Journal of Contemporary Management
Submitted on 20/07/2018
Article ID: 1929-0128-2018-04-14-15
Fangyao Liu, Xixi Feng, and Qinge Ouyang

Factors Exploration on Alumni Donation: A Case Study of Creighton University

Dr. ***Fangyao Liu*** (Correspondence Author)
Department of Information Systems and Quantitative Analysis,
University of Nebraska at Omaha, Nebraska, U.S.A.
E-mail: fangyaoliu@unomaha.edu

Xixi Feng
College of Business Administration,
University of Nebraska at Omaha, Nebraska, U.S.A.

Qinge Ouyang
Department of Advancement Services
Creighton University, Nebraska, U.S.A.

***A*bstract:** Donation from university alumni contributes a lot to undergraduate and graduate level student' s academic success. Alumni donation is very supportive for easing the financial burden of attending college for both prospective and current students. This research paper is going to explore which kinds of factors will affect alumni's giving. The main methods are factor analysis, logistic regression and generalized linear model (GLM), other steps also play valuable and significant roles in the modeling framework. This research paper uses a real university case. Some factors have been discovered as positive impact factors, all the positive factors could influence alumni donation, which further will impact university alumni engagement, events planning and other missions. Some other insights have been also discovered in the conclusion section.

Keywords: Alumni; Donation; Data modeling; Factor analysis; Logistic regression; Generalized linear model (GLM)

JEL Classification: C02, C30, C38

1. Introduction

Creighton University is a private, coeducational, Jesuit, Roman Catholic university in Omaha, Nebraska, United States. Founded by the Society of Jesus in 1878, the school is one of 28-member institutions of the Association of Jesuit Colleges and Universities. The university is accredited by the North Central Association of Colleges and Secondary Schools. Creighton is Nebraska's largest private religious university.

Advancement Services at Creighton University is primarily responsible for data reporting and analysis in University Relations, a very important division of the university with a mission to secure maximum financial and volunteer support for the University by conducting fund-raising programs that focus on annual giving, capital giving and deferred giving. To better serve the needs of programs development and the effectiveness of gifts solicitation, the identification of a wide range

of factors that may impact constituents on giving to the University is especially crucial to the daily operations of the division. This research paper intends to embark this effort and hopes the analysis will be helpful in decision making and better targeting the constituents. So, our research question is: Which kinds of factors impact Creighton Alumni's giving probability?

Specifically, the main idea is to explore the factors that have impacts on the probability of making a gift to the university from the alumni.

By 2003, alumni donations across all US universities have become on average the largest source of donations and in 2005 have risen to 26.6 percent of university donations (Gottfried, 2006). There are many macro and micro level factors could impact alumni's giving probability. Such as, GDP growth rate, employment rate, and others. Also, gender could be another factor. Although some researchers conclude that it is not a significant factor. The covariance regression model results indicate lack of statistically significant difference between gift-giving women and men (Okunade, 1994; Sun, *et al*., 2007; Brooker and Klastorin, 1981).

2. Methodology

Data modeling is keeping making a difference for insight discovering. This paper is to explore which factors will contribute to alumni' s donation giving. So, factor analysis associated with other methods are the main activities in this designed modeling framework. For results verification, training and the testing split will also be adopted.

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors (Cattell, 1952). Basically, it is a process to discover which factors can explain the main effect. For example, for a group of 20 variables, 5 variables are enough to explain the effects. For some cases, one factor can be a single variable, for others, one factor can be a few variables that share the same information.

The factor analysis model can be written algebraically as follows. If you have p variables X_1, X_2 , . . . , X_p measured on a sample of n subjects, then variable I can be written as a linear combination of m factors F_1, F_2, . . . , F_m where, as explained above, m is smaller than p. Thus,

$$X_i = ai_1 \times F_1 + ai_2 \times F_2 + \ldots + ai_m \times F_m + e_i \qquad (1)$$

where the "ai" are the factor loadings (or scores) for variable i and e_i are the part of variable Xi that cannot be explained by the factors.

The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage (Agresti, 2002). Logistic regression will be a great choice when the dependent variable is categorical (Peng, 2008; Saldana, 1984; Tatham, 1998).

The model can be expressed as

$$\text{Logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k \qquad (2)$$

where p is the probability of the presence of the characteristic of interest. The logit transformation is defined as the logged odds:

Odds = p/(1-p) =
(probability of presence of characteristic)/(probability of absence of characteristic)

Generalized linear model (GLM) is flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution (Hastie and Tibshirani, 1990; Peng, 2011). Unlike the OLS regression, the GLM is more useful when the model not meet the normal distribution assumption (Thompson, 2004; Thurstone, 1947).

This following modeling framework (Figure 1) is designed to show the steps of transferring research question to a conclusion:

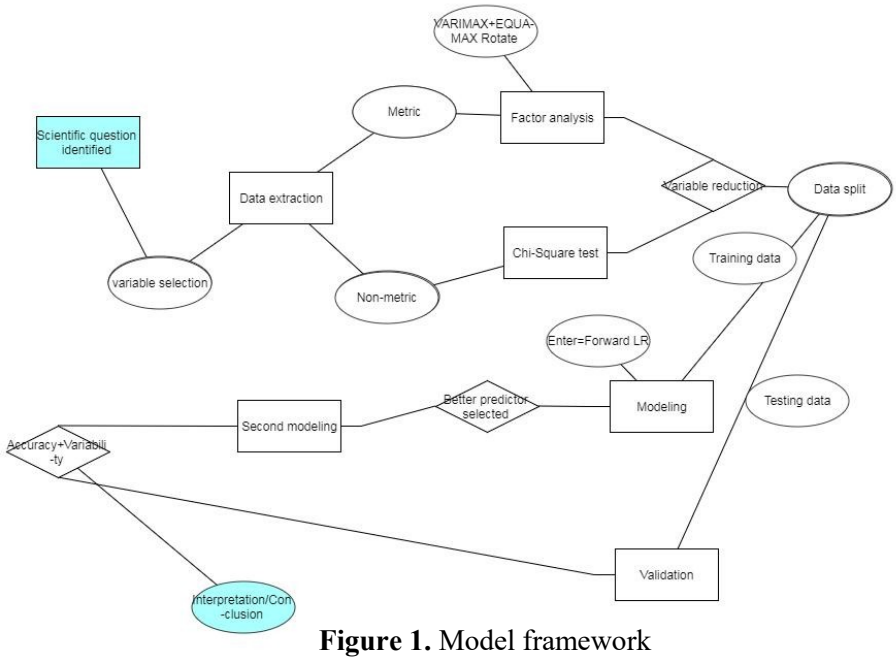

Figure 1. Model framework

3. Data and Definition

Based on the research question, 50 variables were selected out of 60 from an internal database, 3 external variables from US Census Bureau were added into this set. The detailed data definition is available for request if interested. These 50 variables describe the constituents primarily from 5 aspects:

- Biographic information: age, graduation year, school attended, number of degrees obtained from Creighton, highest degree from Creighton, distance to Creighton, employment status, number of active phones, emails, addresses recorded, number of record types
- Giving history: lifetime giving statistics (amount, quantity, average amount, frequency, largest, smallest gift amount), number of emails, letter, and phone appeals received
- Affiliation with Creighton: involvement in number of committees, number of a affiliations, number of student activities, number of volunteer activities, and number of events participated, event recency, student activity recency, volunteer activity recency, committee recency, affiliation recency, award/honor received, sports participated

- Giving capacity: Reeher Networth value
- Demographic information: neighborhood median household income, neighborhood average household size, and neighborhood average family size

Considering the purpose and scope of this research to reduce the correlations among the independent variables, some variables from giving history section were excluded from the analysis, and only responses to appeals, number of emails, letter, and phone appeals received were kept.

Missing data were handled carefully to keep the data integrity and consistency (Norusis, 2008). Specifically,

- Data missing in median household income were filled with 2013 US median household income value;
- Data missing in Average family size were filled with 2013 US average family size;
- Data missing in Average household size were filled with 2013 US average family size;
- Data missing in Age were filled with this formula based on their graduation year and the median graduation age in the degree level the alumni obtained: Age = 2015-grad year + median grad age (Table 1).

Table 1. Median graduation age in the degree level

Degree Level	Median Grad Age
Certificate	22
Bachelor	23
Undergraduate-Non-Degree	24
Doctor	27
Master	30
Associate	30
Graduate-Non-Degree	42
PhD	45
Honor	67

A summary of the data set is shown in Table 2 below.

The percentage of missing data in this data set is (839+33) / (68603*31) = 0.04%. The missing values exist in Reeher Networth and the Distance to CU, which is caused by incompleteness of data collection and will not be a concern in this analysis.

Table 2. Data set summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
RECORD_TYP_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
GENDER * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
AGE * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
YEARS_GRAD * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
NBR_OF_DEGREE * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
NBR_ACT_EMAIL * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
NBR_ACT_PHONE * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
NBR_ACT_ADDR * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
DISTANCE_CU * DONOR_IND	67764	98.8%	839	1.2%	68603	100.0%
NETWORTH_2015 * DONOR_IND	68570	100.0%	33	0.0%	68603	100.0%
RELATIONS_NBR * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
NBR_OF_COMM * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
NBR_OF_AFFILIATION * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
STUDENT_ACT_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
VOLUNTEER_ACT_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
EVENT_PARTICIPATION * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
RECENT_EVENT_DT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
RECENT_STUACT_DT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
RECENT_VOL_DT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
RECENT_AFFL_DT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
RECENT_COMM_DT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
MAIL_APPEALS_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
EMAIL_APPEALS_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
TEL_APPEALS_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
AWARD_HONOR_REC * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
AWARD_HONOR_CNT * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
SPORTS_PART * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
HC01_VC21_avg_hhsize * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
HC01_VC22_avg_familysize * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%
HD01_VD01_Median_household_income * DONOR_IND	68603	100.0%	0	0.0%	68603	100.0%

4. Analysis Employed

The majority of variables in this dataset are interval (metric), only gender and marital status are categorical. The very first step in the analysis process is to reduce or group the possible factors. For metric variables, factor analysis is employed; for categorical variables, the chi-square test is used to identify the possible correlations to the giving/not giving behavior. After the possible factors are identified from the first step, logistic regression is used to predict the likelihood of giving. The primary reasons for using logistic regression in this research are because:

- The research question is to know the probability of giving based on the set of factors and the dependent variable is binary (donor or non-donor)
- The variables are a mixture of metric and non-metric variables

- The majority of variables are not normally distributed and the outliers cannot be excluded due to the nature of data collection (Table 3) .

Table 3. Test of normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
RECORD_TYP_CNT	.507	67757	.000
AGE	.063	67757	.000
YEARS_GRAD	.080	67757	.000
NBR_OF_DEGREE	.520	67757	.000
NBR_ACT_EMAIL	.273	67757	.000
NBR_ACT_PHONE	.251	67757	.000
NBR_ACT_ADDR	.539	67757	.000
DISTANCE_CU	.208	67757	.000
NETWORTH_2015	.239	67757	.000
RELATIONS_NBR	.220	67757	.000
NBR_OF_COMM	.529	67757	.000
NBR_OF_AFFILIATION	.524	67757	.000
STUDENT_ACT_CNT	.426	67757	.000
VOLUNTEER_ACT_CNT	.532	67757	.000
EVENT_PARTICIPATION	.388	67757	.000
RECENT_EVENT_DT	.304	67757	.000
RECENT_STUACT_DT	.078	67757	.000
RECENT_VOL_DT	.461	67757	.000
RECENT_AFFL_DT	.400	67757	.000
RECENT_COMM_DT	.385	67757	.000
MAIL_APPEALS_CNT	.090	67757	.000
EMAIL_APPEALS_CNT	.127	67757	.000
TEL_APPEALS_CNT	.091	67757	.000
AWARD_HONOR_REC	.541	67757	.000
AWARD_HONOR_CNT	.535	67757	.000
SPORTS_PART	.540	67757	.000
HC01_VC21_avg_hhsize	.062	67757	.000
HC01_VC22_avg_familysize	.076	67757	.000
HD01_VD01_Median_household_income	.095	67757	.000

a. Lilliefors Significance Correction

Based on the assumptions, logistic regression is the most appropriate technique in this case.

5. Analysis and Outcomes

5.1 Variables Extraction-Chi-square test

The dependent variable is Donor-ind. In order to know whether there are correlations between Donor-ind and gender, Chi-Square test is performed on these two independent variables in SPSS through Analyze, Descriptive Statistics, Crosstabs:

```
CROSSTABS
/TABLES=GENDER MARTIAL STATUS BY DONOR IND
/FORMAT=AVALUE TABLES
```

```
/STATISTICS=CHISQ CC CORR /CELLS=COUNT ROW
COLUMN TOTAL /COUNT ROUND CELL.
```

Below is the output of the Chi-Square tests (Table 4):

Table 4. Chi-square tests

	Value	d.f.	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-square	370.473 [a]	1	0.000		
Continuity Correction [b]	370.179	1	0.000		
Likelihood Ratio	370.852	1	0.000		
Fisher's Exact Test				0.000	0.000
No. of Valid Cases	68603				

a. 0 cells (0.0%) have expected count less than 5; The minimum expected count is 15829.87.

b. Computed only for 2×2 table.

Noticed that p-value in the test is less than $\alpha = 0.05$, which means that gender is correlated to Donor Ind. So, we will keep the variable for the next step of the analysis.

5.2 Variables Extraction-factor Analysis on Metric Variables

Factor analysis is performed to extract metric variables for next step of the analysis. Since it is uncertain whether there are correlations among the independent variables, two rotation methods were used to compare the factors extracted in both ways: VARIMAX Rotate (Table 5) and EQUAMAX Rotate (Table 6).

Table 5. VARIMAX rotate results

KMO and Bartlett's Test [a]		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.656
Bartlett's Test of Sphericity	Approximate Chi-square	267136.45
	d.f.	378
	Sig.	0.000

a. Only cases in which NONOR_IND = 1 are used in the analysis phase.

Table 6. EQUAMAX rotate results

KMO and Bartlett's Test [a, b]		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.656
Bartlett's Test of Sphericity	Approximate Chi-square	267136.45
	d.f.	378
	Sig.	0.000

a. Only cases in which NONOR_IND = 1 are used in the analysis phase.

b. Results in Tables 5 and 6 are the same; And they should be the same.

The factor analysis results show that only one variable may be removed in the next step analysis: Nbr_act_phone was not loaded at all. KMO test is 0.656, which means that it is appropriate to use factor analysis in this case, although it is not ideal. Since factor analysis is employed here solely for variable reduction, we will use the results as a reference for the logistic

regression. Bartlett's test also shows that the variables entered are not highly correlated, which is good to use these variables in the next step. One thing we have to be careful is that the cross loading among the variables. NBR_OF DEGREE, NBR_ACT_EMAIL, RECENT_AFFL_DT, RECENT_ COMM_DT, and MAIL_APPEALS_CNT were cross loaded. But again, since our purpose here is to extract the variables, we will keep the cross loaded variables for next step.

5.3 Training and modeling

After the factor analysis and chi-square test, there are still 27 variables left. We take these variables into logistic regression. Before the logistic regression is run, the dataset is split into two parts: 50% for training, and 50% for testing. Now we proceed with Binary Logistic Regression in SPSS.

5.3.1 Method 1: Entry

First, we use enter method to include all remained independent variables from previous steps:

```
LOGISTIC REGRESSION VARIABLES DONOR_IND
  /METHOD=ENTER RECORD_TYP_CNT MARTIAL_STATUS GENDER AGE
YEARS_GRAD NBR_ACT_EMAIL NBR_ACT_ADDR
DISTANCE_CU NETWORTH_2015 RELATIONS_NBR NBR_OF_COMM
NBR_OF_AFFILIATION VOLUNTEER_ACT_CNT
EVENT_PARTICIPATION RECENT_EVENT_DT RECENT_STUACT_DT
RECENT_VOL_DT RECENT_AFFL_DT RECENT_COMM_DT
APPEAL_REPS_CNT MAIL_APPEALS_CNT EMAIL_APPEALS_CNT
TEL_APPEALS_CNT AWARD_HONOR_REC AWARD_HONOR_CNT
SPORTS_PART HC01_VC21_avg_hhsize HC01_VC22_avg_familysize
HD01_VD01_Median_household_income
    /CONTRAST (GENDER)=Indicator
    /CONTRAST (MARTIAL_STATUS)=Indicator
    /SAVE=PRED PGROUP
    /PRINT=GOODFIT ITER(1) SUMMARY CI(95)
    /CRITERIA=PIN(0.1) POUT(0.2) ITERATE(20) CUT(0.5).
```

With this method, the output is shown below in Table 7:

Table 7. Results of the entry method

Model Summary				Hosmer and Lemeshow Test		
Step	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2	Chi-square	d.f.	Sig.
1	33172.850 [a]	0.332	0.443	53.858	8	0.000

a. Estimation terminated at iteration number 7 because parameter estimates changed less than 0.001.

The model summary table shows that the model is important. But the variability is explained by the model is not very good, only up to 44.3%. We noticed that in the Variables in the Equation table, GENDER(1), AVG_FAMILYSIZE, AVG_HHSIZE, AWARD_HONOR_REC, RECENT_ VOL_DT, RECENT_AFFL_DT, RECENT_COMM_DT are not significant in the model; in addition to that, Median_household_income, DISTANCE_CU, NETWORTH_2015's beta values are close to 0, which means they do not really have an effect on the dependent variable. So, these variables can actually be removed from the model.

Our final first model equation can be written as:

Ln(Oddsdonor-or-nondonor) = - 3.934 + 0.283 * RECORD_TYP_CNT + 0.21* AGE - 0.24 * YEARS_GRAD+ 0.285 * NBR_ACT_EMAIL + 0.236 * NBR_ACT_ADDR + 0.093 * RELATIONS_NBR + 0.626 * NBR_OF_COMM + 0.726* NBR_OF_ AFFILIATION + 0.767* VOLUNTEER_ACT_CNT + 0.151* EVENT_PARTICIPATION - 0.045* RECENT_EVENT_DT + 0.031* RECENT_STUACT_DT + 0.046* MAIL-APPEALS-CNT + 0.011* EMAIL-APPEALS-CNT + 0.016*TEL-APPEALS-CNT + 0.866*AWARD-HONOR-CNT + 0.190*STUDENT-ACT-CNT

With this model, 75.9% of constituents were grouped into the right categories (Table 8).

Table 8. Classification results of the entry method [a]

	Observed		Predicted		
			DONOR_IND		Percentage Correct
			0	1	
Step 1	DONOR_IND	0	13921	3826	78.4
		1	4343	11760	73.0
	Overall Percentage				75.9

a. The cut value is 0.500.

5.3.2 Method 2: Forward LR

Since the independent variables extraction process is not so effective that we still have plenty of them remained, we are going to try another method in logistic regression, which is Forward LR. Forward LR Stepwise is a variable selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates. We are hoping with this method; additional independent variables can be excluded while keeping a similar level of classification accuracy.

The output with Forward LR method is shown below (Table 9):

Table 9. Results of forward LR method

Model Summary				Hosmer and Lemeshow Test		
Step	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2	Chi-square	d.f.	Sig.
21	33172.850 [a]	0.332	0.443	67.008	8	0.000

a. Estimation terminated at iteration number 6 because parameter estimates changed < 0.001.

Model Equation can be written as:

Ln(Oddsdonor - or - nondonor) = - 3.905 + 0.279*RECORD_TYP_CNT + 0.021*AGE - 0.024 * YEARS_GRAD + 0.288 * NBR_ACT_EMAIL + 0.236 * NBR_ACT_ADDR + 0.092 * RELATIONS_ NBR +0.541 * NBR_OF_ COMM + 0.858 * NBR_OF_AFFILIATION + 0.788 * VOLUNTEER_ACT_CNT + 0.142 * EVENT_PARTICIPATION - 0.048 * RECENT_EVENT_DT + 0.031 *

RECENT_STUACT_DT + 0.045* MAIL_APPEALS_CNT + 0.011 * EMAIL_APPEALS_CNT + 0.016 * TEL_APPEALS_CNT + 0.654 * AWARD_HONOR_CNT - 0.168 * HC01_VC22_ avg_familysize + 0.189 * STUDENT_ACT_CNT

Figure 10. Classification results of forward LR [a]

	Observed		Predicted		
			DONOR_IND		Percentage Correct
			0	1	
Step 21	DONOR_IND	0	13911	3836	78.4
		1	4331	11772	73.1
	Overall Percentage				75.9

a. The cut value is 0.500.

We see that the outcomes of these two methods are similar (Table 10), except that the weights for each independent variable are slightly different.

6. Remodeling

6.1 Repeat method 1 with fewer predictors

To confirm that the independent variables that have a close to 0 beta values can be safely removed from the model, we use Enter method again to run the logistic regression. But this time, we run the regression with the variables only in Variables in the Equation (Forward LR method) table whose B value in the table is not .000 (Table 11; Table 12).

```
LOGISTIC REGRESSION VARIABLES DONOR_IND
  /METHOD=ENTER RECORD_TYP_CNT AGE YEARS_GRAD NBR-
  _ACT_EMAIL NBR_ACT_ADDR RELATIONS_NBR NBR_OF_COMM
NBR_OF_AFFILIATION VOLUNTEER_ACT_CNT EVENT_PARTICIPATION
RECENT_EVENT_DT RECENT_STUACT_DT
      MAIL_APPEALS_CNT EMAIL_APPEALS_CNT TEL_APPEALS_CNT
AWARD_HONOR_REC AWARD_HONOR_CNT
      HC01_VC22_avg_familysize STUDENT_ACT_CNT
    /SAVE=PRED PGROUP
    /PRINT=GOODFIT ITER(1) SUMMARY CI(95)
    /CRITERIA=PIN(0.1) POUT(0.2) ITERATE(20) CUT(0.5).
```

Table 11. Remodeling results

Model Summary				Hosmer and Lemeshow Test		
Step	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2	Chi-square	d.f.	Sig.
1	33725.053 [a]	0.329	0.439	67.999	8	0.000

a. Estimation terminated at iteration number 6 because parameter estimates changed < 0.001.

Table 12. Remodeling classification results [a]

	Observed		Predicted		
			DONOR_IND		Percentage Correct
			0	1	
Step 1	DONOR_IND	0	14130	3949	78.2
		1	4415	11766	72.7
	Overall Percentage				75.6

a. The cut value is 0.500.

So, with one less predictor in the model, AWARD HONOR REC, our penalty is only 0.3%. Considering the efforts and resources it will take to collect data for each measurement, it is desirable to use the model with fewer predictors, but, a similar power of prediction. With 75.6% of cases are classified into the right groups, the model is satisfactory. Although the variability explained is only 43.9%, the model is still significant, according to the Hosmer and Lemeshow Test. Therefore, in this case, we will go with the last option we have here. Our final model equation is now:

Ln(Oddsdonor - or - nondonor) =
- 4.01 + 0.256 * RECORD_TYP_CNT + 0.02 * AGE - 0.026 * YEARS_GRAD + 0.289 * NBR_ACT_EMAIL + 0.259 * NBR_ACT_ADDR + 0.098 * RELATIONS_NBR + 0.563 * NBR_OF_COMM + 0.956 * NBR_OF_AFFILIATION + 0.812 * VOLUNTEER_ACT_CNT + 0.146 * EVENT_PARTICIPATION - 0.047 * RECENT_EVENT_DT + 0.034 * RECENT_STUACT_DT + 0.046 * MAIL_APPEALS_CNT + 0.011 * EMAIL_APPEALS_CNT + 0.016 * TEL_APPEALS_CNT - 0.233 * AWARD _HONOR_REC + 0.869 * AWARD_HONOR_CNT - 0.12 * HC01_VC22_avg_- familysize + 0.197 * STUDENT_ACT_CNT

6.2 Testing

To examine the validity and stability, the final model generated above is now used to test the other half of the data, and the model testing outputs are shown below (Table 13):

Table 13. Remodeling testing outputs

Model Summary				Hosmer and Lemeshow Test		
Step	-2 Log likelihood	Cox & Snell R^2	Nagelkerke R^2	Chi-square	d.f.	Sig.
1	33878.066 [a]	0.327	0.436	51.180	8	0.000

Classification Table [b]	Observed		Predicted		
			DONOR_IND		Percentage Correct
			0	1	
Step 1	DONOR_IND	0	14397	3884	78.8
		1	4462	11600	72.2
	Overall Percentage				75.7

a. Estimation terminated at iteration number 6 because parameter estimates changed < 0.001.

b. The cut value is 0.500.

The prediction accuracy and variability explained by the model seem to be very consistent with the results during the training phase. However, a couple of predictors AWARD-HONOR-CNT and YEARS-GRAD got dropped in the testing phase. The removal of YEARS GRAD is easily understand-able because it was not that significant in the final model. Its beta value was only -0.026 and odds ratio is 0.975, which means it does not really have a significant impact on the dependent variable outcome. The exclusion of AWARD-HONOR-CNT is questionable here since it used to be important in predicting the outcomes. It might be because of the sampling issue. Further analysis would need to be done to find out why. Other than this, the model is overall satisfactory and consistent throughout training the testing phases. And the model is significant in predicting the likelihood of giving.

6.3 GLM method

The current model gives us the valid variables corresponding its coefficients. Next, a GLM method will be used. Only selected variables will be used as input variables. The GLM result for all the selected variables is below (Table 14). It shows that the RECORD TYP CNT is not significant.

Table 14. GLM model results with the variable RECORD_TYP_CNT

Dependent Variable: DONOR_IND

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	.193	.168	1.149	.250	-.136	.523
AGE	.003	.000	10.895	.000	.002	.003
YEARS_GRAD	-.003	.001	-3.062	.002	-.005	-.001
NBR_ACT_EMAIL	.052	.002	26.901	.000	.048	.056
NBR_ACT_ADDR	.039	.007	5.851	.000	.026	.052
RELATIONS_NBR	.015	.000	30.847	.000	.014	.016
NBR_OF_COMM	.041	.000	7.265	.000	.030	.052
NBR_OF_AFFILIATION	.060	.014	4.230	.000	.032	.088
VOLUNTEER_ACT_CNT	.070	.010	6.755	.000	.050	.090
EVENT_PARTICIPATION	.008	.001	7.655	.000	.006	.011
RECENT_EVENT_DT	-.009	.000	-22.001	.000	-.009	-.008
RECENT_STUACT_DT	.006	.001	5.217	.000	.003	.008
MAIL_APPEALS_CNT	.008	.000	53.291	.000	.008	.009
EMAIL_APPEALS_CNT	.002	.000	18.403	.000	.002	.002
TEL_APPEALS_CNT	.003	.000	15.214	.000	.003	.003
AWARD_HONOR_REC	.158	.022	7.230	.000	.115	.201
AWARD_HONOR_CNT	-.060	.019	-3.127	.002	-.097	-.022
HC01_VC22_avg_familyzise	-.031	.006	-5.575	.000	-.042	-.020
STUDENT_ACT_CNT	.025	.002	15.658	.000	.022	.029
[RECORD_TYP_CNT=1]	-.306	.167	-1.836	.066	-.633	.021
[RECORD_TYP_CNT=2]	-.276	.167	-1.657	.098	-.603	.051
[RECORD_TYP_CNT=3]	-.179	.167	-1.070	.285	-.507	.149
[RECORD_TYP_CNT=4]	-.012	.173	-0.067	.947	-.350	.327
[RECORD_TYP_CNT=5]	0 [a]	.	.	.	.	.

Then, another model result which excluded the RECORD TYP CNT variable is shown below in Table 15. Based on the results in Tables 14 and 15, each corresponding coefficient has the same negative or positive sign, which means the results are valid.

Table 15. GLM model results without the variable RECORD_TYP_CNT

Dependent Variable: DONOR_IND

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-.118	.021	-5.652	.000	-.159	-.077
AGE	.003	.000	12.178	.000	.003	.004
YEARS_GRAD	-.004	.001	-3.505	.000	-.006	-.002
NBR_ACT_EMAIL	.053	.002	27.435	.000	.049	.057
NBR_ACT_ADDR	.038	.007	5.700	.000	.025	.051
RELATIONS_NBR	.016	.000	32.582	.000	.015	.017
NBR_OF_COMM	.040	.000	7.103	.000	.029	.051
NBR_OF_AFFILIATION	.076	.014	5.383	.000	.048	.104
VOLUNTEER_ACT_CNT	.074	.010	7.095	.000	.053	.094
EVENT_PARTICIPATION	.010	.001	8.807	.000	.008	.012
RECENT_EVENT_DT	-.009	.000	-21.640	.000	-.009	-.008
RECENT_STUACT_DT	.006	.001	5.439	.000	.004	.008
MAIL_APPEALS_CNT	.008	.000	52.984	.000	.008	.009
EMAIL_APPEALS_CNT	.002	.000	18.621	.000	.002	.002
TEL_APPEALS_CNT	.003	.000	14.758	.000	.003	.003
AWARD_HONOR_REC	.151	.022	6.905	.000	.108	.193
AWARD_HONOR_CNT	-.050	.019	-2.609	.009	-.087	-.012
HC01_VC22_avg_familyzise	-.031	.006	-5.666	.000	-.042	-.021
STUDENT_ACT_CNT	.025	.002	15.539	.000	.022	.028

The new final model is to use the GLM coefficient results.

7. Interpretation

Now let's take a closer look at the final model and find out how the variables we selected impact the giving likelihood of the constituents. Again, the model equation adopted in this research is specified as:

Ln(Oddsdonor - or - nondonor) = - 0.118 + 0.003*AGE - 0.004*YEARS_GRAD + 0.053*NBR_ACT_EMAIL + 0.038*NBR_ACT_ADDR + 0.016*RELATIONS_NBR + 0.04*NBR_OF_COMM + 0.076*NBR_OF_AFFILIATION + 0.074* VOLUNTEER_ ACT_CNT + 0.1*EVENT_PARTICIPATION - 0.009 * RECENT_EVENT_DT + 0.006* RECENT_STUACT_DT + 0.008 * MAIL_APPEALS_CNT+0.002* EMAIL_APPEALS_CNT + 0.003 * TEL_APPEALS_CNT +0.151* AWARD_HONOR_REC - 0.05*AWARD_HONOR_CNT + 0.31* HC01_VC22_avg_familysize + 0.025 * STUDENT_ACT_CNT

Overall, the increase of these variables will lead to an increase in the likelihood of giving: AGE, NBR_ACT_EMAIL, NBR_ACT_ADDR, RELATIONS_NBR, NBR_OF_AFFILIATION, VOLUNTEER_ACT_CNT, EVENT_PARTICIPATION, NBR_OF_COMM, RECENT_ STUACT_DT, MAIL_APPEALS_CNT, EMAIL_APPEALS_CNT, TEL_APPEALS_CNT, AWARD_HONOR_REC, STUDENT_ACT_CNT.

Among them, the change on NBR_OF_COMM, NBR_OF_AFFILIATION, VOLUNTEER_ ACT_CNT, AWARD_HONOR_REC have a larger impact on the likelihood of giving than the change on other variables. Specifically, 1 unit increase on the committees involved, the likelihood of giving will in-crease about 1.8 times; 1 unit increase on the affiliations, the likelihood of giving will increase about 2.6 times; 1 unit increase in the volunteer activities involved, the likelihood of giving will increase about 2.3 times, 1 unit increase on award/honors received from Creighton, the likelihood of giving will increase about 2.4 times.

The change in the following variables will cause a reduction in the likelihood of giving from the constituents: YEARS GRAD, RECENT EVENT DT, AWARD HONOR CNT, AND HC01-VC22 AVG FAMILYSIZE. Which means that, the longer they graduated from Creighton, the less likelihood of giving to the university; the further the last event attended from today, the less likelihood of giving; the further the last honor/award they received from today, the less likelihood of giving; the bigger the family size is, the less likelihood of giving.

It is very surprising that net worth and household income do not have effects on the likelihood of giving; mail solicitation, email solicitation, and phonation efforts seem to have little impact on the giving or not; however, as we expected before, the engagement of alumni in committees, events, and volunteer activities seems to make big difference in the likelihood of giving.

We need to explore more the non-donors who are classified into donor group. This may indicate that we could turn them into donors based on some characteristics they already have. However, for the donors who are classified into nondonors group, we would need to pay a closer attention to it to find out why. We do not want this happen. So, the rate in this category needs to be reduced by a better model. More work will need to be done in this regard.

8. Conclusion

This analysis is very meaningful to the division. It has the application in guiding our daily operations and decision making. The identification of the factors that impact the giving likelihood from our alumni will help the division:

- The selection of variables is very important. The process needs to involve management level and domain experts to identify possible variables in the project. More variables are better than less. We spent most of our time on identifying, selecting, and standardizing the variables for the project.
- The data quality and completeness are the keys to a successful analysis project and a reliable conclusion. We had to exclude a few variables in which the data is incomplete, such as marital status. Too many unknown values will cause bias in the analysis.
- The goal of the research should be achievable within the time frame. We initially had two correlated research questions for this project. Then we realized the complexity of the search and scaled it down to only one.

References

[1] Agresti, Alan. (2002). *Categorical Data Analysis*. New York: Wiley-InterScience. ISBN: 0-471-36093-7 .

[2] Brooker, G. W., Klastorin, T. D. (1981). "College Athletics and Alumni Giving", *Social Science Quarterly*, 62(4): 744-750.

[3] Cattell, R. B. (1952). *Factor Analysis*. New York: Harper .

[4] College Stats.org (2011). "Largest Colleges in Nebraska (NE)", [Online] Available at: https://collegestats.org/ (Retrieved 2011-04-23).

[5] Creighton University (2015). "Mission of University Relations at Creighton". [Online] from http://www.creighton.edu/giving/meetthesta (Retrieved on December 17, 2015).

[6] Gottfried, M. A., Johnson, E. L. (2006). "Solicitation and donation: An econometric evaluation of alumni generosity in higher education", *Inter-national Journal of Educational Advancement*, 6(4): 268-281.

[7] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. (1998). *Multivariate Data Analysis* (5th Ed.), Upper Saddle River, NJ: Prentice Hall.

[8] Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC. ISBN: 978-0-412-34390-2.

[9] IBM Knowledge Center (2015). "Logistic Regression Variable Selection Methods". [Online] Available at: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/regression/logistic_regression_methods.html (Retrieved on December 17, 2015).

[10] Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, 23(3): 187-200.

[11] Miles, M. B., Huberman, A. M., and Saldana, J. (1984). *Qualitative data analysis: A sourcebook*. Beverly Hills.

[12] Norusis, M. (2008). SPSS 16.0 statistical procedures companion. Prentice Hall Press.

[13] Okunade, A. A., Wunnava, P. V., Walsh, R. (1994). "Charitable Giving of Alumni", *American Journal of Economics and Sociology*, 53(1): 73-84.

[14] Peng, Y., Kou, G., Shi, Y., Chen, Z. (2008). "A descriptive framework for the field of data mining and knowledge discovery", *International Journal of Information Technology Decision Making*, 7(04): 639-682.

[15] Peng, Y., Kou, G., Wang, G., Wu, W., Shi, Y. (2011). "Ensemble of software defect predictors: an AHP-based evaluation method", *International Journal of Information Technology and Decision Making*, 10(01): 187-206.

[16] Sun, X., Hoffman, S. C., Grady, M. L. (2007). "A multivariate causal model of alumni giving: Implications for alumni fund-raisers", *International Journal of Educational Advancement*, 7(4): 307-332.

[17] Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC, US: American Psychological Association.

[18] Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.