Mohamamdhossein Jafari

mohammadhossein.jafari@studenti.unimi.it

# Supervised Learning :

## Abstract:

This research will do a binary classification task to generate a model which can determine the people who will donate blood. The Dataset is based on the information of a health center in Taiwana. The Logistic Regression and Support Vector Machines are the Statistical Learning Algorithms which are used to generate the appropriate models. The findings of the research show the logistic Regression can have better performance in producing a model which can predict whether a person will donate blood or not according to the features in this research.

## Goal Of The Analysis And Description Of The Data Set:

### Data Explanation and the purpose of the analysis:

The Blood Transfusion Service Center Dataset has been used in this paper. Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwana and includes 748 people. It is a binary classification task. The attributes are:

- Recency - months since last donation
- Frequency - total number of donations
- Monetary - total blood donated in c.c.
- Time - months since first donation),
- The target attribute is a binary variable representing whether he/she donated blood in March 2007 (2 stands for donating blood; 1 stands for not donating blood).

The binary variable changes to be 1, if the person donated, and 0 if the person did not donate. The Histogram and the density figures of the original datasets.

## Findings of the Research:

- Monetary, total blood donated in c.c, is a variable which increasingly has collinearity with other variables. If this variable is removed from model, it is possible to have more accurate models.
- "Months since last donation" and "Months since first donation" has a negative relationship with the blood donation while "total number of donations" has a positive relationship with blood donation according to Logistic Regression
- According to different kinds of kernel methods, which was tested in this research, of Support Vector Machnes, the polynomial of degree 3 can produce better results in this Dataset
- By running accurately, the Logistic regression, it is possible to decrease the misclassification error to less than 16 percent.
- The Variance Inflation Factor of the Recency, Time, and Frequency are small.

# The Analysis of Models and Explanations:

In this research we try to answer to the question of whether it is possible to generate a model which can accurately predict whether a person will donate blood by having the above-mentioned features such as total number of donations, total blood donated and so on. In this research the performance of two broad family of Machine learning techniques which are Support Vector Machines and Linear Regressions are analyzed. It is clear that the features are extremely different in their largeness and both family of models force us to do feature scaling.

We split the dataset to train and test set. In the first level we allocate 70 percent of the data to the training set and 30 percent to test set.

## Logistic Regression:

logistic regression is run on all independent variables of training set. The Recency, Frequency, Time has small p-values, and they are acceptable. But the Monetary has a large collinearity with the other variables and it must be omitted from the model. The coefficient for Recency is -1.3559, and for Time is -0.4881 while total number of donations has a positive relationship, 0.6847, with the possibility of donating blood.

In the next step, the "total blood donated in c.c." is omitted from the variables and Logistic Regression model is run again. All the estimated coefficients are acceptable form the statistical viewpoint. The only variable which has a positive relationship with donation is frequency, with 0.69, while time and recency has negative relationships with -0.4881, and -1.3559 respectively. Null deviance is 585.46 and the Residual deviance is 511.71 and the AIC is 519.71.

Pseudo-R2 measures

It can be a good approach to have a look at different Pseudo-R2 measures. The McFadden's pseudo r-squared is 12.5 percent while the Cragg and Uhler's pseudo r-squared is about 50 percent larger and is near to 20 percent. The Maximum likelihood pseudo r-squared is about 13 percent.

Variance Inflation Factor:

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model. Fortunately, the Variance Inflation Factors in the generated model is almost low. The vif for Time is 2.217922 while for recency is 1.058702 and for frequency is 2.185240 .

Variable importance:

The varImp method for calculating variable importance. Recency has the largest Variable importance by 4.848789. Then is Frequency by 4.200220 and time by 2.982727.

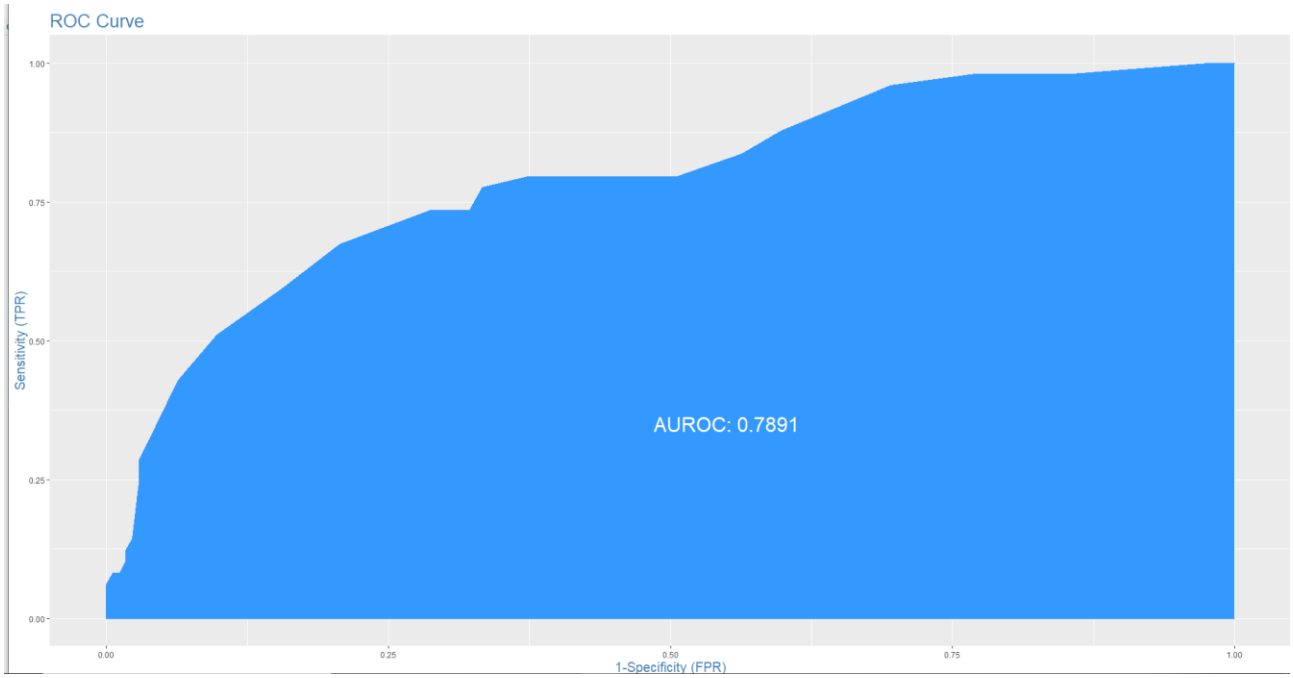
When using Logistic Regression, it is vital to choose the best cut-off value. By using the InformationValue, it is possible to choose the optimalCutoff, and this amount will be 0.4145246 for the generated Logistic Regression Model.

The performance of Logistic Regression:

Confusion Matrix:

For checking the performance of the Logistic Regression, it is critical to check the confusion matrix on Test Set. The model can correctly determine 171 people who will not donate blood. It also can accurately predict 6 people who will donate. The model predict 43 people will donate blood while it is wrong and forecast 3 people will not donate while it is also wrong.

The sensitivity of the forecast is 0.122449 and the specificity is 0.9827586. The miscalculation error is about 15 percent.




An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. The figure of AUROC is 79 percent which is almost acceptable for a classification task.

Support Vector Machines:

Support Vector Machines are among the families which widely use for binary classification.

Choosing an appropriate kind of kernel is a vital topic in Support Vector Machines. When we run the model for all the variables, including the including “total blood donated in c.c.”. The result is that the number of True Positive is 174 while the number of False Positive is 49 which is quiet high.

When the variable which has collinearity with others is omitted, the model has the same accuracy with the previous model which is again not really accurate, because the number of error number 2 is high.

It is critical to choose an appropriate method of kernel. If the method of kernel change from linear to polynomial it can improve the results. Through the default degree, which is 3, the number of error type 2 decrease from 49, in linear kernel, to 46 which is step forward action while the number of True positive remains without any changes, and the number of true negatives becomes 3.

Changing the degree of polynomial SVM to 5 and then will not change a lot the accuracy of the model.

The next method of kernel after polynomials is radial basis. In this method, the error type 1 is 5 and the error type 2 is 46 while the true positive and true negatives are 169 and 46 respectively.

The final kernel method is sigmoid. In this method the number false negative increases decrease to 31 which is a good news, but the error type one increases a lot to 1.

## Conclusions:

This research tried its best to produce a model which can predict whether a person will donate blood. These are the findings of the research:

- It is vital to do feature scaling to have better results
- The total blood donated in c.c was the attribute which has the large collinearity with other factors, and it was deleted.
- optimal Cut off figure is about 41 percent.
- It is vital to allocate more data to training set than test set, because the Statistical learning algorithm needs more data to produce an accurate model
- Cragg and Uhler's pseudo r-squared has the largest amount of pseudo r-square among the tested ones
- The VIF figure for all the features were below 2.3
- The recency is the most important feature among all the others by having varImp of 4.8
- The sensitivity of the Logistic Regression Model is about is about 12 percent
- The specificity of the Logistic Regression Model is about 98 percent
- The misClassError the Logistic Regression Model is about 16 percent
- The AUROC of the Logistic Regression Model is about 79 percent
- The SVM family can have different accuracy in binary classification tasks
- Among all the SVM algorithms, the model generated by kernel polynomial of degree 3 has the best accuracy on this dataset.
- It is vital to work other different Statistical Learning Methods and check the performance of those models in this dataset and compare them with the produced results.