Shortage Blood Donation Prediction

Thip Rattanavilay, DSC 630

Bellevue University

Abstract

Blood donation has been around for a long time. The first successful recorded transfusion was between two dogs in 1665, and the first medical use of human blood in a transfusion occurred in 1818. Even today, donated blood remains a critical resource during emergencies. This dataset is from a blood donation in Taiwan. The Blood Transfusion Service Center collects blood as part of a blood drive. I want to predict whether a donor will give blood the next time around. This is a prediction if the donor will give in March 2007. The goal is to predict whether people wil donated blood in March 2007 and predicting Blood Donations from given parameters such as number of donations, total volume of donation etc.

Shortage Blood Donation Prediction

# Intro/Background

Blood donation is an integral and essential part of the healthcare system. Without blood banks, many of the medical procedures that I otherwise take for granted could not take place. The modern lifestyle, ever-increasing mobility and accompanying higher accident rates, and incidences of natural and human-made disasters (such as wars, earthquakes, etc.) have led to an ever-rising demand for blood transfusions. A constant supply of blood is needed to help ensure that hospitals have access to enough blood to meet their current and future needs. One of the most important factors for a stable supply is the retention of donors, as return donors allow blood banks to not spend additional efforts and resources looking for new ones.

In this study, I focus on building a data-driven system for tracking and predicting potential blood donors. I investigate the use of various binary classification techniques to estimate the probability that a person will donate blood in March 2007 or not based on his past donation behavior. There is a time lag between the demand of blood required by patients suffering extreme blood loss and the supply of blood from blood banks. I try to improve this supply-demand lag by building a predictive model that helps identify the potential donors.

## Method

In this method, the number of clusters or groups that the data needs to be divided is determined beforehand and it assigned to the variable K. Randomly k points are chosen as the centroids of the cluster. Each data point is assigned to one of the K clusters based on the vicinity of the point to the centroid of the Kth cluster. The assignment is determined by calculating the least Euclidean distance of the data point to all the K centroids. Use information about each donor's history 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. The predictions are the probability that a donor made a donation in March 2007.

The method I have select is the Classification and Regression Techniques (CART) coined by Leo Breiman is used to refer to decision trees used for classification and regression techniques in predicting modelling.

**CART Method**

1. **C5.0 Decision Tree** - J48/C4.5 and C5.0 are the successive versions of CART which accept both continuous and discrete features along with missing data points.
2. **Artificial Neural Network** - Artificial neural networks (ANNs) are learning algorithms inspired by human brains. The main architecture of ANN is the input layer, the hidden layer and the output layer
3. **Support Vector Machines** - Support vector machines (SVMs) are supervised classification or regression techniques widely used for non-linear datasets. The kernel trick allows the user to deal with non-linear data without having to worry about its linear separability.
4. **Logistic regression** - One of the benefits of the traditional logit is it is a parametric model that allows one to interpret the effect each variable has on the response. Often public health researchers use this model to estimate odds ratios which provide a meaningful statistic for interpretation

## Cross-Validation Models

Cross-validation (CV) is one of the techniques used to test the effectiveness of machine learning models, it is also a resampling procedure used to evaluate a model if we have limited data. To perform CV we need to keep aside a sample/portion of the data on which is not used to train the model, later use this sample for testing/validating.

This project needed a good amount of data pre-processing. Firstly, the amount of blood donation feature was removed because of singularity. This was an important part of this project. This was tested using a simple linear model. A new feature "blood donation frequency" was added using three other features. Train & test dataset were scaled. Here, random forest model has been used with ntree = 200, and type = probability.

## Results

Months since Last Donation Number of Donations Total Volume Donated (c.c.) Months since First Donation Made Donation in March 2007. The result predicts if the donor will give in March 2007 and Blood Donations from given parameters such as number of donations, total volume of donation etc.

There was 748 donors were randomly selected from the donor database for the study. The features measured include R (Recency - months since last donation), F (Frequency - total number of donation), M (Monetary - total blood donated in c.c.), T (Time - months since first donation), and a binary variable representing whether the donor donated blood in March 2007 (1 stands for donating blood; 0 stands for not donating blood)

Table

Description automatically generated

Exploratory Data Analysis

Chart, bar chart

Description automatically generated

Months since First Donation

Chart, bar chart

Description automatically generated

Years Since Last Donation

Chart, bar chart

Description automatically generated

Months Since Fist Donation by Year

Chart, box and whisker chart

Description automatically generated

Months Since Last Donation

Chart

Description automatically generated with medium confidence

Numbers of Donations

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Volume of Blood Donated

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Prediction between correlation

Text

Description automatically generated

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, line chart

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated



## Discussion/Conclusion

This study has provided the ability to identify important factors that influence the decision of donating blood. I have compared the performance of various binary classification algorithms which invested previously on clustered data and non-clustered data to see if we can better predict if a person is going to donate blood or not. Among the algorithms examined, the cluster (k=4) ANN model performed the best based on the test set AUC, and C.50 based on accuracy. The results on accuracy at 83%- and first-time donor are more like to return. So, I'm quite happy with this result. Using the CART method to calculate probability. The amount of data we will collect the better the model will predict which will result in saving more lives. This is the main reason Data is so powerful and can save lives.

## Acknowledgment

Driven Data - <https://www.drivendata.org/competitions/2/warm-up-predict-blood-donations/>

 This is the smallest, least complex dataset on Driven Data. That makes it a great place to dive into the world of data science competitions. Get your blood pumping and try your hand at predicting donations.

References

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura,

Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital

Blood donor returning Rates: Analysis of 70,000 Clinical Database Blood donor Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

Blood donationss Reduction Program (HRRP). (n.d.). Retrieved September 11, 2020, from

https://www.cms.gov/Medicare/Medicare-Fee-for-Service-

Payment/AcuteInblood donorPPS/Blood donor returnings-Reduction-Program

Leonhardt, M. (2019, November 04). Rising health-care costs stall Americans' dreams of buying

homes, building families and saving for retirement. Retrieved from

https://www.cnbc.com/2019/11/04/health-care-costs-are-preventing-many-americans-

from-hitting-life-milestones.html