Predicting Blood Donations in Taiwan
Using Machine Learning Algorithms

Report prepared for Professor Amir Jafari

By

Chun-Wei Lo
Yi Le

GWU Machine Learning I

May 1, 2017

EXECUTIVE SUMMARY

This project tends to find out what algorithm is best to predict blood donation by three numerical predictor variables, and if it's necessary to cluster dataset before building models. Based on the data obtained from UCI Machine Learning Repository, support vector machine (SVM), logistic regression, random forest and artificial neural network (ANN) are employed to solve supervised classification problem. After the comparison between all models, logistic regression model is implemented in both clustering dataset and non-clustering dataset. The result shows that logistic regression performs better on solving this problem and clustering before building model is not able to enhance statistical power. More detailed conclusions and corresponding suggestions are also given.

1. Introduction

The donation of blood is crucial since most often people requiring blood do not receive it on time causing loss of time. Based on the report of American Red Cross, though an estimated 38 percent of the U.S. population is eligible to donate blood at any given time, less than 10 percent of that eligible population do each year. The situation in Taiwan is worse, only 5% of the eligible donor population actually donate.

One interesting aspect about blood is that it's stochastic behavior of supply to the system as compare to the more deterministic nature of typical commodity supply, such as milk supply. In addition, according to the American Red Cross, blood has a short life of approximately 42 days, and the perishable nature of blood make it more challenging to predict accurately.

The motivation for this report is that blood demand is gradually increasing by the day because of needed transfusions due to accidents, surgeries, diseases etc. Building a data-driven system for tracking and predicting potential blood donors can decrease the risk humans are faced of losing of life due to the shortage of blood. We focus on multiple machine learning techniques and making comparison between those models for predicting blood donations accurately.

1.2 Dataset

The dataset is obtained from UCI Machine Learning Repository. The data was adopted from Blood Transfusion Service Center in Hsin-Chu, Taiwan. The center passed their blood transfusion service blood donated about every three months.

1.2.1 Variables Description

Each observation contains 5 features: R (Recency - months since last donation), F (Frequency - total number of donation), Monetary - total blood donated in c.c.), T (Time - months since first donation), and a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood).

Variable	Type	Description
Predictors		
Recency	Integer	The number of month since the first donation
Frequency	Integer	The number of donations that the donors has
Monetary	Integer	The total amount of blood that the donors has

Time	Integer	The number of month since the last donation
Response variable		
Donation in March 2007	Binary	whether he/she donated blood in March 2007

2. Methodology

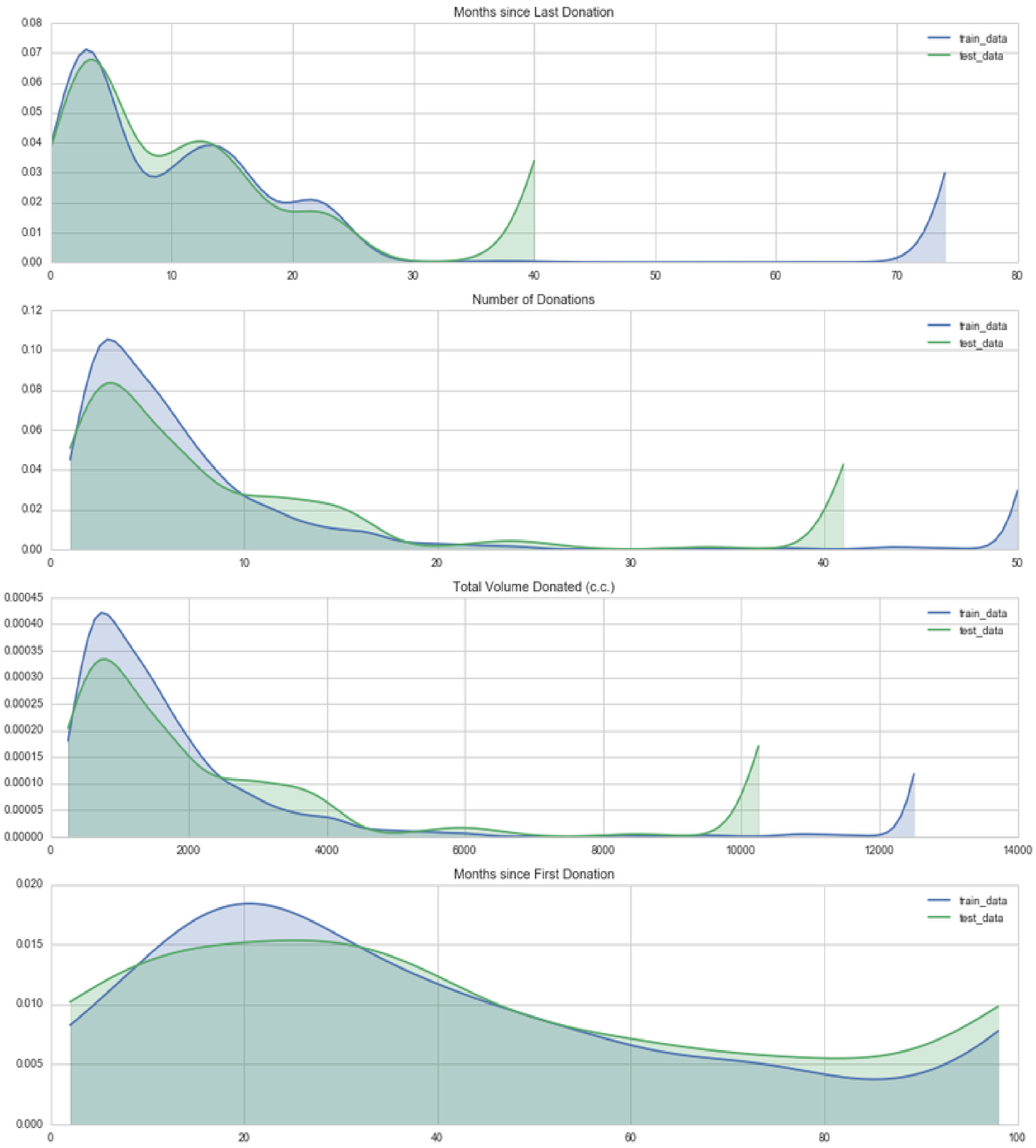
Our analysis mainly consists of three parts. The first part is exploratory data analysis. The second part is to employ multiple machine learning algorithms, including Support Vector Machine, Logistic regression, Random Forest and Artificial Neural Network and make comparisons based on accuracy rate and Area Under Curve (AUC). The third part is to find out if clustering before building models can improve the performance of model picked based on the results of second part. All the three parts are done with R and Python.

2.1 Exploratory Data Analysis (EDA)

Before any modeling process, exploratory data analysis (EDA) is necessary, as it can give us a general view of the structure and characteristics of the data and may inspire us about how to do the following modeling part. As EDA is always done through visual methods, in this section we mainly show the density plots and the correlation plots.

2.1.1 Density Plot

Density plot shows how the data is distributed. It can show the skewness and kurtosis of the distribution of each variable much more comprehensively and vividly. Here we construct a density plot for each of the four predictors. As we can see from the figures below, our sample distributions of predictors have skewness and outliers far from the rest of data. Thus, a transformation is needed for four response variables so we can get more appropriate models

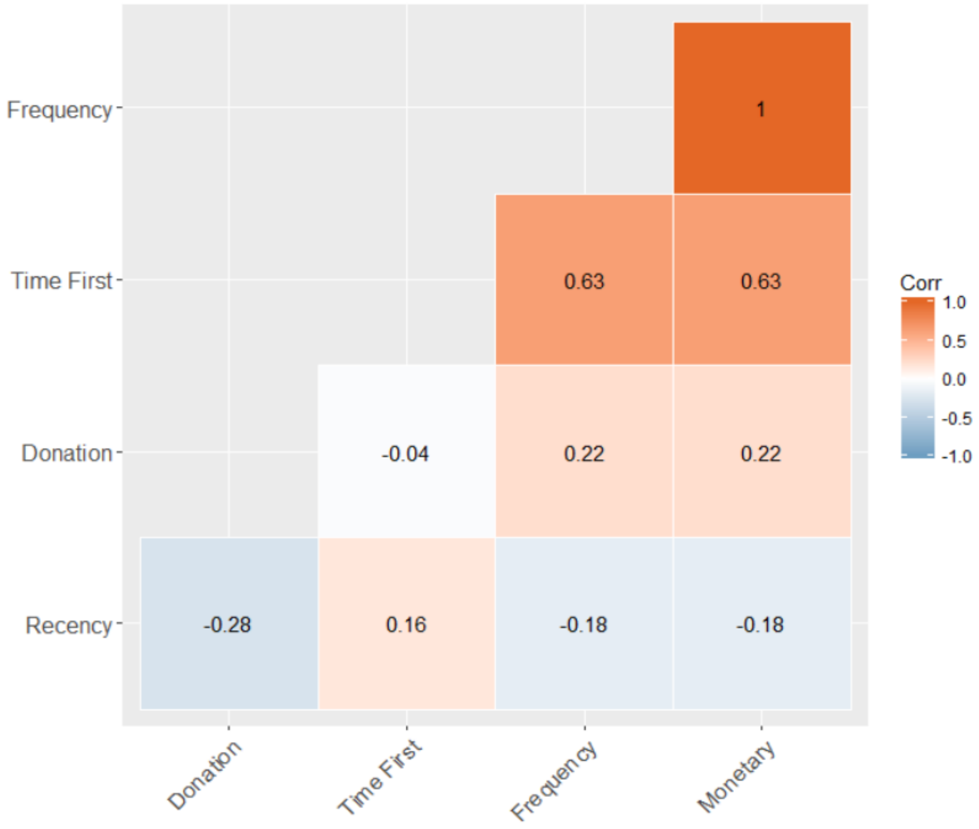


2.1.2 Correlation Plot

Correlation plot is a visualization of the correlation matrix of a dataset. A correlation matrix contains the correlation coefficient between each two of the variables in a given dataset, which indicates how well each pair of variables are related to each other. In statistics, there are different methods to calculate the correlation coefficient. One of the most commonly used is Pearson Correlation Coefficient, which is calculated as the covariance of the two variables divided by the product of the variance of each of the variable.

That is, If the coefficient is positive, we can say that the two variables are positively related. If the coefficient is negative, we can say that the two variables are negatively related. Otherwise, they are not related with each other. Based on the correlation plot, we can find two interesting in our data. First, it's noticeable that there is a perfect correlation between the Monetary (Total Volume Donated (c.c.)) column and the Frequency ("Number of Donations") column. It would be safe to say that each donation is of the same amount, which means that donors are not able to give any more or less blood.

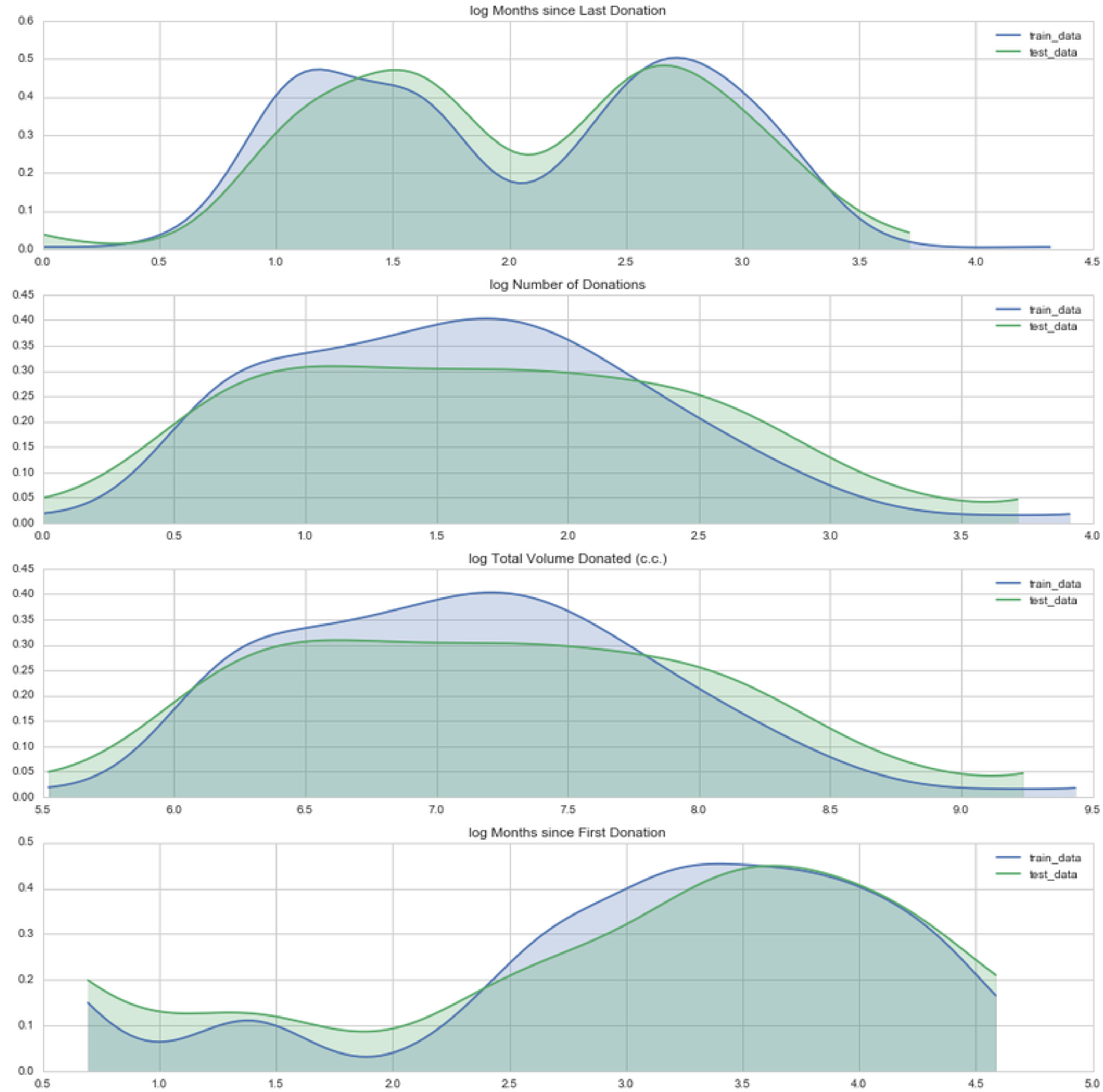
In addition, the next highest correlation is between Time First (Months since First Donation) and the "Number of Donations" which is 0.63. This means that as the number of months since the first donation increases, so does the number of donations. This makes sense as there is more time allowed to make more donations than those with less time.



2.1.3 Data Transformation

After normality check, we found that all predictor variables are not normally distributed. Thus, a transformation needs to be done that logistic regression model is appropriate.

Log transformation is adopted in this case, and the distribution after transformation can be seen below.



The log transformation can turn skewed distribution to be closer to normal distribution, and the final result is satisfying.

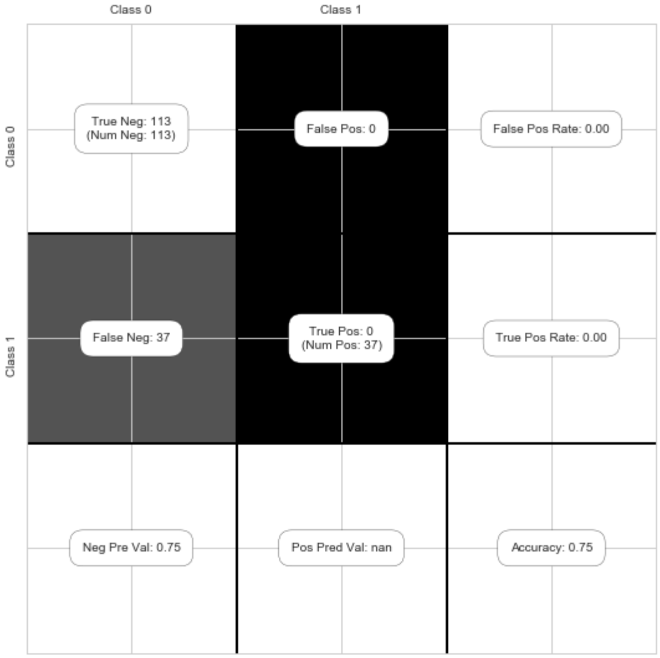
2.2 Machine Learning Algorithms

Our problem is a supervised classification problem. Therefore, we employed four machine learning algorithms and evaluate our models based on accuracy rate and AUC. In addition, we also use 10-fold cross validation in each model since traditional validation approach (train / test split) may lead to higher bias and variability and we would loss some information without cross validation.

2.2.1 Support Machine Vector

Support Vector Machine (SVM) are supervised classification or regression techniques widely used for non-linear datasets. In addition to solve linear separable problems, Kernel function allows us to deal with non-linear data. They are attempting to find a hyperplane that divides the two classes with the largest margin. The support vectors are the points which fall within this margin. However, the drawback is that a very complex hyperplane would lead to overfitting issue.

We build two SVM models since we want to compare performance between original one and one with 10-fold validation. The accuracy of original SVM is 0.753 and it increases from 0.753 to 0.764. It seems that SVM with 10-fold perform better than the one without cross validation.



	Class 0	Class 1	
Class 0	True Neg: 113 (Num Neg: 113)	False Pos: 0	False Pos Rate: 0.00
Class 1	False Neg: 37	True Pos: 0 (Num Pos: 37)	True Pos Rate: 0.00
	Neg Pre Val: 0.75	Pos Pred Val: nan	Accuracy: 0.75

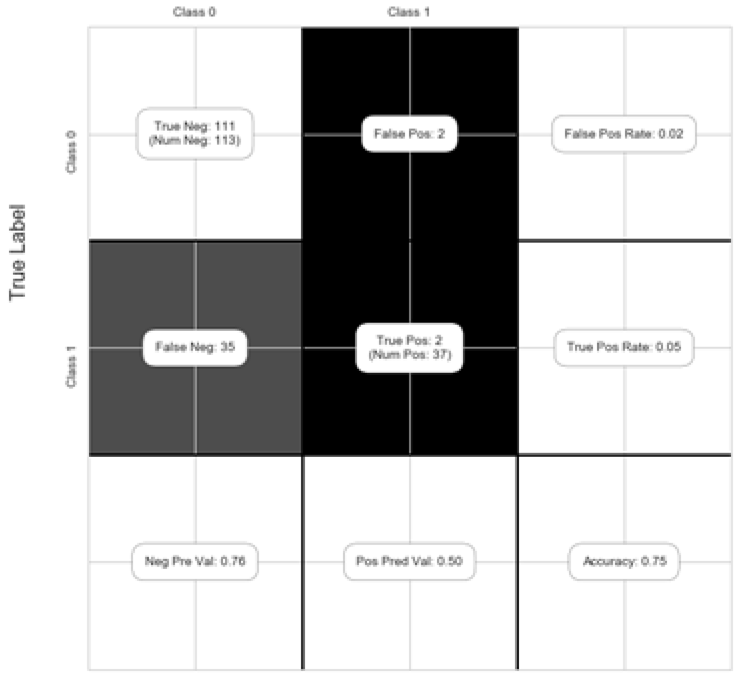
2.2.2 Logistic Regression

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. Logistic regression is a binary class classification algorithm which is used to predict a binary outcome. Given a set of independent variables, it gives the probability of a data point by fitting it to a logit function. In generalized linear model (GLM) terms, the logit is the link function that transforms the response into a logistic s-curve.

The initial exploratory analysis revealed that there was a high correlation between total volume donated and number of donations made. The first logit model estimated included all variables, and we found that the effect of total volume donated was not able to be estimated in the software we used. We then created two: one models removed total volume, the other model removed total number of donations and kept the other variables in model. We found both the variables are statistically significant when not estimated together, but are when estimated together, indicating an issue of multicollinearity. We concluded to use total number of donations, months since first donation, and months since last donation as predictors.

Model: Made Donation in March 2007 ~ f (Total Number of Donations, Months Since First Donation + Months Since Last Donation)

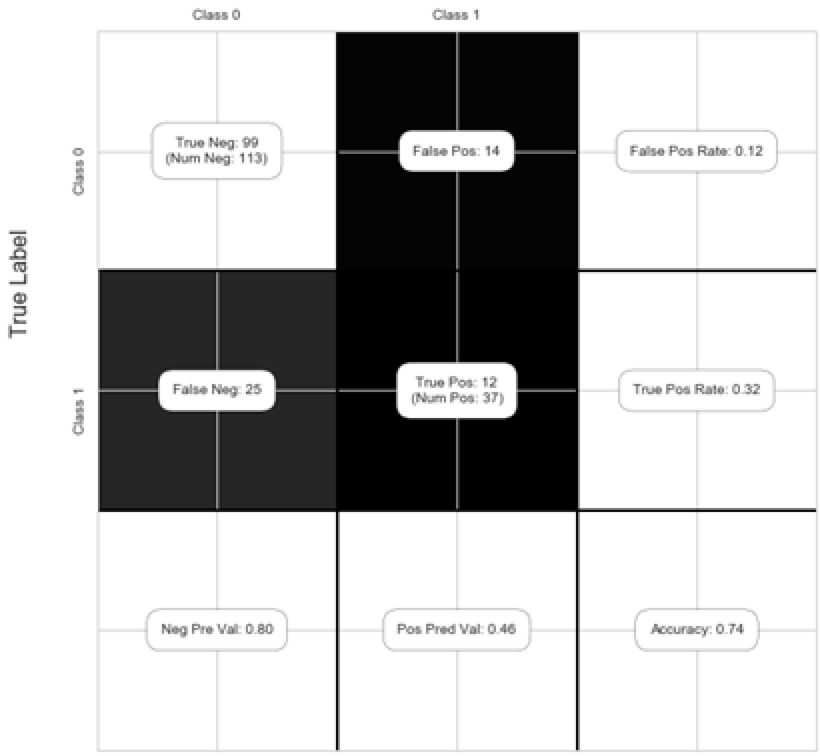
The accuracy of logistic regression is 0.753 before 10-folds, and 0.774 after 10-folds.



2.2.3 Random Forest

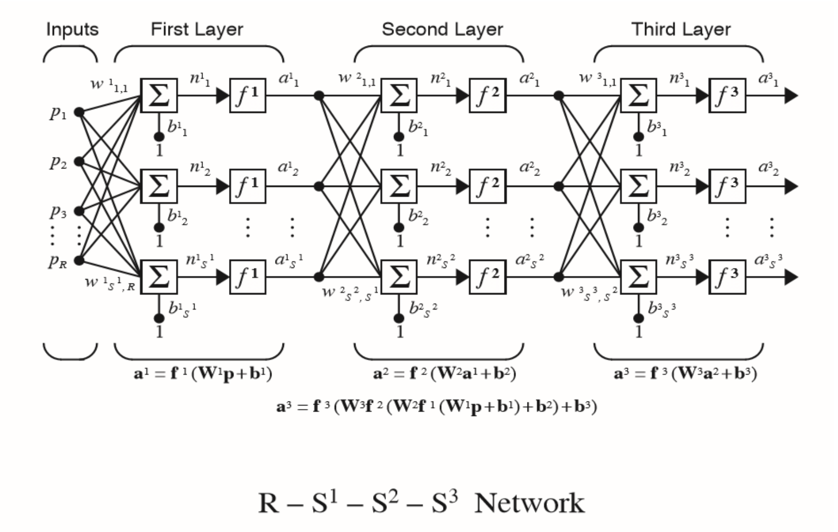
Random forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. The advantage of random forest. It runs efficiently on large databases, and the same as logistic regression, it gives estimates of what variables are important in the classification.

The accuracy of random forest is 0.74 before 10-folds, and 0.737 after 10-folds.

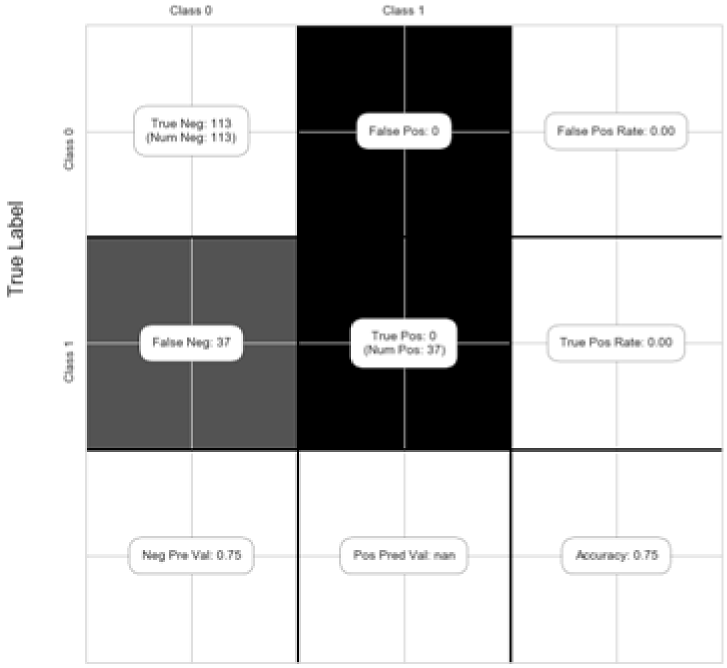


2.2.4 Artificial Neural Network

The artificial neural network, ANN for short, can be described by the figure below


Inputs
First Layer
Second Layer
Third Layer
p_1
p_2
p_3
p_R
$w^1_{1,1}$
$w^1_{S^1,R}$
$w^2_{1,1}$
$w^2_{S^2,S^1}$
$w^3_{1,1}$
$w^3_{S^3,S^2}$
$\mathbf{a}^1 = \mathbf{f}^1(\mathbf{W}^1\mathbf{p}+\mathbf{b}^1)$
$\mathbf{a}^2 = \mathbf{f}^2(\mathbf{W}^2\mathbf{a}^1+\mathbf{b}^2)$
$\mathbf{a}^3 = \mathbf{f}^3(\mathbf{W}^3\mathbf{a}^2+\mathbf{b}^3)$
$\mathbf{a}^3 = \mathbf{f}^3(\mathbf{W}^3\mathbf{f}^2(\mathbf{W}^2\mathbf{f}^1(\mathbf{W}^1\mathbf{p}+\mathbf{b}^1)+\mathbf{b}^2)+\mathbf{b}^3)$
$R – S^1 – S^2 – S^3$ Network

Different number of layers and neurons may produce different results, and both overfitting and invalid decision boundary should be avoided. After the comparison between several group different parameters, we adopted 5 neurons and 2 layers. The accuracy of ANN is 0.75 without 10 folds, and will increase to 0.7743 with 10 folds.

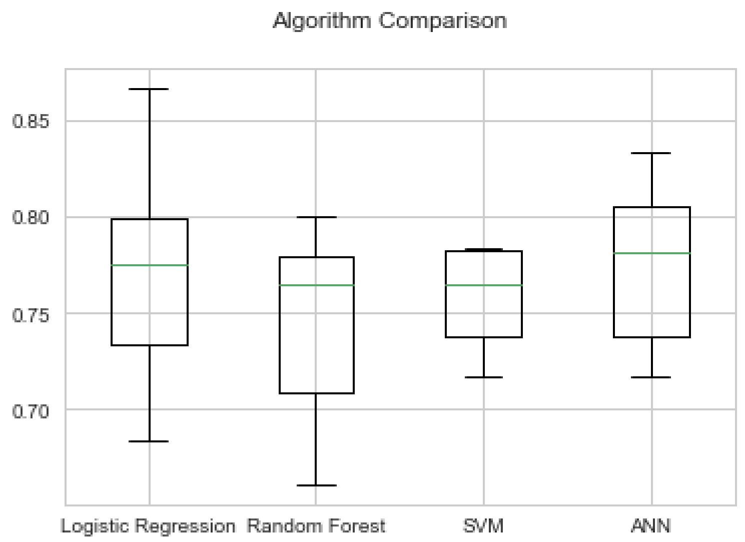


True Label	Class 0	Class 1	
Class 0	True Neg: 113 (Num Neg: 113)	False Pos: 0	False Pos Rate: 0.00
Class 1	False Neg: 37	True Pos: 0 (Num Pos: 37)	True Pos Rate: 0.00
	Neg Pre Val: 0.75	Pos Pred Val: nan	Accuracy: 0.75

It should be noticed although ANN and logistic regression can be used to solve same type of questions (classification), they are totally different. ANN focus on acquire decision boundaries to separate all points to different targets, while logistic regression is a non-linear regression model, and emphasizes on the statistical relationship between predictor variables and response variable. Besides, logistic regression can only be applied to binary response variable, while ANN can solve more problems.

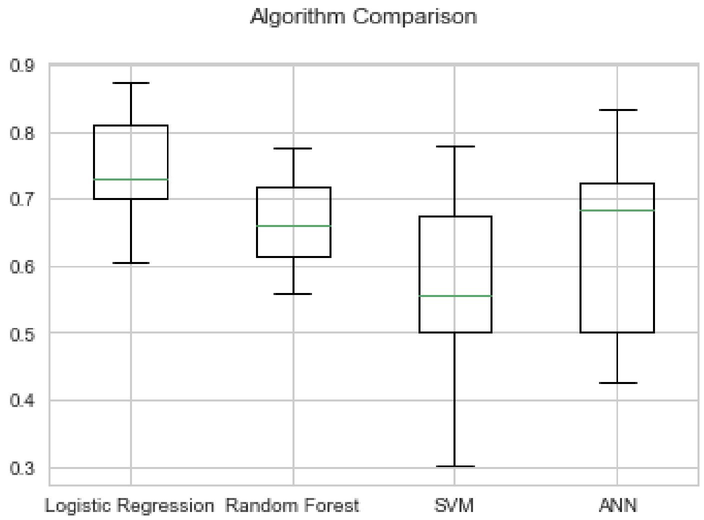
2.3 Comparison between Different Models

First, we made a comparison between the accuracy of these four models, and we found no significant difference between them.



model	Accuracy (mean)	Accuracy (s.d)
Logistic Regression	0.771045	0.052760
Random Forest	0.744040	0.049105
SVM	0.764266	0.038009
ANN	0.774322	0.037791

Then we turned to AUC and tried to figure whether we can find a best model to make prediction for the future blood donation.



Model	AUC(mean)	AUC (s.d)
Logistic Regression	0.742735	0.084677
Random Forest	0.662260	0.067164
SVM	0.577382	0.139380
ANN	0.631463	0.139380

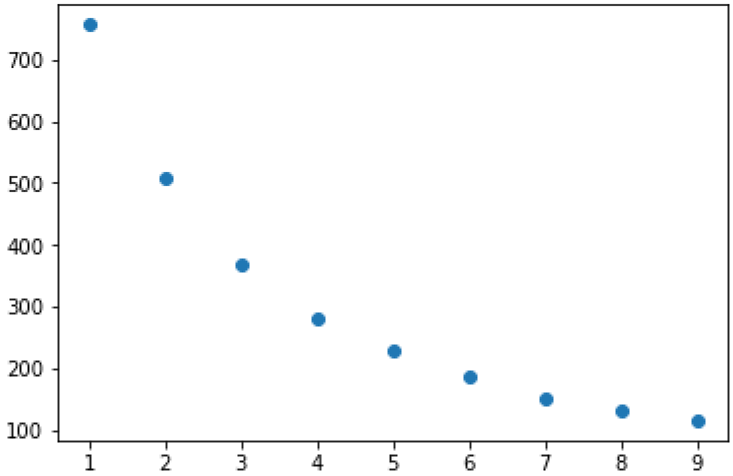
it's reasonable that we can conclude logistic regression performs better than other models.

2.4 Clustering Before Building Model

We adopted k-means to make clustering for the whole raw dataset.

2.4.1 Choose a K Value

The process of choosing the best k value is easy to understand. We split the whole dataset to train and test data, then considered k from 1 to 9, and calculate the loss array for test dataset. Finally, we can generate a figure to observe how loss changes with different k value.



The derivative will come down as k increases, and then 5 is a good choice to make loss low enough, and not too many groups for raw data.

2.4.2 Logistic Regression After Clustering

For each clustering group, 10-fold logistic regression was conducted individually and a table of accuracy was generated.

	accuracy
0	0.7273
1	0.7333
2	0.7273
3	0.8462
4	0.9375

Meanwhile we know the accuracy for non-clustering data set is 0.7710. It's obvious that clustering can't enhance the statistical power in building model.

3. Discussion

Logistic regression performed best among five algorithms. Each algorithm has its own advantage, and the problem we tried to solve is to predict a binary variable by three numeric variables. Logistic regression model is designed to solve such problem, that's why it worked better than others. We can conclude from this result that data type can almost determine which algorithm should be used.

Another important lesson is clustering can't enhance statistical power in further model building. Try doing clustering, and machine learning model will work better among similar group. This though seems reliable at first glance. However, after careful comparison between clustering and non-clustering dataset, this method proves to be useless.

Lastly, we believe the statistical performance found to date would likely be deemed as "good" among blood banks practitioners in the field. We believe assessing the models from a cost-benefit perspective where the financial cost of misclassification and reward of correct classification is incorporated into the confusion matrix and assessment statistics likely provide additional insights to the blood bank than just statistical performance measures alone.

Reference

[1] Ashoori, M., et al. (2017). "Exploring Blood Donors' Status Through Clustering: A Method to Improve the Quality of Services in Blood Transfusion Centers." Journal of Knowledge & Health 11(4): page: 73-82.
[2] Khalid, C., et al. (2013). Classification Techniques in Blood Donors Sector–A Survey. EProceeding of Software Engineering Postgraduates Workshop (SEPoW) 2013, Universiti Teknikal Malaysia Melaka.
[3] Lee, W.-C. and B.-W. Cheng (2011). "An intelligent system for improving performance of blood donation." 品質學報 18(2): 173-185.
[4] Testik, M. C., et al. (2012). "Discovering blood donor arrival patterns using data mining: A method to investigate service quality at blood centers." Journal of medical systems 36(2): 579594.

Website:
http://www.iaeng.org/publication/WCE2014/WCE2014_pp49-52.pdf
https://onlinecourses.science.psu.edu/stat501/node/347
Wikipedia C4.5 algorithm