Blood Donation Shortage Prediction

Thip Rattanavilay, Master of Data Science

Bellevue University

Summer 2021

Table of Contents

Abstract

The inspiration for this project is that blood request is steadily growing day by day due to the need of transfusions of blood because of surgeries, accidents, diseases etc. Accurate forecast of the number of blood donors can help medical professionals know the future supply of blood and plan consequently to attract volunteer of blood donors to fulfill the demand. I found that the Logistic Regression, Random Forest will work the best and by using these models' prediction led to the best test set performance accuracy of (77%), which is better than other data prediction.

Blood Donation Shortage Prediction

Background

Blood donation is an integral and essential part of the healthcare system. Without blood banks, many of the medical procedures that I otherwise take for granted could not take place. The modern lifestyle, ever-increasing mobility and accompanying higher accident rates, and incidences of natural and human-made disasters (such as wars, earthquakes, etc.) have led to an ever-rising demand for blood transfusions. A constant supply of blood is needed to help ensure that hospitals have access to enough blood to meet their current and future needs. One of the most important factors for a stable supply is the retention of donors, as return donors allow blood banks to not spend additional efforts and resources looking for new ones.

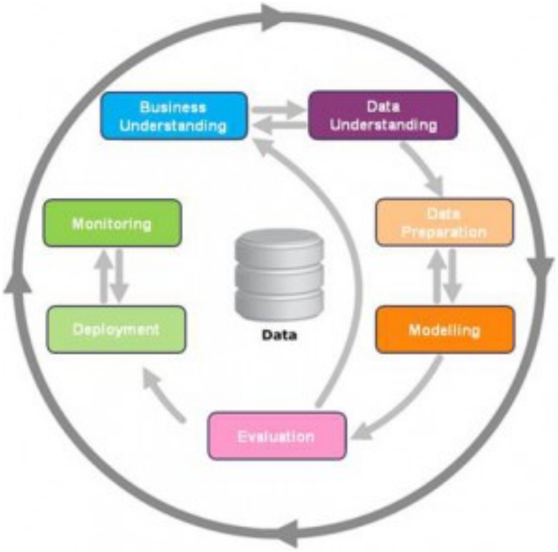
Amazingly, only around 5% of the eligible donor population actually donate. This low percentage highlights the risk humans are faced today as blood and blood products are forecasted to increase year-on-year. This is likely why so many researchers continue to try to understand the social and behavioral drivers for why people donate to begin with. The primary way to satisfy demand is to have regularly occurring donations from healthy volunteers.

Problem Statement

In my finding, I focus on building a data-driven system for tracking and predicting potential blood donors. I investigate the use of various binary classification techniques to estimate the probability that a person will donate blood in March 2007 or not based on his/her past donation behavior. There is a time lag between the demand of blood required by patients suffering extreme blood loss and the supply of blood from blood banks. I try to improve this supply-demand lag by building a predictive model that helps identify the potential donors.

Data Understanding

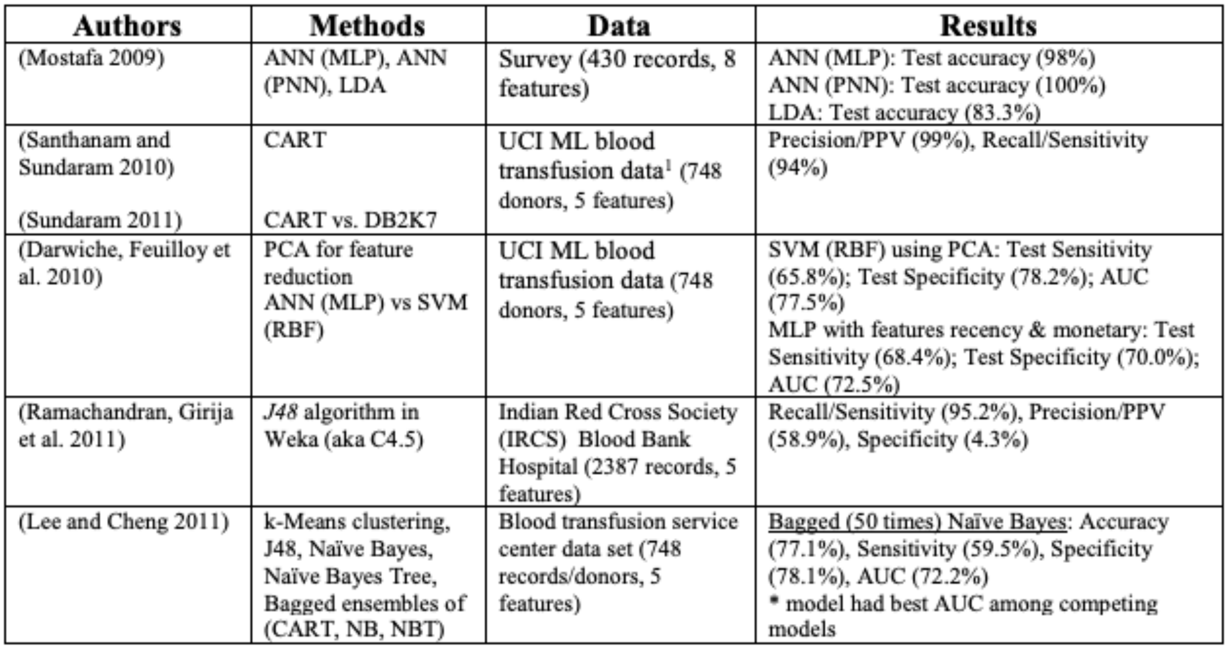
Based on my understanding of the problem, I follow a structured analytical process widely known in the data mining, called the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman, Clinton et al. 2000). The idea behind this analysis framework is to develop and validate a model (or solution) that satisfies the requirements of problem and needs of stakeholders. I used guidance in the academic literature to get ideas of how others have modeled this problem and followed a similar process. Some authors clustered data before building their predictive models and some did not. I tried both and used some algorithms that others have not yet investigated to see if my solution was as good or better than what others have found. I structured this paper as follows. I performed a review on the background, and the data on blood donation to see what methodologies have found to be successful at understanding this problem. I discuss the data set used in my study. Next, I discuss the methodology/design I implemented and discuss the models I investigated. Lastly, I present my results, discuss my conclusions, and how I plan to extend this research.



Data Preparation

As stated previously, only around 5% of eligible donor population actually donate. The reasons for this are regularly reviewed by social and behavior scientists to help improve population participation (Ferguson, France et al. 2007). The focus of this data is to understand the performance that using traditional machine learning techniques can have at predicting future blood donation. Table 1 outlines what I believe is an exhaustive list of all published studies in this domain, the data set used, methods employed, and results achieved. (https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center)

Table 1:

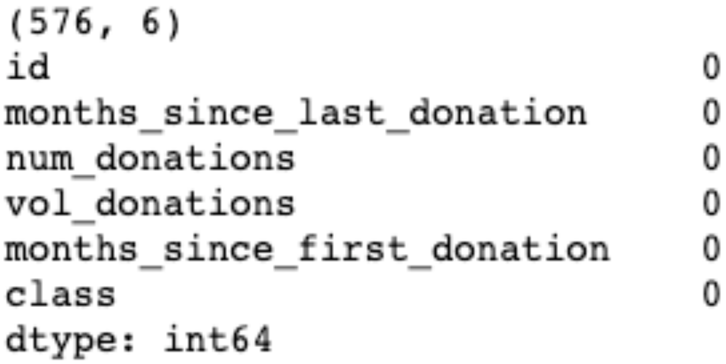


Authors	Methods	Data	Results
(Mostafa 2009)	ANN (MLP), ANN (PNN), LDA	Survey (430 records, 8 features)	ANN (MLP): Test accuracy (98%) ANN (PNN): Test accuracy (100%) LDA: Test accuracy (83.3%)
(Santhanam and Sundaram 2010) (Sundaram 2011)	CART CART vs. DB2K7	UCI ML blood transfusion data[1] (748 donors, 5 features)	Precision/PPV (99%), Recall/Sensitivity (94%)
(Darwiche, Feuilloy et al. 2010)	PCA for feature reduction ANN (MLP) vs SVM (RBF)	UCI ML blood transfusion data (748 donors, 5 features)	SVM (RBF) using PCA: Test Sensitivity (65.8%); Test Specificity (78.2%); AUC (77.5%) MLP with features recency & monetary: Test Sensitivity (68.4%); Test Specificity (70.0%); AUC (72.5%)
(Ramachandran, Girija et al. 2011)	*J48* algorithm in Weka (aka C4.5)	Indian Red Cross Society (IRCS) Blood Bank Hospital (2387 records, 5 features)	Recall/Sensitivity (95.2%), Precision/PPV (58.9%), Specificity (4.3%)
(Lee and Cheng 2011)	k-Means clustering, J48, Naïve Bayes, Naïve Bayes Tree, Bagged ensembles of (CART, NB, NBT)	Blood transfusion service center data set (748 records/donors, 5 features)	Bagged (50 times) Naïve Bayes: Accuracy (77.1%), Sensitivity (59.5%), Specificity (78.1%), AUC (72.2%) * model had best AUC among competing models

The source datasets have been taken from blood donor database of the Blood Transfusion Service Center. 748 donors were randomly selected from the donor database for the study. There were no missing values and it have 576 rows and 6 Columns Table 2. The features measured include R (Recency - months since last donation), F (Frequency - total number of donation), M

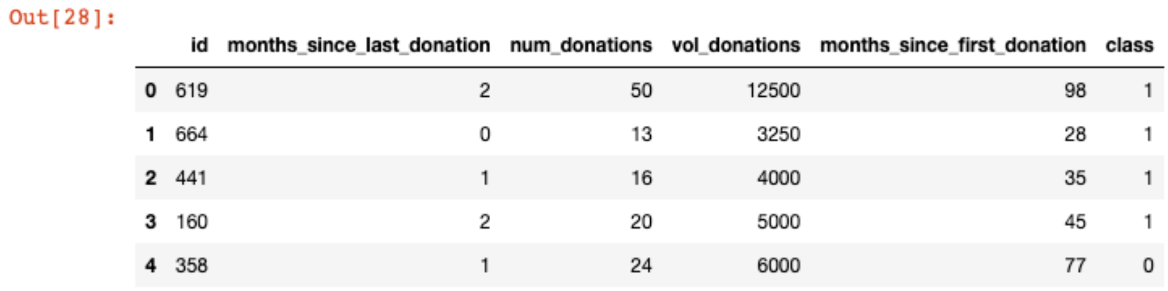
(Monetary - total blood donated in c.c.), T (Time - months since first donation), and a binary variable representing whether the donor donated blood in March 2007 (1 stands for donating blood: 0 stands for not donating blood) as shown in Table 3.

Table 2:



```
(576, 6)
id                              0
months_since_last_donation      0
num_donations                   0
vol_donations                   0
months_since_first_donation     0
class                           0
dtype: int64
```

Table 3:

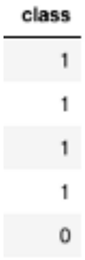


Out[28]:

	id	months_since_last_donation	num_donations	vol_donations	months_since_first_donation	class
0	619	2	50	12500	98	1
1	664	0	13	3250	28	1
2	441	1	16	4000	35	1
3	160	2	20	5000	45	1
4	358	1	24	6000	77	0

In the class column there are two classes Table 4

Table 4:



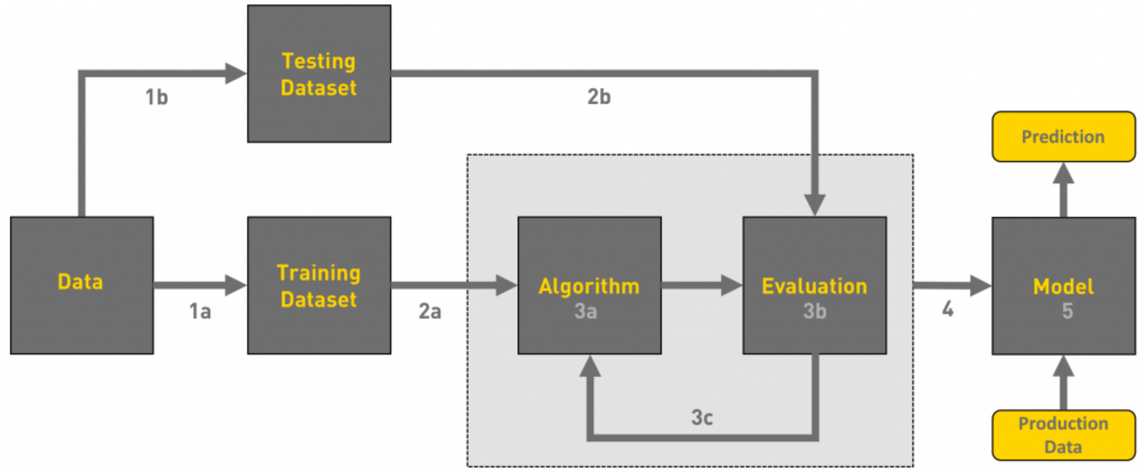
class
1
1
1
1
0

- class 1 : The donor donated blood in March 2007.
- class 0 : The donor did not donate blood in March 2007.

Note: Assuming that 1 means donated and 0 means not donated

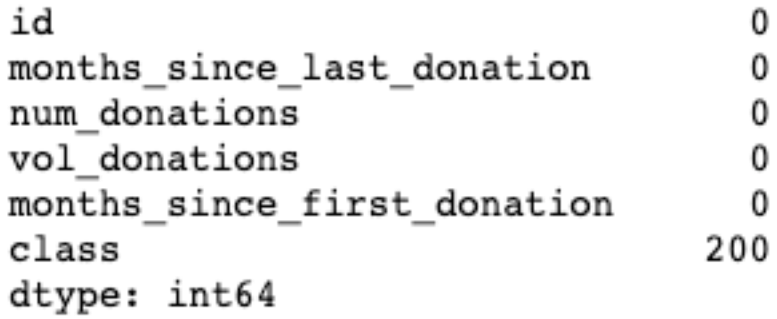
Methods

I present 5 predictive modeling methods: K-nearest neighbors, decision tree, GradientBoosting, logistic regression, and random forest. 80% of the data was used to train the models. To understand the model accuracy with unseen data, the remaining 20% of the data was reserved for the testing set. The same training and testing set was used across all models for consistency.

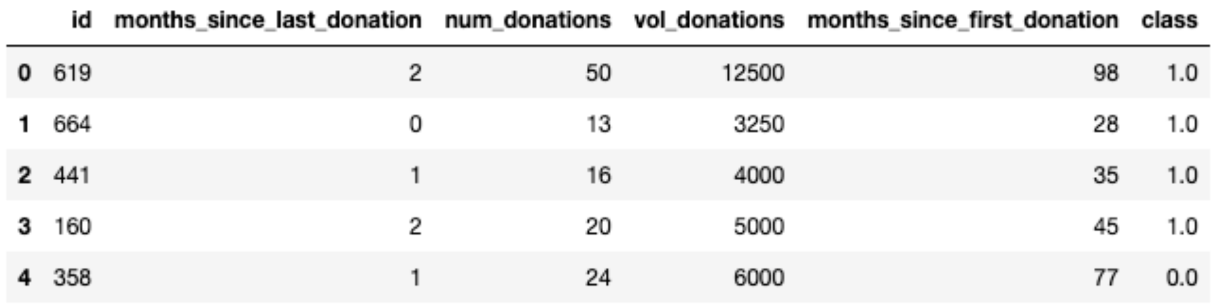


The dataset was randomly partitioned into training set and testing set using a 70/30 train/test partition and I join the training and test data in order to obtain the same number of features during categorical conversion Table 5. Models are trained using various algorithms using the entire training set, as well as trained on each model generated within the training set. Each model was trained once using what is sometimes referred to as a validation-set approach.

Table 5:



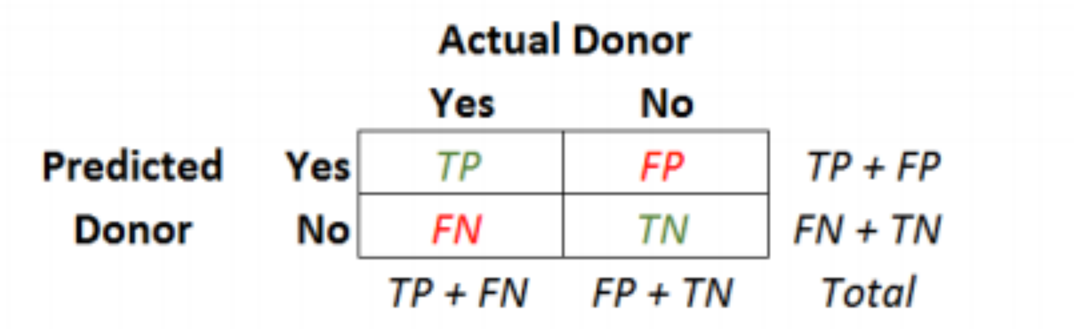
```
id                              0
months_since_last_donation      0
num_donations                   0
vol_donations                   0
months_since_first_donation     0
class                         200
dtype: int64
```



	id	months_since_last_donation	num_donations	vol_donations	months_since_first_donation	class
0	619	2	50	12500	98	1.0
1	664	0	13	3250	28	1.0
2	441	1	16	4000	35	1.0
3	160	2	20	5000	45	1.0
4	358	1	24	6000	77	0.0

Once models are trained, the test data is fed into each trained model to measure model performance. These measures allow us to gauge the generalizability of the remaining subset of data not used in the data and provides us a feel to the degree of how overfit any models are to the training data.

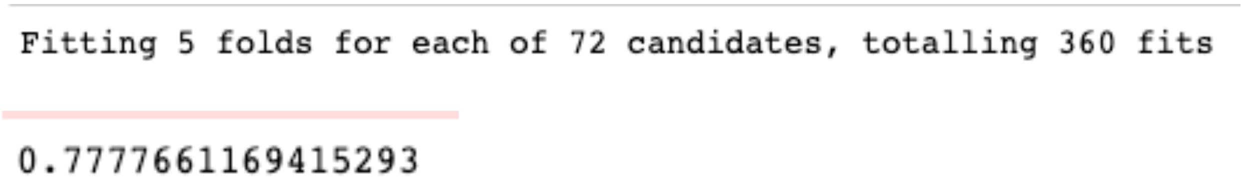
The statistical performance measures I obtained were overall accuracy, sensitivity, specificity, and area under the curve (AUC). The first three measures are easily calculated using a confusion matrix as shown in results section. The overall accuracy measures how well you classify donors versus non-donors (TP+TN/Total). Sensitivity measures how well I are able to correctly predict donors who have actually donated (TP/(TP+FN)). Specificity allows us to gauge how well I are able to predict non-donors among those who did not donate (FP/(FP+TN)).



		Actual Donor		
		Yes	No	
Predicted	Yes	TP	FP	TP + FP
Donor	No	FN	TN	FN + TN
		TP + FN	FP + TN	Total

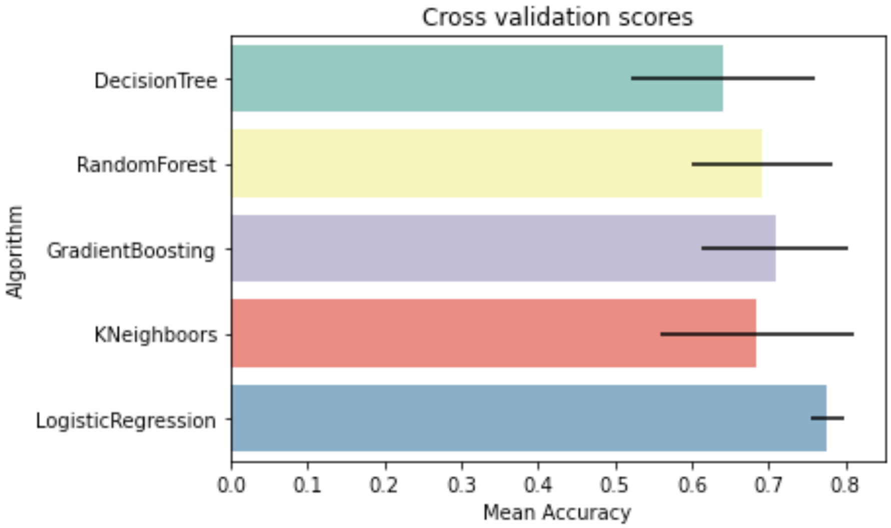
I did try 5-fold cross-validation on two of the models I researched. The idea here is to estimate a model over multiple folds (see result section) instead of just one random training set.

Score Results with Logistic Regression



```
Fitting 5 folds for each of 72 candidates, totalling 360 fits

0.7777661169415293
```

Cross-validation averages model fit performance measures such as prediction error to correct for the optimistic nature observed from the training error and thus provide an estimate of prediction risk that is more transparent. The reason I only tried this on a couple models is I have a very small data sets when the data sets are joined. Folds will make these training sets even smaller which might not provide any algorithm enough examples to learn.



Ensembling approaches are a family of machine learning algorithms which tend to convert weak learners to stronger ones. Random Forest is the most popular ensembeling technique used today. This algorithm ensembles decision trees which can be used for both classification and regression problems. In this method, multiple decision tree models are built on smaller samples. The final output of a classifier is determined by the mode of output of all trees and mean of the outputs if it is a regression problem. Many have found random forests to be one of the more competitive approaches in machine learning

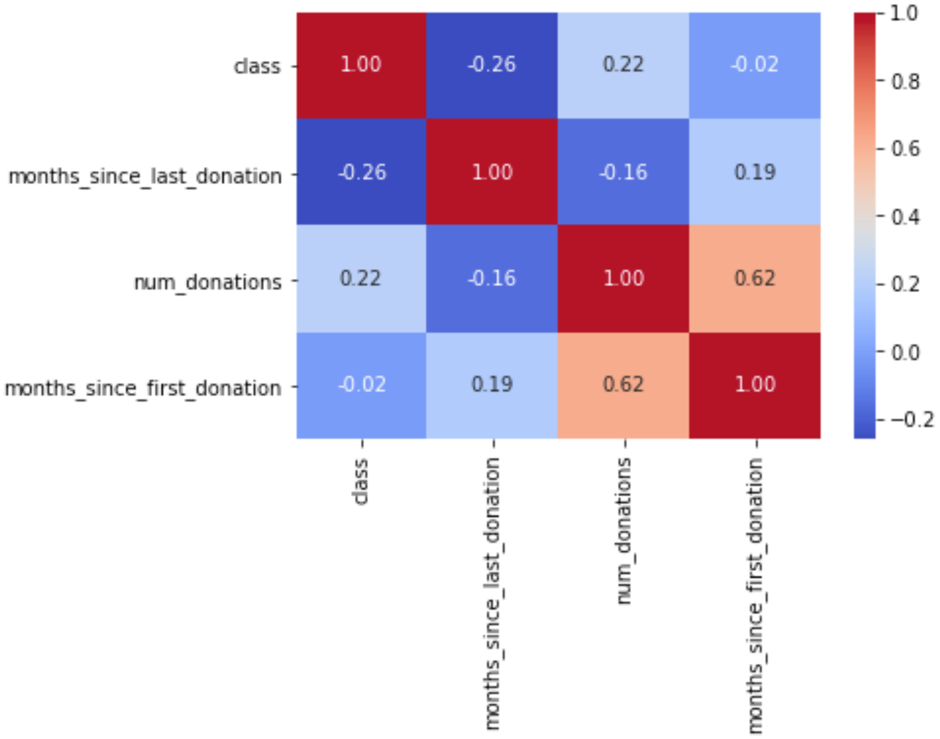
Results

In the initial exploratory analysis phase of this blood donation dataset, I tried to find a visible line of distinction between donors and non-donors. With the following graph of months since first donation, number of donations and months since last donation. I found that the groups of donors and non-donors were not visibly distinct. Since the number of features available for modeling was so few, exploratory data analysis (EDA) was very limited. I investigated interactions among features as well as tried two-way and three-way interactions as model inputs but either led to the same performance or poorer performance. In such cases, I decided to use the main effects and no interactions in any of the methods I investigated. I followed the commonly accepted philosophy in predictive modeling that a simpler, less complex model is preferred when the statistical performance measures are no different.

Correlation matrix results

Correlation matrix between numerical values (SibSp Parch Age and Fare values) and survived.

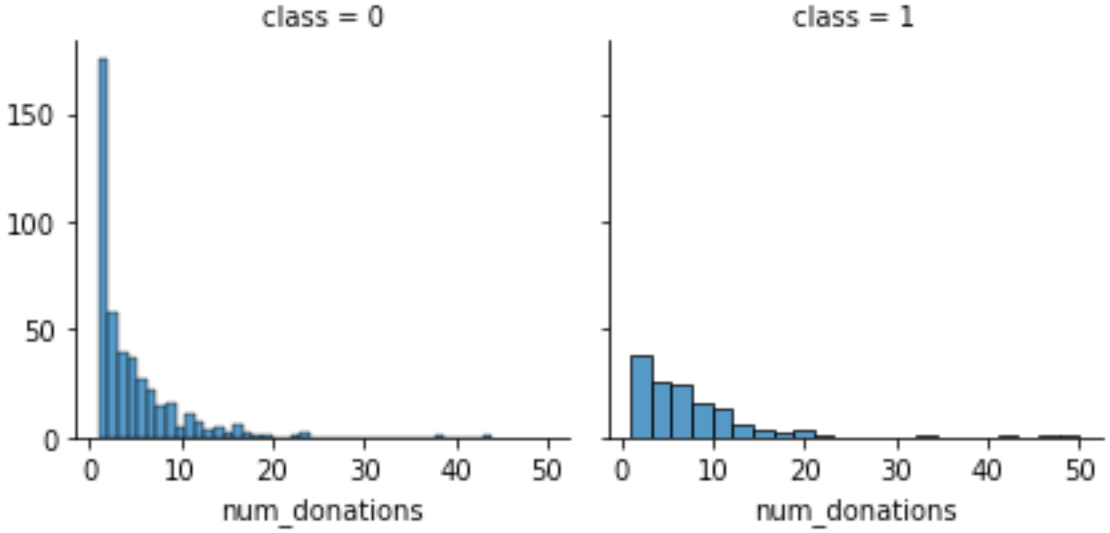
Only months_since_first_donation seems to have a significative correlation with the class probability. It doesn't mean that the other features are not useful. num_donations in these features can be correlated with the class. To determine this, I need to explore in detail these features.



Num donations results

I notice that num_donations distributions are not the same in the class 1 and class 0 subpopulations. Indeed, there is a peak corresponding to the people who have donated only 0-1 time will not donate blood and who have donated 2-3 will likely donate.

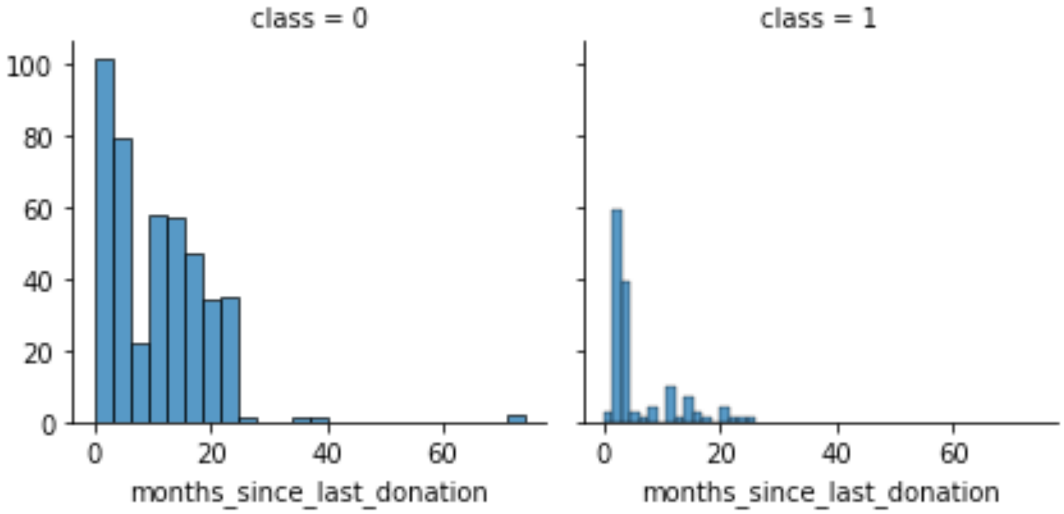
It seems that people have donated a greater number of times are more likely to donate blood.



Months since last donation results

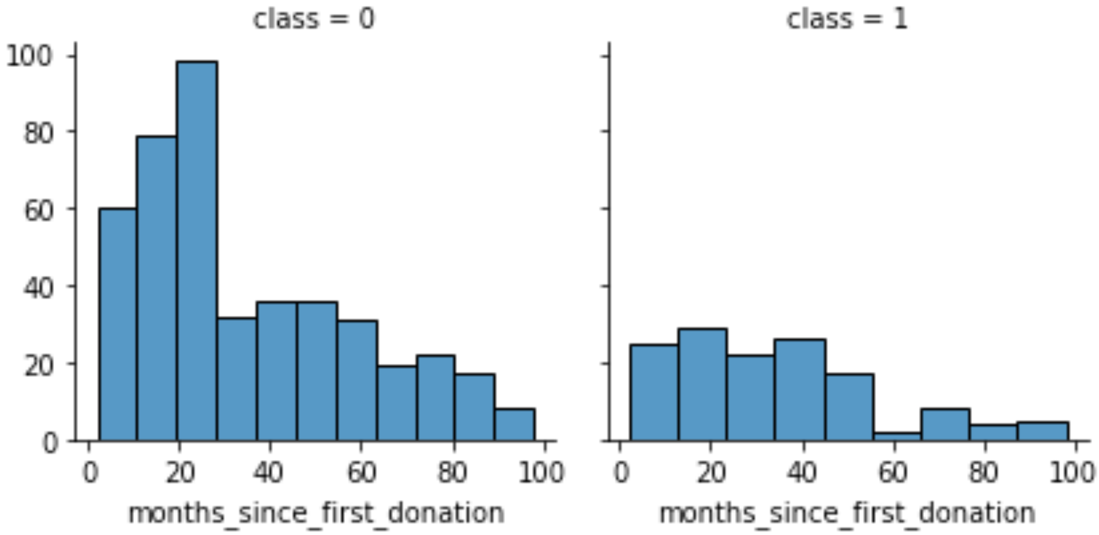
I notice that months_since_last_donation distributions are not the same in the class 1 and class 0 subpopulations. Indeed, there is a peak corresponding to the people who have donated recently (in 1-2 months) will donate blood.

It seems that people have donated recently are more likely to donate blood.



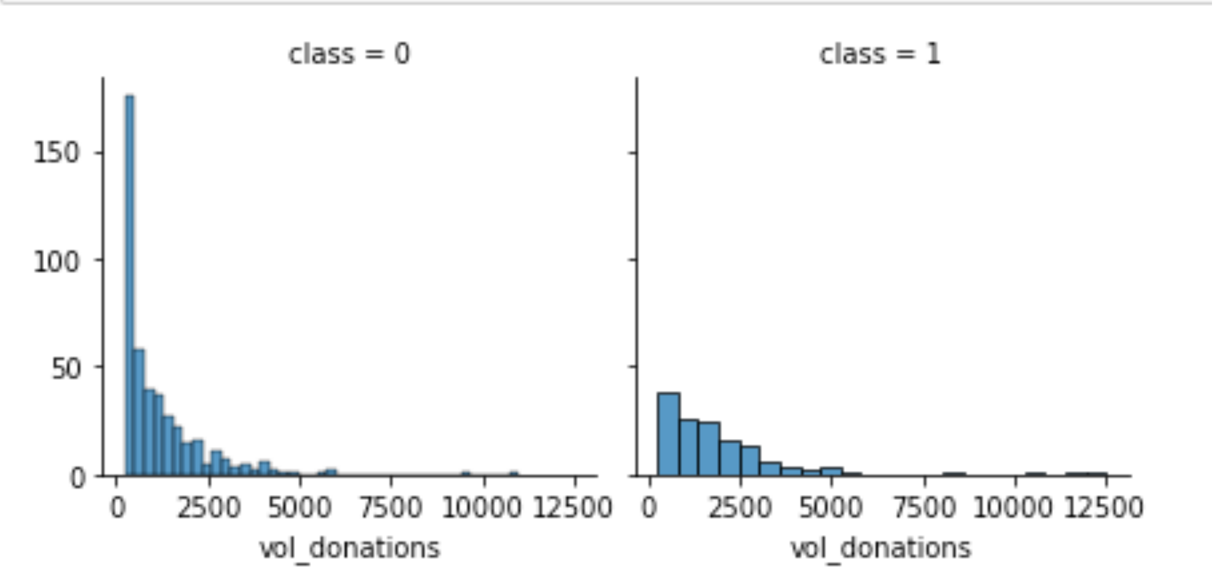
Months since first donation results

I notice that months_since_first_donation distributions are not the same in the class 1 and class 0 subpopulations. Indeed, there is a peak corresponding to the people who have just donated recently (in 6-20 months) will not donate blood.



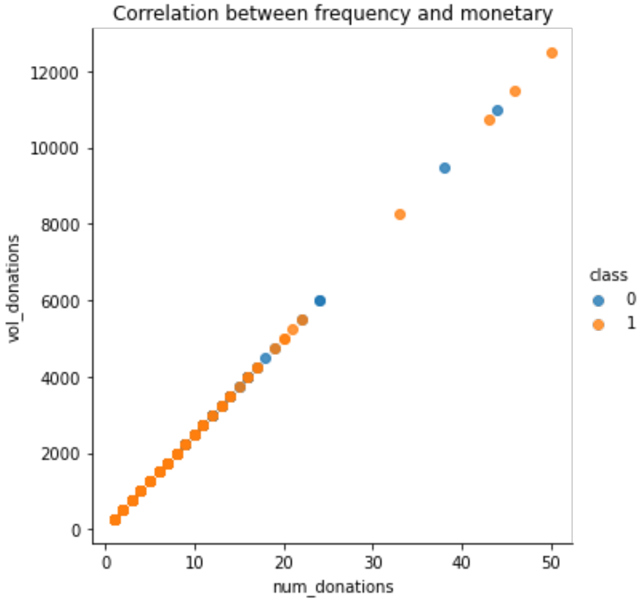
Volume donated results

Volume donated is also a good feature to know whether the donor will donate or not.



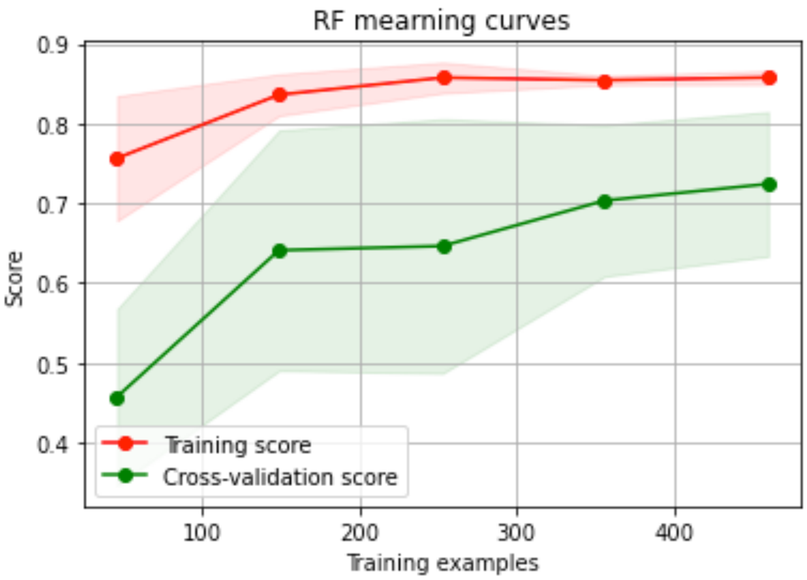
Correlation between frequency and monetary results

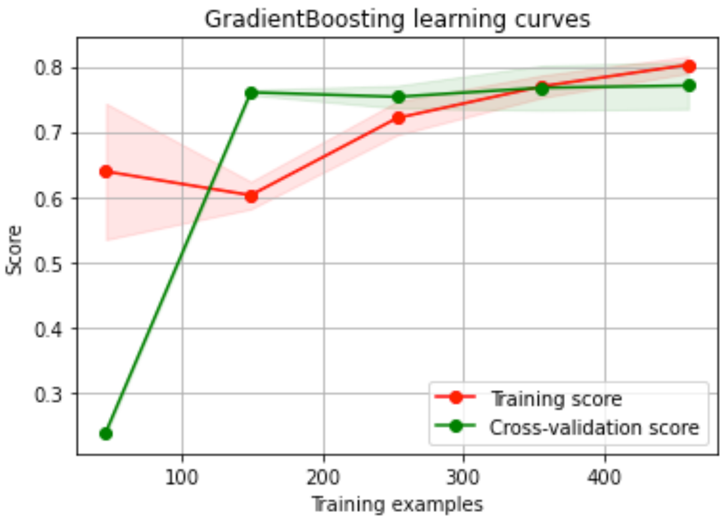
From the graph we can see that Frequency and monetary values are highly correlated. So, we can use only the frequency.



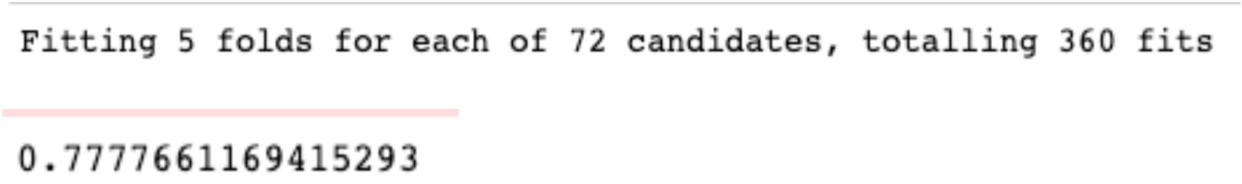
By looking at the learning curve GradientBoosting and Random Forest classifiers tend to overfit the training set. According to the growing cross-validation curves Random Forest classifier seems to better generalize the prediction since the training and cross-validation curves are close together.

ROC curves for each model

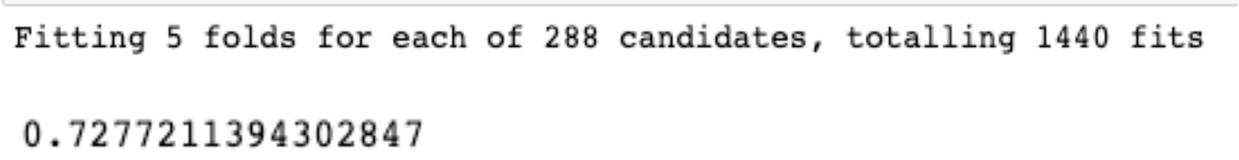




Score Results with Logistic Regression and Random Forest

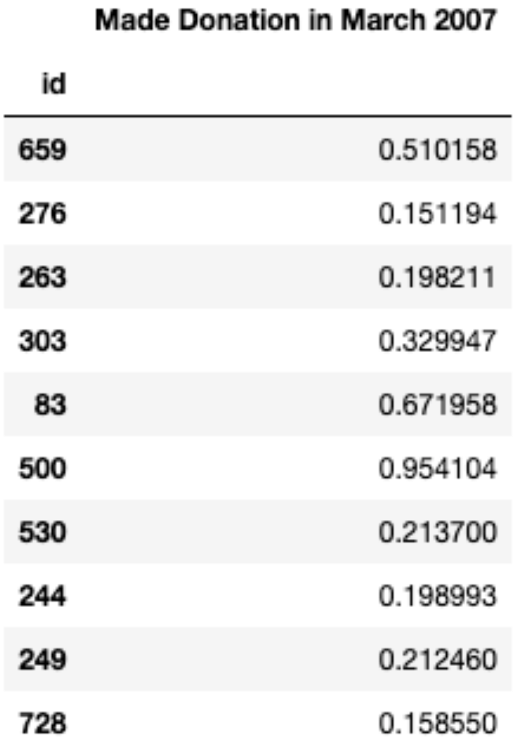


```
Fitting 5 folds for each of 72 candidates, totalling 360 fits

0.7777661169415293
```



```
Fitting 5 folds for each of 288 candidates, totalling 1440 fits

0.7277211394302847
```

Now we can target the people who are interested in donating blood, and which will result in getting more volunteers and we can save more people.



	Made Donation in March 2007
id	
659	0.510158
276	0.151194
263	0.198211
303	0.329947
83	0.671958
500	0.954104
530	0.213700
244	0.198993
249	0.212460
728	0.158550

Conclusion

We have compared the performance of various binary classification algorithms and data to see if we can better predict if a person is going to donate blood or not. Among the algorithms examined, the non-clustered 5-fold cross-validated logistic regression and random forest model performed the best based on the test set AUC. However, AUC alone may not be best performance measure with respect to likelihood to predict blood. That is because AUC considers the area determined by True Positive Rate (TPR)/sensitivity and False Positive Rate (FPR)/(1-Specificity). Our model could be used for targeted advertisement. In such a case, we are more interested in the class 1 which would be to target the actual donors who would be interested in donating blood regularly. Hence, our performance would focus more on sensitivity leading us to recommend maybe a cluster model for this type of prediction.

References

Chapman, P., et al. (2000). "CRISP-DM 1.0 Step-by-step data mining guide."

Ferguson, E., et al. (2007). "Improving blood donor recruitment and retention: integrating theoretical advances from social and behavioral science research agendas." Transfusion 47(11): 1999-2010. Godin, G., et al. (2007). "Determinants of repeated blood donation among new and experienced blood donors." Transfusion 47(9): 1607-1615. James, G., et al. (2013). "An Introduction to Statistical Learning." Katsaliaki, K. (2008). "Cost-effective practices in the blood service sector." Health policy 86(2): 276-287.

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.

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): 579-594. Veldhuizen, I., et al. (2011). "Exploring the dynamics of the theory of planned behavior in the context of blood donation: does donation experience make a difference?" Transfusion 51(11): 2425-2437. Wikipedia C4.5 algorithm