

## Capstone One: Milestone Report

### Introduction:

*Blood.* We all have it, we all need it, and the demand for it never ceases. According to the American Red Cross, approximately 36,000 units of blood are needed every day, and someone in the United States needs blood every two seconds. On top of that, donated blood must be used within 42 days, and it can be difficult to get donors in the chair with a needle in their arm (even with the promise of free cookies afterwards). Setting up blood drives is also costly, and often put on by non-profits with limited funds. So, given the costs and importance of blood supply, predicting the likelihood that individuals will return to donate blood based on past donation patterns is a helpful metric.

High-Yield Platelet Obtainment (HYPO) is a small, fictitious non-profit that operates mobile blood donation buses and is interested in utilizing its limited budget effectively while not losing the opportunity of donations due to lack of supplies or staffing. This can be broken down into the following objectives and costs:

1. HYPO would like to target its marketing outreach to those individuals most likely to return on a specific future date.
  - a. Combined labor, supply, and printing cost: \$10 per individual
2. The non-profit would also like to appropriately staff and supply the upcoming drive. HYPO would prefer to overstaff and oversupply. The cost of losing important blood donations due to inadequate supplies/staff is deemed greater than monetary loss. However, budgetary concerns are still considered.
  - a. Bus operation costs: \$200 per drive
  - b. Nurse costs: \$500 per nurse
    - i. Each nurse can handle 2 stations
    - ii. A bus can hold 4 stations.
    - iii. Each station can handle 40 donations per drive
  - c. Donation supplies cost: \$30 per donation
    - i. Supplies hold value if unused
  - d. Aftercare supplies (food and drink): \$10 per donator
    - i. Aftercare supplies cannot be saved and excess is loss
3. HYPO would also like to give as accurate as possible projected blood donations to the local blood bank.

### Dataset:

Dataset: <https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center>

To answer the problems set out by HYPO, this project will use blood donation data collected in Taiwan obtained from the UCI Machine Learning Repository. It contains information from 748 randomly selected donors and records months since last donation, total

number of donations, total volume donated, months since first donation, and whether or not the individual donated in March 2007. The last is the binary response variable.

### **Data Wrangling:**

The dataset was downloaded relatively clean and little wrangling was necessary. There were no null values to be accounted for, but the column names were renamed for clarity and ease of use:

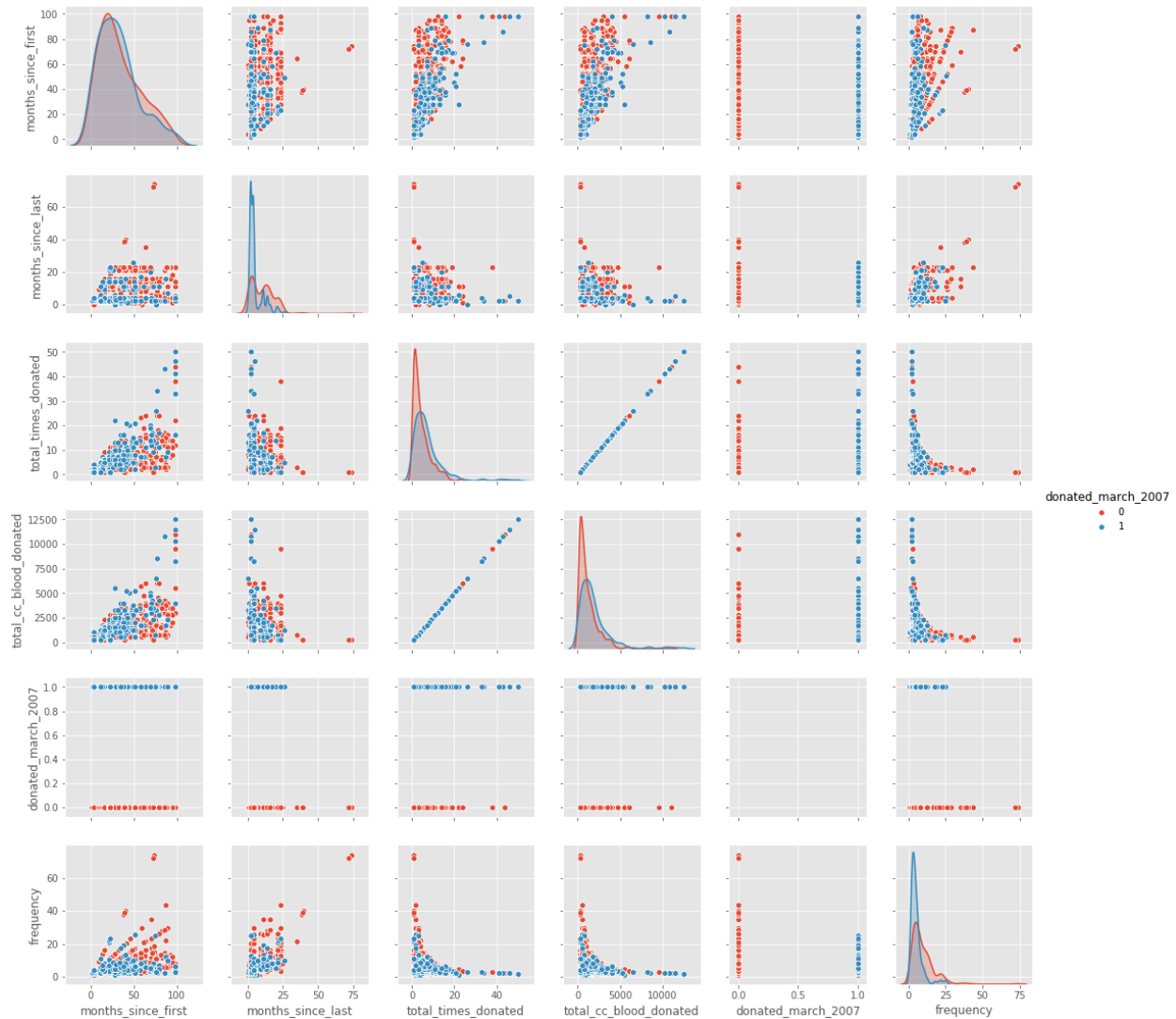
- “Recency (months)” to months\_since\_last
- “Frequency (times)” to total\_times\_donated
- “Monetary (c.c. blood)” to total\_cc\_blood\_donated
- “Time (months)” to months\_since\_first
- “whether he/she donated blood in March 2007” to donated\_march\_2007

A frequency column was also calculated ( $\text{total\_times\_donated} / \text{months\_since\_first}$ ) to provide another possible feature. Outliers were detected, but it was determined to be valid data that could be useful for a predictive model.

### **Data Insights:**

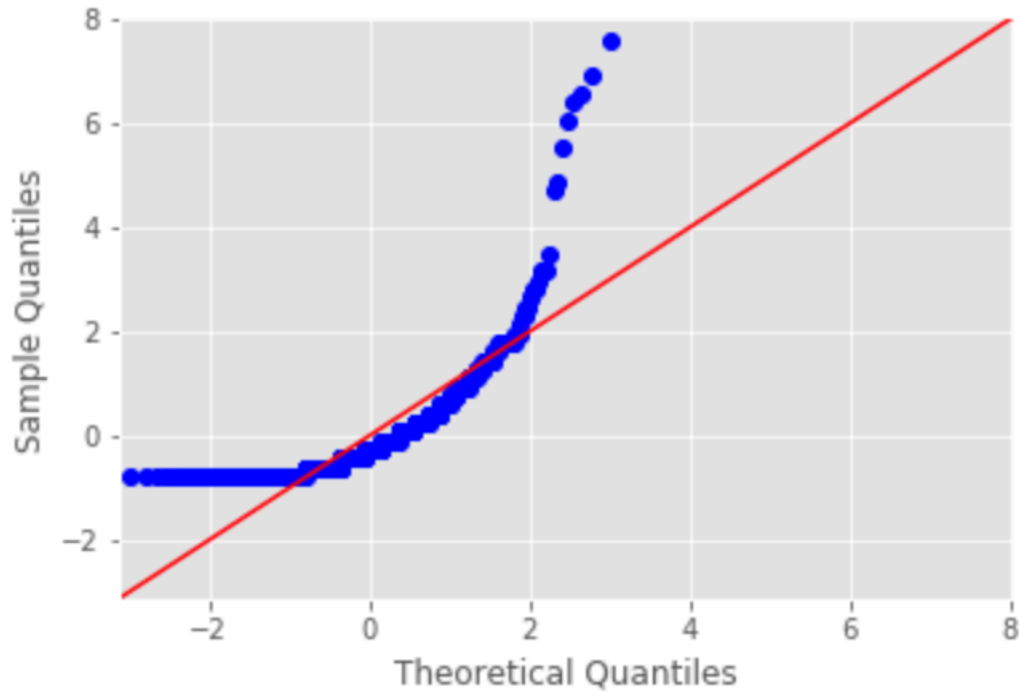
The main objective of this project is to create a model to predict the response column (donated\_march\_2007) which is a two-level categorical variable. As this is a classification problem, one of the first steps was to look at the distribution of the response variable which turned out to be about 0.238. This translates to about 23.8% of individuals recorded returning in March 2007. This is an unbalanced classifier, and steps will need to be taken during in-depth analysis to account for that imbalance. An out-of-the-box model could easily reach 76% accuracy just by classifying all as negatives (0, non-returns), but the precision and recall for the positive response variable (1, returns) would be fairly poor.

The next step was to look at the correlations between the variables of the dataset. This was accomplished by using a pair plot, shown below.



**Figure 1:** The above figure is a pair plot of all the variables of the dataset. The red hue represents non-returns, and the blue hue represents returns.

In looking at Figure 1 above, a few things stand out. Looking at the independent variables themselves initially, months since the first visit seems to have little significance on the response variable visually, but months since last, total times donated, total blood donated, and frequency all seem to have a noticeable difference between returners and non-returners. Total blood donated and total times donated are directly colinear, however, and a calculation confirmed that exactly 250ccs were donated each visit. Therefore, total blood donated will not be useful in the predictive model, but it will be useful in answering the total blood estimates portion of the problem statement. Other noticeable strong correlations include frequency and months since last. Also, frequency is a statistic calculated from months since first and total times donated. This created some odd lines in the scatterplots. It is possible frequency should not be used as a feature due to this as it adds both strong correlations between independent variables as well as extra dimensionality. Further testing during in-depth analysis will confirm. Another aspect of the independent variables to note is all have a distinct right skew indicating that the majority of the dataset contains individuals who donated blood only one to a few times (Figure 2).



**Figure 2:** The above figure is the QQ plot for total times donated. The shape indicates a right skew. The QQ plots for all independent variables were of similar shape.

The following hypotheses were tested for statistical significance using t-tests:

1. Is frequency of visits a significant factor on returning?
  - a.  $H_0$ : There is no difference in mean frequencies between returners and non-returners.
  - b.  $H_A$ : Those who returned in March 2007 have a lower mean frequency (more often).
  - c.  $\alpha < 0.05$ 
    - i. The null hypothesis was rejected. There is a statistically significant difference between means. Returners have a lower mean frequency. This aligns with the visual analysis above.
2. Is time since the last visit a significant factor on returning?
  - a.  $H_0$ : There is no difference in mean time since last visit between returners and non-returners.
  - b.  $H_A$ : There is a difference in mean time since last visit.
  - c.  $\alpha < 0.05$ 
    - i. The null hypothesis was rejected. There is a statistically significant difference between means. This aligns with the visual analysis above.
3. Is time since the first visit a significant factor on returning?
  - a.  $H_0$ : There is no difference in mean time since first visit between returners and non-returners.
  - b.  $H_A$ : There is a difference in mean time since first visit.
  - c.  $\alpha < 0.05$

- i. The null hypothesis could not be rejected. There is no significant difference between means. This is corroborated by the visual analysis above.
- 4. Is the total number of visits a factor on returning?
  - a.  $H_0$ : There is no difference in mean total number of visits between returners and non-returners.
  - b.  $H_A$ : There is a difference in mean total number of visits.
  - c.  $\alpha < 0.05$ 
    - i. The null hypothesis was rejected. There is a significant difference in means. The above visual analysis corroborates this.

The statistical tests corroborated the initial visual analysis. Total number of visits, time since last visit, and frequency are all strong predictors of the response variable. The next steps in this project are to complete an in-depth analysis and fit models to find the most effective one. As this is a classification problem, initial models tested will be : logistic regression, k-nearest neighbor, random forest, and support vector machine.