PREDICTING BLOOD DONATIONS

A Machine Learning Data Analysis

Faisal Fattani
Intern Data Analyst

Feb. 28th, 2021

CONTENTS

OVERVIEW

Blood supplies are vital for a very wide range of usage; from emergency trauma procedures to chronic patients that continuously depend on transfusions and many more. However, the sensitivity of blood and the conditions by which it needs to be stored in, along with the fact that it can only be stored for a limited time deem constant need for blood donations. The red cross reported the need for about 38,000 daily donations, and predictability ensures supply levels staying afloat, preventing many medical catastrophes and even life loss.

This report details an analysis targeted to predict blood donation instances through machine learning. In this report, I will share my approach to the problem and the technical methodology by which I do so, its implementation, and what the conclusions of the analysis entail.

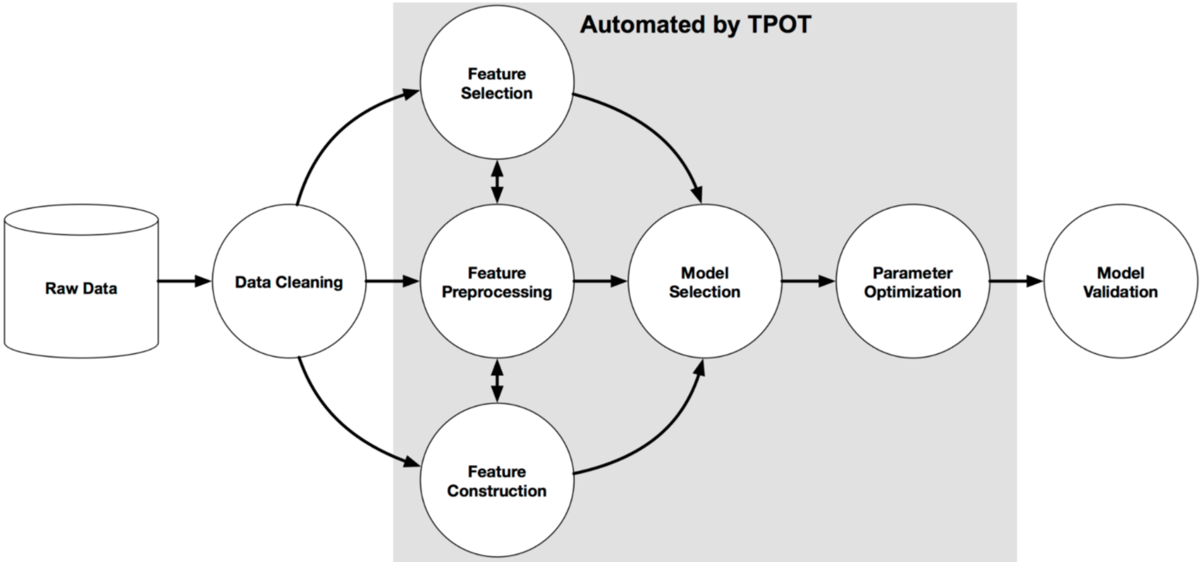
About MTE

MedTourEasy is a global healthcare company that provides the informational resources needed to evaluate global options. MTE helps find the right healthcare solutions based on specific health needs, and affordable care while meeting the quality standards that are expect in healthcare - providing an easy to use platform and service that helps patients get medical second opinions and schedule affordable, high-quality medical treatment abroad.

METHODOLOGY

The project sets up a pipeline to dynamically configure and assess models to find the most accurate fit to predict if a previous donor would donate at a given instance.

The overall work flow of the project is structured as follows:



Data Source

The project uses data sourced from the Blood Transfusion Service Center in Taiwan. The dataset contains 847 blood donors, including the following information about each individual:

- Months since last transfusion
- Total number of donations
- Total blood donated
- Timespan since first donation

Along with the above-mentioned information, the dataset also included weather each individual donated for a specific month (March 2007) - our target variable.

TPOT Pipeline

This project used the **TPOTCLASSIFIER** from the TPOT package, an automated pipeline exploring different algorithms and models; the **ROC_AUC_SCORE** from SKlearn was specified as the validation metric to compare models and choose the best fit based on.

Prediction Model

As TPOT found logistic regression to fit the data best for prediction, the **linear_model.LogisticRegression** implementation from the Sklearn package was used.

IMPLEMENTATION

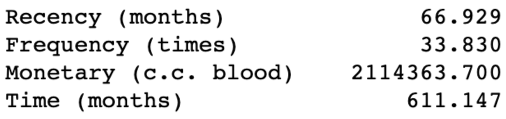
DATA PROCESSING AND BASIC EXPLORATION

The dataset was relatively clean with no missing values, nor were there any misinterpreted values or any implicit biases that were identified. Here are some of the summary statistics from the data:

	Recency (Months)	*Times donated*	*Total Amount (CC)*	*Timespan (Months)*
Mean	9.51	5.51	1378.68	34.28
Std. Deviation	8.10	5.84	1459.83	24.38
25th Percentile	2.75	2.00	500.00	16.00
75th Percentile	14.00	7.00	1750.00	50.00

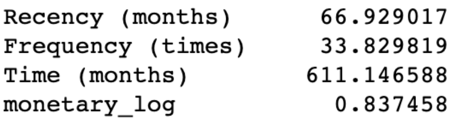
Upon an overview of variance in the data, one variable seemed to stand out with high variance: donated blood amounts. Data for this variable were then log-normalized to decrease variance and improve the model's performance.

DATA VARIANCE



```
Recency (months)                 66.929
Frequency (times)                33.830
Monetary (c.c. blood)       2114363.700
Time (months)                   611.147
```

VARIANCE AFTER LOG-NORMALIZATION



```
Recency (months)       66.929017
Frequency (times)      33.829819
Time (months)         611.146588
monetary_log            0.837458
```

After making sure the variance for each variable in the data was within a magnitude of order, the data were split into training and testing splits (a **0.75/0.25** split) while stratifying on the targeted variable, ensuring a consistent distribution across the splits. The variable of weather an individual donated at the given instance (March 2007) will be the target for the model to predict a future donation instance.

SETTING UP THE TPOT PIPELINE

As mentioned earlier in the methodology section, a TPOTCLASSIFIER was instantiated; using the built-in ‘tpot-light’ configuration for fast processing. A seed was specified for debugging and reproducibility, along with supplying arguments like the population size and SKlearn’s Compute Area Under the Receiver Operating Characteristic Curve (ROC_AUC) score.

TPOT found **Logistic Regression** to be the best model with an AUC score of **0.7850.**

TRAINING THE MODEL

The logistic regression model from the SKlearn package was imported and trained on the training set using the ‘liblinear’ solver. Even though an AUC score was previously given as the pipeline compared models, the model was validated again to test its accuracy on the testing split from the data; the model predicted donations with an accuracy of **0.76**.

CONCLUSIONS

Given an individual’s history of the four given variables – time since last donation, number of times donated, overall donated amount, and timespan since the first donation – a model was developed to predict if the individual would donate or not with an accuracy of 0.76. The model produced the following regression coefficients for the variables;

Recency	:	- 0.09
Frequency	:	0.96
Amount	:	- 0.03
Timespan	:	0.29

Generally associating more established donors to have higher probabilities to donate again, with frequency and timespan having the most impact.

Medical facilities and healthcare providers having a better ability to estimate blood donations ensures a steady supply of blood and the treatment that is contingent, periods of relative scarcity can be avoided if foreseen early either by taking direct measures or by coordination within a health network and its supply chains.

LIMITATIONS

The model was validated at the mentioned accuracy using the testing split, which also targeted the same instance (March 2007); However, the model's utility to predict an individual donating (or not) at a future instance may not consider intangible biases that may not have been accounted for; weather explicitly within the data, or generally about our social tendencies and how they have developed – as demonstrated by the current pandemic. The following questions (among many more) may shed light on what might limit this model's predictions with less supervised raw data:

- Do blood supply levels experience seasonal fluctuations?
- How much does geographical location affect donation tendencies?
- How has the current COVID pandemic affected hygiene standards and policy?
- Have efforts about spreading awareness on blood donations changed with time or have they remained steady?

REFERENCES

Below are links to sources that were referenced through this report:

DATA SOURCE
https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

SciKitLearn PACKAGE: LOGISTIC REGRESSION
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression

SciKitLearn PACKAGE: ROC_AUC_SCORE
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

TPOT PACKAGE: TPOTCLASSIFIER
http://epistasislab.github.io/tpot/using/

RED CROSS BLOOD DONATION ESTIMATES
https://www.umms.org/-/media/files/ummc/community/blood-facts