Name:: Shyni David

Project:: WOSU Donor Database Cluster Analysis & Segmentation

Original Questions:

*Q. What Issue do you want to investigate?*

The issues I want to investigate are:

1. Which individuals or institution and geographical location most profitably contribute to WOSU?

2. Is there a relationship between engagement and giving? Who are the donors and how long the donation is sustained?

Overview

WOSU’s primary source of funding comes from individual community support. Their development team is seeking to develop a deeper understanding of the constituencies by using data analytics in understanding the past and projecting the future. When donors approach philanthropy, their gifts are thoughtful and has intended purposes and seek a return on their philanthropic investments. A pertinent question is why do they give to WOSU? WOSU has a wide variety of fundraising programs through events, radio, TV , web, emails and personal contact. Which of these tasks contribute to increased giving? Which tasks detract from the fundraising success? What is the spread of the donors across the country? What should the campaign goal be? What are the factors for their success?

The exploratory data analysis will be used to investigate the trends. The statistical tools and techniques will help to interpret the data and build models to increase the predictability, thus empowering WOSU to strategically invest their resources in increasing the scalability and sustainability of donations.

1. **The Datasets available.**

**There are 3 Datasets provided by WOSU:**

1. 2000-2009 – 33740 observations (represents Original Donations), 22 variables/attributes
2. 2010-2018- 36361 observations(represents Original Donations), 22variables/attributes
3. Event files – 22 Text file corresponding to some events since 2014 with the names and date and other related information.

Selected Variables

1. EnvSalu - Donor Name - Char
2. LetterSalu- Letter Salutation - Char
3. City - Char
4. StateProv - State - Char
5. AcctSts - Account Status - Char
6. OrigGiftDate - Original Gift Month Year - Date
7. OrigGiftAmt - Original Gift Amount - Numeric
8. OrigGiftSrc - Original Gift Source - Char
9. OrigGiftMode - Original Gift Mode - Char
10. SolType - Original Gift Solicitation Type - Char
11. SolMeth - Original Gift Solicitation Method -Char
12. OrigGiftHasPremFlag - Original Gift Has Premium – Char

The following variables are not needed for my analysis.

1. PBS Digital Token
2. Passport Activation Date

Redundant Variables:

1. AcctID - Account ID - Char
2. Zip Code+4 digits
3. SeqName -Sequence Name
4. Addr1 - Address Line1
5. Addr - Address Line 2
6. EmailAddr - Email address
7. OrigGiftPlgType - Original Gift Pledge Type
8. OrigGivingLevelAmt - Original Gift Level Amount
9. **Attribute Information:**

* Donor Name (qualitative)

Name of the donor as a first name and Last name.

* Letter Salutation (qualitative)

Salutation to individual donors with prefix as Mr. or Mrs. and Organization names.

* City(qualitative)

The data spans around many states in US and territories.

* State (qualitative)

* Account Status (qualitative)

The statuses are expired, contributor and active.

For example, if the first gift was on April 2, 2004, it will be in the Original Gift 2000-2009 file.  People/organizations who stop giving are considered as expired.  People/organizations who are still donating are considered as active/contributor. These files are based on original giving date.

Per Rob Walker, Active member give $60, Contributor gives $59 or less, Expired status refers that the individual or organization is no longer a member.

* Original Gift Month Year (numerical )

The Original giving date for the donor.

* Original Gift Amount(numerical)

The Original donation amount or the donations received for the first time.

* Original Gift Source (qualitative)

Has the detailed information of the event that initiated the donation. Per Rob Walker, TOOP is an internal phrase for a one off TV pledge drive and can be categorized with TV drives. LAPS is a Lapsed member who has not given in the past 12 months before that gift. Some of the Solicitation descriptions in Original Gift Source can be replaced with the descriptions in the variable Original Gift Solicitation Method.

* Original Gift Mode(qualitative)

This field describes how the payment has been made like by CC, Cheque, Cash or EFT.

* Original Gift Solicitation Type(qualitative)

Generalized description of Solicitation type specially to identify Additional Gifts.

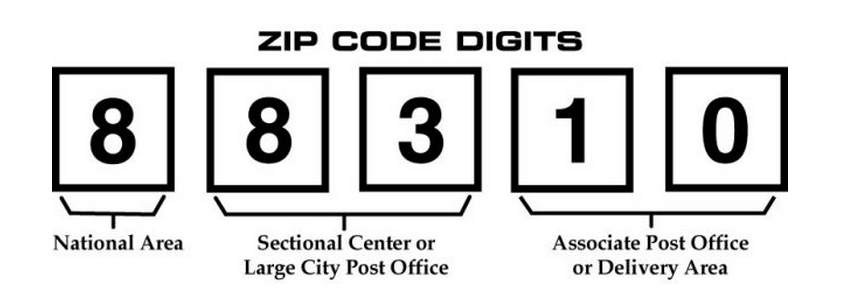
* Solicitation Method (qualitative)
* Original Gift Has Premium (qualitative)

This mean that when the donor made their first gift, they ask for a thank you gift.

Redundant Variables :

* Account ID (qualitative) - This is a unique ID that is randomly assigned when donors make their first donation. These are the people who made their first ever gift in those years.
* Original Gift Level Amount(numerical) -This is the amount of their first gift ever. Is a redundant field.
* Original Gift Pledge Type(qualitative) - This field is same as Solicitation Type
* Sequence Name(qualitative) – This field is same as Donor name.
* Addr1(qualitative) - Address Line1
* Addr(qualitative) - Address Line 2
* Zip code (qualitative)

Zipcode + 4 digits(qualitative) - Zip codes were devised entirely for the USPS mailing efficiency. They do not necessarily end at the county line, city line or any political boundary except the national border. The State and the City variables are more informative of the geographical location.



1. **Data Wrangling**

The Exploratory Data Analysis will be segmented for Individuals and organization for in state(OH) and out of state(all other US states).

The Original Gift Data file has to be wrangled for the following:

* + 1. Split Year, Month, Day
    2. Eliminate redundant variables
    3. Lower case for all char fields
    4. Solicitation source mutate to another root word field as a categorical variable.
    5. Split Donor name as first name and last name to link Event file and Original Gift Source.
    6. Mutate the Letter Salutation to create Gender field. Use gender R package.
    7. Identify Organizations with key words as company, group, corporation, foundation, organization, campaign , firm, establishment, agency, office, enterprise, operation, institution, venture, undertaking, practice, society, league, club, network etc. And keep the set of observations as a separate file.
    8. The State variable has data that spans across many states in US and territories . Ohio is the exceptional contributor so will consider separately for analysis from the rest of states. Will add additional field of Regions and Divisions as a categorical variable.

Additional Information:

Regional divisions used by the United States Census Bureau:

• Region 1: Northeast

• Division 1: New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont)

• Division 2: Mid-Atlantic (New Jersey, New York, and Pennsylvania)

• Region 2: Midwest (Prior to June 1984, the Midwest Region was designated as the North Central Region.)[6]

• Division 3: East North Central (Illinois, Indiana, Michigan, Ohio, and Wisconsin)

• Division 4: West North Central (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota)

• Region 3: South

• Division 5: South Atlantic (Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, and West Virginia)

• Division 6: East South Central (Alabama, Kentucky, Mississippi, and Tennessee)

• Division 7: West South Central (Arkansas, Louisiana, Oklahoma, and Texas)

• Region 4: West

• Division 8: Mountain (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming)

• Division 9: Pacific (Alaska, California, Hawaii, Oregon, and Washington)

* + 1. The Original Gift source is a huge vital list that has to be transformed to root words for analysis. For modelling this field can be split as dummy variables.
    2. Create a new file by adding 2 new fields for the count of donors and Gift amount, grouped by State, Solicitation Method and Original gift source for the 12 months for all 20 years.

New file- Original\_gift\_new\_file

Variables: Original Gift Month, Original Gift Year, Count of donors, Gift Amount aggregate , Division, Region, Gender, Original Gift source, Solicitation Method

* + 1. The 22 Event text file has to be combined into one file with variables as First Name, Last name, Event month, Event year, Event day , Event name and Original Gift Source(root word).
    2. Check for Null Values
    3. Merge the value “Additional Gift” from variable Orig\_Gift\_Pledge\_type into variable Solicitaion\_type.
    4. Do Exploratory data analysis both univariate and bivariate.

1. **Incorporate covariates to explain context.**

Context1: Analyze the first-time donations based on geographical location, individuals or organization by using 2 sets of files one for individuals and one for organizations. If every year a certain level of first-time donors is sustained it will be cumulative income to the contribution by sustained donors for WOSU.

File: Original\_gift\_new\_file (for individuals/organizations)

Response : Gift Amount aggregate

Covariates : Original Gift Month, Original Gift Year, Count of donors, Gift Amount aggregate , Division, Region, Gender(only for individuals), Original Gift source, Solicitation Method

Context2: To determine the relationship between engagement and first time giving. Analyze what was the effect of the Events organized by WOSU on the actual first time donation made. Analyze events that are most effective in materializing the first-time donation and events that were least effective for individuals and organizations.

This will need the 2 files : Original\_gift\_new\_file and Event file

Response : Gift Amount aggregate

Covariates: Event month, Event Year, Solicitation Source, Event name, Region, Division, Count of donors

Context3: Which mode of payment was most commonly used for first time donation? Did the mode of payment had an effect on first time donation in different regions?

This can help to suggest if mode of payment needs improvisation to be more effective in quicker transactions and thus increase scalability. There are tremendous changes in the trend in mode of payment in last 20 years especially with mode of digital payment systems for non-profit organizations.

File: Original\_gift\_new\_file (for individuals/organizations)

Response : Gift Amount aggregate

Covariates : Original Gift Month, Original Gift Year, Count of donors, Gift Amount aggregate , Division, Region, Original Gift source, Original Gift Mode

Context4: Which Solicitation methods (like on-air, web, direct mail, email etc.) were most effective for first time donors and at what time period of the year?

File: Original\_gift\_new\_file (for individuals/organizations)

Response : Gift Amount aggregate

Covariates : Original Gift Month, Original Gift Year, Count of donors, Gift Amount aggregate , Division, Region, Gender(only for individuals), Original Gift source, Solicitation Method

Context5: Analysis of the cost expense to gifts for first time donors. What is the percentage of first time donors who have asked for a thank you gift. The result will inform if the cost for WOSU is trivial, average or high. It is beneficial for WOSU when the first-time donors who requested for thank you gifts has maintained the status as Active/Contributor for a period of time.

File: Original Gift file(individuals/organizations)

Response : Original Gift Amount

Covariates: Region, Division, Original Gift Month, Original Gift Year, Original Gift has Premium

Context6: - The effect of Original Gift Source in sustaining both in state and out of state(OH) after first time donations. Are the donations status active or expired?

File: Original Gift file(individuals/organizations)

Response : Original Gift Amount

Covariates: Region, Division, Original Gift Month, Original Gift Year, Original Gift has Premium, Original Gift source, Solicitation Method, Account Status

Context7:- What is proportion of the additional gifts ?

File: Original\_gift\_new\_file (for individuals/organizations)

Response : Gift Amount aggregate

Covariates : Original Gift Month, Original Gift Year, Count of donors, Gift Amount aggregate , Division, Region, Original Gift source, Solicitation Method, Solicitation\_Type

Functional form – Multiple Regression Model

Performance metric- RMSE

Y =β0+β1x1+β2x2+...+βkxk