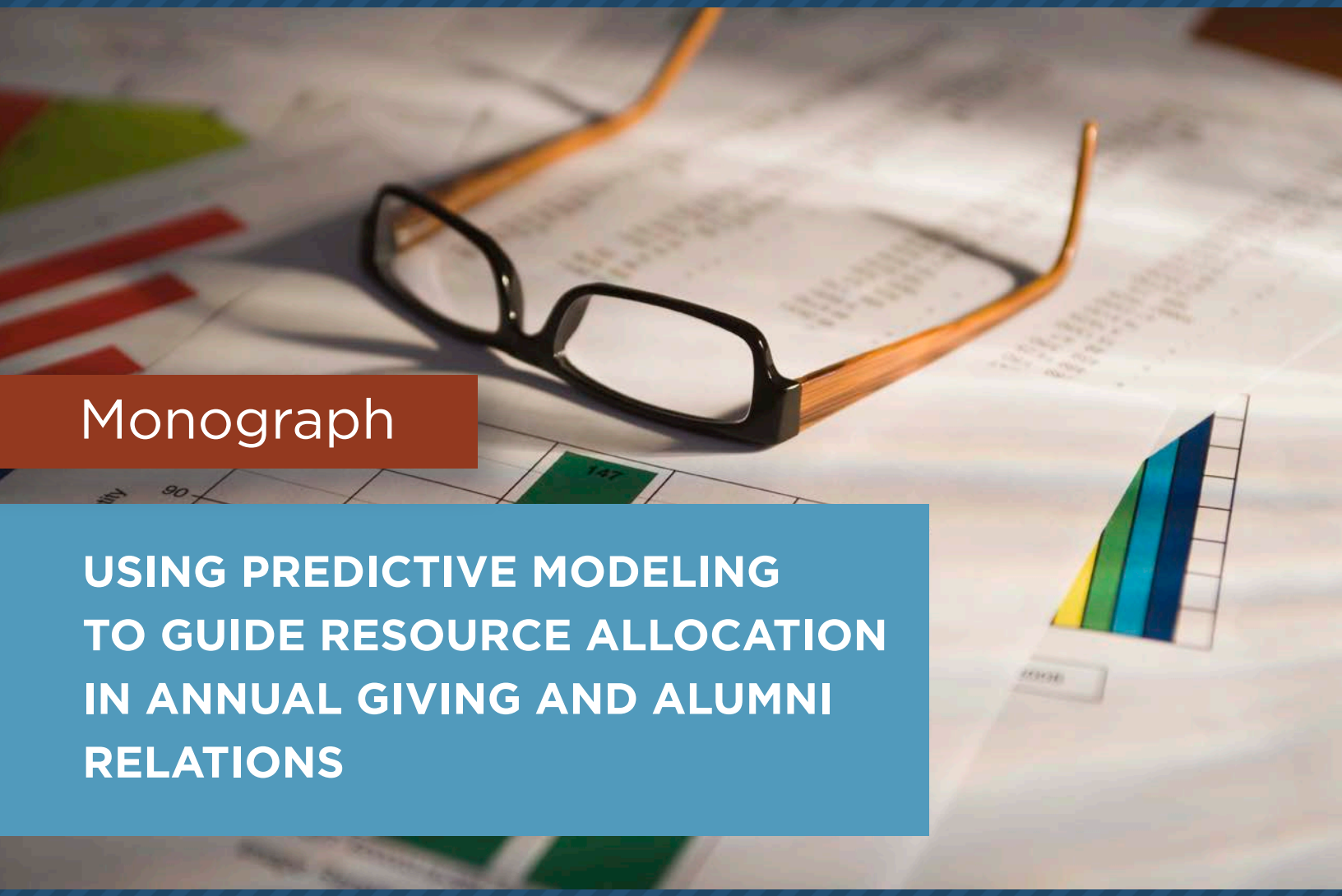
ACADEMIC IMPRESSIONS

ACADEMIC IMPRESSIONS: HIGHER ED IMPACT



Monograph

USING PREDICTIVE MODELING TO GUIDE RESOURCE ALLOCATION IN ANNUAL GIVING AND ALUMNI RELATIONS

Authors: Bob Baer and Kylie Mitchell
Business Analytics
Development and Alumni Relations
University of Pennsylvania

HIGHER ED IMPACT

Delivered free to your inbox, *Higher Ed Impact* provides you with a full tool kit to help you monitor and assess the trends and strategic challenges likely to have an impact on your institution's health and competitiveness.

DAILY**PULSE**

Scan current events, timely research, and notable practices at other institutions.

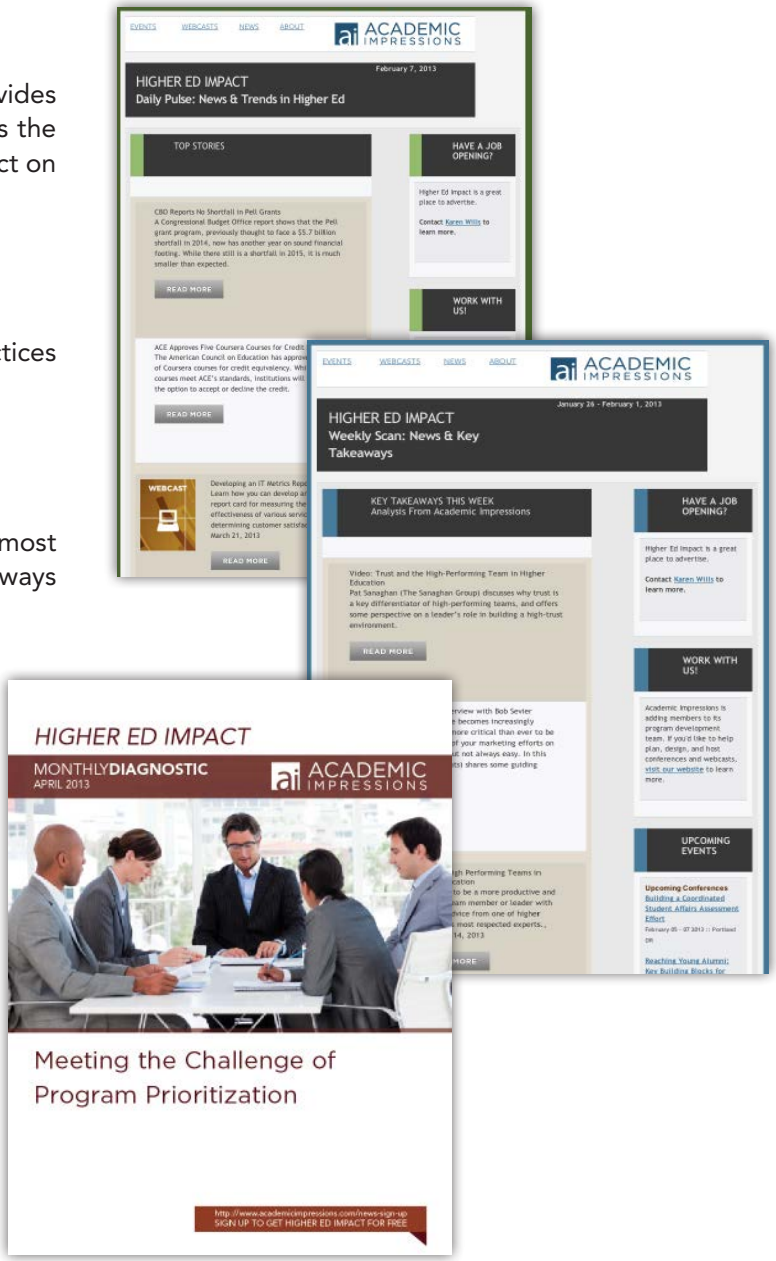
WEEKLY**SCAN**

Review the week's most significant events and the most timely research in higher education, with key takeaways suggested by higher education's leading experts.

MONTHLY**DIAGNOSTIC**

Get an enterprise-wide and in-depth look at a current, strategic challenge; identify steps to take and critical questions to address.

Learn more or sign up to receive *Higher Ed Impact* at: http://www.academicimpressions.com/news-sign-up



ABOUT ACADEMIC IMPRESSIONS

We are an organization that exclusively serves higher education professionals. We offer focused and intentionally crafted learning experiences to help you address your most pressing challenges.

Our work addresses a range of issues related to student recruitment and retention, faculty support and development, alumni engagement and development, and increasing organizational productivity.

Learn more at **www.academicimpressions.com.**

USING PREDICTIVE MODELING TO GUIDE RESOURCE ALLOCATION IN ANNUAL GIVING AND ALUMNI RELATIONS

AUTHORS: Bob Baer and Kylie Mitchell
Business Analytics
Development and Alumni Relations
University of Pennsylvania

TABLE OF CONTENTS

Foreword

Most alumni relations and annual giving operations have limited intelligence about who will be a strong reunion volunteer, annual giver, or alumni travel prospect—if the person has not previously participated in any of those activities. But rather than pulling a random database query and then reaching out at random to the contents of the entire resulting list, applying predictive analytics can help provide a more targeted allocation of your resources and more targeted messaging.

An In-Depth Look at Predictive Modeling

In one example of such predictive analytics, the University of Pennsylvania has found a series of model variables that identify strong prospects. We reached out to Bob Baer and Kylie Mitchell, who worked extensively on this project. In this monograph, Baer and Mitchell address the methodology they used (to help you build on their findings) and then walk you through the independent variables that Penn found to be statistically significant and helpful in guiding decisions in annual giving and alumni relations. With these variables in hand, you'll be able to more effectively **scour your own database to find the best prospects** for various alumni relations and annual giving functions.

In this report, you'll read about using predictive modeling to target:

- → Alumni for reunion-year annual giving
- → Alumni for giving in a non-reunion year—especially looking for prospects among lapsed and never givers
- → Institutional "friends" for annual giving
- → Prospective alumni reunion volunteers
- → Alumni who are most interested in alumni tours.

A Low-Cost Approach

Software. While more sophisticated modeling software has been used for the applications in this report, most of the predictive models that Baer and Mitchell discuss have also been run using the data analysis tools provided in Microsoft Excel.

Staff. The projects described here are the collaborative work of only two analytics professionals—both of whom had other analytics project responsibilities.

And you can get started even with fewer resources.

Your institution may have students and faculty with the background and expertise to use the techniques described here. A partnership between fundraising professionals and academics could be an ideal low-cost approach to get started in predictive modeling.

To Use this Monograph Effectively

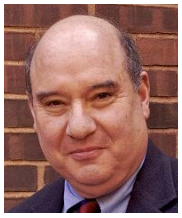
This monograph provides a detailed primer on predictive analytics, and the examples given are very technical.

- → If you already have an **advanced knowledge** of predictive analytics, read this report for insights on variable definition and selection, model structure, data conditioning, and assessment of model "validity" and performance.
- → If you have **little or no prior background** in predictive modeling, read this report to gain an understanding of how one uses predictive modeling, the data requirements, the results you can obtain, and the potential impact for your organization.

About the Projects in this Monograph

The projects described in this monograph were completed between 2010 and 2012 in Development and Alumni Relations at the University of Pennsylvania. Bob Baer was responsible for the earliest of these projects, the Penn Fund Reunion Donor Participation model. All of the other projects were the product of collaboration of Kylie and Bob. On the Reunion Planning Committee Volunteer model, Reunion Gift Committee Volunteer model, and the Penn Vet Annual Giving model projects, Kylie was the principal investigator. On the Non-Reunion Donor Participation model, the Wharton MBA Reunion Donor Participation model, and the Targeting for Alumni Travel model, Bob was the principal investigator.

Authors



Bob Baer

Dr. Robert L. Baer is senior director of business analytics in Development & Alumni Relations at the University of Pennsylvania. For nearly four months in 2011, he was also interim director of development and alumni relations for Penn Athletics while Penn was recruiting a new executive director.

Prior to coming to Penn over thirteen years ago, Bob was a consultant in private practice. For twenty-four years, he worked for the DuPont Company in product marketing, brand marketing, marketing communications, marketing and sales management and marketing research.

His work in linear programming models, simulation modeling, and predictive analytics began in the late 1960s using large mainframe computing systems at the University of Pennsylvania and at the DuPont Company.

Bob is an adjunct faculty member in the Marketing Department at Drexel University in Philadelphia, teaching Introduction to Marketing, Statistics, Business Communications, and Marketing Research courses.

Bob has a BS degree in Chemical Engineering and MS and PhD degrees in Systems Engineering & Operations Research all from Penn.

Kylie Mitchell

Until mid-July 2012, Kylie Mitchell was assistant director of Business Analytics in Development & Alumni Relations (DAR) at the University of Pennsylvania. During her three years in DAR Business Analytics, she worked on a wide range of data mining, data analysis, survey research, and predictive modeling projects.

Kylie graduated cum laude with a BA in Chemistry from the University of Pennsylvania in May 2012 and is currently pursuing a PhD in Inorganic Chemistry at the University of Florida.

OVERVIEW

Development and Alumni Relations (DAR) at the University of Pennsylvania plays an important role in helping to provide the resources to support the mission of the university and its vision for the future. DAR operates as a decentralized organization within a centralized management structure, with over 600 employees across the university in more than 25 Schools, Centers and Central functions. The Business Analytics group consisting of 2–3 professionals, supports these DAR areas with data analysis, data mining, surveys, focus groups, and, of course, predictive modeling projects. The analytical work described in this monograph was performed solely by the authors working both collaboratively and independently depending on the particular project.

Defining Predictive Modeling

Predictive modeling as used in this monograph is the application of statistical analysis to predict future outcomes from an investigation of historical data. More specifically, we are trying to identify relationships between the variable whose outcome we are trying to predict and the "explanatory" variables which most strongly influence or "predict" that outcome.

We refer to the variable we are trying to predict as the dependent variable and the "explanatory" variables as the independent variables. The dependent variable is also referred to as the target variable. After considerable data filtering and conditioning, every individual included in the analysis has independent variables each with a numerical value and a dependent variable with an actual (numerical) value. The individuals are referred to as observations, records or subjects.

Most of the work described in this monograph uses multiple regression analysis, a statistical technique described later in more detail. The output of a multiple regression analysis is a predicted numerical value for the target variable. For this reason, we sometimes refer to this type of predictive model as a "scoring" model, and the activity is referred to as "scoring" each record in the dataset. The higher the model score, the greater the likelihood that individual will exhibit the behavior we are trying to predict.

The output of a multiple regression analysis also includes the equation to calculate the model score and a goodness of fit measure. Although the approach is different, modeling with neural net methodology also yields a model score, an equation and a goodness of fit metric. Decision tree modeling develops a set of decision rules that can be applied to the independent variables to separate individuals likely to have the target variable behavior we are seeking. Goodness of fit can be calculated to compare alternative decision tree models, but there is no equation, per se.

Why is Predictive Modeling Important?

Organizations seeking to improve operating performance are turning to predictive analytics and predictive modeling to either increase revenues, decrease costs, or both. The size of our alumni population and database at Penn make focusing resources critical to fundraising success. The objectives of the applications described in this monograph are to reduce fundraising related marketing costs and/or increase the dollars raised, or improve business performance in some other way.

Several examples in the monograph produce results that enable us to focus on prospects most likely to be donors. Targeting the 40% mostly likely to donate can save mailing costs, phone solicitation costs and even development officer time and travel expenses. Another example described below involves target marketing for Alumni Travel programs. Here, predictive modeling focuses the direct mail campaigns on those most likely to take a Penn Alumni trip, saving mailing costs and/or increasing alumni response rates. Recruiting reunion planning and gift committee members each year starts with an invitation to 25,000+ alumni to a Reunion Planning Leadership Conference. Because this "call" for volunteers may not yield a sufficient number of volunteers, predictive modeling has been used to identify another 300–500 of the best potential volunteers for personal contact by alumni relations and/or reunion annual giving staff.

How We Do Predictive Modeling

The steps employed in the predictive modeling process are to:

- Identify the business need
- Define the problem
- Assemble the data
- Condition the data
- Build the model
- Revise and test
- Implement the results

Identification of the business need and a proper problem definition are critical for a successful project. In most data warehouse systems, the data is encoded in ways that are not immediately useful for predictive modeling. Basically, the data as it exists in the data warehouse will require transformation and conditioning to be useful for predictive modeling. Assembling and conditioning the data can easily account for 90% of the total project effort.

Building the model once the data has been prepared is relatively easy using available software tools. Revising and testing is also relatively easy compared to the work of conditioning the data. Implementation will depend on how the organization will use the model results. Some models are used by only a few people, while others will be used across the entire organization and will need to be available in the operations database.

Although this monograph is intended to provide a reasonably detailed description of the methodology and several specific Annual Giving and Alumni Relations projects, areas like decision trees and neural nets are only covered briefly. An excellent reference for further study is *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, by Michael J.A. Berry and Gordon S. Linoff. (1)

METHODOLOGICAL CONSIDERATIONS

Selecting Model Variables

Choosing the variables to include in a predictive model is a critical first step in every modeling project. The dependent variable which is what we are trying to predict is usually defined by the project objectives. In the examples in this monograph, we are trying to target people most likely to be a donor, a volunteer, or a participant in an alumni tour. For major gift prospects ($25,000 or more) we may be targeting prospects likely to give at or above their gift capacity over a five year period. We might even try to predict giving relative to gift capacity. A modeling project in academic admissions may target applicants most likely to graduate in four years from the institution, or perhaps even predict their final grade point average at graduation. A dependent variable can be either binary (1 or 0) or continuous like grade point average.

There are two different approaches to the selection of independent variables. One approach is to include all the available variables and let the modeling software "automatically" determine those that are predictive or statistically significant, excluding the other variables that are not predictive. Using that approach it would be easy to start with 75–100 different variables. We recommend a more purposeful, deliberate approach to selecting independent variables described below.

As a general rule, to be useful as an independent variable, a "goodly" number of records should have the property measured by that variable. A binary independent variable, where all the records have 0's or 1's is not practically useful as a predictor. In a binary donor participation model, for example, a variable which measures whether an individual is a member of the board of trustees, an overseer board or the alumni executive board (=1) or not a member of these boards (=0), is very predictive of annual giving donor participation (greater than 95% are donors), but fewer than 1% of alumni have ever been on one or more of these volunteer boards.

An independent variable should also have at least an implied causal relationship with the dependent variable. Looking again at the volunteer board variable discussed above, the causal relationship is in the wrong direction. Donor participation is an "implied" qualification for board membership. Board membership does not "cause" donor participation in a predictive modeling sense and is not really useful as an independent variable.

Independent variables should not be highly correlated with other independent variables. When two independent variables are highly correlated, the one that is less correlated with the dependent variable is usually excluded. We will say more about this later.

Finally, independent variables must be measureable in terms of the practical application of the model. Attendance at the reunion is a very strong predictor of donor participation in a reunion year. Alumni who participate in their reunion are very likely to be donors that year. Unfortunately at Penn, this independent variable is virtually useless. Class reunions and Alumni Weekend at Penn are held in mid-May. By the time we know who attended their reunion, the fiscal year is almost over; we know who is a donor in their reunion year and don't need to predict. To be useful at Penn, reunion year donor participation predictive model results are need 10–11 months before the reunion.

Data Conditioning

As mentioned above, assembling and conditioning the data for a predictive modeling project can easily account for most of the total project effort. Because the data is encoded in most data warehouse systems in ways that are not immediately useful for predictive modeling, extensive transformation and conditioning is usually required.

Many variables used in predictive modeling are simple binary variables, taking on the values of yes or no (1 or 0 in the actual model dataset). Converting and conditioning warehouse data can be time consuming and sometimes complicated. For example, on one individual's record, our database will tell us the name of their spouse, if they have one. Whether the spouse is a Penn alumna or alumnus requires looking at the spouse's data record—if they have one—and some additional data processing. Translating that into a binary (1/0) variable where 1 means the individual has an alumni spouse and 0 means they do not have an alumni spouse requires additional processing.

When we consider an individual's giving behavior, we have a number of different possiblities. If we are interested in whether an individual was a donor in FY2010, it is a simple yes/no or 1/0. On the other hand sometimes a measure of donor consistency might be useful. For example, for each individual, we can have string of 15 binary (1/0) variables representing whether the individual gave in each of the last fifteen years. The sum of these binary variables can take on values from zero to fifteen years of donor participation. This donor consistency variable is like a continuous variable, but can only have discrete integer values.

In some of our modeling work, we have expanded these donor consistency variables into a series of binary variables as shown in Table 1.

Table 1. Recoding a Donor Consistency Variable

Value of original variable	Recoded variable
15 years of giving out of 15 years	Giving 15 out of 15 (yes/no or 1/0)
11–14 years of giving out of 15 years	Giving 11–14 out of 15 (yes/no or 1/0)
6–10 years of giving out of 15 years	Giving 6–10 out of 15 (yes/no or 1/0)
1–5 years of giving out of 15 years	Giving 1–5 out of 15 (yes/no or 1/0)
No years of giving out of 15 years	Giving None out of 15 (yes/no or 1/0)

Sometimes variables need to be transformed in other ways to be useful. In some of our modeling work, we have used lifetime giving of the individual as an independent variable. This variable ranges from $0 to over $100 million. A useful transformation here is to take the logarithm to the base 10 to "flatten" this variable. Table 2 below illustrates this approach:

Note: The logarithm to the base 10 of a number is the exponent by which 10 has to be raised to produce that number. For example, 10^6 = 1,000,000, hence the logarithm to the base 10 of 1,000,000 is 6.0.

Table 2. Logarithmic Transformation of Lifetime Giving

Lifetime giving original variable	Recoded variable ($\log_{10}$ of lifetime giving)
$200,000,000	8.3
$100,000,000	8.0
$10,000,000	7.0
$1,000,000	6.0
$100,000	5.0
$10.000	4.0
$1,000	3.0
$100	2.0
$10	1.0
$1	0.0
Under $1	Recoded as 0

Sometimes we have grouped binary variables into an aggregated binary variable. An example is the set of variables that characterize relationships to alumni. Originally these might have been included in the model separately, but after model testing it was decided to combine them.

Table 3. Recoding of Alumni Relationships

Original relationship variables	Recorded relationship variable
Alumni spouse (1/0)	Have an alumni spouse and/or alumni childred and/or alumni parents and/or alumni siblings and/or other alumni relatives (1/0)
Alumni children (1/0)	
Alumni parents (1/0)	
Alumni siblings (1/0)	
Other alumni relatives (1/0)	

There are many other possibilites for transformation of variables, but these are a few we have used most.

Data Exploration Using Crosstabs

The use of cross-tabs (two-way frequency distributions or contingency tables) can provide insights as to relationships between the actual values of the variable we will try to predict (the dependent variable) and one of the explanatory variables (independent variable). Cross-tabs have been used in many of the projects described here for data exploration, and several examples are provided in the section on "Targeting Undergraduate Reunion Year Donors." Cross-tab results may, in some situations, be sufficient to eliminate the need for multiple regression analysis, especially if statistical analysis support is not readily available in an organization.

The example used in Table 4 is a cross-tab for 119,194 individuals showing the relationship between giving in their last reunion year and whether they gave between their last two reunions. Overall, 26% of these individuals gave in the year of their last reunion. When we look only at those 38,268 individuals who gave between their last two reunions, however, 65% gave in the year of their last reunion vs. only 7% of alumni who did not give between reunions gave in the year of their last reunion.

Table 4. Giving at Last Reunion vs. Giving Between Reunions

	Totals	Did not give between last two reunions	Gave between last two reunions
Gave last reunion year	30,451	5,576	24,875
Did not give last reunion year	88,743	75,350	13,393
Totals	119,194	80,926	38,268

	Totals	Did not give between last two reunions	Gave between last two reunions
Gave last reunion year	26%	7%	65%
Did not give last reunion year	74%	93%	35%
Totals	100%	100%	100%

If we were dealing with samples rather than the entire undergraduate population with two or more reunions, the statistical method for the difference of proportions (7% vs. 65%) would be appropriate. Later we will discuss using the Chi-square statistical test on cross-tab or contingency tables. Although Chi-square and difference of proportions are quite different statistical methods, similar results are produced.

Because many times we are comparing two binary variables, the dimensions of the cross-tabs are 2 X 2. If, however, we were exploring giving in the last reunion year vs. undergraduate school (Arts & Sciences, Wharton, Engineering and Nursing) our crosstab would be 2 X 4 (donor/non-donor X four undergraduate schools).

Cross-tab data analysis, in addition to simplicity, has the advantage of being able to accommodate to missing data. Instead of a simple Yes/No variable, we can have Yes/No/Don't Know. The subject of missing data will be discussed below in more detail.

Sampling Considerations

Once your data is downloaded, cleaned, and conditioned, sample size must be considered. The first question that is often asked is, "How many records does it take to create a good model?" This question, unfortunately, is not a simple one. There really is no magic number of records that will create a good model. Obviously, the more data you have to work with, the better the model is likely to be, even if you do not use all of the data. A generally accepted rule of thumb is that there should be ten times the number of records as there are variables being considered to account for all possible patterns in the outcome. This decreases the possibility that any of the outcomes are due to chance. Unfortunately, we do not always have access to this volume of data. A good predictive model can still be developed on a smaller amount of data. However, the effect of sample size can best be gauged by running models of different sample size and comparing the R Square value as well as the beta weights (coefficients) for the variables and statistical significance of the beta weights.

Is it possible to have too much data? In most situations, it is best to use all the records available for model building, testing and validation. There are, however, some cases when this is not true. There may just be too many records, and a subset selected at random may be sufficient, and more practical.

The Penn Reunion Volunteer Recruiting model is a situation where non-representative sampling is necessary. A typical undergraduate reunion cohort has 30,000 alumni. reunion planning is done by approximately 250 volunteers working closely with Alumni Relations staff. Our target (dependent) variable in our multiple regression model has nearly 30,000 observations with a value of 0, and only 250 with a value of 1, a ratio of more than 100:1. With a ratio this high, the observations with a 1 are overwhelmed by the 0's and the resulting model will not be reliable.

The solution is to retain all the records with a target variable value of 1, and select a random sample of records with a value of 0. Many authors and statisticians often recommend that a 10:1 ratio of 0's to 1's is ideal. We have found that the ratio can even be as large as 25:1 and some statisticians do recommend a sample in this range.

Again, to determine the appropriate ratio, it is best to run the model with different sample size ratios and compare R Square values, and the beta weights and their statistical significance. Table 5 compares multiple regression model outputs for the entire population (100:1 ratio) and a non-representative sample (13:1). Clearly the 13:1 model is a major improvement. The goodness of fit is much better, the coefficient values are larger, and the range of predicted model scores will be larger.

Table 5. Balancing Target Variable 0's and 1's Using a Holdout Random Sample (from the Penn Reunion Volunteer Recruiting Model)

	Entire population	Random sample of 0's
Observations with Target value = 1	242	242
Observations with Target value = 0	30,141	3,258
Ratio of 0's to 1's	~100:1	~13:1
Multiple R	0.1983	0.461
R Square	0.0393	0.2125
Adjusted R Square	0.039	0.2105
Standard Error	0.0871	0.2255
Observations	30,383	3,500
Attended Any Homecoming Weekend (1/0)	0.0745	0.3393
Harrison Society member (1/0)	0.0471	0.1726
Alumni Child (1/0)	0.027	0.1355
Alumni Census 1998 – Penn Top 4 Giving Priorities (1/0)	0.0168	0.1181
Student Activity Service (1/0)	0.0128	0.0972
Giving to Penn 3 or More Years out of last 5 (1/0)	0.0118	0.0774
Fraternity/Sorority (1/0)	0.0069	0.0422
Alumni Spouse (1/0)	0.0044	0.0234
Penn Lifetime Giving Flag (1/0)	0.0014	0.022
Intercept	-0.002	-0.011

Once the model is run with the new sample, the equation can then be applied to the hold-out sample in order to have a model score for every record in the population.

Modeling Tools - Multiple Regression

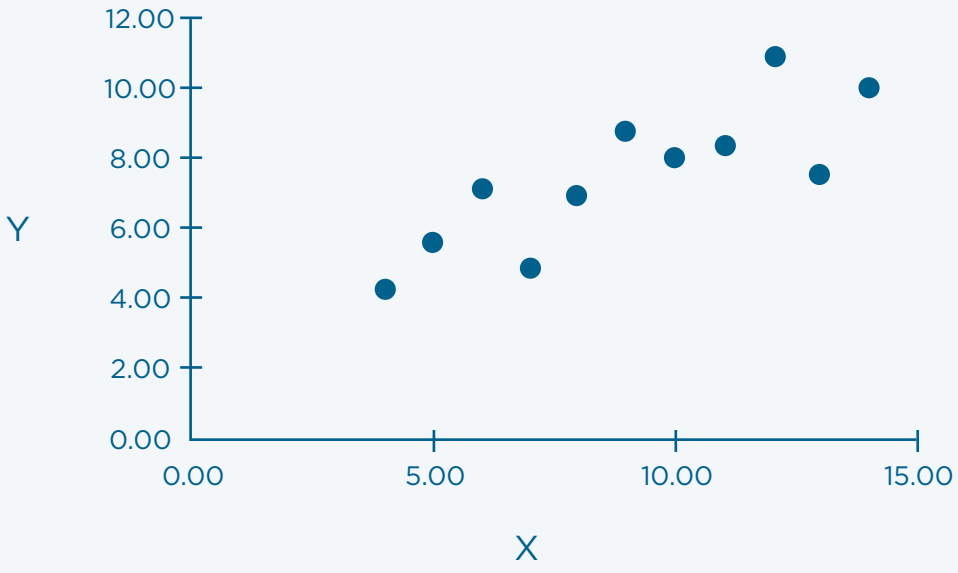
Regression analysis encompasses a number of techniques for analysis and predictive modeling that focus on the relationship between a dependent variable and one or more independent variables. Regression has many variations and goes by many names including linear regression, simple regression, ordinary least squares, and multiple linear regression. The regression equation developed in regression analysis relates the independent or explanatory variables to the dependent or target variable.

The following example, taken in part from a book by Edward B. Tufte (2), will help explain the methodology and the reason for the various similar names. We have a data set consisting of 11 observations where X is the independent variable and Y is the dependent variable (Table 6). We can think of the observations as people, records, or even subjects. The data is shown graphically in Figure 1.

Table 6. Sample X, Y Data

	1	2	3	4	5	6	7	8	9	10	11
X	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Figure 1. Sample X, Y Data



The goal of our regression analysis is to fit a straight line described by the equation:

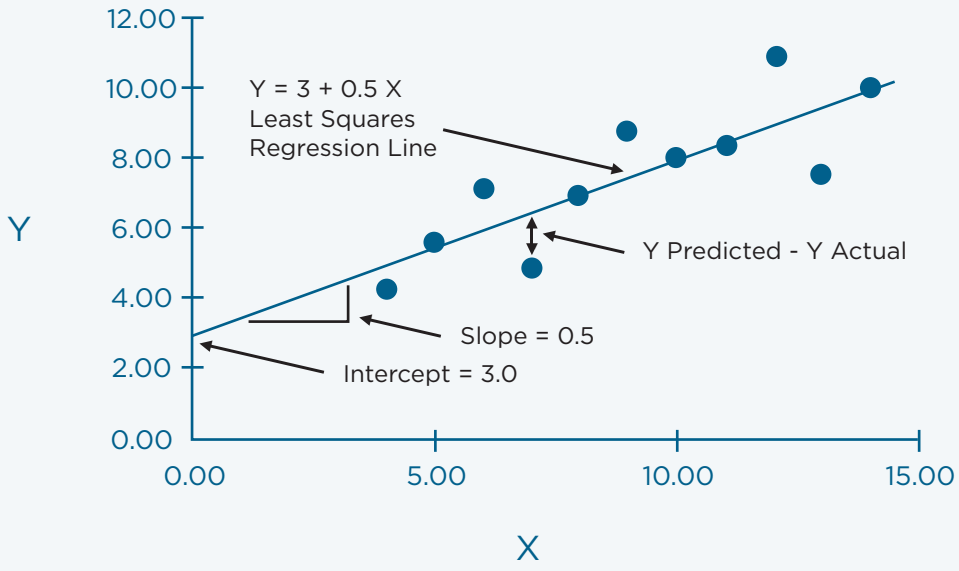
Y = A +B X

by "selecting" the values of A (the intercept) and B (the slope) so as to minimize the sum of the squared differences between Y-predicted and Y-actual. This analysis is called linear regression because we are fitting a straight line. It is simple regression because there are two variables and we can visualize the graphical relationship. It is sometimes called least squares because we seek to minimize the sum of the squared differences between Y-predicted and Y-actual.

The resulting least squares equation is

Y = 3 +0.5 X

Figure 2. Sample X, Y Data with Least Squares Regression Line



The regression statistics from MS Excel are:

Table 7. Regression Results for Sample X, Y Datasets

Regression Statistics	
Multiple R	0.816421
R Square	0.666542
Adjusted R Square	0.629492
Standard Error	1.236603
Observations	11

	Coefficients
Intercept	3.000091
X Variable 1	0.500091

The adjusted R Square, which can have values between -1 and +1, is a measure of goodness of fit of the least squares equation to the underlying data. Although the adjusted R Square in this case is quite good, the regression equation still does not "explain" all the variation in Y values. The R Square value is often used as the criterion to compare different models to select the "best" model. As with any methodology, there are some cautions with regression modeling and R Square that need consideration.

One important caution has been well illustrated by Edward B. Tufte (2) in the introduction to his book *The Visual Display of Quantitative Information*. In Table 8 there are four datasets, each with eleven X, Y observations. We have used Dataset 1 in the earlier description of simple linear regression. The adjusted R Square and least squares regression line are identical for each dataset (Table 9).

Table 8. Four X, Y Datasets

Dataset 1	
X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

Dataset 2	
X	Y
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

Dataset 3	
X	Y
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

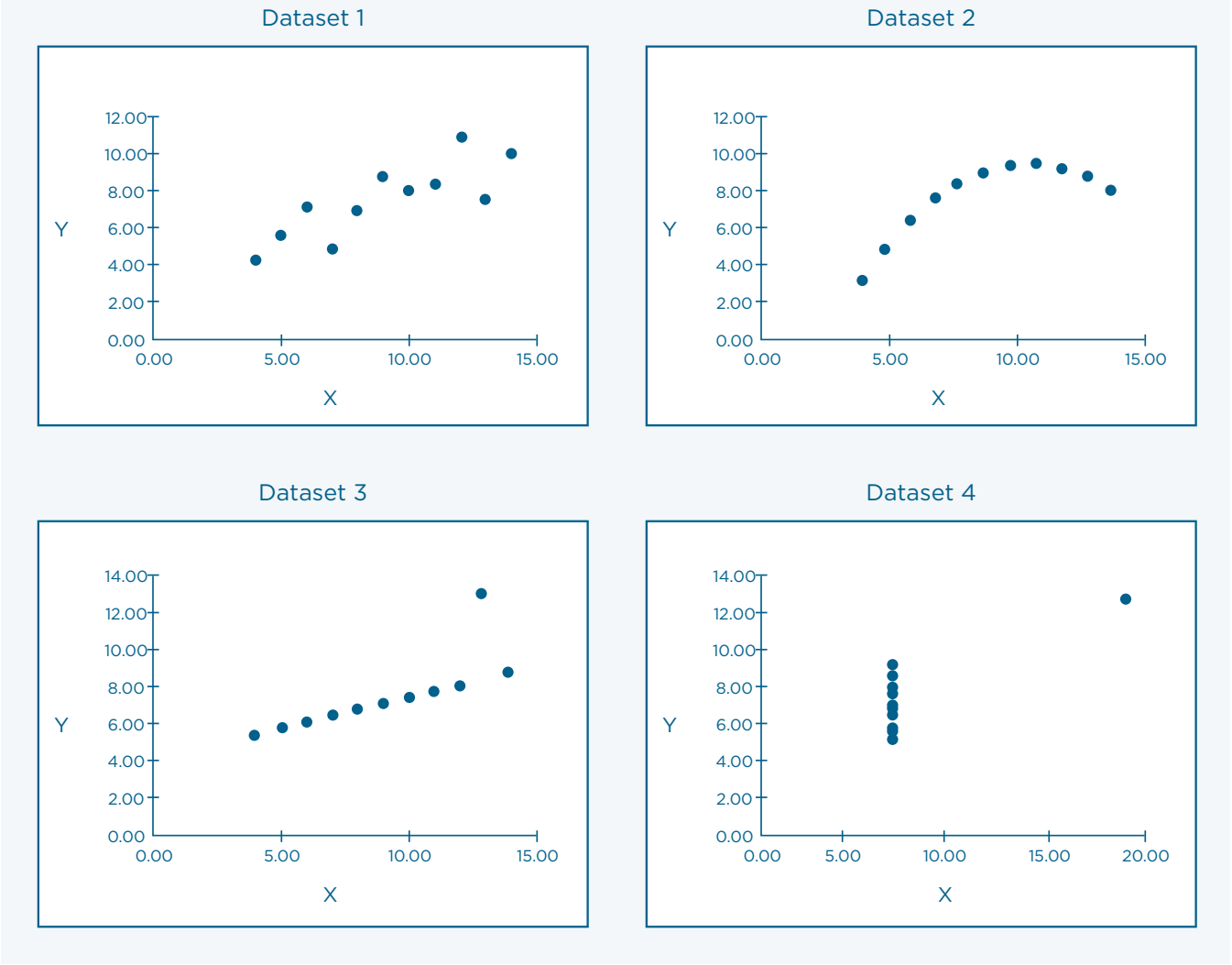
Dataset 4	
X	Y
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

Table 9. Regression Results for Four X, Y Datasets

Regression Statistics	
Multiple R	0.816421
R Square	0.666542
Adjusted R Square	0.629492
Standard Error	1.236603
Observations	11

Coefficients	
Intercept	3.000091
X Variable 1	0.500091

Figure 1. Sample X, Y Data



These four datasets yield the same regression equation and R Square, but are clearly different when viewed graphically, illustrating the need to understand your data. We can visualize data graphically in two or even three dimensions, but this gets difficult when we have more variables.

Multiple regression requires that each individual (observation, record, or subject) have values for all independent variables and the dependent variable. Otherwise the record must be discarded or some method for dealing with missing data must be applied.

Another difficulty that arises in regression analysis is when the independent variables are themselves correlated. This correlation of independent variables, called co-linearity, if not corrected, can produce models that are unstable. Calculation and examination of the correlation matrix with all variables under consideration can help identify highly correlated independent variables and some can be eliminated from the model.

In the section on Data Conditioning, expanding the donor consistency variable into a series of binary variables as shown below was discussed.

Table 10. Using Recoded Donor Consistency Variables

Value of original variable	Variable	Recoded variable
15 years of giving out of 15 years	1	Giving 15 out of 15 (1/0)
11–14 years of giving out of 15 years	2	Giving 11–14 out of 15 (1/0)
6–10 years of giving out of 15 years	3	Giving 6–10 out of 15 (1/0)
1–5 years of giving out of 15 years	4	Giving 1–5 out of 15 (1/0)
No years of giving out of 15 years	5	Giving None out of 15 (1/0)

If we included all five of these variables we would have an unstable model because collectively, these variables are correlated with each other. If for example, an individual has no giving in fifteen years, then the first four variables are always 0 and the fifth variable is always 1. One variable must be excluded (usually variable five) in order to avoid the problem of "over-specification" and the resulting co-linearity.

Modeling Tools - Decision Trees and CHAID

A decision tree is a data mining structure used to divide a large number of records or observations into smaller and smaller subsets. With each division, members of a subset become more similar to each other and less similar to members in other subsets. Decision trees are used for classification purposes, but can also be used for individual record-scoring purposes. The output is highly visual and the result includes a set of decision rules which can be applied to score and group records.

CHAID (Chi-Square Automatic Interaction Detector) is a decision tree analytical technique based on the Chi-Square distribution we mentioned earlier in connection with cross-tabs or contingency tables. Essentially the goal is to detect relationships between the dependent or target variable and independent, explanatory variables. The Chi-Square statistic measures the degree of relationship between the row and column variables in a contingency table. Larger values of Chi-square indicate a stronger relationship between the two variables. If Chi-square is zero, the two variables are statistically independent.

Decision trees and CHAID use categorical variables and, therefore, work especially well with binary variables. For continuous variables, the data is split into ranges, a technique called "discretizing" or binning. Most CHAID algorithms will split data at a point so as to create the largest differences between groups. If one is dissatisfied with the CHAID binning, the analyst can make their own binning choices.

One key advantage of decision trees is the ability to deal with missing data as another category value for a category variable. A disadvantage is the large sample sizes that can be required to produce a stable CHAID model.

To demonstrate the use of a CHAID model, let's try to predict giving in 2010 of alumni with a Penn undergraduate degree in a year ending in a 0 or a 5. The example is based on one of the models for "Targeting Undergraduate Non-Reunion Year Donors" discussed in detail later. Using this dataset and the multiple regression results enables us to compare multiple regression with CHAID. For our discussion purposes, we restrict the variables entering the CHAID model to those that were statistically significant in the multiple regression. In actual application we would start with a larger number of variables and let the CHAID analysis determine which ones are "predictive."

Figure 4 shows the variables from the multiple regression equation used in our CHAID example. The coefficients are the multipliers applied to the independent variables to determine the model score. The higher the score, the greater the likelihood that individual will be a non-reunion year donor. Furthermore, since all the independent variables have the value 1 or 0, variables with larger coefficients contribute more to the model score. To state it a little differently, variables with large coefficients are more important for predicting donor participation. The chart below the table of variables shows the regression model in equation form.

Throughout this monograph, larger coefficients are associated with variables that are more important and more predictive of the dependent variable for the behavior we are trying to predict.

Figure 4. Multiple Regression Variable Definitions and Equation from Targeting Undergraduate Non-Reunion Year Donorswith a Penn degree in a year ending in a 0 or a 5

Variable Name	Variable Definition	Model Component	Coefficient
YN_giving_2010	Gave to Penn in 2010 (Yes/No or 1/0)	Dependent variable	
YN_giving_2009	Gave to Penn in 2009 (Yes/No or 1/0)	Independent variable	0.4123
YN_giving_2006_2008	Gave to Penn in 2006–2008 (Yes/No or 1/0)	Independent variable	0.1905
YN_giving_2005	Gave to Penn in 2005 (Yes/No or 1/0)	Independent variable	0.1336
Alumni_SCPSO	Have an Alumni Spouse, Child, Parent, Sibling, or Other (Yes/No or 1/0)	Independent variable	0.03129
AR-AW-HC	Attended AR Alumni Weekend or Homecoming (Yes/No or 1/0)	Independent variable	0.04393
Other DAR Events	Attended Any Other DAR Events (Yes/No or 1/0)	Independent variable	0.2769
01_IC_Sports	Played Intercollegiate Sport(s) (Yes/No or 1/0)	Independent variable	0.02039
03_Stu_Performing_Arts	Student Performing Arts Participant (Yes/No or 1/0)	Independent variable	0.03088
04_Honor_Society	Honor Society Member (Yes/No or 1/0)	Independent variable	0.02298
		Intercept	0.06979

Multiple Regression Model for Non-Reunion 0's and 5's

YN_giving_2010=

YN_giving_2009* 0.4123+
YN_giving_2006_2008* 0.1905+
YN_giving_2005* 0.1336+
Alumni_SCPSO* 0.03129+
AR-AW-HC* 0.04393+
Other DAR Events* 0.2769+
01_IC_Sports* 0.02039+
03_Stu_Performing_Arts* 0.03088+
04_Honor_Society* 0.02298+
0.06979

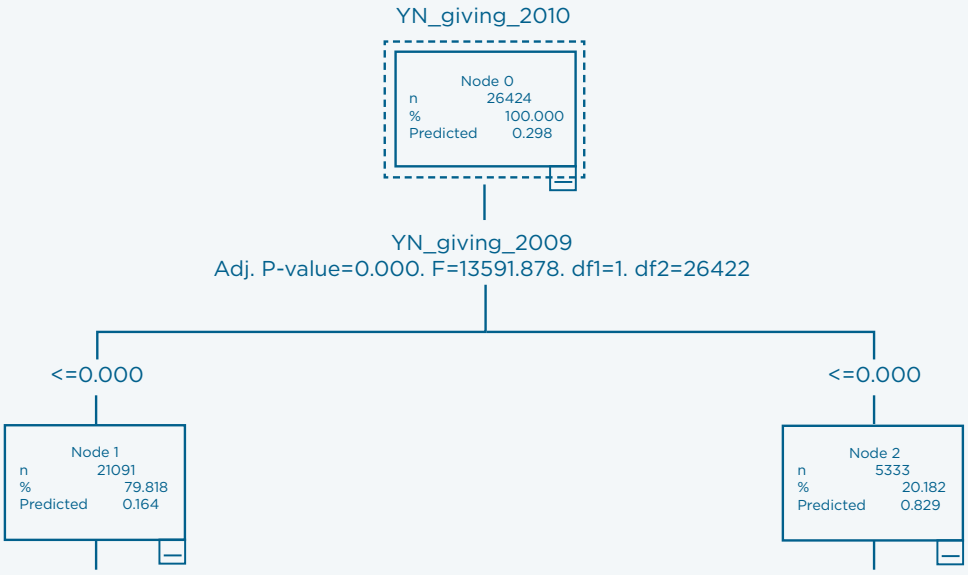
The data records and variables that produced the multiple regression results shown in Figure 4 were used as input to the CHAID modeling algorithm. Figure 5 shows the entire decision tree generated by the CHAID algorithm. For readability Figure 6 shows the first branch node details of the decision tree.

Figure 5. CHAID Decision Tree for Non-Reunion 0's and 5's

CHAID Decision Tree Model for Non-Reunion O's and 5's

Figure 6. CHAID Decision Tree for Non-Reunion 0's and 5's

CHAID Decisions Tree Model for Non-Reunion O's and 5's - First Branch



At Node 0 in Figure 6, the dataset contains 26,424 alumni with an undergraduate degree in a year ending in 0 or 5. Furthermore 29.8% (7,883) of these alumni were donors in FY2010. In the CHAID analysis, cross-tabulations each of the nine independent variables vs. the dependent variable (giving in FY2010) are developed.

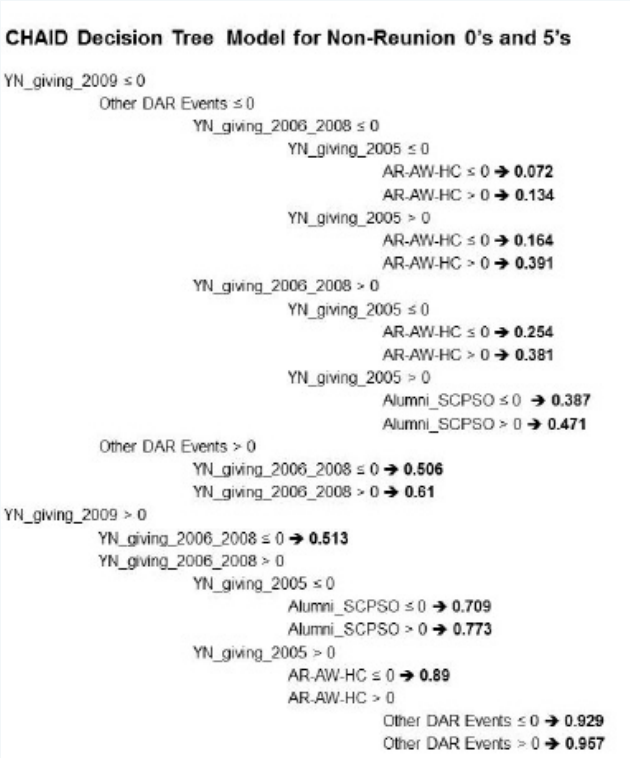
In this example, the variable associated with giving in FY2009 most effectively separates the total population into two groups, one with a larger number of donors and one with a small number of donors. Specifically using giving in FY2009, at Node 1 we have a population of 21,091 alumni, 16.4% (3,464) of whom are FY2010 donors. At Node 2 we have a population of 5,333 alumni, 82.9% (4,419) of whom are FY2010 donors. The following chart summarizes these results.

	Total	Not a Donor in 2009	Donor in 2009
Not a Donor in 2010	18,541	17,627	914
Donor in 2010	7,883	3,464	4,419
Total	26,424	21,091	5,333

Using only one independent variable, giving in FY2009, we can focus on 20.2% of the alumni (5,333 out of 26,424) and isolate 58% of the FY2010 donors (4,419 out of 7,883).

This process of splitting the population at each node continues until splitting yields no improvement in separation or all the variables have been examined. Figure 7 shows the entire decision tree including the percent of alumni at the end of each branch who were donors in FY2010.

Figure 7. CHAID Decision Model Rules for Non-Reunion 0's and 5's



In the bottom branch in Figure 7, 95.7% of the alumni in that branch were FY2010 donors. The following chart summarizes the results for this branch.

	Total	Not a Donor in 2009	Donor in 2009
YN_giving_2009 =1 and YN_giving_2006_2008 =1 and YN_giving_2005 =1 and AR-AW-HC =1 and Other DAR Events =1.	469	20	449 (95.7%)

Some other features typical of a decision tree output are also shown in Figure 7. Only five of the original nine independent variables are used in the decision tree. Not all branches have the same length (number of variables used). The same decision can appear in several places in the tree and variables do not necessarily appear in the same order in all branches.

In general, at Penn we have opted for multiple regression models instead of CHAID because it has been easier to implement in our DAR operational database.

Modeling Tools – Neural Nets

Neural networks are a class of generalized tools for prediction that have been applied to a wide range of problems. Problems most amenable to neural net modeling are those in which:

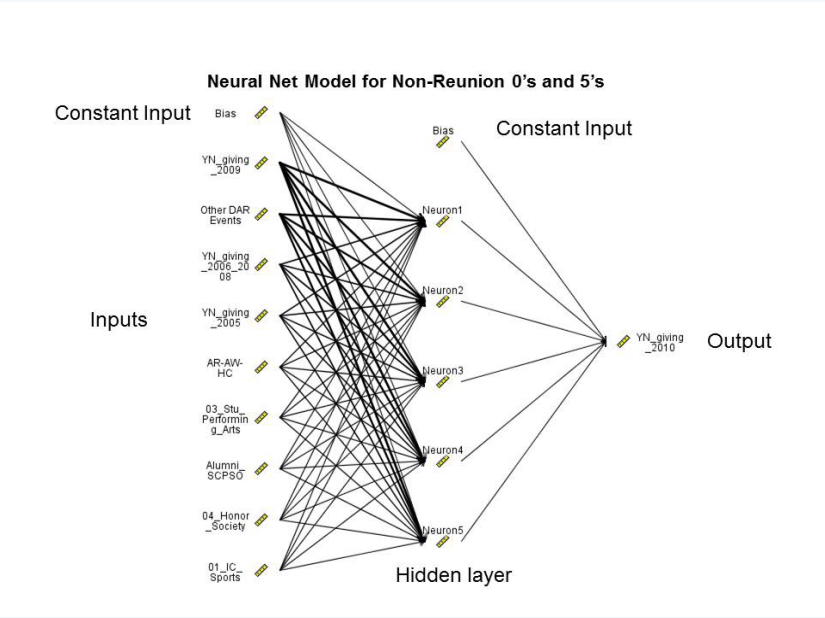
- The inputs are well understood,
- The output is well understood,
- Experience is available, but
- How the inputs are combined to produce an output is not as well understood.

Berry and Linoff (1) have an excellent example of the application of neural nets in determining real estate values based on the particulars of each property. Real estate agents and real estate appraisers are examples of human experts who perform such valuation tasks very well. In describing neural nets they determine the extent to which a neural net can be "trained" to be an expert real estate appraiser.

To demonstrate the use of neural net modeling, we will take the same approach used in the preceding CHAID example (See Figure 4). Again we will use variables and data from a multiple regression analysis developed to predict giving in 2010 of alumni with a Penn undergraduate degree in a year ending in a 0 or a 5.

The neural net generated from SPSS Modeler is shown in Figure 8. Here we see the familiar input and output variables and two new features, the hidden layer and bias. Unlike multiple regression where MS Excel can be used at least for models with fifteen or fewer variables, to use CHAID or neural nets requires a more sophisticated software packages like SPSS or SAS.

Figure 8. Neural Net Model for Non-Reunion 0's and 5's

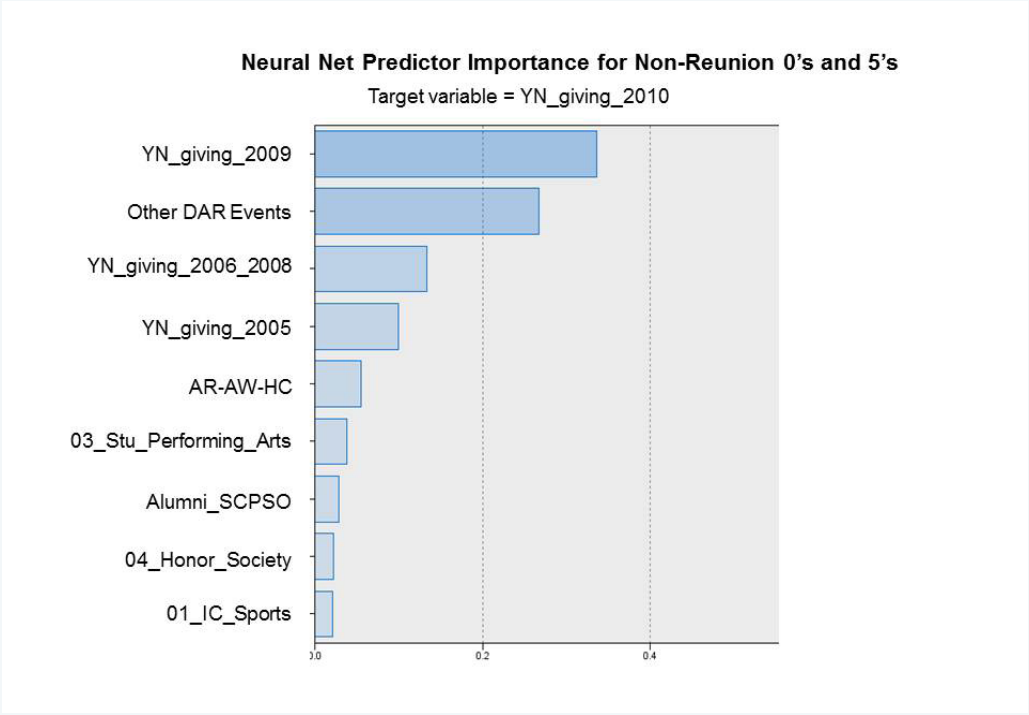


The middle or hidden layer is added to enable the network to recognize more patterns. Bias allows for an offset by a constant amount. In this example, SPSS Modeler added five nodes or neurons in the hidden layer. Each input node is connected to each neuron in the hidden layer and each input-neuron connection has a weight (sometimes called a transform). Likewise, each neuron is connected to the output and has a weight. The input values for a particular observation (record) multiplied by the weights yield a numeric value at each neuron. The values at each neuron multiplied by the weights between the neurons and the output node produce a predicted value of the output. The bias is a constant that provides for a global offset. The SPSS Neural Net algorithm has two offsets—one affecting the node values and the other affecting the output values.

Initially the weights used in the model are only "guesses" in the range of 0 to 1. Each iteration that passes the training set of records through the network produces a new output that depends on the input values, the weights, and the bias values. The network weights and biases are adjusted in a way that improves the correlation between the predicted values of the output and the actual values of the output. When the input weights and biases from the current iteration do not improve the correlation of predicted and actual output variables, by some pre-specified amount, we have our "optimal" neural net solution.

Figure 9 shows the relative importance of the input variables and predictors of the target variable. These can be used as multiplier weights in a linear equation that will function similar to a regression equation. As with CHAID, the order of variable importance in this model is different from the multiple regression results with the same input variables.

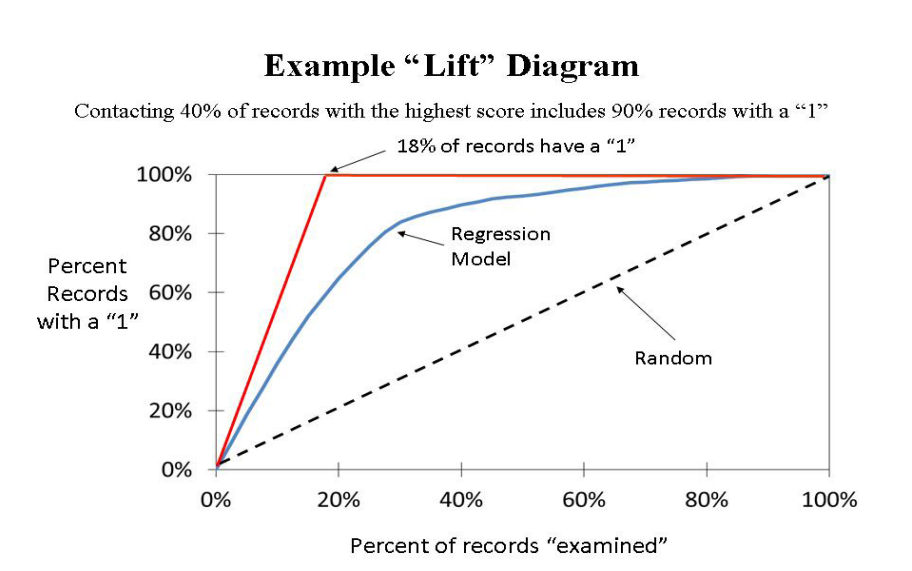
Figure 9. Relative Importance of Input Variable in the Neural Net Model



The Lift Diagram – Characterizing Predictive Capability

The discussion of multiple regression described the use of R Square as a measure of goodness of fit of a model. Another quite different but useful construct for assessing the predictive ability of a model with a binary dependent variable (1/0) is the lift diagram (see Figure 10). Model scores (the predicted value of the dependent variable) are sorted highest to lowest. The x-axis of the lift diagram is the cumulative percent of the records or observations with that model score or higher. The y-axis is the cumulative percent of records with a target variable value of 1 with that model score or higher.

Figure 10. Example Lift Diagram



A model which has no predictive ability would require, for example, that 50% of the records be examined to find 50% of the 1's. The graphical representation of this "random" model is the 45 degree line marked "random" and connecting (x=0, y=0) and (x=100 %, y=100 %)

For our example, the actual percent of dependent variable observations equal to 1 is 18%. The "perfect" model would require that only 18% of the population be examined to find 100% of the observations with a dependent variable value equal to 1. The graph of the "perfect" model is a line connecting (x=0, y=0) and (x=18%, y=100 %) and a line connecting (x=18%, y=100%) and (x=100%, y=100%).

A good predictive regression model will more closely resemble the "perfect" model than the "random" model. Here the "Regression Model" curve shows that with the 20% highest model scores we reach 65% of those with a dependent variable value of 1. With the top 40% of the scores we reach 90%.

In addition to showing the predictive ability of the model, the lift diagram helps in operational implementation by helping set the "optimal" model score cut-off.

Strategies for Dealing with Missing Data

Missing data can cause difficulties with some of the techniques used for predictive modeling. Multiple regression, for example, requires that every independent variable and the dependent variable have values. If data is missing, the records must be removed from the dataset or values must be added where necessary.

Like most development operations we have very complete and accurate giving data for constituents. When creating giving variables, we usually include both hard and soft dollars. Most often we will include pledges (commitments) and gifts (receipts), since we are trying to create variables that indicate the presence or absence of giving transaction(s) for an individual. For binary variables about giving, missing data, therefore, is not a problem for us.

Nearly all of the models described in this monograph are binary models (0's and 1's) and missing data is easier to deal with than in models with continuous variables. For demographic variables like "married to an alumna or alumnus," we assume that if the data is not in our development database, the individual does not have an alumni spouse.

We continually try to improve the accuracy of missing relationships, student activities, sports participation, and other demographics with alumni surveys and a robust online community database where alumni can update their information. Despite these efforts when data is missing we treat it as if that demographic attribute does not exist for that individual. Of course, this negatively impacts the correlation of independent variables with the dependent variable.

With continuous demographic variables we use other approaches. Birth date and age data is only 80% complete in our database. Year of first or last Penn degree is available for all alumni and often used as a surrogate for age (e.g. years since first Penn degree). Unfortunately for alumni from the 1940's we know WWII disrupted or delayed college for many alumni. This reinforces how important it is to "know your data."

Dealing with missing data requires serious consideration to balance the practical limits of available data and predictive modeling methods. Fortunately, unlike multiple regression, methods like neural nets and CHAID allow for individual records with missing values for some of the independent variables.

TARGETING ANNUAL GIVING DONORS

The more than twenty Annual Giving programs at Penn raised over $60 million in the fiscal year ending June 30, 2011. In that year, Annual Giving programs accounted for approximately 14% of the total support raised. The largest Annual Giving program is the Penn Fund, which targets undergraduate alumni, and raised over $30 million in FY2011. Second largest at just over $9 million is the Wharton School of the University of Pennsylvania Annual Giving program which focuses on MBA alumni.

Both of these Annual Giving programs have a very strong, but not exclusive, emphasis on quinquennial (every five years) reunion classes. Alumni Relations programming in general and especially in support of class reunions is extensive. Alumni Weekend at Penn is held in mid-May on the Friday, Saturday, and Sunday before commencement which is on Monday. Not including the commencement ceremonies, each year more than 10,000 alumni return to campus for Alumni Weekend.

Alumni Relations, the Penn Fund, and Programs and Special Events staffs work much of the year planning for Alumni Weekend. Each reunion class also has Reunion Planning and Gift Committees staffed by alumni volunteers who are recruited at least nine months before Alumni Weekend.

Four different modeling projects that focus on annual giving donors discussed here are:

- Targeting Penn Fund Undergraduate Reunion Year Donors
- Targeting Wharton MBA Reunion Year Donors
- Targeting Penn Fund Undergraduate Non-Reunion Year Donors
- Targeting Vet Hospital Client Annual Giving Program Donors

Successful predictive modeling that target donors requires a history of giving data. At Penn, giving transaction history is available beginning in 1978. Predictive modeling variables used for undergraduate and graduate alumni are different, mostly in terms of the number of non-giving variables available. Undergraduate alumni have a large number of student activities in which they could have participated, including intercollegiate sports, fraternities/sororities, performing arts, service groups, senior societies, and academic honor societies. Predictive modeling for graduate alumni may be more difficult unless other non-giving variables are available.

Several annual giving programs at Penn target constituencies other than Penn alumni. Annual giving programs of the Hospital of the University of Pennsylvania and affiliates and the Veterinary Hospitals are a few examples. Our fourth project demonstrates how donor targeting in these situations is different than for Alumni Annual Giving programs.

The Ryan Veterinary Hospital for small animals provides veterinary care and support for clients. Important variables include the frequency of hospital visits, specific services delivered, the customer experience the client receives, and the outcome of the visit.

Targeting Undergraduate Reunion Year Donors

There are more than 130,000 undergraduate alumni of record at Penn. Each year more than 25,000 undergraduate alumni celebrate a quinquennial class reunion. These undergraduate class reunions are important milestone events, providing the opportunity for alumni to reconnect with Penn and with each other. Although all alumni are welcomed and encouraged to participate in Alumni Weekend activities, reunion classes receive special attention with respect to programming and gift solicitation. A wide range of resources involving both staff and volunteers are used to contact and solicit reunion class alumni.

Discussions with Penn Fund staff in January 2009 identified a need to determine which reunion class alumni were most likely to make a reunion gift to Penn. There are several reasons that a predictive model would be useful, including:

- Annual giving at Penn has a strong reunion focus, and alumni are much more predisposed to give in reunion years in terms of both donor participation and average gift size.
- Both donor participation and total giving by class are important metrics for the Penn Fund staff.
- Reunion gift solicitation by email, direct mail, phone, volunteers, and staff requires commitments of scarce resources. With the possible exception of email solicitation, costs increase with the number of alumni solicited.
- The ability to predict who will be a donor can focus resources and perhaps reduce costs with little to no loss in amounts raised.

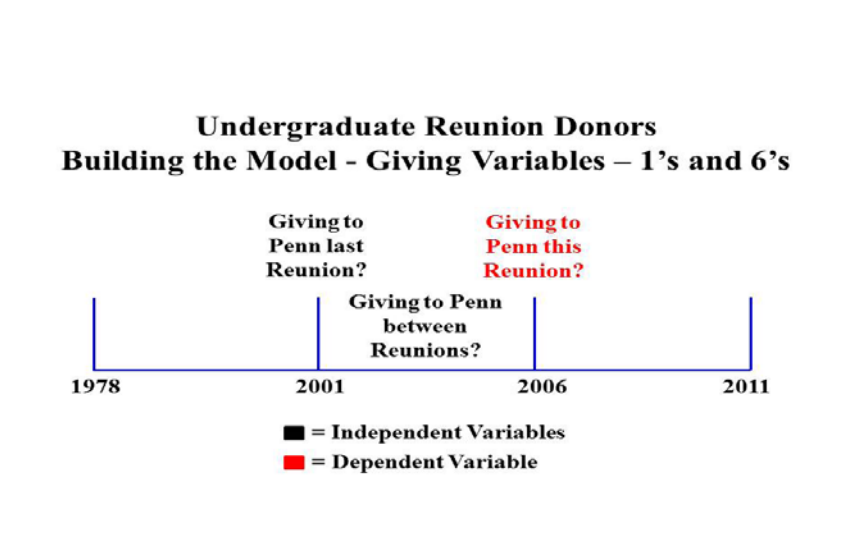
The Objective

The objective of this predictive modeling project was to develop a scoring model that predicts donor participation in a reunion year. A model score is produced for each alumna or alumnus; the higher the score, the higher the likelihood that they will be a donor.

The Process

A data set with 119,194 records for alumni with degrees before 1999 was created, cleaned, and conditioned. Everyone in the data set had at least two reunions in which they could have participated. The class of 1996, for example, had reunions in 2001 and 2006. Figure 11 shows the model time frame for reunion classes ending in 1 or 6.

Figure 11. Model Time Frame for Reunion Class Years Ending in 1's and 6's



Variables selected for consideration included:

- Giving variables
- Relationships
- Reunion participation
- Other event participation
- Academic variables
- Involvements as a student
- Involvements as an alumna/alumnus
- Penn interests

Data Exploration with Crosstabs

Before developing and testing the multiple regression models, data exploration was conducted using pair wise cross-tabulation. These crosstab results gave us great confidence that we would later get useful results from regression modeling.

The dependent variable (what we are trying to predict) in the model is an individual's donor participation at their last (current) reunion. It turned out that the most powerful independent variable (most predictive) in the model was whether the individual gave between the last two reunions. The crosstab in Table 11 shows the relationship between giving in the year of their last reunion and giving between the last two reunions. Only 7% of alumni who did not give between reunions gave in the year of their last reunion vs. 65% of those who gave between reunions.

Table 11. Giving at Last Reunion vs. Giving Between Reunions

	Did not give between last two reunions	Gave between last two reunions
Gave last reunion year	5,576	24,875
Did not give last reunion year	75,350	13,393
Total	80,926	38,268

	Did not give between last two reunions	Gave between last two reunions
Gave last reunion year	7%	65%
Did not give last reunion year	93%	35%
Total	100%	100%

The crosstab in Table 12 shows the relationship between giving in the year of their last reunion and attending the last reunion. Only 22% of alumni who did not attend the last reunion gave in their last reunion year vs. 68% of those who attended their last reunion.

Table 12. Attendance at last Reunion

	Did not attend last reunion	Attended last reunion
Gave last reunion year	23,606	6,845
Did not give last reunion year	85,462	3,281
Total	109,068	10,126

	Did not attend last reunion	Attended last reunion
Gave last reunion year	22%	68%
Did not give last reunion year	78%	32%
Total	100%	100%

Unfortunately, this strongly correlated independent variable is not useful for predictive modeling, because attendance at the last (current) reunion is not known until the end of May after the reunion and most of the giving for that year has taken already place.

Crosstabs with other independent variables are summarized in Tables 13 through 22.

Table 13. Giving Reunion Year Prior to last Reunion

	Did not give in reunion year before last reunion year	Gave in reunion year before last reunion year
Gave last reunion year	14%	70%
Did not give last reunion year	86%	30%
Base	93,627	25,567

Table 14. Attendance at Reunion Prior to last Reunion

	Did not attend reunion before last reunion	Attended reunion before last reunion
Gave last reunion year	23%	60%
Did not give last reunion year	77%	40%
Base	110,383	8,811

Table 15. Penn Children

	No Penn Children	Penn Children
Gave last reunion year	23%	55%
Did not give last reunion year	77%	45%
Base	110,763	8,431

Table 16. Penn Spouse

	No Penn Spouse	Penn Spouse
Gave last reunion year	23%	48%
Did not give last reunion year	77%	52%
Base	107,025	12,169

Table 17. Penn Siblings

	No Penn Siblings	Penn Siblings
Gave last reunion year	24%	42%
Did not give last reunion year	76%	58%
Base	107,964	11,230

Table 18. Penn Parents

	No Penn parents	Penn parents
Gave last reunion year	24%	38%
Did not give last reunion year	76%	62%
Base	109,443	9,751

Table 19. Penn Graduate/Professional Degree

	Do not have Penn grad/prof degree	Have Penn grad/prof degree
Gave last reunion year	24%	38%
Did not give last reunion year	76%	62%
Base	105,315	13,879

Table 20. Sports

	Did not play sport(s)	Played sport(s)
Gave last reunion year	24%	40%
Did not give last reunion year	76%	60%
Base	105,990	13,204

Table 21. Fraternity or Sorority

	Not a fraternity or sorority member	Fraternity or sorority member
Gave last reunion year	22%	36%
Did not give last reunion year	78%	64%
Base	90,491	28,703

Table 22. Other Student Activities

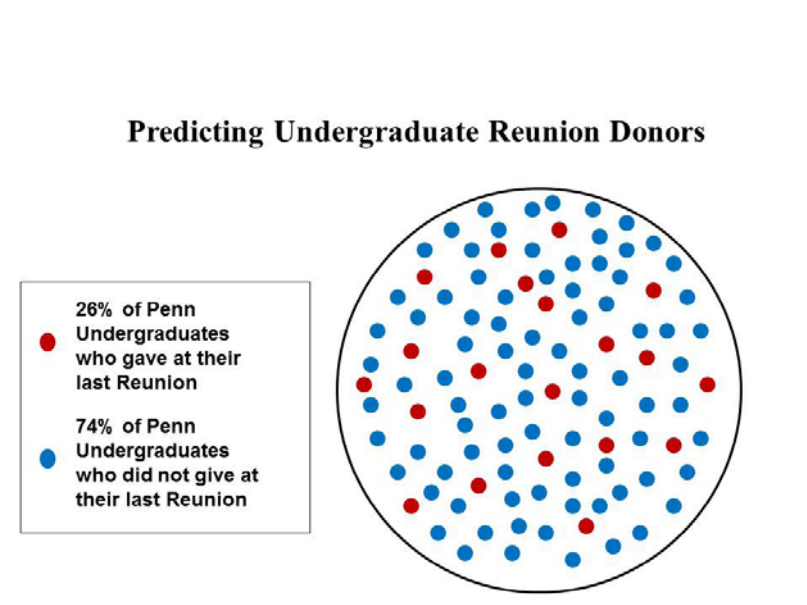
	No other student activities	Had other student activities
Gave last reunion year	24%	35%
Did not give last reunion year	76%	65%
Base	106,754	12,440

Multiple Regression Scoring Model

The use of multiple regression to develop a model score for each alumna or alumnus may best be understood graphically before looking at the resulting final equation.

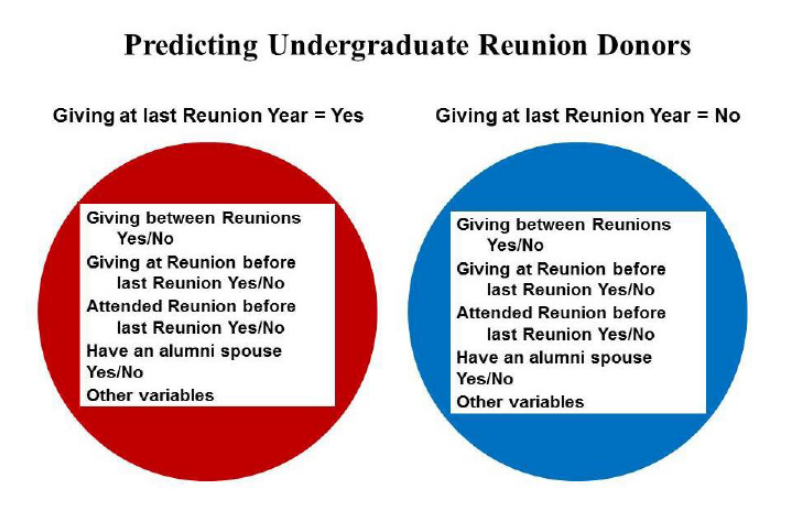
In Figure 12, the 26% of the 119,194 Penn undergraduates who gave during their last reunion year are represented by red dots and the 74% who did not give during their last reunion year are represented by blue dots. Unfortunately, without a scoring model, we have no precise way of separating the 119,194 dots into two subsets, one rich in red dots and the other deficient in red dots.

Figure 12. Graphical Representation of Penn Reunion Donors and Non-Donors



With but one difference, each dot has the same characteristics. The red dots represent donors to Penn at their last reunion and the blue dots represent Penn alumni who did not make a gift during the year of their last reunion (See Figure 13).

Figure 13. Graphical Representation of Individual Penn Alumni



The multiple regression model results from Excel 2007 are shown below. The adjusted multiple R Square of 0.439 is quite good for this type of model. The multiple R Square is a measure of goodness of fit of the regression equation to the data. Later we will discuss the lift diagram and associated measures that are perhaps better indicators of the predictive power of scoring models.

The variables in the multiple regression have been sorted in order of the coefficient values, largest to smallest. Because all the variables have binary values of 0 or 1, the coefficients represent weights that indicate the importance of the variable to each individual's model score. Variables with larger coefficients or weights are more important for predicting undergraduate alumni donor participation in a reunion year.

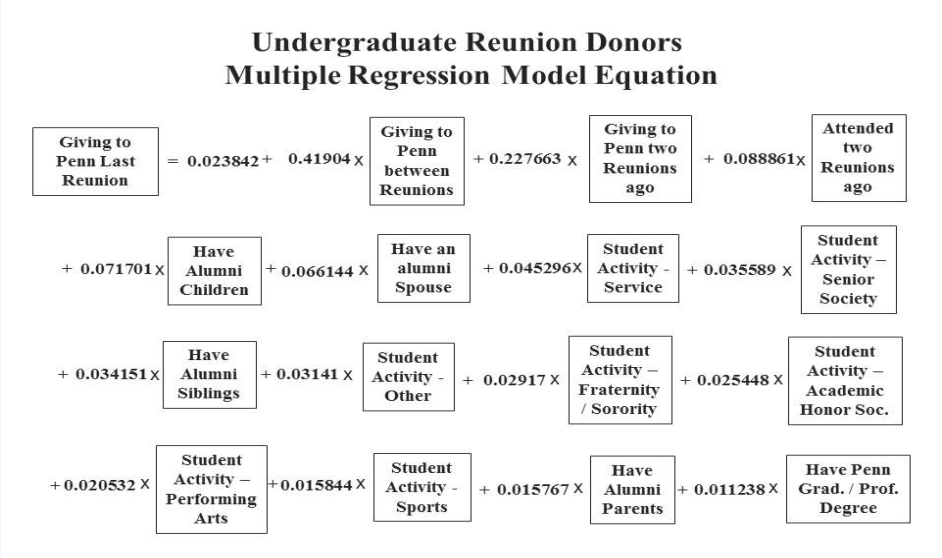
Table 23. Multiple Regression Model Results

Regression Statistics	
Multiple R	0.662991756
R Square	0.439558069
Adjusted R Square	0.43948753
Standard Error	0.32651858
Observations	119194

	Coefficients
Intercept	0.023842
Giving to Penn between reunions (1/0)	0.419040
Giving to Penn two reunions ago (1/0)	0.227663
Attended two reunions ago (1/0)	0.088861
Have Alumni Children (1/0)	0.071701
Have Alumni Spouse (1/0)	0.066144
Have Student Activity - Service (1/0)	0.045296
Have Student Activity - Senior Honor Society (1/0)	0.035589
Have Alumni Siblings (1/0)	0.034151
Have Student Activities – Other (1/0)	0.031410
Have Student Activity - Fraternity or Sorority (1/0)	0.029170
Have Student Activity - Academic Honor Society (1/0)	0.025448
Have Student Activity - Performing Arts (1/0)	0.020532
Have Student Activity - Sports (1/0)	0.015844
Have Alumni Parents (1/0)	0.015767
Have Penn Graduate / Professional Degree (1/0)	0.011238

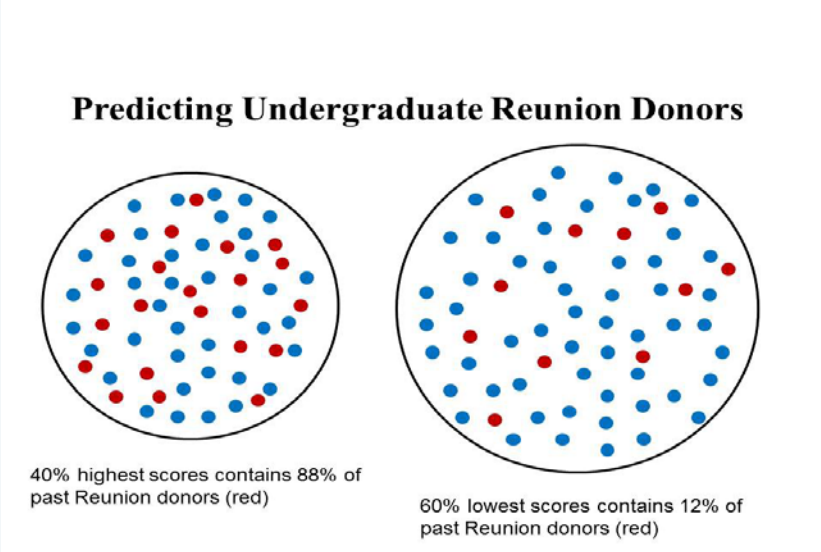
Graphically, the resulting model equation is shown in Figure 14. Note that the results are all based on a reunion year that has already taken place. Later we will shift the equation five years forward to predict who will be a donor at their next reunion.

Figure 14. Multiple Regression Model



Looking at the result in terms of red dots and blue dots, model scores enable us to separate the original 119,194 undergraduate alumni into a group of 40% with the highest model scores containing 88% of donors (red dots). The remaining 60% contains just 12% of the red dots (See Figure 15).

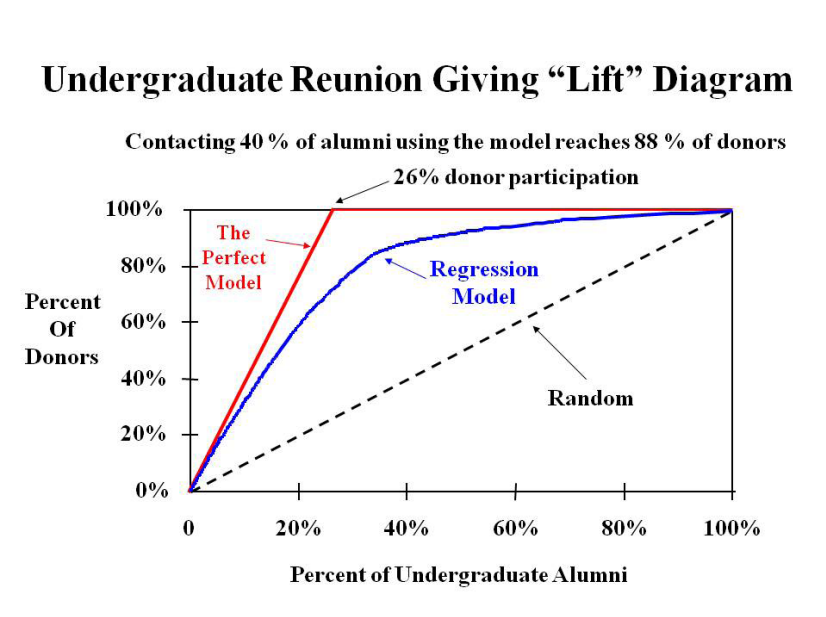
Figure 15. Undergraduate Alumni Segment with a High Concentration of Donors



The lift diagram for the undergraduate reunion giving model shown in Figure 16 characterizes the predictive ability of our scoring model differently than looking at the multiple R Square. Since we know from our data that the donor participation is 26%, the perfect model would be able to select only donors and exclude non-donors. Using the "random" model would require marketing to the entire population to get the 26% that are donors. A good predictive model more closely resembles the "perfect" model than the "random" model.

For the multiple regression model developed here, the highest 40% of scores contains 88% of the donors. The top 60% of scores includes 90+ % of donors. The lift diagram gives a practical measure of the model's predictive power.

Figure 16. Undergraduate Reunion Giving Model – Lift Diagram



Predicting Donors in Their Next Reunion Year

To predict who will be a donor at their next reunion, we shift the time frame of the equation five years forward to apply the multiple regression equation. Figure 17 shows the time shift for reunion classes ending in 1 or 6.

Figure 17. Application of Multiple Regression Equation in FY2011

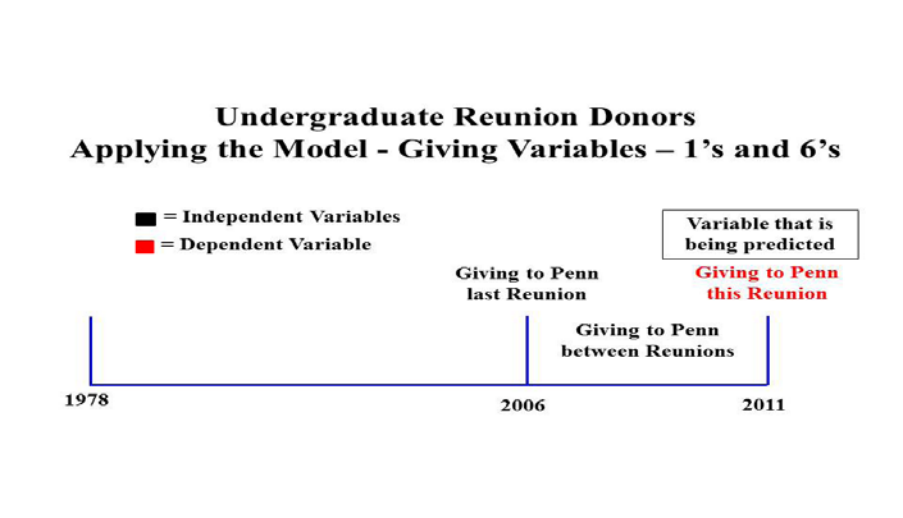
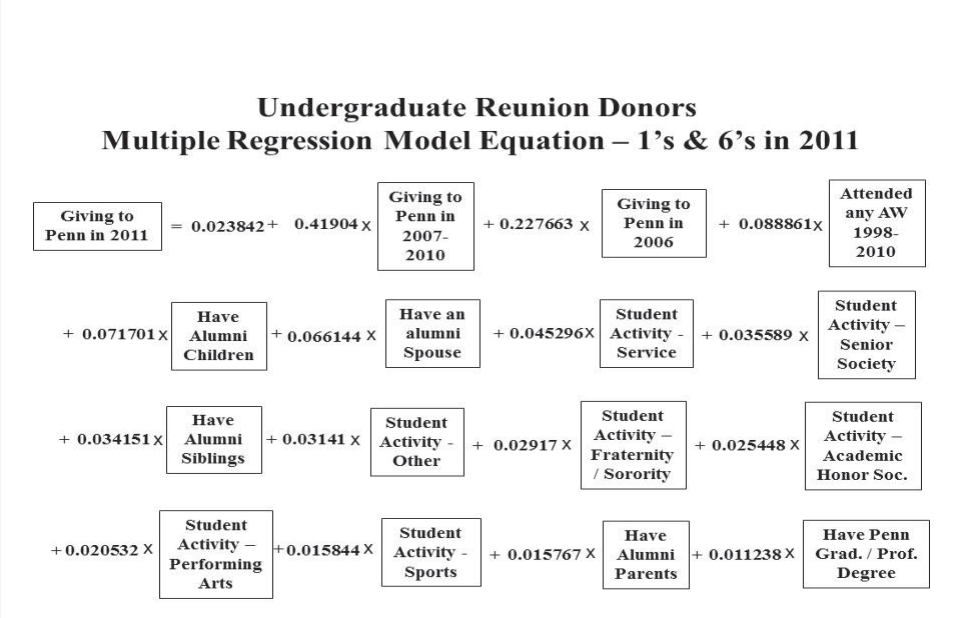


Figure 18 shows the multiple regression model equation for 1's and 6's in 2011. This model was used to calculate a score to predict the likelihood of being a donor in FY2011.

Figure 18. Multiple Regression Model Equation for 1's and 6's in 2011



Model Validation (July 2011)

When FY2011 ended on June 30, 2011, we determined which alumni in the classes ending in 1 or 6 were actually donors during the preceding year. Among these 28,113 alumni there were 6,709 donors for an overall donor participation rate of 24% (See Figure 19). Model scores were ranked and grouped into 10 "deciles." Because the "deciles" are uneven in terms of number of alumni, they might better be called quantiles. For the 476 alumni with the highest model scores, actual FY11 donor participation was 91%. As model scores decrease actual donor participation decreases.

Figure 19. Model Validation Results Using FY11 Actual Giving

Undergraduate Reunion Donors
Model Validation Using FY11 Actual Giving

Model "Decile"	UG 2011 Reunion Alumni	FY11 Actual Donors	FY11 Donor Participation
1	476	432	91%
2	1,371	1,117	81%
3	2,716	1,957	72%
4	1,774	1,073	60%
5	1,134	536	47%
6	1,826	667	37%
7	542	128	24%
8	1,314	188	14%
9	2,609	189	7%
10	14,351	422	3%
	28,113	6,709	24%

In Figure 20, we see that the top 40% of predictive model scores identified 91% of the actual FY11 donors.

Figure 20. Model Validation Results Using FY11 Actual Giving

Undergraduate Reunion Donors
Model Validation Using FY11 Actual Giving

Focusing on top 40% of UG reunion alumni would have reached 91% of FY11 donors

Model "Decile"	UG 2011 Reunion Alumni	Percent of UG 2011 Reunion Alumni	Cumulative Percent of UG 2011 Reunion Alumni	FY11 Actual Donors	Percent of Actual FY11 Donors	Cumulative Percent of Actual FY11 Donors
1	476	2%	2%	432	6%	6%
2	1,371	5%	7%	1,117	17%	23%
3	2,716	10%	16%	1,957	29%	52%
4	1,774	6%	23%	1,073	16%	68%
5	1,134	4%	27%	536	8%	76%
6	1,826	6%	33%	667	10%	86%
7	542	2%	35%	128	2%	88%
8	1,314	5%	40%	188	3%	91%
9	2,609	9%	49%	189	3%	94%
10	14,351	51%	100%	422	6%	100%
	28,113			6,709		

Summary of Results and Implementation Recommendations

The variables in this model most likely to predict giving to Penn in an undergraduate reunion year were associated with previous giving behavior:

- Giving to Penn between the last reunion year and the current reunion year, and
- Giving to Penn in the previous reunion year.

Once these two giving variables are included in the model, other giving variables like giving in the four years prior to the last reunion do not add to the predictive capacity of the model.

Alumni who attended their previous reunion are more likely to give in a reunion year than those who did not attend.

Penn alumni relationships were the next strongest predictors of giving in a reunion year, especially:

- Having alumni children, and
- Having an alumni spouse.

Less important were:

- Having alumni siblings, and
- Having alumni parents.

Also contributing to predicting giving in a reunion year was participation in student activities:

- Service organizations
- Senior honor societies
- Fraternity or sorority membership
- Academic honor societies
- Performing arts organizations
- Intercollegiate sports teams

Last, but still significant as a predictor was having a Penn graduate or professional degree in addition to an undergraduate degree.

The variables and the hierarchy of importance are not surprising and provide insight as to where to begin a predictive modeling project. To predict giving in a reunion year, focus on prior giving behavior, attendance at the previous reunion, family relationships with the institution, involvements as a student, and graduate Penn degree(s).

Although the predictive model results discussed here are generally intuitive, the value of the regression model is the equation which weights each of the variables to produce model score for each individual.

Finally, using predictive modeling in a reunion year uses existing data on giving at the last reunion as the dependent variables. When the model equation is applied to the current reunion year the time frame is advanced five years to calculate model scores.

Targeting Wharton MBA Reunion Year Donors

There are more than 39,000 living Penn alumni with an MBA degree from the Wharton School of the University of Pennsylvania. Each year, more than 7,500 Wharton MBA alumni celebrate a quinquennial class reunion. Because most of these Wharton alumni have undergraduate affiliations other than Penn, these class reunions are important ways for alumni to reconnect with Wharton and with each other.

Discussions with the Wharton director of Alumni Affairs & Annual Giving in March 2010 identified a strong interest in developing a predictive model to identify Wharton MBAs most likely to be donors in a reunion year. Reasons such a model would be useful are similar to the reasons mentioned earlier for undergraduate alumni:

- Donor participation and total giving by class are important metrics for Wharton Annual Giving.
- Reunion gift solicitation by email, direct mail, and phone, volunteers and staff uses scarce resources. Other than for email solicitation, costs increase with the number of alumni solicited.
- Focusing resources on likely donors may reduce costs with little to no loss in revenue.

The Objective

The objective of this predictive modeling project was to develop a scoring model that predicts donor participation of Wharton MBAs in a reunion year.

The Process

A data set with 28,771 records was created for Wharton MBA alumni with degrees before 2001. Everyone in the data set had at least two reunions in which they could have participated, but dependent variables are different for each reunion cohort.

Table 24. Model Dependent Variables by Cohort

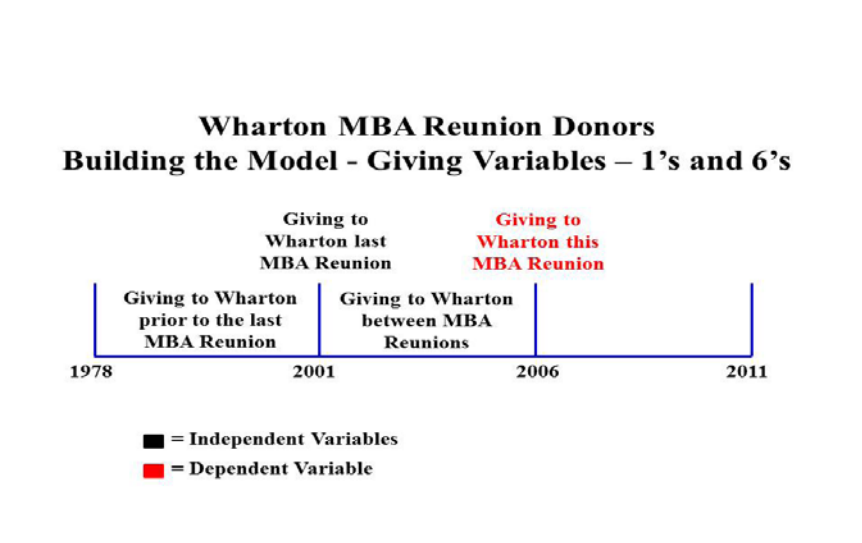
Reunion Cohort	Dependent Variable
0's and 5's	Giving to Wharton 2010 – Y/N
4's and 9's	Giving to Wharton 2009 – Y/N
3's and 8's	Giving to Wharton 2008 – Y/N
2's and 7's	Giving to Wharton 2007 – Y/N
1's and 6's	Giving to Wharton 2006 – Y/N

Variables selected for consideration included:

- Giving variables
- Relationships
- Academic variables
- Involvements as a student
- Involvements as an alumna/alumnus
- Penn interests
- Event participation (other than reunions)

Participation in Wharton MBA reunions would have been considered but has only been recorded systematically in the past few years. Figure 21 shows the model time frame for the reunion 1's and 6's cohort.

Figure 21. Model Time Frame for Reunion Class Years Ending in 1's and 6's



Multiple Regression Scoring Model

The multiple regression model results from Excel 2007 are shown in Table 26. The variables are sorted by coefficient value, largest to smallest. The coefficients represent weights that indicate the importance of the variable to each individual's model score. Variables with larger coefficients or weights are more important for predicting Wharton MBA alumni who give to Wharton in a reunion year.

Table 25. Multiple Regression Model Results

Regression Statistics	
Multiple R	0.585441
R Square	0.342741
Adjusted R Square	0.34249
Standard Error	0.311584
Observations	28,771

	Coefficients
Intercept	-0.00584
Gave to Wharton between MBA reunions (1/0)	0.360046
Gave to Wharton in the previous MBA reunion year (1/0)	0.175908
Gave to Wharton before last two reunions (1/0)	0.032577
Wharton is a School Center / Personal Interest (1/0)	0.030750
Attended Other Penn Events (1/0)	0.030382
Have a Penn Alumni Spouse (1/0)	0.024822
Attended Other Wharton Events (1/0)	0.022584
Attended One or more Wharton 125th Receptions (1/0)	0.019628
Wharton Alumni Volunteer (1/0)	0.018732
Participated in Wharton Follies As a Student (1/0)	0.017416
Have a Wharton Executive MBA Degree (1/0)	-0.037200

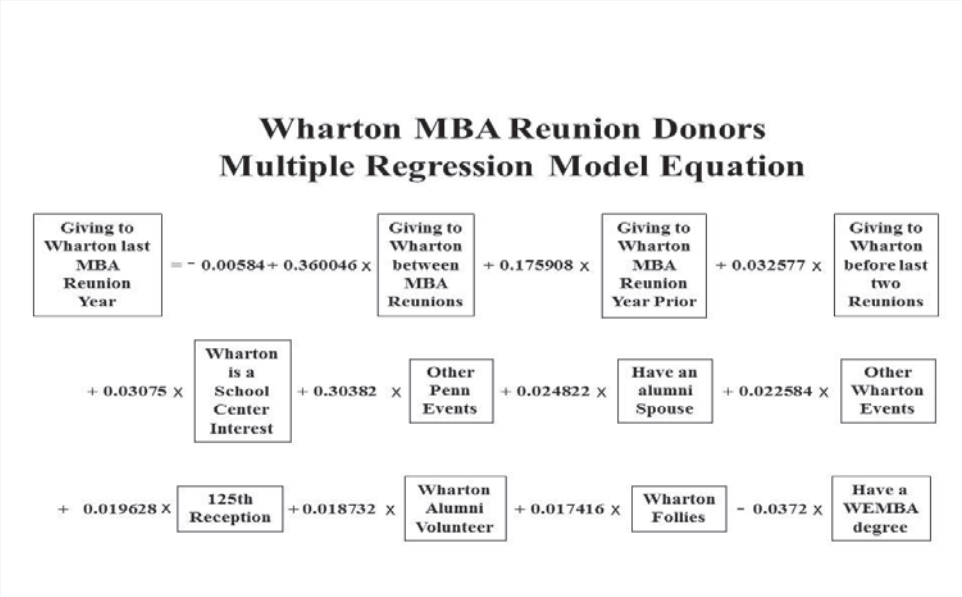
As with the undergraduate reunion scoring model, giving between the last two reunions is more important than giving at the previous reunion, although both are important. Giving prior to the last two reunions also adds to the model score.

The coefficient for a Wharton Executive MBA (WEMBA) degree is negative and deducts from the model score. All other things being equal, alumni of the WEMBA program are slightly less likely to donate in a reunion year than regular MBAs. WEMBA students meet the same admissions requirements, receive the same program services, follow the same curriculum, and study with the same faculty as students in the traditional, full-time MBA program. Sponsored mostly by their employer, however, they attend classes every other week on Fridays and Saturdays as a group. It is understandable, that the propensity to give during their reunion might be less for WEMBA alumni than for traditional MBAs.

Graduate and professional alumni have far fewer student activities than do undergraduates. Wharton Follies, however, is one of the most active MBA student clubs at Wharton. Each year, Follies members write and perform an original musical comedy that presents the amusing side of business school and corporate life. Participation in Wharton Follies adds positively to the model score.

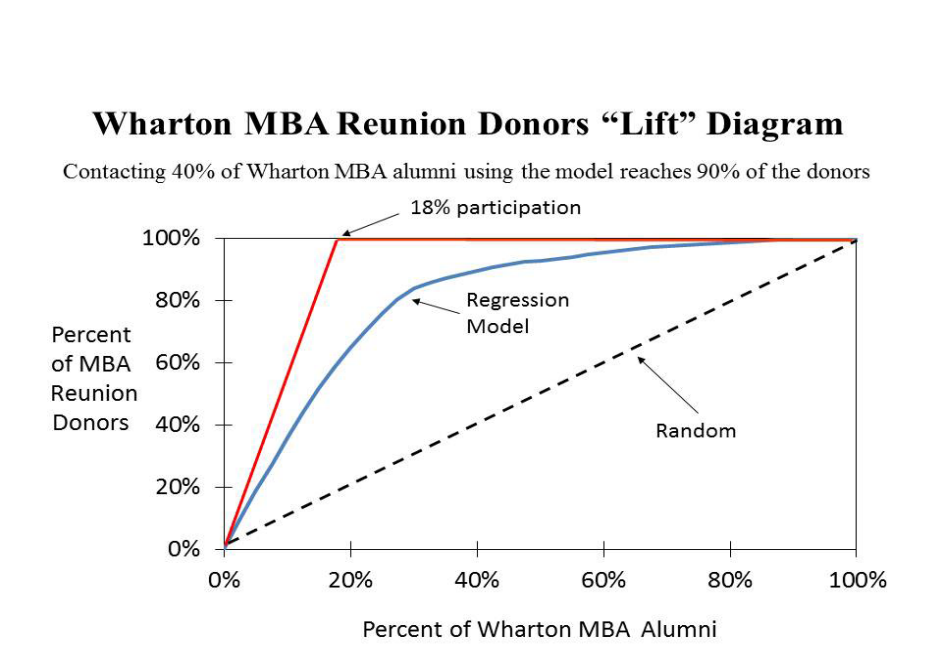
Graphically, the resulting model equation is shown in Figure 22. Note that the results are all based on a reunion year that has already taken place. Later, we will shift the equation five years forward to predict who will be a donor at their next reunion.

Figure 22. Multiple Regression Model



The lift diagram for the Wharton MBA Reunion Donor Participation model shown in Figure 23 characterizes the predictive ability of our scoring model differently than looking at the multiple R Square. Since we know from our data that the donor participation is 18%, the perfect model would be able to select only donors and exclude non-donors. Using the "random" model would require marketing to the entire population to get the 18% that are donors. For the multiple regression model developed here, the best 40% of scores contains 90% of the donors.

Figure 23. Wharton MBA Reunion Donors Lift Diagram



Predicting Donors in Their Next Reunion Year

To predict who will be a donor at their next reunion, we shift the time frame of the equation five years forward to apply the multiple regression equation. Figure 24 shows the time shift for reunion classes ending in 1 or 6.

Figure 24. Application of Multiple Regression Equation in FY2011

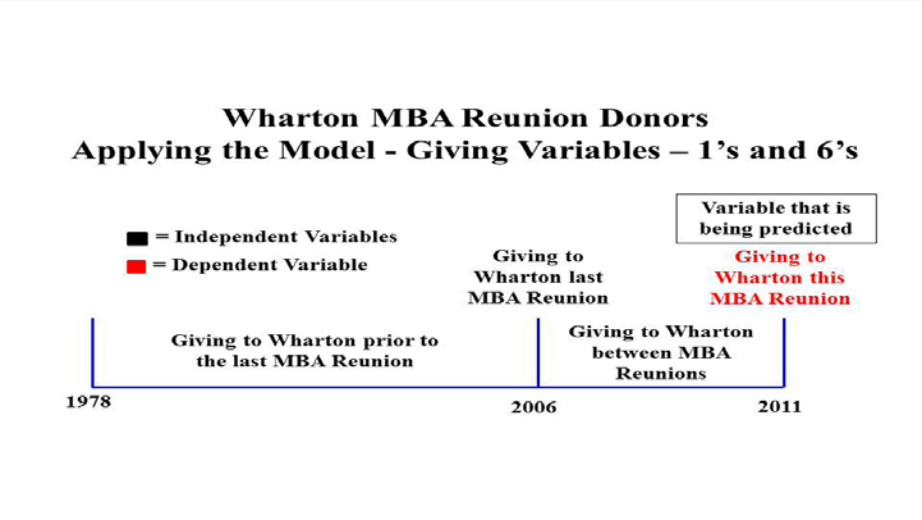
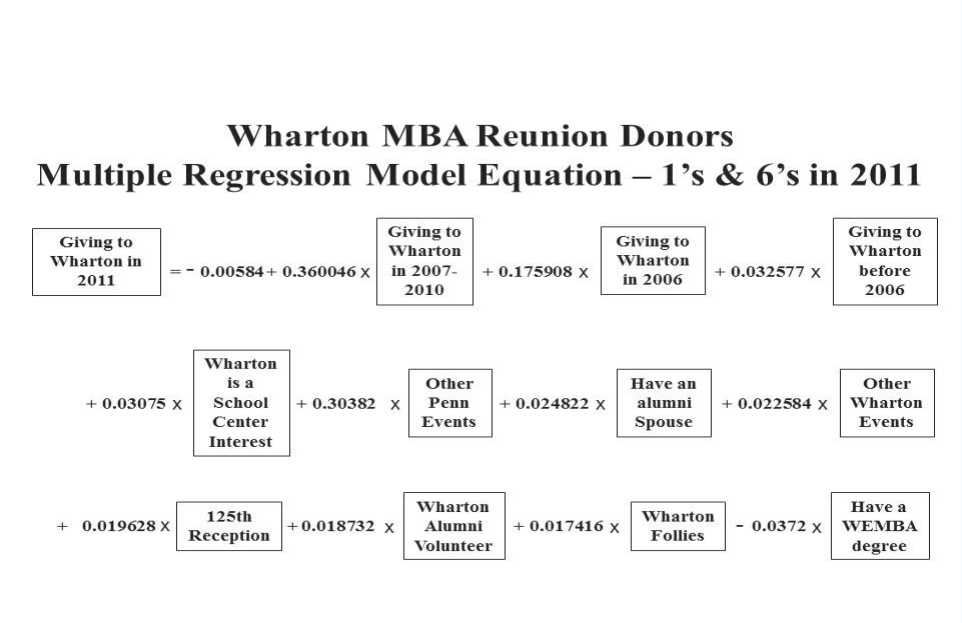


Figure 25 shows the multiple regression model equation for 1's and 6's in 2011. This model was used to calculate a score to predict the likelihood of being a donor in FY2011.

Figure 25. Multiple Regression Model Equation for 1's and 6's in FY2011



Model Validation (July 2011)

When FY2011 ended on June 30, 2011, we determined which Wharton MBA alumni in the classes ending in 1 or 6 were actually donors to Wharton during the preceding year. Among these 7,413 alumni there were 1,468 donors for an overall donor participation rate of 20% (Figure 26). Model scores were ranked and grouped into 10 "deciles." For the 717 alumni with the highest model scores, actual FY11 donor participation was 61%. As model scores decrease actual donor participation decreases.

Figure 26. Model Validation Results Using FY11 Actual Giving

Wharton MBA Reunion Model Validation

Model "Decile"	MBA Reunion 2011 Alumni	FY11 Actual Donors	FY11 Donor Participation
01	717	434	61%
02	665	340	51%
03	725	241	33%
04	707	214	30%
05	742	95	13%
06	568	35	6%
07	725	31	4%
08	885	45	5%
09	817	19	2%
10	862	14	2%
	7,413	1,468	20%

In Figure 27, we see that the top 38% of predictive model scores identified 84% of the actual FY11 donors. The top 48% of model scores identified 90% of actual FY11 donors.

Figure 27. Model Validation Results Using FY11 Actual Giving

Wharton MBA Reunion Model Validation

Model "Decile"	MBA Reunion 2011 Alumni	Percent of MBA Reunion 2011 Alumni	Cumulative Percent of MBA 2011 Reunion Alumni	FY11 Actual Donors	Percent of Actual FY11 Donors	Cumulative Percent of Actual FY11 Donors
01	717	10%	10%	434	30%	30%
02	665	9%	19%	340	23%	53%
03	725	10%	28%	241	16%	69%
04	707	10%	38%	214	15%	84%
05	742	10%	48%	95	6%	90%
06	568	8%	56%	35	2%	93%
07	725	10%	65%	31	2%	95%
08	885	12%	77%	45	3%	98%
09	817	11%	88%	19	1%	99%
10	862	12%	100%	14	1%	100%
	7,413	100%		1,468	100%	

Summary of Results and Implementation Recommendations

The variables in this model most likely to predict giving by Wharton MBA alumni in a reunion year were associated with previous giving behavior:

- Giving to Wharton between MBA reunions
- Giving to Wharton in the previous MBA reunion year
- Giving to Wharton before the last two reunions

Non-giving variables contributing to predicting giving to Wharton in an MBA reunion year were:

- Wharton was identified in alumni surveys as a Penn Interest
- Attendance at Wharton's 125th Anniversary Reception event(s)
- Attendance at other Wharton event(s)
- Attendance at other Penn event(s)
- Involved as a Wharton Alumni volunteer
- Married to a Penn alumna or alumnus
- Participated in Wharton Follies performance(s) while an MBA student

In most of our modeling work, variables that are predictive have a positive weight. Here we have an exception. If an individual has a Wharton Executive MBA (WEMBA) degree there is a slight deduction from their model score. This is not surprising considering the details of the WEMBA program described above. WEMBA's are somewhat less connected to Wharton than traditional MBA's.

With undergraduate alumni predictive modeling, we have a large number of student-activity-related variables to consider—few of which apply to graduate and professional degree alumni. As a result, more event and involvement variables come into play.

In summary, to predict giving in an MBA reunion year, focus on prior giving behavior, event attendance, Penn spouse relationship, and involvements as a student if available.

Finally, using predictive modeling in a reunion year uses existing data on giving at the last reunion as the dependent variable. When the model equation is applied to the current reunion year the time frame is advanced five years to calculate model scores.

Targeting Undergraduate Non-Reunion Year Donors

Each year, there over 100,000 alumni who are not having a class reunion; four times larger than the more than 25,000 alumni who are having a reunion. Penn undergraduate reunion classes receive special attention with respect to Alumni Weekend programming and gift solicitation, but not at the expense of annual giving solicitation for non-reunion year alumni.

In August 2010, we were asked by a member of the Penn Fund staff to develop a predictive model for non-reunion alumni (those who would not be having a reunion in May 2011). Of special interest was the need for a method to prioritize approximately 59,000 non-reunion alumni who were identified as lapsed or never givers in FY2010.

The Objective

The objective of this predictive modeling project was to develop scoring models for each of four class cohorts that will be used to predict donor participation in 2011, a non-reunion year for these classes. As with earlier scoring model projects a model score will be produced for each alumna or alumnus; the higher the score, the higher the likelihood that they will be a donor.

Although such models are used to identify likely donors with high model scores, we expected that the model scores could also be used to prioritize the 59,000 lapsed donors and never givers who probably have lower model scores.

Methodology

Datasets each with approximately 25,000 alumni records were created, cleaned, and conditioned. The lapsed donors/never givers represented slightly less that 60% of the records.

Table 26. Model Dependent Variables by Cohort

Reunion Cohort	Population	Lapsed Donors/ Never Giver List	Dependent Variable
0's and 5's	26,424	14,827	Giving to Penn 2010 – Y/N
4's and 9's	26,945	15,089	Giving to Penn 2010 – Y/N
3's and 8's	26,958	14,932	Giving to Penn 2010 – Y/N
2's and 7's	25,881	14,379	Giving to Penn 2010 – Y/N
Totals	106,208	59,227	

A different scoring model was developed for each reunion cohort using the variables shown in Table 27A and 27B. The models differed only in the particular giving variables used. Demographic, student activity, event participation, Alumni Census 1998, and other variables were the same for all models. For alumni in the reunion cohort of 0's and 5's, FY2010 was a reunion year. Because the results of these models will be implemented in FY2011, the model developed for alumni with a reunion in 2010 is structurally similar to the other cohort models, so we refer to all of these as non-reunion year modeling.

Table 27A. Giving Data Used for Each Reunion Cohort

Y = dependent variable, X = independent variables

1 = yes, 0 = no

Variable Definition	Reunion Cohort			
	0's – 5's	4's – 9's	3's – 8's	2's – 7's
ID	Information Only	Information Only	Information Only	Information Only
Preferred Class Year				
Last Reunion Year				
Last Non-Reunion Year				
Reunion Cohort				
Gave to Penn in 2010 (1/0)	Y	Y	Y	Y
Gave to Penn in 2009 (1/0)	X	X	X	X
Gave to Penn in 2008 (1/0)			X	X
Gave to Penn in 2007 (1/0)				X
Gave to Penn in 2006 (1/0)				
Gave to Penn in 2005 (1/0)	X			
Gave to Penn in 2004 (1/0)		X		
Gave to Penn in 2003 (1/0)			X	
Gave to Penn in 2002 (1/0)				X
Gave to Penn in 1978–2004 (1/0)	X			
Gave to Penn in 1978–2003 (1/0)		X		
Gave to Penn in 1978–2002 (1/0)			X	
Gave to Penn in 1978–2001 (1/0)				X
Gave to Penn in 2006–2008 (1/0)	X			
Gave to Penn in 2005–2008 (1/0)		X		
Gave to Penn in 2004–2007 (1/0)			X	
Gave to Penn in 2003–2006 (1/0)				X

Table 27B. Non-Giving Data Used for All Reunion Cohorts

Variable Definition	Variable Definition
Degree Complete (1/0)	Attended Events Recorded As Actions (1/0)
Graduate Degree (1/0)	Intercollegiate Sports (1/0)
Have an Alumni Spouse, Child, Parent, Sibling, or Other (1/0)	Fraternity Sorority (1/0)
Have an Alumni Spouse (1/0)	Student Performing Arts (1/0)
Have an Alumni Child (1/0)	Honor Society (1/0)
Have an Alumni Parent (1/0)	Student Activity Culture (1/0)
Have an Alumni Sibling (1/0)	Student Activity Service (1/0)
Have other Alumni Relatives (1/0)	Student Award (1/0)
Attended AR Alumni Weekend or Homecoming (1/0)	Student Academic Award (1/0)
Attended Any AR Alumni Weekend (1/0)	Other Student Sports (1/0)
Attended Any AR-Homecoming (1/0)	Other Student Activities (1/0)
Attended Any Penn Reunion Leadership Council (1/0)	Had Favorable Opinion of Penn As a Student According to Census 1998 (1/0)
Attended Any Other DAR Events (1/0)	Had Current Favorable Opinion of Penn According to Census 1998 (1/0)
Attended Any School Events (1/0)	Penn Was One of Top 4 Giving Priorities According to Census 1998 (1/0)
Attended Any Center Events (1/0)	Had Fulfilling Penn Area Mentioned in Census 1998 (1/0)
Attended Any Other Events (1/0)	Responded to Arts & Culture Survey in 2004 (1/0)

Scoring Model Results

Scoring model equations for each reunion cohort are shown in Table 28A. The more recent giving variables have higher weights than "older" giving variables; giving variables, in general, have higher weights (regression coefficients) than demographic, student activity, event participation and other variables. In other applications, one might consider dropping all variables but the giving variables. In this situation where we are trying to identify the best lapsed and never givers, the non-giving variables play an important role in the model score for these constituents.

Table 28A. Multiple Regression Coefficients by Reunion Cohort

1 = yes, 0 = no

Variable Definition	Reunion Cohort			
	0's – 5's	4's – 9's	3's – 8's	2's – 7's
ID	Information Only	Information Only	Information Only	Information Only
Preferred Class Year				
Last Reunion Year				
Last Non-Reunion Year				
Reunion Cohort				
Intercept	0.06979	-0.01744	0.00545	0.01103
Gave to Penn in 2009 (1/0)	0.41229	0.37953	0.47132	0.4108
Gave to Penn in 2008 (1/0)			0.18174	0.22427
Gave to Penn in 2007 (1/0)				0.10585
Gave to Penn in 2005 (1/0)	0.13361			
Gave to Penn in 2004 (1/0)		0.10158		
Gave to Penn in 2003 (1/0)			0.04834	
Gave to Penn in 2002 (1/0)				0.03029
Gave to Penn in 1978–2003 (1/0)		0.0381		
Gave to Penn in 1978–2002 (1/0)			0.01942	
Gave to Penn in 1978–2001 (1/0)				0.01277
Gave to Penn in 2006–2008 (1/0)	0.19054			
Gave to Penn in 2005–2008 (1/0)		0.21617		
Gave to Penn in 2004–2007 (1/0)			0.10272	
Gave to Penn in 2003–2006 (1/0)				0.06439

Variable Definition	Reunion Cohort			
	0's – 5's	4's – 9's	3's – 8's	2's – 7's
Graduate Degree (1/0)		0.01153		
Have an Alumni Spouse, Child, Parent, Sibling, or Other (1/0)	0.03129	0.02747	0.01568	0.01674
Attended AR Alumni Weekend or Homecoming (1/0)	0.04393			
Attended Any Other DAR Events (1/0)	0.27693			0.04038
Attended Any School Events (1/0)		0.03194	0.0433	0.03417
Attended Any Center Events (1/0)		0.10771	0.10249	0.13286
Attended Events Recorded As Actions (1/0)				-0.0187
Intercollegiate Sports (1/0)	0.02039			
Student Performing Arts (1/0)	0.03088			
Honor Society (1/0)	0.02298			
Penn Was One of Top 4 Giving Priorities According to Census 1998 (1/0)		0.02733	0.02665	

Variables with larger coefficients or weights in this table are more important for predicting undergraduate alumni donor participation in a non-reunion year.

The adjusted R Square, which measures the model's goodness of fit, is shown in Table 28B for each model. Models with more giving variables have slightly better adjusted R Square values. This suggests a consideration of some alternate model structures in future work.

Table 28B. Multiple Regression Statistics by Reunion Cohort

Variable Definition	Reunion Cohort			
	0's – 5's	4's – 9's	3's – 8's	2's – 7's
Multiple - R	0.66015	0.67517	0.72057	0.732228
Adjusted R Square	0.43560	0.45567	0.51907	0.535961
Alumni Count	26,424	26,945	26,958	25,881

Lift diagrams for each model are shown in Figures 28A through 28D.

Figure 28A. Non-Reunion Scoring Model for 0's and 5's– Lift Diagram

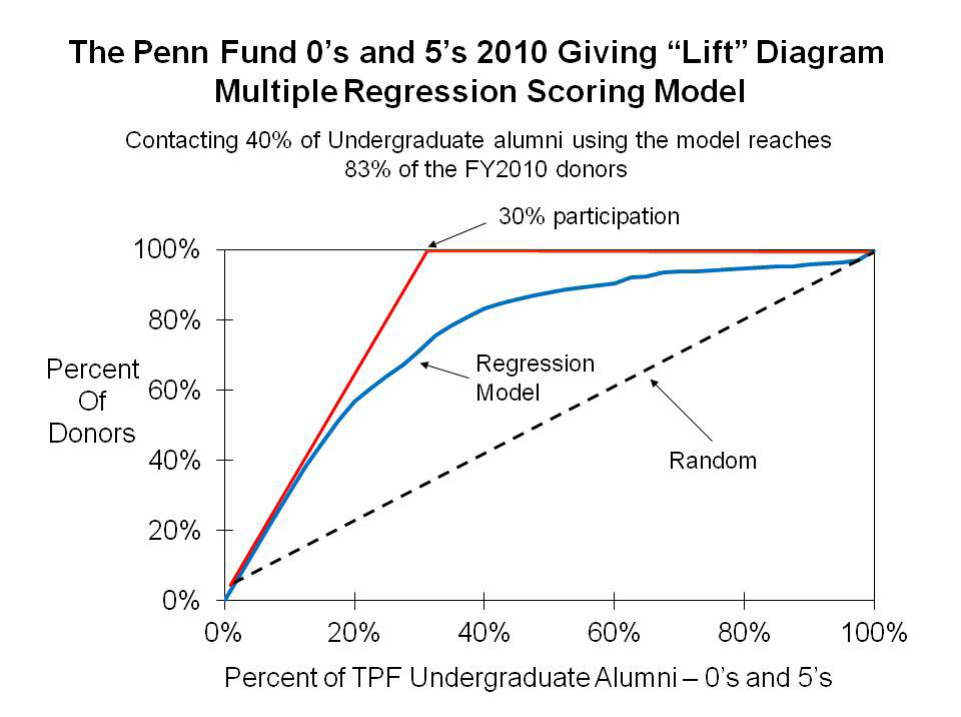


Figure 28B. Non-Reunion Scoring Model for 4's and 9's– Lift Diagram

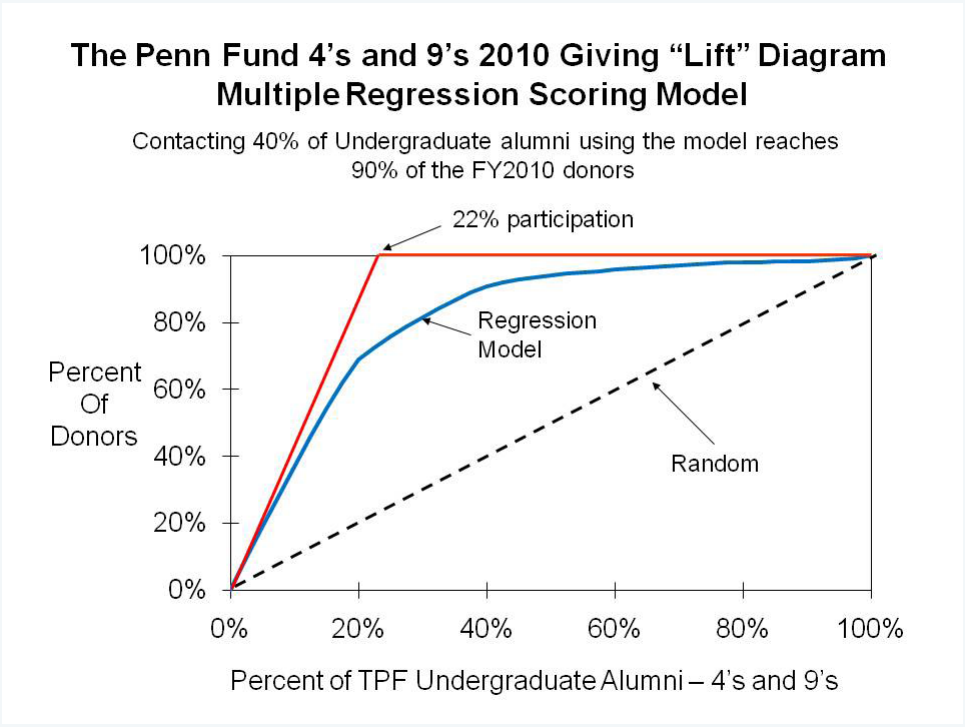


Figure 28C. Non-Reunion Scoring Model for 3's and 8's– Lift Diagram

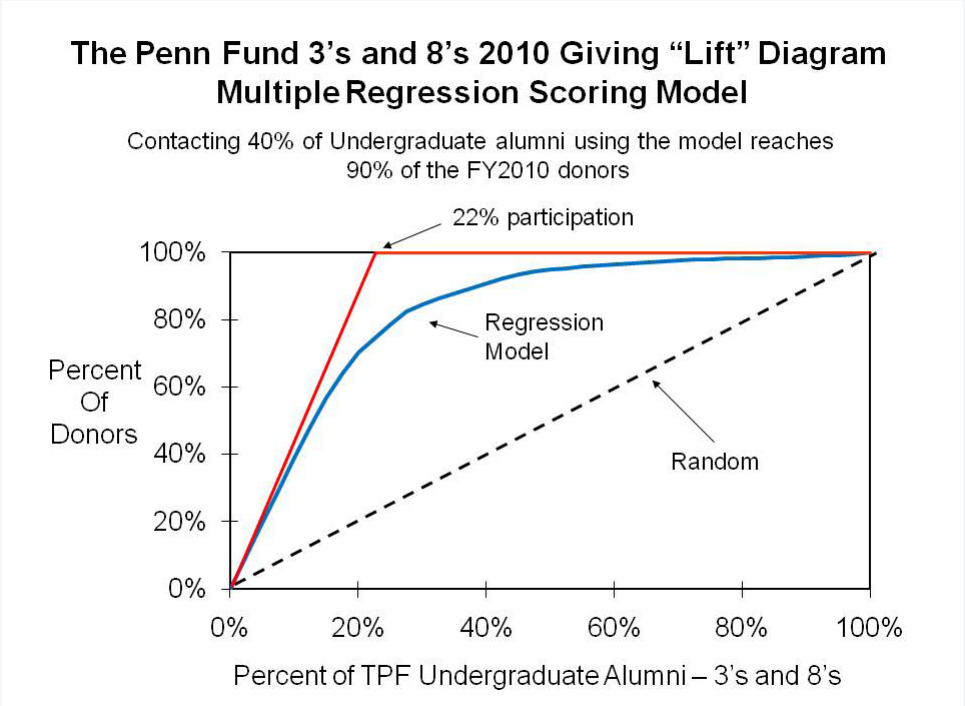


Figure 28D. Non-Reunion Scoring Model for 2's and 7's– Lift Diagram

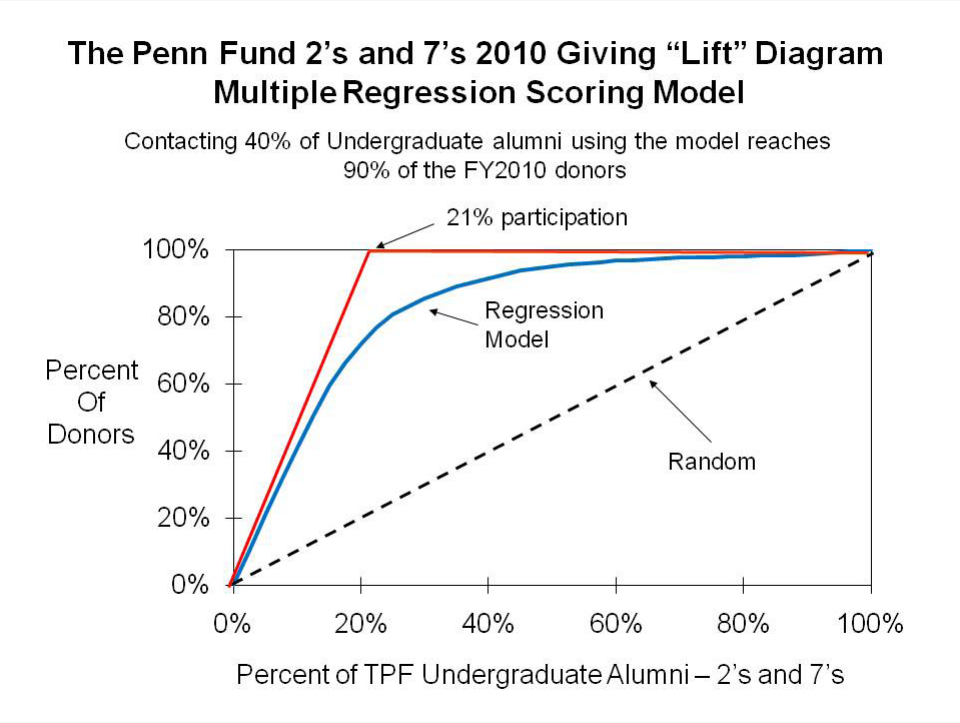


Table 29 shows a summary of Lift Diagram results. Although the predictive ability of all the models is high based on a sorting of model scores, those models with more giving variables in the model structure perform better especially when looking at the top 10% or 20% of model scores.

Table 29. Summary of Lift Diagram Model Score Results

Lift Diagram Model Score	Percent of FY2010 donors			
	Reunion cohort			
	0's – 5's	4's – 9's	3's – 8's	2's – 7's
10% highest scores	31%	37%	39%	41%
20% highest scores	57%	69%	70%	72%
40% highest scores	83%	90%	90%	90%
Alumni base	26,424	26,945	26,958	25,881

In a typical application individuals in a reunion cohort would be divided into groups based on model score. This model score grouping for the alumni with a reunion year ending in a 0 or 5 is shown in Table 30. For selecting likely donors overall, this grouping approach works well.

Table 30. Model Score Groups to Identify Likely Donors in a Reunion Class (Example - Reunion Cohort – 0's – 5's)

Group code	Count	Percent	Maximum model score	Minimum model score	FY10 donors	Percent of donors
1	5,282	20%	1.232623	0.555543	4,466	57%
2	3,773	14%	0.548994	0.301182	1,676	21%
3	4,774	18%	0.299005	0.113164	858	11%
4	3,863	15%	0.101085	0.090184	365	5%
5	8,732	33%	0.069795	0.069795	518	7%
Total	26,424	100%			7,883	100%

Remembering that the objective is to prioritize non-reunion lapsed and never givers, Table 31 shows that this grouping is not especially effective in segmenting these alumni.

In order to make these non-reunion scoring model results more useful, model groups were constructed for only the lapsed and never givers in a reunion cohort (See Tables 32A through 32D).

Table 31. Model Score Groups to Identify Likely Donors among Lapsed and Never Givers (Example - Reunion Cohort - 0's – 5's)

Group code	Count	Percent	FY10 donors	Percent of donors	Lapsed and Never Givers	Percent of Lapsed and Never Givers
1	5,282	20%	4,466	57%	501	3%
2	3,773	14%	1,676	21%	1,102	7%
3	4,774	18%	858	11%	3,152	21%
4	3,863	15%	365	5%	3,102	21%
5	8,732	33%	518	7%	6,970	47%
Total	26,424	100%	7,883	100%	14,827	100%

Table 32A. Lapsed and Never Givers Model Score Groups
Reunion Cohort - 0's – 5's

Score group	Count	Percent	Maximum score	Minimum score
1	1,467	10%	1.209643	0.342892
2	1,495	10%	0.335552	0.188380
3	1,793	12%	0.175886	0.113164
4	3,102	21%	0.101085	0.090184
5	6,970	47%	0.069795	0.069795
Total	14,827	100%		

Table 32B. Lapsed and Never Givers Model Score Groups
Reunion Cohort - 4's – 9's

Score group	Count	Percent	Maximum score	Minimum score
1	1,675	11%	0.896582	0.226059
2	1,450	10%	0.210255	0.079934
3	4,185	28%	0.075464	0.020669
4	1,902	13%	0.014500	-0.005911
5	5,877	39%	-0.017436	-0.017436
Total	15,089	100%		

Table 32C. Lapsed and Never Givers Model Score Groups Reunion Cohort - 3's – 8's

Score group	Count	Percent	Maximum score	Minimum score
1	1,526	10%	0.990472	0.150503
2	1,418	9%	0.143278	0.067205
3	1,699	11%	0.064433	0.032101
4	3,791	25%	0.024875	0.021132
5	6,498	44%	0.005451	0.005451
Total	14,932	100%		

Table 32D. Lapsed and Never Givers Model Score Groups Reunion Cohort - 2's – 7's

Score group	Count	Percent	Maximum score	Minimum score
1	1,357	9%	1.064833	0.120375
2	1,468	10%	0.118483	0.055990
3	2,818	20%	0.054098	0.026480
4	2,254	16%	0.023806	0.021823
5	6,482	45%	0.011033	-0.007686
Total	14,379	100%		

Scoring Model Implementation

Using the model groupings in Tables 32A through 32D, the Penn Fund staff was able to select the number of lapsed and never givers in each reunion cohort and reunion class for solicitation. For email solicitations, the entire group might be targeted. For more expensive and labor intensive methods such as mail and phone, the best 30% or 40% might be a more appropriate target.

Summary of Results and Implementation Recommendations

The variables in these four models most likely to predict giving to Penn in a non-reunion year were associated with previous giving behaviors in some way, typically:

- Giving to Penn in the previous year
- Giving to Penn in the last reunion year
- Giving to Penn in years between reunions either separately or in a grouping of years
- Giving to Penn in years prior to the last reunion.

Other important predictors of donor participation were event participation variables including attendance at:

- Any Alumni Relations Alumni Weekends or Homecoming Weekends
- School-Sponsored events
- Center-Sponsored events
- Other Development and Alumni Relations events.

Penn alumni relationships in aggregate were also important predictors of giving in a non-reunion year. Here we combined the existence of an alumni spouse, child, parent, sibling, or other relative into a single variable.

Participation in intercollegiate sports, student performing arts groups, and honor society membership were statistically significant variables in one of the models.

In 1998, Penn conducted an alumni census to gather contact, demographic, and attitudinal information. Answers to where Penn is ranked as a philanthropic priority has been useful as an independent variable in two models here and also in other predictive models not discussed in this monograph.

One other significant variable as a predictor in one model was having a Penn graduate or professional degree in addition to an undergraduate degree.

Similar to predictive models discussed earlier, to predict giving in a non-reunion year, focus on prior giving behavior, attendance at events, family Penn relationships, involvements as a student, graduate Penn degree(s), and available attitudinal variables like Penn as a philanthropic priority.

One interesting outcome of this modeling project was the ability of the models to prioritize lapsed and never givers. Usually because giving variables are so important in our predictive models, the most likely donors identified with high model scores are previous donors—the so-called best of the best prospects for donor participation. Here we are using model scores to identify the best donor prospects among alumni with less giving or no giving history.

Targeting Penn Vet Client Annual Giving Program Donors

The School of Veterinary Medicine at the University of Pennsylvania was founded in 1884. Penn currently has two veterinary teaching hospitals: The Matthew J. Ryan Veterinary Hospital for companion and small animals and The George D. Widener Hospital for large animals at New Bolton Center. Both hospitals see very large caseloads and have financial needs like all non-profit institutions. The work of these hospitals is made possible largely by private donations. Because of this, the Penn Vet Annual Giving program is especially crucial to the success of the hospitals.

The challenge for the Penn Vet Annual Giving program is that the majority of their donors are previous clients, of which little information is known except for the services performed at the hospital. Also, many of the clients have no other affiliation to Penn. With over 30,000 patients per year at the Ryan small animal hospital, Penn Vet can only solicit a fraction of these clients to make an annual gift due to mailing costs. With such a loosely affiliated target audience, it is difficult to improve on dollar amounts raised with messaging and other techniques. If, however, solicitation mailing costs can be reduced without decreasing the dollar amount raised, it would greatly benefit the program.

A predictive model was created for the Ryan small animal hospital to help target the most likely donors, and cut solicitation mailing costs, with little to no loss in the dollar amounts raised.

The Objective

The objective of this predictive modeling project was to develop a scoring model to identify and prioritize previous clients who are most likely to make an annual gift to Penn Vet. A model score was produced for each client; the higher the score, the higher the likelihood that they will be a donor to Penn Vet.

The Process

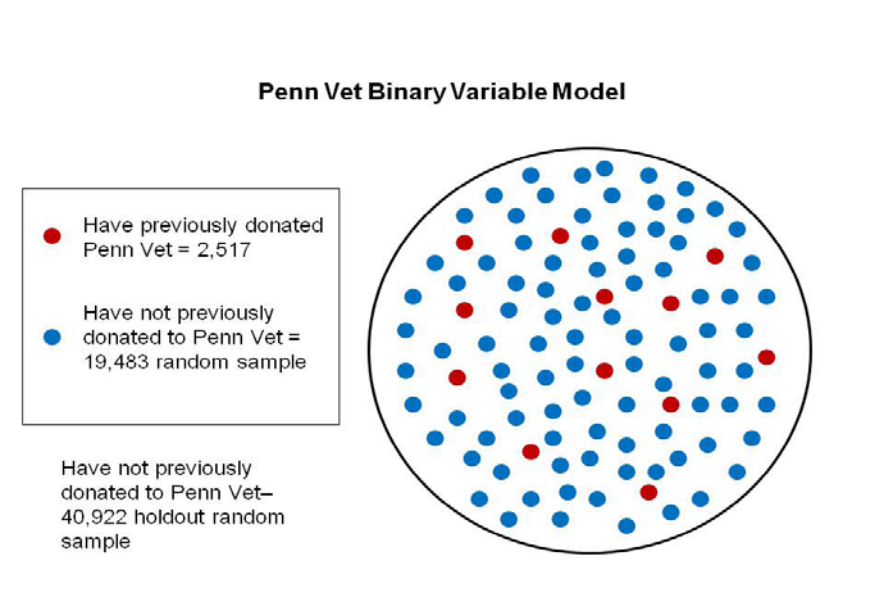
A data set with 62,922 records for clients to the Ryan small animal hospital was cleaned and conditioned. In order to begin to build this model, many variables were considered, including:

- New client (yes/no)
- Service(s) performed
- Total visits lifetime
- Service charges to client's account
- Discharge condition of animal
- Total animals treated lifetime
- Time period since first/last visit
- Emergency service (yes/no)

Multiple Regression Scoring Model

In this situation only a very small percentage (4%) of records represent donors to Penn Vet Annual Giving (target variable value =1). As discussed in the section on sampling methodology, unless we use non-representative sampling, the population is far too saturated with non-donors to create a reliable model using multiple regression. The solution was to draw a random sample of 19,483 non-donor clients from the Ryan small animal hospital database to combine with the 2,517 donor clients. The concept is illustrated in Figure 29. The red dots represent the clients who have made a donation to Penn Vet Annual Giving funds. The blue dots represent the random sample of clients who are non-donors. For the model, we now have a ratio of approximately 8:1 non-donors to donors. The remainder of the records (40,922) were held-out of the model-building process. A score was calculated for each of these records later using the resulting model.

Figure 29. Graphical Representation of Penn Vet Clients



The multiple regression results from IBM SPSS Modeler 14.1 are shown below in Table 33. The variables are sorted by coefficient value, largest to smallest. The coefficients represent weights that indicate the importance of the variable in each individual's model score.

Table 33. Multiple Regression Model Results

Regression Statistics	
Multiple R	.219
R Square	.048
Adjusted R Square	.048
Standard Error	.310622
Observations	22,000

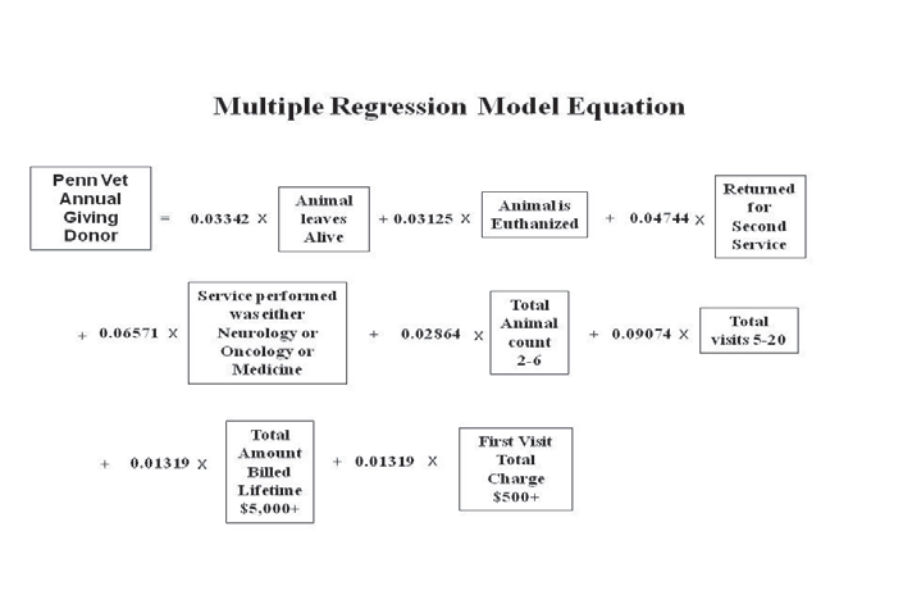
	Coefficients
Intercept	0.00000
Total Amount Billed Lifetime $5,000+ (1/0)	0.09476
Total Visits 5–20 (1/0)	0.09074
Service Performed Was Neurology, Oncology, or Medicine (1/0)	0.06571
Returned for Second Service (1/0)	0.04744
Animal Leaves Alive (1/0)	0.03342
Animal is Euthanized (1/0)	0.03125
Total Animal Count 2–6 (1/0)	0.02864
First Visit Total Charge $500+ (1/0)	0.01319

The independent variables demonstrate the importance of aspects of connectivity to the hospital, such as total amount billed lifetime of $5,000 or more, more than five visits to the hospital, and the type of service received. Three services, in particular, that showed higher donor participation rates than others were neurology, oncology, and medicine.

A very interesting variable was the discharge condition of the animal. Whether the animal was released from the hospital alive or was euthanized both positively contribute to the model score. It is hypothesized that even those clients whose animal must be euthanized feel a connection to the hospital for the care their animal received and the compassion they were shown.

Graphically, the resulting model equation is shown in Figure 30.

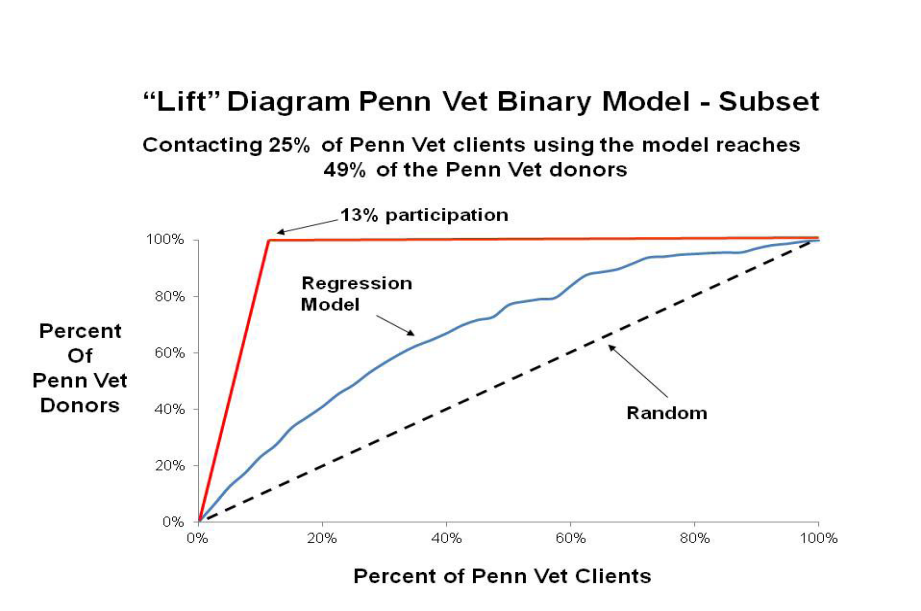
Figure 30. Vet Annual Giving Multiple Regression Model



The R Square for this regression model is very low, compared to alumni annual giving models. The lift diagram, however, shows in a different way the predictive capability of the model. This demonstrates that R Square does not necessarily define the predictive power of a model.

The perfect model would be able to select only the 11% who are Penn Vet Annual Giving donors. The random model requires contacting the entire population to get the 11% that are donors. With the regression model developed, contacting the clients with a model score in the top 25% would reach 49% of Penn Vet Annual Giving donors.

Figure 31. Penn Vet Annual Giving Model Lift Diagram



Summary of Results and Implementation Recommendations

The predictive modeling project for the Vet Client Annual Giving program represents a very different situation than what was discussed earlier in this monograph. To predict Vet Clients who are most likely to donate in response to an annual giving appeal, the only giving variables available were used as the dependent variable.

Operationally, the Penn Vet development office will mail an annual giving appeal to everyone who has been a donor in the past five or more years. The purpose of the predictive modeling is to use information about the hospital visit experience to identify the clients who are most likely to become new donors.

Most important predictive variables are related to amount of service used:

- Total amount billed to the lifetime is $5,000 or more
- Total number of visits in the lifetime are between five and twenty
- The first visit total charges were $500 or more
- The number of animals treated for this client are between two and six
- Whether or not the client returned for a second service.

Another important variable is if the service was performed by Neurology, Oncology and/or Medicine departments.

Finally, variables related to the service outcome are important. Either having the animal discharged alive or having the animal euthanized is a positive predictor of donor participation. The lesson here is to be careful not to exclude variables that might turn out to be predictive even if, intuitively, they seem not to be.

We also used the technique of random sampling of the clients who had not yet made a gift to Vet Annual Giving in creating the dataset for predictive model. Without random sampling, the number of records with a dependent variable value of 1 (giving) would be overwhelmed by those records with a dependent variable value of 0, causing difficulties in obtaining a stable regression equation.

This project required more data exploration than most of our other work because it was new territory for us. Before starting on the predictive modeling work, a large number of crosstabs were constructed to examine relationships between the dependent variable (giving to the Vet Annual Giving program) and each of the independent variables. For a detailed description of data exploration with cross-tabs see the discussion of Targeting Undergraduate Reunion Year Donors earlier in this monograph.

REUNION VOLUNTEER RECRUITING – PLANNING COMMITTEES

Up until this point in the monograph, the models we discussed have been focused around financial goals and predicting donor participation. We can, however, also predict other non-financial behaviors.

Penn's Alumni Relations (AR) department focuses much of their resources on quinquennial reunion classes. For several years, Penn Development and Alumni Relations has held the Penn Reunion Leadership Conference (PRLC) to attract, train, and motivate alumni to be Reunion Planning Committee volunteers and/or Reunion Gift Committee volunteers. For each class reunion, AR recruits a Reunion Planning Committee to assist in planning the reunion celebration for its classmates. The challenge that AR faces is that current committee members only recruit the classmates they know, instead of reaching out to other class members; this creates a lack of breadth on the committee. Furthermore, the Alumni Weekend post-event survey demonstrates the importance of contact from a classmate on reunion participation. A lack of committee breadth may also limit class participation.

Given the importance of volunteer recruiting, AR staff asked if we could build a model to identify the characteristics of alumni who are most likely to be Reunion Committee volunteers. With such a predictive model, AR could begin to reach to a broader audience of new, more diverse alumni to recruit and build a pipeline for the future.

In the summer of 2011, we found that there were 242 Reunion Planning Committee volunteers out of 30,383 undergraduate alumni in the reunion cohort ending in 2 or 7. It would be very helpful to AR to have a predictive model that would effectively identify those alumni most likely to volunteer to help plan their reunion in 2012.

The Objective

The objective of this predictive modeling project was to develop a scoring model to identify and prioritize potential Reunion Planning Committee volunteers in the reunion classes ending in 2 or 7 for additional outreach and follow up. A model score was produced for each alumna or alumnus. The higher the model score for an individual, the higher the likelihood that they are willing to be a Reunion Planning Committee volunteer, if recruited more proactively.

The Process

A data set with 30,383 records for alumni with an undergraduate degree from Penn in the years ending in 2 or 7 was created, cleaned, and conditioned. In order to begin to build this model, many variables were considered. After looking at the characteristics of the previous volunteers, we realized that independent variables that demonstrated a recent connectivity to Penn would most likely have the largest impact on whether or not an alumna or alumnus would be a volunteer.

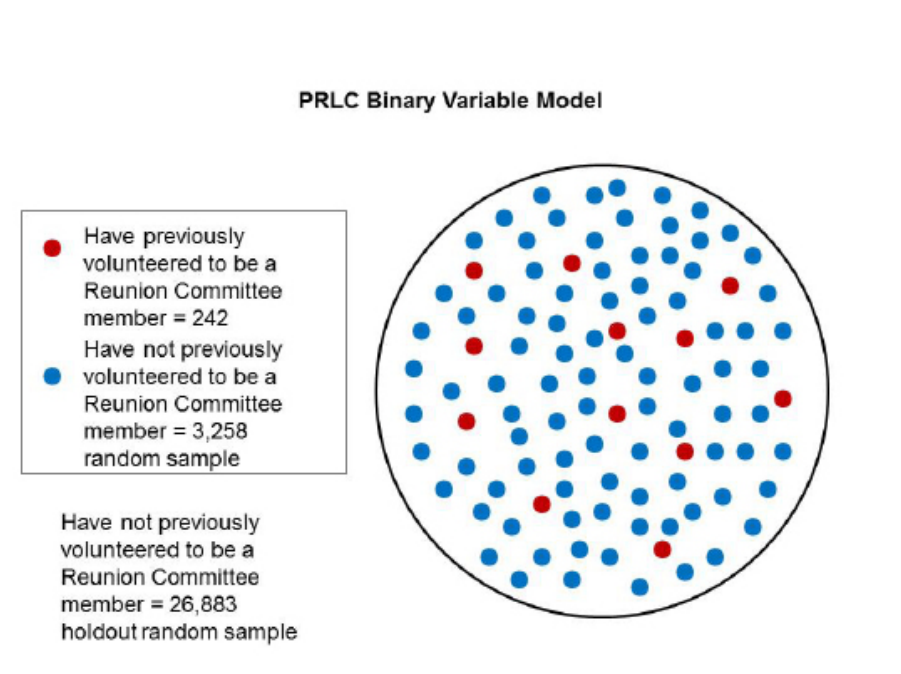
Variables that were considered are:

- Giving to Penn - amounts and consistency measures
- Event participation
- Harrison Society member
- Student activity participation
- Alumni Census 1998—opinions about Penn and Penn's philanthropic rank
- Previous other alumni volunteer participation
- Rated pool prospect
- Relationships with alumni and non-alumni

Multiple Regression Scoring Model

Since only 0.8% of alumni previously served on a Reunion Planning Committee, without sample adjustment, the population would be far too saturated with non-volunteers to create a reliable model using multiple regression. The sampling concept is illustrated in Figure 32. The red dots represent the 242 alumni who have been Reunion Planning Committee volunteers. The blue dots represent a random sample of 3,258 alumni who have not been volunteers. The remaining population of 26,883 was held-out of the process of building the regression model. A model score was calculated and applied later to these records using the resulting equation. The ratio of previous volunteers to non-volunteers in the regression model was 13:1.

Figure 32. Sampling of Penn Reunion Committee Volunteers and Non-Volunteers



The multiple regression results from IBM SPSS Modeler 14.1 are shown below in Table 34. The variables are sorted by coefficient value, largest to smallest. The coefficients represent weights that indicate the importance of the variable in each individual's model score.

Table 34. Multiple Regression Model Results

Regression Statistics	
Multiple R	.461
R Square	.213
Adjusted R Square	.211
Standard Error	.225451
Observations	3,500

	Coefficients
Intercept	-0.01143
Attended a Homecoming Weekend (1/0)	0.3393
Harrison Society Member (1/0)	0.1726
Alumni Child (1/0)	0.1355
Alumni Census 1998 - Penn Top 4 Giving Priorities (1/0)	0.1181
Student Activity Service (1/0)	0.09724
Giving to Penn 3 or More Years out of 5 (1/0)	0.07743
Alumni Spouse (1/0)	0.02343
Fraternity/Sorority (1/0)	0.02343
Penn Lifetime Giving (1/0)	0.02202

The variables used in the model show a strong emphasis on recent involvement with Penn through multiple channels. This is an intuitive and practical result considering what would make someone more likely to volunteer for Penn. The variable with the largest coefficient is attendance at Homecoming Weekend, which only has very reliable data for the past three years. This is an indicator that the individual has been engaged with Penn very recently. We did not use attendance at Alumni Weekend as a variable since every volunteer would have that variable; they are planning their reunion, which takes place during Alumni Weekend, and it is expected that they will attend and participate as well.

Another strongly predictive variable is Harrison Society member. The Harrison Society was founded as a way to provide ongoing thanks and to acknowledge the generosity of alumni, parents and friends who have named Penn as a beneficiary of a will, living trust, retirement plan or life insurance policy, or have set up a life income gift that benefits Penn in the future. Therefore, members of this society have shown a connection and commitment to Penn, making them more likely to be a volunteer.

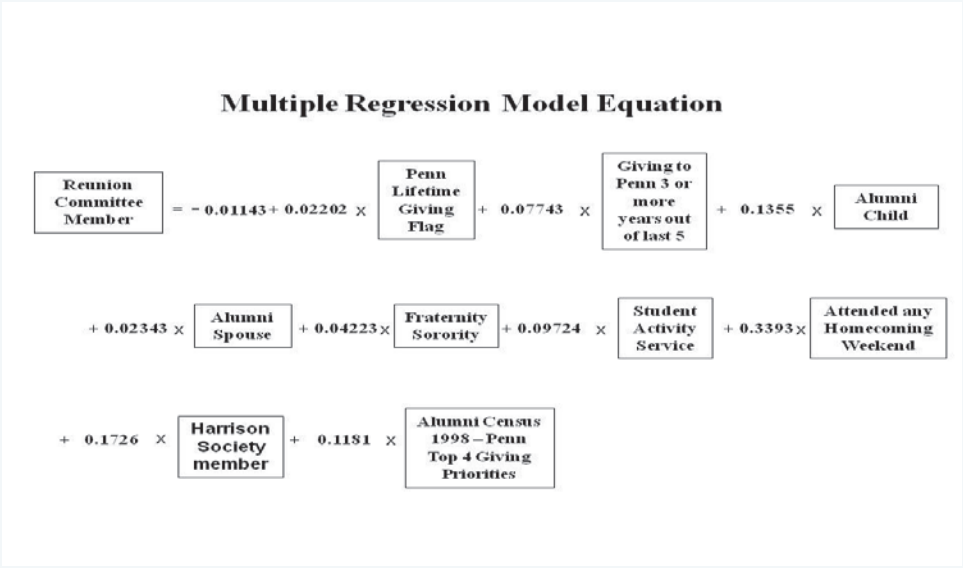
In 1998, Penn conducted an all-alumni census to gather and update biographical information. In this census, there were questions about an individual's attitudes and opinions about Penn. In one question, respondents were asked to rank their top four giving priorities—with Penn as one of the options. In this model, an individual choosing Penn as any one of their top four giving priorities was shown to be predictive.

As in many of our models, relationships to other Penn alumni indicate a stronger connectivity to Penn and are often predictive. Also, involvements while they were a student are shown to be predictive. In this case whether or not they participated in a fraternity or sorority and if they were involved in a student group focused around community service were strongly related to being a Penn Reunion Planning Committee volunteer.

Finally, in this model there are donor participation and giving consistency variables, but they are binary variables and not based on any specific dollar amount. They are not major contributing variables to the model score. Even someone without giving, therefore, could still have a relatively high model score.

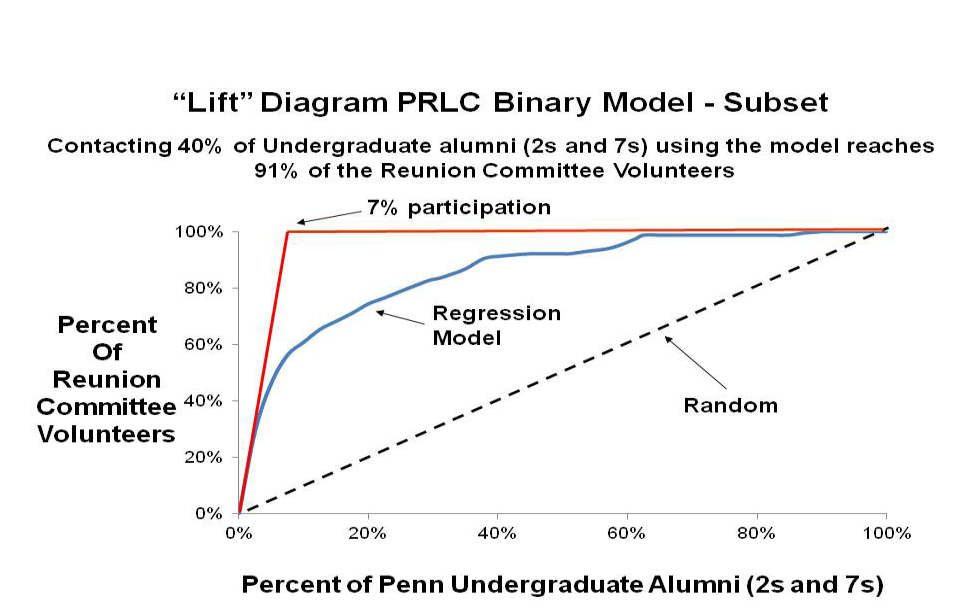
Graphically, the resulting model equation is shown in Figure 33.

Figure 33. Reunion Planning Committee Volunteer Multiple Regression Model



The lift diagram for the Reunion Planning Committee Volunteer model shown in Figure 34 characterizes the predictive ability of our scoring model differently than looking at the multiple R Square. The R Square is relatively low compared to alumni annual giving models. The lift diagram, however, shows a strong predictive power for the model. Specifically, contacting 40% of undergraduate alumni in the reunion cohort of 2's and 7's using the model selects 91% of the previous Reunion Planning Committee volunteers. The perfect model would be able to select exactly those alumni who were previous volunteers (7% of the sample). The random model requires contacting the entire population to get all the previous volunteers.

Figure 34. Reunion Planning Committee Volunteer Model Lift Diagram



The regression model developed here provides an effective tool for AR staff to identify prospective volunteers using characteristics that separate volunteers from non-volunteers.

To most effectively use the results of the model, records need to be sorted first by class year and then by model score from highest to lowest. Each reunion class year has their own Reunion Planning Committee to organize their specific reunion dinner and other class events. Also, the scores are relative to the class group. For example, the older alumni classes have more Harrison Society members than the younger alumni classes. Therefore, the older alumni classes have higher scores than the younger alumni. This is not a problem in this situation though because AR will be comparing model scores for individuals in the same reunion class, not the overall population. The difference in scores for the classes is illustrated in Figure 35 by looking at the maximum and minimum score for each class.

Figure 35. Reunion Planning Committee Volunteer Class Scores

UG Binary Variable Model Scores by Class

Class Year	Count	Reunion Committee Volunteers	Max. Score	Min. Score
1952	1,643	39	0.97419	-0.01143
1957	1,437	15	0.70229	-0.01143
1962	1,463	10	0.91918	-0.01143
1967	1,890	17	0.89832	-0.01143
1972	2,168	15	0.85609	-0.01143
1977	2,203	13	0.71802	-0.01143
1982	2,484	56	0.72572	-0.01143
1987	2,474	8	0.58625	-0.01143
1992	2,530	34	0.54799	-0.01143
1997	3,316	11	0.66609	-0.01143
2002	2,888	10	0.56679	-0.01143
2007	3,066	1	0.59022	-0.01143
Total	27,562	229		

Each class contains anywhere from 1,500 to 3,000 alumni, and a volunteer committee can have between 10 and 50 volunteers. Therefore, it is only realistic for Alumni Relations staff to try to contact about 10% of the total class members in order to get the desired number of volunteers. Figure 36 demonstrates that with the model, even by only contacting 10% of the class, one is able to identify a large portion of previous volunteers. Others in the top 10% groups are individuals who have the same characteristics as the previous volunteers and have the best likelihood of being new Reunion Planning Committee volunteers.

Figure 36. Model Class Year Segmentation Results

Model Results

	% Volunteers in Top 10% of Scores		% Volunteers in Top 10% of Scores
1952	51%	1982	54%
1957	87%	1987	50%
1962	60%	1992	32%
1967	71%	1997	45%
1972	53%	2002	80%
1977	77%	2007	100%

Summary of Results and Implementation Recommendations

Unlike the models for Annual Giving programs described earlier, the variables in this volunteer recruiting model that are most important at predicting which alumni are likely to volunteer for a role on their class reunion planning focus on connectivity with Penn in a variety of ways. Important variables were related to more recent connections:

- Attendance at a Homecoming Weekend
- Membership in the Harrison Society (i.e., planned giving commitment to Penn)
- Having a child who went to Penn
- Married to an alumna or alumnus.

Other variables indicate connections as a student:

- Service-Oriented student activities
- Membership in a fraternity or sorority.

Still other variables relate to an "overall relationship of giving to Penn":

- Having lifetime giving to Penn
- Giving to Penn three or more years out of five
- Penn is ranked as one the top four philanthropic priorities (Census 1998).

Each quinquennial reunion class has over 25,000 Penn undergraduate alumni. Only one percent of these alumni have ever been a Reunion Planning Committee volunteer. Here again, the technique of random sampling of the alumni who had never been a volunteer was used to create the dataset for predictive modeling.

One concern with this model that needs to be addressed is that the independent variables "favor" older alumni. Simply said, older alumni have more time to join the Harrison Society, marry an alumna or alumnus, have a Penn child, have lifetime giving, and are old enough to have been a respondent to a survey conducted in 1998. Because we look at model scores class-by-class, all alumni in a reunion class have the same opportunity to have lifetime giving to Penn, etc. so we can still use ranked model scores within each class to identify likely Reunion Planning Committee volunteers.

Finally conducting a predictive modeling project like this may require the consideration of variables and use of random sampling which are not usually needed in donor participation models.

REUNION VOLUNTEER RECRUITMENT – GIFT COMMITTEES

The Penn Fund (TPF) Annual Giving program focuses much of their resources on quinquennial reunion classes. Effective recruiting of Reunion Class Gift Committee volunteers is especially important. For several years, Penn Development and Alumni Relations has held the Penn Reunion Leadership Conference (PRLC) to attract, motivate, and train alumni to be Reunion Planning Committee volunteers, Reunion Gift Committee volunteers, or both. With over 150,000 living Penn undergraduate alumni and approximately 30,000 in each reunion class, the ability to focus fundraising resources on those most likely to be a volunteer is very important.

TPF encounters the same challenges that Alumni Relations (AR) has in recruiting new committee volunteers because the current volunteers typically recruit only the classmates they know, creating a lack of breadth on the committees. Therefore, TPF was looking for a way to identify reunion class alumni who were most likely to be a Reunion Gift Committee volunteer.

The PRLC Volunteer model was originally created to predict the likelihood an individual will be a Reunion Planning Committee volunteer. The Penn Fund Reunion Donor Participation model was created to predict the likelihood an individual makes a gift to Penn in their reunion year. We hypothesized that the combination of these two models should determine which alumni have a high inclination to volunteer and to give to Penn in their reunion year. These are, of course, the desirable characteristics for a Reunion Gift Committee volunteer.

The Objective

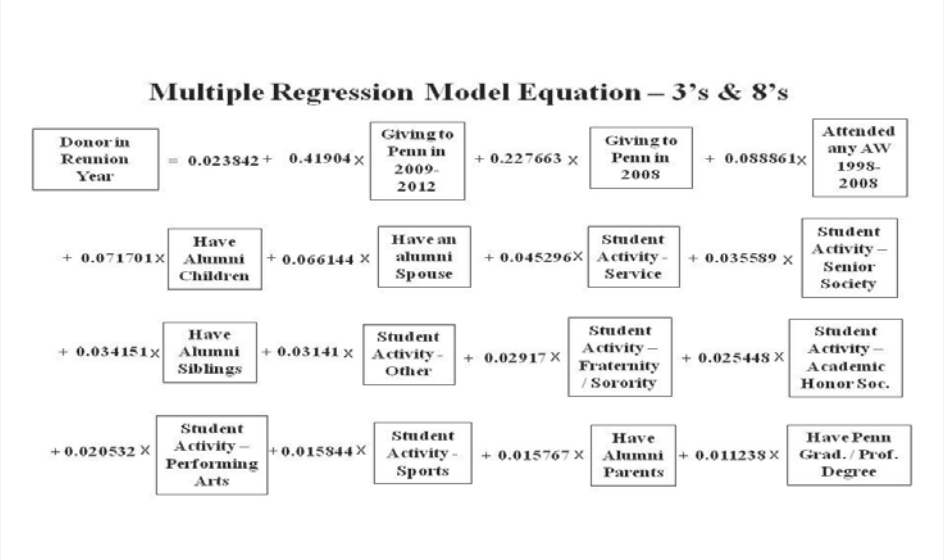
The objective of this predictive modeling project was to combine two scoring models to identify and prioritize potential Reunion Gift Committee volunteers in the reunion classes ending in 3 or 8 for additional outreach and follow up.

To demonstrate the validity and operational value of these models for Penn Fund reunion planning, the models were applied retrospectively to alumni who just had their reunion at Alumni Weekend in May 2011.

The Process

A data set with 32,454 records for alumni with an undergraduate degree from Penn in the years ending in 3 or 8 was assembled. The Penn Fund Reunion Donor Participation model equation was applied and scores were calculated for this group. The graphical representation of this model is shown in Figure 37.

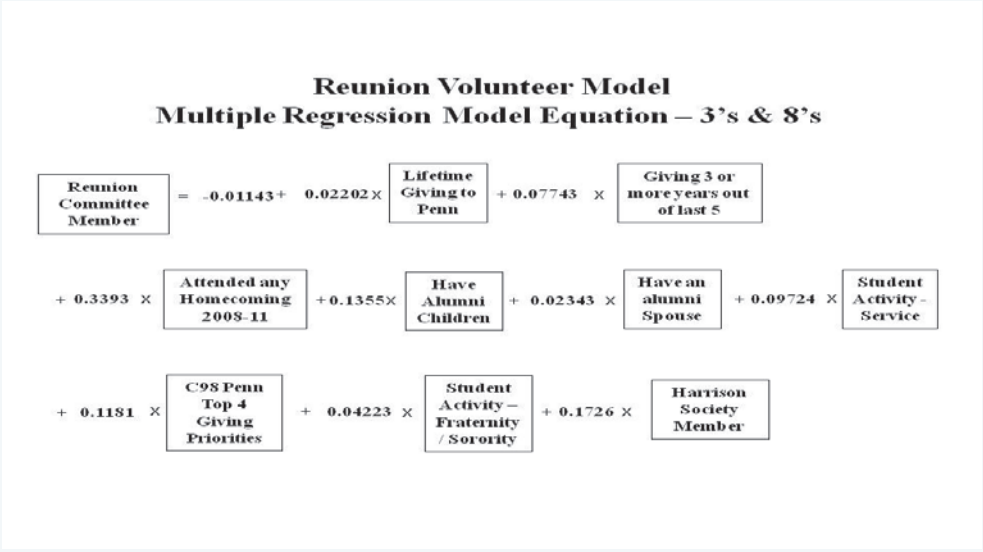
Figure 37. The Penn Fund Donor Participation in a Reunion Year Model Equation



Each record was then assigned to a group ranging from 1 to 10 depending on the model score, with 1 being the group most likely to donate (largest model score).

The PRLC Reunion Planning Committee Volunteer model equation was also applied to this group. The graphical representation of the model equation is shown in Figure 38.

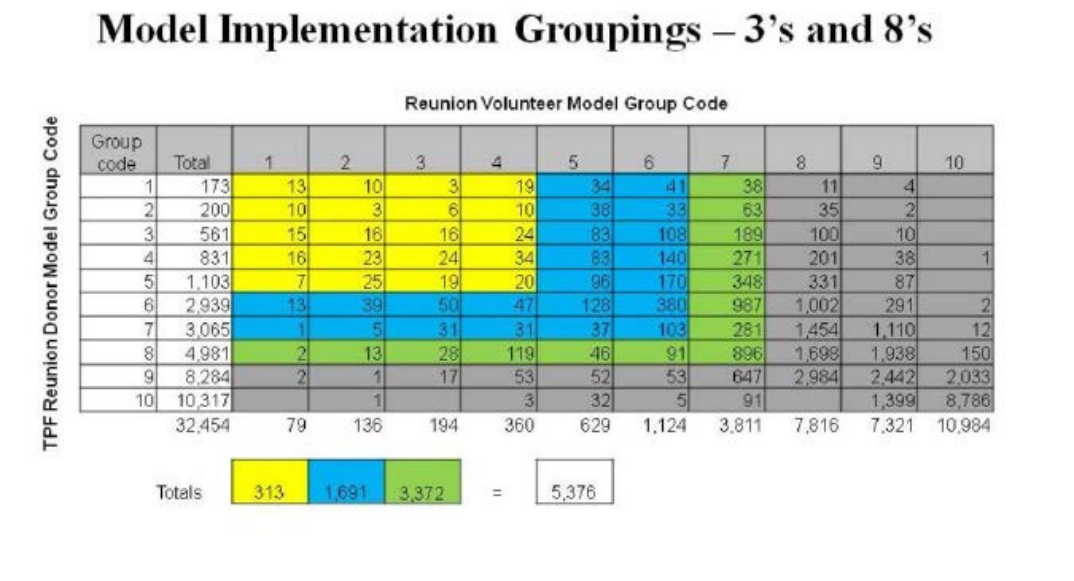
Figure 38. PRLC Reunion Planning Committee Volunteer Model Equation



Each record from this model was also assigned to a group ranging from 1 to 10 depending on the model score, with 1 being the group most likely to volunteer.

Combining the two model groupings as shown in Figure 39, illustrates the process for identifying alumni most likely to volunteer for the Reunion Gift Committee among the 3's and 8's.

Figure 39. Implementation of Reunion Gift Committee Volunteer Model – 3's and 8's



Using the model we can identify alumni most likely to volunteer for the Reunion Gift Committee as shown in Table 35.

Table 35. Target of Reunion Gift Committee Volunteers

Reunion planning committee model groups	Reunion donor model groups	Number of alumni identified in both groups	Percent of all alumni in 3's and 8's identified in both groups
1–4	1–5	313	1%
5–6	6–7	1,691	5%
7	8	3,372	10%
Totals		5,376	17%

Since recruiting Reunion Gift Committee volunteers using the model results would represent a change in approach for existing committee members, we needed some way to validate the effectiveness of these models in combination.

The most recent Alumni Weekend at that time was May 2011. For these reunion classes we could run planning committee and donor models for the 1's and 6's and develop the groupings the same as was done for the 3's and 8's. Because, by this time, we knew who was a volunteer in FY2011, we could determine how well the combination model identified FY2011 volunteers. Both models were applied to the 28,090 alumni in the reunion classes ending in a 1 or 6. Results of the combined models are shown in Figure 40 using the same combination grouping and color coding as for the 3's and 8's.

Figure 40. Application of Reunion Gift Committee Volunteer Models to 1's and 6's

Model Validation – 1's and 6's

Reunion Volunteer Model Group Code

TPF Reunion Donor Model Group Code

Group Code	Total	1	2	3	4	5	6	7	8	9	10
1	113	16	13	9	11	16	25	14	8	1	
2	154	8	15	5	8	30	25	32	29	2	
3	241	8	16	13	18	40	48	42	48	8	
4	556	14	29	12	23	68	113	137	135	25	
5	1,358	6	33	46	33	120	234	307	467	112	
6	2,874	5	36	53	36	122	301	616	1,287	418	
7	3,995	2	18	48	50	85	243	464	1,614	1,452	19
8	3,081		13	32	59	26	87	400	999	1,320	145
9	6,692		7	16	21	22	44	379	1,481	2,991	1,731
10	9,026				15	33	4	79	49	2,061	6,785
	28,090	59	180	234	274	562	1,124	2,470	6,117	8,390	8,680

Totals	336	1,718	2,229	=	4,283

The Penn Fund staff provided the actual list of Reunion Gift Committee volunteers for the reunion classes ending in 1 or 6. These volunteers are placed on the table based on their two model scores as shown in Figure 41. In total there were 310 volunteers. The model placed 215 of the volunteers in the top three tiers, 81 in the first tier, 74 in the second tier and 60 in the third tier. In summary, if the model had been used for FY2011, contacting 15% of the 28,090 alumni would have identified 215 or 68% of the volunteers. This validation provided the necessary confidence to implement it to recruit committee volunteers among the alumni who will be having a reunion in May 2013.

Figure 41. Validation of Reunion Gift Committee Volunteer Model to 1's and 6's

Model Validation – 1's and 6's

Reunion Volunteer Model Group Code

TPF Reunion Donor Model Group Code

Group Code	Total	1	2	3	4	5	6	7	8	9	10
1	31	13	7	2	3	2	4				
2	28	3	5	2	2	6	3	5	2		
3	29	5	6	2	1		4	5	5	1	
4	56	3	12	2		7	10	9	12	1	
5	74	3	6	3	1	6	13	19	17	6	
6	45	2	1	3	1		6	11	15	6	
7	36		1	3		1	1	9	18	3	
8	6				1		1			4	
9	1									1	
10	4									2	2
	310	29	38	17	9	22	42	58	69	24	2

81 74 60 = 215

Percent Alumni Contacted	Percent Volunteers Reached
1%	26%
7%	50%
15%	69%

Summary of Results and Implementation Recommendations

The Penn Fund (TPF) Reunion Gift Committee Volunteer Recruiting model is unique in that it brings together two different models already discussed earlier in this monograph. The purpose of the Alumni Relations Reunion Planning Committee Volunteer Recruiting model is to identify alumni most likely to volunteer to be a planning committee member. The model for Targeting Penn Fund Undergraduate Reunion Year Donors identifies alumni most likely to make a gift to Penn in their reunion year. A combination of a high score in both models defines an alumna or alumnus that would be an ideal Gift Committee volunteer.

The use of a cross-tab of model scores from two existing models to identify likely Gift Committee volunteers represents an innovative way to address the needs of the Penn Fund without building yet another model.

ALUMNI TRAVEL TARGET MARKETING

Penn Alumni Relations manages a comprehensive program of communications, activities, outreach, and services designed to attract, inform, and involve Penn's alumni. According to the Alumni Relations Website, the Alumni Travel Program sponsored by the University of Pennsylvania Alumni Society is designed to combine travel and study in exotic locations in the company of other Penn alumni and faculty. Tours are "sponsored" or provided by outside vendors that typically market the same tour to alumni of various colleges and universities with comparable alumni demographics.

For more than ten years, Alumni Relations has been developing a list of Penn alumni who either have taken Penn-sponsored tours or are interested in such tours. The current marketing mailing list of nearly 6,000 is the foundation of the direct marketing program for Alumni Travel. Tour providers often prefer a target marketing mailing list or email list of between 15,000 and 25,000 alumni. The key question is how to effectively supplement the basic list with quality prospects for alumni tours. In the past, alumni to add to the list were typically chosen by year of Penn degree between 1945 and 1975, for example.

Predictive modeling provides a more comprehensive approach to targeting alumni. In 2003, a predictive model was developed for Alumni Travel marketing. In August 2010, this model was replaced with a more robust model. Using the model structure developed in August 2010, the model was refreshed in May 2012 with updated independent variable values.

Objectives

The objective of this August 2010 project was to use predictive modeling to develop two scoring models that can be used to supplement the basic Alumni Travel marketing lists. One model was for undergraduate alumni and the other was for graduate and professional alumni.

Undergraduate Alumni Scoring Model

A data set consisting of 65,850 living undergraduates with first Penn degrees from 1930 through 1979 was developed, transformed and conditioned. The dependent variable was a binary (1/0) variable representing whether that alumna or alumnus was on the basic Alumni Travel list. Independent variables included both binary (1/0) variables and continuous variables. The variable definitions and statistical results are in Table 36.

Table 36. Multiple Regression Model Results - Undergraduate Alumni

Regression Statistics	
Multiple R	0.220829
R Square	0.048765
Adjusted R Square	0.048577
Standard Error	0.213091
Observations	65,850

Variable Descriptions	Definition	Coefficients
Intercept		-0.050560
Arts and Culture 2004 Survey Travel Interest	1 = yes 0 = no	0.072429
Attended a Homecoming Weekend	1 = yes 0 = no	0.047504
Census 1998 Travel Interest	1 = yes 0 = no	0.037888
Attended Other Events	1 = yes 0 = no	0.032341
Decade of First Penn Degree Before 1990	Decimal values allowed (e.g. 1972=1.8)	0.017056
Attended an Alumni Weekend	1 = yes 0 = no	0.016775
Log_{10} of Lifetime Hard and Soft Commitments	Gifts>$10 = Log_{10} (gifts), $0<Gifts <$10 = 1.0, and No Gifts = 0.0	0.014268
Member of a Student Performing Arts Group	1 = yes 0 = no	0.011031
Member of Any Other Student Activity	1 = yes 0 = no	0.010717
Degree Complete	1 = yes 0 = no	0.009394
Alumni Relatives (Spouse, Children, Parents, Siblings, Other)	1 = yes 0 = no	0.007351
Undergraduate Alumni with a Penn Graduate degree	1 = yes 0 = no	0.005790
Member of Student Fraternity or Sorority	1 = yes 0 = no	0.004674

Graphically, the resulting model equation is shown in Figure 42.

Figure 42. Alumni Travel Undergraduate Model Regression Model

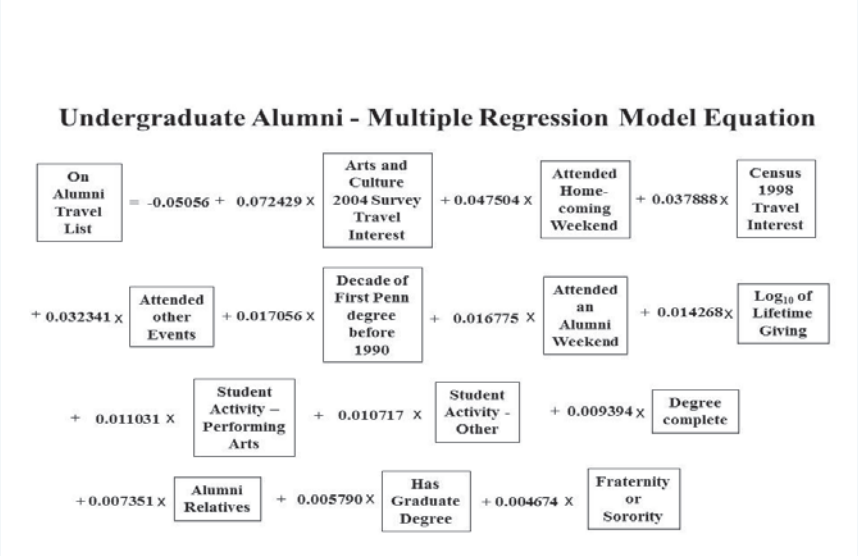
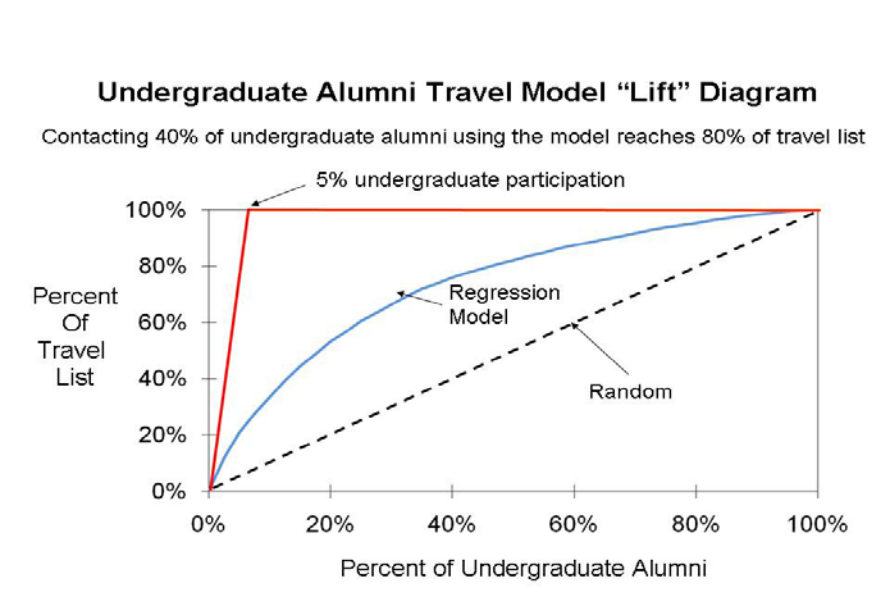


Figure 43 is the Undergraduate Alumni Travel Model lift diagram. Using the Undergraduate Alumni model scores, contacting the 40% of the alumni with the highest scores will identify 80% of the total undergraduate population already on the Alumni Travel List.

Figure 43. Alumni Travel Undergraduate Model Lift Diagram



Graduate and Professional Alumni Scoring Model

A data set consisting of 49,538 living graduate-level alumni with first Penn degrees from 1930 through 1979 was developed, transformed, and conditioned. The dependent variable was a binary (1/0) variable representing whether that alumna or alumnus was on the basic Alumni Travel list. Independent variables included both binary (1/0) variables and continuous variables. The variable definitions and statistical results are in Table 37.

Table 37. Multiple Regression Model Results - Graduate Alumni

Regression Statistics	
Multiple R	0.165401
R Square	0.027358
Adjusted R Square	0.027161
Standard Error	0.161972
Observations	49,538

Variable Descriptions	Definition	Coefficients
Intercept		-0.04565
Attended an Alumni Weekend	1 = yes 0 = no	0.062161
Alumni Census 1998 Travel Interest	1 = yes 0 = no	0.021333
Attended Other Events	1 = yes 0 = no	0.020392
Decades Before 1990	Decimal values allowed (e.g. 1972=1.8)	0.015613
Degree Complete	1 = yes 0 = no	0.013691
Log_{10} of Lifetime Hard and Soft Commitments	Gifts>$10 = Log_{10} (gifts), $0<Gifts <$1 = 1.0, and No Gifts = 0.0	0.009974
Penn Medicine Degree	1 = yes 0 = no	0.008663
Alumni Relatives (Spouse, Children, Parents, Siblings, Other)	1 = yes 0 = no	0.00845
Penn Law Degree	1 = yes 0 = no	0.002586
Had Any Student Activities	1 = yes 0 = no	0.002529

Graphically, the resulting model equation is shown in Figure 44.

Figure 44. Alumni Travel Graduate Model Regression Model

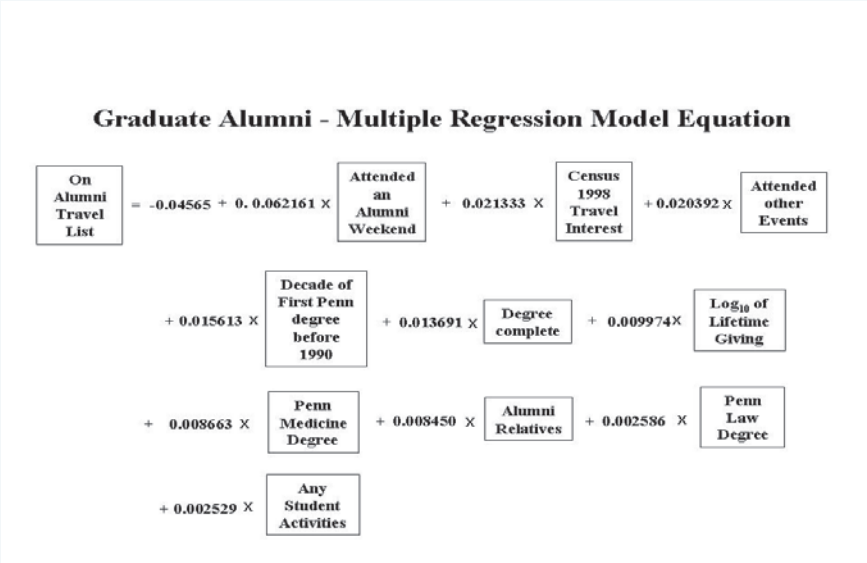
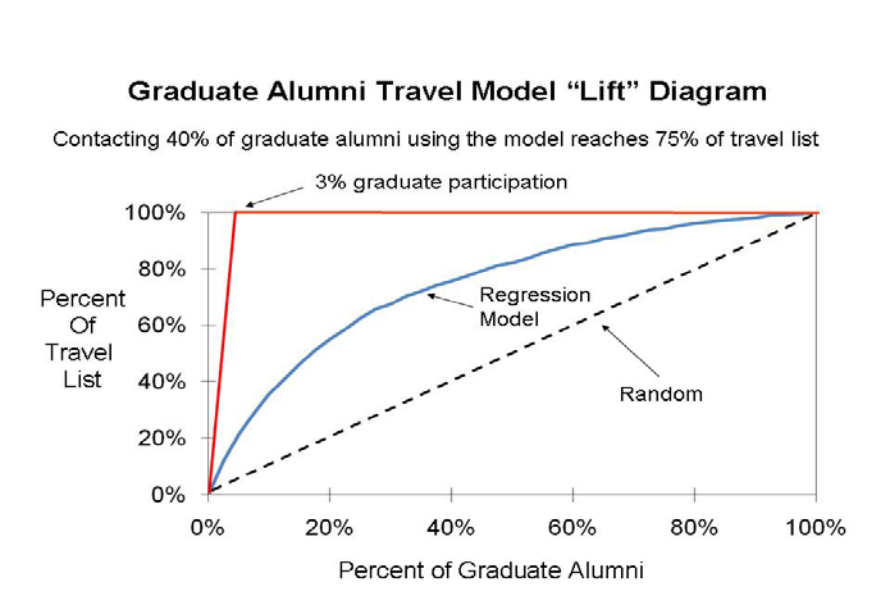


Figure 45 is the Graduate Alumni Travel model lift diagram. Using the Graduate Alumni model scores, contacting the 40% of the alumni with the highest scores will identify 75% of the total graduate alumni population already on the Alumni Travel List.

Figure 45. Alumni Travel Graduate Model Lift Diagram



Scoring Model Groupings

To facilitate using the model to add to the basic Alumni Travel list, alumni were aggregated by their model score into 10 unequal sized groups (See Tables 38 and 39). The group sizes for higher model scores are smaller to allow for more precise selection of list additions. The groups are unequal in size so as not to separate alumni with the same model score into different groups. It is unlikely the alumni with low model scores (groups 7, 8, 9 and 10) will ever be selected to supplement an Alumni Travel mailing list.

Table 38. Scoring Model Grouping – Undergraduate Alumni

Score group	Count	Percent	Maximum model score	Minimum model score
1	3,233	4.9%	0.340073	0.14252
2	3,249	4.9%	0.142484	0.1155
3	3,214	4.9%	0.115493	0.098502
4	3,285	5.0%	0.098498	0.085503
5	6,714	10.2%	0.085499	0.066502
6	6,259	9.5%	0.0665	0.053503
7	6,708	10.2%	0.053499	0.042505
8	6,376	9.7%	0.042495	0.0325
9	13,614	20.7%	0.032498	0.010529
10	13,198	20.0%	0.010484	-0.0318
Total	65,850	100%		

Table 39. Scoring Model Grouping – Graduate Alumni

Score group	Count	Percent	Maximum score	Minimum score
1	2,448	4.9%	0.2071638	0.0765111
2	2,309	4.7%	0.0764907	0.0635002
3	2,336	4.7%	0.0634927	0.0555053
4	2,803	5.7%	0.0554869	0.0485005
5	4,624	9.3%	0.0484993	0.0395008
6	5,409	10.9%	0.0394926	0.0315002
7	4,507	9.1%	0.0314993	0.0255035
8	5,217	10.5%	0.0254990	0.0185004
9	9,560	19.3%	0.0184988	0.0045096
10	10,325	20.8%	0.0044570	-0.0284771
Total	49,538	100%		

For ease of use in implementation, it was also necessary to combine the undergraduate and graduate alumni lists, even though the models and resulting scores are different from each other. The undergraduate and graduate alumni model lists were merged using the prospects list rank number as a percent of the total list length. As a result, undergraduate and graduate alumni who are in the top 5 % of their original lists are also in the top 5% of the combined list (Table 40).

Table 40. Scoring Model Grouping – All Alumni - combined

Score group	All alumni		Undergraduate		Graduate	
	Count	Percent	Count	Percent	Count	Percent
1	5,769	5%	3,292	5%	2,477	5%
2	5,769	5%	3,293	5%	2,476	5%
3	5,770	5%	3,292	5%	2,478	5%
4	5,769	5%	3,293	5%	2,476	5%
5	11,539	10%	6,585	10%	4,954	10%
6	11,539	10%	6,585	10%	4,954	10%
7	11,539	10%	6,585	10%	4,954	10%
8	11,538	10%	6,585	10%	4,953	10%
9	23,078	20%	13,170	20%	9,908	20%
10	23,078	20%	13,170	20%	9,908	20%
Total	115,388	100%	65,850	100%	49,538	100%

Model Implementation

To add to an Alumni Travel list for a particular mailing, the following steps would be followed:

1. Select the basic Alumni Travel list.
2. Select the groups needed to supplement the basic list (additional count needed).
3. Combine the lists and remove duplicate records, deceased, lost, mail keep-offs, and probably those prospects with a preferred mailing address outside the U.S.

Summary of Results and Recommendations

The regression model developed here provides an effective tool for Alumni Relations staff to identify alumni most likely to be interested in participating in an alumni tour. This project provides another example of non-giving predictive modeling.

Different models were developed for undergraduate alumni and graduate and professional alumni. Two other features of this model are the use of a mix of binary and continuous independent variables, and the use of "dummy" 0/1 variables to represent graduate alumni with degrees from specific schools.

The continuous variables are the level of lifetime giving where we use the logarithm transformation to "flatten" the curve as discussed at the beginning of the monograph, and the number of years before 1990 of the first Penn degree of the alumna or alumnus.

Undergraduate Model

The variables in the undergraduate model most likely to predict and interest in Penn Alumni Travel are:

- Decades before 1990 of the first Penn degree
- Log^{10} of Lifetime Hard and Soft Gift commitments

Other variables that predict interest in Alumni Travel are:

- An interest in travel in a response to an Arts and Culture 2004 Survey question
- An interest in travel in a response to a Census 1998 Survey question.

Event participation variables that predict interest in Alumni Travel are:

- Attended a Homecoming Weekend
- Attended an Alumni Weekend
- Attended other events

Academic variables that predict interest in Alumni Travel are:

- Degree complete
- Undergraduate Alumni with a Penn Graduate degree

Student activities that predicts an interest in Alumni Travel are:

- Member of a student performing arts group
- Member of any other student activity
- Member of student fraternity or sorority

Finally having alumni relatives (Spouse, Children, Parents, Siblings, Other) predict interest in Alumni Travel.

Graduate Model

The variables in the graduate model most likely to predict and interest in Penn Alumni Travel are:

- Decades before 1990 of the first Penn degree
- Log^{10} of lifetime hard and soft gift commitments

Another variable that predicts interest in Alumni Travel is an interest in travel in a response to a Census 1998 Survey question.

Event participation variables that predict interest in Alumni Travel are:

- Attended an Alumni Weekend
- Attended other events

Academic variables that predict interest in Alumni Travel are:

- Degree complete (from any Penn school)
- A degree from Penn Medicine
- A degree from Penn Law

Participation in any student activities predict interest in Alumni Travel.

Finally, having alumni relatives (spouse, children, parents, siblings, other) predicts interest in Alumni Travel.

CONSIDERATIONS FOR IMPLEMENTING PREDICTIVE MODELING

Successful implementation of predictive modeling project results is critical. Without it, the project remains a mere analytical exercise. A good approach is to plan for implementation as a part of business needs identification and problem definition.

In this section we will explore some strategies for implementation and tactics for data management and model updating.

Model Output - Individual Scores and Cohorts

The output of the predictive modeling projects in this monograph is a database where each dataset record (person) has a model score. The model scores are produced by applying the regression equation to each person's input variables (independent or explanatory variables). An individual's model score is only useful, however, in comparison with other people and their model scores. For this reason in each project, we sort the entire dataset by model score and create groups (quartiles, quintiles, or deciles depending on the number of groups). Each record now has a score and a model group number – the lower the group number, the higher the model score.

The groups are unequal in size so as not to separate people with the same model score into different groups. Because the models have binary independent variables, there are usually distinct step changes or breaks in the sorted model scores.

Implementation Strategies – Using Direct Mail vs. Email Marketing

Usually the implementation of model results focuses marketing efforts on those people with the highest scores. These people are most likely to exhibit the response behavior we are seeking (becoming a donor, volunteering, or booking an alumni tour). This decision, however, is mostly driven by costs associated with a particular marketing tactic. Direct mail costs, for example, typically consist of fixed costs for design, printing, etc. and variable costs that are related to the number of pieces printed and mailed. In this situation, focusing on people with the highest model scores produces good results in terms of the behavior we desire at lower costs than contacting all prospects.

Marketing using email, on the other hand, has fixed costs associated with design and preparation, but distribution costs are small and mostly constant regardless of the size of the email list. In this situation, we would email our marketing appeal to the entire target audience because the costs are virtually the same as targeting a smaller group using the model. With an email appeal we can measure the response rate for each predictive model cohort and assess the model validity. Once a person responds to an email appeal, their re-calculated model score will likely increase such that in the future they may be in a cohort that receives direct mail as well.

Implementation Strategies – Cohorts with Lower Model Scores

Although we focus our direct mail, volunteer, and other marketing resources more on the groups with higher predictive model scores, we don't completely ignore groups with lower scores. We have already described how we would market by email to these groups with lower scores, but might not send them a USPS mail piece.

The objective of the predictive model for Targeting Undergraduate Non-Reunion Year Donors described earlier in the monograph was to prioritize lapsed and never givers. Although the model scores and cohorts were based on all alumni in a graduating class who were not currently having a reunion, people with the highest scores were targeted in other ways. Here the issue was how to identify the best prospects among the less consistent givers.

In general, with donor participation models, giving variables are the most important (they have the largest coefficient values in the regression equation). Individuals with the highest scores, therefore, usually exhibit giving in prior time periods (as measured by the independent variables). Lapsed and never givers, if they have high scores, it is for other reasons. They have alumni relatives (spouse, children, parents, siblings, etc.), were involved as a student (sports, fraternity/sorority, student activities), have recently attended events, etc. Consideration of homogeneous segments of lapsed and never givers with similar independent variable characteristics (1's and 0's) may help in developing targeted re-engagement messaging. Furthermore, the database with model variable values and model scores is useful for experimentation because of the ability to draw random samples of segment members. This is especially important, if the cost of marketing experimentation is high (e.g., direct mail vs. email marketing.)

Storing Model Group Codes

When a predictive modeling application produces results of interest across an entire organization, it is necessary to find a way to store Group Codes in the database used for development and alumni relations operations. Issues such as mass record updating and user training become critical for implementation. In other situations, predictive modeling results can be stored in a data warehouse or with the business analytics organization.

Model Updating Considerations

After a predictive model has been used by an organization for a while, an important consideration is how and when to update. The timing depends on the dynamics of the environment in which the model is used. Some industry applications may update a person's model score any time there is a "trigger" event involving that individual (a purchase, complaint, marriage, new child, etc.).

In the Annual Giving and Alumni Relations applications discussed in the monograph, an annual update of the independent variables and a recalculation of the model scores using the equation developed earlier should be sufficient. Some variables don't change (e.g. year of first Penn degree, member of a fraternity or sorority, played a sport). Others may change infrequently (e.g. now married to an alumna or alumnus). Some variables change within a year (e.g. donor/non-donor status, attending an event).

When a model has been used for three to five years, a new modeling effort may be needed to see if other variables have become important or if the weighting of explanatory variables has changed. This rerun will produce a new model equation. An example is the Alumni Travel Target Marketing Model. First developed in 2003, it was rerun completely in 2010 producing slightly different equations for undergraduate and graduate alumni. In 2012, model scores were updated using new data and the 2010 version of the equations.

WORKS CITED

(1) Berry, Michael J. A., and Gordon S. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* Second ed. Indianapolis, IN: Wiley Pub., 2004. Print.

(2) Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire: Graphics, 1983. Print.