Blood Donation Shortage Prediction

Thip Rattanavilay,

DSC 630 – Predictive Analytics

Bellevue University

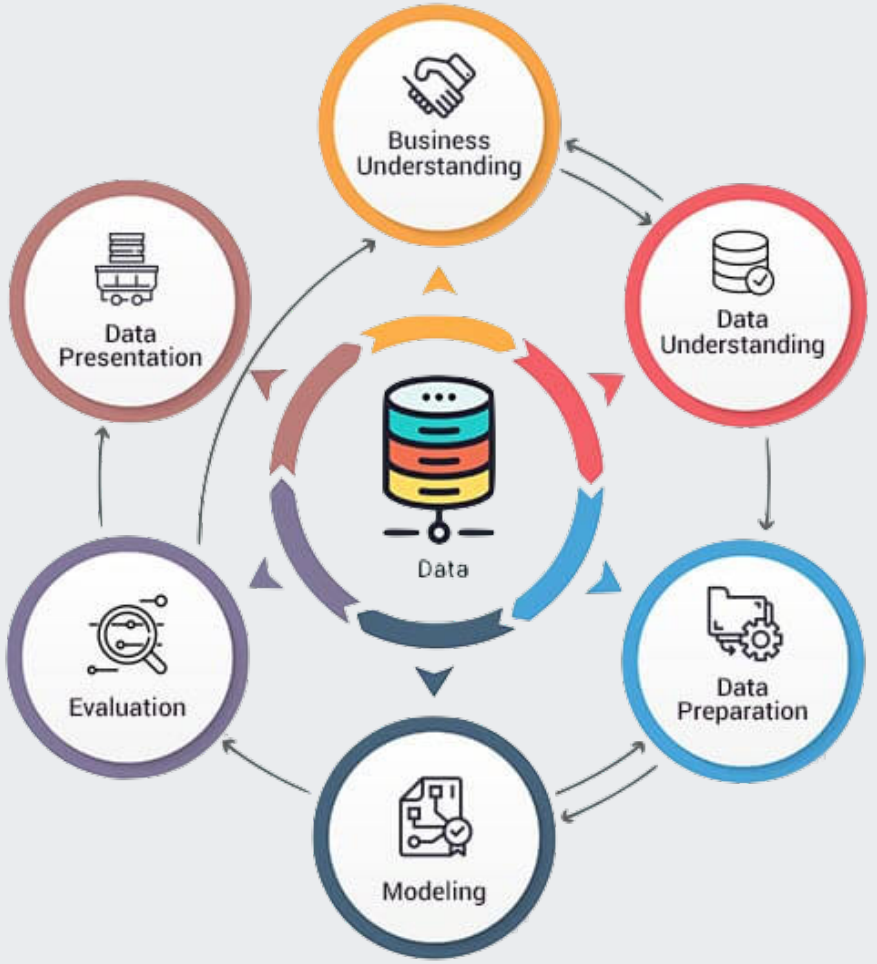
INTRODUCTION

- Most of us have donated blood at least once in our lifetime or has not.
- Blood Banks faces supply and demand and shortage
- We want to know often do the donors return to donate blood.
- There are variety of factors that can impact a blood donation.
- This project uses a dataset with various features to solve this problem
- We want to find how many donations are being made and attempt to predict blood donations.
- This gathering will be useful in targeting the people who are interested in donating blood, which results in getting more volunteers to save more lives.

METHOD & STEPS

- Understanding the business
- Data understanding
- Exploratory analysis of the dataset
- Data cleaning and preparation
- Split the data set into train and test set
- Training and Test the models
- Feature selection
- Model evaluation
- Results

CRISP-DM



Transfusion dataset #1

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	250.000000	2.000000	0.000000
25%	2.750000	2.000000	500.000000	16.000000	0.000000
50%	7.000000	4.000000	1000.000000	28.000000	0.000000
75%	14.000000	7.000000	1750.000000	50.000000	0.000000
max	74.000000	50.000000	12500.000000	98.000000	1.000000

DATA UNDERSTANDING & EXPLORATORY DATA ANALYSIS

This data frame has 748 rows and 5 variables.

- **Key variables are**
  - **Recency months**
  - **Frequency times**
  - **Monetary c.c blood**
  - **Time months**
  - **Whether he/she Donated blood in March 2007**

Transfusion dataset #2

	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

DATA UNDERSTANDING & EXPLORATORY DATA ANALYSIS

- This data frame has 576 rows and 5 variables.
- Key variables are
  - id
  - Months since last donation
  - Num donations
  - Vol donations
  - Months since first donation
  - Class

Features and Joined datasets

Joining these features as one dataset

- Months since Last Donation
- Number of Donations
- Total Volume Donated
- Months since First Donation

- Recency
- Frequency
- Time
- Monetary

Training and test dataset

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)
132	2	2	500	10
294	11	5	1250	35
522	4	13	3250	39
291	16	12	3000	50
106	0	8	2000	59

```
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

	id	months_since_last_donation	num_donations	vol_donations	months_since_first_donatio	class
0	619	2	50	12500	9	1
1	664	0	13	3250	2	1
2	441	1	16	4000	3	1
3	160	2	20	5000	4	1
4	358	1	24	6000	7	0

DATA UNDERSTANDING & EXPLORATORY DATA ANALYSIS

In the class column there are two classes

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

Let us assume that *class 1* means donated, and *class 0* means not donated

DATA PREPARATION & CLEANING

Inspecting transfusion dataset

	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

Checked the data frame for null, or missing value

```
Recency (months)                                    0
Frequency (times)                                   0
Monetary (c.c. blood)                               0
Time (months)                                       0
whether he/she donated blood in March 2007          0
dtype: int64
```

No missing values, and we have 576 rows and 6 Columns with clean data.

The features are 'Months since Last Donation', 'Number of Donations', 'Total Volume Donated', 'Months since First Donation'.

TRAINING & TEST MODELING

Training is the process of applying the available data to the chosen algorithms

For the project I will be using 5 prediction models

I will also find Correlation between Features

Training multiple models allow comparing and choosing the best performing model.

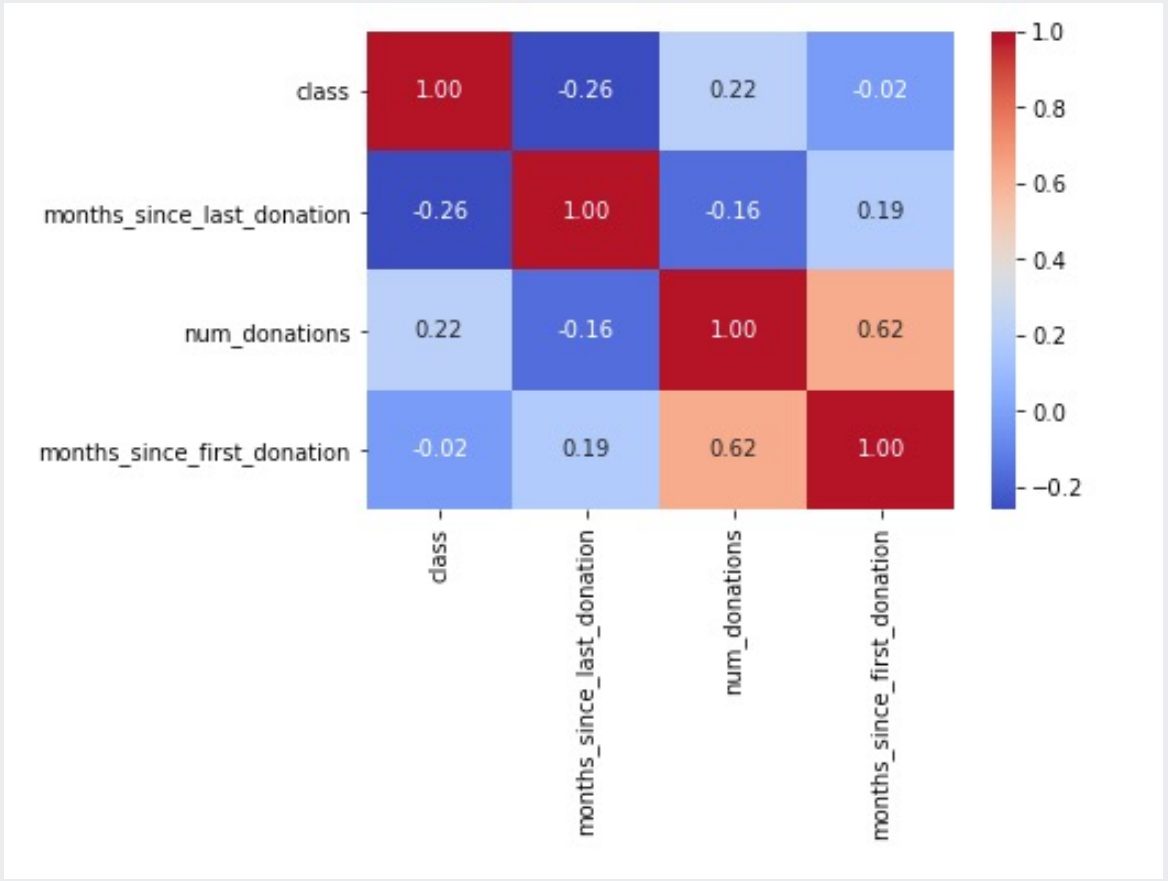
Using 5 models in total with cross-validation

FEATURE SLECTION

Correlation matrix between numerical values.

Only months_since_first_donation seems to have a significative correlation with the class probability.

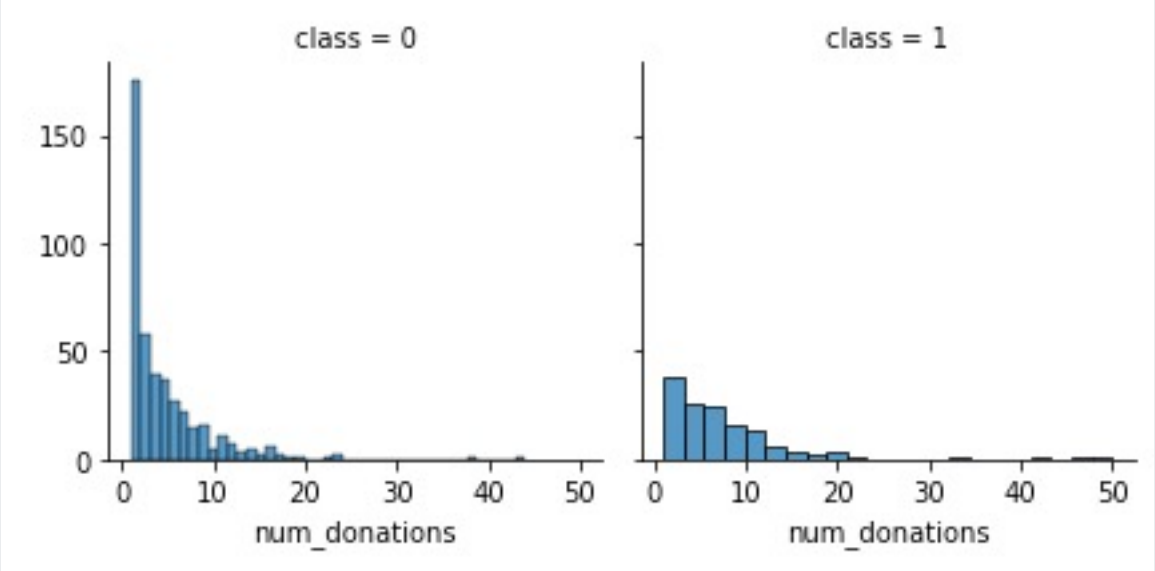
It doesn't mean that the other features are not useful. num_donations in these features can be correlated with the class.



FEATURE SLECTION

We 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. We can see that class 0 has a higher number than class 1. There will be no donation class 0 and fewer donation for class 1.

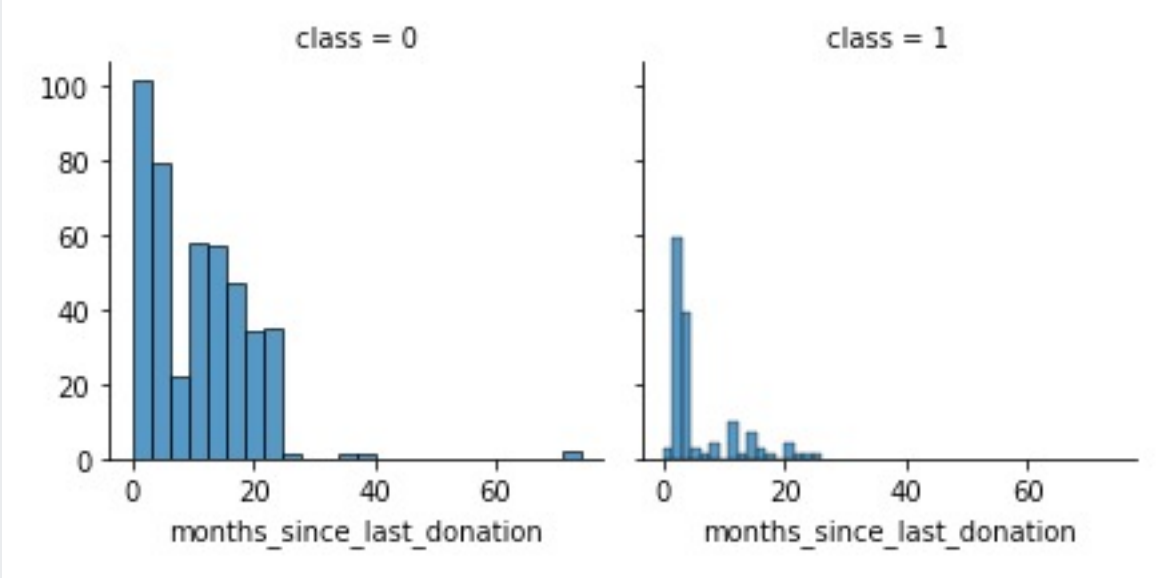


FEATURE SLECTION

We 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.

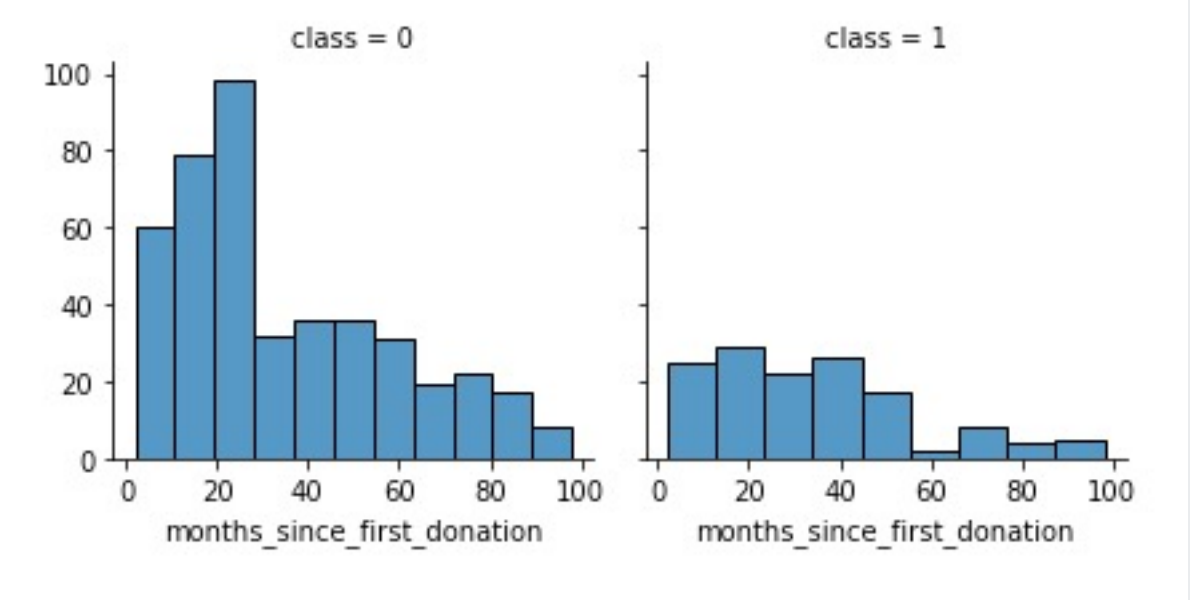
It seems that people have donated recently are more likely to donate blood.



FEATURE SLECTION

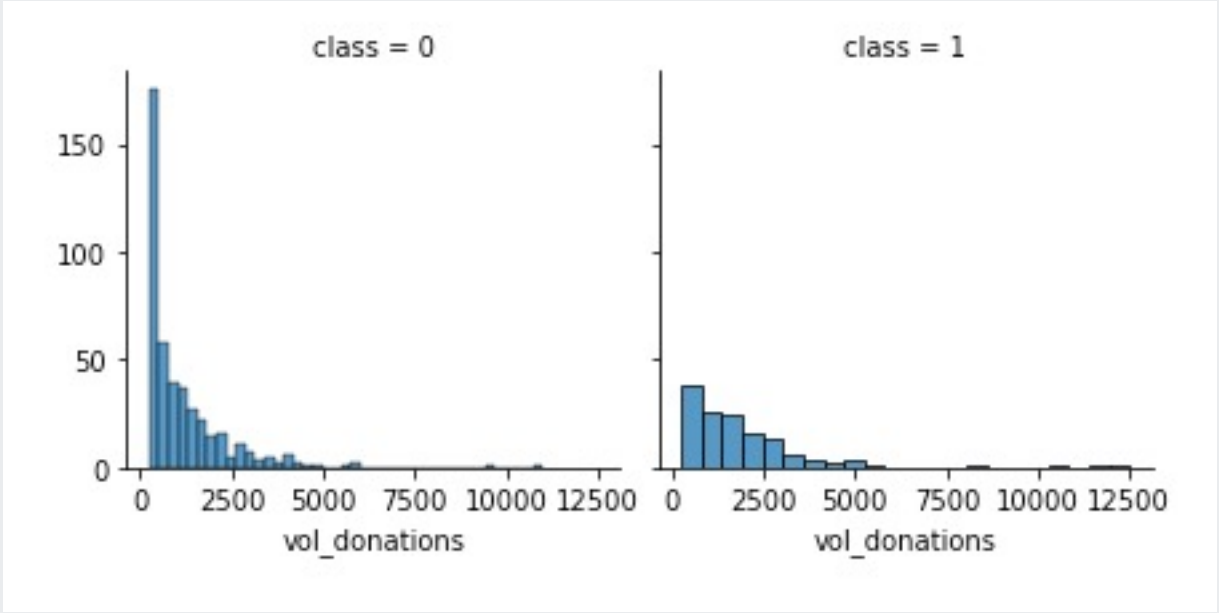
We notice that months_since_first_donation distributions are not the same in the class 1 and class 0 subpopulations.

Indeed, there is a peak corresponding to the people who have just donated recently (in 6-20 months) will not donate blood.



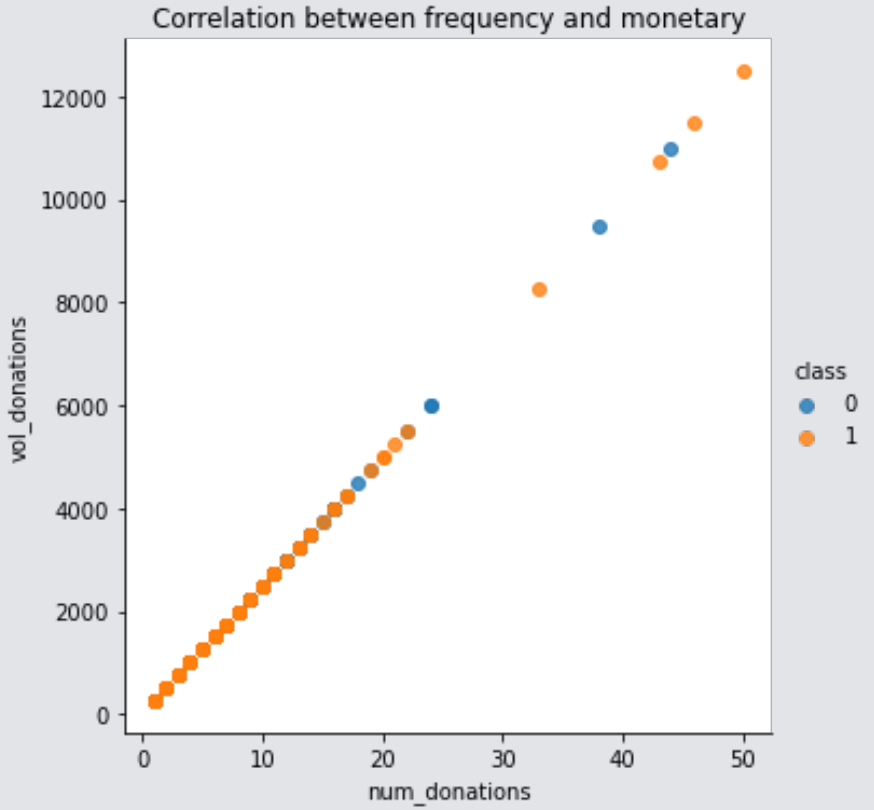
FEATURE SLECTION

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



Correlation between frequency and monetary

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



Modeling

Cross Validation Models

I will compare five (5) popular classifiers and I will evaluate them using the mean accuracy of each of them using the kfold cross validation procedure.

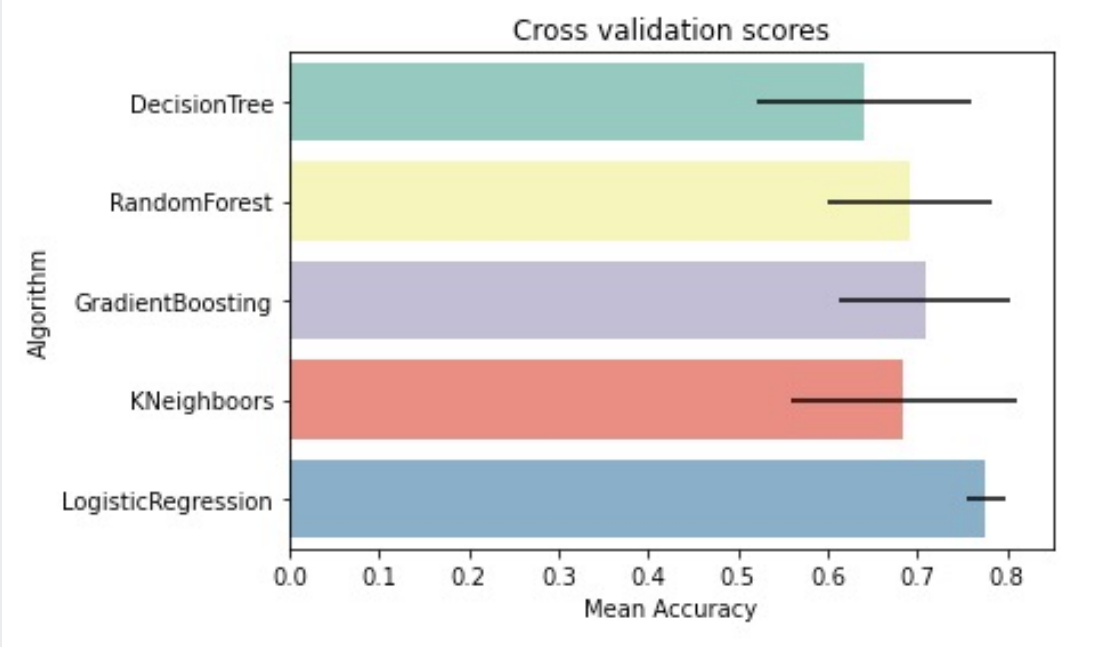
- Decision Tree
- Random Forest
- Gradient Boosting
- KNN
- Logistic regression

Modeling

Cross Validation Models

I compared five (5) popular classifiers and evaluate the mean accuracy of each of them by a stratified kfold cross validation procedure.

By seeing the figure, we can see that what model will work the best.



Modeling

Hyperparameter tunning for best models

I performed a grid search optimization for Random Forest, Gradient Boosting classifiers and Logistic Regression.

The computation time is clearly reduced.

Random Forest

```
Fitting 5 folds for each of 288 candidates, totalling 1440 fits
[Parallel(n_jobs=4)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done  62 tasks       | elapsed:    8.5s
[Parallel(n_jobs=4)]: Done 212 tasks       | elapsed:   29.7s
[Parallel(n_jobs=4)]: Done 462 tasks       | elapsed:  1.1min
[Parallel(n_jobs=4)]: Done 812 tasks       | elapsed:  2.2min
[Parallel(n_jobs=4)]: Done 1262 tasks      | elapsed:  3.4min
[Parallel(n_jobs=4)]: Done 1440 out of 1440 | elapsed:  3.9min finished
0.7277211394302847
```

Gradient Boosting

```
Fitting 5 folds for each of 72 candidates, totalling 360 fits
[Parallel(n_jobs=4)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done  76 tasks       | elapsed:    2.2s
[Parallel(n_jobs=4)]: Done 360 out of 360 | elapsed:   10.4s finished
0.7777661169415293
```

Logistic Regression

```
Fitting 5 folds for each of 28 candidates, totalling 140 fits
[Parallel(n_jobs=4)]: Done 123 tasks       | elapsed:    2.1s
[Parallel(n_jobs=4)]: Done 140 out of 140 | elapsed:    3.3s finished
0.7708095952023988
```

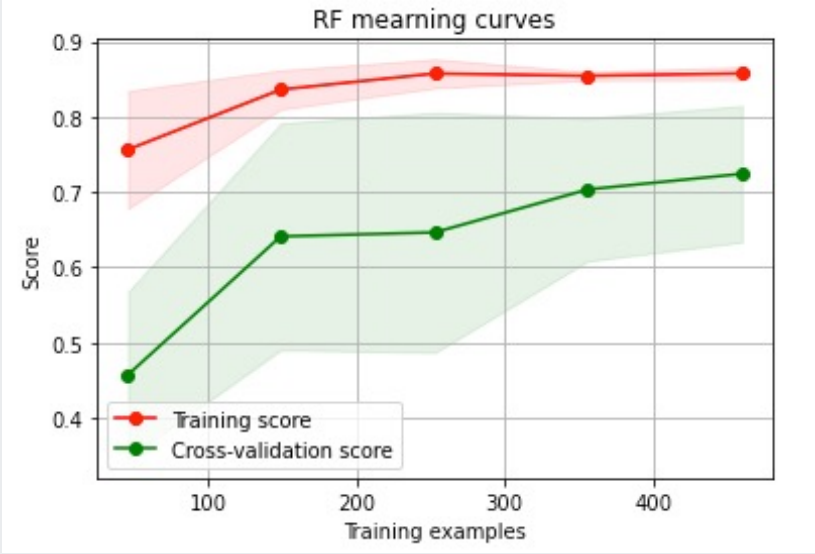
Modeling

Plot learning curve

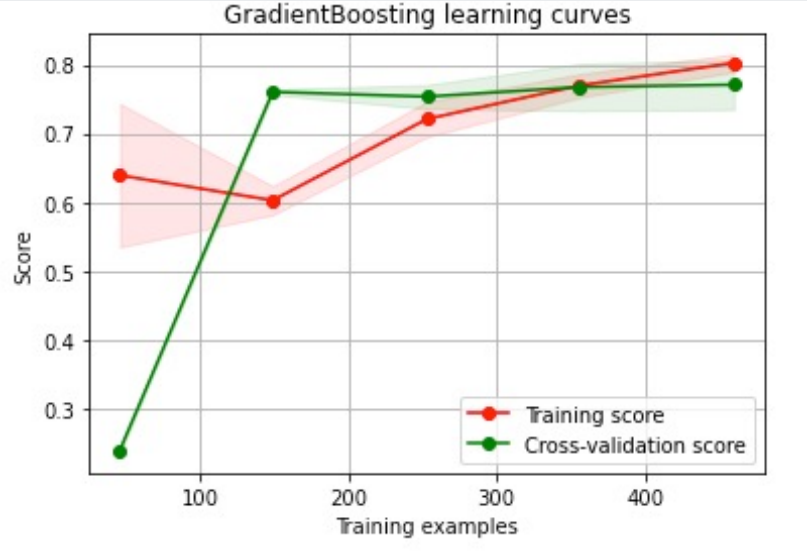
I performed a learning curve for Random Forest, and Gradient Boosting classifiers.

According to the cross-validation curves Random Forest classifier seems to be the better prediction model. Because the training and cross-validation curves are close together.

Random Forest



Gradient Boosting



RESULTS

	precision	recall	f1-score	support
0	0.80	0.89	0.84	114
1	0.43	0.28	0.34	36
accuracy			0.74	150
macro avg	0.62	0.58	0.59	150
weighted avg	0.71	0.74	0.72	150

```
LogisticRegression : 0.793
LogisticRegression AUC score : 0.755


Supportvector : 0.767
Supportvector AUC score : 0.722


GaussianNB : 0.747
GaussianNB AUC score : 0.700


rfc : 0.740
rfc AUC score : 0.710
```

```
Voting Classifier accuracy: 0.78

VotingClassifier AUC score: 0.750
```

	Feature	F statistic	p value
0	Recency (months)	46.977265	0.0000
1	Frequency (times)	23.608148	0.0000
2	Monetary (c.c. blood)	23.608148	0.0000
3	Time (months)	0.373632	0.5413

RESULTS PREDICTION

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

id	Made Donation in March 2007
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

Thank you!

- In results with prediction, we measured accuracy on donations donated each months
- We have achieved nearly 78% accuracy and hoping to get a better result from more cross-validation and methods
- Donors are more likely to donate and become a repeat donor

COME AND DONATE TODAY

REFENCES

- https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/
- https://www.usatoday.com/story/news/health/2021/06/23/blood-banks-urge-donations-us-blood-supply-drops-demand-increases/5312985001/
- https://machinelearningmastery.com/tpot-for-automated-machine-learning-in-python/
- https://rdrr.io/cran/simpleNeural/man/UCI.transfusion.html