Executive Summary

Blood donation shortage prediction

With increasing emphasis on using data / predictive analytics in support in blood drive / healthcare organizations, I describe in my analytics collecting and analyzing data for Blood donation given at a blood drive. I describe the process in identifying the Months since Last Donation from Number of Donations at the blood donation center based on past donation data. My analysis is divided into two parts. The first part focuses on data analysis which studies the context of data itself. I have identified the relationship among donor's and classify them into class 1 and frequency of donation that are donated in March 2007, the relationship between donation frequencies and how often they donated. The second part which is class 0, it focuses on analysis donor that did not donated blood in March 2007 the findings derived from the first part in identifying new blood donor and correlation between frequency and monetary.

My approach follows the CRISP-DM methodology where I have gathered the necessary data and cleanup/preparation was done on the data as needs. Based on graph analysis, blood donation time does have influence on prediction of repeat blood donors. We notice that num_donations distributions are not the same in the class 1 and class 0 subpopulations. Indeed, there is a peak corresponding to the people who have donated only 0-1 time will not donate blood and who have donated 2-3 will likely donate. I notice that months_since_last_donation distributions are not the same in the class 1 and class 0 subpopulations. Indeed, there is a peak corresponding to the people who have donated recently (in 1-2 months) will donate blood.

It seems that people have donated recently are more likely to donate blood and it seems that people have donated a greater number of times are more likely to donate blood.

list describes the variables in the data set

- recency: Months since the last donation.
- frequency: Total number of donations.
- quantity: Total blood donated.
- time: Months since the first donation.
- donation: True if the person donated in the last campaign, false otherwise.

The total number of instances is 576. From that, we set 60% for training, 20% for selection, and 20% for testing.

Risk/Opportunity

- Months since Last Donation: this is the number of months is since this donor's most recent donation.
- Number of Donations: this is the total number of donations that the donor has made.
- Total Volume Donated: this is the total amount of blood that the donor has donated in cubic centimeters.
- Months since First Donation: this is the number of months since the donor's first donation. Competition

Random forest 72%

Logistic Regression 77%

Conclusions

Although no individual factors were found to be statistically significant, the identification of optimal time intervals and total number of donations at which donors are more likely to return may allow for more strategic scheduling of blood drives, increasing the likelihood of a donor returning while also increasing the total number of donations for that individual. It turns out that the Random Forest, Logistic regression will work the best.