Joseph Petersen
Springboard Data Science Career Track
Oct. 2018 Cohort

Capstone One: Final 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 goals:

1. HYPO would like to target its marketing outreach to those individuals most likely to return on a specific future date.
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.
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 uses 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 (total_times_donated/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-returners), but the precision and recall for the positive response variable (1, returners) 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.

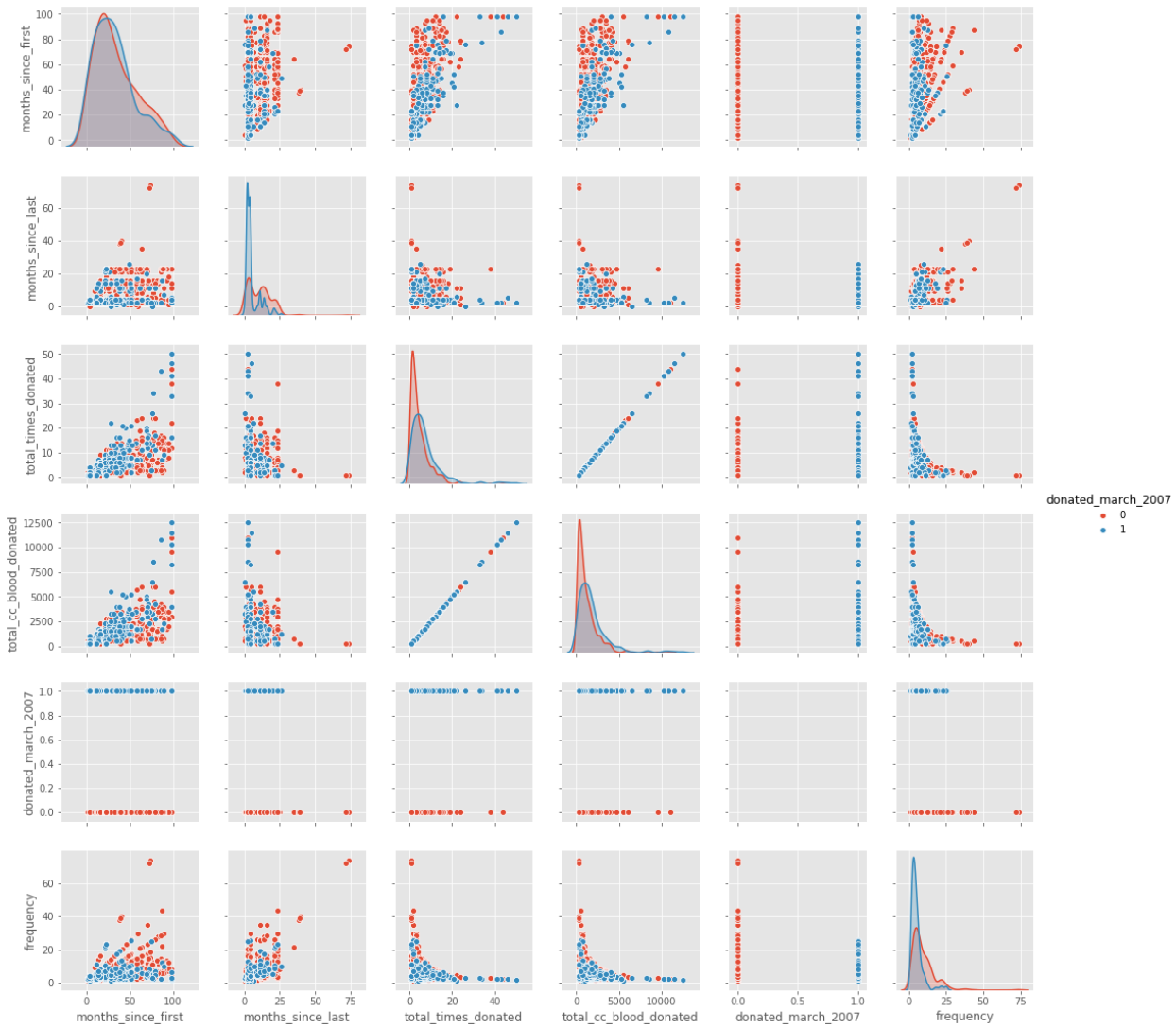


Figure 1: The above figure is a pair plot of all the variables of the dataset. The red hue represents non-returners, and the blue hue represents returners.

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. It is also possible that frequency, being a ratio of the total times donated and months since the first visit, could be used instead of those two features thereby reducing 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).

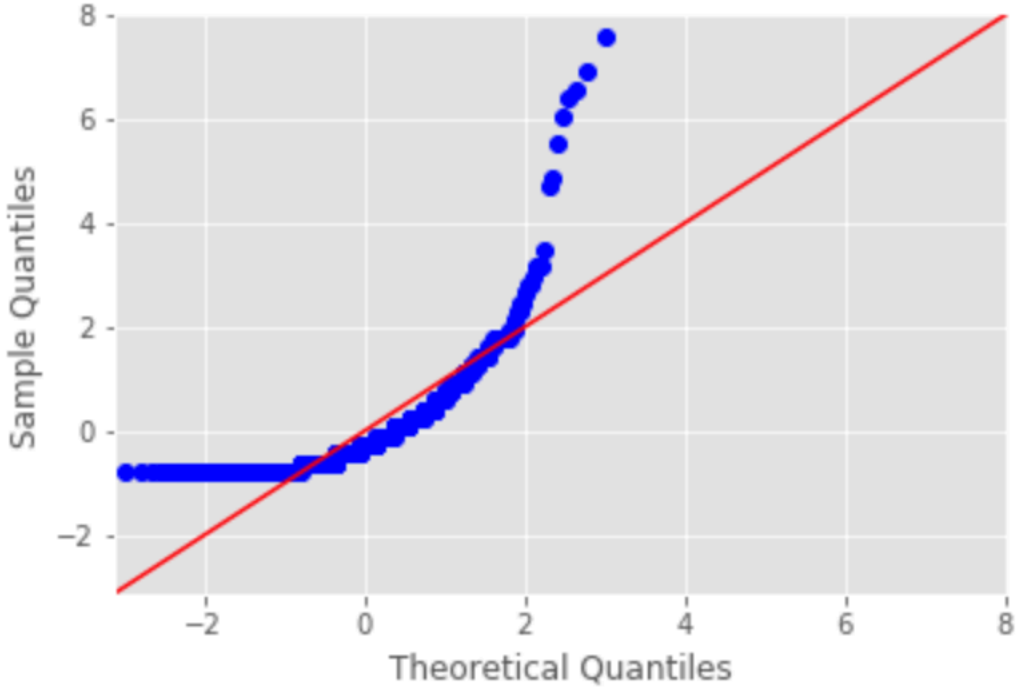


Figure 2: The above figure is the QQ plot for total times donated. The shape indicates a right skew. The QQ plots for all independent variables were of similar shape.

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: Those 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.

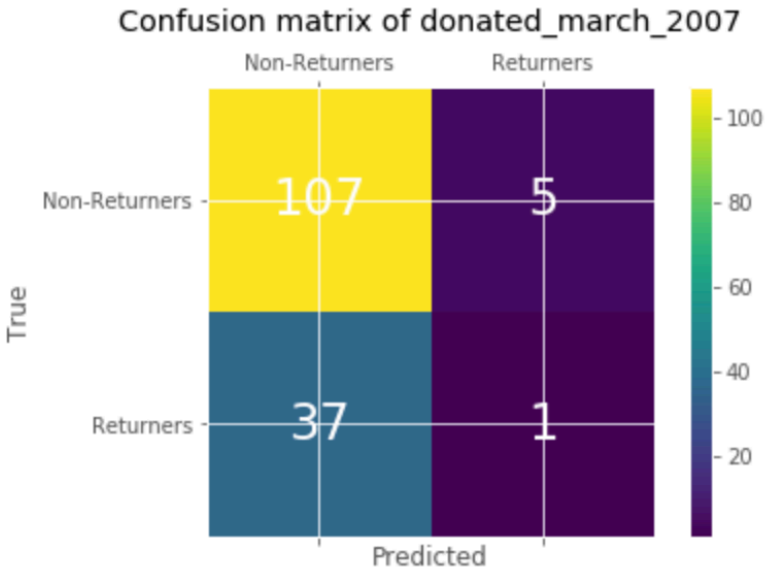
c. $\alpha < 0.05$
    i. The null hypothesis could not be rejected. There is no significant difference between means. This is corroborated by the visual analysis above.

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, machine learning 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.

Machine Learning Analysis:

Unbalanced Logistic Regression:

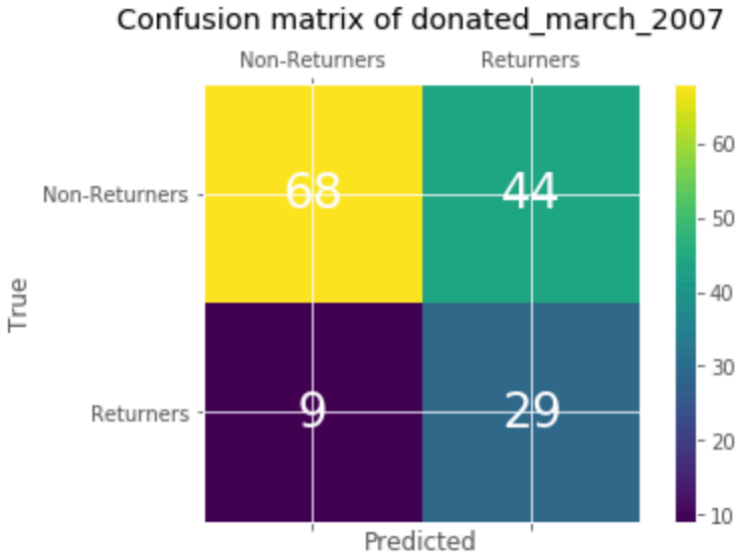


	precision	recall	f1-score	support
0	0.74	0.96	0.84	112
1	0.17	0.03	0.05	38
micro avg	0.72	0.72	0.72	150
macro avg	0.45	0.49	0.44	150
weighted avg	0.60	0.72	0.64	150

Figure 3: The above figure is a logistic regression analysis with tuned parameters but without a balanced response variable.

As figure 3 shows above, balancing the response variable is necessary. With so few 1's, the model can achieve relatively high accuracy by just predicting 0's. Both the LogisticRegression and LinearSVM functions have built-in balanced class weight parameters, but as other models were also tested it was necessary to up-sample the training data using Synthetic Minority Over-sampling Technique (SMOTE). The models tested were K-nearest neighbors, random forest, support vector machines, and logistic regression. Of those, the only ones with comparably good results were logistic regression and linear support vector machines, and both models performed better with their built-in class weight parameter set to 'balanced' than through up-sampling.

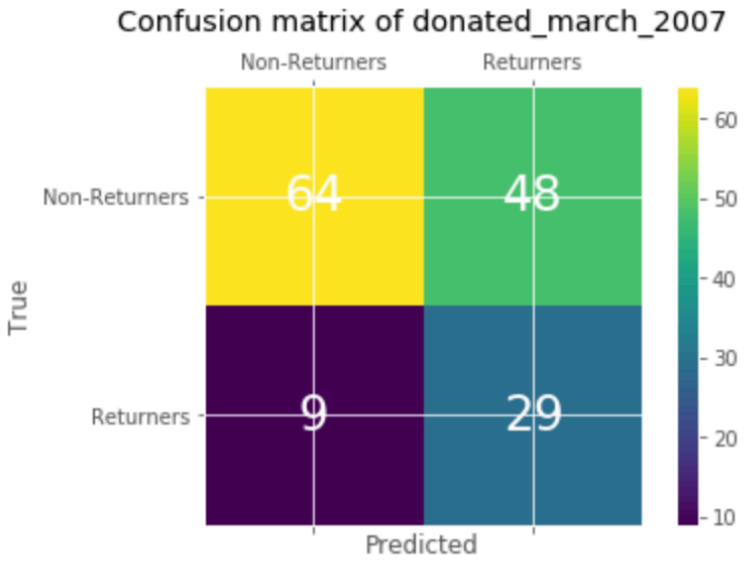
Balanced Logistic Regression:



	precision	recall	f1-score	support
0	0.88	0.61	0.72	112
1	0.40	0.76	0.52	38
micro avg	0.65	0.65	0.65	150
macro avg	0.64	0.69	0.62	150
weighted avg	0.76	0.65	0.67	150

Figure 4: The above figure shows the logistic regression with the class weight parameter set to 'balanced'.

Balanced Linear Support Vector Machines:

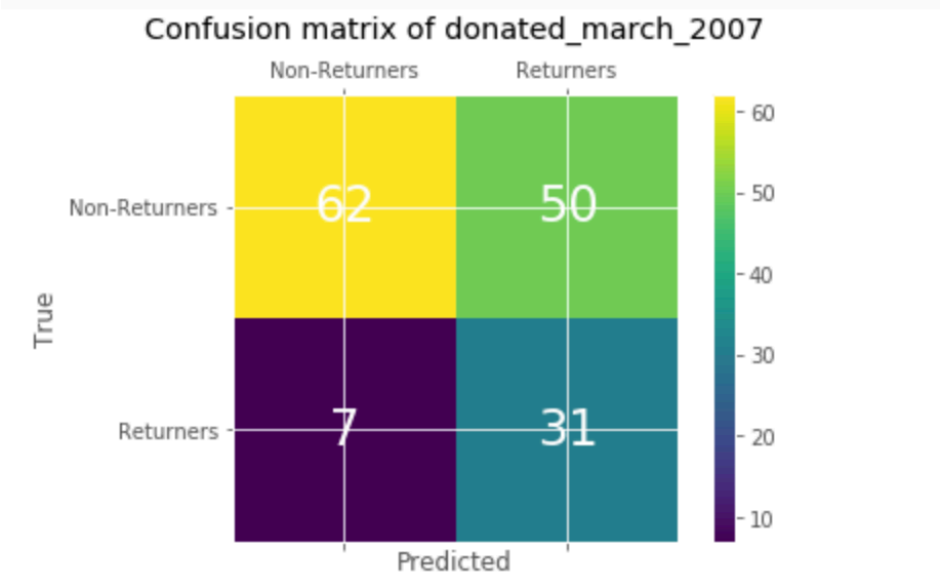


	precision	recall	f1-score	support
0	0.88	0.57	0.69	112
1	0.38	0.76	0.50	38
micro avg	0.62	0.62	0.62	150
macro avg	0.63	0.67	0.60	150
weighted avg	0.75	0.62	0.64	150

Figure 5: The above figure shows a linear support vector machines model with class weight set to 'balanced'.

The results of these two models were comparable. To further finetune these models, the dimensionality of the original dataset was reduced. The frequency column of the dataset is calculated from the total times donated and months since first features. By reducing the dataset to just include frequency and months since last as features, we retain the information of the lost columns in frequency but reduce the dimensionality. Adjusting the prediction threshold also allowed for more finetuning of the model to minimize the number of false negatives which is the priority. The best model was logistic regression with a threshold of 0.45 (predictions of 0.45 or greater are classified as 1's) (Figure 6). The threshold was very sensitive to false positives. Decreasing false negatives tended to increase false positives by a larger scale.

Balanced Logistic Regression (Reduced Dataset):



	precision	recall	f1-score	support
0	0.90	0.55	0.69	112
1	0.38	0.82	0.52	38
micro avg	0.62	0.62	0.62	150
macro avg	0.64	0.68	0.60	150
weighted avg	0.77	0.62	0.64	150

Figure 6: The above figure is the logistic regression on the reduced dataset with threshold tuned to 0.45.

Recommendations:

The logistic regression model with adjusted threshold performed the best on this dataset when prioritizing recall of the positive response variable. It will likely perform the best on data of HYPO donators. Further, if HYPO would like cleaner predictions, collecting more than this dataset provided such as anonymous personal information like sex and age could provide greater insight. As it stands, the individuals falsely classified as returners could be very good targets for educational information about the importance of blood donation and future locations of blood drives.