Joseph Petersen - October 2018 Springboard Data Science Career Track Cohort

Capstone 1 Exploratory Data Analysis Report

Introduction:

The primary aim of this project is to create a model that predicts whether or not an individual will return to donate blood based off previous donation data. The predictor variables available are months since the first visit (months_since_first), months since the last visit (months_since_last), total times a person donated (total_times_donated), the total amount of blood donated (total_cc_blood_donated), and a calculated frequency statistic (frequency). The response variable is whether or not the person donated in March 2007 (donated_march_2007), a categorical variable with two levels, those who returned (1) and those who didn't (0).

Methods and Aims:

In this initial exploratory analysis, the aim is to determine the strength of the predictor variables and examine the correlations between them. To that end, the data will be subset based on the response variable, and statistical significance between the two samples will be tested for each independent variable using t-tests. A scatter matrix of all variables will show any correlations between the independent variables.

Findings:

The amount of blood donated during each session was strictly controlled as the amount of blood donated is directly correlated with the total amount of visits. Therefore, the amount of blood donated is a product of time and highly collinear with the number of visits. The t-tests revealed that the difference in mean frequency between returners and non-returners is significant. Returners have a lower mean frequency as they donate blood with greater regularity. The tests also showed that the mean total number of visits was significantly different between returners and non-returners. This was also the case for months since the last visit. This aligns with the frequency findings as those that have a pattern of returning continue to do so and likely have been recently. The t-test for months since first, however, showed no significant difference between the two samples.

Conclusions:

The number of months since the last visit, total times donated, and the frequency of visits are all strong predictors. The number of months since the first visit is not, and the total amount of blood donated is too collinear with the total times donated.