**Introduction**

The use of statistics and analytics to develop insight and solve real world problems continues to increase in popularity. Specifically, with the emergence of technological and advances and improves in computational analysis, the branch of predictive analytics known as machine learning has taken a firm hold in core application of data analysis. Use cases range from a variety of approaches and can employ real-time analytical insights to produce both predictive and explanatory information. This write-up will explore the use of machine learning by a fictional charitable organization to develop a cost-effectiveness model by exploring the data, developing insights and applying them to a wide range of models.

**Objective**

Through the use of machine learning practices, the intent is to develop the most cost-effective model related to a mailer marketing campaign. There is, on average, a 10% response rate to the marketing and the average donation for each response is $14.50 while each mailer has a cost of $2.00. The organization has collected data on mailer response rate and the average donations and through statistical analysis hopes to develop a classification model that effectively targets donors most likely to respond and in turn will maximize profits.

**Data**

The dataset consists of 3984 training observations, 2018 validation observations, and 2007 test observations. Fortunately, weighted-sampling has been employed so that each grouping of observations has a proportional inclusion of both donors and non-donors. The data consist of 17 variables (not including index nor either dependent variables), including both continuous and categorical. The variables and descriptions are shown in Table 1.0:

|  |  |
| --- | --- |
| **Name** | **Description** |
| **ID** | Index of observation |
| **REG1-4** | Geographic Region |
| **HOME** | Homeowner |
| **CHILD** | Number of Children |
| **HINC** | Household income |
| **GENF** | Gender |
| **WRAT** | Wealth rating |
| **AVHV** | Average Home Value |
| **INCM** | Median Family income |
| **INCA** | Average Family Income |
| **PLOW** | Percent low income in neighborhood |
| **NPRO** | Number of promotions received |
| **TGIF** | Dollar amount of lifetime donation |
| **LFIG** | Dollar amount of largest donation |
| **RGIF** | Dollar amount of most recent donation |
| **TDON** | Months since last donation |
| **TLAG** | Months between first and second donation |
| **AGIF** | Average dollar amount of donation |
| **DONR** | Class response variable |
| **DAMT** | Prediction Response variable |

*Table 1.0*

**Approach**

Prior to model evaluation, a through Exploratory Data Analysis (EDA) will be conducted. This will include an evaluation of each variable to determine normality, missing values, type, outliers, relation to dependent variable and perceived useful need for modeling. The process will include observing multiple plotting and graphical evaluations. There will also be a need to evaluate opportunities of both variable transformations, imputing, and creation of additional variables. Following EDA, multiple models will be built, including: (1) Linear Discriminate Analysis, (2) Logistic Regression, (3) K-Nearest Neighbor, (4) Least Squares Regression, (5) Best Subset Using K-Crossfold, (6) Ridge Regression, and (7) LASSO.

Following the model building, a final comparison of each model using stastical measures such as Mean Prediction Error and residual analysis, a final classification model for the dependent variable DONR and a final prediction model for the dependent variable DAMT will be selected for use by the charitable organization to maximize profits in the marketing campaign.

**Exploratory Data Analysis**

**Models**

**Analysis of Models and Final Selections**

**Conclusion**