Philanthropy Statistical Analysis Report

T. Lang

QSO 510 – Quantitative Analysis

Dr. D. Babb

2 April 2023

**Introduction**

The philanthropic organization Paralyzed Veterans of America (PVA) is launching a donation campaign to help disabled veterans. One of the campaigns used by PVA sends greeting cards and address labels to potential donors. PVA would like to know the amount that a donor is likely to give based on the data they have collected from past donations.

**Analysis Plan**

**Factors**

The dataset provided by PVA for analysis includes 26 predictor variables and one response variable for 3648 donors. The response variable is GIFTAMNT (gift amount), which indicates the amount of money donated in the last solicitation campaign. The predictor variables are divided into two categories: measures based on zip code and measures specific to individual donor.

The measures that are based on zip code are based on the percentage of the target population that live in the same zip code as the donor. These include male veterans, Vietnam veterans, WWII veterans, local government employees, state government employees, and federal government employees. Individual donor information includes demographic information, such as age, number of children in the home, and having a published home number, as well as history of giving to various promotions and a marketing cluster code. One data point is the key identifier, CONTROLN.

Individual factors or a combination of the factors may influence the amount of the donation. A preliminary assessment of the data can be seen in Appendix A. The measures of central tendency for the target variable shows the mean gift amount is 15.716623. The median gift amount is 14. The mode, or most frequent, gift amount is 10. There were two donations in the amount of $200 and one of $150, which appear to be potential outliers. Further modeling will determine whether those three donations may negatively influence the outcomes and need to be addressed (Ferguson, 2018). These values may be deleted, depending on the analysis as outliers, or imputed with another value, such as a mean for the variable of gift amount.

The histogram of the frequency of each of the donation amounts show a significantly greater number of smaller donations than larger (see Appendix A, figure 2). There are no clear patterns in the scatterplots for each of the predictor variables that lend to a clear correlative relationship with the outcome variable, though some suggest there may be significance, as most recent amount given appears to have a near-linear relationship with amount given in the most recent and average donations (Appendix A, figures 22 and 23).

There may be a relationship with greater donations when fewer government employees live in the same zip code as the donor (Appendix A, figures 8-10). The fewer card promotions received appears to have potential correlation with a greater amount of a donation, with a slight right skew (Appendix A, figure 14). These variables would all be considered for further analysis as well as any interactions.

**Problem Statement**

The PVA would like to identify the factors that contribute significantly to the amount a person donates in order to predict the amount of a donation received based on these inputs.

**Strategy**

In order to predict the amount of a donation based on variables given, a predictive model will need to be created. Given the number of variables and unclear impact each has on the outcome variable of GIFTAMT, a multiple regression model will be built. Since the outcome variable is a continuous variable a linear regression will be used, as a logistic regression would be appropriate for a categorical outcome variable. Following the CRISP-DM methodology, the data will be prepared and modeled (Rodrigues, 2020). Missing variables, outliers, and inconsistencies in the data set will be noted and handled.

The variables will be evaluated for relationships to the outcome variable as well as potential interactions. The final model will allow values for the predictor variables to be entered for future potential donors and produce a predicted value for the donation expected. The number of predictor variables in the model will be sufficient to produce statistically significant results, yet not excessive in number, as this would reduce reliability of the final model (Vidyashri, 2022). The correlation coefficients will determine the strength of the relationship between the predictor variables and the outcome variable in the final model, which will also help guide targeted marketing in the future.

**Statistical Methods**

**Method**

The statistical model that would be most appropriate for this data would be a multiple regression. Using the CRISP-DM method for evaluating this problem, the data will be understood, prepared, and a model created (Shearer, 2000). Knowing that the model created is attempting to predict the amount a donor would give based on data from previous donation campaigns, different predictive models were compared.

Various predictive models can be divided into two categories based on the expected outcomes: classification or forecasting. In the case of PVA, the goal of the model is to identify the amount of the donation for the person, a forecast, and not whether or not they will donate, a classification. For forecasting models, four different types are identified (Schmidt, 2023): straight line, moving average, simple linear regression, and multiple linear regression.

As described in the article by Schmidt, straight line and moving average are forecasting methods using historical data to anticipate the trend. The regression models use observations to compare variables based on observations. Further, the difference between the simple linear regression and multiple linear regression is how many variables are compared

**Method Justification**

For the model for PVA, the goal is to create a model that can identify the statistically significant variables, apply a weight to those values, and be input into a formula that outputs a predicted donation value with a specified level of confidence. An assumption for using this model is that there is a linear relationship between the predictor and outcome variables (Sharpe, et al., 2019). Additionally, the sample is assumed random and variables independent of each other, not collinear. The variables are also assumed to have similar variances and the results of the model distributed normally (Bommae, 2015). Also pointed out by Bommae (2015), this allows for valid results to be drawn from the hypothesis testing.

**Data-Driven Decisions**

**Process**

A multiple regression model develops coefficients for each of the significant predictor variables, allowing inputs for these variables to produce the response variable (Sharpe, et al., 2019). For PVA, the response variable is the donation amount. Predictor variables will be narrowed down from the full dataset provided.

Stepwise regression aims to reduce the number of independent variables that are used in the model to only those who can predict accurately and reliably through the stepwise process, which is outlined by Zaiontz (2018). A significance level is established and linear regression models are built for each of the independent variables. Provided there is at least one variable with statistical significance, the variable with the lowest p-value is identified as the starting point of the model.

The model starts with the independent variable of the lowest p-value. The model is then tested with that variable (for the purposes of this description, noted as x1) and each of the other variables, so x1 + x2, x1 + x3, etc. Should one of those combinations improve upon the model, resulting in a p-value lower than x1 alone, that will be the start of the next round. This continues until the addition of another variable does not improve upon the model.

Some sources note limitations of using a stepwise regression. . Ravelo (2022) indicates a greater Type 1 error with stepwise regression. Issues with overfitting the model, not accounting for interaction of variables, and limited accountability for exchange of variables that may be weighted equivalently in other multiple regression models, such as year over year change compared to last year’s value and this year’s value.

Smith (2018) points out that big data is more problematic using this method. The data set for PVA is not anticipated to encounter these issues to a significant level as the data set is much smaller than present with big data and will be mitigated by a p-value that allows a confidence level that is tolerable to the organization.

**Data Mining**

Data mining as a process is important to the overall reliability of the project. As Sharpe, et al. (2019) point out, the key components of data mining that make it different from the analysis itself include: the size of the data set that is pulled, the mining process is exploratory and not results-driven, the data is “happenstance”, the results are actionable, and the modeling choices are automatic. The process of data mining are to, without bias, gather the data from the source without any expected direction or outcome.

The metadata of the dataset is also important. As pointed out by Sharpe, et al. (2019), knowing when the data was pulled, the way that it was collected, and other background information on the variables, or the metadata, is important to the data understanding and the ability to prepare the data. Knowing if the data was collected through surveys, telephonic or mail, the year the data was collected, whether there were other mitigating factors impacting the data collection, all can be relevant to the interpretation of the information. For example, should the information be gathered by interviewing individuals at a political rally or a protest, the answers may be impacted.

**Structured vs. Unstructured**

As further classification of the problem presented by PVA, the question that is presented is a structured problem. The question that is posed will create a process that can be used repeatedly. This is not a novel and infrequent problem that requires further investigation to get a better understanding of it (Gordon, 2022), such as some ethical decisions and more ambiguous decisions.

**Variables**

The variables that are being evaluated for the model are those that have been pulled for the data set by PVA. In running the initial analysis, the variables will identify whether they can adequately predict the amount of the donation. Since “the actual variables used in the final prediction model should be determined by analysing the data,” according to Chowdhury and Turn (2020), it is to be determined through the analysis whether there are variables that are significantly impacting the prediction of a donation amount. The specific variables may show significance as they are able to identify a myriad of factors about individuals and their history of donating. Should these factors identified correlate, the model will identify which predictive variables and the degree to which each does influence the outcome.

In the data set there were no missing values. All minimum values were valid, in that there were no negative values for any variables in which that would be irrational. All maximum values and means were also in a reasonable range.

**Recommended Operational Improvements**

**Summary of Results**

A stepwise regression was completed to create a predictive model that indicates the predicted amount of a donation for PVA. The resulting formula is:

GIFTAMNT = 5.2850428 + 0.60251041 LASTGIFT + 0.38301912 AVGGIFT + -0.099172504 NGIFTALL + -0.20620422 MINRAMNT + -0.75261254 HOMEOWNER + -0.016161821 CLUSTER2 + 0.052265251 STATEGOV. This formula allows for the values of the variables to be plugged in and a predicated amount of gift produced as a result.

The regression identified the significant variables as the amount of the last gift, the average gift given, the number of gifts given, the minimum amount, whether they were a homeowner, a cluster of variables around the location of where they lived, and whether they lived in proximity of a state government official. The coefficients of those variables indicate the degree to which they influenced the amount of the gift predicted.

The variables that were chosen have statistically significant values compared to a p of 0.05. The chart of p-values can be seen in Appendix B, Figure 1, which shows all p-values remained less than 0.05. In the same appendix, Figure 2 indicates the estimated influence of the parameter for each of the variables, which become the coefficients in the equation, as well as the t-statistic and p-values for those variables. The Analysis of Variance (ANOVA) for the model is depicted in Appendix B, Figure 3. The F-statistic for this model is 620.21431. This value indicates there is a statistically significant difference between the variance between groups than within the same group (Feldman, 2018). For our model, this indicates that the group of variables chosen for the model are statistically significant compared to other groups of variables.

**Solution**

For PVA, the predictive model indicates that with the values of the specific data points a predicted amount can be identified for an individual to donate. Using this formula the gift amount can be estimated for a person whose last gift was for $100, gives an average gift of $50, has given 3 gifts in their lifetime, with a smallest gift of $20, is a homeowner, is categorized in a Cluster2 of 30, and lives near 5 state government officials by plugging those values into the formula: GIFTAMNT = 5.2850428 + 0.60251041 LASTGIFT + 0.38301912 AVGGIFT + -0.099172504 NGIFTALL + -0.20620422 MINRAMNT + -0.75261254 HOMEOWNER + -0.016161821 CLUSTER2 + 0.052265251 STATEGOV. The resulting value for a predicted gift amount is: $67.80.

Compared to an individual who has not given a gift before, is not a homeowner, does not live near a state government official, the value of a gift would be at most $5.29 and decreased by 0.16 times the Cluster2 assignment. Knowing this information, the individuals who are more likely to donate a significant amount could be identified for the next giving campaign. This would reduce wasteful spending for solicitation of donations. Additionally, PVA would be able to predict the total of the donations received from a given campaign.

The Paralyzed Veterans of America would benefit from having a predictive model such as this one for the more effective marketing and donation campaigns. Using stepwise regression they will be able to reduce cost and anticipate the donation sums. The model created indicates the values predicted align with the actual donations received in the past, as shown in Appendix B, Figure 5.

**Appendix A: StatCrunch Calculations and Plots**

**Figure 1.**

Central Tendency Measures on Gift Amount

**Table

Description automatically generated**

**Figure 2.**

Histogram of Number of Gifts by Gift Amount

**Chart, histogram

Description automatically generated**

**Figure 3.**

Scatterplot of Gift Amount by Homeowner Status**Chart, line chart, box and whisker chart

Description automatically generated**

**Figure 4.**

Scatterplot of Gift Amount by Hit Rate

Chart, scatter chart

Description automatically generated

**Figure 5.**

Scatterplot of Gift Amount by Variable MaleVet

Chart, scatter chart

Description automatically generated

**Figure 6.**

Scatterplot of Gift Amount by Variable Vietnam Veterans

Chart, scatter chart

Description automatically generated

**Figure 7.**

Scatterplot of Gift Amount by Variable WWII Veterans

Chart, scatter chart

Description automatically generated

**Figure 8.**

Scatterplot of Gift Amount by Local Gov Variable

Chart, scatter chart

Description automatically generated

**Figure 9.**

Scatterplot of Gift Amount by State Gov Variable

Chart, scatter chart

Description automatically generated

**Figure 10.**

Scatterplot of Gift Amount by Fed Gov Variable

Chart, scatter chart

Description automatically generated

**Figure 11.**

Scatterplot of Gift Amount by Lifetime Card Promotions

Chart, scatter chart

Description automatically generated

**Figure 12.**

Scatterplot of Gift Amount by Date of Most Recent Promotion

Chart

Description automatically generated

**Figure 13.**

Scatterplot of Gift Amount by Total Promotions Received

Chart, scatter chart

Description automatically generated

**Figure 14.**

Scatterplot of Gift Amount by Number of Card Promotions Received in Past Year

Chart, scatter chart

Description automatically generated

**Figure 15.**

Scatterplot of Gift Amount by Number of Promotions in Past Year

**Chart, scatter chart

Description automatically generated**

**Figure 16.**

Scatterplot of Gift Amount by Number of Gifts Given in Lifetime

Chart, scatter chart

Description automatically generated

**Figure 17.**

Scatterplot of Gift Amount by Number of Gifts Given to Card Promotions in Lifetime Chart, scatter chart

Description automatically generated

**Figure 18.**

Scatterplot of Gift Amount by Amount of Smallest Gift Given

**Chart, scatter chart

Description automatically generated**

**Figure 19.**

Scatterplot of Gift Amount by Date of Smallest Gift Given

**Chart, scatter chart

Description automatically generated**

**Figure 20.**

Scatterplot of Gift Amount by Amount of Largest Gift Given

**Chart, scatter chart

Description automatically generated**

**Figure 21.**

Scatterplot of Gift Amount by Date of Largest Gift Given

**Chart, scatter chart

Description automatically generated**

**Figure 22.**

Scatterplot of Gift Amount by Most Recent Gift Amount

Chart, scatter chart

Description automatically generated

**Figure 23.**

Scatterplot of Gift Amount by Average of Gifts Given

Chart, scatter chart

Description automatically generated

**Figure 24.**

Scatterplot of Gift Amount by Control Number (Unique Record Identifier)

Chart, scatter chart

Description automatically generated

**Figure 25.**

Scatterplot of Gift Amount by Whether Donor Has a Published Home Phone Number

Chart

Description automatically generated

**Figure 26.**

Scatterplot of Gift Amount by Marketing Cluster Code

Chart, scatter chart

Description automatically generated

**Figure 27.**

Scatterplot of Gift Amount by Number of Children Living at Home

Chart

Description automatically generated

**Figure 28.**

Scatterplot of Gift Amount by Age

Chart, scatter chart

Description automatically generated

**Appendix B**

**Figure 1.**

Chart of variable significance values, Root Mean Square Error, R-squared, and R-squared adjusted.

Table

Description automatically generated

**Figure 2**.

Parameter estimates of variables with t-statisticsTable

Description automatically generated

**Figure 3**.

ANOVA table

Table

Description automatically generated

**Figure 4.**

Summary of Fit statistics

Text

Description automatically generated

**Figure 5.**

Plot of Actuals vs. Predicted

Chart, scatter chart

Description automatically generated

**References**

Bommae, K. (2015, September 14). *Should I always transform my variables to make them normal?* Research Data Services + Sciences. Retrieved March 19, 2023, from https://data.library.virginia.edu/normality-assumption/

Chowdhury, M. Z., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1). https://doi.org/10.1136/fmch-2019-000262

Feldman, K. (2018, November 7). *F-value (ANOVA)*. isixsigma.com. Retrieved April 1, 2023, from https://www.isixsigma.com/dictionary/f-value-anova/#:~:text=The%20F%2Dvalue%20is%20the%20ratio%20of%20your%20between%20group,difference%20in%20your%20group%20means.

Ferguson, K. (2018, March 6). *When should you delete outliers from a data set?* Humans of Data. Atlan. Retrieved February 19, 2023, from https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/

Gordon, J. (2022, April 16). *Structured problem vs unstructured problems - explained*. The Business Professor, LLC. Retrieved March 19, 2023, from https://thebusinessprofessor.com/en\_US/management-leadership-organizational-behavior/structured-problem-vs-unstructured-problem

Ravelo, C. (2022, November 23). *Stepwise regression: What is it and should you use it?* Statistics Solutions. Retrieved March 19, 2023, from https://www.statisticssolutions.com/stepwise-regression-what-is-it-and-should-you-use-it/

Rodrigues, I. (2020, February 17). *CRISP-DM methodology leader in Data Mining and Big Data*. Medium. Retrieved February 19, 2023, from https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781

Sharpe, N. R., De Veaux, R. D., & Velleman, P. F. (2019). *Business statistics* (4th ed.). [e-book] Pearson.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22. Retrieved March 17, 2023, from https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Smith, G. Step away from stepwise. *J Big Data* 5, 32 (2018). https://doi.org/10.1186/s40537-018-0143-6

Vidyashri, M. H. (2022, January 20). *Advantages and disadvantages of regression model*. VTUPulse. Retrieved February 19, 2023, from https://www.vtupulse.com/machine-learning/advantages-and-disadvantages-of-regression-model/

Zaiontz, C. (2018). *Stepwise Regression*. Real Statistics. Retrieved March 19, 2023, from https://real-statistics.com/multiple-regression/stepwise-regression/