Prediction of blood donor volunteers based on blood transfusion service center data set

Asha Umapathi, Department of Computer Science, Hood College, Frederick, MD 21701
Email: au2@hood.edu

Abstract

Data collection and it's handling always cumbersome in blood transfusion service center. The blood transfusion service center collects verity of data set, which include month since last donated, total number of blood donation, total volume of blood collected, month since first donation and whether an individual who donated blood in March 2007. Based on the above dataset, I am utilizing the data mining techniques and predicting the outcomes of data. Data evaluation has been performed through visualization, later using classification models, I obtained the accuracy of the outcome. These data mining models will help to achieve the objective in predicting whether an individual will donate blood or not when next collection happens.

Keywords: *Data mining, Classification, Model, Logistic Regression, Decision Tree, K-Neighbors, random Forest, Resampling, Log Normalization Prediction.*

I. INTRODUCTION (*HEADING 1*)

In today's world, dealing with large data sets are always challenging, therefore, there is a pressing need for new technologies. Data mining was introduced in 1990s, and it revolutionized the data evaluation in many ways. Specifically, identifying the patterns and analyzing the discovered patterns to see how they can be effectively used in current and future. Furthermore, classification is a technique, which separates data points into various categories. The goal of using the classification models is to predict the target class for each case in the data.

Every drop of blood count when it comes to blood transfusion for a needy patient/individual. Therefore, profit, and non-profit organizations aim in collecting and storing the blood and they work under the theme 'donate the blood and save the life'. At present, everyday lifesaving blood transfusion happens in hospitals, emergency facilities, medical centers at universities or institutions. To meet the unmet need, various blood transfusion center collects and store various groups of blood and they work closely with above mentioned centers.

Blood transfusion center apart from collecting, storing the lifesaving blood, and it gathers huge data. Here, one such data set is from blood transfusion service center in Taiwan will being analyzed. The service center works closely with different universities in Hsin-Chu city, as a part of blood donation drive the service center collects blood. Service center maintains various data set such as month since last donated, total number of blood donation, total volume of blood collected, month since first donation and whether an individual who donated blood in March 2007. But the above data set cannot accurately predict whether an individual who likely to donate the blood or not next time.

Data mining technology will certainly help to predict whether a donor will likely donate blood for next time collection. Here, I used the classification models, which help blood transfusion center to predict whether he/she donated blood in March 2007. This modeling will assist in stocking up the blood to the needy person for life saving transfusion. All attributes in this dataset are dependent on each other which help to predict the target variable.

A. Data set

The dataset is from the blood transfusion service center in Hsin-Chu city in Taiwan. This data has no missing values, so no cleaning is done. The data set is multivariate which means containing two or more variables, in our case it has 5 variables, and these variables contains real values.

1. There are 748 instances(rows) and 5 attributes(columns) in the dataset.

2. Each instances include:
   - R(Recency) - month since the person last donation
   - F(Frequency) – total number of donations done

- M(Monetary) – total amount of blood donation in c.c.
- T(time) – months since the persons first donation

3. Target variable:

We must predict weather he/she donated blood in March 2007 which is a binary variable. Here 1 stands for donating blood and 0 stands for not donating blood.

B. Link to data set

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

C. Methods

By Visualizing the data set it has no missing values, so no cleaning is necessary. Then divide the data set into two parts for training and testing, and then use the classification model on these two data sets. The main purposes of using classification algorithm are to increase the accuracy of the prediction obtained from the classification models. I will use some of the classification models like

- Logistic Regression
- Decision tree
- K-Neighbors
- Random Forest

on the data set to predict the target variable will donate the blood when the next donation happens.

a. Logistic Regression: It is a supervised learning used when the target variable is categorical (1 or 0). It evaluates the probability of a dependent variable as a function of independent variable. In this project the dependent variable is the target variable that is to be predicted whereas the independent variable is the once which helps in predicting the target.

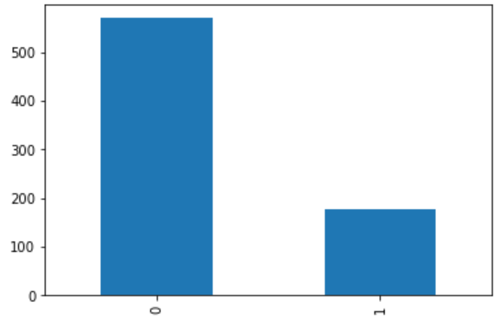
b. Decision tree: It is a supervised learning used in data mining and machine learning for classification problems. The goal is to create a model that uses a set of rules to make decisions on the target variable. A tree like structure where you create a yes or no question to continuously split the dataset until all the data of each class are separated. The first node is called root node. The end node is called leaf node and each leaf node has a class label. In our case the class label will be yes (1) or no (0).

c. K-Neighbor: it is a non-parametric, supervised learning. It uses proximity to make prediction of the target variable based on the assumption that similar points are found near one another. In k-neighbor, the value of k defines how many neighbors will be checked to determine the boundary. the value of k depends on the data. A data with more noise will perform better with larger values of k, so it’s better to choose odd values to k to prevent ties in classification.

d. Random Forest: it is supervised learning meta ensemble where the aggregation of collection of models is used to make prediction instead of one model and hence the problem of overfitting is taken care. Decisions trees are built on different samples and majority vote is taken for classification. When building each individual tree, random forest uses bagging(each individual tree is allowed to randomly sample from the given dataset with small changes to the training set resulting in different trees) and feature randomness(each tree pick from a random subset of features which forces more variation within the trees in the model resulting in lower correlation and more diversification across trees) to create an uncorrelated forest of trees where the prediction is more accurate when compared to individual tree.

D. Visulaization:

The Bar graph of target variable in the dataset

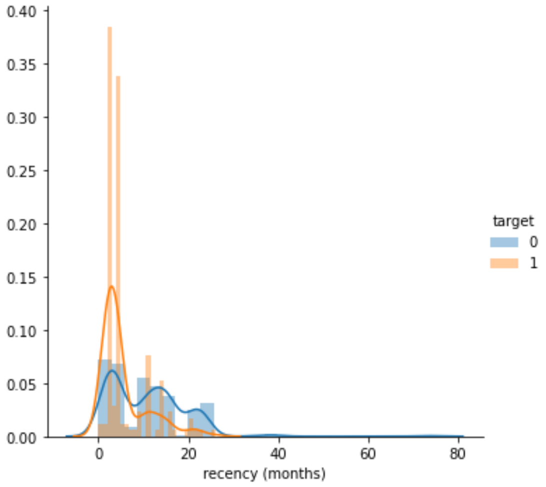


0   570
1   178
Name: target, dtype: int64
whether he/she donated blood in March 2007: [1 0]

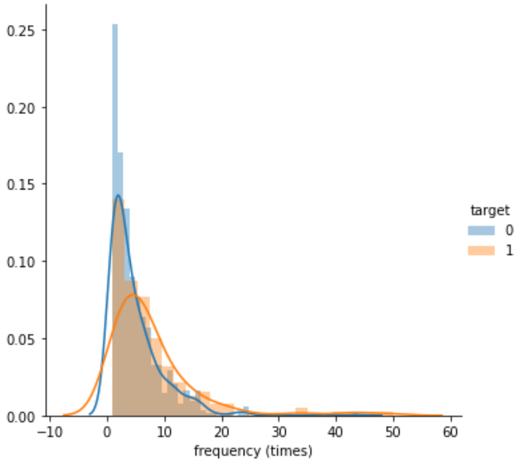
The first 5 rows of the data set

Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	Whether he/she donated blood in March 2007
2	50	12500	98	1
0	13	3250	28	1
1	16	4000	35	1
2	20	5000	45	1
1	24	6000	77	0

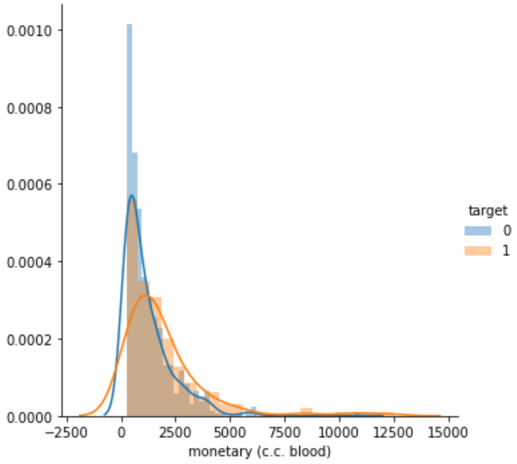
From the below graph the peak corresponding to the people who donated blood recently will likely donate



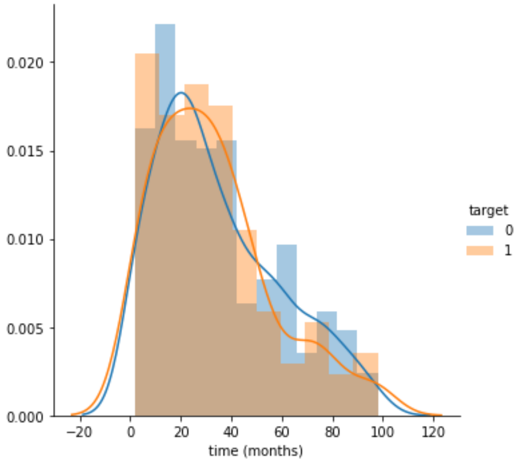
From the below graph the peak corresponding to the people who donated blood only 0-1 times will be less likely to donate than the people who donated 2-3 times more



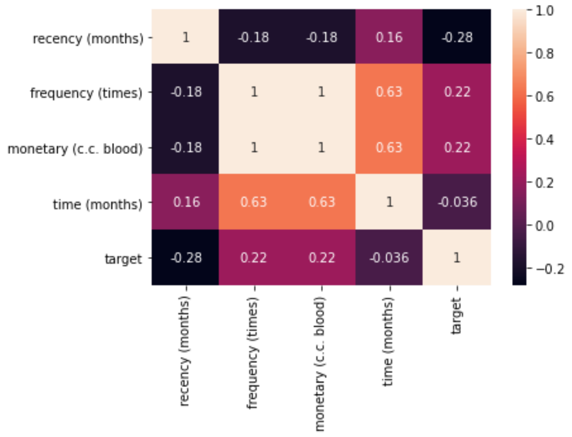
The peak corresponds to the people who donated blood more than 5 times.



The peak corresponding to the people who donated blood recently (6-20 months) will not donate.



From the heat map we can see that frequency and monetary features are correlated well, that is the donor will donate certain amount of blood for each donation. In this dataset each time the donor donates 250 c.c.



recency (months)	target	waiting period
2	1	1.920000
0	1	2.153846
1	1	2.125000
2	1	2.150000
1		3.166667
4	0	0.000000

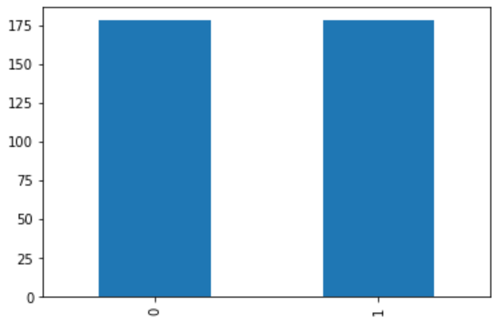
recency (months)	target	waiting period
23	0	0.000000
23	0	7.250000
23	0	0.000000
23	0	9.285714
16	0	23.333333
23	0	7.500000

The waiting period which is calculated by subtracting months since last donation from months since first donation and dividing the whole by number of donations. If the waiting period is approximately equal to months since last donation that is recency (months) then the person will donate blood otherwise no. some of the outliers are seen in the above first table.

E. *Resampling:*

When the dataset has high imbalanced classes, then the classifier will always predict the majority class without analyzing the features, even though it will have high accuracy rate which is illusionary. So, resampling is the technique used to deal with imbalanced dataset. We can resample the data in many ways. But here I taken two types. Removing random samples from the majority class to match the minority class is called under sampling. Adding more random samples to the minority class to match the majority class is called over sampling. Both types have weaknesses, under sampling can cause loss of information and over sampling can cause overfitting. For this project I am using under sampling.

```
0   178
1   178
```



F. *Checking for variance and apply normalization:*

If one of the attributes in our dataset is much greater than the other attribute, then it will affect the model ability to learn from the other attributes. Correcting this high difference helps the model's ability to learn from other attributes and is called log normalization. Log normalization applies log transformation to the values which transforms the values on to a scale that approximates normality. Normalization should be done before training the model otherwise the attribute with greater value may get more importance by the model. In this dataset we can see monetary value is much greater than other attributes.

```
recency (months)      51.22
frequency (times)     47.09
monetary (c.c. blood) 2943098.19
time (months)            629.11
```

After applying normalization

```
recency (months)      51.22
frequency (times)     47.09
```

time (months)	629.11
monetary_log	0.84

G. Results:

Accuracy of different models on the given dataset
Logistic regression accuracy: 0.76
K-Neighbor accuracy: 0.77
Decision tree accuracy: 0.66
Random forest accuracy: 0.73

Accuracy of different models after resampling
Logistic regression accuracy: 0.71
K-Neighbor accuracy: 0.62
Decision tree accuracy: 0.67
Random forest accuracy: 0.68

Accuracy of different models after resampling and normalization
Logistic regression accuracy: 0.69
K-Neighbor accuracy: 0.67
Decision tree accuracy: 0.69
Random forest accuracy: 0.71

H. Retrospective studies forecast blood donation

Studies made by WHO and other organizations shows that 118.54 million blood donations are collected worldwide, out of them, 40% are from high-income countries. The average amount of blood donation is 31.5 donations per 1000 people in high-income countries, 16.4 donations per 1000 people in upper-middle-income countries, 6.6 donations per 1000 people in lower-middle-income countries, and 5.0 donations per 1000 people in low-income countries. About 13,300 centers in 169 countries reported 106 million donations. Over 79 countries collect 90% of the blood supply from the unpaid volunteers and 54 countries collect over 50% from family members or paid donors. Globally 33% of blood donors are female. When age is considered in donors then young generation are more likely to donate blood. Blood donor’s demographic information is more important to be considered. All these features about the blood donors like where they are from (developed countries), which age group they belong, gender, and paid/unpaid volunteers will help for more accurate prediction of frequent donors or one-time donors.

I. Conclusion:

The need for the blood varies throughout the year. So, an accurate information about the future supply of the blood allows to prepare ahead of time for blood storage because the blood donations may slow down in the holiday seasons. In this project I used different data mining models to predict the accuracy of the of the donors who will donate the blood when next donations happen. From all these analyses, I can conclude that the blood donors who donate frequently in the past are more likely to donate, and who donated within 2-5 months are more likely to donate than the ones who donated in the past 6 months. So, based on this observation, we can contact the people who are interested in donating the blood, this will eventually garner more volunteers which can save many lives.

J. References:

[1]. UCI Machine Learning Repository. https://archive.ics.uci.edu/

[2]. IBM. Documentation. Algorithms-classification https://www.ibm.com/docs/SSEPGG_10.1.0/com.ibm.im.model.doc/c_classification.html

[3]. Mr. Sudhir M. Gorade, Prof. Ankit Deo, Prof. Pritesh Purohit. A Study of Some Data Mining Classification Techniques-IRJET Journal.

[4]. https://scikit-learn.org/stable/modules/tree.html

[5]. Resampling strategies for imbalanced datasets Kaggle https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook

[6]. Log normalization-Python-DataCamp

[7]. Blood safety and availability- WHO https://www.who.int/news-room/fact-sheets/detail/blood-safety-and-availability