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

Asha Umapathi, Department of Computer Science, Hood College, Frederick MD
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.

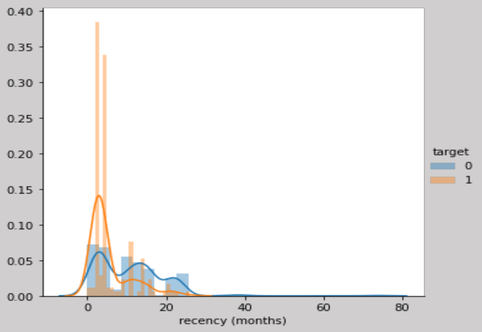
Dataset

The data is from the blood transfusion service center in Hsin-Chu city in Taiwan. 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.
The data contains 748 instances(rows) and 5 attributes(columns).
Each instances include:

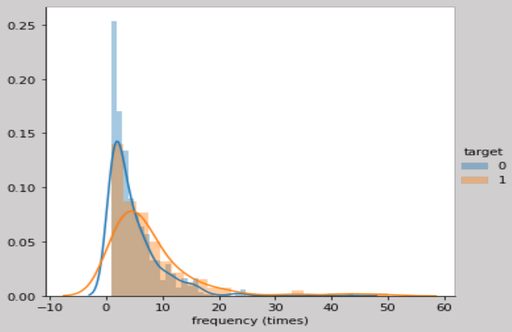
- R(Recency) - month since last donation
- F(Frequency) – total number of donations
- M(Monetary) – total blood donation in c.c.
- T(time) – months since first donation

Target variable:
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.

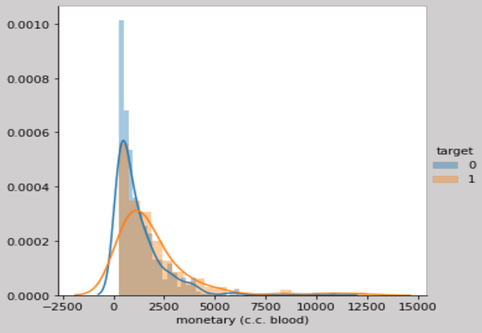
Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	Whether he/she donated blood in March 2007
2	50	12500	98	1
0	13	3250	28	1
1	16	4000	35	1
2	20	5000	45	1
1	24	6000	77	0



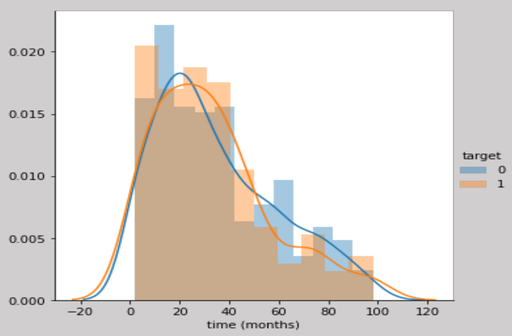
The above peak corresponds to the people who donated blood recently will likely donate the blood



The peak corresponds 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 more than 5 times

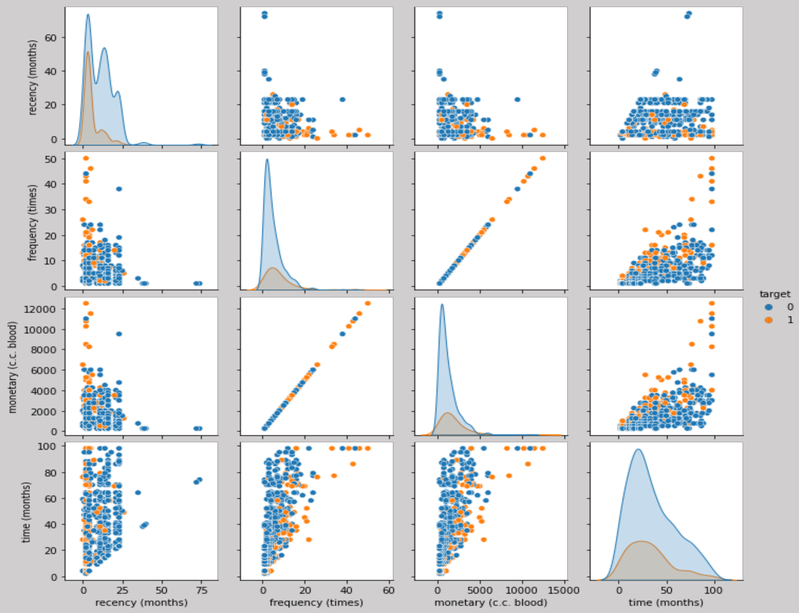


The peak corresponding to the people who donated blood recently (6-20 months) will not donate

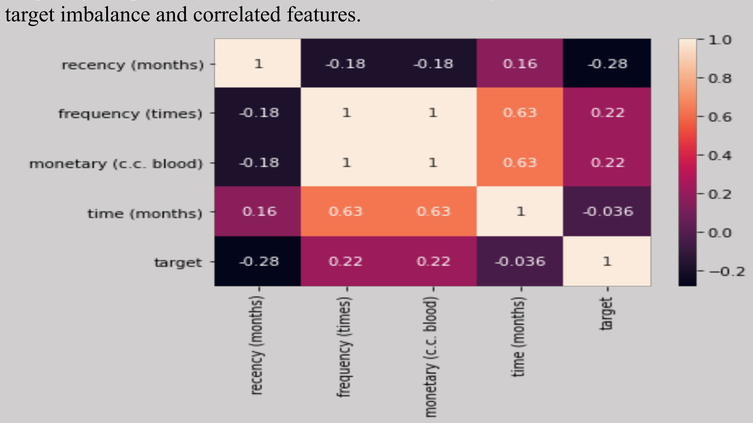
Index	recency (months)	target	waiting period
0	2	1	1.920000
1	0	1	2.153846
2	1	1	2.125000
3	2	1	2.150000
4	1	0	3.166667
5	4	0	0.000000

Index	recency (months)	target	waiting period
738	23	0	0.000000
739	23	0	7.250000
740	23	0	0.000000
741	23	0	9.285714
742	16	0	23.333333
743	23	0	7.500000

The above two tables clearly shows that if the waiting period is approximately equal to recency (months) than the person will donate the blood. We can also see some outliers in the fist table.



We don't see any features that individually could help at separating the two targets and there is no striking combinations between pair of features. But monetary and frequency are perfectly correlated. From this we can say that dataset suffers from target imbalance and correlated features.



The above 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.

Methods

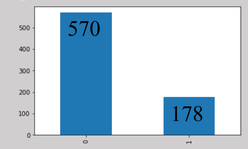
By Visualizing the data set it has no missing values, so no cleaning was necessary. Then divide the data set into two parts for training and testing, and then use the classification model on these two data sets. The main purposes of using classification algorithm are to increase the accuracy of the prediction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

- 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.
- Decision tree: It is a supervised learning used in data mining and machine learning for classification and regression 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).
- 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.
- 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.

Results

Accuracy of different models on the given dataset

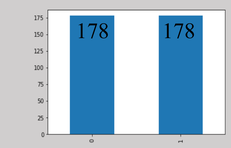
Logistic regression accuracy: 0.76
K-Neighbor accuracy: 0.77
Decision tree accuracy: 0.66
Random forest accuracy: 0.73



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 have 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.

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



Log 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

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

Summary

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 we 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 who donated in the past 6 months. So based on this we can contact the people who are interested in donating the blood resulting in more volunteers which can save many lives.

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. https://scikit-learn.org/stable/modules/tree.html
4. Resampling strategies for imbalanced datasets Kaggle, https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook
5. Log normalization-Python-DataCamp