Donor Sentiment and Characteristic Analysis Using SAS® Enterprise Miner™ and SAS® Sentiment Analysis Studio

Ramcharan Kakarla, Dr. Goutam Chakraborty; Oklahoma State University, Stillwater, OK

ABSTRACT

It has always been a million-dollar question, "What inhibits a donor to donate?" Many successful universities have deep roots in annual giving. We know sentiment is a key factor in drawing attention to engage donors. This paper is a summary of findings about donor behaviors using textual analysis combined with the power of predictive modeling. In addition to identifying the characteristics of general donors, the paper focuses on identifying the characteristics of a first-time donor. It distinguishes the features of the first-time donor from the general donor pattern. A data set containing 247,000 records was obtained from a University Foundation alumni database, Facebook, and Twitter. Solicitation content such as email subject lines sent to the prospect base was considered. Time-dependent data and time-independent data were categorized to make unbiased predictions about the first-time donor. The predictive models use inputs such as age, educational records, scholarships, events, student memberships, and solicitation methods. Models such as decision trees, Dmine regression, and neural networks were built to predict the prospects. SAS® Sentiment Analysis Studio and SAS® Enterprise Miner™ were used to analyze the sentiment.

INTRODUCTION

Fundraising is an important activity for any foundation or charitable trust. It plays a pivotal role in development of the university propelling the growth and research. Solicitation involves cost; therefore it is important to keep a check on the dollar amount raised. It is often important to find out the characteristics of first time donor since it provides us insights about donor behavior and also allows us to track the prospective donors. This paper discusses the findings of first time donor by capturing the snapshot of donors when they first make a donation along with distinguishing donors and non-donors. Textual data, i.e., posts made by university foundation in social media have been gathered to mine the sentiment of donors. It often happens that people react by what they see and hear from other people in social media.

The popularity of social media and the widespread availability of opinions in them, means organization must track and measure what are being said about them in such media. This gives an opportunity for the Foundation to identify areas of improvement. Foundation's presence in social media is critical for mass communication and attracting and engaging with the general donor base. A great amount of unstructured data are available in these platforms. For this research, data has been obtained from Twitter and Facebook to verify and explore trends in data for making meaningful decisions.

The main objective of analyzing unstructured data in this paper is to find sentiments and relate the findings back to the structured analysis. Since the textual data is culled from variety of sources the connection of each text with an individual donor has not been made. Rather a generalized study of the unstructured data is done to find insights into donor behavior and study the impact it has on the predictive models. The project is focused on annual giving prospects for XYZ university foundation who wishes to remain anonymous. The main objectives discussed in the paper are

- Distinguish donors from non-donors

- Identify the snapshot of first time donors
- Recognize the sentiment of the organization in social media

DATA DICTIONARY

Some of the sample variables that were used in the data analysis are presented below:

Variable	Data Type
Constituent ID	Nominal/ID
Donor Age	Interval
Graduation Year	Nominal
Marital Status	Nominal
Bachelor's Indicator	Binary
Graduate Indicator	Binary
Doctors Indicator	Binary
City	Nominal
Response	Binary
State	Nominal
County	Nominal
Title	Nominal
Parent from same university	Binary
Athlete	Binary
Scholarship	Binary
Attended Alumni Events	Binary
Homecoming	Binary
Spring graduate	Binary
Fall graduate	Binary
Summer graduate	Binary
Clicked	Binary
Donor Gender	Binary
Donor Spouse Gender	Binary
Donor Spouse Age	Interval
No of Solicitations	Interval
Number of Educational Records	Interval
Groups Membership	Binary
Total Number of Immediate Family Relations	Interval
Facebook and Twitter	Text
Email Subject Line	Text
Solicitation Type	Nominal

DATA PREPARATION

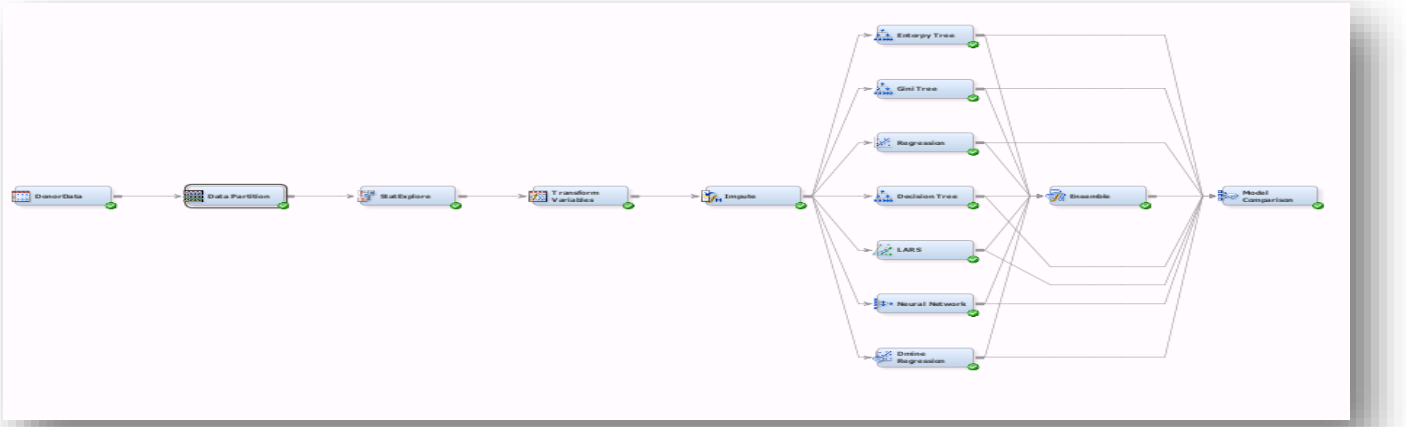
Data Preparation is divided into three segments. For the first objective, distinguishing donors from non-donors, the data has been obtained from several internal tables. Few tables have multiple instances due to the nature of transactional data. Each table has been flattened before joining with other tables using an identifier key. External data has been collected from the US census at a county level and merged back with the internal tables. Different external datasets were combined to form a single external data set at a county level. Internal and External Data sets are merged to create the final dataset. The final dataset contained 221 variables with 247000 records. Input space reduction techniques were used to further reduce the number of input variables by selecting the best variables from the available inputs. This needed transformations, recoding and binning of the variables. Complete data was available for only 10% of the records. Data has been imputed for some variables using decision tree techniques.

For the second objective, identifying the characteristics of first time donor, the previous data had been tweaked to retain the essential information. Since there was no data warehouse structure in place for the organization, all the details were not readily available. For example the marital status of the donor might be different at the time of first donation than later ones; Age may not be treated as a constant. All these time dependent data are crucial in building the predictive model. Few data fields had to be obtained from Meta data fields and few were obtained from the unstructured format. Due to lack of information for all the years, a year cutoff was set arbitrarily. To build a predictive model, we need data from both donors and non-donors. Many fields available for donors such as first giving date, gift history were not available for non-donors. For the purpose of modeling, many fields were flattened out to set up common fields. Since there were dependencies with the time lines, no external data was used for this project. Since the time independent data has been created already for the previous project, the same fields have been used. The final dataset has been obtained after several phases of iterations and transformations of the variables. After data preparation, data exploration was done to identify the leads in distinguishing the characteristics. After data exploration, predictive models were built using different algorithms. The models include Decision Trees, Regressions, Neural Networks, Partial least squares, Dmine Regression. Several combinations of data partition were tried such as 50/50, 80/20, 70/30, 60/40 for training and validation respectively. All these models were run in these combinations. After modeling, validation and assessment of these models was done to verify the results. The results are presented in the conclusion section.

For the final objective, which is to identify sentiment expressed about the organization in social media, the data has been gathered from Facebook, Twitter and LinkedIn for text analysis. The unstructured data has been transformed to a semi structured format. R Language scripts and in-house tools were used to grab the social media data. Data has been processed using SAS Sentiment Analysis studio for determining the sentiment. SAS Sentiment Analysis has been used to classify the comments into positive and negative sentiments. Rule-based models have been used to determine the sentiment. These models use CLASSIFIER rules identifying the key words. Weights are given according to the prominence of the word and position of word from the feature description. Statistical rule based models are also available in the Sentiment Analysis Studio. These models require pre-classification of text into positive or negative sentiments. Models are built using the examples provided by experts which classify the comments. To quickly identify the key features and words, Text Miner in the Enterprise Miner has been used using the approach advocated by Chakraborty, Pagolu and Garla (2013). Parsing node has been used which identifies the parts of speech and does stemming. The variants of the words are categorized in this step. It is then passed through filter node which ensures that valuable information is retained. Similar process has been applied to the email subject lines. When passed through clustering node, 1000 email subject lines resulted in 8 different clusters.

METHODOLOGY

The modeling approach followed for the project is SEMMA (Sample, Explore, Modify, Model and Assess). The data was portioned into to two stratified samples (training, validation). The training data is used to build the model. Validation data is used for testing the accuracy of the model. This provides an honest assessment of the models built.



For Sentiment Analysis, the data had been parsed, filtered and clustered using SAS enterprise miner to understand the features in the data. Then, a sample of Twitter feeds were classified into positive and negative categories using expert opinions this sample was used to train the statistical models in the sentiment analysis studio which was later used to classify the remaining text.

MODEL ASSESMENT

Different models have been built for distinguishing donors from non-donors. Dmine Regression is selected based on validation misclassification rate.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE
Y	DmineReg	DmineReg	Dmine Regression	resp	Resp	0.23553	123567	0.232611	0.986524	38969.62	0.157686	0.397097	247134
	Ensmbl	Ensmbl	Ensemble	resp	Resp	0.236817	123567	0.232319	0.966035	38848.01	0.157194	0.396477	247134
	Tree	Tree	Decision Tree	resp	Resp	0.239981	123567	0.239846	0.997986	40944.24	0.165676	0.407034	247134
	Tree2	Tree2	Gini Tree	resp	Resp	0.241705	123567	0.238972	0.997986	40760.95	0.164935	0.406121	247134
	Neural	Neural	Neural Network	resp	Resp	0.241834	123567	0.238874	0.986774	40485.85	0.163821	0.404749	247134
	Reg	Reg	Regression	resp	Resp	0.241964	123567	0.2383	0.989489	40649.55	0.164484	0.405566	247134
	Tree3	Tree3	Entorpy Tree	resp	Resp	0.246755	123567	0.24161	0.997986	42045.83	0.170134	0.412473	247134
	LARS	LARS	LARS	resp	Resp	0.250486	123567	0.246652	0.981183	41311.51	0.167162	0.408855	247134

In case of first time donor classification decision tree model emerged to be the winner amongst all the models.

Findings of the analysis are presented in the results section.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Tree4	Tree4	Misclassific...	resp	Resp	0.244505
	Ensmbl	Ensmbl	Ensemble	resp	Resp	0.262285
	MBR2	MBR2	MBR (2)	resp	Resp	0.337571
	Reg4	Reg4	Regression...	resp	Resp	0.38128
	Reg3	Reg3	Regression...	resp	Resp	0.543134

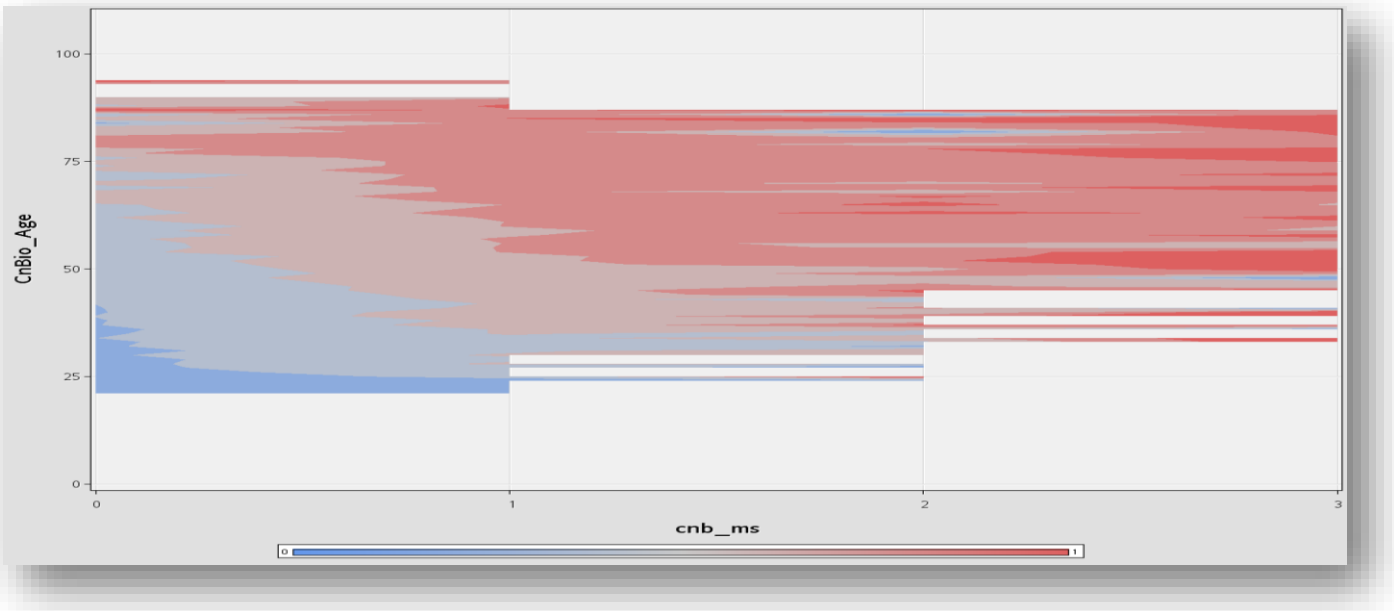
Predecessor Node

RESULTS

Distinguishing Donors from Non Donors

It is found that donation patterns vary widely across age groups and marital status. Attendance at events conducted by university was also found to be the influencing factors in distinguishing donors from non-donors for Annual Giving. Single are least likely to give. Widowed and divorced are most likely to make a donation starting from their late 30s.

Below graph represents the donor's propensity across age and marital status. Blue represents least likely and red represents most likely to give. On the horizontal axis, 0 represents single, 1 represents married, 2 represents divorced and 3 represents widowed. Age is represented on the vertical axis. There is a clear distinguishing pattern based on age and marital status.



Married are likely to make donations around age groups of 45 and above. It is also found that if the constituent is a volunteer there is a greater chance of donation. This classification provides us power to promote and design programs to the specific groups. Event attendance has been significant driving factor for

annual giving. The chances of donation are likely to go up if the constituent had attended an event conducted in the past 6 to 12 months.

Two in every three constituents are likely to make a donation if the constituent and spouse are from the university. Alumni pairs account for the major chunk of the annual giving programs. The chances of donations are directly related to the number of educational degrees of the constituent. Donors with post graduate or higher degree have a higher probability of giving than the remaining. Couples of same gender have a high likelihood of giving.

Scholarship is not a strong distinguisher among the constituents who donate versus those who don't. There has been speculative hypothesis that scholarship recipients would give to annual giving section. But, this is not supported in this data. University Foundation solicits the donors through four different channels namely phone, postal mail, email and personal solicitation. Out of the four different types of solicitations, direct mail solicitations are found to be most influencing followed by email solicitations. Not surprisingly, counties with better median incomes have a better prospect base of donors. Especially counties with median income greater than $50K were found most promising.

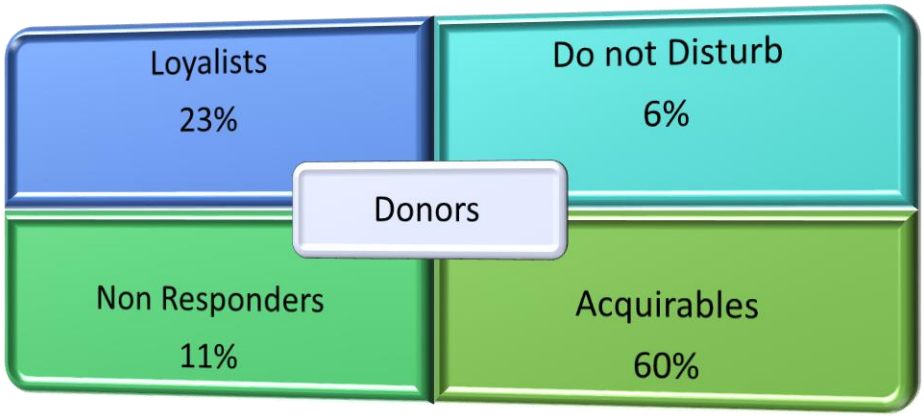
External data from US census has also been used in the analysis. Apart from the median incomes the only other variable that made to the final model was percent of smokers increase in a county. As the count of smokers increase in the county chances of donation go down. There also has been strong seasonality effect on the pattern of donations. It is found that best months in terms of donation are December, January and May. One plausible explanation for this might be the festive season at the end of the year. The other interesting fact that was found was average age of people who donate is significantly different from those who don't (11 years gap 51 vs. 40). This could also be observed in the graph presented above especially in case of married segment.

There is a 90% likelihood for a single to donate if they have multiple degrees and are a net community member. Net communities are any alumni association memberships or online activities that constituents are actively participating. The person is likely to donate with a 50% chance of probability if he had multiple bachelors or master's degree from the University. As the number of immediate family relationships go up a person is more likely to donate. Gender had no impact on the donation. There are no significant pattern change behaviors between males and females in donations. If the constituent is a member of any one of the college group there is a fair chance of donation (65-70%). Constituent would be a very good prospect if he/she had a membership in two or more groups then the chances of donation increase to 90%.

Another interesting question that was answered during the analysis was the status of donation pattern if the constituent has a parent who has attended the university. It was found that there was no significant impact in the donation pattern if the alumni had a parent who attended the university. A constituent who had studied in University along with other state colleges had 67% chance of donation. Chances improve if the latest degree is from the university. Another important fact uncovered was who is least likely to donate. It was found that if the donor age is between 25 and 45 and is either single or married without a membership in anyone of the groups the likeliness of donation is 7 in 100 constituents.

Donor Classification

For the purpose of donor analysis, Donors were divided into four segments. Loyalists are the set of donors who have made their donations even before solicitations are made. Do not disturb segment of constituents are those who preferred no solicitations of any form from the university. Non responders are the segment where there is no response from the constituent after multiple solicitations over the span of 10 years or greater. Remaining fall under the category of acquirable.



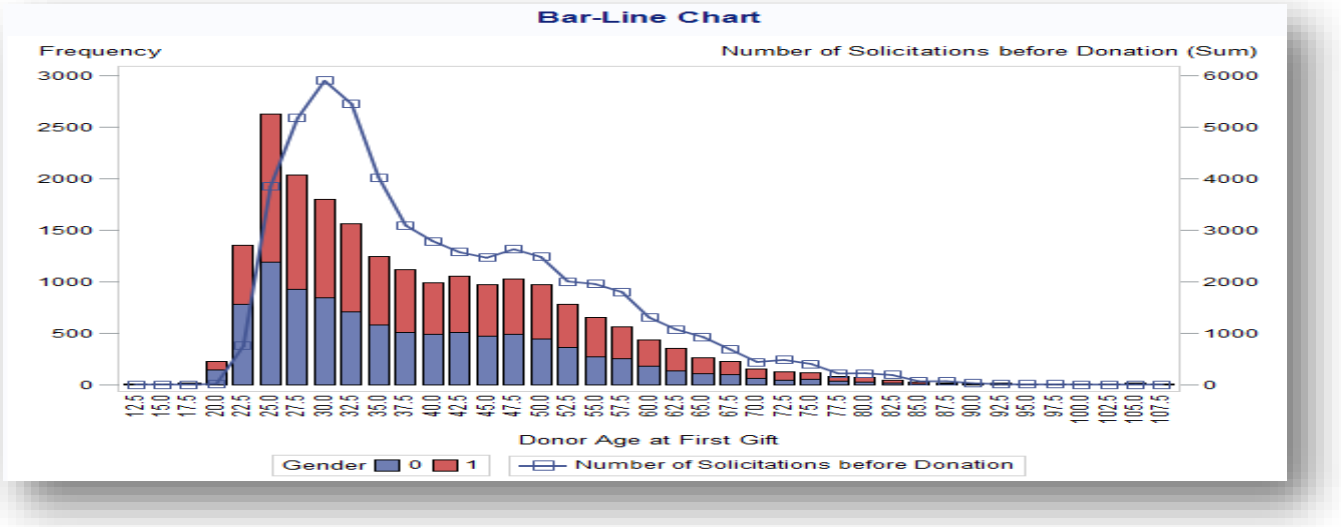
First Time Donor Characteristics

The mean age of first time donor is 36 years. Although mean shows a higher age, majority of the constituents make their first donations in the age groups of 22 to 30. Especially the prospect group above 40 is likely to respond positively after 2 or more solicitations. For the first time donors, November and April are the favorite months followed by October, December and May. Plausible explanations for these months are Thanksgiving Day, tax day and graduating months. Statistically there is a difference in number of first time donations made in these months.

There were 40,564 new donors since 2003. Out of these donors, 13,304 members made donations only after solicitations. The most responsive group after solicitations is Age group 20-30 followed by 30-40. 67.33% of prospects in age group of 20-30 require a minimum of 2 solicitations to donate. Volunteers are the constituents who volunteer in university activities. The volunteers who donated, made their donations even before they assumed their positions. Volunteer segment was also found to be distinguishing donors from non-donors. Scholarship is not a strong distinguisher in case of first time donors.

In case of first donation, the number of solicitations required to be made to males are marginally greater than the females before they turn into donors. Majority of first time donors are the alumni who graduated in the past 12 years. Below graph shows the distribution of first time donors across different age groups and gender

classification. The line shows the total number of solicitations that were made to each of these groups



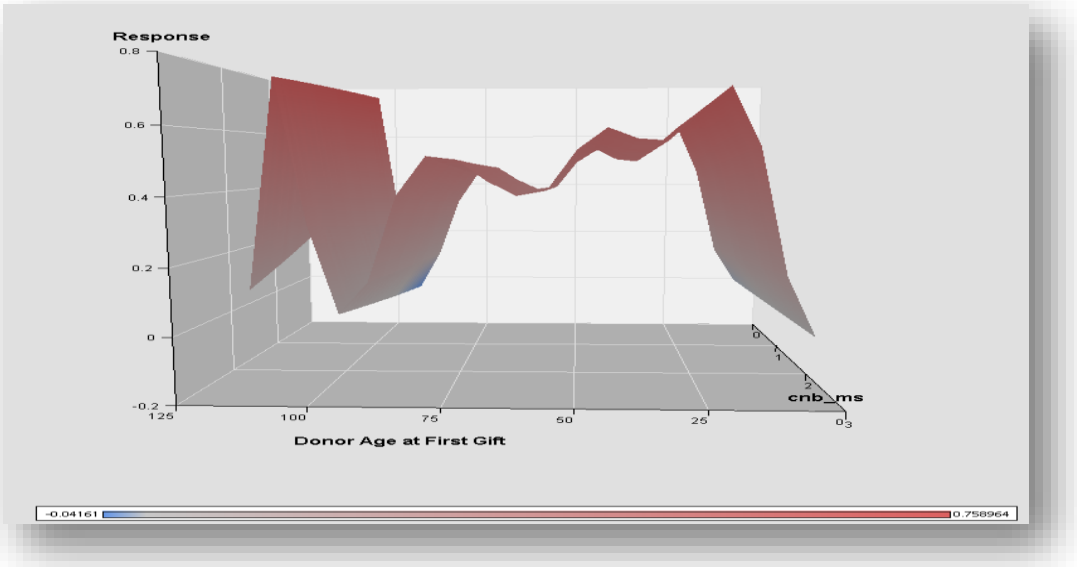
Direct mail solicitations are the most important followed by Email solicitations statistically even in case of first time donors. 90% of students who donated during graduation year have done the donation voluntarily. These Alumni are more loyal and 63% of them made multiple donations, Conversions could be much easier with these Alumni. These alumni can be classified as the loyalists. Two-thirds of the student athletes who made a donation, will go on to make multiple donations.

On an average 4 Solicitations were done before a first donation is made. 22 to 33 age group is the most affected by solicitations, 36-40 age group is the least affected apart from older age groups. From age 40 and above major first time donors are males and continue to be completely dominant in the later age groups in case of athletes. Numbers show the number of first time donations is high in the age group of 22 to 24. From late 20s there is a decrease in number of first time donors.

Optimum Solicitation combination when 4 solicitations are made is 2 Email + 1 Direct mail+ 1 Tele Solicitation based on the response rates. Next best combination is 2 Tele Solicitations + 1 Direct Mail + 1 Email.

Among PhD graduates who donate, 50% are likely to make their first donations in the age group of 30 to 50. In graduates who donate 52% are likely to make their first donations in the age group of 23 to 41. Among MBA graduates who donate 55% are likely to make their first donations in the age group of 23 to 38. Prospects who have never made a donation and have attended an event are most probably likely to make a donation within 10 to 11 months. 24-36 age is the major age group attending these events and attendance falls after age 38.

Below graph shows the first time donor response pattern according to the donor age. The spikes observed to the end of the age group is caused to due to fewer observations.



Email Subject Line Analysis

There were over 1,000+ different email subject lines that were used for communicating with the prospects and donors. In order to improve the reachability towards the donors, these subject lines were observed to see if few subject lines have better impact than others in viewership of the mail. It is observed that 'Video' in the subject line is most likely to increase the likelihood of opening the mail greater than any other word. It refers to the fact that video messages attract more viewership than normal textual mails.

'Happy' in the subject line is also likely to increase the likelihood of opening the mail. Numbers addressed in the beginning of subject line would have a greater influence in opening the mails than numbers elsewhere. Campaign details with dollar amounts at the start of the email have a better viewership than those addressed in the middle of the subject line. Shorter subject lines have a greater impact in the conversion. Subject lines with number of words less than 6 have a better viewership. 'Change' and 'Family' is a significant word both in opening of the mails and conversion. Donors have an affinity towards these words. All the donations from mail channel had these words in the subject lines. Conversion rate is 0.01% through email channel. All Subject lines could be divided into 8 clusters

- Cluster1: Video + President + Event
- Cluster2: Welcome Reception+ Reminder+ Celebration
- Cluster3: Holiday+ Cowboys+ Foundation
- Cluster4:Happy Thanksgiving
- Cluster5: Gift+ Announcement+ Charitable
- Cluster6:Change+ Acknowledgement+ Community
- Cluster7: Homecoming+ Football+ Ticket
- Cluster8: Newsletter+ Scholarship+ Application

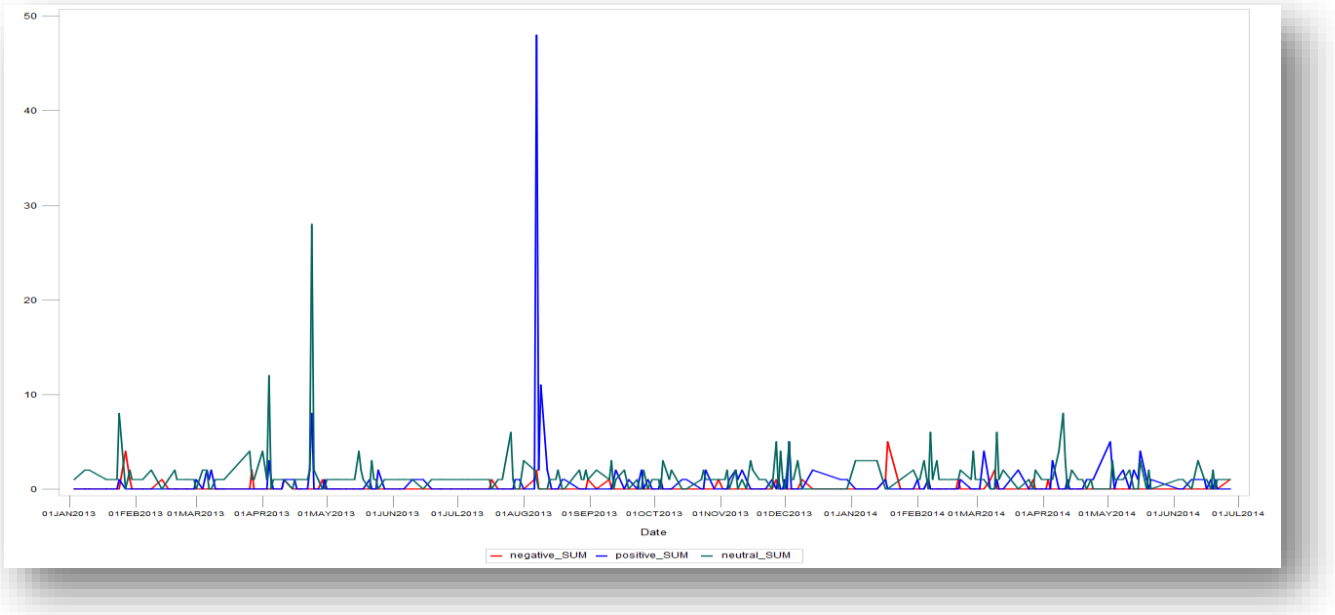
Sentiment Analysis Inferences:

600+ distinct comments and twitter feeds were used in the analysis. Both statistical and rule based models were used to identify the polarity of the comments. It is found that social media is widely as platform for making announcements and acknowledgements. Since January 2013, events were positively received by the donors. It is found that there were positive trends towards the food items that are being distributed or used in the events. Constituents identify themselves when invited to an event. It is also found that employees working as part of the foundation have a strong positive affinity towards the organization.

Few constituents have responded positively towards the calls from of the foundation. Few people admitted giving just because of calling. Certain Segment of people like when being recognized on the social networks. Branding awareness program conducted by university foundation on American Airlines, magazines showed positive trends. Although circulated vividly in Twitter, neutral trends were recorded when university announced its $1 Billion goal. Negative trends were recorded on the time of the call. Solicitations during major games telecast were not well received by the constituents. Overall 6% of negative comments, 32% positive comments and 62% of neutral comments were observed.

TimeLine Sentiment Trends

All the comments have been classified as positive, negative or neutral. They are represented below on a timeline. Blue represents the sum of positive comments, red represents negative and green represents the neutral comments. It is observed that in August 2013 there was a spike in the positive trend when the University was announced as one of the best in the region. The surge happened due to the euphoria expressed as comments and retweets during the time period.



CONCLUSION

The comprehensive analysis on structured and unstructured data has provided deep insights to better understand the donor behavior. The age, marital status and event attendance are important predictors in case of donations. The optimized combination of solicitation that may likely initiate a donation was figured out from the available data. The key words that draw the attention of donor were understood by email subject line analysis. Deeper insights to the donor preferences was understood using the textual analysis. Event attendance is found to be a significant factor that can drive alumni towards donations. Text mining reasserts this fact that constituents identify themselves with university when being invited to the event and that the odds of donation would improve upon the event attendance.

REFERENCES

- Goutam Chakraborty, Murali Pagolu and Satish Garla. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®, 2013.
- Birkholz, Joshua. 2008. Fundraising Analytics: Using Data to Guide Strategy. Hoboken, New Jersey : John Wiley & Sons, Inc., 2008
- Sargeant, Adrian. Marketing management for nonprofit organizations. Oxford: Oxford University Press, 1999
- Baker, Michael John, and Susan J. Hart, eds. The marketing book. Routledge, 2008.
- Sun, Xiaogeng, Sharon C. Hoffman, and Marilyn L. Grady. "A multivariate causal model of alumni giving: Implications for alumni fundraisers." International Journal of Educational Advancement 7.4 (2007): 307-332.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at

Ramcharan Kakarla
Oklahoma State University
Stillwater, OK, 74075
ramcharan.kakarla@okstate.edu

Ramcharan Kakarla is a recent graduate in Management Information Systems from Oklahoma State University. Before joining the graduate program he worked as a Performance Test Engineer. He has an undergraduate degree in Electrical Engineering. He has two year experience with SAS Data Mining and Analytical tools. He is a Advanced SAS certified programmer and SAS Certified Predictive Modeler using SAS Enterprise Miner 7.1. In 2014 he received SAS and OSU Data Mining Certificate.

Dr. Goutam Chakraborty
Oklahoma State University
Stillwater, OK, 74078
goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

SAS® and all other SAS® Institute Inc. product or service names are registered trademarks or trademarks of SAS® Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.