Direct-Mail Fundraising Modeling

Victor Feagins

Business Objectives and Goals.

The context of this work is modeling wither or not someone will donate or not for a direct mail fundraising campaign. We would like to mail to people who will donate so we don't waste money sending mail to people who do not donate. We would like to predict donation status with at least 50% accuracy.

Data Sources and Data used.

The data used for this task is from The American Legion a national veterans' organization that would like to improve their cost effectiveness of their marketing campaign. It is a history of their mailing efforts that has been weighted sampled to have 50% donors and 50% non donors. The reason for having a weighted sample instead of a random sample is because we would like our model to learn both donors and non donors characteristics. Since generally there are more non -donors that would skew our predictive model.

Type of Analysis performed: what, why, findings

The beginning of analysis I accessed the quality of the variables in the dataset and examined the relationship with donor. First, I examined contingency tables with the categorical variables. Summary of my results with the categorical variables

Categorical variables

Techniques used.

- Contingency tables: seeing if there is direction to variables
- Chi squared tests: testing if there was a relationship between variables

Summary

No or very little influence on Donations

- Zipcode variables were found not to be associated with donors. They are independent of donors

Small Influence on Donations

- Homeownership was found to have a weak relationship with donors.
- Females were found to donate more but it is also a weak relationship

Numeric variables

Techniques used.

- Histograms: Using to examine the frequency of the data
- QQplots: For accessing normality
- Shapiro Wilk test: Testing for normality
- Box-cox transformations: Transforming the data to make it more normal

- ANOVA: Testing if donor and non donors are different in relation to a variable
- Kruskal-Wallis test: Non parametric testing if donor and non donors are different in relation to a variable
- Mahalanobis distance: For outlier detection

Summary

Small influence on Donations

- Number of children: People who have more children tend to donate less but very weak relationship.
- Income: People who have more income lead to more donations but very weak
- Wealth: People who are wealthy will donate more but not all the time

No or very little influence on Donations

- Neighborhood Average Value: Does not have a strong relationship with donors.
- Median Family Income: Does not have a strong relationship with donors.
- Average Family Income: Does not have a strong relationship with donors.
- Percent Earning less then 15k: Does not have a strong relationship with donors.
- Number of months between first and second gift: Does not have a strong relation with donors.

Small Influence on Donations

- Number of promotions received: Seems that more promotions received will increase the chances of donating.
- Dollar amount of lifetime gifts: A greater lifetime of gifts slightly increases the chances of donating
- Largest gift: It is possible that the people who a huge donation are slightly less likely to donate again. The reason behind it is because they are one and done donator.
- Dollar amount of recent gift: If the recent dollar amount is large then people are less likely to donate again.
- Number of Months Since last donation: Harder to interpret. It seems that people wait longer donate more or that people donate seasonally.
- Average dollar amount of gifts to date: A larger amount means less likely to donate again.

None of the variables had very strong discriminatory power. Models that rely strong boundaries like discriminant analysis would not work. Logistic regression works well when classes are closer together and might perform well on this data set.

Exclusions

To begin our modeling efforts, I created 6 datasets.

Only influential variables are the variables that were found to have an effect with donations from the Type of analysis performed section. The Transformation will be in the Transformation. Outliers an outlier which is defined having a very high or very low Mahalanobis distance and that is only in some of the datasets trained.

Data Name	Variables	Transformation	Outliers

df.r	All of them	None	Included
df.r.o	All of them	None	Not included
df.m	Only influential	None	Included
df.m.t	Only influential	Yes	Included
df.m.o	Only influential	None	Not included
df.m.t.o	Only influential	Yes	Not included

Variable Transformations

The following Transformations were done in the transformation datasets.

Wealth.rich	If wealth > 5 then it is 1 or if no then 0
num_prom.log	Natural Logarithm of number of promotions
Lifetime_gifts.bc	BoxCox Transform of lifetime gifts
Largest_gift.bc	BoxCox Transform of largest_gift
Last_gift	Natural Logarithm of Last_gifts
Months_since_donate.sqrt	sqrt(max(months_since_donate+1) - months_since_donate
Avg_gift.bc	BoxCox.Transform of Average Gift amount

All transformations were done because all the variables listed here were not normal and right skewed. Except for Months Since Donation which was left skewed. And Wealth.rich which is bucketed into rich and poor.

Methodology used, background, benefits

With the 6 datasets I created I will be training and testing 3 types of models: Logistic regression, Decision Tree, and Random Forest. The main metric I will be assessing is classification accuracy.

Each model with have the same training and set test that is created with the associated dataset. Logistic regression will be the baseline model to compare other models to because it is the simplest model. Decision Tree and Random Forest will be considered because there might be variables or non-parametric relationship that those methods will capture.

Model performance and Validation Results

The following are the results of training and testing multiple models on the various data sets.

		Test Accuracy					
		df.r	df.r.o	df.m	df.m.t	df.m.o	df.m.t.o
Logistic regression		0.5433	0.5476	0.55	0.55	0.5663	0.5612
Decision Tree		0.5383	0.5646	0.5383	0.5383	0.5493	0.5493
Random Forest		0.5667	0.517	0.5383	0.55	0.5578	0.5578

We can see the highest accuracy is with the Random forest with an accuracy of .5667. Though all the models are very similar to each other in terms of accuracy. These all use the cutoff of .5 probability.

When using the respective models on the data set part of the contest we get:

Logistic Regression: Score .65

Decision Tree: Score .6

Random Forest: Score .591667

It seems logistic regression is the best. I believe the reason this is the case is because in terms of space the two classes are very close together. The tree method split the data up by space but if the data is very close together then it will underperform. It didn't seem to be important to transform the variables here but getting rid of outliers was a good decision for the logistic regression. Also, variable selection was good idea. The logistic regression performed better with the smaller subset of variables. Whereas the tree methods work better with more variables. There is probably more that could be done like standardization as well as trying penalized logistic regression.

Recommendations

I recommend gathering more variables that can discriminate the data more. As it stands the variables do not explain the relationship between donors and not donors a tremendous amount. For example, new variable that could be extracted are perhaps about veteran populations or military bases nearby. As it stands the current model logistic regression is currently predicting above 50% which was the goal. This could be further improved by perhaps considering other transformations or in my approach I either did no transformation or all transformation and it is possible a mix of variables could improve the model.