Blood Donation Shortage Prediction

Thip Rattanavilay,

DSC 630 – Predictive Analytics

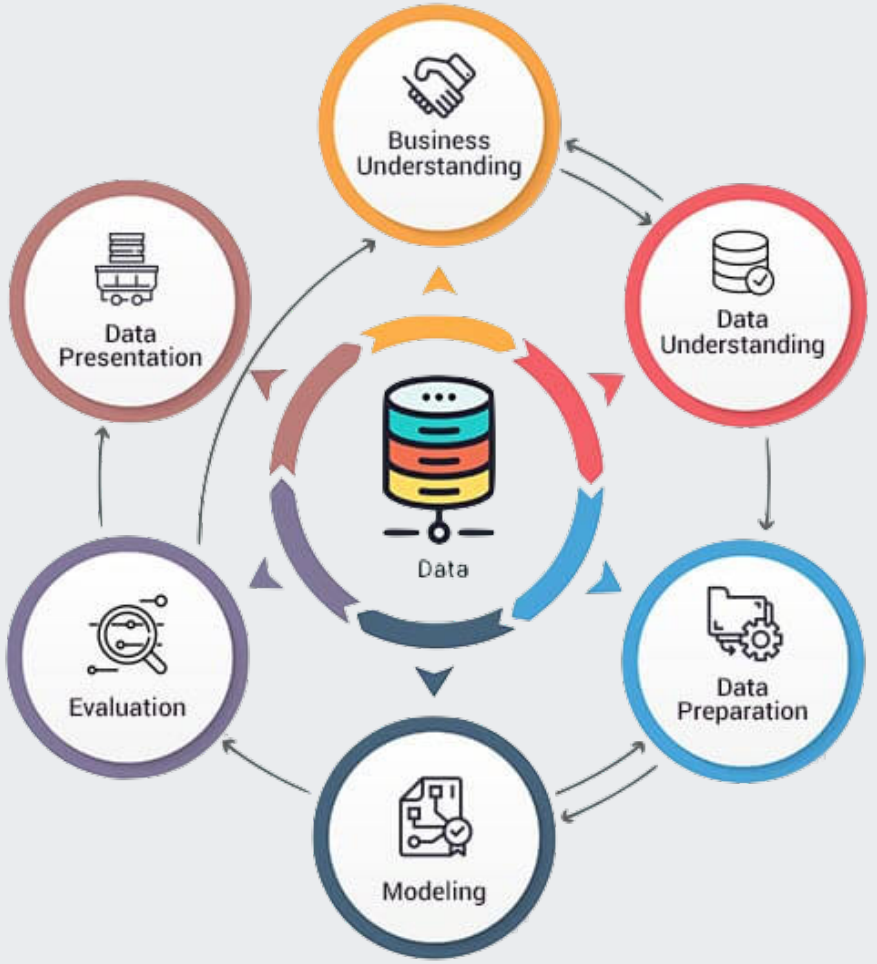
Bellevue University

INTRODUCTION & PROBLEM STATEMENT

- Most of us have donated blood at least once in our lifetime.
- How often do you return to donate in order to save more lives?
- Blood Banks problem is supply and demand.
- There are variety of factors that impact the blood donation is repeat donations.
- This project uses a dataset with various features about the donation being made the attempt to predict blood donations.
- This 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
- Discuss the best Model deployment



	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 748 entries, 0 to 747
Data columns (total 5 columns):
 #   Column                                           Non-Null Count  Dtype
---  ------                                           --------------  -----
 0   Recency (months)                                 748 non-null    int64
 1   Frequency (times)                                748 non-null    int64
 2   Monetary (c.c. blood)                            748 non-null    int64
 3   Time (months)                                    748 non-null    int64
 4   whether he/she donated blood in March 2007       748 non-null    int64
dtypes: int64(5)
memory usage: 29.3 KB
```

DATA UNDERSTANDING & EXPLORATORY DATA ANALYSIS

- The dataset used in this analysis is from UCI.
- Inspecting the transfusion data frame from March 2007.
- 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
- Want to split the transfusion into train and test dataset
- Correlation between features
- Probability Distribution Curves
- Feature Importance
- Train the model

DATA PREPARATION & CLEANING

Inspecting transfusion dataset

	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

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
```

```
Index(['Recency (months)', 'Frequency (times)', 'Monetary (c.c. blood)',
       'Time (months)', 'whether he/she donated blood in March 2007'],
      dtype='object')
```

TRAINING & TEST MODELING

- Split data into three sets, 5%, 20% and 2%
- Training is the process of applying the available data to the chosen algorithms
- For the project I will be using TPOT and Linear Regression
- I will also find Correlation between Features
- Training multiple models allow comparing and choosing the best performing model.
- I ended up training 25 models in total with cross-validation

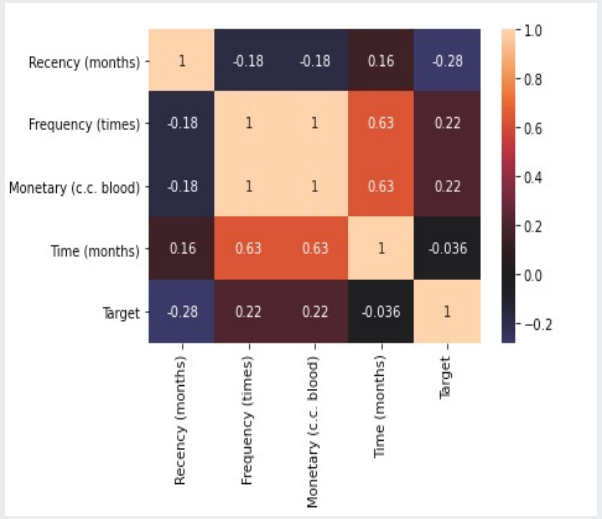
	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)
132	2	2	500	10
294	11	5	1250	35
522	4	13	3250	39
291	16	12	3000	50
106	0	8	2000	59

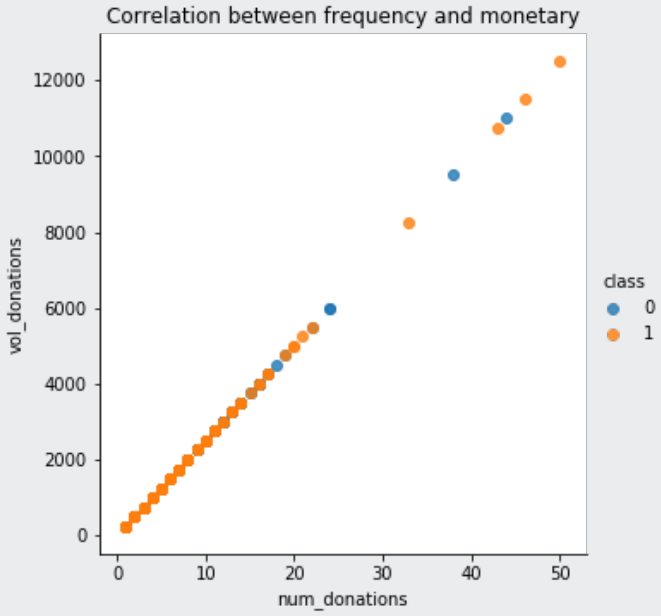
```
Generation 1 - Current best internal CV score: 0.740869179713229
Generation 2 - Current best internal CV score: 0.7420690542167971
Generation 3 - Current best internal CV score: 0.7438028147056875
Generation 4 - Current best internal CV score: 0.7457258916287643
Generation 5 - Current best internal CV score: 0.7457258916287643
Best pipeline: DecisionTreeClassifier(MinMaxScaler(input_matrix), criterion=gini, max_depth=6, min_samples_leaf=20, m
in_samples_split=12)

AUC score: 0.7474

Best pipeline steps:
1. MinMaxScaler()
2. DecisionTreeClassifier(max_depth=6, min_samples_leaf=20, min_samples_split=12,
                          random_state=1)
```

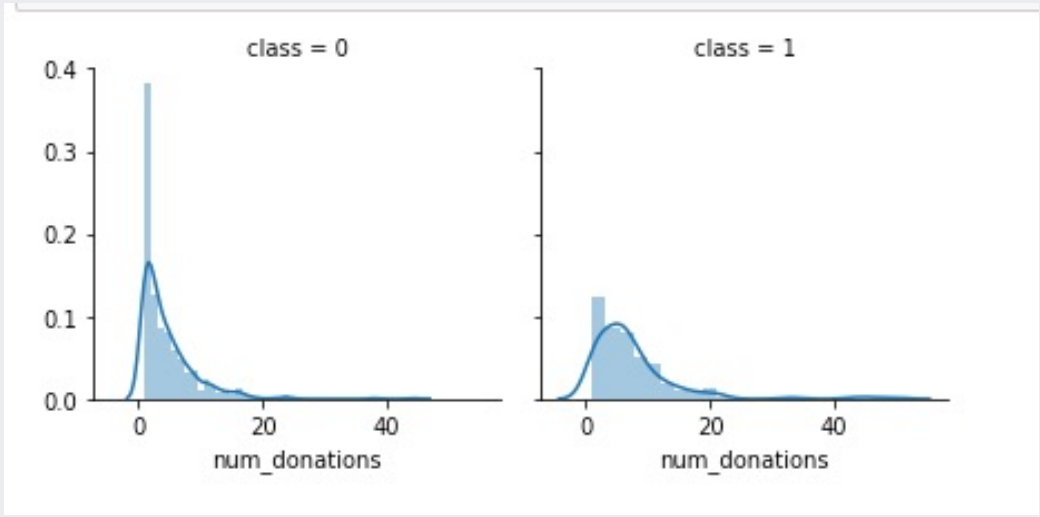
```
Recency (months)      62.281
Frequency (times)     33.797
Time (months)        607.188
monetary_log           0.850
dtype: float64
```

FEATURE SLECTION

- We have a significative correlation with the class probability
- notice that months 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
- You can see that Frequency and monetary values are highly correlated. So we can use only the frequency.







EVALUATION

- Compared AUC, Accuracy, Precision, F1 and Recall scores between train and validation for generalization and between models for performance
  - Overall, scores were similar between both except for Precision, as expected, because the train data set was balanced, and the validation data set was not.

    - logistic regression
    - Support Vector Classification
    - Naive biar classification

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

RISK & NEXT STEPS

- Results in other slide shows prediction are accurate or close to donations donated each months
- At this point of the analysis have achieved nearly 80% accuracy and hoping to get a better result from more cross-validation and methods
- Exploring more options to use ensemble method to gain better accuracy
- I will work on features as well to improve blood donations prediction

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