Predicting Blood Donors

Scott Herman

July 30th, 2017

Prepared for Predict-413: Time Series Analytics and Forecasting
Northwestern University Masters in Science, Predictive Analytics
Data-Driven File Submission: **sherman_413_nw.csv**

Introduction

The objective for this analysis to develop a classification model that will accurately predict whether or not an individual donates blood within a given window of time. The dataset collected is a mondified version of the donor database of Blood Transfusion Service Center in Hsin-Chu City, Taiwan, and contains a total of 576 observations along with an initial set of five variables describing various characteristics specific to each unqiue record.

Significance

According to the Taiwan Blood Services Foundation, the country is currenlty experiencing a severe shortage of blood of all types, and the national average of blood reserves are limited to a six day supply level. Additionally, tighter screening of blood donors in recent years has futher eroded the pool of potential blood donors across the globe. In places like Taiwan, the amount of blood donation is far fewer than what is required, and there might not be enough blood in reserve in the event of an emergency. In order to keep up with this demand, data-driven systems for monitoring and predicting donations can improve the entire supply chain, ensuring that more patients get the blood transfusions they need - 1. The results of this analysis aim to better understand how to effectively predict potential blood donors' behavior which would help increase the number of blood donors, along with the available blood supply, helping to save thousands of lives every day.(Holdershaw, 2005)

Literature

Previous research on modeling blood donation behavior does exist on the subject. Cheng Yeh, King-Jang Yang, and Tao-Ming Ting expanded upon the Recency, Frequency, and Monetary Value (RFM) model, which is a behavior-based model generally utilized for customer relationship marketing. Their reserach introduced a comprehensive methodology to discover knowledge for selecting targets for direct marketing from a database, and utilized the Bernoulli sequence in probability theory, to derive a formula that estimated the probability that a customer would buy the next time, and the expected value of the total number of times that the customer will buy in the future.

Another approach presented a profiling system that profiles indviduals based upon their donations patterns. This methodology identifies different donor groups based upon the donation frequency. These groups are mutually exclusive and present donors in different stages of the donor lifecycle. The donor types range from first-time, regular, returning, lapsing, inactive, and all the way to stopped donor. This type of classification apprach enables blood banks to understand the current activity level of their database of donors.(Veldhuizen, 2013). An additional methodology examined was a Dynamic Logistic Regression for Binary Classification which is a procedure proposed for cases when there is uncertainty about the model to use and the parameters within a model change over time (McCormick, 2011).

This type of model acounts for model uncertainty and enables estimates to be updated as additional data become available. This type of methodology seems to be particularly relevant and applicable to our blood donor data in predicting donation behavior.

The table below gives the definitions of each variable within our initial data set. Note, we have modified the variable names for ease of understanding and interpretation. It appears that we have four quantitative variables that are continuous in nature describing various units of time. Our respose variable, on the other hand, appears to be categorical.

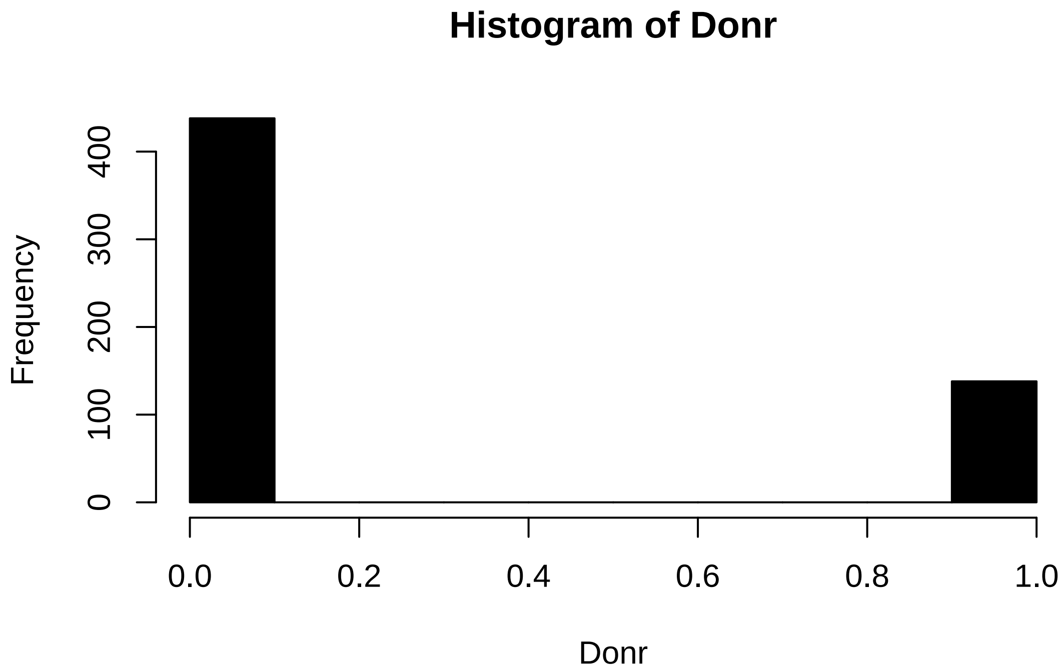
Table 1: Data Definitions

Variable	Description
ID	Unique ID of current/former donor
Recency	Number of months since the last donation
Frequency	Number of donations
Volume	Total volume donated
Time	Months since first donation
Donr	Made donation in March, 2007

Data Exploration

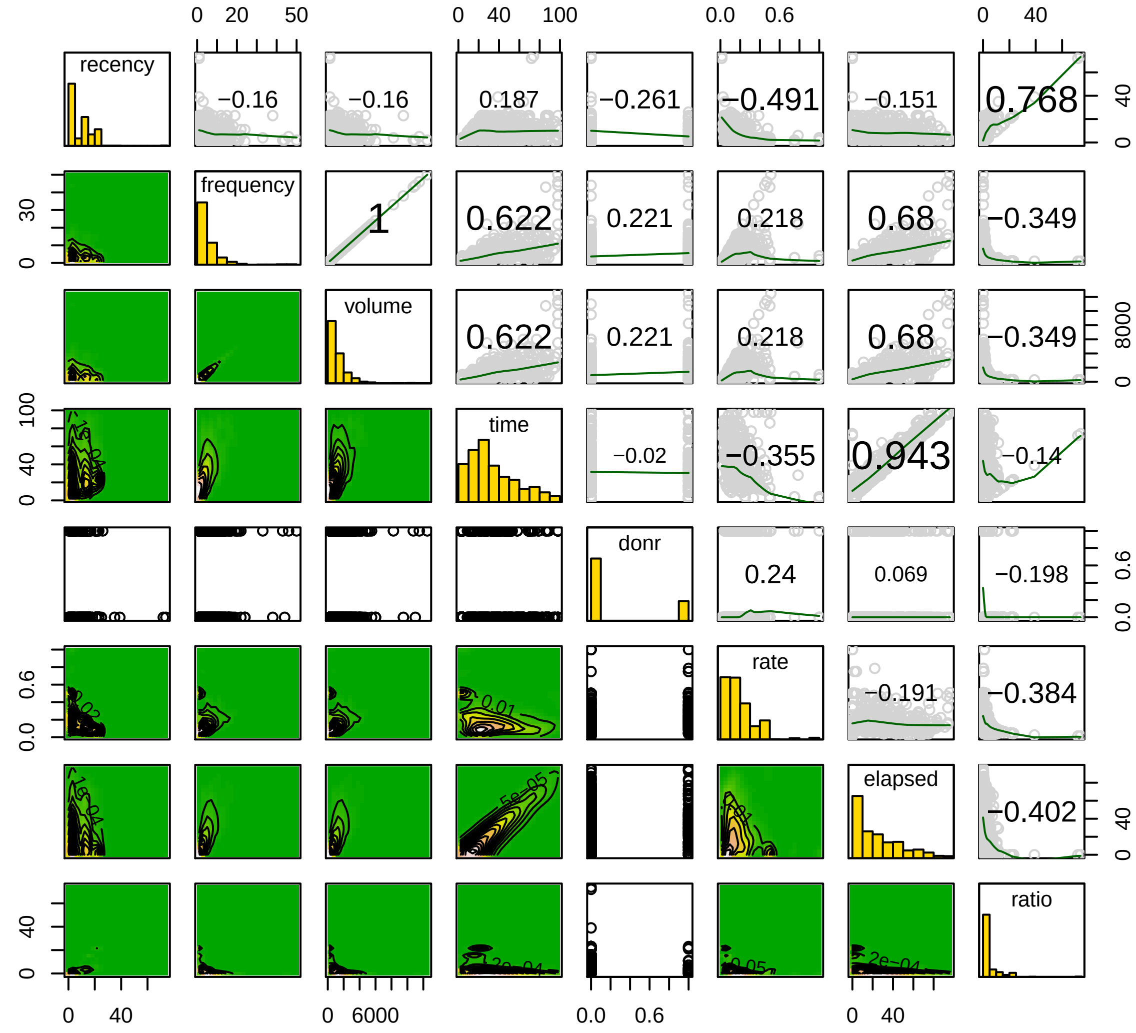
This analysis begins by understanding the variables within our data set along with their corresponding observations in an attempt to identify the structure and quality of our data set. First, we will examine the distribution of our response variable. Then, we will plot the distributions of our predictor varibles, check for any missing of influential observations, and examine their correlation to our response.

The histogram below, reveals the distribution of our target variable. Right off the bat, we can see that our response variable, Donr, is binary in nature, and only contains two potential responses; a one, or a zero. Those records with a "1" indicate that an indvidual did donate blood in March, 2007, while records represented by a "0" mean that he/she did not donate blood during this frame. Additionally, this visual reveals that Donr posesses a zero-inflated distribution with only 24% of the observations indicating a positive response for donating blood. These are important distinctions that we will need to account for once we begin developing our classification models.



Next, we examine the relationships between each of our variables which is given by the Scatterplot Matrix, below. The first thing that stands is out the nearly perfect linear relationship between Volume and Frequency. This makes sense, as the Volume of the total blood donated for each indvidual is going to increase each time they frequent the donation center, yet is fixed for each given visit. This also indicates that these variables have a multicollinear relationship and tells us that we probably want to remove Volume from our modeling efforts to avoid any unnecessary error in the final results.

Additionally, this visual indicates some interesting correlations to our response variable. Here, we see that Recency has a -0.261 correlation to Donr, while Frequency shows a 0.221 correlation. This means that those who donate more frequently are 22% more likely to donate again, while the recency since their last donation tend not to donate 26% of the time. We also note that Time shows very little correlation to our response variable. Lastly, we can confirm that there were no missing observations identified in our data.



Data Preparation

After our initial data exploration, we move on to preparing our data for our model development. Although, we have decided to drop Volume from our predictor set, and have decided not to adjust the remaining set of variables to maintain the consistency of our initial data set. In an attempt to increase the accuracy of our predictions, we have moved forward with creating three new additional variables. These newly created variables are Rate, Ratio, and Elapsed. Rate was calculated by dividing the Frequency by Time, giving an indication of how often, on average, a given person donates. Ratio was calculated by dividing Recency by the Frequency. Finally, Elapsed was created experimentally, to try and boost the predictive power of the duration measured of our donor base. The table below summarizes our final set of predictor variables.

Predictor Variables	Correlation to Donr
Recency	-0.2612337
Frequency	0.2206153
Time	-0.01981889
Rate	0.2402734
Ratio	-0.197649
Elapsed	0.0687556

Model Development and Results

In proceeding with our model development, we will attempt to build four separate classification models in predicting whether or not an individual will donate blood on the given date. For this stage in the process, we created separate training and validation sets for evaluating each model. We will begin by fitting each model using the training set, and evaluate their results on the validation data. The logarithmic loss metric will be used as our evaluation criteria in assessing the strength of each model developed. The logarithmic loss provides a steep penalty for predictions that are both confident and wrong. Therefore, the goal is to minimize the log loss score, and the model indicating the lowest score on our validation set will then be deployed on the test set as our final predictions for the Datadriven competition submission.

The first model developed was a Logistic Regression Model. This model was chosen because Logistic Regression models are well suited for classification problems and allow us to identify the probability that a given record will fall into either the 'Yes', or 'No' category for the response. Next, we developed a Decision Tree model as this type of procedure also works well when dealing with qualitative responses. In attempting to improve upon these initial results, we utilized a random forest Bagged model, and lastly a boosted model. For each model developed, we utilized the same set of six predictors to gain an understanding of the impacts seen with each given procedure. The results from each of these classification models is listed below.

Model	Log Loss Score
Logistic	0.5044102
Tree	0.4750621
Bag	0.294752
Boost	0.3331708

Conclusion

After reviewing the performance of each of our models, it appears that the bagged model achieved the lowest logarithmic loss when evaluated against the validation data we partitioned from our initital data set. Despite performing well on the validation data, the boosted model predictions actually performed better in the final submission with the test data. Although we might have assumed the bag model possessed the strongest predictive accuracy, this result was not too surprising given that the Log Scores of these two models were fairly close. The final boosted model achieved a log loss of 0.5005 on the final test set, while the model bagging procedure resulted in a score of 0.5574.

The results from this analysis could be utilized to identify further improvements that promote additional accuracy in predicting blood donations. In building upon these models it might make sense to test out a number of adjustments including transformations to both the set of predictors, as well as looking at new calculations to create new variables that may increase each model's predictive power.

Citations

1. Yeh, I-Cheng, King-Jang Yang, and Tao-Ming Ting. "Knowledge discovery on RFM model using Bernoulli sequence." Expert Systems with Applications 36.3 (2009): 5866-871.

2. Chen, Yen-Liang, Mi-Hao Kuo, Shin-Yi Wu, and Kwei Tang. "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data." Electronic Commerce Research and Applications 8.5 (2009): 241-51.

3. Veldhuizen, I. J. T. "Blood donor profiling using donation patterns." ISBT Science Series 8.1 (2013): 233-37.

4. Raghuraj, Rao, and Samavedham Lakshminarayanan. "Variable predictive models-A new multivariate classification approach for pattern recognition applications." Pattern Recognition 42.1 (2009): 7-16.

5. Mccormick, Tyler H. "Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification." Biometrics 68.1 (2011): 23-30.