International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 07 | July -2020 ISSN: 2582-3930

Predict Blood Donation

Karan Aggarwal*, Shobhit Kumar

Galgotias University, Greater Noida, Uttar Pradesh

Abstract

Blood donation is very important to save other's life in the time of crisis. In the process of blood donation, the blood is directly collected from the donor and then processed and after that stored safely in blood banks. The whole process of blood donation requires an intelligent system for automation. This type of system would involve machine learning algorithms for efficient blood donors. Accurate prediction of the number of blood donors can help medical professionals know the future supply of blood and plan accordingly to entice voluntary blood donors to meet demand. To predict whether individual person is a donor or not from the data given by the person, Naive Bayes technique and K-nearest neighbors (KNN) algorithm are used. The results indicate that the accuracy value for KNN is higher than the Naive Bayes algorithm. The database can be used to track potential blood donors.

Keywords: Blood Donation, Machine Learning, Healthcare, KNN

1. Introduction

Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. According to WebMD, "about 5 million Americans need a blood transfusion every year".

Our dataset is from a mobile blood donation vehicle in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus.

The data is stored in datasets/transfusion .data and it is structured according to RFMTC marketing model (a variation of RFM).

TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming.

TPOT will automatically explore hundreds of possible pipelines to find the best one for our dataset. Note, the outcome of this search will be a scikit-learn pipeline, meaning it will include any pre-processing steps as well as the model.

We are using TPOT to help us zero in on one model that we can then explore and optimize further.

2. Literature Review

We examined the academic literature and grouped what we found into a couple different categories. First, blood banks often will survey donor volunteers to try and understand the factors that led them to donate. For example Godin, Conner et al. (2007) found that the important factors that lead to repeated blood donation among experienced donors were intention, perceived control, anticipated regret, moral norm, age, and past donation frequency. Moreover, the factors leading to repeated blood donation among new donors were only intention and age.

Others have designed studies to understand one's motives for donating blood. Sojka and Sojka (2008) surveyed over five hundred donators and found that the most commonly reported motivator among their participants was friend influence (47.2%), followed by media requests (23.5%). Lastly, they found that altruism (40.3%), social responsibility (19.7%), and friend influence (17.9%) were the primary drivers for blood donors to continue to be blood donors in the future.

As stated previously, only around 5% of eligible donor population actually donate (Katsaliaki 2008). The reasons for this are regularly reviewed by social and behavior scientists to help improve population participation (Ferguson, France et al. 2007).

Blood Donation System

The blood supply chain in general consists of three main roles: blood donors, transfusion agencies (such as hospitals), and blood centers (which attempt to coordinate the balance between supply and demand). The system can be described in three stages as shown in Fig. 1.

Stage 1: Blood Collection

Blood centers invite donors to donate blood in a number of different ways, such as recruitment campaigns, direct mailings and phone calls. The donor can be a volunteer, a replacement (a patient's relative or friend), or a paid donor. The blood collection could be for whole-blood, plasma, or platelet. The collection process is performed in government institutions, companies, and hospitals [7].

Stage 2: Blood test and processing

After collection, the blood centers apply serological and immunohematological tests on the collected blood bags. One of these tests is to determine the blood group of each bag, which can be one of the four different blood group (A, B, AB, and O) and is either Rh-positive or Rh-negative. Then, the bags are sent to processing units to extract and store blood components [7].

Stage 3: Blood distribution

Requests from hospitals to fill their blood bank needs are met by centers releasing and distributing blood components.

B. Blood Donation around the World

Blood centers all over the world are suffering a high shortage in the blood supply. Moreover, the demand for blood replacement is continually increasing in all countries. The American Red Cross states that there is someone in need of blood every two seconds.

However, the blood centers rely on volunteer donors, and most do not return to donate again.

Unfortunately, not all donors are eligible, since a lot of donated blood bags are rejected after the test stage.

Another factor causing a decline of donation is that many returning donors are aging, whereas younger donors are rare.

Although the United States has a strong coordinated effort between the government and the Red Cross, the blood supply remains below the demand. In the U.S., the daily need of blood donors is around 44000. According to the Blood Transfusion Service in Northern Ireland, the yearly need of new donors is 10000 donors [10]. A study by Gibbs and Corcoran reported that "80% of developing countries depend totally or partially on replacement donors, 15% on voluntary/non-remunerated and 25% on paid donations".

According to the Ministry of Transportation, 13221 car accidents were recorded in 2018 [4]. As a result, many blood banks in all the regions are highly needed. He stated that 60% of the need in the blood banks in the Kingdom is usually covered by the donation of relatives and friends. The voluntary donation of blood meets only 40% of the total need, which is too low, as he would like to see that closer to 100%.

As a result of the increased global demand for blood, there is a serious need to keep an adequate supply of blood that is readily available. Finding a way to recruit new donors and encourage previous donors to return is a major challenge for blood centers. In order to address this challenge, many studies have been conducted to apply different data mining tasks in the blood donation field. For instance, predicting return donors and forecasting the number of donors over a short time interval.

Return donors prediction: Most of the previous studies applied logistic regression to donor demographic information in order to predict returning donors and to understand the underlying factors affecting this prediction.

A study on the REDS donors' dataset, which had been collected by six blood bank centers around the U.S. from 2003 to 2004, was conducted along with a survey of the donors. The donors' dataset contained information of donation dates, donation status (first-time versus repeat), donation frequency, donation type, and other demographic characteristics (age, sex, and race). The class attribute (1: return donor, 0: non-return donor) was defined as 1 if the donor returned to donate during one year after completing the survey, and 0 otherwise. A binary logistic regression was used to determine the most significant predictive attributes ($p < 0.05$). The study found that predicting donor return was highly dependent on the donation frequency, the convenience of the donation place, and a good donation experience.

A study on demographic data, frequency of return, and the time interval between donations was conducted by collecting this data from the Shahrekord Blood Transfusion Center for five years. To conduct this study, researchers created a list of first-time donors for a single year (2008-2009), then the additional variables of frequency of return, the time interval between donations, and the class attribute return to donation (1: return at least once, otherwise 0: non-return) were collected for the next four years. The three response variables (return to donation, frequency of return and the time interval between donations) were analyzed using logistic regression, negative binomial regression. The results of the logistic regression showed that donor return was mainly affected by ***sex, weight, and career***

3. DATA

The dataset used in our study is one used by others researchers studying the problem posted on the UCI Machine Learning Repository 3. The source data has been taken from blood donor database of the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. 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).

Variable	Type	Description
X	Integer	Donor ID
Months since Last	Integer	This is the number of months since this donor's most
Number of	Integer	This is the total number of donations that the donor has
Total Volume	Integer	This is the total amount of blood that the donor has
Months since First	Integer	This is the number of months since the donor's first
Donated blood in	Binary	This gives whether person donated blood in March 2007

4. Methodology

- First, we used k-Means clustering to cluster the data into similar groups.
- The dataset was randomly partitioned into training set and testing set using a 70/30 train/test partition.
- Models are trained using various algorithms using the entire training set, as well as trained on each cluster generated within the training set.
- Once models are trained, the test (i.e. holdout) data is fed into each trained model to measure model performance.
- The statistical performance measures we obtained were overall accuracy, sensitivity, specificity, and area under the curve (AUC).
- AUC is generated from a receiver operating characteristic (ROC) curve.

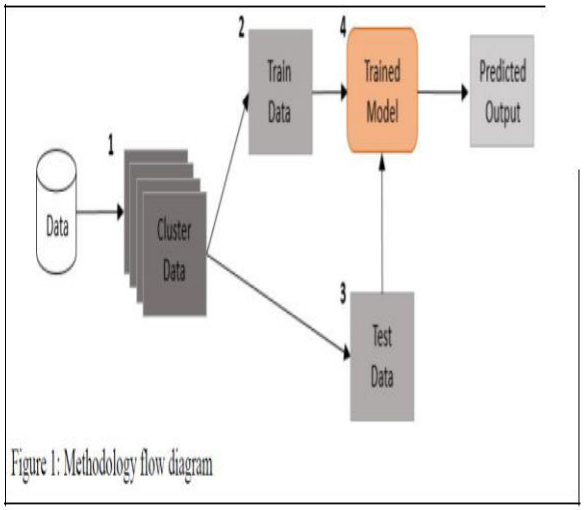


Figure 1: Methodology flow diagram

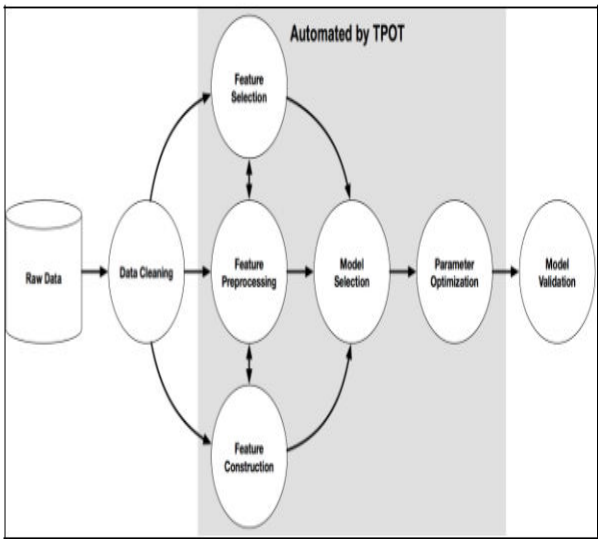
5. Model

TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming.

TPOT will automatically explore hundreds of possible pipelines to find the best one for our dataset. Note, the outcome of this search will be a scikit-learn pipeline, meaning it will include any pre-processing steps as well as the model.

We are using TPOT to help us zero in on one model that we can then explore and optimize further.

K-means clustering is an unsupervised learning method which is used when labeled data is not available. 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. After formation of the clusters, the process is repeated several times until the centroids converge, that is they stop moving. This is determined by calculating the Euclidean distance between the new centroid and the old centroid of the same cluster.



6. Conclusion

The demand for blood fluctuates throughout the year. As one prominent example, blood donations slow down during busy holiday seasons. An accurate forecast for the future supply of blood allows for an appropriate action to be taken ahead of time and therefore saving more lives.

In this notebook, we explored automatic model selection using TPOT and AUC score we got was 0.7850. This is better than simply choosing 0 all the time (the target incidence suggests that such a model would have 76% success rate). We then log normalized our training data and improved the AUC score by 0.5%. In the field of machine learning, even small improvements in accuracy can be important, depending on the purpose.

Another benefit of using logistic regression model is that it is interpretable. We can analyze how much of the variance in the response variable (target) can be explained by other variables in our dataset.

In this study we have compared the performance of various binary classification algorithms not investigated 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.

References

- Linden, J. V., et al. (1988). "An estimate of blood donor eligibility in the general population." Vox sanguinis 54(2): 96-100.Masser, B. M., et al. (2009). "Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model." Transfusion 49(2): 320-329.Quinlan, J. R. (1993). C4. 5: programs for machine learning, Morgan kaufmann. Datacamp.com

- Darwiche, M., et al. (2010). Prediction of blood transfusion donation. Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on, IEEE.

- Katsaliaki, K. (2008). "Cost-effective practices in the blood service sector." Health policy 86(2): 276-287.

- Ashoori, M., et al. (2015). "A model to predict the sequential behavior of healthy blood donors using data mining."