefst/100)*12-mod(datefst,100); /* base year is
1900 and 9506 = sas number 1146*/
rcdatlrg=1146-int(datelrg/100)*12-mod(datelrg,100); /* base year is
1900 and 9506 = sas number 1146*/
```

```
RCSLDAT1=1146-int(SLDAT1/100)*12-mod(SLDAT1,100); /* base year is 1900
and 9506 = sas number 1146*/
```

References

1. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Knowledge Discovery and Data Mining, 1998. in press.
2. Boslaugh, Sarah and Paul Andrew Watters, Statistics in a nutshell, O'Reilly Media, Inc., Sebastopol, CA, 2008.
3. Cole, Karen, Rachel Dingle, Rajesh Bhayani, Pledger modeling: Help the Aged case study, International Journal of Nonprofit and Voluntary Sector Marketing. London: Feb 2005. Vol. 10, Iss. 1; p.43-52.
4. Cui, Dapeng and Curry, David, Prediction in Marketing Using the Support Vector Machine, Marketing Science, Fall 2005, 24, 4; ABI/INFORM Global p.595.
5. Chang, C.C., Lin, C.J., 2001. LIBSVM: A library for support vector machine. Software available from: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
6. DMEF, Direct Market Education Foundation, Non-Profit Organization Data Set One, http://www.directworks.org/.
7. Deichmann, Joel; Eshghi, Abdolreza; Haughtom, Dominique; Sayek, Selin; Teebagy, Nicholas; Application of Multiple Adaptive Regression Splines (MARS) in Direct Response Modeling. Journal of Interactive Marketing; Autumn 2002; 16, 4; ABI/INFORM Global p. 15.
8. Diamond, W. D., and Noble, S. M. "Defensive Responses to Charitable Direct Mail Solicitations." Journal of Interactive Marketing, 2001, 15 (3), p. 2–12.
9. Goodman, Steve and Gary Plouff, Neural Network Modeling: Artificial Intelligence Marketing Hits the Non-Profit World, Fund Raising Management, June 1997, p.19-20.
10. Haughton, Dominique and Oulabi, Samer Direct Marketing Modeling with CART and CHAID, Journal of Direct Marketing, Vol. 11, No. 4, Fall 1997, p.42-52.

11. Ha, Kyoungnam, Sungzoon Cho, and Douglas MacLachlan, Response models based on bagging neural networks, Journal of Interactive Marketing, Winter 2005, Vol. 19, Iss. 1, p 17-30.
12. Hansotia, Behram, and Brad Rukstales, Incremental Value Modeling, Journal of Interactive Marketing, Hoboken, Summer 2002, Vol. 16, Iss. 3, p. 35-46.
13. Key, Jennifer, Enhancing fundraising success with custom data modeling, International Journal of Nonprofit and Voluntary Sector Marketing. London: Nov 2001. Vol. 6, Iss. 4; p. 335-346.
14. Lanckriet, Gert R. G., Cristianini, Nello, etal., Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research 5 (2004) p. 27-72.
15. Malthouse, Edward C., Assessing the performance of direct marketing scoring models, Journal of Interactive Marketing, Winter 2001, Vol. 15, Iss. 1, p. 49-62.
16. Malthouse, Edward C., Performance – based variable selection for scoring models, Journal of Interactive Marketing, Autumn 2002, Vol. 16, Iss. 4, p. 37-50.
17. Mangasarian. Linear and nonlinear separation of patterns by linear programming. Operations Research, 13: p. 444-452,1964.
18. Mangasarian. Nonlinear Programming. McGraw-Hill New York, NY, 1969.
19. Menard, Scott W. , Applied logistic regression analysis, 2nd edition, Quantitative Applications in Social Sciences, SAGE University Paper, v. 106, 2002.
20. Neter, John, Kutner, M., Nachtsheim, C. J., and Wasserman, W., 1996, Applied Linear Statistical Models 4th edition, New York, Irwin.
21. Smola, Alex J. and Bernhard Schölkopf, A tutorial on support vector regression, September, 2003, Retrieved on March 20, 2008 [http://www.svms.org/regression/SmSc98.pdf].

22. Suhr, Diana, Selecting a Stratified Sample with PROC SURVEYSELECT, Retrieved on April 22 2010 [http://www.wuss.org/proceedings09/09WUSSProceedings/papers/cod/COD-Suhr.pdf].
23. Shtatland, E. S., Kleinman K., and Cain E. M. (2004). A new strategy of model building in PROC LOGISTIC with automatic variable selection, validation, shrinkage and model averaging. SUGI '29 Proceeding, Paper 191-29, Cary, NC: SAS Institute, Inc.
24. Vapnik and A. Lerner. Pattern recognition using generalized portrait Method. Automation and Remote Control, 24, 1963.
25. Vapnik and A. Chervonenkis. A note on one class of perceptrons. Automation and Remote Control, 25, 1964.
26. Vapnik, Vladimir. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.
27. Wang, Jiaqi, Wu, Xindong, Zhang, Chengqi, Support vector machines based on K-means clustering for real-time business intelligence systems, Int. J. Business Intelligence and Data Mining, Vol. 1, No. 1, 2005.
28. Yu, Hwanjo and Kim, Sungchul, SVM Tutorial: Classification, Regression, and Ranking, Retrived on 4/20/2010 at [http://hwanjoyu.org/publication/svmtutorial.pdf].
29. Zahavi, J, and Levin, N (1997), Applying Neural Networking Computing to Target Marketing, Journal of Direct Marketing, 11(1), p. 5-22.