Capstone Milestone Report

KDD Cup 1998

1. Data Source

This project uses the dataset in The Second International Knowledge Discovery and Data Mining Tools Competition, held in conjunction with KDD-98.

The dataset was provided by the Paralyzed Veterans of America (PVA), a not-for-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. PVA is also one of the largest direct mail fund raisers in the country. The analysis dataset includes:

* A subset of the 3.5 million donors who received fund raising appeals from PVA in 1997.
* A flag to indicate respondents to the appeal and the dollar amount of their donation
* PVA promotion and giving history
* Overlay demographics, including a mix of household and area level data.

This project uses the Training dataset in the competition, which includes 95412 rows and 481 columns.

* Column TARGET\_B: a target variable, binary variable indicating whether or not the record responded to the promotion of interest (‘97NK’ mailing)
* Column TARGET\_D: contains donation amount (dollar), and is only observed for those that responded to the promotion.

1. Project objectives

The objectives of this project are two-fold:

* Predict who will respond to the 97NK mailing promotion (a classification problem)
* For those who are predicted to be donors, predict the amount of their donation (a regression problem)

1. Deep Dive into the Data:
2. Challenges with the data set:

* The data set is highly imbalanced, with only 5% response rate.

This creates an issue for classification models, as the positive records are overpowered by the negative records, thus causing all predictions to be negative (in the case of Logistic Regression).

* About a quarter of the columns have more than 30% missing values. Imputing the missing values might introduce bias, thus I decided to drop all these columns.
* The data set has too many features, thus feature selection is crucial in order to have a good model.

1. Data cleaning and wrangling tasks

* Create a balanced subset of the data set by keeping all the positive cases and sample an equal subset of the negative cases.
* Check columns with missing values and drop those with >30% missing values
* Drop some variables which are believed to be unimportant
* Split some categorical variables which were coded as matrix into components
* Impute missing values with the mode for numeric variables, and with the most frequent value for categorical variables
* Create some new variables based on existing ones, i.e., some date columns.
* Create dummy variables from categorical variables
* Normalize the data set in order to perform feature selection using variance threshold.
* Performed feature selection using variance threshold, then recursive feature elimination (RFE) to reduce the number of features used in the model.

1. Preliminary Findings

* Using a balanced subset (where the response rate is almost 50%), I compared some characteristics between donors and non-donors.

It appears that donors and non-donors have very similar demographic attributes (age, gender, income, socio-economic status, etc…) as well as giving history (last date and first date of donation, total number of donations to date, amount of most recent donation).

* I tried several classifiers (Logistic Regression, Decision Trees, Naïve Bayes, Random Forests) on the reduced data set, and it appears that the Logistic Regression performs the best in predicting who will be donors (F1 score is 0.56).
* Using the samples who are predicted to be donors by the Logistic Regression model, I continued to predict the amount of their donations. Tried several regression models (Linear Regression, Ridge, Lasso, Elastic Net, Regression Tree), and Regression Tree performed the best with RMSE of 0.12.