John Fields

Dr. Gregory Block

IST652 - Homework #1

7/31/19

Introduction

Organizations such as schools, humanitarian groups and other not for profit entities rely on charitable donations to function. These organizations and groups are increasingly using tools such as data analysis and machine learning to target donors and increase giving.

The purpose of this assignment is to explore a data set of donors to look for trends and giving patterns that would be useful to better target potential users and communicate more effectively with existing donors.

Data and Source

The data set provided for this assignment includes donor information in a csv file format that was downloaded from the Syracuse 2U platform. No information was provided on the source of this data but a data definition document was provided to explain each of the variables in the dataset. Below is a summary of the data including new column names that will be created in Python.

New Column Name	Data Type	Donor_data Name	Northwestern PREDICT422 Name	Definition
DROP		Row Id	INDEX	INDEX
DROP		Row Id.	ID number [Do NOT use this as a predictor variable in any models]	ID number [Do NOT use this as a predictor variable in any models]
Region	Qualitative, Nominal	zipconvert_2,zipconvert_3, zipconvert_4, zipconvert_5	REG1, REG2, REG3, REG4	Region (There are five geographic regions; only four are needed for analysis since if a potential donor falls into none of the four he or she must be in the other region. Inclusion of all five indicator variables would be redundant and cause some modeling techniques to fail. A "1" indicates the potential donor belongs to this region.)
Homeowner	Qualitative, Nominal	homeowner dummy	HOME	(1 = homeowner, 0 = not a homeowner)
Number of Children	Quantitative, Ratio (Continuous)	NUMCHLD	CHLD	Number of children
Income	Qualitative, Nominal	INCOME	HINC	Household income (7 categories)
Gender	Qualitative, Nominal	gender dummy	GENF	Gender (0 = Male, 1 = Female)
Wealth Rating	Qualitative, Nominal	WEALTH	WRAT	Wealth Rating (Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and 0 being the lowest.)
Home Value	Quantitative, Ratio (Continuous)	HV	AVHV	Average Home Value in potential donor's neighborhood in $ thousands
Median Income	Quantitative, Ratio (Continuous)	Icmed	INCM	Median Family Income in potential donor's neighborhood in $ thousands
Average Income	Quantitative, Ratio (Continuous)	Icavg	INCA	Average Family Income in potential donor's neighborhood in $ thousands
Percent Low Income	Quantitative, Ratio (Continuous)	IC15	PLOW	Percent categorized as "low income" in potential donor's neighborhood
Number of Promotions	Quantitative, Ratio (Discrete)	NUMPROM	NPRO	Lifetime number of promotions received to date
Lifetime Gifts($'s)	Quantitative, Ratio (Continuous)	RAMNTALL	TGIF	Dollar amount of lifetime gifts to date
Largest Gift($'s)	Quantitative, Ratio (Continuous)	MAXRAMNT	LGIF	Dollar amount of largest gift to date
Last Gift($'s)	Quantitative, Ratio (Continuous)	LASTGIFT	RGIF	Dollar amount of most recent gift
Last Donation(months)	Quantitative, Ratio (Discrete)	totalmonths	TDON	Number of months since last donation
First-Second Gift (months)	Quantitative, Ratio (Discrete)	TIMELAG	TLAG	Number of months between first and second gift
Average Gifts to Date	Quantitative, Ratio (Continuous)	AVGGIFT	AGIF	Average dollar amount of gifts to date
Donor (0 or 1)	Qualitative, Nominal	TARGET_B	DONR	Classification Response Variable (1 = Donor, 0 = Non-donor)
Donation Amount($'s)	Quantitative, Ratio (Continuous)	TARGET_D	DAMT	Prediction Response Variable (Donation Amount in $).

Table 1 - donors.csv data variables, descriptions and data types

Data Exploration and Cleaning

The requirements for this assignment provided two options for the data structure in Python: (1) A list of dictionaries or a combination of lists, dictionaries and numpy arrays, (2) A pandas dataframe. Since the dataset is in a "rectangular" format (rows and columns), a dataframe seemed a better choice so this structure will be used for the assignment.

The first step in the process was to import the csv file to Python and explore the data to understand what changes and cleaning needed to occur. The following observations needed to be addressed before starting the analysis:

1. The column names did not provide easy to understand descriptions of the data and were changed to those shown in Column 1 of Table 1.
2. The row id and row id. columns did not provide any useful information for the analysis and these were deleted using the df.drop function.
3. Several of the variables are boolean (0 or 1) and could be converted from integer (int64) to boolean (bool). After consulting with Neal Bates (a Python programmer in the Applied Data Science program), the decision was made to keep in integer format because it is easier to sum columns compared to locate and count in Python.

With the data cleansing and formatting complete, summary statistics were created using the df.describe function. Below is the resulting table after transposing the orientation.

	count	mean	std	min	25%	50%	75%	max
Region1	3120.0	0.214423	0.410487	0.000000	0.000000	0.0	0.000000	1.000000
Region2	3120.0	0.185256	0.388568	0.000000	0.000000	0.0	0.000000	1.000000
Region3	3120.0	0.214423	0.410487	0.000000	0.000000	0.0	0.000000	1.000000
Region4	3120.0	0.384615	0.486582	0.000000	0.000000	0.0	1.000000	1.000000
Homeowner	3120.0	0.770192	0.420777	0.000000	1.000000	1.0	1.000000	1.000000
Number of Children	3120.0	1.069231	0.347688	1.000000	1.000000	1.0	1.000000	5.000000
Income	3120.0	3.893910	1.636186	1.000000	3.000000	4.0	5.000000	7.000000
Gender	3120.0	0.609295	0.487987	0.000000	0.000000	1.0	1.000000	1.000000
Wealth Rating	3120.0	6.402244	2.539978	0.000000	5.000000	8.0	8.000000	9.000000
Home Value	3120.0	1141.361859	946.642162	0.000000	556.000000	822.0	1338.750000	5945.000000
Median Income	3120.0	388.217308	172.815950	0.000000	278.000000	356.0	465.000000	1500.000000
Average Income	3120.0	432.088141	168.195104	0.000000	318.000000	396.0	516.000000	1331.000000
Percent Low Income	3120.0	14.702885	12.079882	0.000000	5.000000	12.0	21.000000	90.000000
Number of Promotions	3120.0	49.089423	22.717130	11.000000	29.000000	48.0	65.000000	157.000000
Lifetime Gifts	3120.0	110.399875	147.299933	15.000000	45.000000	81.0	134.625000	5674.900000
Largest Gift	3120.0	16.651397	22.223521	5.000000	10.000000	15.0	20.000000	1000.000000
Last Gift	3120.0	13.522917	10.581439	0.000000	7.000000	10.0	16.000000	219.000000
Last Donation	3120.0	31.136859	4.132952	17.000000	29.000000	31.0	34.000000	37.000000
First-Second Gift	3120.0	6.861859	5.561209	0.000000	3.000000	5.0	9.000000	77.000000
Average Gifts to Date	3120.0	10.690713	7.443980	2.138889	6.356092	9.0	12.811652	122.166667
Donor	3120.0	0.500000	0.500080	0.000000	0.000000	0.5	1.000000	1.000000
Donation Amount	3120.0	6.499612	10.597849	0.000000	0.000000	0.5	10.000000	200.000000

Table 2 - Summary Statistics for the donors.csv data set

During the exploratory data analysis, four visualizations were created in Figures 1-4 to better understand the data and explore any patterns that exist.

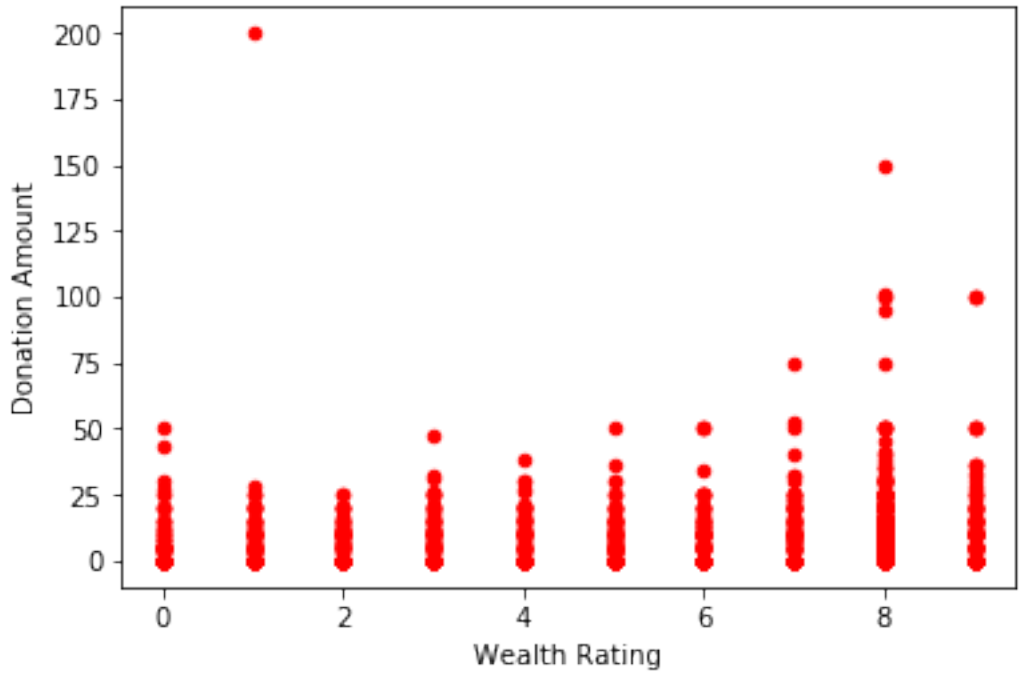

Figure 1 - Wealth Rating and Donation Amount

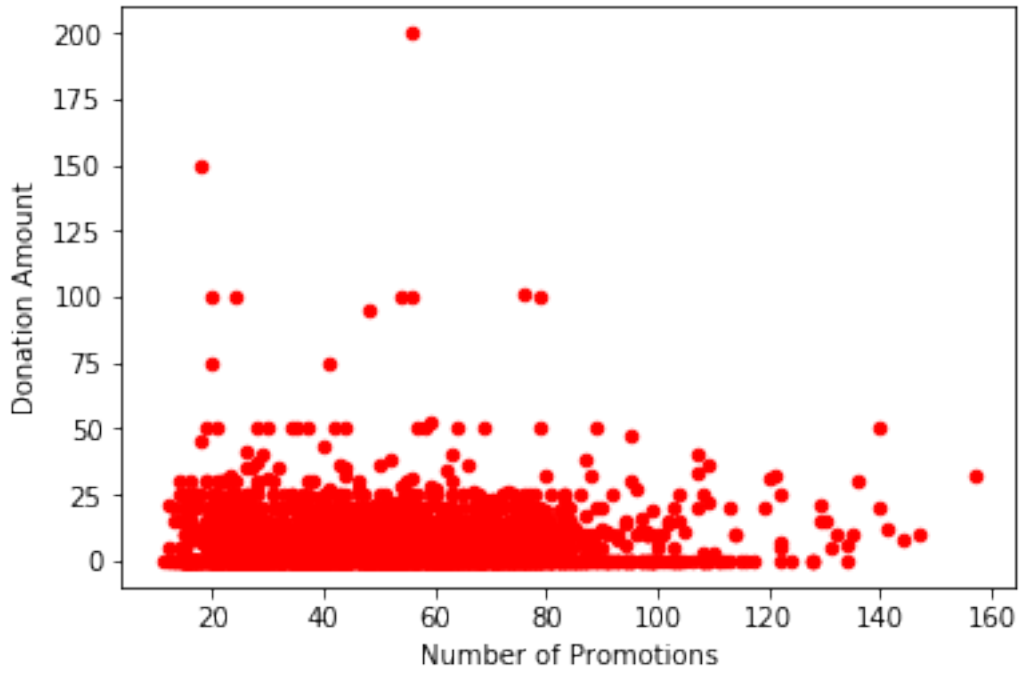

Figure 2 - Number of Promotions and Donation Amount

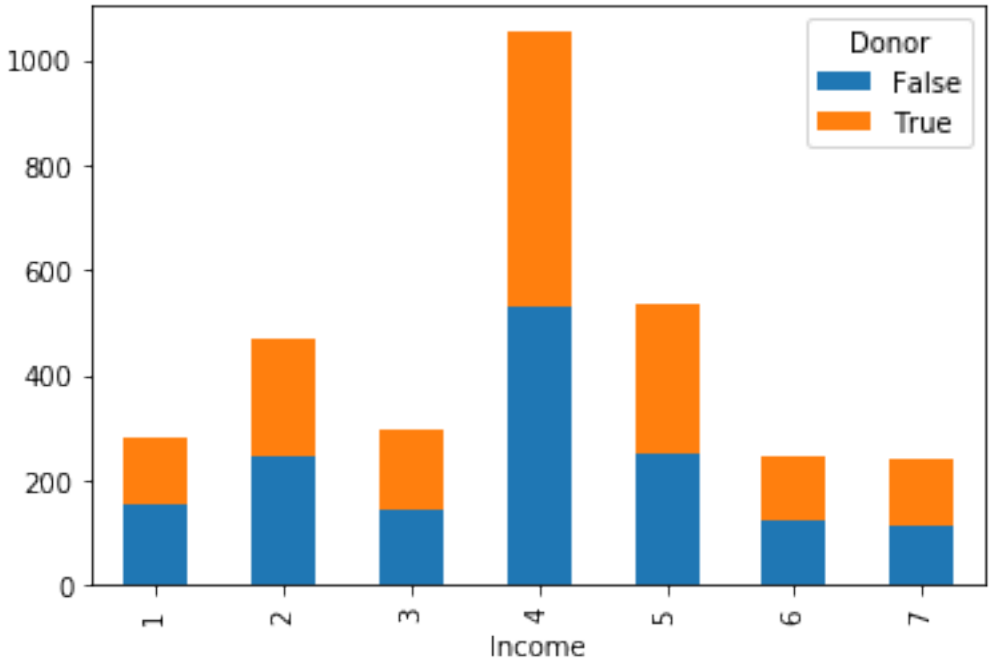

Figure 3 - Income Category and Donor Status (True or False)

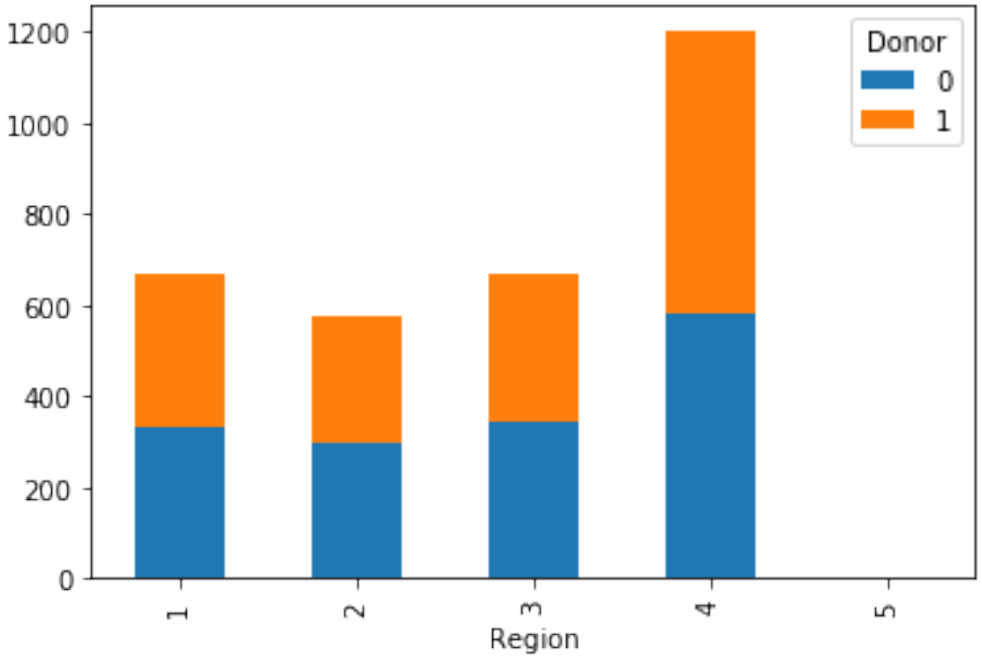

Figure 4 - Region and Donor Status (0 or 1)

Comparison Questions

The first comparison question explores the correlation between the variables in the dataset. Figure 5 below was created to visualize the correlations with yellow indicating higher values (except the diagonal yellow line which shows when a value has a perfect correlation to itself)

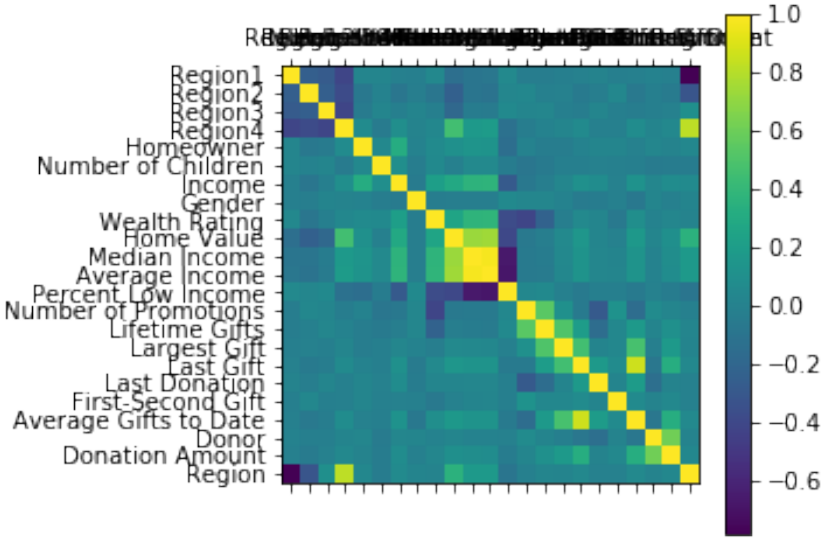

Figure 5 - Correlation of Variables (x labels are not visible due to formatting)

A yellow square in the lower right quadrant was of interest with an 86% correlation between Last Gift and Average Gift. So, the question to explore is: Why are these correlated and how can this information could be used to better understand donor behavior? Figure 6 below shows the relationship between these two variables. This could indicate that recent promotions may be encouraging people to give higher amounts.

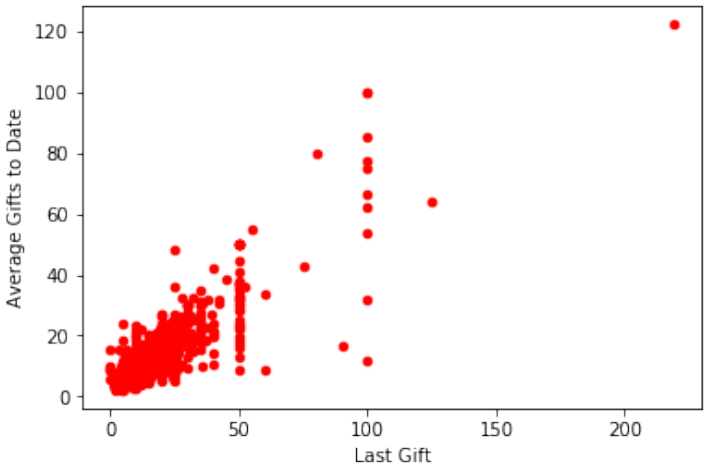

Figure 6 - Last Gift Amount ($'s) and Average Gifts to Date ($'s)

A second question to explore is the high number of people in Region 4 (the total is 1200). The column with (4,0) is Region 4, non-donors and (4,1) is Region 4, donors.

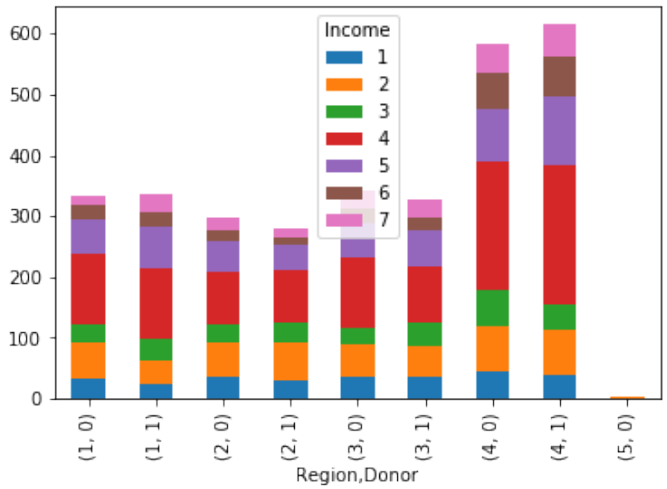

Figure 7 - Region, Donor Status by Number of Donors

Exploring further in Region 4 reveals some interesting data about the lifetime gifts by income level in Region 4. Wealth Rating 3 has the highest lifetime gifts with a mean of $203.58, while Rating 8 has the lowest lifetime gifts with a mean of $69.86. Since Region 4 has a disproportionate number of overall people, it would be interesting to investigate further why Rating 3 (lower income) and Rating 8 (higher income) have such a disparity in overall giving.

Wealth Rating	Lifetime Gifts
0	191.191818
1	158.225918
2	167.825500
3	203.578431
4	145.578947
5	175.380685
6	152.207031
7	169.071587
8	68.864701
9	166.650541

Table 3 - Region 4 Lifetime Gifts by Wealth Rating

Description of the Program and Output Files

The Python program to import the dataset, format the data and perform the analysis/visualizations was developed in Jupyter Notebook. The file Homework1.ipynb was uploaded to 2U with this Word document. Please see the Python code and comments in this file for additional details.

Conclusion

The donor dataset contains rich information that could be used by an organization to target existing donors and potential new donors. One observation that was revealed in the data is that Region 4 has the highest number of donors and non-donors (1200). Performing a deeper investigation into this region and developing marketing programs could deliver better results than "mass marketing" to all donors.